Estadística y Biometría

Ilustraciones del Uso de InfoStat en Problemas de Agronomía

Autores
Mónica Balzarini
Julio Di Rienzo
Margot Tablada
Laura Gonzalez
Cecilia Bruno
Mariano Córdoba
Walter Robledo
Fernando Casanoves

© by Balzarini Mónica; Di Rienzo Julio; Tablada Margot; Gonzalez, Laura; Bruno Cecilia; Córdoba Mariano; Robledo Walter; Casanoves Fernando.

©Editorial Brujas
1º Edición
Primera Impresión
Impreso en Argentina
ISBN:
Queda hecho el depósito que prevé la ley 11,723

La presente edición corresponde a una versión actualizada de la obra "Introducción a la Bioestadística. Aplicaciones con InfoStat en Agronomía" de Balzarini *et al*. 2011.

Queda prohibida la reproducción total o parcial de este libro en forma idéntica o modificada por cualquier medio mecánico o electrónico, incluyendo fotocopia, grabación o cualquier sistema de almacenamiento y recuperación de información no autorizada por los autores.

Aprendiendo a leer entre números

Organigrama

Prólogo

Este libro tiene un doble propósito: presentar principios y conceptos básicos de la Bioestadística que consideramos necesarios para comprender trabajos de investigación y desarrollo en Agronomía y, por otro lado, ilustrar cómo pueden usarse herramientas estadísticas clásicas para efectuar análisis de datos en problemas de investigación en Ciencias Agropecuarias. Los análisis se realizan con soporte computacional usando el software estadístico InfoStat desarrollado por nosotros en la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba (Di Rienzo et al., 2008). InfoStat permite realizar una amplia gama de análisis estadísticos y la versión estudiantil y su Manual de Usuario (Balzarini et al., 2008) pueden obtenerse gratuitamente (www.InfoStat.com.ar). No obstante, el objetivo de la obra no está focalizado en el "manejo" del software sino en la presentación comentada, más que formal, de conceptos teóricos (que subyacen los procedimientos de análisis de datos) y en la ilustración de estrategias de análisis e interpretación de resultados, con distintas aplicaciones de herramientas bioestadísticas en problemas de la Agronomía usando archivos que se encuentran disponibles en la carpeta de datos de InfoStat.

La obra, se organiza en capítulos en función de núcleos temáticos comunes en los programas introductorios de Estadística en carreras de Agronomía. Los autores de los capítulos son, en mayoría, docentes investigadores de la Cátedra de Estadística y Biometría de la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba. Esperamos que el libro sea de utilidad para quienes se introducen en el mundo del análisis de datos y sus aplicaciones.

Los autores

¿Qué es la Biometría?

La Bioestadística, también conocida como Biometría en algunas áreas, es una rama de la Estadística que se ocupa de problemas planteados dentro de las Ciencias Biológicas como es la Agronomía. Debido a que las cuestiones a investigar, cuando se trabaja con personas, animales, plantas u otros organismos vivos, son de naturaleza muy variada, la Biometría es una disciplina en constante desarrollo. Incluye no sólo herramientas para el análisis estadístico descritivo de datos biológicos sino también el uso de numerosos procedimientos y algoritmos de cálculo y computación para el análisis inferencial, el reconocimiento de patrones en los datos y la construcción de modelos que permiten describir y analizar procesos aleatorios.

Se dice que un fenómeno es de naturaleza aleatoria cuando los resultados del mismo no se pueden predecir con exactitud. Es decir, cuando la respuesta observada puede tener una componente de azar de manera tal que datos colectados para una carcaterística de interés, sobre distintos casos individuales o unidades de análisis, varian.

Por ejemplo, el rendimiento de plantas de olivos para una determinada región y sistema de manejo puede tener un valor esperado de 30 kg/planta; no obstante plantas de un mismo lote, aún siendo de la misma variedad y recibiendo idéntico manejo no rendirán exactamente lo mismo. Una desviación en más o menos 2 kg/planta puede ser común (error aleatorio). El valor de tal desviación en una planta particular es imposible de predecir antes de que se realice su producción, es decir, antes que se coseche. Luego, predecir un volumen de cosecha es un problema de naturaleza aleatoria y por tanto la respuesta deberá ser estadística, deberá contemplar ésta y posiblemente otras componentes de error asoaciadas a la variabilidad propia del fenómeno. Numerosos problemas de importancia agronómica se estudian a través de modelos que incorporan esta componente aleatoria.

La palabra Biometría hace alusión a que el centro de atención está puesto en la medición de aspectos biológicos. El nombre proviene de las palabras griegas "bios" de vida y "metron" de medida. Comprende también el desarrollo y aplicación de métodos y de técnicas de análisis de datos (cuanti y cualitativos) para extraer información desde conjuntos de datos que pueden ser obtenidos desde estudios experimentación u observacionales.

Las herramientas Bioestadísticas son claves en la generación de nuevos conocimientos científicos y tecnológicos. La estrecha relación de la Estadística con el método científico hace de la disciplina una componente de gran valor en proyectos de investigación e innovación en numerosas áreas. En las Ciencias Agropecuarias, el pensamiento estadístico se encuentra presente durante todas las etapas de una investigación; es importante reconocer la naturaleza aleatoria de los fenómenos de interés durante el diseño del estudio, durante el análisis de los datos relevados y, más aún, durante la interpretación de los mismos y la elaboración de conclusiones.

La Estadística nos provee de herramientas no sólo para transformar datos en información sino también para ser buenos consumidores de ésta, saber interpretar lo que escuchamos o leemos y poder decidir criteriosamente sobre la confiabilidad de la información. Resulta fundamental comprender que la naturaleza variable del fenómeno se traduce en un margen de error en la conclusión y que algunas conclusiones son más validas que otras cuando se trabaja con muestras de procesos variables. Así se podrá apreciar la importancia de contar con buenas herramientas estadísticas en los proceso de toma de decisión bajo incertidumbre.

La Estadística se comenzó a desarrollar en las primeras civilizaciones como una Ciencia Social, a partir de la necesidad de mediciones que tenía el Estado para estudiar la población, de ahí deriva su nombre. En esta etapa, la disciplina estaba acotada a realizar cálculos que resumieran los datos recogidos, construir tablas y gráficos con medidas de resumen tales como promedios y porcentajes. Este tipo de Estadística es aún hoy de gran importancia para la sociedad y en la mayoría de los países está a cargo de instituciones oficiales, como es el caso del Instituto Nacional de Estadística y Censos (INDEC) en Argentina. No obstante, la Estadística experimental, que es la que nosotros abordaremos en el libro, es conceptualmente diferente a la Estadística que se usa en Demográfica y Ciencias Sociales. La Estadística como herramienta para acompañar desarrollos científicos fue desarrollada desde diversas motivaciones, principalmente por físicos y astrónomos para concluir a partir de datos que inevitablemente acarreaban errores de medición y por biometristas, formados en las Ciencias Biológicas y en Matemática Aplicada, para explicar la variabilidad debida a diferencias entre individuos, a diferencias entre parcelas de ensayos, entre animales, es decir, entre las unidades biológicas en estudio. Numerosas técnicas estadísticas de fuerte impacto en la generación de conocimiento en Ciencias Biológicas, de la Salud y del Ambiente fueron desarrolladas para investigadores interesados en la observación de la naturaleza y por ensayos de campo, como son los trabajos de Wright, Pearson y Fisher, de claro corte agronómico. Actualmente el análisis de grandes bases de datos biológicos, generados por nuevas biotecnologías, demanda algoritmos informáticos específicos. Así, la Bioestadística se encuentra en su expansión con la Bioinformática. La automatización de procedimientos de capturas de datos como la instalada ya en monitores de rendimiento, en los secuenciadores de ADN, en los sensores de propiedades del suelo y en las imágenes satelitales de áreas de cultivo, generan importantes volúmenes de datos y nuevos desafíos tanto estadísticos como informáticos para su almacenamiento, análisis y uso en tiempo real.

Índice de contenidos

Descriptiva

Capítulo 1

Análisis Exploratorio de datos

Margot Tablada
Mónica Balzarini
Mariano Córdoba

Análisis exploratorio de datos

Motivación

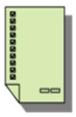
Experimentar la Agronomía desde la búsqueda de información nos permite comprender desarrollos científicos y tecnológicos en su lenguaje. Leer y comunicar artículos sobre Ciencias Agropecuarias involucra saberes relacionados a entender y crear distintos tipos de representación de información. Las herramientas bioestadísticas que conforman el núcleo conceptual denominado Estadística Descriptiva o Análisis Exploratorio de Datos, constituyen preciados instrumentos para organizar, representar y analizar información naturalmente variable como la proveniente de procesos biológicos. A través de medidas de resumen y gráficos conformados por la combinación de puntos, líneas, símbolos, palabras y colores en sistemas coordenados, se muestran de manera sintética las cantidades relevadas en diversos tipos de estudios (poblacionales/muestrales, experimentales/observacionales). Los estadísticos descriptivos bien seleccionados para cada estudio particular representan la vía más simple, y a la vez más potente, de analizar y comunicar información en ciencia y tecnología. El saber usar correctamente herramientas de la Estadística no sólo es útil para la generación de información científica desde proyectos basados en datos, sino también para evaluar resultados de estudios que realizan otras personas y se publican en diversos medios, para detectar estadísticas que consciente o inconscientemente son engañosas y para identificar conjuntos de datos que no resultan buenos para tomar decisiones.

Este Capítulo provee conceptos para comprender medidas resumen y gráficos, principales herramientas del análisis estadístico exploratorio, y enseñar, desde la práctica con software y casos reales, aspectos relevantes a la representación tabular y visual de información estadística. Se presentan los principios para ver y crear gráficos estadísticos simples para una variable, hasta gráficos multivariados útiles para representar casos de estudio sobre los que se han registrados múltiples variables.

Conceptos teóricos y procedimientos

La búsqueda de nueva información generalmente comienza con un proceso de exploración de datos relevados sobre una cantidad previamente determinada de unidades de análisis. Para caracterizar uno o más atributos o variables de interés, es necesario realizar mediciones de esa variable en varias unidades de análisis. Los datos relevados, para cada caso o unidad, se usan para construir una tabla o base de datos que será objeto de exploración o análisis estadístico.

Para llevar adelante un buen análisis cuantitativo sobre un problema, es importante elaborar un protocolo o proyecto. Éste debe incluir suposiciones a priori, definición clara del proceso a estudiar, los objetivos y la finalidad del análisis, las mediciones a ser obtenidas (variables), el origen de las fuentes de datos, la explicitación de fuentes de variación conocidas (factores y covariables), el tipo de diseño del estudio (observacional o experimental), la planificación de la estrategia de análisis estadístico a realizar, el tipo de resultado esperado y, de ser posible, los mecanismos para evaluar su impacto.



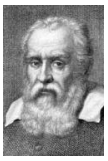
Proyectos sin objetivos claros, claramente no alcanzan sus objetivos.

Los resultados del proceso de análisis estadístico exploratorio de un conjunto de datos, provenientes de mediciones repetidas sobre distintas unidades de análisis, resultan familiares cuando pensamos en los promedios y porcentajes que comúnmente se publican en formato de tablas y gráficos en diversos medios. Estas medidas (denominadas medidas resumen) tratan de describir, de manera resumida, las características más importantes del conjunto de datos.

Los datos son la materia prima de los análisis estadísticos y más aún de los análisis exploratorios o descriptivos. Las características a las que se refieren estos datos se conocen como **variables** ya que pueden asumir distintos valores sobre distintas unidades de estudio.

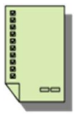
El concepto opuesto al de variable es el de **constante**, una característica que asume siempre el mismo valor para todos los casos o unidades de estudio.

No todas las variables que se relevan son luego analizadas estadísticamente como **variable respuesta** o variable de interés. Algunas variables se relevan simplemente para clasificar a los individuos en grupos (**variables o factores de clasificación**) mientras que otras (**covariables**) se miden porque pueden relacionarse con la variable de interés y por tanto su variación sirve para comprender mejor la variación de la variable de interés.



Lo que no es medible, hazlo medible. Galileo Galilei (1562-1642)

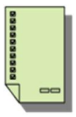
Las variables respuestas pueden ser obtenidas desde unidades de análisis que se encuentran bajo condiciones a las que fueron expuestas intencionalmente (esto sucede en **estudios experimentales**) o bajo condiciones en las que no hubo ningún tipo de intervención por parte del investigador y por tanto se registran u observan los valores de la variable tal cual se dan en la realidad (**estudios observacionales**). En los primeros, el investigador modifica las condiciones y decide bajo qué valores de éstas desea registrar la respuesta. Así es posible estudiar relaciones causales; es decir identificar bajo qué condición o valor de un factor experimental se registran determinadas respuestas. En los estudios experimentales el concepto de **aleatorización** juega un rol importante. Usualmente, el azar (por algún procedimiento de aleatorización) se utiliza para decidir qué unidades de análisis se expondrán bajo cada una de las condiciones de interés (o tratamientos). Así, la aleatorización ayuda a evitar el **confundimiento** de efectos de factores que podrían modificar el valor de la variable de análisis. La importancia de los estudios experimentales aleatorizados y repetidos radica en que, al obtener las respuestas, es posible pensar que éstas se deben a la condición asignada y no a otro factor.



La validez de extender los resultados de un estudio a otros estudios similares depende de la asignación aleatoria de tratamientos a cada unidad de análisis en los estudios experimentales y del azar que hay existido en la toma de muestras, en los estudios observacionales.

En el área de la Agronomía muchos experimentos se llevan a cabo para decidir cuáles prácticas de manejo son más favorables para una determinada producción. Se conducen ensayos a campo, o en laboratorio, en los que se eligen las condiciones en las que se registra una variable de interés; por ejemplo si se desea saber bajo qué condiciones o tratamientos conviene realizar un cultivo, se mide una variable respuesta como el rendimiento. Las condiciones experimentales suelen estar dadas por distintas densidades, fechas de siembras, distintas dosis y/o tipos de fertilizante o distintas frecuencias de riego. Éstos son factores que el investigador decide qué valores asumirán cuando se realiza el estudio experimental, luego aleatoriza la asignación de los mismos a las distintas unidades de análisis y controla que los efectos de un factor no enmascaren los efectos de otro. Por ejemplo para no confundir el efecto del factor fecha de siembra con el efecto del factor variedad, podría decidir sembrar todas las variedades que desea evaluar en una misma fecha de siembra. Por el contrario, en estudios observacionales, no se imponen condiciones sobre el cultivo y se observa lo que ocurre en la realidad sobre cada unidad de análisis. Así, enun estudio observacional,

se podría observar el rendimiento logrado por distintos productores de una zona y la superficie cultivada por cada uno de ellos. Si bien podría detectarse una relación entre ambas variables, es claro que ésta no se puede atribuir como causa del rendimiento logrado a la superficie cultivada, porque los productores pueden estar usando distintas variedades, fechas de siembra, fertilizantes, o mostrar diferencias en otros factores que impactan el rendimiento. El valor de rendimiento relevado en un estudio observacional puede ser consecuencia de factores que no se han medido o no se han controlado y por tanto no se pueden establecer relaciones causales a partir de estudios observacionales.



En ambos tipos de estudios estadísticos (experimentales y observacionales) cada condición de inetrés es observada y valorada repetidamente sobre distintas unidades para poder aplicar técnicas sustentadas en la variación de la respuesta a través de las unidades.

Cada unidad de análisis que forma parte de un estudio, manifestará una respuesta a la condición bajo la que se encuentra y esta respuesta será registrada como un valor de la variable de estudio. Así, la variable asumirá un valor, dentro de sus valores posibles, para cada unidad de análisis.

En las variables de naturaleza **cuantitativa** cada valor será un número que puede ser interpretado como tal, mientras que en variables de naturaleza **cualitativa** el valor será una categoría o cualidad. Si los valores posibles de una variable cuantitativa son números enteros y provienen de un proceso de conteo, la variable se dice de tipo **discreta**. Por ejemplo: cantidad de frutos por planta, número de yemas por estaca, cantidad de insectos por trampa o número de crías por parto. Si los valores que puede asumir una variable cuantitativa corresponden potencialmente a cualquier número real, por supuesto en el rango de variación de la misma, la variable se dice **continua**. Las variables continuas surgen a partir de procesos de medición como pueden ser pesadas o determinaciones de longitudes, tiempos, áreas y volúmenes. Por ejemplo: rendimiento de soja en qq/ha, longitud de espigas de trigo en centímetros, aumento de peso en kilogramos, diámetro de granos de maíz en milímetros, temperatura máxima diaria en grados centígrados, son variables que clasificamos como cuantitativas continuas.

Cuando la variable es cualitativa, los valores posibles son categorías o clases en las que pueden clasificarse las unidades de análisis de manera excluyente; es decir cada unidad pertenece a una y sólo una de las clases o categorías de la variable. Para este tipo de variables, es importante también que las clases sean exhaustivas es decir que cubran todas las clases posibles en las que puede asignarse una unidad de análisis. Por ejemplo, si la variable cualitativa es “máximo nivel de estudio alcanzado por el encargado del establecimiento”, los valores de la variable deberían ser analfabeto o ninguno, primario, secundario, terciario, universitario y posgrado. Si cuando se **operacionaliza** la variable, es decir cuando se decide cuantas categorías tendrá para el estudio de interés, se establecen las categorías primario, secundario y universitario, no se sabrá qué valor asignar a la variable en establecimientos donde el encargado tenga estudios terciarios o de posgrado.

Dos tipos diferentes de variables cualitativas o categorizadas son las variables **nominales** y las **ordinales**. En ambos casos, las categorías representan a diferentes clases como es propio de las variables categorizadas. No obstante, en una variable nominal cada clase representa una cualidad que no tiene ningún sentido ordenar (como mayor o menor) respecto a otra de las clases de la variable. Por ejemplo, en un estudio observacional realizado sobre 30 establecimientos lecheros se podría relevar la variable "estación de concentración de partos" según las categorías: verano, otoño, invierno y primavera. Si bien podríamos usar códigos para relevar la información, asignando un valor numérico a cada categoría (verano=1, otoño=2, invierno=3 y primavera=4), éstos valores no son interpretados estadísticamente como números; sólo podemos decir que un establecimiento al que le fue asignado el valor 1 tiene los partos concentrados en una época distinta al que tuvo un valor de 2, 3 o 4, pero no que 1<2<3<4 con algún sentido de ordinalidad. Ejemplos de variable nominales son: sexo (hembra/ macho), resultados del tacto que se realiza a una vaca (preñada/ vacía), tenencia de la tierra (alquilada/ prestada/ propia/usurpada/ otra), tipo de labranza (convencional/ directa/ reducida). En el caso particular de variables nominales con dos categorías, como los dos primeros ejemplos, también suele usarse el nombre de variables **binarias** o **dicotómicas**.

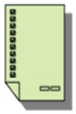
En las variables cualitativas ordinales, las categorías indican un orden de la clasificación y si se usan códigos es posible establecer un orden jerárquico entre los mismos, diciendo por ejemplo 1<2<3 para la variable "severidad de una enfermedad" registrada como leve=1, moderada=2, alta=3; contrariamente 1>2 para la variable "nivel de ataque de insectos en lotes" que asume los valores por encima del umbral económico=1 y por debajo del umbral=2. Para ninguna de las variables cualitativas es estrictamente necesario usar códigos numéricos, pueden usarse directamente los nombres de las categorías como valores de variable ya que en ningún caso los códigos serán usados como números.

En variables ordinales como nominales, las clases o categorías podrían estar representadas por valores numéricos, por ejemplo macho=1 y hembra=2, clorosis baja= 1, clorosis alta=2 y clorosis muy alta=3, pero las diferencias entre tales no reflejarían diferencias aritméticas; esto es, en las variables nominales los valores sólo representan estados mientras que en las ordinales éstas dan cuenta del órden de las categorias. Clorosis 2 representa mayor clorosis que el nivel 1 pero no significa el doble de clorosis que en el nivel 1.

Población y muestra

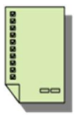
En la obtención de datos hay varios aspectos a considerar por lo que el investigador debe planificar su estudio de manera tal que con los datos que obtenga, y un adecuado análisis, logre información relevante para sus objetivos. Relevante se refiere a aquella

información que permite elaborar conclusiones, que aporten conocimiento, que respondan una pregunta de investigación o que resuelvan un problema de interés. Usualmente las preguntas están referidas a una o más variables de un conjunto de unidades de estudio que se denomina **población**. Para que la pregunta quede mejor definida, la población deberá estar acotada en el tiempo y el espacio.
La proposición anterior pone de manifiesto que, por ejemplo, los rendimientos obtenidos en la última campaña agrícola por todos los productores de maíz de la provincia de Córdoba, conforman una población. A su vez, podemos pensar que podríamos estar interesados en todos los rendimientos obtenidos en la última campaña por todos los productores de maíz del país, vale decir, en una nueva población: la producción de maíz a nivel nacional. En la práctica de la investigación cuantitativa de poblaciones, éstas pueden ser demasiado grandes y por tanto no se pueden obtener todos los datos de la población.



*Las limitaciones para acceder a la población pueden ser de diferente índole. Puede que no se cuente con los recursos necesarios como para obtener datos para todas las unidades de estudio o que éstas sean prácticamente infinitas (**población infinita**).*

En la mayoría de las situaciones de la práctica profesional agronómica, los estudios se llevan a cabo examinando una parte o porción de la población objetivo. Al subconjunto de elementos de la población que será analizado se le llama **muestra**. La cantidad de unidades de estudio en la muestra se denomina **tamaño muestral** y usualmente se simboliza con la letra **n**. Mantengamos presente la idea de que para estudiar fenómenos biológicos aleatorios, detectar diferencias entre grupos de unidades o estudiar relaciones entre variables, será necesario medir más de un individuo o caso de interés, y que la cantidad de casos en la muestra depende de varios factores como lo son la variabilidad de las mediciones, la magnitud de las diferencias que se estudian, o el grado de asociación entre variables. Cuando la variabilidad de los datos es baja, o las diferencias que se esperan encontrar son grandes, o las relaciones muy obvias, el análisis de pocos casos (bajo tamaño muestral) podría ser suficiente para logar una buena conclusión. Por el contrario, cuando se estudian variables que cambian mucho su valor de unidad a unidad, o cuando se desean estudiar diferencias entre grupos, o asociaciones entre variables, que pueden ser muy sutiles, es necesario aumentar el tamaño de la muestra, es decir observar más casos.



La muestra es una parte del todo, es la parte que será analizada unidad por unidad para finalmente inferir o especular el comportamiento de la variable de interés en la población. Por lo tanto, es importante conseguir una buena muestra.

El **diseño del muestreo**, es decir el planificar cómo se tomará una muestra, usualmente se relaciona con preguntas tales como: ¿cuántas unidades conformarán una muestra?,

¿cómo se seleccionarán estas unidades desde la población? Como el objetivo es concluir sobre la población a través de lo observado en una parte de ella, todas estas preguntas persiguen un mismo fin: obtener **muestras representativas** de la población. Esto implica que la muestra seleccionada para llevar a cabo el estudio, nos permitirá conocer acertadamente características de la población de la que ha sido extraída. Muy raramente nos interesa sólo analizar la muestra sin pensar lo que ésta nos dice de la población.

El tamaño de la muestra es una característica a considerar para lograr buena representatividad. Los procedimientos de selección de muestra o de **muestreos basados en el azar** (procedimientos aleatorios) son preferibles a los procedimientos de muestreos **basado en el juicio** del investigador, sobre cuáles elementos considerar en la muestra y cuáles no. Los muestreos aleatorios son **muestreos probabilísticos** ya que es posible conocer la probabilidad que tiene cada muestra de ser seleccionada. En el **muestreo aleatorio simple**, uno de los más utilizados, todas las unidades tienen la misma posibilidad de formar parte de la muestra. Si bien existen fórmulas para calcular los tamaños muestrales necesarios para una situación particular de análisis, fracciones de muestreo de un 10% de la población simple (sin estructura), proveen usualmente de buena cantidad de datos como para estimar lo que sucede en la población. Obviamente, el 10% de una población grande, puede implicar un tamaño muestral inmanejable.

No descuidemos los procedimientos involucrados en la selección de unidades de análisis desde la población para conformar una muestra. Una muestra es como una ventana a través de la cual observamos a la población; la ventana tendrá que tener un tamaño suficiente que nos permita ver bien a la población. El mecanismo más recomendado para mejorar la representatividad de una muestra tomada al azar desde una población es aumentar su tamaño, es decir aumentar el número de casos en análisis y usar una buena técnica de muestreo (basada en procedimientos aleatorios).

Técnicas de muestreo

Hay numerosos métodos de muestreo probabilístico y la elección del mismo depende de características de la población a muestrear. Entre los más usados se encuentran el muestreo aleatorio simple, el muestreo estratificado, el muestreo sistemático y el muestreo por conglomerados.

Muestreo aleatorio simple (MAS)

El muestreo aleatorio simple se lleva a cabo de manera tal que todas las unidades que componen la población tengan igual probabilidad de ser elegidas para conformar una muestra. Este muestreo puede hacerse con o sin reposición.

Sin reposición: Una unidad seleccionada no es devuelta a la población hasta que no se hayan elegido todos los elementos que conformarán esa muestra. Por lo tanto no puede ser nuevamente elegida para formar la muestra.

Con reposición: Una unidad seleccionada es devuelta a la población y por lo tanto puede ser nuevamente elegida para formar una misma muestra.

Las características de un estudio llevan a elegir cómo se obtendrán las muestras. Por ejemplo, en el caso de realizar una encuesta de opinión no se realizará un muestreo con reemplazo.

*Cuando se hace un experimento, por ejemplo medir el contenido de proteínas en fardos de alfalfa, éste se repite **n** veces, bajo las mismas condiciones, y esas repeticiones conforman una muestra.*

Muestreo aleatorio estratificado

En este muestreo se reconoce a priori que la población en estudio se divide en diferentes estratos, o grupos, de unidades de análisis. Los estratos son formados de modo que la variabilidad dentro de un estrato sea menor a la variabilidad entre estratos, para una covariable o factor que puede modificar la respuesta de interés. Por ejemplo, si la variable de interés es la adopción de tecnología, la cual puede ser influenciada por el tipo de productor, primero los productores se estratificarán según su tipo y luego en cada estrato las unidades de análisis se eligen usando un MAS. Este muestreo puede ser más conveniente que el basado en la elección de una muestra aletoria de personas, ya que un estrato podría estar representado en exceso y otro estrato estar ausente en la muestra.

Muestreo por conglomerados

En este muestreo se reconoce a priori que la población está conformada por un conjunto de conglomerados o aglomerados. Los conglomerados son grupos de unidades de análisis heterogéneas de modo que cada conglomerado pueda representar a la población. Es decir la mayor variabilidad se produce entre unidades de un mismo conglomerado y no entre conglomerados. Conformados los conglomerados, se selecciona una muestra aleatoria de los mismos y dentro de cada uno de ellos se observan todas las unidades que lo componen (censo). Por ejemplo supongamos un estudio socio-demográfico donde se quiere estimar la conformación de la pirámide poblacional etaria de una comunidad rural y se tiene un listado de las personas y

familias u hogares en las que viven. Conviene seleccionar una muestra aleatoria de hogares y registrar la edad de sus integrantes, más que seleccionar una muestra de personas individuales, en vez de hogares, para así evitar un exceso de niños o adultos mayores en la muestra.

Muestreo sistemático

En este muestreo se establece una regla para la forma en que se eligen las unidades de análisis. La regla hace referencia a la cantidad de unidades que no serán elegidas pero que se presentan entre dos unidades que serán seleccionadas. El muestreo comienza eligiendo al azar una unidad de análisis y a partir de dicha elección habrá k unidades disponibles que no se seleccionarán. De este modo, las unidades que conforman la muestra son elegidas cada k unidades. El procedimiento suele ser usado para el monitoreo de plagas en un cultivo. Si la unidad de muestreo es un metro lineal de surco un muestreo sistemático de k pasos igual a 80 permitirá, por ejemplo, identificar las unidades de muestreo sobre las que se harán las mediciones. Se comienza desde un punto elegido al azar dentro del lote y cada 80 pasos se registran las observaciones en un metro lineal de surco.

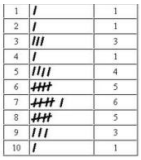
Estadística descriptiva

Generalmente, y sobre todo cuando se cuenta con importante cantidad de datos, es necesario comenzar el análisis estadístico con un proceso de exploración o **minería de datos**. En la etapa exploratoria se utilizan métodos para estudiar la distribución de los valores de cada variable y las posibles relaciones entre variables, cuando existen dos o más variables relevadas. La idea es poder visualizar el comportamiento de las variables a través del uso de tablas, gráficos y medidas de resumen. Éstas son las principales herramientas de la **Estadística Descriptiva** y se aplican casi indistintamente según se tengan los datos de toda la población o de una muestra. Aunque, como se dijera anteriormente, lo más usual en Bioestadística es analizar una muestra ya que la mayoría de las poblaciones de interés son de tamaño prácticamente infinito.
La adecuada obtención y organización de los datos, son el punto de partida de cualquier análisis estadístico. Por eso es importante contar con registros adecuados, datos de calidad o con poco error de medición, y bien sistematizados en bases de datos que se puedan procesar fácilmente.

En el caso del software InfoStat las bases de datos se organizan en tablas de doble entrada, donde usualmente cada fila contiene datos de una unidad de análisis y cada columna corresponde a una variable relevada (variable de clasificación, variable respuesta o covariable). Los valores de cada variable observados en cada unidad se ubican en las celdas de la tabla.

Frecuencias y distribuciones de frecuencias

Las frecuencias asociadas a valores o rango de valores de una variable aleatoria indican la cantidad de veces que un valor de la variable fue observado en el conjunto de unidades en análisis. Las frecuencias sirven para conocer cómo se distribuyen los datos o valores de la variable, permitiendo aproximar la distribución de frecuencias a alguna función o modelo teórico para posteriores análisis y cálculos probabilísticos. Analizando las frecuencias es factible identificar datos extremos (es decir poco frecuentes por ser muy pequeños o muy grandes), y valores, o conjuntos de valores, que aparecen con mayor frecuencia. Las frecuencias en que se presentan los valores de una variable se pueden tabular o graficar.



Es importante tener presente que para aproximar la verdadera distribución de una variable (es decir la distribución en la población), a partir de los datos de una muestra, es necesario contar con una cantidad importante de datos en la muestra.

Tablas de distribuciones de frecuencias

Una **tabla de frecuencias** organiza los datos de manera tal que en una columna de la tabla aparecen los valores de la variable, según el tipo de variable, y en sucesivas columnas se muestran diferentes tipos de frecuencias asociadas a esos valores (frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas y frecuencias relativas acumuladas). Veamos algunos ejemplos de distribuciones de frecuencias y su presentación a través de tablas.

El Cuadro 1.1**¡Error! No se encuentra el origen de la referencia.** muestra la distribución de frecuencias de 50 datos de una **variable cuantitativa discreta "**número de años de agricultura continua en 50 lotes extraídos al azar de una población de lotes en producción agrícola para una región en un año particular". La primera columna (clase) indica cuántos valores diferentes se registraron para la variable (en este ejemplo 11); la segunda columna (**MC** o marca de clase) indica cuáles son los valores que se registraron para la variable, sugiriendo que los lotes muestran de 5 a 15 años de agricultura continua. Las restantes columnas muestran las frecuencias absolutas (**FA**=cantidad de lotes con un valor determinado de años en agricultura continua), las **frecuencias relativas** (**FR=**a cada FA expresada como proporción, es decir referida al total de unidades de análisis), frecuencias absolutas acumuladas (**FAA**) y frecuencias relativas acumuladas (**FRA**) que, para una fila de la tabla, corresponden a la suma de las frecuencias absolutas y relativas de las filas anteriores hasta la fila actual, respectivamente.

Cuadro 1.1. Frecuencias del número de años de agricultura continua/lote

Clase	MC	FA	FR	FAA	FRA
1	5	1	0,02	1	0,02
2	6	1	0,02	2	0,04
3	7	3	0,06	5	0,10
4	8	6	0,12	11	0,22
5	9	4	0,08	15	0,30
6	10	4	0,08	19	0,38
7	11	9	0,18	28	0,56
8	12	8	0,16	36	0,72
9	13	7	0,14	43	0,86
10	14	4	0,08	47	0,94
11	15	3	0,06	50	1,00

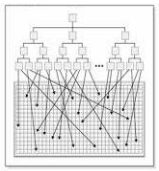
En el Cuadro 1.2 se encuentran las frecuencias para 707 datos de la **variable continua "**pesos de cabezas de ajo blanco".

Cuadro 1.2. Frecuencias de pesos (g) de cabezas de ajo blanco

Clase	LI	LS	MC	FA	FR	FAA	FRA
1	7,70	21,66	14,68	91	0,13	91	0,13
2	21,66	35,63	28,64	228	0,32	319	0,45
3	35,63	49,59	42,61	182	0,26	501	0,71
4	49,59	63,55	56,57	119	0,17	620	0,88
5	63,55	77,51	70,53	66	0,09	686	0,97
6	77,51	91,48	84,49	17	0,02	703	0,99
7	91,48	105,44	98,46	3	4,2E-03	706	1,00
8	105,44	119,40	112,42	1	1,4E-03	707	1,00

A diferencia de una tabla de frecuencias para una variable discreta, los valores registrados para la variable peso (que teóricamente pueden ser muchos y todos distintos por ser continua) han sido agrupados en **intervalos de clase** cuyos límites se indican con LI=límite inferior y LS=límite superior. En cada intervalo de clase se han contabilizado o agrupado, para el cálculo de frecuencias, aquellos datos comprendidos entre los límites de dicho intervalo.

Se puede observar que el límite superior de una clase tiene el mismo valor que el límite inferior de la clase siguiente, sin embargo un dato coincidente con dicho valor será incluido en uno de los dos intervalos según se definan los límites de cada intervalo como cerrados o abiertos; en este ejemplo, los límites superiores son cerrados y los inferiores abiertos, por tanto un valor exactamente igual a un LS será incluido en el primero de los dos intervalos que tengan este valor como límite.



El agrupamiento de los datos continuos es necesario a los fines de conocer la distribución de frecuencias puesto que si no son agrupados es muy probable, por la naturaleza de la variable, que cada valor de la misma aparezca una sola vez en el conjunto de datos y por tanto las frecuencias absolutas serán 1 para la mayoria de los valores.

La determinación de la cantidad y amplitud de los intervalos es generalmente arbitraria pero existe consenso en que deberían usarse entre 5 y 15 intervalos puesto que si no hay suficientes intervalos habrá demasiada concentración de datos y si hay demasiados, puede suceder que algunos no contengan observaciones. Existen expresiones matemáticas recomendables para calcular el número de intervalos que podría resultar más conveniente para un determinado conjunto de datos.

Tanto en el Cuadro 1.1 como en el Cuadro 1.2, la primera columna solo enumera las clases, sin tener significado estadístico. La columna MC o **marca de clase**, para una variable discreta es directamente un valor de la misma, mientras que en una variable continua contiene el valor medio del intervalo de clase. La MC para tablas de variables continuas debe interpretarse como un valor que representa a todos los valores incluidos en cada intervalo de clase. La MC es calculada como la suma de los límites de cada intervalo dividida por 2.

Como puede observarse los nombres de los diferentes tipos de frecuencias son los mismos sin importar el tipo de variable. **FA** es la **frecuencia absoluta** e indica las veces que se registró cada valor de la variable discreta, o la cantidad de datos que hay en cada intervalo de clase de la variable continua. Las FA responden a preguntas del tipo: ¿qué cantidad de unidades de análisis asumieron un valor o valores en un intervalo de clase determinado? (respuesta: 1 unidad, 5 unidades, etc.), ¿qué cantidad de cabezas de ajo tienen un peso aproximado entre 36 g y 49 g? (respuesta: 182 cabezas). La suma de todas las FA debe coincidir con el total de datos, es decir con el tamaño poblacional si se está analizando una población entera o con el tamaño muestral si el estudio se realiza a partir de una muestra.

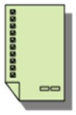
Con **FR** se obtienen las **frecuencias relativas** al total de datos, es decir, el cociente entre la correspondiente FA y el total de datos. Estas FR nos remiten a la idea de proporciones, que multiplicadas por 100 pueden ser interpretadas como porcentajes. Brindan respuestas a preguntas tales como ¿qué proporción o porcentaje de lotes tuvieron 10 años de agricultura continua? (respuesta: 4 lotes en un total de 50 lotes, o 0,08 u 8%), ¿qué proporción o porcentaje representan las cabezas de ajo con pesos entre 36 g y 49 g? (respuesta: 182/707, o 0,26 o 26%). La suma de las FR debe ser igual a 1.

Tanto las FA como las FR, pueden ser acumuladas (**FAA** y **FRA**, respectivamente) permitiendo conocer, por ejemplo, la cantidad de lotes con 10 o menos años de agricultura continua (19 lotes) o con más de 10 años (50 – 19 = 31 lotes), o el porcentaje de cabezas de ajo con peso menor o igual a 91 g (el 99%).

En el caso de variables cualitativas o **categorizadas nominales**, las frecuencias de individuos que pertenecen a cada una de las clases, pueden presentarse en una tabla similar a las anteriores, sólo que para este tipo de variables no se usan frecuencias acumuladas porque la relación de mayor o menor carece de sentido entre sus valores o categorías.

Cuadro 1.3. Frecuencias de las categorías de la variable migración en una zona rural

Sentido de la migración	FA	FR
No migró	33	0,17
Temporal rural-urbana	14	0,07
Definitiva rural-rural	58	0,30
Definitiva rural-urbana	89	0,46
Total	194	1,00



Las variables ordinales usualmente se tratan como las nominales, aunque la frecuencia acumulada podría tener sentido.

El Cuadro 1.3 es una **tabla de contingencia** de una única variable o a un criterio de clasificación. Es común cuando se trabaja con datos categorizados confeccionar tablas de contingencia (o **tablas de clasificación cruzada**) a dos o incluso a tres criterios o vías de clasificación. Una tabla de contingencia con dos criterios de clasificación permite ver simultáneamente dos variables cualitativas. Su distribución conjunta provee información sobre la posible asociación o no de las variables. Para construir la tabla de contingencia se presentan las frecuencias de individuos que son clasificados en grupos definidos por la combinación de una clase de una variable y otra clase de la otra variable. De este modo, si trabajamos con 2 variables, las *r* clases de una de ellas se usan como filas de la tabla y las *c* clases de la otra variable se disponen en las columnas, obteniéndose una tabla de *r*×*c* celdas que contienen las frecuencias de cada combinación.

En el Cuadro 1.4 la tabla de contingencia se construyó con las frecuencias absolutas de cada combinación; también podría haberse realizado con las frecuencias relativas y en ese caso es importante especificar si las frecuencias absolutas se relativizarán con respecto a los totales filas, a los totales columnas o al total de unidades de análisis. Las frecuencias relativas pueden expresarse como proporción, pero es común expresarlas como porcentajes (es decir en base 100).

Cuadro 1.4. Tabla de contingencia asociando tratamiento (vacunado o no vacunado) con estado sanitario en un conjunto de 300 unidades de análisis. Frecuencias absolutas

	Estado sanitario		
Tratamiento	Sanos	Enfermos	Total
No vacunados	29	71	100
Vacunados	144	56	200
Total	173	127	300

El Cuadro 1.4 contiene en las filas a las clases (no vacunados y vacunados) de una variable cualitativa nominal y en las columnas a las clases (sano o enfermo) de otra variable cualitativa nominal. En las celdas aparecen las frecuencias absolutas, o cantidad de unidades de análisis, bajo cada condición.

En el ejemplo, una de las variables (Estado Sanitario) pareciera ser una variable respuesta y la otra (Tratamiento) una variable de clasificación. No obstante, estas tablas pueden construirse con cualquier par de variables cualitativas aún si no existe esta relación de causa-efecto entre ellas. Por ejemplo, si a un conjunto de personas encuestados se les pregunta: 1) si en el fútbol simpatiza con "River", "Boca", "otro equipo" o "con ninguno" y 2) se registra el género: "femenino" o "masculino", interesa la asoaciación entre ambas variables sin necesidad de clasificar una como causa y otro como efecto.

Con el menú Estadística>datos categorizados>tablas de contingencia de InfoStat, se pueden obtener las frecuencias relativas, al total de datos, de cada categoría de cada variable y su intersección como se muestra en el Cuadro 1.5. Las frecuencias también pueden calcularse en relación al total de las filas o al total de las columnas.

Cuadro 1.5. Frecuencias relativas al total de unidades de análisis (animales) según el tipo de tratamiento que recibe y su estado sanitario

	Estado sanitario		
Tratamiento	Sanos	Enfermos	Total
No vacunados	0,10	0,24	0,33
Vacunados	0,48	0,19	0,67
Total	0,58	0,42	1,00

Las tablas de contingencia se usan tanto en estudios experimentales como observacionales. En los primeros es común que los totales filas (suponiendo que en las filas se representan las condiciones experimentales) sean fijados por el investigador y por tanto se suelen usar frecuencias relativas por filas.. En los estudios observacionales, los totales marginales (filas o columnas) usualmente son aleatorios o no fijados por el investigador y todos los tipos de frecuencias tienen sentido de ser calculados.

En nuestro ejemplo sería de interés presentar las **frecuencias relativas por fila**. Esto es, la proporción de animales sanos y la proporción de animales enfermos en relación al total de animales no vacunados (total de la fila 1) y en relación al total de animales vacunados (total de la fila 2). Estas proporciones obtenidas en relación a los totales de las filas se denominan **perfiles filas** y permiten conocer la distribución de las categorías de la variable columna (variable respuesta) en cada categoría de la variable fila (variable de clasificación). Los perfiles filas en los animales no vacunados y en los vacunados, se muestran en el Cuadro 1.6.

Cuadro 1.6. Frecuencias relativas de animales sanos o enfermos según hayan sido o no vacunados

	Estado sanitario		
Tratamiento	Sanos	Enfermos	Total
No vacunados	0,29	0,71	1,00
Vacunados	0,72	0,28	1,00
Total	0,58	0,42	1,00

En el grupo de animales vacunados el porcentaje de animales sanos fue de 72%, mientras que en el grupo no vacunado fue de solo 29%.

El escenario en el que se obtuvieron los datos de la cantidad de animales sanos o enfermos ilustra una situación común en el ámbito de la agronomía. Se cuenta con un grupo de individuos (100 animales) que han recibido un tratamiento (vacunados) y con otro grupo de individuos (200 animales) que no han sido tratados (controles). Cada grupo de individuos se interpreta como una muestra que representa a una población en estudio (en ese ejemplo, las poblaciones en estudio son dos: la población de animales vacunados y la población de animales a los que no se vacuna). El objetivo del estudio es determinar si bajo diferentes tratamientos, se obtienen respuestas diferentes. Dicho de otra manera: ¿se puede decir que esas dos poblaciones no son idénticas?

Las distribuciones de frecuencias y los valores de las variables en estudio no solo pueden presentarse mediante tablas. En numerosas ocasiones se prefiere utilizar gráficos de barras o de sectores para las variables cualitativas o cunatitativas discretas e histograms para las variables cuantitativas continuas. Estos permiten complementar la información tabular.

Gráficos de distribuciones de frecuencias

Las frecuencias de variables **discretas** se grafican utilizando **gráficos de barras**. En el eje X se representan los valores de la variable y en el eje Y, la frecuencia. Cada barra se levanta sobre un punto del eje X que representa a un valor de la variable y la altura de la barra señala la frecuencia para dicho valor.

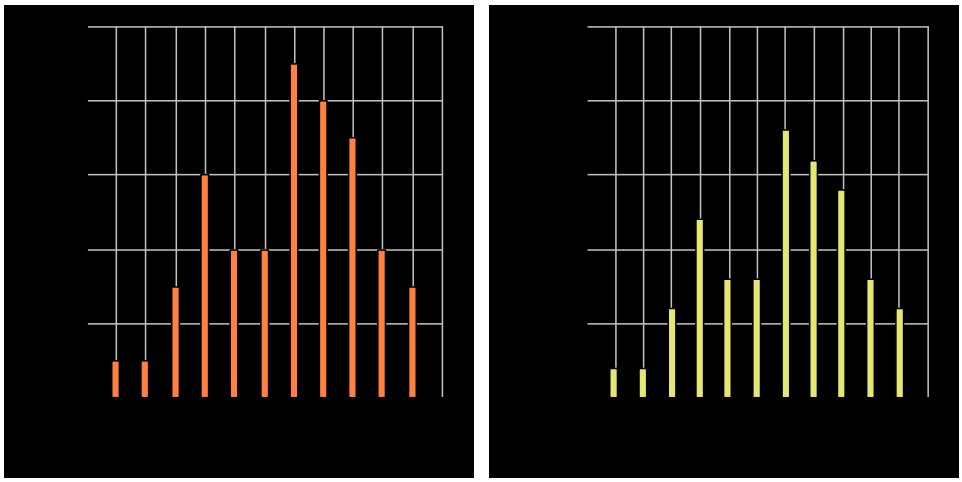


Figura 1.1. Frecuencias absolutas y frecuencias relativas del número de flores por planta

En estos gráficos puede leerse la misma información que observamos en las columnas FA y FR de una tabla de frecuencias. Observemos que la distribución de los datos es la misma en ambos gráficos, solo que se encuentra representada en diferentes escalas.
Otro gráfico que podría utilizarse para observar frecuencias absolutas de una variable es el gráfico de **densidad de puntos o dispersograma.**

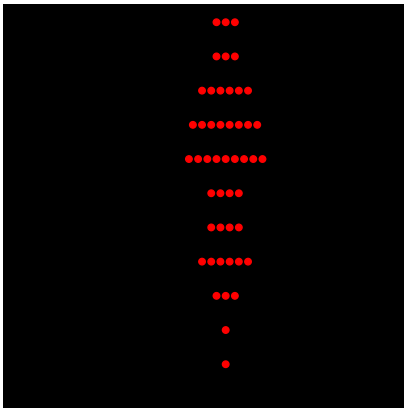


Figura 1.2. Gráfico de densidad de puntos de la variable número de flores por planta.

Las frecuencias de variables **continuas** se grafican más comúnmente utilizando **histogramas** y/o **polígonos de frecuencias**. En el eje X se representan los valores de la variable y en el eje Y, la frecuencia. En un **histograma** se observan "clases" sucesivas. Cada barra se levanta sobre un conjunto de puntos del eje X (una clase o un intervalo de clase). La altura de la "barra" señala la frecuencia relevada para la clase. Las barras se dibujan pegadas, y no separadas como en las variables discretas, para indicar que la variable continua puede asumir cualquiera de los valores comprendidos entre la primera y la última clase.
El **polígono** de frecuencias es una gráfica construida a partir de segmentos de línea que unen las marcas de clase (MC) de los intervalos de clase si se usan FA o FR, o los límites superiores de cada clase en el caso de usarse FAA o FRA. Los polígonos de frecuencias relativas acumuladas también se conocen como **ojivas.** En la Figura 1.3 se muestran histogramas y polígonos de frecuencias para los datos representados en el Cuadro 1.2.

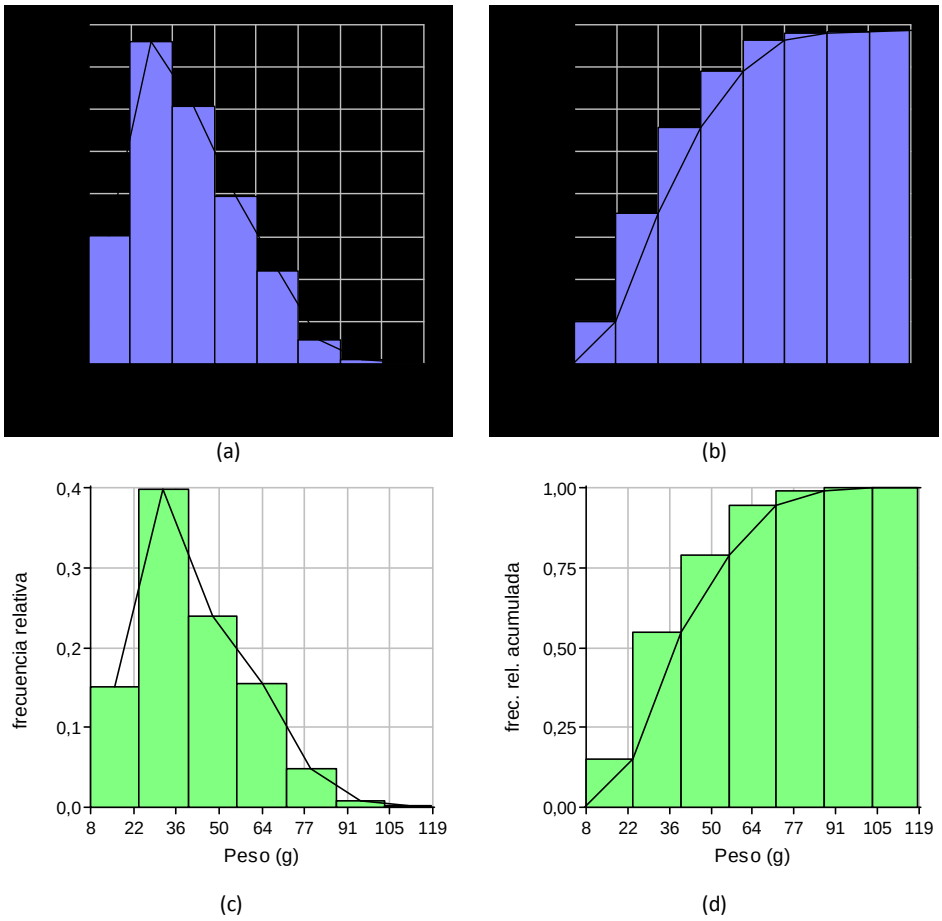
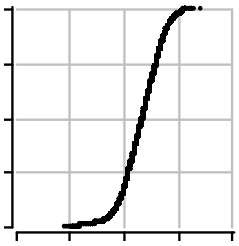


Figura 1.3. Histograma y polígono de frecuencias absolutas (a), frecuencias absolutas acumuladas (b), frecuencias relativas (c) y frecuencias relativas acumuladas (d) de pesos (en g) de cabezas de ajo blanco.



*En InfoStat se pueden obtener las ojivas directamente, es decir sin realizar un histograma previo, seleccionando gráfico de la **distribución empírica**. Para construirlos, el softwaare, ordena los valores de menor a mayor y a cada uno le asigna una FR calculada como el cociente entre su orden o ranking en la lista de datos ordenados y el total de casos. En el eje X se muestran los valores observados de X y en el eje Y la función de distribución empírica evaluada en cada valor de X. Los polígonos de frecuencias acumuladas (ojivas) se usan para leer más directamente la proporción de valores que son menores o iguales a un valor determinado de X.. También dado un valor de porporción se puede saber cual es el valor de la variable (cuantil) para el cual la poporción de valores menores o iguales es igual al valor dado.*

El siguiente gráfico corresponde a la **distribución empírica** (**ojiva**) de los datos de pesos de cabezas de ajo blanco.

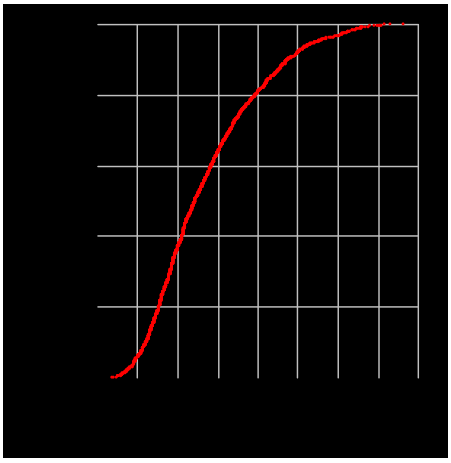


Figura 1.4. Gráfico de distribución empírica de la variable pesos (en g) de cabezas de ajo blanco.

El énfasis en conocer empíricamente (es decir a partir de los datos) la distribución de una variable se relaciona con la necesidad de poder luego aproximar, razonablemente, los valores observado de la variable con modelos matemáticos teóricos que permitirán calcular probabilidades para comprender mejor los fenómenos aleatorios y concluir bajo incertidumbre.

Para representar datos de variables **categorizadas** se pueden utilizar el **gráfico de barras** (presentado para las variables discretas), el **gráfico de sectores** y el **gráfico de barras apiladas**.

Tanto en el caso del gráfico de sectores como en el de barras apiladas, la idea es tomar una figura cuya área representa al total de casos y dentro de tal área ubicar sectores o porciones que permiten visualizar la proporción de casos en cada categoría de la variable. La Figura 1.4 y la Figura 1.5 muestran estos tipos de gráficos para los perfiles filas presentadas en el Cuadro 1.6.

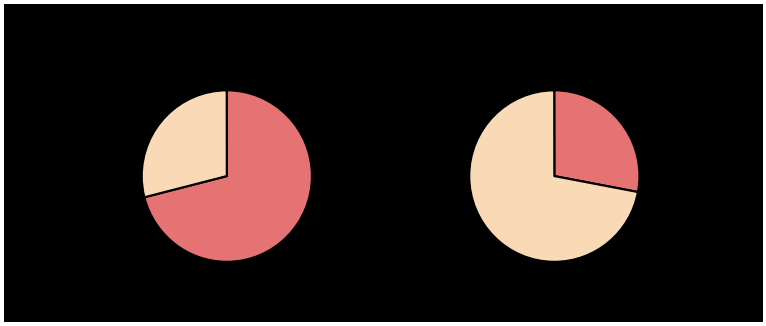


Figura 1.5. Gráfico de sectores para las frecuencias relativas de animales sanos y enfermos según el tratamiento aplicado.

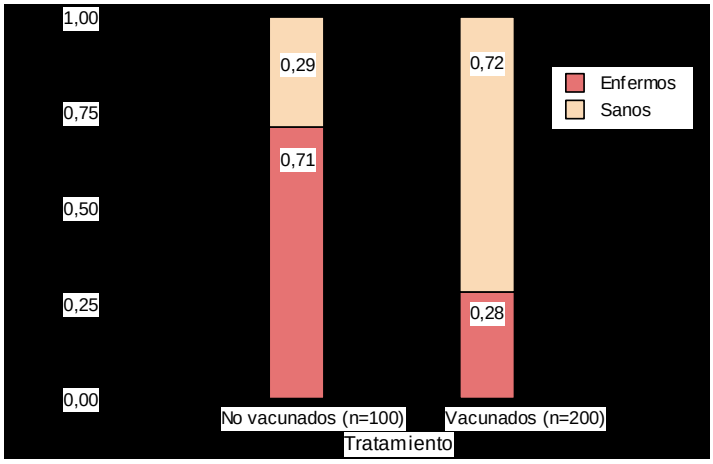


Figura 1.6. Gráfico de barras apiladas para las frecuencias relativas de animales sanos y enfermos según el tratamiento aplicado.

En un gráfico de sectores o barras apiladas resulta oportuno agregar el valor de n, es decir la cantidad de casos que se analizaron para obtener los porcentajes o proporciones que se muestran. Imaginemos un estudio que se realiza por encuesta donde se indaga a cada individuo sobre si consume o no drogas; si el individuo contesta que sí se le pregunta luego, si consume marihuana u otra clase de drogas. Luego de hacer el estudio se registran sobre el total de encuestas, digamos n=100, que 20 consumen drogas y que de ellos 15 consumen marihuana. Un gráfico mostrando que el

75% (15/20) de los individuos se droga con marihuana, sin decir que de 100 fueron 20 los casos de consumo de drogas, podría ser muy engañoso.

En los gráficos hay que ser cuidadoso de no mostrar información engañosa. Para ello, hay que acompañarlos con la mayor cantidad de información sobre su construcción.

Los ejes de un gráfico deben siempre tener nombres (aunque consideremos que es obvia la información que el eje contiene). Las unidades de medida deben estar explicitadas; los mínimos y máximos de los ejes deben ser seleccionadas criteriosamente para no magnificar ni minimizar diferencias y para que el valor inicial y final del eje sea un número entero de rápida lectura. Por ejemplo, aunque igualmente se puedan representar rendimientos en una escala que va desde 8,3 a 28,35 qq/ha, resulta más fácil de visualizar la gráfica si éstos se muestran en un eje cuyo mínimo es 0 y máximo 30 qq/ha. El uso de decimales de más (o de menos) puede dificultar la lectura de la gráfica. La cantidad de "ticks" o marcas sobre cada eje no debe ser demasiada pero tampoco escasa y debe estar asociada a la variación en la serie de valores que se grafica. Los tamaños, los colores y la simbología usada para representar la información deben permitir diferenciar datos que son distintos. Cuando existen más de una serie gráfica es importante incorporar leyendas claras.

Gráficos para dos variables

En la presentación de las tablas de frecuencias para variables categorizadas, se mencionó su uso para el estudio de asociaciones o relaciones y en el ejemplo de los gráficos de sectores o de barras apiladas se observa cómo pueden ser usados para representar las dos dimensiones de las tablas de contingencia.

Cuando el objetivo es estudiar relaciones entre variables cuantitativas, es común utilizar **diagramas de dispersión** para observar la tendencia de la relación (Figura 1.7).

Los gráficos de dispersión muestran los valores de una variable en el eje X y los valores de la otra variable en el eje Y. Si se piensa que los valores de una de las variables dependen de los valores de la otra, se las denomina variable dependiente y variable independiente, respectivamente. En estos casos la variable dependiente o respuesta ocupa el eje Y y la variable independiente se ubica en el eje X. De lo contrario, es indistinto colocar cualquier variable en cualquier eje.

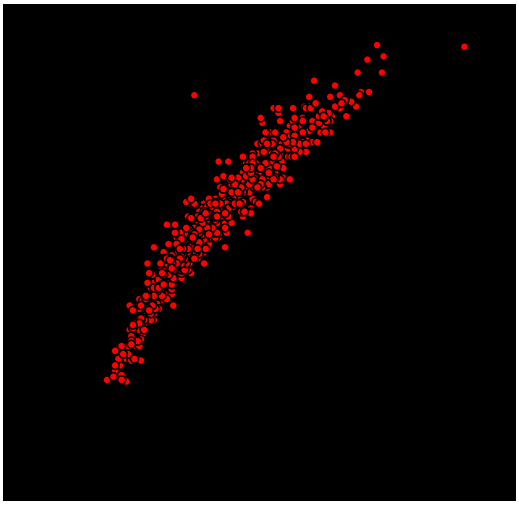


Figura 1.7. Gráfico de dispersión entre perímetro (cm) y peso (g) de cabezas de ajo blanco.

En el caso de representar relaciones entre una variable cuantitativa y otra cualitativa puede utilizarse un gráfico de barras (Figura 1.8).

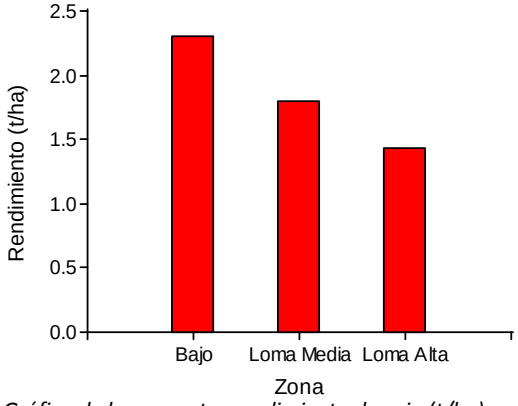


Figura 1.8. Gráfico de barras entre rendimiento de soja (t/ha) y zona productiva de un lote en producción.

Gráficos multivariados

Los gráficos presentados hasta este momento son gráficos uni o bivariados ya que permiten visualizar la distribución de una variable o de dos variables. En el caso de dos variables, puede resultar de interés analizar la distribución conjunta de las dos o la distribución de una de las dos condicionada a niveles fijados para la otra variable, es decir para determinados valores de la segunda variable. En este último caso como en los análisis univariados se dice que la respuesta es unidimensional..

Por el contrario, existen respuestas multidimensionales; éstas se generan cuando sobre una misma unidad de análisis se miden varias variables. Un ejemplo de esta situación se produce cuando se toman muestras de suelo y en cada una se realizan múltiples análisis y por tanto se tienen múlltiples datos (materia orgánica, carbono, nitratos, capacidad de intercambio catiónica, conductividad eléctrica, pH, entre otros). El análisis estadístico multivariado se usa en bases de datos que tienen más de una variable medida para cada unidad de análisis; puede ser que alguna variable sea respuesta y otras explicativas, o bien que todas sean respuestas, es decir tengan la misma "jerarquía".
En esta sección ilustramos el uso de herramientas gráficas que pueden resultar de utilidad en problemas multivariados. Los principios y conceptos teóricos del análisis multivariado no serán discutidos en este libro; ellos pueden ser estudiados en los siguientes libros y materiales: Peña (2002), Johnson & Wichern (2007), Balzarini (2008).

Matriz de diagramas de dispersión: es útil para casos donde se miden más de una variable pero no tantas como para impedir visualizar todas las relaciones de a pares. El siguiente gráfico (Figura 1.9) fue construido con datos del archivo *[Salinidad]* Las variables, sobre un conjunto de 45 macetas fueron biomasa de la planta que crece en cada maceta, pH, zinc, potasio y salinidad del suelo usado como sustrato. Al observar las principales correlaciones, pareciera que la biomasa se correlaciona positivamente con el pH (es decir a medida que aumenta el pH, aumenta la biomasa) y negativamente con el zinc (es decir a medida que aumenta el zinc, disminuye la biomasa).

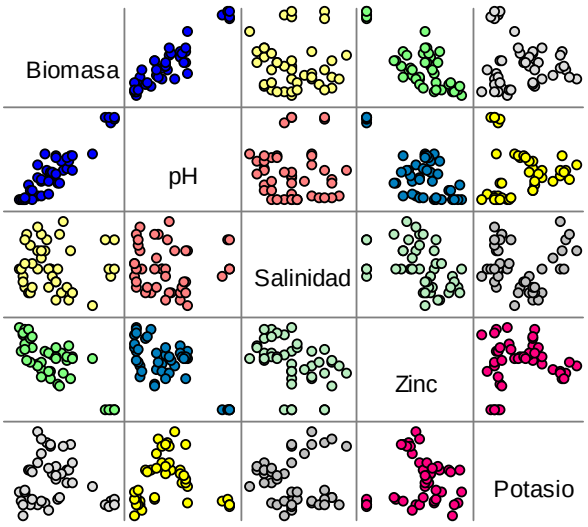


Figura 1.9. Matriz de diagramas de dispersión para las variables biomasa, pH, salinidad, zinc y potasio.

Para elaborar este gráfico en InfoStat en el menú *Gráficos* seleccionamos el submenú *Matriz de diagramas de dispersión (SPlotM)* y dentro de esta ventana seleccionamos las

variables Biomasa, pH, Salinidad, Zinc y Potasio. Accionamos *Aceptar* y aparecerá la ventana *Gráficos* y junto a ella la ventana *Herramientas gráficas*. Esta última ventana muestra un diálogo que permite modificar los atributos del gráfico obtenido.

Gráfico de estrellas: se utiliza para situaciones donde se miden muchas variables y hay pocas unidades de análisis o el interés es representar grupo de unidades. Se construye una estrella para cada unidad o para cada grupo de unidades. Los rayos de las estrellas representan las variables. Las estrellas muestran las variables con mayor valor (rayos más largos) y con menor valor (rayos más cortos) en cada caso. La comparación gráfica de las formas de las estrellas permite visualizar las principales diferencias entre unidades.

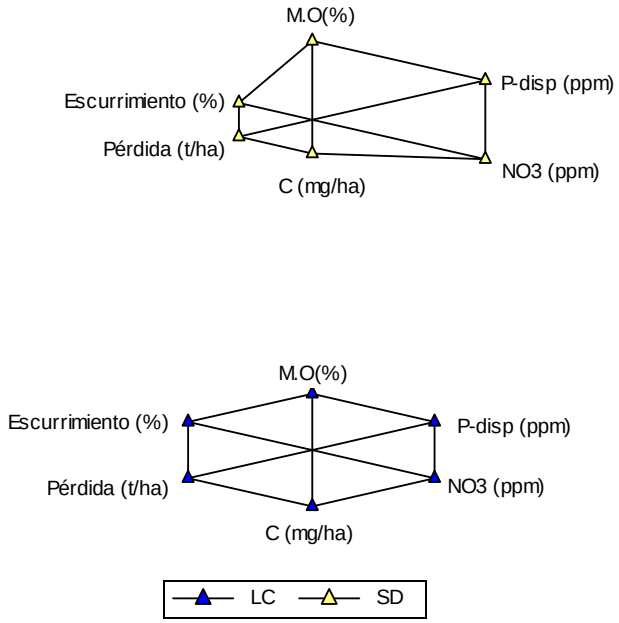


Figura 1.10. Gráfico de estrellas para las variables materia orgánica (MO), C, nitrato (NO3), fósforo disponible (P-dis), escurrimiento y pérdida de suelo evaluados en sistemas de siembra directa (SD) y labranza convencional (LC).

El gráfico de la Figura 1.10 fue construido con datos del archivo *[Estrellas]*. Las variables analizadas fueron medidas durante 10 años y corresponden a los contenidos promedio de materia orgánica (MO), carbono (C), fósforo disponible (P-disp.), nitratos (NO3), pérdida de suelo y escurrimiento de un lote dividido en dos partes, una bajo un sistemas de siembra directa (SD) y la otra utilizando labranza convencional (LC). Se observa que el contenido de MO, C, P-disp, NO3 es más alto en SD, mientras las pérdidas de suelo y escurrimiento son mayores con LC.
Para obtener este gráfico la especificación de las variables en el selector de variables es similar a la realizada con la matriz de diagramas de dispersión.

Biplot del Análisis de Componentes Principales (ACP): se utiliza para situaciones de observaciones multivariadas donde todas las variables son de naturaleza cuantitativa. Se realiza un Análisis de Componentes Principales para combinar las variables en índices y luego se construyen diagramas de dispersión usando estos índices para definir los ejes. Los índices o variables sintéticas se llaman Componentes Principales (CP). Se pueden construir varios índices o combinaciones de variables. No obstante el gráfico más difundido es el basado en las dos primeras componentes principales (CP1 y CP2) porque estas combinaciones son las que explican mejor las diferencias entre unidades de análisis. El grafico se llama Biplot, porque en el mismo espacio (que conforman la CP1 y CP2) se representan las unidades de análisis y las variables, es decir las dos dimensiones de la tabla de datos.

El siguiente gráfico (Figura 1.11) fue construido con datos del archivo *[Proteínas]*. La base de datos contiene datos estadísticos para distintos países europeos referidos al porcentaje de la dieta proteica de sus habitantes, que proviene del consumo de carne de cerdo, carne de vaca, huevos, leches, frutas y vegetales, embutidos, cereales, frutos secos y pescado; vale decir 9 variables.

Los gráficos Biplot siempre se inspeccionan primero sobre el eje CP1 (y luego sobre el CP2). El valor de las CP no es importante como tal, ya que éstas son índices cuya escala depende de la combinación particular de variables que representen. El valor del eje solo es importante para identificar qué observaciones tienen mayores valores positivos y cuáles más negativos. Esto implica que esas unidades de análisis son las más diferentes ("las más opuestas"). Unidades de análisis con valores de CP parecidos, son más parecidas entre sí que unidades con valores más distantes y por tanto más alejados en el plano de representación. El "parecido" implica similitud de todo el perfil de variables, es un parecido en sentido multivariado. Los vectores que representan las variables surgen del centro de la grafica y se puede inferir que: 1) vectores que van para el mismo lado del grafico, es decir con ángulos agudos conformados entre ellos, sugieren variables correlacionadas positivamente; 2) vectores que oponen su sentido, es decir que tienden a formar ángulos llanos, sugieren variables correlacionadas negativamente y 3) vectores que forman ángulos rectos, sugieren variables no correlacionadas. Los vectores de variables que se dirigen hacia valores altos de la CP indican que esa variable asume valores altos en las unidades de análisis que tienen también los valores más altos para la componente. Análogamente se concluye respecto a los vectores que tienen valores bajos de la componente. Luego, el Biplot de componentes principales permite:

1) Analizar variabilidad entre unidades de análisis
2) Analizar correlación entre variables
3) Analizar correlación entre valores de variables y unidades de análisis.

A partir del índice CP1 (que representa un 44,5% de la variabilidad total contenida en la base de datos), se observa que los países (unidades de análisis) Yugoslavia, Albania, Bulgaria, Rumania (parecidos entre ellos en cuanto a las fuentes proteicas usadas) son diferentes de Irlanda, Dinamarca y Alemania O. Estas diferencias se deben principalmente a que los mencionados primeros consumen más cereales y frutos secos,

mientras que Irlanda, Dinamarca y Alemania O., tienen mayores consumos de huevos, leche y carnes. A partir del índice CP2 (que representa un 18,2% de la variabilidad total), se observa que Portugal y España se diferencian del resto de los países; las variables de mejor representación sobre ese eje son el consumo de pescado, frutas y vegetales y embutidos. Consecuentemente, se infiere que en Portugal y España los consumos de proteínas vía estas fuentes alimenticias son mayores que en los otros países.
Usualmente, los gráficos biplot del ACP representan bien la estructura de la tabla de datos cuando la suma de los porcentajes de variabilidad explicados por cada eje es mayor al 60 o 70%.

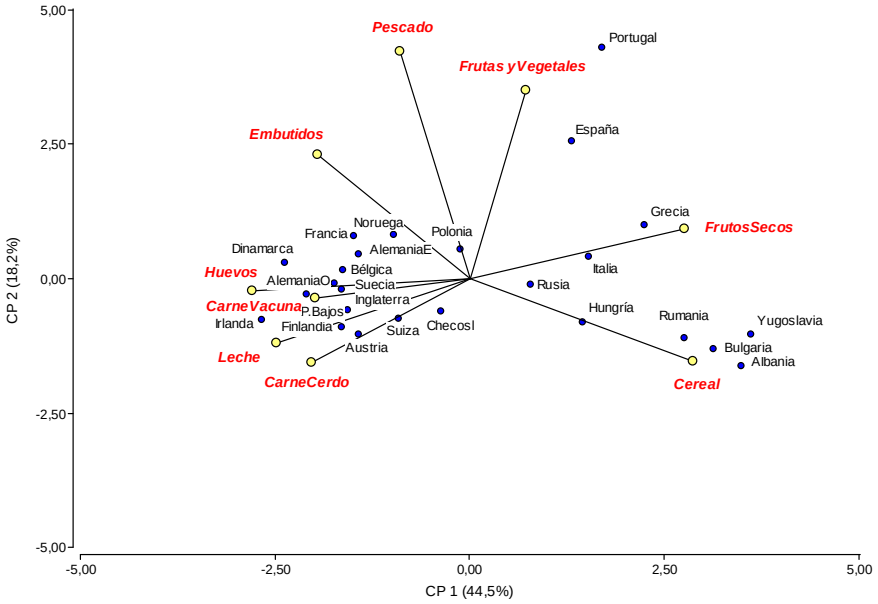


Figura 1.11. Biplot obtenido mediante un análisis de componentes principales usando el consumo de nueve fuentes de proteínas de 25 países de Europa (en la década del 60). Archivo Proteínas.

Para obtener el Biplot en InfoStat, seleccionamos en el menú *Estadísticas* el submenú *Análisis Multivariado* y dentro de este, *Análisis de componentes principales.* En la ventana Análisis de componentes principales seleccionamos las variables CarneVacuna, CarneCerdo y las demás variables que representan la fuente de proteínas, como Variables, y País como Criterio de clasificación. Se dejan activas las opciones que están por defecto y se activa la opción Biplot.

Biplot del Análisis de Correspondencias Múltiples (ACM): se utiliza para situaciones de observaciones multivariadas donde todas las variables son de naturaleza cualitativa. Se realiza un Análisis de Correspondencias Múltiples para estudiar, vía tablas de contingencia, la asociación o correspondencia entre todos los pares de variables. A cada

categoría de cada una de las variables categorizadas se le asigna un peso (o inercia) para cada uno de dos nuevos ejes o variables sintéticas que se usarán para la representar el total de asociaciones. Modalidades con pesos grandes (alejados del cero) y cercanos en un eje, se encuentran asociadas; es decir aparecen juntas con alta frecuencia (en la tabla de contingencia entre las dos variables, la frecuencia para la celda referida a la presentación simultánea de las dos modalidades, es alta o también cuando es baja. Los Biplot de ACM también se leen primero sobre el Eje 1 u horizontal (eje que explica mayor porcentaje de variación) y luego sobre el Eje 2 o vertical.

El siguiente Biplot de ACM (Figura 1.12) se realizó con el archivo *[Autos].* Los datos corresponden a una encuesta realizada en un negocio de ventas de autos en USA, donde se le pregunta a cada cliente cuál es el origen del auto que actualmente tiene (Europeo/Japonés/Americano), cual es su estado civil (soltero/casado/casado con hijo), el tipo de propiedad de la vivienda (dueño/alquila), el tipo de auto (sport/familiar/trabajo), género (hombre/mujer), tamaño del auto (Chico/Grande) y cantidad de ingresos en el hogar (ingreso 1/ingreso 2). La distribución de las modalidades indica que la modalidad soltero (para la variable estado civil) se asoció frecuentemente con las modalidades: alquila, tiene un solo ingreso en la casa, auto chico, sport, de origen japonés y, hombre. Mientras que se opone a este perfil de unidad de análisis (cliente) el de las personas casada-hijo, con auto grande, con dos ingresos en el hogar, que son mujeres y usan autos familiares y de origen americano. Así el grafico permite, de manera muy rápida identificar los principales tipos de cliente que tiene la empresa para orientar mejor sus estrategias de venta.

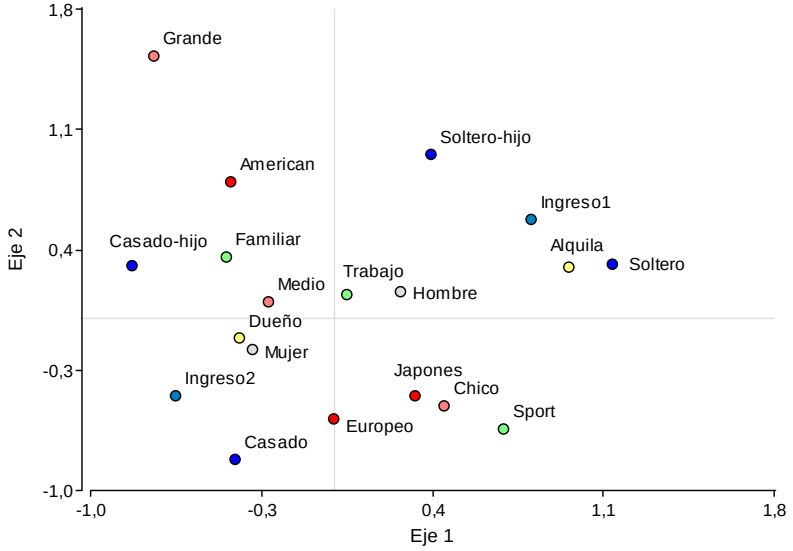


Figura 1.12. Biplot obtenido a partir del análisis de correspondecias múltiples. Archivo Autos.

Para obtener este gráfico en InfoStat, seleccionamos en el menú *Estadísticas* el submenú *Análisis Multivariado* y dentro de éste, *Análisis de correspondencias*. En Criterios de clasificación seleccionamos todas las variables, accionamos *Aceptar* y en la siguiente ventana se dejan las opciones por defecto.

Medidas resumen.

Para resumir la distribución de un conjunto de datos de naturaleza cuantitativa, aparte de gráficoss, se calculan medidas de posición, de variación y de forma de la distribución asociada. La obtención de estas medidas permite complementar y acompañar a la información contenida en una tabla de frecuencias o a la distribución mostrada en un gráfico.

Media, mediana y moda

Tomemos un gráfico de la distribución de la **variable discreta** número de flores por planta, que hemos presentado anteriormente.

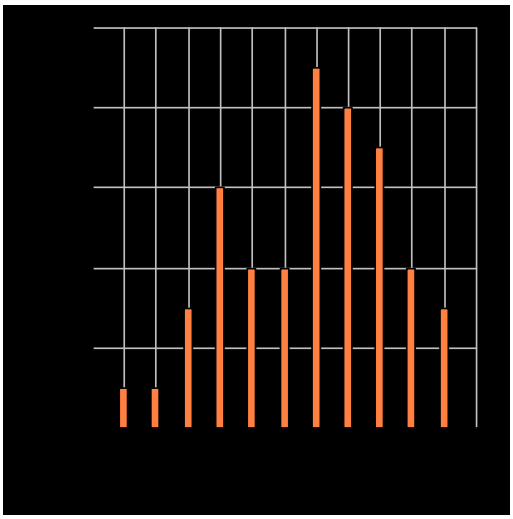


Figura 1.13. Gráfico de barras para la variable número de flores por planta.

La distribución de un conjunto de datos se encuentra situada en un intervalo de valores, ya que en todo conjunto de datos hay un **valor mínimo** y un **valor máximo**. La diferencia entre dichos valores es el **rango** o **recorrido de la distribución**.

- *el valor mínimo observado del número de flores por planta es 0 y el máximo es 10. La distribución tiene un rango de 10.*

Todos los valores de una variable no están igualmente distribuidos dentro del rango de variación; esto es, los valores se presentan con diferentes frecuencias. Al valor que aparece con mayor frecuencia se lo denomina **modo** o **moda**. Una distribución puede tener más de un valor modal.

- *La moda del número de flores por planta, en el ejemplo, es 6.*

Hay valores que se ubican en el centro de la distribución, o cercanos a éste, y otros que se encuentran en los extremos. Aquel valor que ocupa exactamente el centro de la distribución, de modo que la mitad de los datos son valores menores o iguales que éste y la otra mitad son valores que lo superan, se denomina **mediana**.

- *La mediana del número de flores por planta en el ejemplo también es 6.*

El valor que representa al conjunto de datos es el promedio o **media aritmética**. La media es un valor que se ubica en el centro o cercana al centro de una distribución. Se obtiene por el cociente entre la suma de todos los datos y la cantidad total de datos.

- *La media del número de flores por planta es 5,86.*

Si bien el cálculo de la media es 5,86, dado que la variable es discreta, es más apropiado informar que en promedio el número de flores por planta es 6 aproximando al entero más cercano. La mediana es una medida de posición "robusta" (soporta varios valores extremos sin modificar su valor). De hecho, ésta no será afectada hasta que el 50% de los datos se contaminen con valores aberrantes. La mediana es resistente a valores extremos pero la media no. Otro estimador robusto de posición es la media podada, *i.e.* después de descartar de la muestra de datos un porcentaje de las observaciones más grandes y más pequeñas. Específicamente una **media podada**-α es la media muestral después de remover desde los valores más grandes y más chicos de la muestra una porción del $100\times\alpha$% de los datos.

En numerosas ocasiones la media aritmética se compara con el centro de gravedad de un cuerpo. La media sería el punto de equilibrio de una distribución. A diferencia de la mediana, que siempre está en el centro de la distribución, en algunas distribuciones la media no coincide con el centro de los datos porque es afectada por valores extremos que causan su desplazamiento. Esto hace que en algunos conjuntos de datos donde existen valores extremos se prefiera a la mediana, antes que a la media, como resumen de la medida de posición del conjunto de datos.

Notemos que la **moda**, la **mediana** y la **media** son valores de la variable que en la serie ordenada de datos ocupan una posición, por lo cual se les llama **medidas de posición**. A su vez, son valores de **tendencia central**. En cambio el **rango** no ocupa una posición sino que describe la **variación** de los datos, ésta es una medida de dispersión.
En las distribuciones que son **simétricas** unimodales los valores de la moda, la mediana y la media son iguales. Si la media es mayor que la mediana, la distribución es **asimétrica derecha**. Si la media es menor que la mediana la distribución es **asimétrica izquierda** (Figura 1.14). Existen coeficientes que miden la simetría y también otro que piden "la picudez" o kurtosis de la distribución. Ambos son considerados medidas de la forma de la distribución.

(a) (b) (c)

Figura 1.14. Gráfico de una función de densidad con simetría (a), asimetría derecha (b) y asimetría izquierda (c).

- *El número de flores por planta presenta una distribución con leve asimetría a la izquierda*

Veamos ahora la distribución de la **variable continua** peso de las cabezas de ajo blanco.

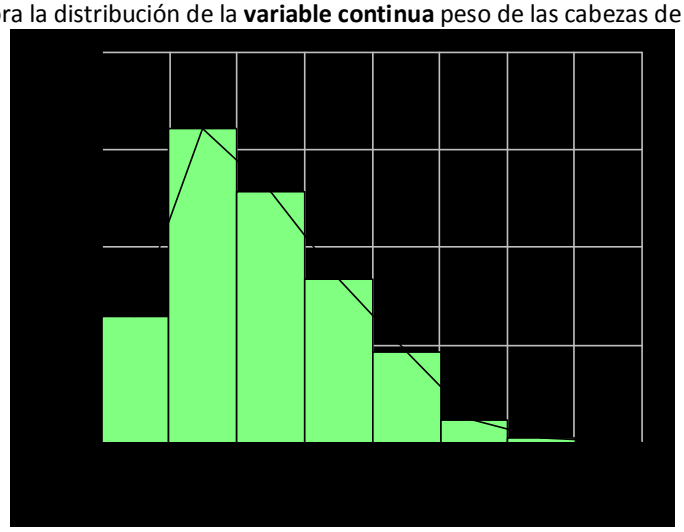


Figura 1.15. Histograma de frecuencias relativas de pesos (en g) de cabezas de ajo blanco

Observemos que en este caso no es tan directo ubicar en el gráfico los valores de las medidas resumen como lo fue para la variable discreta. Esto se debe al agrupamiento de los datos en intervalos de clase.

- *Los valores **mínimo** y **máximo** (7,70 g y 119,40 g, respectivamente), no se leen exactamente en el gráfico debido a que se ha modificado la escala a los fines de lograr una mejor presentación sobre el eje X. Sin embargo la escala utilizada muestra claramente el intervalo de valores de la muestra analizada.*
- *Como los datos son agrupados en intervalos de clase, para reportar la **moda** se hará referencia al intervalo que la contiene. En este caso fueron más frecuentes las cabezas de ajo con pesos entre 22 g y 36 g.*

Para observar la **mediana** es más sencillo trabajar con el polígono de las **FRA**. En el eje Y debe ubicarse el valor 0,50 y se trazará una línea recta, paralela al eje X, hasta llegar al

polígono; luego se leerá en el eje X el valor correspondiente al punto del polígono. Dicho valor de X es la mediana. El procedimiento se muestra a continuación.

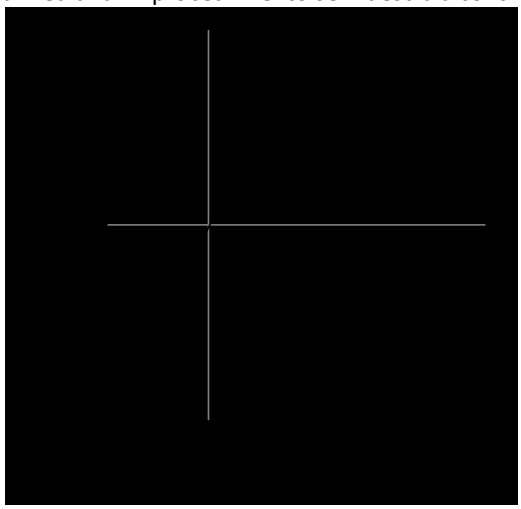


Figura 1.16. Aproximación del valor de la mediana del peso (en g) de cabezas de ajo blanco utilizando el polígono de frecuencias relativas acumuladas

- *La mediana del peso de las cabezas de ajo es 37g.*

El valor calculado de la mediana es 37,6 g. Vemos que a través del método gráfico se logra una buena aproximación. La mediana también puede obtenerse creando una lista de todos los valores en análisis, que muestre a los mismos de menor a mayor y seleccionar el valor posicionado en el medio de la lista (o el promedio de los dos valores posicionados en el medio de la lista si el número de valores listados es par).
El valor de la media (40,77 g) supera al valor de la mediana (37,6 g)

- *La distribución es asimétrica a la derecha*

Cuantiles y percentiles

En la distribución de los valores de una variable, los cuantiles son medidas de posición. Un **cuantil** es un valor de la variable cuya ubicación en la distribución, deja por debajo una proporción del total de los datos. El nombre del cuantil hace referencia a dicha proporción. De otro modo, en la distribución de una variable hay una proporción de valores, en relación al total de datos, menores o iguales a un valor determinado. Por ejemplo, en el caso del peso de las cabezas de ajo vimos que una proporción de 0,50 son valores de peso menores o iguales que 37,6 g; entonces, el valor 37,6 es el cuantil 0,50. Este ejemplo, ilustra que para la proporción 0,50 la palabra cuantil es sinónimo de mediana. No obstante, podemos estar interesados en otros cuantiles, digamos el cuantil 0,05 o el cuantil 0,75, por ejemplo.
Los cuantiles pueden obtenerse, o aproximarse, utilizando el **polígono** de la distribución de **FRA**. Debemos proceder en forma similar a la antes indicada para obtener la

mediana: ubicar en el eje de las **FRA** el valor de la proporción a la que hace referencia el nombre del cuantil, cortar al polígono y luego bajar al eje X, leyendo el valor del cuantil.

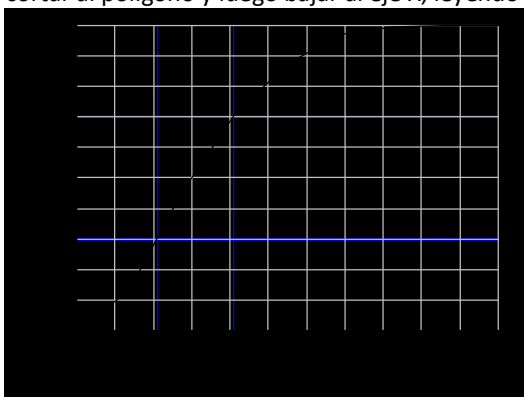
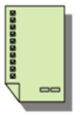


Figura 1.17. Aproximación de los cuantiles 0,30 y 0,70 de la distribución del peso (g) de cabezas de ajo blanco utilizando el polígono de frecuencias relativas acumuladas

En el polígono de FRA de los pesos de las cabezas de ajo (Figura 1.17), se muestra la aproximación para los cuantiles 0,30 y 0,70. El cuantil 0,30 es 29 g y el cuantil 0,70 es 49 g. Estos valores indican que en la muestra de datos, una proporción de 0,30 son cabezas con peso menor o igual a 29 g. De forma similar, una proporción de 0,70 corresponden a cabezas con pesos de hasta 49 g.
El nombre **percentil** se usa si en el eje de las **FRA** la escala se expresa en **porcentaje**. Así, el cuantil 0,30 se corresponde con el percentil 30 y el cuantil 0,70 es sinónimo de percentil 70. Se puede decir que un 30% de cabezas de ajo tienen pesos menores o iguales a 29 g y un 70%, pesan hasta 49 g o que un 30% pesan más que 49 g.



En capítulos posteriores veremos que los cuantiles 0,05 y 0,95 son de amplio uso en la construcción de intervalos de confianza y en el contraste de hipótesis.

Asociados a la obtención de cuantiles, se suelen obtener los llamados **cuartiles**. Estos no son más que los cuantiles 0,25; 0,50 y 0,75 (denotados como Q1, Q2 y Q3, respectivamente). Es decir, se divide la distribución en cuartos y se calcula el primer, segundo y tercer cuartil.

*La diferencia entre el tercer cuartil y el primer cuartil (Q3 – Q1), se denomina **rango intercuartílico** y es una medida robusta de dispersión que no es afectada por valores extremos (los menores al cuantil 0,25 y los mayores al cuantil 0,75).*

Algunos cuantiles pueden ser identificados en el **gráfico de caja o *box-plot*** que representa a una distribución señalando, además de los cuantiles y la presencia de valores extremos o aberrantes, la posición de la media y de la mediana (Figura 1.18).

Medidas resumen

Resumen	peso
n	707,00
Media	40,77
Mín	7,70
Máx	119,40
Mediana	37,60
P(05)	17,20
P(25)	27,90
P(50)	37,60
P(75)	52,50
P(95)	72,60

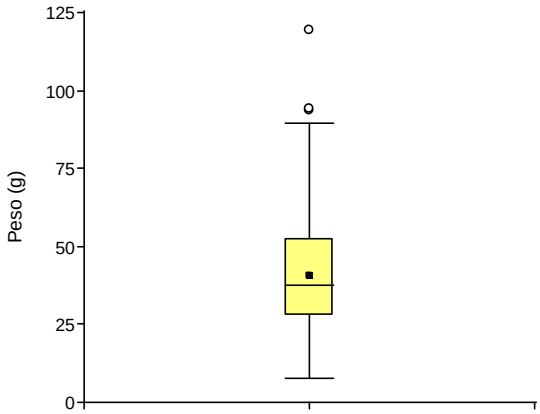


Figura 1.18. Distribución del peso (en g) de cabezas de ajo blanco utilizando un gráfico box-plot. Se acompaña con las medidas resumen que se pueden ubicar en el gráfico

Varianza y desviación estándar

Hemos visto que un conjunto de datos tiene una distribución y que se pueden obtener medidas para caracterizarla. De las medidas presentadas solo el rango nos da idea de la variación de los datos. Sin embargo, estudiar la variación de los datos es uno de los aspectos de fundamental importancia en Bioestadística. Por ello, analizaremos otras medidas que permitan explorar variación. Veamos el siguiente ejemplo.

Los siguientes histogramas (Figura 1.19) muestran distribuciones de rendimientos de trigo obtenidos usando tres diferentes cultivares.

Podemos ver que las distribuciones tienen similares medias, cercanas a los 4000 kg/ha para cada cultivar. Si usamos sólo la media como medida resumen para caracterizar la distribución de valores, concluiríamos que los cultivares muestran iguales rendimientos. Sin embargo, la Distribución 1 presenta mayor dispersión, los datos se concentran más alrededor de la media en la Distribución 2 y la Distribución 3 tiene una dispersión intermedia.

Distribución 1

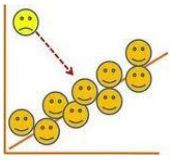
Distribución 2

Distribución 3

Figura 1.19.Distribuciones de rendimientos de tres cultivares de trigo (kg/ha) con diferente dispersión

Expresar la **dispersión** de un conjunto de datos **en relación a su media**, puede realizarse con distintos estadísticos o cálculos: la **varianza** (Var) y el **desvío estándar** (DE) son los más usados. La varianza se obtiene en base al promedio de las distancias o desvíos de los datos respecto de la media. Como la media se encuentra en el centro de una distribución, la suma de estas distancias es nula, siendo necesario calcular el estadístico sumando los cuadrados de los desvíos más que los desvíos puros. Pero esto conlleva a cambiar la magnitud en la que se obtiene la información. Por ejemplo, en las distribuciones anteriores, las varianzas se expresarían en $(kg/ha)^2$, lo cual carece de sentido práctico. Por ello, para expresar la variabilidad en la unidad de medida original se obtiene la raíz cuadrada de la varianza, a la que se denomina **desvío** o **desviación estándar** (medida también conocida como desviación típica)

Los valores de las desviaciones estándares de los rendimientos de los cultivares de trigo en las distribuciones 1, 2 y 3 son 327 kg/ha, 260 kg/ha y 280 kg/ha, respectivamente. Estos valores indican que si bien bajo los diferentes cultivares el comportamiento promedio es casi el mismo, con el 2 se obtienen rendimientos más uniformes; la variabilidad de lote a lote será menor, los rendimientos serán más homogéneos o más parecidos al promedio.



*La desviación estándar es comúnmente utilizada para identificar valores extremos o para establecer valores que se consideran extremos. Datos que se encuentran muy por encima o por debajo de la Media+4*DE o la Media-4*DE son considerados como valores extremos o "outliers", para cualquier tipo de distribución.*

Es común representar valores **medios y desviaciones estándares** mediante gráficos de puntos o gráficos de barras, como se muestra en la Figura 1.20 .
El **gráfico de puntos** muestra que el promedio (puntos) de los rendimientos fue mayor en lotes fertilizados y que, a su vez, se observó menor desvío estándar (líneas por encima y por debajo de los puntos que representan a las medias).
El **gráfico de barras** muestra los promedios de materia seca en floración en parcelas de maíz fertilizadas según la localidad. Las líneas por encima de cada barra representan a los desvíos estándares. El desvío estándar fue mayor en la localidad de Córdoba.

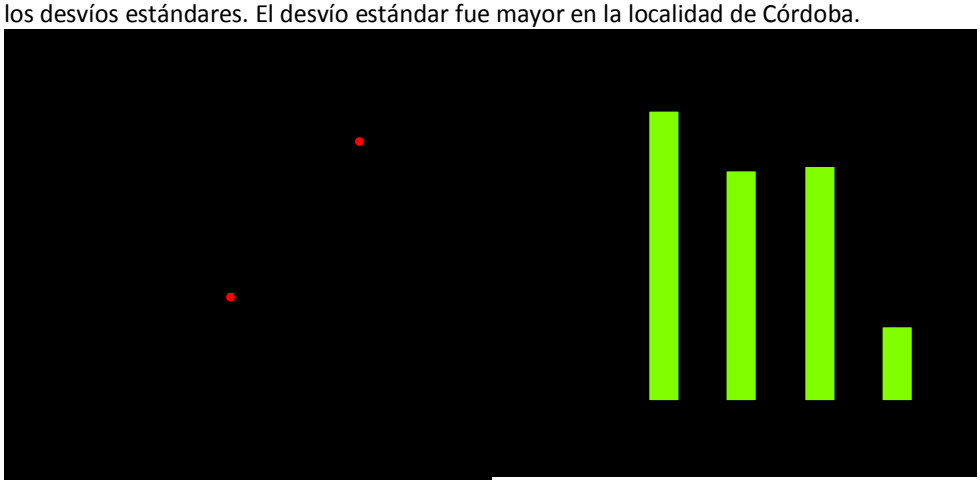


Figura 1.20.Gráfico de puntos de rendimientos promedios de trigo (izquierda) y gráfico de barras de los promedios de materia seca en floración (derecha), obtenidos en maíz bajo diferentes condiciones experimentales. Se muestran las desviaciones estándares.

Coeficiente de variación

Esta es una medida que también permite estudiar la dispersión de los datos. Si bien la desviación estándar es muy útil para comparar la dispersión de dos o más distribuciones, el problema se presenta cuando se desea comparar distribuciones de variables medidas en diferentes magnitudes. Por ejemplo, podemos estar interesados en determinar si el peso de las cabezas de ajo es más variable que el perímetro. El peso expresado en (g) y el perímetro expresado en (cm) no admiten comparación.
El **coeficiente de variación (CV)** es el cociente entre el desvío estándar y la media, por lo que es una medida adimensional de la dispersión relativa a la media. Se suele expresar

en porcentaje. Si un conjunto de datos tienen menor coeficiente de variación, indica comportamiento más homogéneo.

El coeficiente de variación también es útil en el caso de comparar conjuntos de datos de iguales magnitudes pero medidas en diferentes unidades como por ejemplo toneladas y gramos. Siempre que los conjuntos de datos tengan una media muy distinta será necesario elegir el CV como medida de dispersión antes que el DE o la Varianza.

Covarianza y coeficiente de correlación

Para estudiar la variación conjunta de dos variables, digamos X y Y, se puede obtener una medida que considere, simultáneamente, los desvíos de los datos respecto de la media de cada conjunto de datos. En la Figura 1.21 se presentan diferentes tipos de relación entre dos variables.

La **covarianza** entre X e Y es positiva, indicando que los valores de ambas variables crecen simultáneamente. Esto es, a valores mayores de X les corresponden mayores valores de Y. Por el contrario, la relación entre X1 e Y1, es inversa; la covarianza será negativa. Hay que tener en cuenta que el valor de la covarianza depende de las magnitudes de medida. Por lo tanto es necesaria una expresión adimensional.

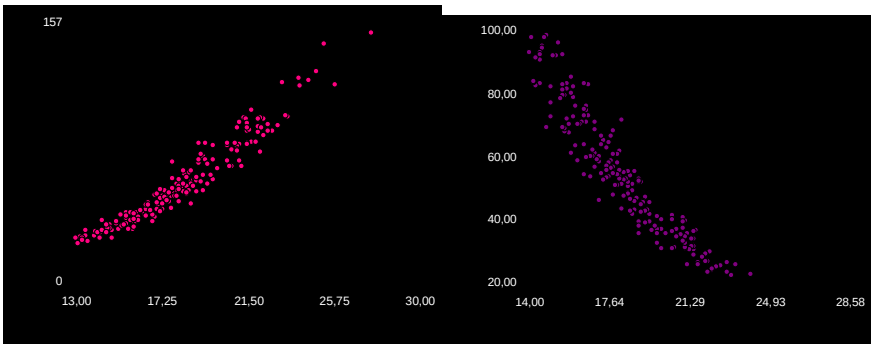


Figura 1.21. Gráficos de dispersión indicando relación directa entre las variables (izquierda) y relación inversa (derecha)

El **coeficiente de correlación lineal** es una medida adimensional que se calcula como el cociente entre la covarianza y el producto de las desviaciones estándar de cada conjunto de datos. El coeficiente toma valores entre -1 y 1. Valores cercanos a -1 indican correlación o covariación inversa. Valores cercanos a 1 indican covariación directa. Valores cercanos a 0 indican falta de covariación.

El coeficiente de correlación indica si las variables se relacionan de forma lineal pero no que existe una relación de causalidad.

Comentarios

En este capítulo hemos presentado conceptos y métodos estadísticos para investigar el comportamiento de diferentes tipos de variables a través del estudio de un conjunto de datos que pueden ser poblacionales o muestrales y provenir de distintos tipos de estudio (experimentales u observacionales). Se pone de manifiesto que el tipo de herramienta estadística a usar es altamente dependiente del tipo de variable que se estudie y de cómo se ha decidido registrar sus valores.
Si bien ahora hemos trabajando con estadística descriptiva, es conveniente resaltar que los estudios que involucran datos, comúnmente, deben transitar por las siguientes etapas:

- *Diseño del estudio incluyendo muestreo y definición de variables*
- *Depuración de bases de datos para el control*
  - Control de tipo de variables
  - Identificación de valores extremos
  - Construcción de nuevas variables
- *Caracterización estadística o análisis exploratorio de datos (Estadística descriptiva)*
- *Inferencia Estadística sobre parámetros (poblacionales) a partir de estadísticos (muestrales)*
  - Estimación de parámetros (esperanza y varianza) y del modelo teórico de distribución de las variables de interés
  - Intervalos de confianza y pruebas de hipótesis sobre los parámetros de una o más distribuciones
  - Exploración de causas de variación
  - Relaciones entre variables respuesta y variables explicativas
  - Relaciones entre variables sin necesidad de especificar causalidad
  - Ajustes de modelos explicativos y finalmente puesta a punto de modelos o herramientas predictivas

Notación

Variables

Letras mayúsculas de imprenta: *X, Y, Z,* etc. Los valores particulares de una variable se indican con letra minúscula y un subíndice que señala el orden de las observaciones: $y_1, y_2, \ldots, y_n$ (primer, segundo y n-ésimo valor de la variable Y, respectivamente).

Medidas resumen

Tamaño muestral: n
Valor mínimo: mín
Valor máximo: máx
Media: $\overline{Y}$
Mediana: me o $Y_{0,50}$
Modo o moda: mo

Varianza (Var): S^2
Desvío estándar (DE): S
Coeficiente de variación: CV
Covarianza entre X y Y: cov(X,Y)
Coeficiente de correlación: r
Percentil *k*: P(*k*) ; Cuantil *p:* Y_p

Definiciones

Definición 1.1: Población

*Una **población** es un conjunto de elementos acotados en un tiempo y en un espacio determinado, con alguna característica común observable o medible.*

Definición 1.2: Tamaño poblacional

*Si la población es finita o contable, diremos que el **tamaño poblacional** es el número de elementos de la misma o número de unidades potenciales de análisis y lo denotaremos con **N**.*

Definición 1.3: Muestra

*Se entiende por **muestra** a todo subconjunto de elementos de la población.*

Definición1.4: Elemento muestral

*Un **elemento muestral** es la entidad de la muestra (unidad de análisis).*

Definición 1.5: Tamaño muestral

***Tamaño muestral** es el número de elementos de la población que conforman la muestra y se denota con **n**.*

Definición 1.6: Variable

*Una **variable** es una característica, propiedad o atributo, con respecto a la cual los elementos de una población difieren de alguna forma.*

Definición 1.7: Frecuencia absoluta

*Se denomina **frecuencia absoluta** al número de veces que el valor de la variable se repite en un conjunto de datos.*

Definición 1.8: Media muestral o promedio

*Si $y_1, y_2, \ldots, y_n$ constituyen una muestra aleatoria de tamaño n, luego la **media muestral o promedio en la muestra** se define como:* $\overline{Y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Definición 1.9: Cuantil muestral

*Si $y_1, y_2, \ldots, y_n$ constituyen una muestra aleatoria de tamaño n entonces el **cuantil p** de su distribución de frecuencias muestral es el valor que en la muestra ordenada en forma ascendente ocupa la posición [p x n] con p tal que $0<p<1$.*

Definición 1.10: Mediana muestral

*Si $y_1, y_2, \ldots, y_n$ constituyen una muestra aleatoria de tamaño n entonces la **mediana** muestral es el cuantil 0,50 de su distribución de frecuencias muestral.*

Definición 1.11: Moda muestral

*Si $y_1, y_2, \ldots, y_n$ conforman una muestra aleatoria, la **moda muestral** es el valor de la variable que ocurre con mayor frecuencia.*

Definición 1.12: Rango muestral

*Dada una muestra aleatoria $y_1, y_2, \ldots, y_n$, el **rango muestral** se define como $r = y^{(n)} - y^{(1)}$, donde $y^{(n)}$ e $y^{(1)}$ corresponden a los valores máximo y mínimo en la muestra respectivamente.*

Definición 1.13: Varianza muestral

*Si $y_1, y_2, \ldots, y_n$ conforman una muestra aleatoria la **varianza muestral** es una función de los desvíos, de cada y_i respecto a la media muestral $\overline{Y}$:*

$$Var(Y) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2.$$

***La desviación estándar muestral** se define como:* $DE = S = \sqrt{S^2}$.

Definición 1.14: Grados de libertad (una aproximación intuitiva)

En una muestra de tamaño ***n****, si calculamos* $Var(Y)$ *,* ***n-1*** *valores de la muestra tienen "libertad" de variar, ya que el último queda determinado por el conocimiento de la media. Por ello, calculada la media se dice que existen* ***n-1*** *grados de libertad.*
Ejemplo: se tiene una muestra de 6 valores que tienen una media de 26, entonces ¿cuál es la mínima cantidad de valores que se requiere para conocer todo el conjunto de valores que dio origen a la media? Respuesta: ***n-1****=5 valores.*
Si n=6 y $\overline{Y} = 26$ *entonces:*

$$\sum_{i=1}^{n} Y_i = 156, \quad \text{ya que:} \quad \sum_{i=1}^{n} Y_i \Big/ n = \overline{Y}$$

Así una vez que se conocen 5 de los 6 valores, el sexto no es necesario ya que puede ser determinado porque conocemos que la suma debe ser 156.

Definición 1.5: Coeficiente de variación muestral

Dada una muestra aleatoria $y_1, y_2, \ldots, y_n$ *con media* $\overline{Y}$ *y desviación estándar S, el* ***coeficiente de variación muestral*** *se define como:* $CV = \frac{S}{\overline{Y}} \cdot 100$.

Definición 1.6: Covarianza

Si $x_1, x_2, \ldots, x_n$ *conforman una muestra aleatoria de una variable X e* $y_1, y_2, \ldots, y_n$ *conforman una muestra aleatoria de una variable Y, la* ***covarianza muestral*** *entre X e Y es una función de los desvíos, de cada* x_i *respecto a la media muestral* $\overline{X}$ *, y de los desvíos de cada* y_i *respecto a la media muestral:*

$$Cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}).$$

Definición 1.7: Coeficiente de correlación muestral

El ***coeficiente de correlación lineal*** *entre las variables aleatorias X e Y es:*

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Aplicación

Análisis exploratorio de datos de agricultura de precisión

La producción de los cultivos varía espacialmente dentro de los lotes como consecuencia de la variación de una diversidad de factores biológicos, edáficos, meteorológicos y de las intervenciones del hombre. Conocer dicha variabilidad permite definir factores limitantes, formas adecuadas para la aplicación de fertilizantes y otros

insumos, y establecer prácticas de manejo y de conservación específicas para cada sitio. Las nuevas tecnologías en maquinarias agrícolas asociadas a la agricultura de precisión proporcionan la oportunidad de medir con mayor nivel de detalle la variabilidad en el rendimiento y en las variables que se correlacionan con éste. El estudio de la variabilidad espacial de las propiedades del suelo y su relación con la distribución espacial del rendimiento de los cultivos dentro del lote, es clave para realizar manejos sitios-específicos. Indicando los patrones espaciales de productividad de los cultivos, se identifican los sitios o subregiones donde los insumos agrícolas son necesarios, mejorando de esta forma la eficiencia en el uso de los mismos, la protección del medio ambiente por el uso adecuado de los agroquímicos y potenciando el rendimiento del cultivos con una producción sustentable en el tiempo.

El archivo *[CE]* contiene datos de mediciones georreferenciadas de conductividad eléctrica aparente (CEa, en mS/m), altimetría (m) y rendimiento de soja (Rto_Sj) y trigo (Rto_Tg) (t/ha) de un lote ubicado al sudeste bonaerense de la República Argentina (Gentileza: Ing. Agr. José L. Costa y N. Peralta, INTA-Balcarce). La medición georreferenciada es una medición donde no sólo se toma el dato de la variable de interés sino que también se mide con algún dispositivo la latitud y la longitud del punto del cual se extrae el dato.

La CEa es una herramienta tecnológica de la agricultura de precisión que permite investigar las propiedades físico-químicas del suelo (*i.e.* humedad del suelo, capacidad de intercambio catiónico, materia orgánica, textura y contenido de sales) que influyen en los patrones de rendimiento de los cultivos. La altimetría es otra propiedad importante que afecta directamente el crecimiento y desarrollo de los cultivos por la acumulación de agua en diferentes partes del terreno, e indirectamente por la erosión y deposición del suelo. Los monitores de rendimiento permiten obtener datos georreferenciados de producción de un lote, con los que se elaboran los mapas de rendimiento. Todas estas herramientas generan grandes cantidades de datos que son analizados teniendo como objetivo de estudio la variación espacial de las variables para delimitar en el lote zonas homogéneas.

Estrategia de análisis

Supondremos que el objetivo de análisis es estudiar el lote del cual se tomaron los datos y por tanto nuestra población objeto de estudio está conformada por todos lo pixeles o puntos que conforman el área del lote. Si bien se dispone de un conjunto de muchos datos porque se han relevado con instrumentos de agricultura de precisión, estos conforman una muestra (n=7577) porque no corresponden a todos los sitios que conforman el lote. El tamaño muestral es grande por lo que estaremos en muy buenas condiciones para realizar análisis estadísticos. En una primera etapa del estudio, etapa exploratoria o descriptiva, resumiremos la información a través de distintas medidas resúmenes y gráficos.

Se obtendrán medidas resumen acorde a la naturaleza cuantitativa de las variables y se realizarán histogramas y box-plot, así como gráficos de la distribución empírica de cada variable, para comprender mejor la variabilidad de las mediciones. En una etapa más

tardía de la investigación seguramente los ingenieros estudiarán la distribución espacial de estos datos dentro del lote y construirán mapas que permitirán definir áreas homogéneas. En la etapa exploratoria, debido a que medimos varias variables cuantitativas, haremos un biplot producto de un Análisis de Componentes Principales para estudiar correlaciones entre variables. También graficaremos en una matriz de diagramas de dispersión, todos los diagramas de dispersión necesarios para estudiar la posible correlación entre pares de variables.

Resultados

Medidas Resumen: para obtener las medidas resumen de los datos del archivo *[CE]* se utiliza el software estadístico InfoStat. Eligiendo el Menú *Estadísticas* y seleccionando el submenú *Medidas resumen*, se abre la ventana *Medidas resumen* y se eligen las variables que se desea analizar (CEa 30, CEa 90, altimetría, Rto_Sj y Rto_Tg). Para continuar, se acciona el botón *Aceptar* y activaremos las siguientes medidas: número de observaciones (n), Media, desviación estándar (D.E), coeficiente de variación (CV), valor mínimo (Mín), valor máximo (Máx), Mediana, cuantil 0,25 o primer cuartil (Q1) y cuantil 0,75 o tercer cuartil (Q3). Dejamos la presentación de los resultados por defecto en forma horizontal. Accionamos el botón *Aceptar* y se obtiene la salida que se muestra en el siguiente cuadro.

Cuadro 1.7. Salida de InfoStat. Medidas Resumen para los datos del archivo CE

Variable	n	Media	D.E.	CV	Mín	Máx	Mediana	Q1	Q3
CEa 30	7577	30,01	8,22	27,38	14,80	61,80	29,50	23,40	35,30
CEa 90	7577	29,88	6,93	23,19	12,40	56,90	29,70	25,50	34,00
Altimetría	7577	141,68	1,82	1,28	134,56	147,05	141,74	140,43	143,00
Rto_Sj	7576	1,85	0,39	21,31	1,04	2,98	1,80	1,55	2,11
Rto_Tg	7576	3,72	0,64	17,08	1,91	5,68	3,65	3,26	4,14

A partir de las medidas resumen, se puede observar que la CEa no cambia mucho entre los 30 y 90 cm de profundidad; que la altimetría es una variable con poca variación relativa como pone en evidencia el bajo CV; que la variable rendimiento de soja, a pesar de tener un menor desvío estándar que la variable rendimiento de trigo muestra mayor variación relativa, pudiendo concluir que los rendimientos de trigo son levemente más uniformes entre sitio y sito del lote, que los de soja. Para todas las variables medidas, la similitud encontrada entre media y mediana sugiere que las distribuciones de frecuencias podrían considerarse como simétricas. Si bien se observaron rendimientos de trigo entre 1,91 t/ha y 5,68, la mayoría de éstos (el 75%) se encontró entre 3,26 y 4,14 t/ha, con un 25% de los valores de rendimiento menores a 3,26 (Q1 o P(25)) y un 25% mayores a 4,14 (Q3 o P(75)).

Tablas de Frecuencias: otra forma alternativa de presentar estos resultados es mediante las tablas de frecuencias y los histogramas. Para ello en el menú *Estadísticas* seleccionamos el submenú *Tabla de frecuencias* y elegimos las variables analizadas anteriormente. Accionamos el botón *Aceptar* y en la siguiente ventana los campos activados por defecto son los límites inferiores (LI) y superiores (LS) de los intervalos de clase, marca de clase (MC), frecuencias absolutas (FA) y frecuencias relativas (FR). Para este ejemplo activamos también frecuencias absolutas acumuladas (FAA) y frecuencias relativas acumuladas (FRA). Modificamos el número de clases en 10 y el resto de las opciones mostradas en la ventana se dejan por defecto. Accionamos *Aceptar* y obtenemos como salidas las tablas de frecuencias para cada variable. Aquí se muestran solo las tablas de frecuencias para las variables CEa 30 y Rto_Sj (Cuadro 1.8 y Cuadro 1.9).

Cuadro 1.8. Salida de InfoStat. Tablas de Frecuencias para la variable rendimiento de soja (Rto_Sj) del archivo CE

Variable	Clase	LI	LS	MC	FA	FR	FAA	FRA
Rto_Sj	1	1,044	1,238	1,141	273	0,036	273	0,036
Rto_Sj	2	1,238	1,432	1,335	883	0,117	1156	0,153
Rto_Sj	3	1,432	1,626	1,529	1324	0,175	2480	0,327
Rto_Sj	4	1,626	1,820	1,723	1428	0,188	3908	0,516
Rto_Sj	5	1,820	2,014	1,917	1238	0,163	5146	0,679
Rto_Sj	6	2,014	2,208	2,111	966	0,128	6112	0,807
Rto_Sj	7	2,208	2,402	2,305	662	0,087	6774	0,894
Rto_Sj	8	2,402	2,596	2,499	472	0,062	7246	0,956
Rto_Sj	9	2,596	2,790	2,693	240	0,032	7486	0,988
Rto_Sj	10	2,790	2,984	2,887	90	0,012	7576	1,000

Cuadro 1.9. Salida de InfoStat. Tablas de Frecuencias para la variable conductividad eléctrica aparente (CEa) del archivo CE

Variable	Clase	LI	LS	MC	FA	FR	FAA	FRA
CEa 30	1	14,800	19,500	17,150	700	0,092	700	0,092
CEa 30	2	19,500	24,200	21,850	1419	0,187	2119	0,280
CEa 30	3	24,200	28,900	26,550	1466	0,193	3585	0,473
CEa 30	4	28,900	33,600	31,250	1588	0,210	5173	0,683
CEa 30	5	33,600	38,300	35,950	1241	0,164	6414	0,847
CEa 30	6	38,300	43,000	40,650	676	0,089	7090	0,936
CEa 30	7	43,000	47,700	45,350	282	0,037	7372	0,973
CEa 30	8	47,700	52,400	50,050	119	0,016	7491	0,989
CEa 30	9	52,400	57,100	54,750	58	0,008	7549	0,996
CEa 30	10	57,100	61,800	59,450	28	0,004	7577	1,000

La tabla de distribución de frecuencias de la variable Rto_Sj sugiere que el 51,6% de los datos son menores a 1,82 t/ha. La marca de clase de este intervalo, 1,723 t/ha, es un

valor que aproxima la tendencia central de la distribución. Este valor puede ser bien aproximado desde el gráfico de la distribución empírica. También podríamos decir que solo en 90 sitios, es decir menos de un 2% de los datos, se registraron rendimientos entre 2,79 y 2,98 t/ha mientras que un alto porcentaje de sitios tienen rendimientos de soja entre 1,432 y 2,014 t/ha.
Para la variable CEa 30 un 47,3% de los datos son menores a 28,9 mS/m. Valores de CEa 30 entre 19,5 y 38,3 mS/m son más frecuentes mientras que valores menores a 19,5 mS/m o mayores a 38,3 mS/m son menos frecuentes de encontrar dentro del lote. El número total de observaciones es de n=7577.

Histogramas: para construir los histogramas de frecuencias en el menú *Gráficos* seleccionamos el submenú *Histogramas* y dentro de esta ventana seleccionamos las variables CEa 30, CEa 90, Altimetría, Rto_Tg y Rto_Sj. Accionamos *Aceptar* y aparecerá la ventana *Gráficos* y junto a ella la ventana *Herramientas gráficas.* Esta última ventana muestra un diálogo que permite modificar los atributos del histograma obtenido. En la solapa *Series* de la ventana *Herramientas gráficas*, hay un menú de opciones de histograma que permite cambiar el número de clases (Clases), realizar ajustes (Ajuste) a una distribución determinada, ingresar los límites inferior (LIPC) y superior (LSUC) para la primera y última clase respectivamente y elegir la frecuencia representada en el histograma (Frec.), entre otras opciones. En este ejemplo activamos la opción polígono, desactivamos la opción marcas de clase (M. clases), elegimos 10 clases y modificamos la frecuencia a representar (frecuencia relativa). Las interpretaciones de los histogramas son similares a las de tablas de frecuencias.

Gráficos de cajas (box-plot): este gráfico permite también visualizar la forma de la distribución de frecuencias de cada variable analizada. En un mismo elemento gráfico se representa la información acerca de la mediana, la media, los cuantiles 0,25, 0,75 y la presencia, si los hubiere, de valores extremos. El "bigote inferior" indica el menor valor observado que es mayor o igual a la diferencia Q1-1,5 RI, donde RI es el recorrido intercuartílico. Dicho valor observado coincide con el mínimo si no hay valores atípicos o extremos. El "bigote superior" coincide con el mayor valor observado que es menor o igual que Q3+1,5RI (coincide con el máximo si no hay valores atípicos o extremos). Los **valores atípicos** inferiores están entre Q1-15RI y Q1-3RI y los superiores entre Q3 + 1,5RI y Q3 + 3RI. Los **valores extremos** aparecen por debajo de Q1-3RI y por encima de Q3 + 3RI.
La especificación de las variables en el selector de variables de este tipo de gráfico es idéntica a la realizada con los histogramas. Para este ejemplo hacemos un gráfico de cajas para cada variable, aunque es posible incluir en un mismo gráfico varias variables y será necesario, entonces, agregar los correspondientes ejes Y para mostrar cada variable en la escala apropiada.
A continuación se presentan los histogramas de frecuencias relativas y gráficos de cajas (box-plot), observe el grado de asimetría que se visualiza con ambos gráficos en las variables CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg (Figura 1.22).

Gráfico de distribución empírica: este gráfico presenta los valores observados de la variable en el eje X y la función de distribución empírica evaluada en cada uno de los puntos observados, en el eje Y.

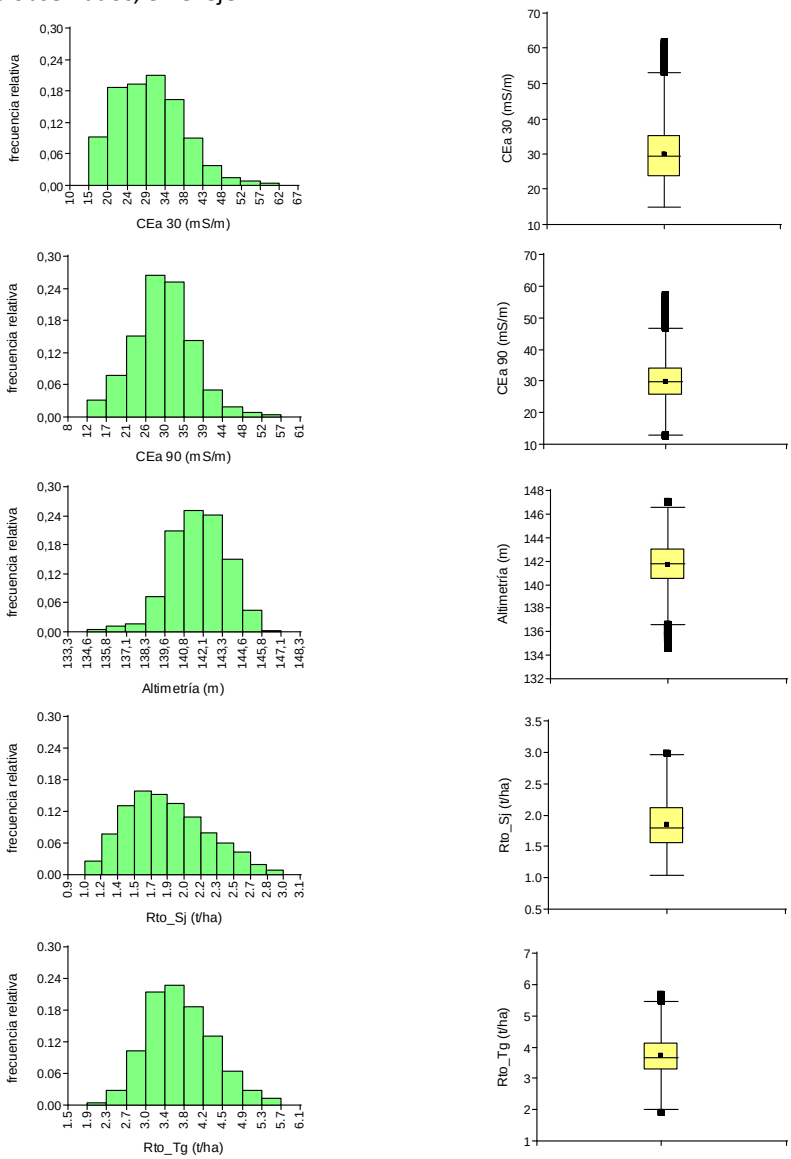


Figura 1.22. Histograma de frecuencias relativas (izquierda) y gráfico de cajas (derecha) para las variables CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg. Archivo CE.

El procedimiento para confeccionar este gráfico es similar al de los anteriores gráficos: menú *Gráficos*, submenú *Gráficos de la distribución empírica* y dentro de esta ventana seleccionamos las variables a graficar (CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg). Accionamos *Aceptar* y aparecerá la ventana *Gráficos* y junto a ella la ventana *Herramientas gráficas*, en la ventana *Gráficos* activamos Mostrar-Ocultar grilla. A continuación se presentan cuatro gráficos de la función de distribución empírica; las variables CEa 30 y CEa 90 se grafican en forma conjunta. Los gráficos de la función de distribución empírica no evidencian en ningún caso, una fuerte anomalía, con respecto a una curva sigmoidea perfecta, que como veremos más adelante corresponde a la función de distribución normal (Figura 1.23).

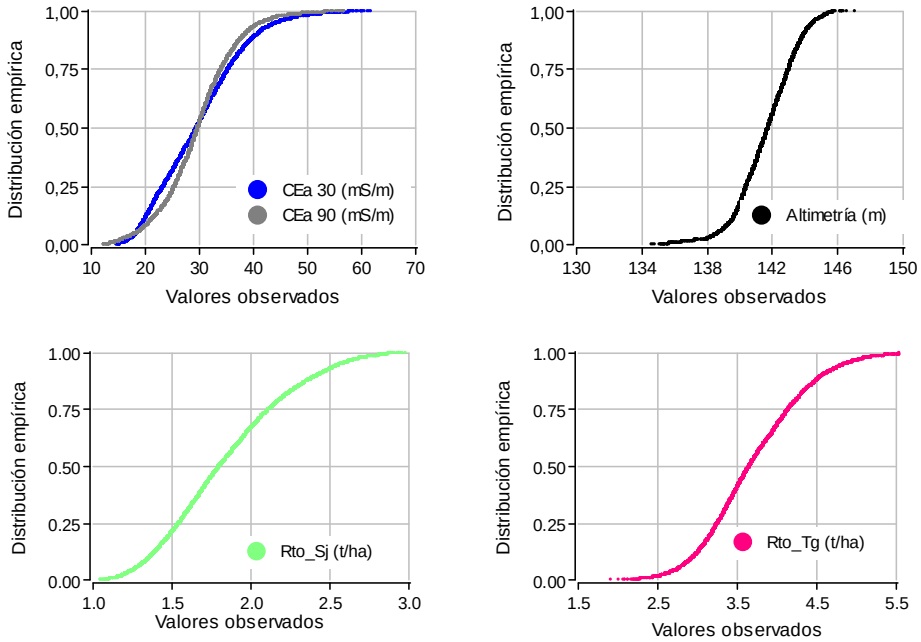


Figura 1.23. Gráficos de la distribución empírica para las variables CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg. Archivo CE.

Matriz de diagramas de dispersión: permite visualizar en un mismo gráfico las relaciones entre un conjunto de variables. La Figura 1.24 muestra esta forma de representación de las relaciones entre las variables CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg. Al observar las correlaciones, pareciera que la CEa 30 se correlaciona negativamente con el Rto_Sj y Rto_Tg y positivamente con la CEa90.

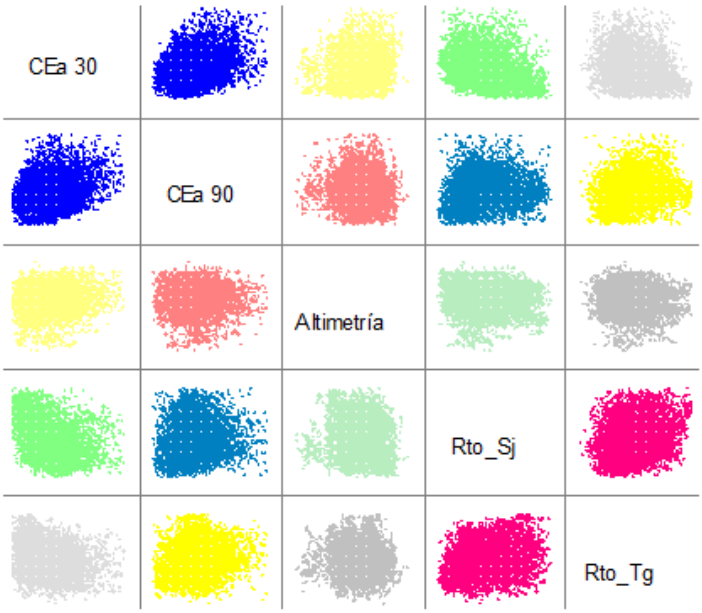


Figura 1.24. Matriz de diagramas de dispersión para las variables CEa 30, CEa 90, Altimetría, Rto_Sj y Rto_Tg. Archivo CE.

Biplot del Análisis de Componentes Principales (ACP): como puede observarse en el biplot (Figura 1.25) la primera componente (CP1) separa dos grupos de variables, uno representado por la CEa 30 y otro por el Rto_Sj y Rto_Tg, por lo tanto, la mayor variabilidad entre datos se explica con estas variables. Con los dos ejes se explicó el 57% de la variabilidad total en las observaciones. La variable Rto_Sj recibe el peso negativo más alto y la variable CEa 30 el peso positivo más alto. Luego se puede interpretar que la CP1 opondrá sitios del lote que tendrán alta medición de CEa 30 a aquellos que tendrán altos rendimientos de soja y trigo. En este ejemplo se podría destacar la variabilidad introducida por la variable CEa 90 analizando la CP2. La CP2 provee nueva información sobre variabilidad respecto a la provista por la CP1.

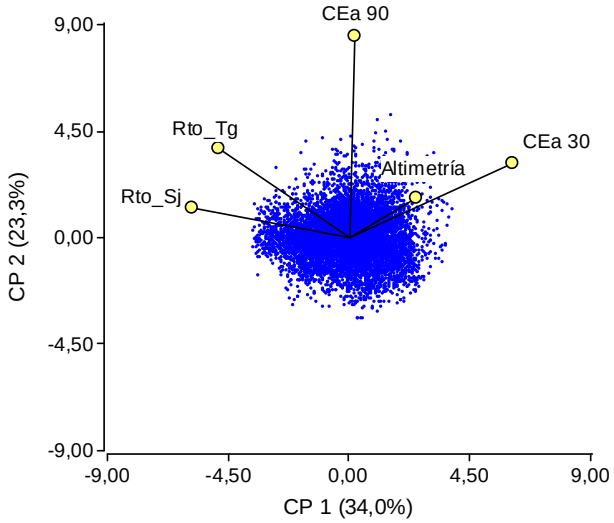


Figura 1.25. Biplot obtenido por análisis de componentes principales. Archivo CE.

Conclusión

Las medidas resumen y los gráficos permitieron observar los valores relevados de las 5 variables cuantitativas de manera más fácil que la que se lograría observando directamente el archivo de datos. Por ahora, hemos podido explorar la base de datos, analizar las distribuciones de las variables, visualizar algunas interesantes correlaciones, detectando que el rendimiento de soja, y el de trigo, se correlacionan con la CEa medida a los 30 cm de profundidad, más que con la altimetría. Por tanto, se podría presuponer que los rendimientos de futuros cultivos en ese lote podrían "copiar" o mapearse según los patrones de variación espacial de Cea 30.

Ejercicios

Ejercicio 1.1: En el cultivo de la papa (Solanum tuberosum L.), el hongo Phytophtora infestans (Mont) de Bary, produce la enfermedad Tizón Tardío. Ésta afecta no solo al rendimiento sino también a la calidad de los tubérculos, ya que produce manchas oscuras en la piel y en el interior de los mismos. Una de las estrategias de control consiste en aplicar fungicida.

En una zona con condiciones ambientales favorables para la presentación del patógeno, se plantea hacer un ensayo trabajando con la variedad de papa Spunta, susceptible a la enfermedad, para comparar el efecto de dos fungicidas (F1 y F2) y, posiblemente, recomendar el uso de alguno de ellos.

Se sembraron tubérculos-semilla de alta sanidad, bajo las condiciones de manejo habituales, en parcelas experimentales de 4 surcos y 5 m de largo cada uno. Para la aplicación de cada fungicida se pulverizó con mochila usando una dosis de 2 kg/ha de producto activo, a intervalos de una semana a partir de los 45 días después de la siembra. De un total de 9 parcelas se seleccionaron al azar un tercio que no fueron pulverizadas, otro tercio en el que se aplicó el F1 y en el tercio restante se usó el F2.

La severidad de la enfermedad se evaluó en base a síntomas en el follaje de una planta tomada al azar de cada parcela, en una escala donde 0= sin síntomas, 1=infección leve, 2= infección moderada, 3= infección severa, 4= infección máxima, al final del periodo de observación.

Después de la cosecha se obtuvo el rendimiento por parcela (kg/ha) de tubérculos y todos ellos fueron clasificados según su destino en: comerciales (con peso igual o mayor a 60 g) y tubérculos que se usarán como semilla (peso menor a 60 g).

Las determinaciones de rendimiento se hicieron sobre los surcos centrales de las parcelas para evitar efectos de bordura y arrastre del fungicida.

De acuerdo a la situación planteada, responda:

a) ¿El estudio es de tipo experimental u observacional?
b) Mencione dos variables podrían ser consideradas como variable respuesta. Clasifíquelas según su naturaleza o tipo.
c) Mencione variables que podrían ser variables de clasificación (o factores). Enumere los valores o niveles de estos factores.
d) ¿Cuáles son las poblaciones sobre las que se desea concluir con el ensayo de fungicida?
e) ¿Cuál es el tamaño de las muestras que serán analizadas en cada población estudiada: n=4 o n=3?
f) ¿Podría estudiarse alguna asociación entre variables?, ¿Cuáles?
g) Al elaborar un análisis estadístico descriptivo: ¿Qué herramientas usaría?

Ejercicio 1.2: Los técnicos de una región de productores de cabras desean identificar las condiciones de manejo que más afectan a la producción de leche. Para ello, cuentan con planillas de 400 productores que contienen datos de los diferentes establecimientos. Como punto de partida del análisis, deciden estudiar la asociación entre el manejo nutricional y la producción de leche. Resuelven considerar a las variables en la siguiente forma:

Manejo nutricional: usa verdeos, usa suplementos, usa verdeos y suplementos, no usa verdeos ni suplementos.

Producción promedio de leche: alta (más de 1,5 kg/día), media (de 1 a 1,5 kg/día) y baja (menor a 1 kg/día).

De acuerdo a la situación planteada:

h) Proponga dos alternativas para realizar este estudio.

i) Suponga n=100 y construya una tabla de contingencia que podría obtenerse, proponiendo frecuencias absolutas razonables.

Ejercicio 1.3: Clasificar las siguientes variables según su naturaleza:

a) Cantidad de vacas en ordeñe por tambo en una cuenca lechera en el año 2011.

b) Estado (preñada o vacía) de una vaquillona (al tacto).

c) Período de tiempo en días transcurridos desde el almacenamiento y hasta que se produce el deterioro del 50% de los frutos almacenados en una cámara.

d) Milímetros de precipitación registrados en una localidad por año.

e) Porcentaje de semillas en dormición en una caja de 50 semillas.

f) Concentración de proteínas (baja, media, alta), en una muestra de leche de cabra.

g) Cociente entre el largo y el ancho de una vaina de soja.

Ejercicio 1.4: Al realizar un inventario forestal en un bosque nativo de la zona chaqueña, se tabularon, entre otros, los datos de la cantidad de especies presentes en el área de muestreo. Represente con un gráfico de sectores la abundancia de las diferentes especies en la muestra, en base al porcentaje de arboles de cada especie respecto del total de arboles presentes.

Especies	Cantidad de árboles
Quebracho blanco	449
Quebracho colorado	401
Guayaibí	224
Itín	176
Palo Santo	112
Otros	241

Ayuda: *cargue los datos en InfoStat, en el menú Gráficos seleccione el submenú Gráficos de sectores, opción Categorías en filas. Luego seleccionar la variable Especies en la ventana Clase y Cantidad de árboles en la ventana Frecuencia. Finalmente accionar Aceptar.*

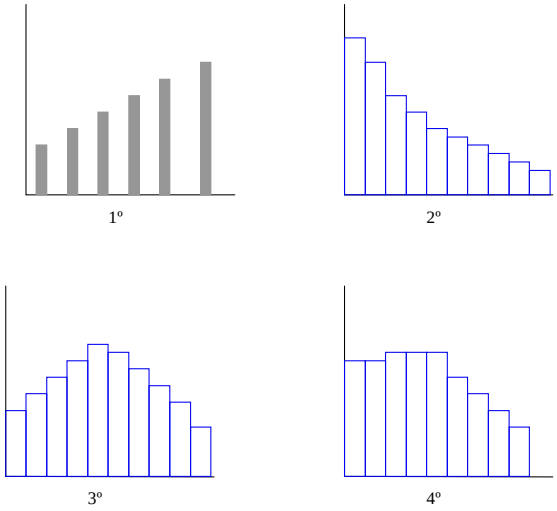
Ejercicio 1.5: A partir de la observación de los siguientes gráficos, ¿Cuál de ellos se asocia con cada una de las siguientes descripciones?

a) Distribución de la población argentina en 2012 según la edad (en años). El rango es de 0 a 90, el tamaño de la clase o amplitud del intervalo es 10.

b) Distribución del número de plantas muertas con relación a la severidad de una enfermedad. La severidad se mide de acuerdo a una escala categórica de 0 a 5 en orden creciente de ataque.

c) Distribución de altura de plantas (en cm) en un cultivo de trigo. Rango de 0 a 50, tamaño de clase 5.

d) Distribución de personas según la distancia (en km) que transitan desde su hogar al trabajo. El rango va de 0 a 50, el tamaño de clase es 5.



Ejercicio 1.6: La siguiente tabla muestra la distribución de frecuencias de la variable producción de papa (en t/ha), según la información obtenida en un muestreo aleatorio de 80 productores:

Producción (t/ha)	Cantidad de productores
(17 - 23]	5
(23 - 28]	21
(28 - 34]	25
(34 - 39]	17
(39 - 45]	9
(45 - 50]	3

Análisis exploratorio de datos

De acuerdo a la situación planteada, responda:

a) ¿En qué porcentaje de la muestra se obtuvieron producciones menores o iguales a 23 t/ha?
b) ¿Qué porcentaje de productores obtuvo una producción mayor a 34 t/ha?
c) ¿Qué cantidad de productores obtuvieron producciones mayores a 39 t/ha?
d) ¿En que intervalo se encuentra el cuantile 0,50? Interprete este valor.
e) ¿En que intervalo se encuentra el cuantile 0,85? Interprete este valor.
f) ¿Qué tipo de gráfico podría usarse para determinar estos cuantiles?

Ejercicio 1.7: Los siguientes datos se refieren al número de dientes por hoja en bulbos de ajo:

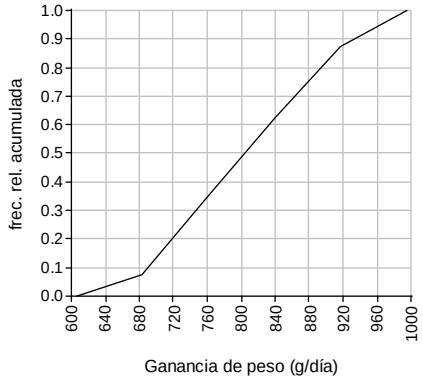
4	2	2	3	3	2	3	3	2	2
3	3	2	1	2	2	2	2	4	2
4	2	3	3	1					

a) Construya una tabla de distribución de frecuencias la variable numero de dientes por hoja.
b) Represente gráficamente la distribución de frecuencias de la variable en la muestra.
c) ¿Cuál es la proporción de hojas con menos de 2 dientes?
d) ¿Cuál es la proporción de hojas con más de 2 dientes?

Ejercicio 1.8: Los siguientes datos corresponden a la ganancia de peso por día (expresada en gramos), de novillos sometidos a una dieta experimental de engorde a corral.

704	890	986	806	798	995	876	705	706	915
801	720	807	960	858	606	798	708	893	906
660	780	615	895	969	880	700	697	804	918
825	809	758	705	800	910	896	708	690	830

a) Obtenga las siguientes medidas resumen: media, mediana, mínimo, máximo rango, varianza (n-1), desviación estándar y coeficiente de variación en la muestra de los datos.
b) Utilizando el gráfico de la distribución de la variable en la muestra, que se muestra a continuación, asignar el valor de Verdadero (V) o Falso (F) a cada una de las consignas del cuadro.



I. La proporción de ganancias de peso diarias entre 720g/día y 800g/día es 0,35.	
II. La proporción de ganancias de peso mayores a 880g/día es igual a 0,75.	
III. Aproximadamente un 35% de las ganancias de peso fueron menores a 760g/día.	
IV. El rango intercuartílico es de aproximadamente 140g/día.	
V. La distribución es asimétrica izquierda	
VI. Si se consideran que ganancias por debajo de los 720g/días son bajas, un total de 8 novillos cumplen esta condición.	
VII. El cuantil 0,5 es igual a 800g/día.	

Ejercicio 1.9: En un estudio en un monte del Chaco Árido se midieron los perímetros basales (en centímetros), de troncos de plantas de quebracho blanco y se obtuvieron los siguientes datos.

138	164	150	132	144	125	149
140	147	136	148	152	144	168
163	119	154	165	146	173	142
140	135	161	145	135	161	145
145	128	157	146	158	126	147
142	138	176	135	153	150	156

a) Utilizando InfoStat, construya los siguientes gráficos que muestren la distribución de la variable: histograma de frecuencias relativas con polígono de frecuencias, grafico de distribución empírica y grafico de cajas (Box-Plot).

b) Compare la información provista por cada grafico. ¿Cuál sería más apropiado para calcular cuantiles?

c) Obtenga las siguientes medidas resumen: media, mediana, X0.25, X0.75, rango, varianza (n-1), desviación estándar y coeficiente de variación.

d) Podría afirmarse que la distribución de la variable es aproximadamente simétrica?

Ejercicio 1.10: Una compañía dedicada a la comercialización de semillas decidió poner a prueba el rendimiento de dos híbridos experimentales de sorgo granífero bajo riego. Se estudiaron dos muestras, una del híbrido A y otra del híbrido B. Los resultados, en qq/ha fueron:

Hibrido A:

110	112	135	140	128	132	123	125	140	142
112	128	152	136	152	139	142	129	150	135
151	113	142	123	118	143	138	135	140	135
119	128	123	142	138	145	136	147	141	137

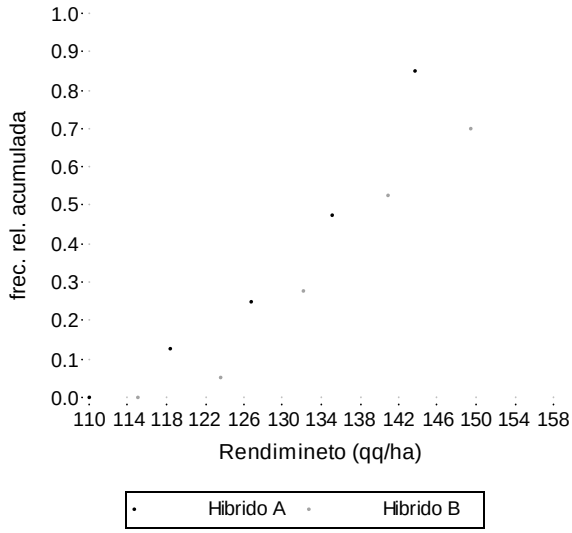
Hibrido B:

115	158	139	143	151	152	148	139	153	125	136
125	130	140	149	150	139	142	138	129	126	137
151	154	139	132	129	146	136	140	150	140	139
128	129	148	146	150	158	153	119	139	154	139

a) En base a medidas de posición, ¿cuál de los dos híbridos recomendaría?

b) En base a medidas de dispersión, ¿cuál de los dos híbridos recomendaría?

c) A partir de las distribuciones de frecuencias graficadas y tabuladas, asignar el valor de Verdadero (V) o Falso (F) a cada una de las consignas del cuadro.



Hibrido	Clase	LI	LS	MC	FA	FR	FAA	FRA
A	1	110,00	118,40	114,20	5	0,13	5	0,13
A	2	118,40	126,80	122,60	5	0,13	10	0,25
A	3	126,80	135,20	131,00	9	0,23	19	0,48
A	4	135,20	143,60	139,40	15	0,38	34	0,85
A	5	143,60	152,00	147,80	6	0,15	40	1,00
B	1	115,00	123,60	119,30	2	0,05	2	0,05
B	2	123,60	132,20	127,90	9	0,23	11	0,28
B	3	132,20	140,80	136,50	10	0,25	21	0,53
B	4	140,80	149,40	145,10	7	0,18	28	0,70
B	5	149,40	158,00	153,70	12	0,30	40	1,00

I. El 30% de los valores obtenidos con el híbrido B son superiores a 149,40 qq/ha.	
II. Con el híbrido A aproximadamente el 80% de los rendimientos fueron superiores a 142 qq/ha.	
III. La proporción de rendimientos entre 134 y 142 qq/ha con el híbrido A es, aproximadamente, de 0,35.	
IV. Con el híbrido B un 53% de los datos de rendimientos fueron mayores a 123,6 y menores o iguales a 140,8 qq/ha.	
V. La proporción de valores de rendimientos por encima de 142 qq/ha fue mayor en el híbrido B que en el A.	
VI. El máximo rendimiento obtenido con el híbrido A fue mayor a 158 qq/ha.	
VII. La mediana híbrido B es de aproximadamente 140 qq/ha.	
VIII. El P(70) de la variedad B es de aproximadamente150 qq/ha.	
IX. El P(60) de la variedad A es de aproximadamente 138 qq/ha.	
X. En ambas distribuciones la diferencia entre el cuantil 0,70 y el cuantil 0,30 es 0,40.	

d) Reproducir, usando InfoStat, el gráfico y las tablas mostradas en el inciso b).

Capítulo 2

Variables aleatorias y probabilidades

Mónica Balzarini
Cecilia Bruno

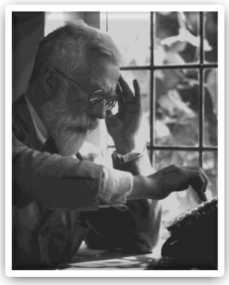
2. Variables aleatorias y probabilidades

Motivación

Hemos usado el término variable para referirnos a una característica de interés en un estudio donde se realizan mediciones. Las mediciones realizadas de la característica varían de unidad a unidad y el valor que asumen en cada una de ellas no puede ser predicho con certeza. Si bien la medición de la característica tiene un "valor esperado", existe una componente de azar que hace a estas mediciones no determinísticas. Tales variables son conocidas como variables aleatorias e interpretadas como una función que relaciona un resultado del estudio con un valor numérico. Las variables aleatorias, por definición están íntimamente asociadas al concepto de probabilidad, término que intuitivamente mencionamos a diario y que es posible calcular. Se puede decir que el descubrimiento de métodos rigurosos para calcular probabilidades ha tenido un profundo efecto en la sociedad moderna. La probabilidad es una medida del grado de incertidumbre sobre el valor que puede asumir una variable aleatoria. A través de probabilidades se puede cuantificar el grado de ignorancia, o certeza, sobre el resultado de un experimento aleatorio. En un universo determinista, donde se conocen todas las condiciones que determinan un evento, no hay probabilidades. En el universo de problemas biológicos, por el contrario, el conocimiento nunca es completo, siendo las probabilidades fundamentales para poder asignar medidas de confiabilidad a las conclusiones. Los conceptos de azar, variable aleatoria y probabilidad están omnipresentes en cualquier aplicación Bioestadística. En este Capítulo presentaremos algunas ideas de su significado sin pretender definirlas formalmente porque, para ello, es necesario recurrir a conceptos matemáticos avanzados de la teoría de la medida.

Conceptos teóricos y procedimientos

El azar



R. Fisher (1890-1962). Nacido en Londres. Científico, matemático, estadístico, biólogo evolutivo y genetista.

La Bioestadística, como una forma de pensar sobre los datos biológicos, es una disciplina científica relativamente nueva, ya que la mayoría de los desarrollos que hoy aplicamos ocurrieron en los últimos 100 años.

Las contribuciones significativas de Ronald Fisher y Karl Pearson se produjeron a principios del siglo pasado para responder a la necesidad de analizar datos en agricultura y biología.

No obstante el núcleo conceptual que sustenta la disciplina formal, el cual está basado en el azar y las probabilidades, se fue moldeando desde muchos años antes; primero por la necesidad de un mundo numérico más fácil de manipular y luego por la necesidad de encontrar o describir patrones estables en observaciones sociales y naturales. Las **leyes del azar** hicieron que el comportamiento social y la naturaleza se vean como menos caprichosos o caóticos.

En 1800 se decía que la palabra **azar** no significaba nada, o bien que designaba una idea del vulgo que señalaba la suerte o "la falta de ley", de manera que debía quedar excluida del pensamiento de la gente ilustrada (Hacking, 1991). La principal creencia del "**determinismo**" o pensamiento determinístico era que todo suceso derivaba de una serie anterior de condiciones.

En oposición, se encontraba la lógica del azar que fue fuertemente influenciada por filósofos franceses e ingleses. Entre la Revolución Industrial y la Revolución Francesa las leyes estadísticas desplazaron el determinismo. En el otro extremo del determinismo, se destaca **Peirce** (1839-1914) quien creía en el **azar absoluto** y en un universo en el que las leyes de la naturaleza, en el mejor de los casos son aproximadas y evolucionan según procesos fortuitos.

"El azar es de todas las cosas la mas entremetida" (Hacking, 1991); El azar siempre está presente y es una componente más a considerar en cualquier problema que involucre variables aleatorias.

Así, el azar ya no era la esencia de la falta de ley sino que estaba en el centro de todas las leyes de la naturaleza y de toda inferencia inductiva racional. Reducir el mundo a una cuestión de probabilidades, es sin duda, una posición extrema, tanto como pensar

que todo está dado y determinado. No obstante la **domesticación del azar** abrió caminos para que las probabilidades y las leyes estadísticas entraran a nuestro mundo.

Al extender las probabilidades a las ciencias de la vida, nació un nuevo tipo de "conocimiento objetivo" producto de nuevas tecnologías estadísticas para obtener información bajo incertidumbre.

Se presentan a continuación algunos conceptos que sustenta la estadística y permiten interpretar y trasladar conceptos abstractos como el de azar y probabilidad en decisiones y respuestas a preguntas sobre variables aleatorias.

Espacio muestral y variables aleatorias

Las variables aleatorias, pueden ser interprtadas como funciones usadas para describir los resultados de un estudio aleatorio. Para el propósito del análisis de datos las clasificamos en cuantitativas y cualitativas y a las primeras en discretas y continuas dependiendo de los posibles valores que la variable pueda asumir (contable o no).
Para la definición formal de variable aleatoria, el tipo de variable es importante. El tipo de variable depende del conjunto de todos los valores que potencialmente pueden asumir en un estudio aleatorio. Tal conjunto de resultados posibles se denomina **espacio muestral** y es usualmente denotado con la letra griega omega (Ω).
Los concpetos de **punto muestral** y **evento aleatorio** de un espacio muestral ayudan a introducir el concepto de variable aleatoria

a) Se denomina **punto muestral** a cada uno de los posibles resultados de un estudio aleatorio, es decir a cada elemento de Ω..
β) Se llama **evento** a cualquier subconjunto de elementos de Ω.

Por ejemplo, supongamos un experimento aleatorio donde se tiran dos dados y se registran los resultados de cada dado. Todos los pares de números del 1 al 6 conforman el espacio muestral. Un evento de Ω, puede ser "que salga un seis en un dado y un seis en el segundo dado"; otro evento puede ser "que salga un seis en un dado y cualquier otro número distinto de seis en el otro dado".
Este segundo evento está constituido por más puntos muestrales y por tanto será **más probable** de ocurrir.

Por esta idea, de que algunos eventos son más probables que otros, es que cuando jugamos al "poker" la "escalera real" otorga más puntos que un "par simple". Esto se debe al hecho de que es más probable obtener un "par simple" . No todas las jugadas de 5 cartas son equiprobables (o igualmente probables)!!

Un sesgo frecuente en el razonamiento probabilístico es pensar que, porque los resultados del experimento son aleatorios, todos los eventos tienen igual probabilidad → NO debemos incurrir en el sesgo de equiprobabilidad!. La probailidad de un evento puede ser, y generalmente lo es, distinta a la de otro evento del mismo espacio muestral.

Definiremos a una **variable aleatoria** como una función que asocia a cada elemento del espacio muestral Ω un número real y luego a cada uno de estos valores le asignaremos probabilidades de ecurrencia. El **tipo de espacio muestral** determina el tipo de variable aleatoria

El espacio muestral asociado a una variable aleatoria de **tipo continua** es no contable, queriendo significar que entre dos valores de la variable, pueden realizarse un número infinito de otros valores.

Además, si el espacio muestral es continuo, la diferencia entre valores de la variable está definida aritméticamente.

Ejemplo de variables aleatorias con espacios muestrales con estas características son los rendimientos, las ganancias de peso, las precipitaciones, entre otras.

Por el contrario, el espacio muestral asociado a una variable de **tipo discreta** es siempre contable, es decir puede ser teóricamente enumerado, aún si éste es infinitamente grande o no está acotado. Por ejemplo, el número de nematodos por hectárea registrado a partir de una muestra aleatoria de hectáreas en producción de papas, podría no tener un valor límite.

En las variables discretas, es posible contar el número de veces que un determinado valor ocurre en el espacio muestral.

Entre las variables discretas es importante distinguir al menos dos subtipos muy comunes en estudios biológicos: las proporciones que provienen de conteos que no puede superar el número de elementos evaluados y los conteos no acotados o sin denominador natural. Ejemplo de una variable discreta expresada como proporción es el número de semillas germinadas en cajas de Petri con 25 semillas cada caja; los resultados se expresan como proporciones porque existe un denominador natural: la

cantidad de semillas por caja. Ejemplo de variable discreta obtenida por un conteo (no acotado) es el número de pústulas de roya por m^2 de cultivo.

Para el caso de proporciones es importante dejar expresado que si bien el valor puede ser continuo en el rango 0-1, el espacio generatriz es discreto, porque la base de la variable es el conteo.

Si el espacio muestral de una variable es discreto pero representado por nombres o códigos que representan categorías excluyentes y exhaustivas de la variable, entonces la variable aleatoria es una variable cualitativa de **tipo categorizada** (nominal u ordinal).

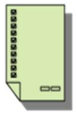
Probabilidad

El concepto de probabilidad puede definirse de distintas formas y con distintos niveles de abstracción. Las definiciones clásica, frecuencial y de Kolmogorov son las más conocidas.

Cuando Ω es finito (el número de puntos muestrales es contable) se puede dar una definición de probabilidad que se basa en la observación de los elementos del espacio muestral. Ésta se desarrolló originariamente estudiando los juegos de azar. y se conoce como el **concepto o enfoque clásico de probabilidad**:

Si A es un subconjunto de puntos muestrales de Ω, entonces la probabilidad de ocurrencia del evento A, denotada por P(A) es:

$$P(A) = \frac{\text{Número de puntos muestrales favorables}}{\text{Número total de puntos muestrales en el espacio muestral}}$$



Dado que el número de puntos favorables es un subconjunto del espacio muestral, se deduce que la probabilidad de un evento siempre será un número positivo, entre 0 y 1.

La **definición frecuencial de probabilidad** es distinta ya que se refiere a una serie repetida de estudios aleatorios. Generalmente se usa cuando el espacio muestral es infinito y por tanto no se pueden enumerar todos los resultados posibles del estudio. Así, se repite el estudio un número grande de veces y se registra la frecuencia relativa de ocurrencia de cada resultado, la que es luego usada como un estimador de probabilidad. La definición frecuencial de probabilidad establece que:

Si A es un evento y n_A es el número de veces que A ocurre en N repeticiones independientes del experimento, la probabilidad del evento A, denotada por **P(A)**, se define como el límite, cuando el número de repeticiones del experimento es grande, de la frecuencia relativa asociada con el evento.

Por ejemplo, consideremos que la germinación de una semilla es un experimento aleatorio (puede germinar o no). Supongamos que con A se representa el evento

"encontrar la semilla germinada". Si se observan 1000 semillas, es decir se repite 1000 veces el ensayo de germinación (N=1000) en condiciones tales que cada observación no afecte a las otras y 600 semillas germinan (n_A=600), se dice que la probabilidad estimada de observar una semilla germinada, está dada por:

$$P(A) = P(\text{observar una semilla germinada}) = \frac{n_A}{N} = 600 / 1000 = 0,6$$

Es claro que, bajo este enfoque, estamos usando un concepto usual en la descripción de datos que hemos discutido en el Capítulo anterior. Ëste es el concepto de frecuencia de ocurrencia de un evento y, entonces, surge la pregunta: ¿Qué diferencia existe entre el concepto de frecuencia relativa y el de probabilidad? Si bien la analogía es fundamental, las frecuencias se entienden como probabilidades sólo cuando N tiende a infinito. Si el número de veces que se repite un experimento no es grande, entonces hablaremos de frecuencia relativa y diremos que ésta "aproxima" una probabilidad.

Otra idea importante para comprender la **medida de probabilidad** es la de **eventos mutuamente excluyentes**.

Se dice que dos eventos son mutuamente excluyentes si cada uno está formado por puntos muestrales distintos, es decir no existe ningún punto muestral en la intersección de los subconjuntos que representan los eventos y , por la teoría de conjuntos, se tiene:

Si A y B son dos eventos de Ω, la **unión de eventos** conforma un nuevo conjunto, que contiene a los puntos muestrales de A y de B. La unión de A y B se denota por $A \cup B$.

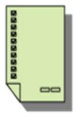
Si A y B son dos eventos de Ω, la **intersección de eventos** conforma un nuevo conjunto, que contiene a los puntos muestrales que simultáneamente pertenecen al subconjunto A y al subconjunto B. Denotaremos la intersección de A y B con $A \cap B$.

Cuando dos eventos son excluyentes, la intersección es cero y por tanto la probabilidad de la unión de esos eventos, $P(A \cup B)$, es la suma de las probabilidades de cada evento. Por el contrario, si la intersección no es vacía, la probabilidad de la unión de eventos es la suma de las probabilidades de cada evento, menos la probabilidad de la intersección.

La definición de probabilidad de Kolmogorov (1937) establece que una función P(.) será considerada una medida de probabilidad si a cada evento de un espacio muestral se le asigna un número real entre 0 y 1 y, además, se cumplen tres axiomas:

c) la probabilidad asociada al evento espacio muestral es igual a 1. Este resultado sugiere que si el evento de interés es todo el espacio muestral, la probabilidad de ocurrencia dado el experimento aleatorio, es 1. Existe certeza de la existencia de un resultado en el espacio muestral.
d) la probabilidad de cualquier evento que sea un subconjunto del espacio muestra es mayor o igual a cero. Si entendemos a la probabilidad como el límite de una frecuencia relativa (cantidad de casos respecto de un total) es claro que las probabilidades nunca pueden ser negativas.
e) Si existen dos o más eventos mutuamente excluyentes, la probabilidad de que ocurra uno u otro evento, es decir la probabilidad de la unión es igual a la suma de la probabilidad de cada uno de estos eventos.

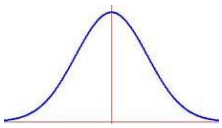
Si los eventos no son excluyentes, el cálculo de la probabilidad de que ocurra uno o el otro evento debe corregirse restando la probabilidad de los elementos en la intersección de ambos eventos. Llegamos a la siguiente proposición:
Dados los eventos A y B, la probabilidad de que ocurra A o B es dada por $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, donde $P(A \cap B)$ denota la probabilidad de que ocurran A y B simultáneamente.
Si A y B son mutuamente excluyentes, $A \cap B$ es vacía y por tanto $P(A \cap B) = 0$.
Un teorema asociado a probabilidades condicionales de eventos, es el Teorema de Bayes. A través de éste es posible encontrar la Probabilidad de un evento de un espacio muestral, dado que otro evento del mismo espacio ya se ha realizado. Por ejemplo, si se estudia la probabilidad de aborto espontáneo en vacas de segunda preñez de un establecimiento ganadero, el cálculo de probabilidad no será el mismo si se condiciona al requerimiento de probabilidad de abortos de vacas de segunda preñez que ya tuvieron un aborto previo. El condicionamiento, restringe el espacio muestral que se usa como referencia en el cálculo de la probabilidad.
El teorema de Bayes establece que $P(A/B) = P(A \cap B)/ P(B)$.
Esta expresión se lee como "la probabilidad condicional del evento A, dado el evneto B (es decir dado que ya ocurrión B), es el cociente entre la probabilidad conjunta de A y B (es decir la probabilidad de que se den ambos eventos) y la probabilidad marginal de B. Cuando la probabilidad de A dado B es igual a la Probabilidad de A, entoces se dice que ambos eventos son independientes, es decir el hecho de que se de B, no afecta la probabilidad de A.



Una idea importante en estadística es la de independencia de eventos. Se dice que dos eventos (A y B) son independientes, si la probabilidad de la intersección de ambos también puede calcularse como el producto de las probabilidades de cada evento, $P(A \cap B)=P(A).P(B)$. En esta situación la probabilidad de A condicional a B es igual a la probabilidad de A (no condicional).

Distribuciones de variables aleatorias

Distinguir el tipo de variable es útil no solo en la etapa exploratoria del análisis de datos sino también en etapas donde se quiera asignar probabilidades a eventos relacionados con la variable.
Para ciertos tipos de variables aleatorias ya se conocen modelos probabilísticos teóricos que ajustan razonablemente bien sus distribuciones empíricas y por tanto se usan estos modelos para el cálculo de probabilidades.



Para una variable continua y de distribución simétrica unimodal, es común el uso del modelo Normal; mientras que para proporciones se piensa en el modelo probabilístico Binomial y para conteos no acotados en el modelo Poisson.

Una vez que se tiene un modelo teórico para la distribución de valores de la variable de interés, es fácil calcular probabilidades.

Hemos visto a una variable aleatoria como un descriptor de eventos aleatorios que tiene asociada una función para asignar probabilidades a esos eventos. La **función de distribución de probabilidad** de una variable aleatoria discreta y la **función de densidad** de una variable aleatoria continua denotada como f(.) contienen exhaustivamente toda la información sobre la variable. La distribución de una variable aleatoria, independientemente del tipo de variable, puede representarse también por su **función de distribución**, denotada como F(y). Esta función asigna a cada valor de la variable un valor entre 0 y 1 que indica la probabilidad de que la variable, observada para un caso particular, asuma un valor menor o igual al valor en que se está evaluando la función. Por ejemplo, si F(30)=0,60 diremos que 0,60 es la probabilidad de que la variable se realice en un caso de análisis particular con el valor de 30 o con un valor menor a 30.

Para ejemplificar los conceptos distribucionales de probabilidad y función de distribución; supongamos un experimento aleatorio donde se tiran dos dados, cada uno de los resultados posibles de la tirada son representados por el par de números que salen:

Ω={(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

Este espacio muestral es finito y discreto y por ello se pueden calcular probabilidades desde el concepto clásico para cualquier variable aleatoria definida sobre el espacio. Por ejemplo, si se quiere estudiar la variable aleatoria y=suma de los puntos en los dos dados, el espacio muestral de esta variable tendrá como elementos las sumas posibles (es decir todos los valores posibles para y).

Ω(y)={2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

Para una variable aleatoria discreta la función de distribución de probabilidades f(.), es aquella que designa una probabilidad de ocurrencia a cada valor de la variable (Tabla 2.1). A diferencia de la función de probabilidad, se tiene la distribución acumulada F(.), la que se puede representar como se muestra en la Tabla 2.1. En la primera columna, se detallan los posibles valores de la variable Y, en la segunda f(y) y en la tercera F(Y).

Tabla 2.1. Distribución de probabilidades y función de distribución de la variable aleatoria Y

y	f(y)	F(y)
2	f(2) = 1/36	F(2) = f(2) = 1/36
3	f(3) = 2/36	F(3) = f(2) + f(3) = 1/36 + 2/36 = 3/36
4	f(4) = 3/36	F(4) = f(2) + f(3) + f(4) = 1/36 + 2/36 + 3/36 = 6/36
5	f(5) = 4/36	F(5) = f(2) + f(3) + f(4) + f(5)= 1/36 + 2/36 + 3/36 + 4/36= 10/36
6	f(6) = 5/36	F(6) = f(2) + f(3) + f(4) + f(5) + f(6) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36= 15/36
7	f(7) = 6/36	F(7) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 = 21/36
8	f(8) = 5/36	F(8) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36= 26/36
9	f(9) = 4/36	F(9) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) + f(9)= 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 + 4/36= 30/36
10	f(10) = 3/36	F(10) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) + f(9) + f(10) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 + 4/36 + 3/36= 33/36
11	f(11) = 2/36	F(11) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) + f(9) + f(10) + f(11) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 + 4/36 + 3/36 + 2/36= 35/36
12	f(12) = 1/36	F(12) = f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) + f(9) + f(10) + f(11) + f(12) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 + 4/36 + 3/36 + 2/36 + 1/36= 36/36= 1

Así, se tiene que:

f) $F(y) = 0$ para valores de $y < 2$
g) $F(y) = 1/36$ para $2 \leq y < 3$
h) $F(y) = 3/36$ para $3 \leq y < 4$
i) $F(y) = 6/36$ para $4 \leq y < 5$
j) $F(y) = 10/36$ para $5 \leq y < 6$
k) $F(y) = 15/36$ para $6 \leq y < 7$
l) $F(y) = 21/36$ para $7 \leq y < 8$
m) $F(y) = 26/36$ para $8 \leq y < 9$
n) $F(y) = 30/36$ para $9 \leq y < 10$
o) $F(y) = 33/36$ para $10 \leq y < 11$
p) $F(y) = 35/36$ para $11 \leq y < 12$
q) $F(y) = 1$ para $y \geq 12$

El gráfico de esta función de distribución acumulada será:

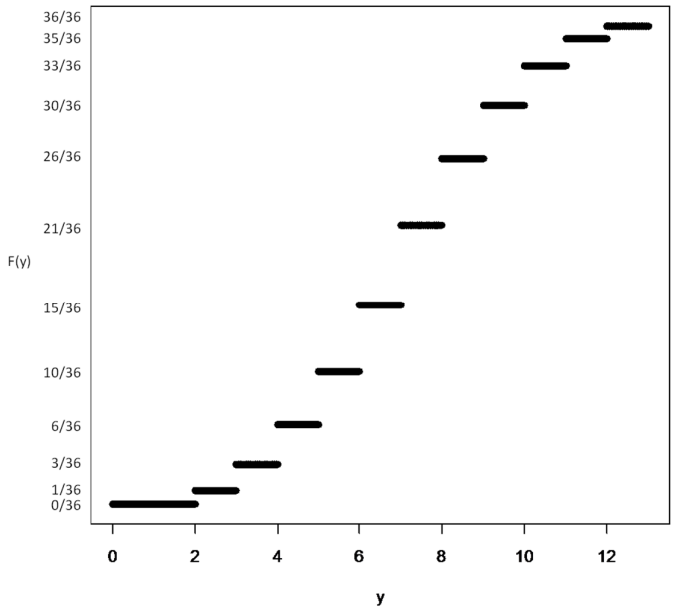


Figura 2.1: Gráfico de la función de distribución de la variable aleatoria "suma de puntos en la tirada de dos dados".

A diferencia de las variables discretas, para las variable continuas pensaremos que los datos son observaciones de una variable aleatoria con función de densidad f(.) más que con función de probabilidad. La función de densidad permite asignar probabilidades a eventos definidos en términos de intervalos. Así, en las variables continuas se podrá conocer la probabilidad de que la variable asuma un valor entre "tanto" y "tanto", mayor a "tanto" o menor a "tanto", pero no exactamente igual a un valor determinado (esta última probabilidad por definición es cero).

Por ejemplo, para la variable rendimiento de soja en qq/ha, esta función podría darnos la probabilidad de que el rendimiento de un lote particular, tomado al azar de una población de lotes donde se ha registrado el rendimiento, asuma un valor entre 30 y 35 qq/ha.

El histograma de la distribución de frecuencias relativas de la variable provee una estimación (aproximación) de f(IC), es decir la probabilidad de que Y asuma un valor en el intervalo de clase IC. Si el número de datos es grande el histograma representa una aproximación buena de la función de densidad teórica ya que las frecuencias relativas pueden interpretarse como probabilidades.

Para una variable continua la función de distribución acumulada, se puede visualizar utilizando un gráfico de dispersión con posibles IC de valores de Y en el eje de las abscisas y la probabilidad acumulada correspondiente a cada IC en el eje de las ordenadas.

La función de distribución empírica en lugar de trabajar con IC, trabaja directamente con los valores observados de Y, relacionando cada valor con la probabilidad de valores menores o iguales. En las gráficas de funciones de distribución empírica, puede leerse la probabilidad de eventos que se expresan en función de desigualdades. Por ejemplo, en la función de distribución de la variable litros de leche producidos por cada lactancia en vacas de establecimientos lecheros de una cuenca lechera, con un valor esperado de 7002 l/lactancia y una desviación estándar de 3975 l/lactancia, podríamos indagar sobre la probabilidad de observar lactancias con producciones menores o iguales a 3000 l o bien con producciones mayores a 3000 l. En la Figura 2.2 se observan los valores F(3000)=0,1 y 1-F(3000)=0,9; por tanto el valor 3000 es el cuantil 0,10 de la distribución de la variable.

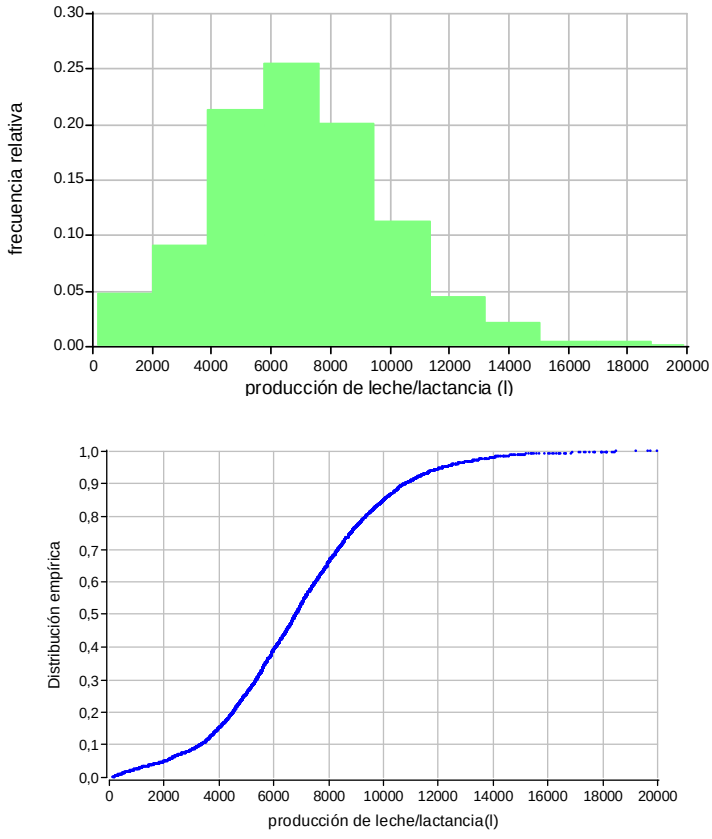


Figura 2.2: Histograma de la variable litros de leche/lactancia (arriba) y gráfico de la función de distribución empírica en una base de datos con 5000 registros (abajo).

Si la distribución teórica no se conoce, las probabilidades acumuladas se pueden aproximar desde las funciones de distribución empírica. Para que las frecuencias que allí se leen puedan ser interpretadas como probabilidades es importante contar con una gran cantidad de datos ya que, como se vio con el concepto frecuencial de probabilidad, las probabilidades deben interpretarse como frecuencias relativas pero en el límite de N tendiendo a infinito.

El concepto de función de distribución acumulada y su aproximación vía la distribución empírica se aplica en gran variedad de situaciones que van desde los juegos de azar hasta el análisis riesgos.

Si bien las funciones de probabilidad y de densidad, de las variables aleatorias discretas y continuas, contienen toda la información sobre los procesos que generan los datos de la variable, usualmente es conveniente resumir las principales características de la distribución. Para todas las distribuciones existen valores numéricos (constantes) que se denominan parámetros de la distribución.

Desde un punto de vista estadístico, un **parámetro** es una función de todos los valores distintos que asume la variable aleatoria en la población. Mientras que una función de los valores la variable, pero en una muestra, se conoce con el nombre de **estadístico**. Luego, los parámetros se derivan de poblaciones y los estadísticos desde muestras.

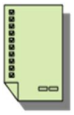
El valor esperado y la varianza son los parámetros más usados en estadística para estudiar y utilizar funciones de distribución de variables aleatorias.

- *El valor esperado, formaliza la idea de valor medio de un fenómeno aleatorio.*
- *La varianza formaliza la idea de incertidumbre y su recíproco la idea de precisión, más varianza indica más incertidumbre sobre el fenómeno y menor precisión de las conclusiones que podemos elaborar desde los datos que lo caracterizan.*

La esperanza matemática de una variable aleatoria, usualmente denotada por E(.) o la letra griega Mu (μ) es, desde un punto de vista intuitivo, un promedio de los valores asumidos por la variable, donde cada valor es ponderado por su probabilidad de ocurrencia.

La esperanza de una variable aleatoria sólo proporciona información parcial acerca de la función de probabilidad (o densidad) ya que explica dónde está posicionada la distribución de valores sobre la recta real. La esperanza es una medida de la tendencia central de la distribución. Pero dos distribuciones con igual esperanza pueden tener distinta dispersión, y por tanto la esperanza puede no ser suficiente para caracterizar completamente de la distribución.

La varianza de una variable aleatoria, denotada por Var(.) o la letra griega Sigma al cuadrado (σ^2), es una medida de dispersión. Su raíz cuadrada, denominada desvío estándar (σ) es usada para expresar la dispersión en término de diferencias (o desvíos) de cada dato respecto a la esperanza.



La varianza es un parámetro que tiene un valor pequeño cuando la mayoría de los valores de la variable se encuentra cerca de la esperanza y crece a medida que éstos se desvían del centro de la distribución. Por ejemplo, la varianza es cero si todos los datos son exactamente iguales.

Existen propiedades de la esperanza y la varianza que son muy usadas en Estadística porque ayudan a comprender la distribución de nuevas variables que han surgido como función de otras variables para las cuales se conoce su Esperanza y su Varianza. Las principales **propiedades de la esperanza** son:

$E(Y+c)=E(Y)+c$

$E(X+Y)=E(X)+E(Y)$

$E(Y+c)=E(Y)+c$

La primera expresión sugiere que si estudiamos una variable aleatoria con determinada esperanza y a cada valor de esa variable se le suma una constante *c*, entonces la esperanza de la nueva variable es igual a la esperanza de la variable original "corrida" o "trasladada" por la constante. La segunda expresión establece que la esperanza de una variable aleatoria obtenida a partir de la suma de otras dos variables, es la suma de las esperanzas de éstas. Finalmente, la tercera propiedad establece que la esperanza de una variable aleatoria que surge de multiplicar cada uno de los valores de una variable original por una constante *c*, es igual a la *c* veces la esperanza de la variable original.

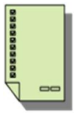
Las principales **propiedades de la varianza** son:

$V(Y)\geq 0$

$V(aY+c)=a^2V(Y)$, dado que a y c son números reales y que la varianza de una constante es cero, es decir, V(c)=0.

$V(Y+X)=V(Y)+V(X)+2Cov(Y,X)$, donde Cov(Y,X) es la covarianza entre la variable Y y la variable X.

$V(Y-X)=V(Y)+V(X)-2Cov(Y,X)$



Las propiedades de la Esperanza y de la Varianza de la distribución de una variable aleatoria premiten establecer cúales serán los parámetros de las distribuciones de "nuevas" variables obtenidas por transformaciones de variables originales con Esperanza y Varianza conocida. Así por ejemplo, si disponemos de la caracterización de la variable rendimiento en qq/ha, podremos saber cuál es la Esperanza y la Varianza de la distribución de los mismos rendimientos expresados en kg/ha ya que entre una y otra variable solo existe la multiplicación por una constante.

Comentarios

En este Capítulo hemos presentado el concepto de variable aleatoria y el de distribución de los valores de una variable aleatoria. La necesidad de definir matemáticamente las funciones que describen la distribución de probabilidad de variables aleatorias proviene del hecho de centrar nuestro interés en fenómenos que no se pueden predecir con exactitud, fenómenos de naturaleza variables donde la componente de azar está siempre presente. Podemos decir que al cuantificar fenómenos aleatorios, hay un valor esperado o un conjunto de valores que con mayor frecuencia se espera que ocurran; no obstante la variable también puede asumir valores alejados del valor esperado. La varianza es una medida de la incertidumbre asociada a la dispersión de los valores de la variable en torno a su valor esperado.

Notación

P(A) → probabilidad del evento A
La esperanza o media de datos poblacionales (distribución) es representada por la letra griega μ, mientras que el estadístico media muestral por la letra que representa la variable con una raya encima de la letra ($\bar{Y}$).
La letra griega σ se usa para representar el parámetro desviación estándar (DE), es decir la desviación estándar calculada con datos de la población o la desviación estándar de la distribución de la variable, mientras que la letra S o la expresión DE se usa para el estadístico desvío estándar muestral.

Definiciones

Definición 2.1: Espacio muestral

Se llama espacio muestral al conjunto de todos los resultados posibles de un estudio aleatorio experimental u observacional. Será denotado con la letra griega omega (Ω).

Definición 2.2: Punto muestral o evento elemental

Se llama punto muestral o evento elemental a cada uno de los elementos del conjunto Ω y será denotado genéricamente como ω.

Definición 2.3: Evento

Dado un espacio muestral Ω se llama evento a cualquier subconjunto de Ω.

Definición 2.4: Eventos mutuamente excluyentes

Se dice que dos eventos A y B de un espacio muestral Ω son mutuamente excluyentes si no contienen elementos en común, o sea si la intersección de A y B es el conjunto vacío ($A \cap B = \varnothing$*).*

Definición 2.5: Medida de probabilidad (Kolmogorov, 1937)

Sea Ω un espacio muestral. La función P(.) que asigna a cada evento de Ω un número real en el intervalo [0,1], se llama medida de probabilidad si satisface los siguientes axiomas:

i. $P(\Omega) = 1$

ii. $P(A) \geq 0$, donde A representa un evento cualquiera de Ω

iii. Si A1, A2, ... es una secuencia de eventos mutuamente excluyentes entonces: $P\left(U_i A_i\right) = \sum_i P\left(A_i\right)$.

Definición 2.6: Probabilidad concepto frecuencial

Si A es un evento y n_A es el número de veces que A ocurre en N repeticiones independientes del experimento, la probabilidad del evento A, denotada por P(A), se define como: $P(A) = \lim_{N \to \infty} \frac{n_A}{N}$.

Definición 2.7: Variable aleatoria

Dado un espacio muestral Ω con un probabilidad asociada, una variable aleatoria Y es una función real definida en Ω tal que $[Y \leq y]$ es un evento aleatorio $\forall y \in \Re$. O sea Y: $\Omega \rightarrow B \subseteq \Re$ es una variable aleatoria si para cualquier $y \in \Re$, $[Y \leq y]$ es un evento aleatorio.

Definición 2.8: Función de distribución acumulada

La función de distribución acumulada, o simplemente función de distribución, de una variable aleatoria Y, denotada por F(.), es una función $F: \Re \rightarrow [0,1]$ tal que:
$F(y) = P([Y \leq y]) \ \forall y \in \Re$.

Definición 2.9: Función de distribución de probabilidad de una variable aleatoria discreta

La función de distribución de probabilidad de una variable aleatoria discreta, denotada por f(.), es una función $f: \Re \rightarrow [0,1]$ tal que:

$f(y) = \begin{cases} P(Y = y) si\ y \in C \\ 0\ en\ caso\ contrario \end{cases}$ *donde C = {y_1, y_2, y_3, ... } es el conjunto de valores que puede tomar la variable aleatoria discreta.*

Definición 2.10: Función de densidad de una v.a.variable aleatoria continua

La función de densidad de una variable aleatoria continua es una función $f(.) \geq 0$ tal que:

$$P([\, y_1 \leq X \leq y_2]) = \int_{y_1}^{y_2} f(y)dy, \forall\ y_1, y_2 \in \Re.$$

Definición 2.11: Esperanza de una variable aleatoria discreta

La esperanza de una variable aleatoria discreta Y, con función de densidad f(.),es:

$$E(Y) = \mu = \sum_{y_i \in C} y_i f(y_i)$$

siendo C el conjunto de valores posibles

Definición 2.12: Esperanza de una variable aleatoria continua

La esperanza de una variable aleatoria continua Y, con función de densidad f(.), es:

$$E(Y) = \mu = \int_{-\infty}^{\infty} y\, f(y)\, dy$$

Definición 2.13: Varianza de una variable aleatoria discreta

La varianza de una variable aleatoria discreta Y se define como:

$Var(Y) = \sigma^2 = \sum_{y_i \in C} (y_i - \mu)^2 f(y_i)$ *donde* μ *= E(Y), f(.) la función de distribución de probabilidad y C =* $\{y_1, y_2, \ldots\}$ *el conjunto de valores posibles.*

Definición 2.14: Varianza de una variable aleatoria continua

La varianza de una variable aleatoria continua Y, se define como:

$V(Y) = \sigma^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) d(y)$ *donde* μ *= E(Y) y f(.) la función de densidad.*

Aplicación

Análisis de datos de velocidad del viento

En un establecimiento agrícola se desea usar la energía eólica como una energía alternativa para bombeo de agua subterránea. El viento, al estar constantemente en movimiento produce energía. Se estima que la energía contenida en los vientos es aproximadamente el 2% del total de la energía solar que alcanza la tierra. El contenido energético del viento depende de su velocidad. Cerca del suelo, la velocidad es baja, aumentando rápidamente con la altura. Cuanto más accidentada sea la superficie del terreno, más frenará ésta al viento. Es por ello que sopla con menos velocidad en las depresiones terrestres y más sobre las colinas. Además, el viento sopla con más fuerza sobre el mar que en la tierra. El instrumento que mide la velocidad del viento es el anemómetro, que generalmente está formado por un molinete de tres brazos, separados por ángulos de 120° que se mueve alrededor de un eje vertical. Los brazos giran con el viento y accionan un contador que indica en base al número de revoluciones, la velocidad del viento incidente. La velocidad del viento se mide en nudos, generalmente en náutica, y mediante la escala Beaufort que, ideada en el siglo XIX por el Almirante Beaufort; esta es una escala numérica utilizada en meteorología

que describe la velocidad del viento en km/h o m/hora. Esta asigna números que van del 0 (calma) a 12 (huracán).

Estrategias de Análisis

Se compararán datos de viento en dos lugares de un establecimiento. Para ello se realizaron tres mil lecturas con anemómetro, en la zona Norte y en la Zona Sur. Para analizar la distribución del viento en cada sitio, se construyeron las distribuciones empíricas de la variable velocidad del viento y se analizaron parámetros de posición y de dispersión de estas distribuciones.

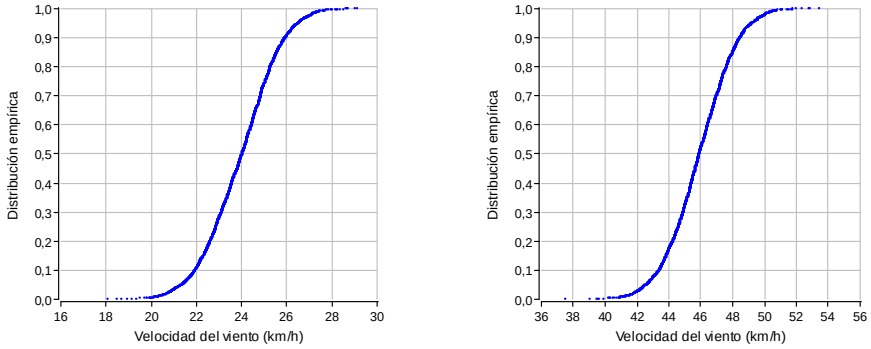


Figura 2.3: Gráfico de la distribución empírica de la velocidad del viento (km/h) en dos zonas de un establecimiento agrícola, denominadas zona sur (izquierda) y zona norte (derecha)

Se considera que un molino de viento para generar electricidad, comienza a funcionar cuando el viento alcanza una velocidad de unos 19 km/h, logran su máximo rendimiento con vientos entre 40 y 48 km/h y deja de funcionar cuando los vientos alcanzan los 100 km/h. Los lugares ideales para la instalación de los generadores de turbinas son aquellos en los que el promedio anual de la velocidad del viento es de al menos 21 km/h. Mientras que si el molino se coloca con fines de extracción de agua subterránea, se espera una velocidad del viento promedio de 26 km/h.Las distribuciones disponibles muestran que en la zona Norte la mediana de la velocidad del viento es aproximadamente de 35 km/h, esto es equivalente a decir que el 50% de las veces, el viento alcanza una velocidad promedio de 35 km/h o menor. El 10% de las veces, la velocidad del viento superó 39 km/h. El rango de velocidades en la zona norte varía entre 26 km/h hasta 44 km/h, mientras que, en la zona sur se registran velocidades del viento que oscilan entre los 20 y 28 km/h. Sólo el 10% de las veces la velocidad del viento supera los 26 km/h.

Conclusión

Se recomendaría la zona norte como aquella con mejores aptitudes en cuanto a la velocidad del viento para poder utilizar la energía eólica para extraer agua.

Ejercicios

Ejercicio 2.1: Supongamos que se toma una muestra aleatoria con reposición de tamaño n=2 a partir del conjunto {1,2,3} y se produce el siguiente espacio muestral con 9 puntos muestrales:

Ω={(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)}

Supongamos además que definimos la variable aleatoria Y=suma de los dos números, que conforma un nuevo espacio probabilístico y que estamos interesados en los siguientes eventos:

El evento A conformado por los puntos muestrales cuya suma sea un número par, es decir, A={(1,1),(1,3),(2,2),(3,1),(3,3)} y P(A)= 5/9.

El evento B conformado por los puntos muestrales cuya suma sea un número impar, siendo B={(1,2),(2,1),(2,3),(3,2)} y P(B)=4/9.

El evento C conformado por los elementos cuya suma es 5.

Preguntas:

a) ¿Qué tipo de concepto de probabilidad aplicaría para calcular probabilidades?
b) Los eventos A y B, ¿son independientes?
c) ¿Cuál es la probabilidad de que ocurra A o B?
d) ¿Cuál es la probabilidad de que ocurra B o C? Representar tabularmente a F(Y).

Ejercicio 2.2: Los siguientes datos corresponden a clasificaciones de 320 lotes en producción de tres grupos o consorcios de productores. Las clasificaciones se realizaron según el nivel de la producción

Nivel producción	Grupo de productores A	Grupo de productores B	Grupo de productores C	Total
Alto	20	10	50	80
Medio	25	18	27	70
Bajo	75	62	33	170
Total	120	90	110	320

Preguntas:

a) Especificar un evento simple relacionado a la variable nivel de producción.
b) Conociendo esta tabla, qué concepto de probabilidad podría aplicar para asignar probabilidad a eventos de interés?
c) Cuál es la probabilidad del evento especificado?
d) Cuál es la probabilidad de obtener un nivel bajo de producción y ser productor del grupo A?

e) Cuál es la probabilidad de un nivel bajo de producción dado que el productor pertenece al grupo A? Cómo se llama este tipo de probabilidad?

Ejercicio 2.3: Los siguientes datos corresponden a la venta de tractores que registra una empresa de maquinarias agrícolas en los días laborables del último año:

Tractores vendidos	Cantidad de días
0	110
1	80
2	35
3	25
4 o más	10
Total	260

Preguntas:

a) Cuál es la variable en estudio?
b) Cuántos resultados posibles tiene la variable? Qué tipo de variable es?
c) Cuál es la probabilidad que hoy no venda ningún tractor?
d) Cuál es la probabilidad que un día, seleccionado al azar dentro de los días laborables del año, venda 3 o más tractores?
e) Cuál es la probabilidad que en los próximos dos días venda 3 tractores?

Ejercicio 2.4: Si los eventos A y B pertenecen al mismo espacio probabilístico y se conoce que P(A/B)=0 , P(A)=0.10 y P(B)0=.50

Preguntas:

a) Ay B son mutuamente excluyentes?
b) A y B son estadísticamente dependientes?

Ejercicio 2.5: Se registró el nivel de educación alcanzado de empleados rurales registrados en una zona según la categoría de edad.

Nivel educación alcanzado	Abreviaturas	Menores de 25 años de edad	Entre 25 y 40 años	Mayores de 40 años de edad	Total
No alcanzó ningún nivel	SE	120	250	340	710
Nivel Primario	P	100	200	300	600
Nivel Secundario	S	50	100	60	210
Nivel Terciario	T	0	30	5	35
Nivel Universitario	U	1	25	10	36
Nivel Posgrado	PG	0	5	0	5
Total		271	610	715	1596

Preguntas:

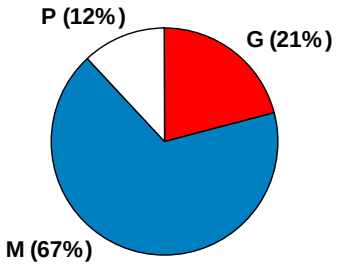
a) Cuál es la probabilidad que un empleado, seleccionado al azar de los registrados en la zona, acredite al menos el nivel secundario de estudio?

b) Cuál es la probabilidad que una persona que se selecciona al azar desde las registradas, sea menor de 25 años?

c) Los eventos: ser menor de 25 años y ser mayor de 40 años, ¿son mutuamente excluyentes? Son estos eventos independientes?

d) Cuál es la probabilidad que teniendo más de 40 años, tenga nivel terciario completo o tenga universitario completo?

Ejercicio 2.6: El gráfico muestra la estructura de productores de una región según la superficie trabajada por cada productor. De un total de 2385 productores, el 21% fue caracterizado como productor grande (G), el 67% como mediano productor (M) y el 12% como pequeño productor (P).



Pregunta:

a) Si se selecciona un productor al azar, cual es la probabilidad que sea un pequeño productor o un productor mediano? Cómo son estos eventos?

Ejercicio 2.7: Se conoce que los niveles de infestación de un cultivo (medido como chinches por metro lineal de surco) en una región se distribuyen según la siguiente función:

Cantidad de chinches por metro lineal de surco	Probabilidad
0	0,35
1	0,25
2	0,10
3	0,20
4	0,05
5 o más	0,05

Variables aleatorias y probabilidades

Preguntas:

a) Graficar la función de probabilidad y la distribución acumulada de la variable.
b) Para un metro lineal elegido al azar, cuál es la probabilidad de encontrar más de 2 chinches?
c) Cuál es el valor esperado del número de chinches por metro? Como se interpreta este valor?
d) Cuál es la varianza de la variable?

Ejercicio 2.8: Se cuenta con datos históricos de rendimiento de lotes de girasol de dos zonas pertenecientes a la región girasolera argentina. Los datos pertenecen a una campaña y están expresados en qq/ha. Una zona es el Sur Oeste de la provincia de Buenos Aires (SO) y la otra zona el Centro de la provincia de Buenos Aires (CBA). En la siguiente figura se muestra la función de distribución empírica de la producción de girasol en cada una de las zonas. Usaremos las FRA para aproximar probabilidades. Estas han sido calculadas con más de 1000 datos por zona.

a) ¿Cuál es la producción de girasol sólo superada por el 10% de los rendimientos en la zona CBA?
b) ¿Con que Probabilidd se supera un rendimiento de 30 qq/ha en SO?
c) ¿En qué zona hay mayor probabilidad de obtener rendimientos altos?
d) ¿Cuál es la zona con mayor varianza en sus rendimientos?

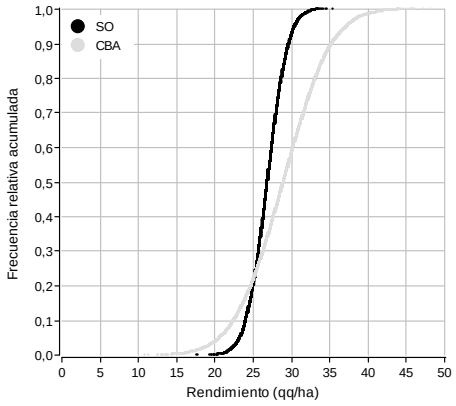


Gráfico de la función de distribución empírica de rendimientos de girasol.

Capítulo 3

Modelos probabilísticos

Fernando Casanoves

3. Modelos probabilísticos

Motivación

Cuando estudiamos una variable aleatoria, es de interés calcular probabilidades sobre la ocurrencia de ciertos valores (eventos). Por ejemplo, podríamos estimar la probabilidad de obtener un rendimiento de maíz superior a 100 qq/ha, de tomar 100 semillas y que no germinen más de 90, o de tomar una muestra de insectos con golpes de red y capturar menos de 20 insectos. Los cálculos de probabilidad pueden hacerse luego de enumerar todo el espacio muestral, cuando esto es posible, usando información sobre las frecuencias con que ocurren los distintos eventos o bien usando un modelo de distribución teórico que ajuste relativamente bien a la distribuón empírica de la variable. Para la elección del modelo de probabilidad teórico, es importante considerar características de la variable tales como la forma en que se cuantifica (medición, proporción, conteo, etc.). La naturaleza de la variable, es decir si es discreta o continua, las condiciones en que se realiza el experimento y el registro de los valores son determinantes para la selección de un modelo probabilístico.

Conceptos teóricos y procedimientos

El concepto de variable aleatoria está íntimamente ligado al de función de densidad y función de distribución. Por lo general la forma o expresión matemática de la función que describe a la variable aleatoria no se conoce, por lo que los técnicos e investigadores suelen proceder a recolectar datos mediante estudios observacionales o experimentales, y a partir de ellos buscar cuál es la función que mejor describe la o las variables aleatorias en estudio.

No cualquier función matemática es útil para caracterizar una variable aleatoria, por el contrario, las funciones de densidad y de distribución acumulada deben reunir una serie de propiedades para que sea posible asignar probabilidades a los eventos de interés a partir de las mismas. Desde el punto de vista teórico se han estudiado con suficiente detalle un conjunto de funciones matemáticas que verifican las propiedades de las funciones de distribución acumulada y de las funciones de densidad tanto para variables discretas como para continuas. Luego, el técnico o investigador que no conoce la

función exacta que caracteriza a la variable aleatoria que está estudiando puede, por conocimiento empírico, proponer alguna de las funciones, del conjunto de funciones antes indicado, para describir el comportamiento de su variable. De la habilidad para escoger una distribución adecuada, depende la calidad de los modelos y las predicciones que se construyan.

Variables aleatorias continuas

Para seleccionar un modelo probabilístico para una variable aleatoria continua cuando se tienen datos de esa variable, resulta recomendable graficar un histograma de frecuencias relativas y observar la forma del mismo. Existen diversos modelos téoricos o funciones matemáticas que podrían ajustar o "aproximar bien" la forma del histograma. Por ejemplo, en la Figura 3.1 se presentan cuatro histogramas de frecuencias relativas diferentes y a cada uno de ellos se les ha superpuesto un modelo teórico que aproxima relativamente bien la forma del histograma. Los nombres de estos modelos de probabilidad son Chi-Cuadrado, Normal, Exponencial y Uniforme.

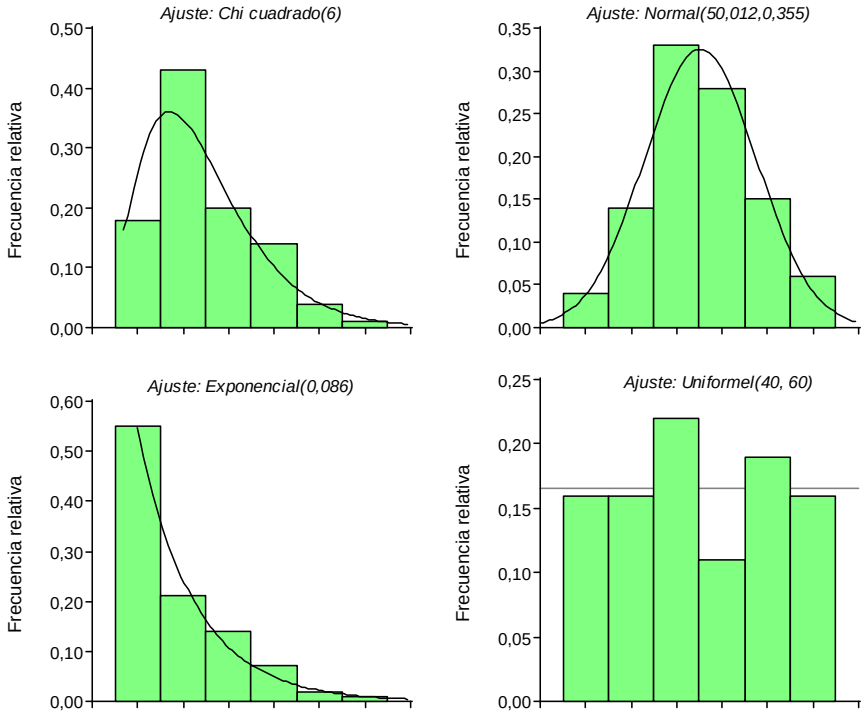


Figura 3.1. Histogramas de frecuencias relativas de variables aleatorias continuas donde se superponen funciones de modelos probabilísticos teóricos que ajustan relativamente bien las formas de los histogramas.

En esta sección se darán ejemplos del modelo de probabilidad Normal o Gaussiano. Esta distribución es, podríamos afirmar, la más usada en las ciencias biológicas, agronómicas y forestales ya que usualmente ajusta bien histogramas de frecuencias de variables como el peso y la altura de seres vivos así como otras mediciones morfométricas además del rendimiento. Estas características, particularmente interesantes en agronomía, son producidas por el resultado de la acción conjunta de muchos factores y por tanto asumen muchos valores distintos (en un continuo de valores posibles) entre las unidades de análisis. No obstante, algún valor o intervalo de valores se repite con mayor frecuencia, mientras que otros muy alejados de estos valores centrales (por ser mucho mayores o mucho menores) aparecen con menor frecuencia.
La distribución normal se usa para el cálculo de probabilidades de variables continuas, cuyos histogramas tienen forma "acampanada", por eso y porque su expresión matemática fue estudiada por Gauss, también se conoce como modelo Gaussiano. El siguiente histograma corresponde a la variable aleatoria perímetro que fue medido sobre numerosas cabezas de ajo, para el cual el modelo Normal con media 17,2 y varianza 10,7 pareciera proveer un buen ajuste (Figura 3.2).

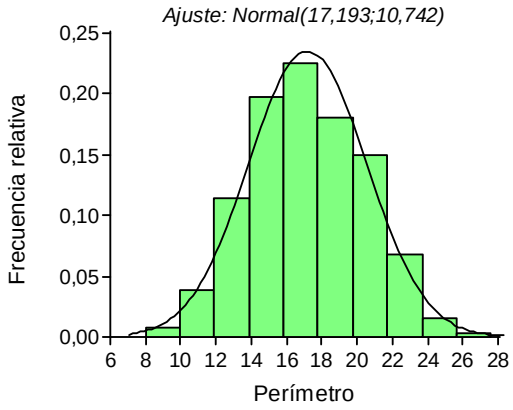
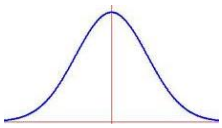


Figura 3.2. Histograma de frecuencias relativas para la variable perímetro de cabezas de ajo (Archivo Ajoblanc).

Como puede apreciarse, la distribución de frecuencias de esta variable tiene ciertas características: es aproximadamente simétrica, posee una gran cantidad de valores cerca del centro. La media, la moda y a la mediana son prácticamente iguales y los valores extremos, tanto inferiores como superiores, tienen menor frecuencia de ocurrencia que los valores centrales. Además la distribución es simétrica, es decir con distribución de valores superiores a la media igual a la de valores por debajo de la media.

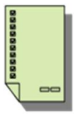


El modelo Normal se usa para calcular probabilidad en variables continuas y de distribución simétrica unimodal.

La **distribución normal** de una variable aleatoria Y tiene la siguiente función de densidad:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

donde μ puede asumir valores entre menos infinito e infinito y σ puede asumir valores entre cero e infinito. La localización del centro de la campana está dado por el parámetro μ (también conocido como esperanza de Y) y la mayor o menor amplitud de la campana viene dada por el parámetro σ^2 (la varianza de Y en la población).
Como la función es simétrica respecto de μ, ésta divide a la gráfica en partes iguales. Está definida para todo $\Re$ y para valores en la abscisa que tienden a infinito y a menos infinito, se aproxima al eje horizontal sin tocarlo (curva asintótica). Como toda función de densidad, el área comprendida entre el eje de las abscisas y la curva es igual a la unidad.



La función de densidad de una variable aleatoria normal tendrá distintas formas dependiendo de sus parámetros que son la esperanza y varianza.

La distribución normal es un modelo de probabilidad y una vez adoptado el modelo es posible responder a las siguientes preguntas:
-¿Cuál es la probabilidad de que la variable en estudio tome valores menores a un valor determinado?.
Por ejemplo, si la variable es el rendimiento de un cultivar, el responder a esta pregunta podría indicar la posibilidad de obtener rendimientos que no justifiquen el costo de producción.
-¿Cuál es la probabilidad de que la variable en estudio tome valores mayores a un valor determinado?.
Si la variable aleatoria en estudio es la cantidad de semillas de maleza en el suelo antes de la siembra, el responder a esta pregunta podría indicar si se necesitará o no aplicar herbicida (este podría ser el caso de modelación de una variable aleatoria discreta como si se tratara de una continua).
-¿Cuál es la probabilidad de que la variable en estudio tome valores entre 2 valores determinados?.
Esta probabilidad es de interés, por ejemplo, al clasificar tubérculos de papa dado que aquellos con volumen entre 59 cm^3 y 80 cm^3 son considerados de valor comercial.

Podemos tener distribuciones normales con iguales valores de varianza pero diferentes valores de esperanza.Supongamos que la producción de leche diaria de las vacas de un tambo se distribuye como el modelo normal, con esperanza 25 l y varianza 9 l^2. Si a las vacas se les da una nueva ración que aumenta en 5 l la producción diaria, pero no modifica las varianzas, la función de densidad de la producción de leche diaria de los animales con la nueva ración tendrá un valor esperado de 30 l (Figura 3.3).

Para hacer una gráfica que represente las densidades en estudio se usó el software InfoStat accionando el menú APLICACIONES ⇒ DIDÁCTICAS ⇒ GRÁFICOS DE FUNCIONES DE DENSIDAD CONTINUAS, se especificaron los parámetros como se muestra en la Figura 3.4 y posteriormente, en la ventana de **Herramientas gráficas**, solapa **Series**, primero se presiona el botón **Clonar**, y luego, a una de las series se le cambió la media a 30 (Figura 3.5).

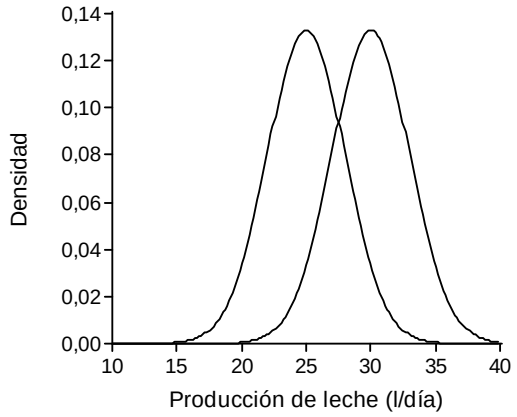
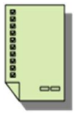


Figura 3.3. Funciones de densidad normal con la misma varianza pero distintas medias ($\mu_1 = 25$ y $\mu_2 = 30$)



El modelo Normal permite aproximar, como se dijo, el comportamiento estadístico de muchas variables continuas pero también incluso de algunas variables discretas cuando los tamaños muestrales con los que se trabaja son grandes.

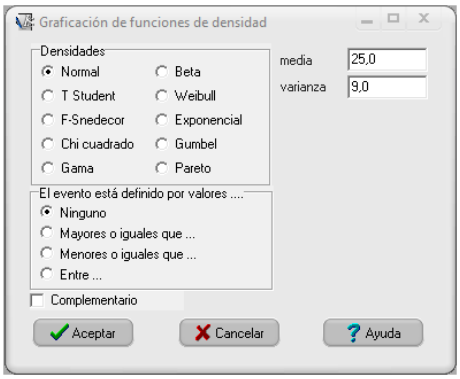


Figura 3.4. InfoStat. Ventana de diálogo para graficar funciones de densidad continua.

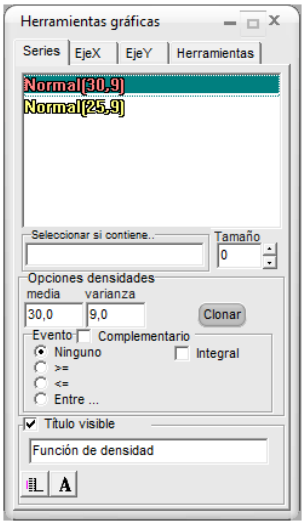


Figura 3.5. InfoStat. Ventana Herramientas gráficas con las especificaciones para obtener las densidades normales de la Figura 3.3.

En un tambo con producciones diarias distribuidas normal con media 25 l y varianza 9 l^2, el productor puede decidir darles más ración a las vacas con menor producción y menos ración a las vacas de mayor producción, ocasionando un cambio en la varianza, pero no necesariamente sobre la media. Se espera que con raciones diferenciales, la varianza disminuya, ya que las vacas que producían poco, al tener más ración se acercarán al promedio de las producciones, y las vacas con mayor producción, al tener una quita se acercarán también al promedio de las producciones, así, la amplitud de las producciones será menor. Si la nueva técnica reduce la varianza a 2, la gráfica que compara las dos condiciones experimentales podría ser como la de la Figura 3.6.

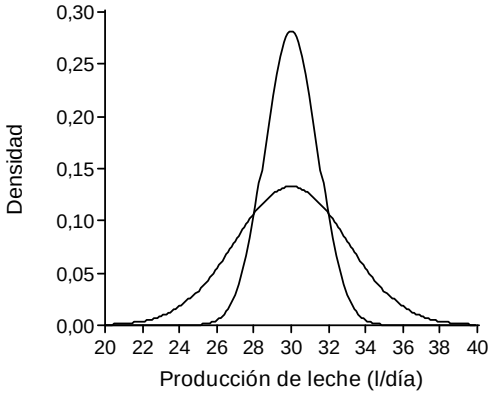


Figura 3.6. Funciones de densidad normal con la misma media pero distintas varianzas ($\sigma_1^2 = 9$ *y* $\sigma_2^2 = 2$)

El cálculo de probabilidades en variables aleatorias continuas, como es el caso de las variables con distribución Normal, puede realizarse gráficamente midiendo el área bajo la curva de la función de densidad correspondiente al intervalo de valores de interés. En cualquier distribución continua si se fijan dos puntos cualesquiera, por ejemplo y_1 y y_2, sobre el eje que representa los valores de la variable (abscisas), la porción del área por debajo de la curva que queda comprendida entre esos dos puntos corresponde a la probabilidad de que la variable aleatoria *se realice* entre y_1 y y_2. Si se llama A a esta área, se puede representar simbólicamente lo expuesto anteriormente como:

$$A = P (y_1 \leq Y \leq y_2)$$

La probabilidad que un dato de rendimiento tomado al azar desde la población esté comprendido en el intervalo 50 a 65 qq/ha, está representada por el área sombreada en la Figura 3.7 y es igual a la proporción de la superficie del área respecto al área total bajo la curva (que por ser una función de densidad vale 1).

Por ejemplo, si Y es el rendimiento de un híbrido de maíz que puede modelarse con una distribución normal, con media de 60 qq/ha y varianza de 49 $(qq/ha)^2$ (esta especificación suele escribirse de manera concisa como Y~N(60; 49)).

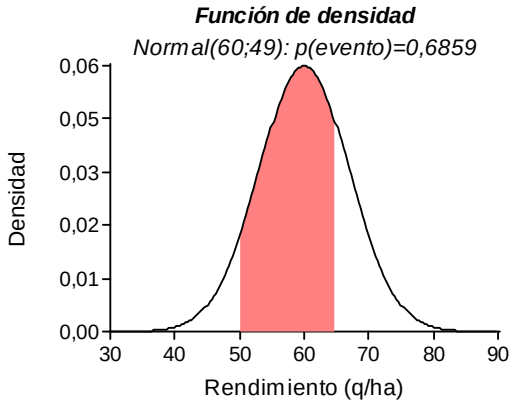
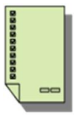


Figura 3.7. Función de densidad normal para el rendimiento de un híbrido de maíz con la probabilidad del evento [50 ≤Y ≤65] representado por el área sombreada.

De esta manera se lee que la probabilidad del evento "observar un rendimiento comprendido entre 50 y 65 qq/ha" es de 0,6859. Esta probabilidad se obtuvo con InfoStat integrando la función de densidad normal (con parámetros media=60 y varianza=49) entre 50 y 65:

$$P(50 \leq Y \leq 65) = \int_{50}^{65} \frac{1}{7\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-60}{7}\right)^2} dy$$



Antes de disponer de software que calculen la proporción relativa de éstas áreas, es decir resulevan estas integrales, se usaban tablas construidas de manera tal de poner a disposición del usuario las probabilidades para una serie de eventos posibles.

Las tablas y software son usados para calcular probabilidades sin necesidad de resolver integrales como el de la función de densidad normal. Para el caso de la distribución normal, las tablas existentes (ver Tablas Estadísticas) tienen las áreas (integrales) correspondientes a valores menores o iguales a un valor particular. Estas áreas son interpretadas como probabilidades acumuladas. No obstante, ellas no están disponibles para cualquier valor de cualquier variable normal ya que existen infinitas distribuciones normales.

La tabla de distribución normal presenta las áreas correspondientes a valores posibles de una normal de media 0 y varianza 1. Esta densidad normal particular, recibe el nombre de **normal estándar**.

Para usar las tablas, debemos expresar nuestra variable como una normal estándar. Para ello usamos una transformación llamada **estandarización** que nos permite llevar

cualquier distribución normal a la distribución normal estándar. La transformación, estandarización, tiene la siguiente forma:

$$Z = \frac{Y - \mu}{\sqrt{\sigma^2}}$$

donde Y es el valor de la variable aleatoria que define el evento de interés, μ y σ^2 son la media y la varianza de la distribución de Y. La nueva variable aleatoria Z, obtenida mediante estandarización de Y, se distribuye normal con media cero y varianza uno, es decir, normal estándar.

Siguiendo el ejemplo del rendimiento de un híbrido, para obtener la probabilidad de encontrar valores de rendimientos entre 50 y 65 qq/ha se deberá calcular:

$$Z_1 = \frac{50 - 60}{\sqrt{49}} \simeq -1,4286$$

$$Z_2 = \frac{65 - 60}{\sqrt{49}} \simeq 0,7143$$

La importancia de esta transformación radica en que las probabilidades que se obtendrían a partir de la distribución original de la variable Y son iguales a las obtenidas luego de estandarizar la variable Y y buscar los valores de probabilidad asociados a los valores de Z producto (hasta valores de Y) de la estandarización de los valores de Y, en una tabla de normal estándar (Figura 3.8).

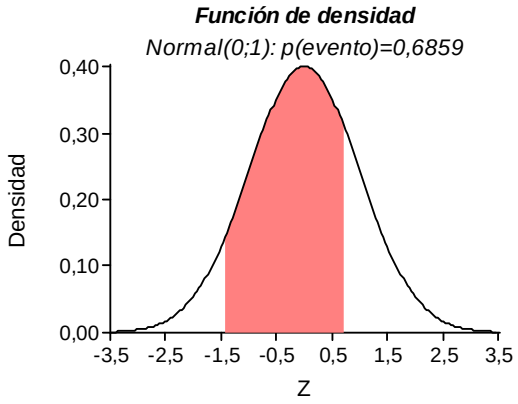


Figura 3.8. Función de densidad normal estándar con la probabilidad del evento [-1,4286 ≤Z ≤0,7143] representada por el área sombreada.

El cálculo puede expresarse de la siguiente manera:

P[50≤Y≤ 65]=F(65)-F(50)=P[Y≤ 65]-P[Y≤50]=P[-1,4286≤Z≤0,7143]=

P [Z ≤ 0,7143] - P [Z ≤ -1,4286]= 0,7625 – 0,0766=0,6859

De esta manera la probabilidad de interés se calcula como la diferencia entre las probabilidades de los eventos [Z ≤ 0,7143] y [Z ≤ -1,4286], es decir, entre dos eventos cuya probabilidad se lee directamente de una función de distribución acumulada que está tabulada (Figura 3.9).

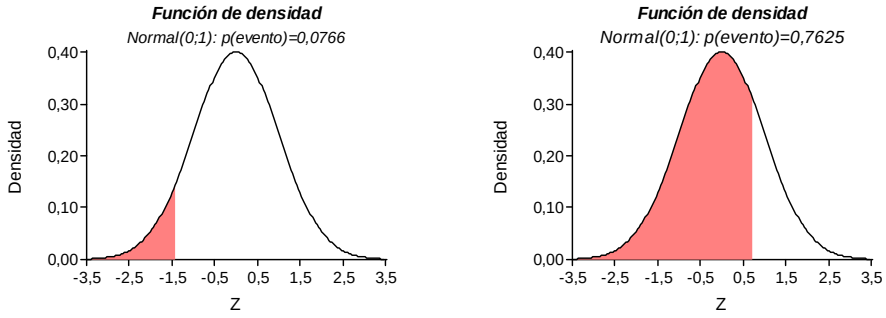


Figura 3.9. Funciones de densidad normal estándar con la probabilidad del evento [Z ≤-1,4286] (izquierda) y [Z ≤0,7143] (derecha) representadas por el área sombreada.

Si se quiere calcular la probabilidad de obtener rendimientos menores a 55 qq/ha, entonces solo necesitamos estandarizar el valor 55 de la variable Y, es decir encontrar que valor en la densidad de la variable Z (normal estándar) es equivalente al valor 55 de la distribución de Y. Luego,

$$Z = \frac{55 - 60}{\sqrt{49}} \simeq -0,7143$$

El cálculo de la probabilida puede expresarse como:
P[Y ≤ 55] = P [Z≤ -0,7143] = 0,2375 (Figura 3.10)

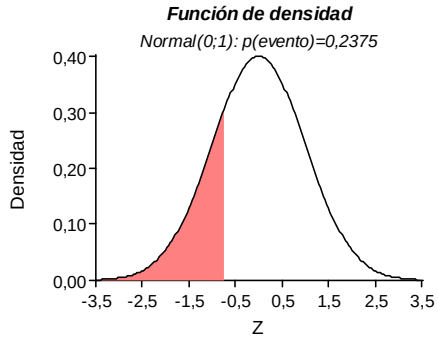


Figura 3.10. Función de densidad normal estándar con la probabilidad del evento [Z≤-0,7143] representada por el área sombreada.

Si se quiere calcular la probabilidad de observar valores mayores a 65 qq/ha en la distribución de la variable Y, entonces debemos estandarizar ese valor para obtener un valor de Z que sea equvalente al 65qq/ha de la distribución de Y:

$$Z = \frac{65 - 60}{\sqrt{49}} \simeq 0,7143$$

Luego, P[Y Ó65] = 1-P [Y ≤ 65] = 1-P [Z ≤ 0,7143] = 1-0,7625=0,2375 (Figura 3.11)

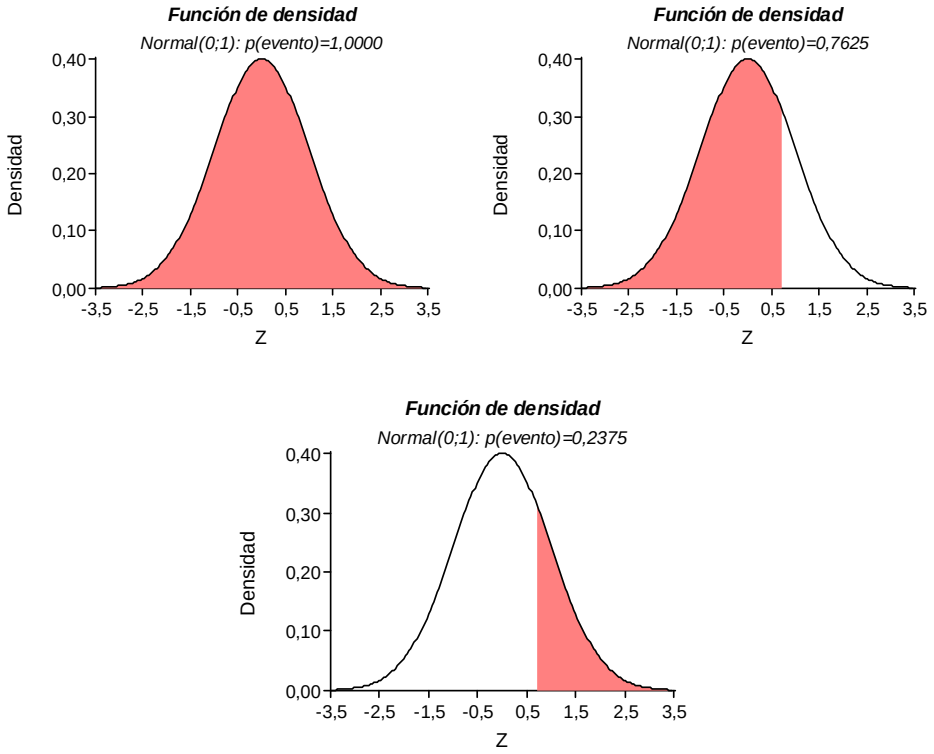


Figura 3.11. Funciones de densidad normal estándar con la probabilidad del evento [-∝ ≤Z ≤∝] (izquierda), [Z ≤0,7143] (derecha) y [Z Ó0,7143] (abajo) representados por el área sombreada.

En síntesis, podemos decir que si Y se distribuye normal con media μ y varianza σ^2, luego la variable Z (la estandarización de Y), se distribuye normal con media 0 y varianza 1, esto es:

$$Y \sim N(\mu,\sigma^2) \Rightarrow Z = \frac{Y - \mu}{\sqrt{\sigma^2}} \sim N(0,1)$$

Se ha reducido el problema de tener muchas distribuciones, a tener una sola. Pero para hallar la probabilidad de que Y tome un valor entre dos valores determinados se deberá

integrar la función de densidad N (0,1). Estas integrales se encuentran resueltas y tabuladas. Por ejemplo, si Y ~ N (μ,σ^2) con μ = 10 y σ^2 = 4 y se desea conocer la P [8 $\leq$ Y $\leq$ 9] se procede de la siguiente manera:

a) Se estandariza de modo que: $z_1 = \frac{8-10}{2} = -1$ y $z_2 = \frac{9-10}{2} = -0.5$
b) Luego: P[8≤X≤ 9]=P[-1≤Z≤ -0.5] y se lee F(-1) y F(-0.5). Desde una tabla se leen las áreas asociadas a estos valores de Z y finalmente se restan esas areas, ya que una cuantifica la probabilidad de tener valores menores a 9 y la otra de tener valores menores a 8. Luego la diferencia entre ambas otorga la probabilidad de que un valor seleccionado al azar de la distribución de interés se encuentre entre 8 y 9.

La variable Z puede ser vista como una desviación de Y en torno a la media, medida en unidades de desviación estándar. Es decir P [-1 < Z < 1] debe entenderse como la probabilidad de que Y tome valores que se alejan de la media en menos o más una desviación estándar, es decir, P[μ-1σ<Y<μ+1σ].

En una distribución normal teórica, esta probabilidad es igual a 0.6827, lo que equivale a decir que en la distribución normal el 68.27% de las observaciones están comprendidas entre la esperanza menos un desvío estándar y la esperanza más un desvío estándar:

[$\mu \pm 1\ \sigma$] incluye al 68.27% de las observaciones

De igual manera se deduce que:

[$\mu \pm 2\ \sigma$] incluye al 95.45% de las observaciones

[$\mu \pm 3\ \sigma$] incluye al 99.74% de las observaciones

Existen pruebas formales para verificar la condición de normalidad como es la prueba de Shapiro Wilks y los gráficos QQ-plot. Más adelante en esta obra, se explicará cómo éstas pueden realizarse usando InfoStat.

Aplicación

Manejo de plantaciones

Una de las estrategias para determinar el manejo de bosques naturales se basa en la reducción de un porcentaje de los árboles presentes (raleo). Los árboles que se cortan son los de mayor diámetro. Si la distribución de los diámetros de los árboles sigue una distribución normal, con media 60 cm y varianza 144 cm^2.

a) ¿qué porcentaje de árboles se removerá si se talan todos los árboles con más de 70 cm de diámetro?
b) Si se quiere remover el 30% de los árboles, ¿cuál será el diámetro mínimo para cortar el árbol?

Estrategia de análisis

Graficaremos una distribución normal y demarcamos el área de interés. Usando el menú APLICACIONES $\Rightarrow$ DIDÁCTICAS $\Rightarrow$ GRÁFICOS DE FUNCIONES DE DENSIDAD

CONTINUAS de InfoStat se obtiene la siguiente ventana de diálogo, donde se deben colocar los parámetros de la distribución (60; 144) (Figura 3.12).

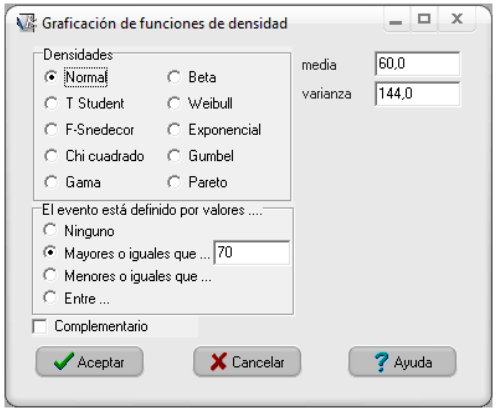


Figura 3.12. InfoStat. Ventana de diálogo para graficar una función de densidad normal con media 60 y varianza 144 y el área correspondiente con valores mayores a 70.

El software nos proporciona directamente la probabilidad de encontrar valores superiores a 70, P(Y>70cm) = 0,2023 (Figura 3.13).

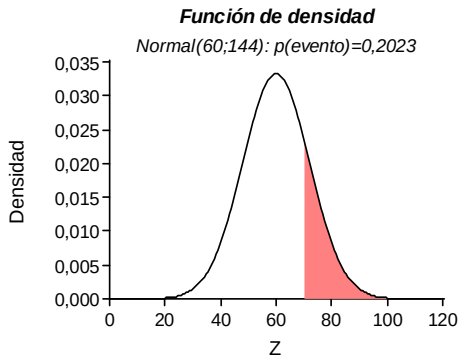


Figura 3.13. Función de densidad normal para los diámetros de árboles con la probabilidad del evento [Y > 70] representado por el área sombreada.

Para calcular esta probabilidad usando tablas, primero hay que estandarizar:

$$Z = \frac{70 - 60}{\sqrt{144}} \simeq 0{,}8333$$

Luego, P[Y Ó70] = 1-P [Y ≤ 70] = 1-P [Z ≤ 0,8333] = 1-0,7977 = 0,2023

Según los cálculos si se ralean árboles con diámetros mayores a 70 cm, se talará un 20% de los árboles presentes en el bosque. Para responder a la segunda pregunta, cuál será

el diámetro mínimo para cortar el árbol si se quiere remover el 30% de los árboles, debemos encontrar el valor de la variable por encima del cual se encuentra el 30% de los diámetros, es decir debemos hallar el percentil 70 o cuantil 0,70 de la distribución de los diámetros. Podemos hacer esto con el calculador de cuantiles y probabilidades de InfoStat del menú ESTADÍSTICAS ⇒ PROBABILIDADES Y CUANTILES, aparecerá una ventana de diálogo donde se deben ingresar los valores de los parámetros de la distribución y el cuantil que se desea calcular, en nuestro caso, $C_{0,70}$. Al presionar el botón **Calcular** tendremos la estimación del cuantil, en este caso X=66,29.

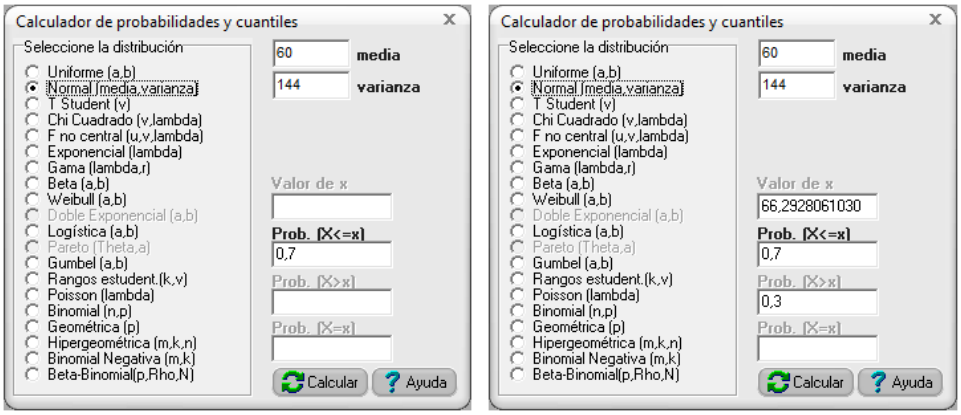
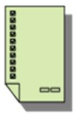


Figura 3.14. InfoStat. Ventana de diálogo para calcular probabilidades y cuantiles de una función de densidad normal para obtener el cuantil 0,70 de una distribución normal con media 60 y varianza 144. Resultado: 66,29

Variables aleatorias discretas

Distribución Binomial

La **distribución Binomial** puede usarse para el cálculo de probabilidades de eventos provenientes de conteos acotados. Se supone que se realizan cierto número (n) de experimentos aleatorios y en cada experimento se registra uno de dos resultados posibles, éxito o fracaso donde el éxito tiene una cierta probabilidad (P) de ocurrencia (este ensayo con resultado binario se conoce como **ensayo Bernoulli**). Se supone además que estos experimentos son independientes (es decir el resultado de un experimento no afecta al resultado de otro) y que la probabilidad de éxito (o fracaso) se mantiene constante a través del conjunto de experimentos. Interesa la variable aleatoria cantidad de éxitos en los *n* ensayos.



Como el número de ensayos es conocido podríamos usarlo como un denominador natural y expresar los valores de la variable de interés como porcentajes.

Por ejemplo, al tirar una moneda y observar el resultado este puede ser cara o cruz. Luego, la tirada de la moneda es un ensayo Bernoulli ya que los resultados posibles son dos, uno con probabilidad *p* y otro con probabilidad *q=1-p*. Si se considera éxito a la cara, la probabilidad de éxito es p=0,5. Si tiramos la misma moneda 20 veces y podemos pensar que cada tirada es un ensayo Bernouilli independiente, podríamos calcular probabilidades en relación a los valores de la variable aleatoria Y= número de caras en las 20 tiradas. Este tipo de variable, Y, donde se contabilizan los éxitos en una serie de ensayos Bernouilli independientes, cada uno con probabilidad de éxito p, tienen una distribución de probabilidades que ajusta al modelo Binomial. En este caso particular, al model binomial con parámetros n = 20 y P= 0,5.
La función de probabilidad de una variable aleatoria Y que se distribuye como una Binomial puede expresarse como:

$$f(y;n,P)=\begin{cases}\binom{n}{y}P^{y}(1-P)^{n-y} \;\; si\; y=0\;,1,...,n\\ 0 \quad en\ caso\ contrario\end{cases}$$

donde P es la probabilidad de éxito y por lo tanto pertenece al intervalo [0;1] y *n* es el número de ensayos Bernouille independientes.

Nota: $\binom{n}{x}$ representa el número de combinaciones posibles de armar en base a *n* elementos en grupos de *x*, siendo $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ y $n! = 1\times2\times...\times n$.

La E(Y) y la V(Y) cuando Y tiene distribución Binomial son:

$\mu = E(Y) = \Sigma_{y=0}^{n} y\, f(y) = nP$

$\sigma^2 = V(Y) = nP(1-P)$

Por ejemplo, si se tira 20 veces una moneda, y se quiere calcular la probabilidad de obtener 12 caras, es decir, P (Y = 12), esta será:

$$p(Y=12)=\binom{20}{12}0,5^{12}(1-0,5)^{20-12}=0,1201$$

La esperanza de la variable Y=número de lanzamientos que resultan en cara es igual a 20x0,5=10 y la varianza de Y es 20x0,5x0,5=5.
Las probabilidades pueden calcularse con la función o bien con software que incluyen la función Binomial o con tablas de la distribución (ver Tablas Estadísticas). Para ilustrar el uso de la función presentamos el siguiente ejemplo. Supóngase que se toman 10 semillas de *Panicum sp* y se registra el evento "germinó" o "no germinó" después de 5 días desde su implantación. En este experimento las semillas están suficientemente

aisladas como para asegurar respuestas independientes. Si la probabilidad de germinación es (para todas las semillas) igual a 0.25 calculemos:

c) Probabilidad que germinen 7 de las 10 semillas,
d) Probabilidad que germinen al menos 3 de las 10 semillas,
e) Probabilidad que germinen a lo sumo 5 semillas.
f) La esperanza de esta variable aleatoria.
g) La varianza.

Si Y~ *Bin*(7; 10, 0.25), luego:

a) $P(Y= 7) = \binom{10}{7} 0.25^7 (1\text{-}0.25)^{(10\text{ -}7)} =$

$$\binom{10}{7} 0.25^7 (1-0.25)^{10-7} = \frac{10!}{7!(10\text{ - }7)!}\ 0.25^7\ 0.75^3 = \frac{0.0185}{6} = 0.0031$$

b) $P(Y \geq 3) = P(Y = 3) + P(Y= 4) + ... + P(Y= 10) =$
$= 1 - (P(Y = 0) + P(Y= 1) + P(Y = 2)) =$
$= 1 - (0.0563 + 0.1877 + 0.2816) = 0.4744$

c) $P(Y \leq 5) = P(Y = 0) + P(Y = 1) + ... + P(Y = 5) =$
$= 0.0563 + 0.1877 + 0.2816 + 0.2503 + 0.1460 + 0.0584 = 0.9803$

d) $E(Y) = 10\ (0.25) = 2.5$

e) $V(Y) = 10\ (0.25)\ (1 - 0.25) = 1.875$

Para citar otro ejemplo (que resolveremos con software), supongamos que un criadero de semillas afirma que el poder germinativo de las semillas de un nuevo híbrido es del 98%. Un técnico decide poner a prueba esta afirmación, y para esto toma 100 semillas del hibrido en forma aleatoria y las coloca en bandejas de germinación lo suficientemente distanciadas como para pensar que cada semilla germina o no independientemente de las semillas vecinas. El técnico realiza la prueba siguiendo los protocolos de ensayos de germinación (cada uno se considera un ensayo Bernouilli) y encuentra que la cantidad de semillas germinadas es de 94.

- *¿Cuál es la probabilidad de la condición de verdad de la afirmación de la empresa vendedora?*

Para el cálculo de la probabilidad es necesario definir los parámetros de la distribución Binomial, que en este caso son *n*=100 (considerando que las semillas germinan independientemente unas de otras) y P= 0,98; luego calcular la P(Y≤94). El cálculo con el software InfoStat se hace siguiendo las instrucciones dada para otras distribuciones.
La probabilidad de obtener valores de poder germinativo menores o iguales a 94% es muy baja (P=0,0154), es decir solo el 1,5% de las veces que se realice este experimento se obtendrán 94 semillas germinadas o menos si es cierta la afirmación del vendedor.

Por la baja probabilidad calculada, aquí se podría deducir que la semilla del híbrido no tiene el poder germinativo que indica el vendedor.

Aplicación

Plagas cuarentenarias

Los mercados internacionales de productos agropecuarios para exportación tienen exigencia estrictas sobre la presencia de plagas cuarentenarias. Una plaga cuarentenaria es un plaga que no está presente en el país que importa productos, y por este motivo se establecen barreras de control y protección en los puertos de entrada. Así es el caso de la exportación de plantas ornamentales, donde un lote completo es rechazado si se encuentra solo una plaga cuarentenaria. Para el control de plagas los organismos de inspección toman muestras de plantas de cada uno de los contenedores que se intentan importar y examinan cuidadosamente cada planta de la muestra.

Se sabe que la probabilidad de éxito (encontrar la presencia de una plaga) en estas especie en nuestro país es P=0,01. Si se examinan 50 plantas, ¿cuál es la probabilidad de encontrar al menos una con la presencia de la plaga? ¿Cuál es la probabilidad de encontrar exactamente 2 plantas de las 50 con la plaga? ¿Cuál es la probabilidad de detectar al menos una planta con la plaga si la probabilidad de éxito del evento de interés cambiase a P=0,1?

Estrategia de análisis

Se observa que el número de plantas con plaga en este experimento está acotado, tienen un máximo. Ya que se realizan 50 observaciones, el máximo valor de la variable de interés es 50 (todas las plantas infectadas) y el mínimo 0 (ninguna infectada). Considerando que las extracciones y observaciones de cada una de las 50 plantas son independientes, es decir, la presencia de una plaga en una planta no depende de lo que sucede en las otras plantas muestreadas, se decide modelar a la variable Y=número de plantas con plaga con la distribución binomial, con parámetros n=50 y p=0,01.

Se desea calacular la probabilidad de encontrar al menos una planta con la presencia de la plaga, es decir, P [Y Ó 1]. Este cálculo se podría realizar sumando P [Y= 1] + P [Y = 2] +…+ P [Y= 50]. Pero es más fácil si se saca por diferencia:

$$P[Y\geq 1]=1-P[Y<1]=1-P[Y=0]$$

Usando el calculador de probabilidades y cuantiles de InfoStat, menú ESTADÍSTICAS ⇒ PROBABILIDADES Y CUANTILES, en la ventana de diálogo se establecen los parámetros de una binomial (50; 0,01) y el valor de Y=0 .

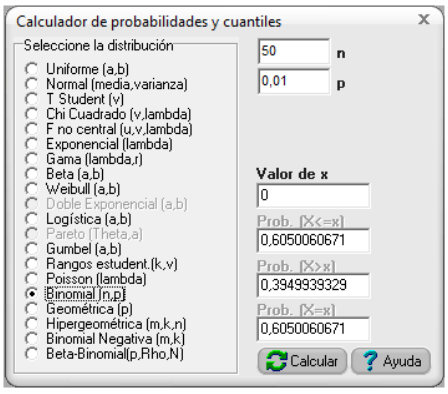


Figura 3.15. InfoStat. Ventana de diálogo para calcular probabilidades y cuantiles para una binomial (50; 0,01) con Y=0.

Al presionar el botón **Calcular** se observa que la P [Y ≤ 0] = P [Y = 0] =0,6050 y la P [Y > 0] = 0,3949.
Luego, P [Y ≥ 1] = 1 – P [Y = 0] = 1 – 0,6050 = 0,3949.
Para calcular la P [Y = 2], ponemos el valor 2 en el calculador de probabilidades Así, la P [Y = 2] = 0,07.
Para responder a la última pregunta, ¿cuál es la probabilidad de detectar al menos una planta con la plaga si la probabilidad de éxito cambia a P=0,1?, debemos cambiar los parámetros de la distribución binomial a (50; 0,1) y calcular esta probabilidad.
Se puede resaltar que por ser la distribución binomial una distribución para variables aleatorias provenientes de conteos (acotados por el número de ensayos Bernoulli) puede asumir como valores los números naturales incluido el cero (es decir, 0≤Y≤n).
La función permite observar que, si la entidad reguladora quiere tener mayor probabilidad de encontrar una plaga en cargamentos donde la probabilidad de éxito es baja, deberá trabajar con un *n* o tamaño de muestra mayor.
Podríamos preguntarnos entonces, cuál debería ser el tamaño de muestra a tomar si la probabilidad de éxito es 0,01 y se quiere tener una probabilidad de 0,80 de encontrar al menos una planta con plaga.
Para esto, usando el calculador de probabilidades establecemos el parámetro P=0,01 y aumentamos *n* hasta obtener una probabilidad de detección de 0,80
Entonces, si la probabilidad de éxito P=0,010, se deberán tomar muestras de tamaño 160 si se quiere tener una probabilidad de 0,7997 de detección de la plaga.

Distribución Poisson

La distribución de Poisson también sirve como modelo probabilístico para variables discretas de tipo conteo. A diferencia de la Binomial, donde el conteo se realizaba sobre *n* experimentos independientes, en el caso de la Poisson, los conteos se refieren al número de veces que un evento ocurre en una unidad de tiempo o espacio dada (hora, kilo, m^2, m^3, planta, etc.) y por tanto los valores de la variable no están acotados. Es

decir, mientras los valores de Y en una Binomial podían pertenecer a los naturales entre 0 y *n* inclusive, en el caso de una Poisson pueden pertenecer a los naturales entre 0 e infinito.

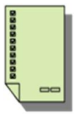
En Agronomía, la distribución Poisson suele usarse para modelar el número de insectos sobre una planta, o en un golpe de red, el número de manchas defectuosas en un mosaico, o en un metro cuadrado de piso, el número de colémbolos en 100 g de suelo, o en 1000 cm^3 de suelo o el número de coliformes en 1 ml de agua, entre otros conteos de interés.

La función de probabilidad de una variable aleatoria Y que se distribuye como una variable Poisson puede expresarse como:

$$f(y,\lambda) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!} \; si \; y = 0,1,2,... \\ 0 \; en \; caso \; contrario \end{cases}$$

Como puede observarse desde la función, el único parámetro de la distribución Poisson es λ. Si una variable aleatoria Y se distribuye como Poissson lo denotamos como: Y~ *Poisson*(λ). Esta distribución tiene un único parámetro, que representa la esperanza y también a la varianza, es decir que cuando Y~ *Poisson*(λ), se cumple:

$$\mu = E(Y) = \lambda$$
$$\sigma^2 = V(Y) = \lambda$$



La propiedad de esperanza igual a varianza de la distribución Poisson implica que al aumentar el promedio de los conteos, aumenta también su varianza. La varianza de una Poisson es así función de la media.

Para ejemplificar un cálculo de probabilidad bajo el modelo Poisson, supongamos que el número promedio de picaduras de gorgojo por semilla es 0.2 (es decir, por ejemplo que, en promedio, cada 100 semillas se cuentan 20 picaduras). El modelo Poisson podría ayudarnos a resolver estas preguntas ¿cuántas de 100 semillas no tendrán picaduras?, ¿cuántas 1 picadura? y ¿cuántas 2 o más?

Para responder se calcula la probabilidad de que una semilla tomada al azar tenga una picadura o ninguna picadura de la siguiente manera:

$$P(Y=0) = \frac{0.2^0 e^{-0.2}}{0!} = 0.819$$

$$P(Y=1) = \frac{0.2^1 e^{-0.2}}{1!} = 0.164$$

y P(Y>1)= 1-[P(Y=0)+P(Y=1)]=1-0.982= 0.018

En consecuencia, si la probabilidad de que una semilla tomada al azar no tenga picaduras es 0.819, deberíamos esperar que, en un grupo de 100, aproximadamente 82 no estén picadas, y si la probabilidad de que tengan solo una picadura es de 0.164,

entonces solo 16 semillas cumplirán esta condición y finalmente, aproximadamente 2 de cada 100 semillas tendrán 2 o más picaduras.

Para dar a otro ejemplo, supongamos que un comerciante que vende arroz fraccionado desea exportar su producto bajo la etiqueta de alta calidad; sin embargo, el producto será aceptado bajo esa denominación sólo si la cantidad de granos de arroz partidos no es mayor a 50 granos por kilo. El comerciante extrajo 50 muestras de 1 kg para determinar el número de grano partidos (Tabla 3.1).

Tabla 3.1: Resultados de calidad de arroz obtenidos a partir de 50 muestras

Número de granos partidos por kilo	Número de muestras con dicha cantidad
10	3
20	6
30	10
40	20
50	6
60	5

El propósito de este muestreo fue estimar el parámetro λ de esta distribución Poisson, que se calcula de la siguiente manera:

$\lambda = [(10 \times 3)+(20 \times 6)+(30 \times 10)+(40 \times 20)+(50 \times 6)+(60 \times 5)]/50 = 870/50 =$ **37**

Es decir, en promedio se esperan 37 granos partidos por kilogramo de arroz.

Una vez estimado el parámetro λ, podemos calcular probabilidades de ocurrencia de eventos bajo una distribución Poisson. Si se define Y como el número de granos partidos por kilo de arroz, podemos responder a las siguientes preguntas:

¿Cuál es la desviación estándar de Y para este comerciante?

La desviación estándar es la raíz cuadrada de la varianza, en este caso: $\sqrt{\lambda} = \sqrt{37} = 6,08$

Usando el software InfoStat para el cálulo de probabilidades, se dieron respuestas a las siguientes preguntas:

¿Cuál es la probabilidad de una partida de arroz con 50 granos partidos?

P(Y=50)=0,0072

¿Cuál es la probabilidad de una partida con más de 50 granos partidos?

P(Y>50) = 0.0167

¿Cuál es la probabilidad de obtener 10 granos partidos en una muestra?

P(Y=10) = $1,13 \times 10^{-7}$, es decir prácticamente cero

Si un exportador más exigente pide a lo sumo 10 granos partidos por kilo, ¿Cuál es la probabilidad de rechazo de la partida?

P(Y≥10) = 0,9999 es decir que, prácticamente con seguridad, la partida será rechazada.

La distribución Poisson facilita el cálculo de probabilidades de variables aleatorias que provienen de conteos no acotados; mientras que la distribución binomial asigna probabilidades a variable aleatorias que cuentan la cantidad de éxitos y donde el máximo de la variable está acotado por n, el número de observaciones de tipo éxito/fracaso que se realicen.

Aplicación

Manejo de acoplados de cosecha

Se conoce a través de registros históricos, que en un establecimiento que produce granos, durante la época de cosecha salen del establecimiento hacia la acopiadora, en promedio, cuatro acoplados con grano por hora. Para organizar el traslado de una nueva cosecha es necesario calcular:
¿Cuál es la probabilidad que salgan más de dos acoplados en media hora?
¿Cuál es la probabilidad que salgan como máximo seis acoplados en una hora?
¿Cuál es la cantidad de acoplados por hora que sólo podría ser superada por el 1% de las horas en observación?

Estrategia de análisis

Para responder a la primera pregunta debemos calcular la P(Y>2) usando una distribución Poisson con parámetro $\lambda = 2$, ya que la unidad de tiempo en la pregunta es la mitad de la unidad de tiempo en la que se expresó el parámetro lambda.
Para esto podemos valernos del calculador de probabilidades y cuantiles de InfoStat. Usando el menú ESTADÍSTICAS $\Rightarrow$ PROBABILIDADES Y CUANTILES, aparecerá una ventana de diálogo donde se debe ingresar el valor del parámetro lambda ($\lambda = 2$) luego de seleccionar la distribución Poisson y el valor 2 como valor de la de la variable (que en infoStat se denota como valor de X). El resultado que se muestra indica que P(Y > 2)=0,3233.
Para responder a la pregunta ¿Cuál es la probabilidad que salgan como máximo seis acoplados en una hora? Usaremos también el calculador de probabilidades y cuantiles de InfoStat pero con $\lambda = 4$ ya que la pregunta esta referida a una hora. Así, se observa que la (Y<=6)=0.8893.
Por último, la tercera consulta hace referencia a la identificación de un cuantil de la distribución y no al cálculo de una probabilidad; se desea conocer el cuantil 0.99 o percentil 99, es decir el valor de la variable tal que el 99% de los valores son menores o iguales a éste y por tanto sólo el 1% de valores de la variable superarán a éste que llamamos percentil 99. En el calculador de probabilidades y cuantiles de InfoStat, se debe ingresar el valor del parámetro ($\lambda = 4$) luego de seleccionar la distribución Poisson. No podemos ingresar el valor de la variable, porque justamente éste es nuestra incógnita, entonces ingresaremos información en las casillas para las cuales tengamos el

dato. Podemos ingresar 0,99 en el espacio reservado para Prob(X<=x) o bien el valor 0,01 en la casilla reservada para ingresar la proporción de valores mayores que la incognita. El resultado que se obtiene indica que 9 acoplados es el percentil 99 de la distribución, es decir sólo en un 1% de las horas de observación se espera que pasen más de 9 acoplados.

Definiciones

Definición 3.1: Variable aleatoria normal

Una variable aleatoria X se define como normalmente distribuida si su función de densidad está dada por:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

donde: los parámetros μ y σ satisfacen $-\infty \leq \mu \leq \infty$ y $\sigma > 0$
e = base de los logaritmos naturales (aprox: 2.7182818), π = constante matemática aproximada por 3.14159 y y $\in (-\infty, \infty)$.

Definición 3.2: Estandarización

Se llamará estandarización a la siguiente transformación:

$$Z = \frac{y-\mu}{\sqrt{\sigma^2}}$$

donde :Z: es la variable aleatoria obtenida de la transformación
Y: la variable aleatoria original
μ y σ^2 son respectivamente, la esperanza y la varianza de la distribución de Y.

Definición 3.3: Distribución Binomial.

Una variable aleatoria Y tiene distribución Binomial si y sólo si su función de densidad, con 0<P<1, es:

$$f(y;n,P) = \begin{cases} \binom{n}{y} P^y (1-P)^{n-y} \ si \ y = 0,1,\ldots,n \\ 0 \quad caso \quad contrario \end{cases}$$

Definición 3.4: Distribución Poisson.

Una variable aleatoria Y tiene distribución Poisson si y sólo si su función de densidad es:

$$f(y,\lambda) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!} \ si \ y = 0,1,2,\ldots \\ 0 \, caso \, contrario \end{cases}$$

Ejercicios

Ejercicio 3.1: Uso de la tabla de cuantiles de la Distribución Normal Estándar

Esta tabla presenta 2 columnas: La primera columna se refiere a la distancia desde un valor a la media medida en número de desviaciones típicas (valores de la variable Z). Por ejemplo el valor 1 indica 1 DE por encima de la media y el valor -1.7 corresponde a 1.7 DE por debajo de la media. La segunda columna contiene el área bajo la curva normal entre $-\infty$ y el valor correspondiente a la primer columna, es decir el valor de la función de distribución normal acumulada. Por ejemplo para el valor 1 de z, el área asociada es 0.8413. Así se puede concluir que la probabilidad de que una variable distribuida normalmente con $\mu = 0$ y $\sigma^2 = 1$ tome valores iguales o menores que 1, es igual a 0.8413 y también se puede decir que el valor 1 es el cuantil 0.8413 de la distribución normal estándar.

Usando la tabla de cuantiles de la Distribución Normal Estándar obtener las siguientes probabilidades:

a) $P(Z \leq 1.3)$ b) $P(Z \leq 4)$ c) $P(Z \geq 1.3)$
d) $P(-1 \leq Z \leq 1)$ e) $P(0.5 \leq Z \leq 1)$ f) $P(Z = 1)$

Ejercicio 3.2: Si X es una variable aleatoria distribuida normalmente con $\mu = 10$ y $\sigma^2 = 4$.

a) ¿Cuál es la probabilidad de que X tome valores menores que 9?.
b) ¿Cuál es la probabilidad de que X tome valores entre 9 y 11?.

Ejercicio 3.3: La variable altura de plántulas para una población dada se distribuye normalmente con media $\mu = 170$ mm y $\sigma = 5$ mm. Encontrar la probabilidad de los siguientes eventos:

a) Plantas con alturas de al menos 160 mm.
b) Plantas con alturas entre 165 y 175 mm.

Ejercicio 3.4: Si la variable espesor de un sedimento en un sustrato de suelo, se distribuye normalmente con media $\mu = 15$ micrones y desviación estándar $\sigma = 3$ micrones.

a) ¿Cuál es el cuantil 0.75 de la distribución de la variable?.
b) ¿Cómo se interpreta este valor?.

Ejercicio 3.5: El caudal de un canal de riego medido en m^3/seg es una variable aleatoria con distribución aproximadamente normal con media 3 m^3/seg. y desviación estándar 0.8 m^3/seg. A partir de estas referencias calcular la probabilidad de los siguientes eventos:

a) Evento A: que el caudal en un instante dado sea a lo sumo de 2.4 m^3/seg.
b) Evento B: que el caudal en un instante dado esté entre 2.8 y 3.4 m^3/seg.

Ejercicio 3.6: La cantidad de microorganismos que tiene un mililitro de leche determina su calidad. Un establecimiento lácteo recibe diariamente leche , con Unidades Formadoras de Colonias (UFC) de micrrorganismos que se suponen se distribuyen normalmente con un promedio de bacterias de 75 UFC/ml y varianza de 200 (UFC/ml)^2. La leche 70 UFC/ml o menos se usa para consumo fresco, la leche con más de 85 se usa para fabricar leche en polvo, y la leche con calidad intermedia se usa para fabricar quesos. Si la empresa recibe 300000 l por día:

a) ¿Qué cantidad de litros se usan para consumo fresco, queso y leche en polvo?

Ejercicio 3.7: El espesor de la cáscara del huevo determina la probabilidad de ruptura desde que la gallina lo pone hasta que llega al consumidor. El espesor, medido en centésimas de milímetro, se distribuye normal y se sabe que: se rompen el 50 % de los huevos con espesor de cáscara menor a 10 centésimas de mm (cmm). Se rompen el 10 % de los huevos cuyo espesor de cáscara está comprendido entre 10 y 30 cmm. No se rompen los huevos con espesor de cáscara mayor de 30 cmm. Si en un establecimiento avícola la media del espesor de cáscara es de 20 cmm y la desviación estándar de 4 cmm:

a) ¿Cuántos, de los 5000 huevos que se producen diariamente, llegan sanos al consumidor?

Ejercicio 3.8: Una empresa exportadora de manzanas necesita encargar 10000 cajones para el embalaje de la fruta. Sin embargo, no todos los cajones son iguales ya que sus especificaciones dependen de la calidad del producto envasado. Así, de acuerdo al diámetro de la manzana se identifican 3 categorías de calidad.

Categoría I: manzanas cuyo diámetro es menor de 5 cm

Categoría II: manzanas cuyo diámetro está comprendido entre 5 y 7 cm

Categoría III: manzanas cuyo diámetro es mayor que 7 cm

Las frutas de mayor calidad son las correspondientes a la categoría II por su tamaño y homogeneidad. Si la distribución del diámetro de las manzanas puede modelarse bien mediante una distribución normal con media $\mu = 6.3$ y varianza $\sigma^2 = 2$, responder:

a) ¿Cuántos cajones se necesitarán para cada categoría de manzanas?

Ejercicio 3.9: Siguiendo con el ejercicio anterior y conociendo el comportamiento cíclico de la demanda de cada categoría de manzanas, se sabe que en la presente campaña va a tener más demanda la manzana de la categoría II (manzanas con diámetro entre 5 y 7 cm), con lo cual las ganancias para el exportador se maximizarían en caso de aumentar el volumen de la cosecha para esta categoría. Una forma de regular el tamaño final de esta fruta es mediante la eliminación temprana de los frutos en formación (raleo). Si se eliminan muchos frutos el tamaño final de las manzanas será mayor que si se eliminan pocos o ninguno.

La experiencia ha permitido establecer las características distribucionales del diámetro final de las manzanas bajo dos estrategias de manejo:

A: no eliminar ningún fruto

B: eliminar 1 de cada 3 manzanas

La estrategia A produce frutos con diámetros distribuidos N (6.3, 2.0) y la estrategia B produce frutos con diámetros distribuidos N (6.8, 0.9).

a) ¿Cuál de las dos estrategias produce mayor proporción de frutos de Categoría II?

Ejercicio 3.10: Por medio de un tamiz de malla de 8 mm de diámetro se zarandean 8000 granos de maíz. El diámetro del grano de maíz sigue una distribución normal con esperanza igual a 9 mm y una desviación estándar de 1.2 mm.

a) ¿Qué proporción de granos serán retenidos por el tamiz?.
b) ¿ Qué proporción de granos no retenidos, serán retenidos por un tamiz de diámetro de malla igual a 7.5 mm?.
c) ¿Qué proporción de granos pasará a través de los dos tamices?.

Ejercicio 3.11: Un fitomejorador desea controlar la variabilidad de los brotes comerciales de espárrago, ya que las normas de embalaje establecen una longitud máxima de cajas de 23.5 cm. Suponiendo que la longitud de los brotes de este cultivo se distribuye normalmente, con una esperanza igual a 21 cm:

a) ¿Cuál debería ser el valor de la desviación estándar del carácter longitud del brote, para que la probabilidad de que existan espárragos que no puedan ser embalados, no sea mayor a 0.05?.

Ejercicio 3.12: Si la variable callos enraizados en cajas de Petri, donde se colocan 5 callos por caja, tiene una distribución binomial con p=0.20

Cantidad de callos enraizados en cajas de Petri	Probabilidad
0	0.32768
1	0.40960
2	0.20480
3	0.05120
4	0.00640
5	0.00032

Preguntas:

a) ¿Cuál es su valor esperado y su varianza?
b) ¿Cuál es la P(X <4)?
c) ¿Cuál es el valor de P(2 <X< 5)?

Ejercicio 3.13: La proporción de productores hortícolas orgánicos en una región es de 0,30. Si un técnico desea realizar una encuesta sobre técnicas de producción orgánica:

a) ¿Qué probabilidad tiene de encontrar al menos 5 productores orgánicos luego de entrevistar a 15?
b) ¿Cuántos campos deberá visitar si desea realizar al menos 10 encuestas a productors hortícolas orgánicos?

Ejercicio 3.14: Un dosificador de producto fitosanitario libera producto a un promedio de 10 gotas por minuto

Preguntas:

a) ¿Cuál es la probabilidad que se liberen menos de 6 gotas en un minuto?
b) ¿Cuál es la probabilidad de que se liberen como máximo 3 gotas en un minuto?
c) ¿Cuál es la probabilidad de que se liberen las 10 gotas en medio minuto?
d) ¿Cuál es la probabilidad que no salga ninguna gota en un periodo de 15 segundos?

Ejercicio 3.15: La transferencia embrionaria en vacas puede ser exitosa con probabilidad 0.70 o no exitosa. Si se selecciona un lote de 10 animales al azar entre aquellos lotes que recibieron transferencia embrionaria,

Preguntas:

a) ¿Qué modelo de distribución de probabilidades puede usarse para calcular probabilidades?
b) ¿Cuantas vacas del lote se espera hayan tenido una transferencia exitosa?
c) ¿Cuál es la probabilidad de lograr una transferencia exitosa en los 10 animales del lote?

Ejercicio 3.16: Un Ingeniero Agrónomo del Servicio de Alerta contra Fitóftora de una región viñatera afirma que 2 de cada 10 lotes afectados por la enfermedad se deben al mal manejo de los mismos. Cuál es la probabilidad de que:

a) en 100 lotes, a lo sumo 10, sean afectados por la enfermedad, por problemas de mal manejo
b) de 100 lotes, ninguno presente la enfermedad por problemas de mal manejo

Ejercicio 3.17: Se quiere encontrar plantas de trigo con propiedades resistentes a los pulgones. Un síntoma de resistencia es la ausencia de pulgones en la planta. Se calcula que la frecuencia de plantas sin pulgones en un cultivo es de alrededor de 1/200 pero solo 1 de cada 10 de estas plantas presentan genes de resistencia.

a) ¿cuántas plantas de trigo deberán revisarse para tener una probabilidad de al menos 0.95 de encontrar una con los genes de resistencia?

Ejercicio 3.18: En una red de computadores asociados a estaciones agroclimatológicas y dedicadas a transmitir la información registrada a un computador central (servidor) vía telefónica, el 1.4% de los llamados desde los computadores al servidor dan ocupado. Determinar la probabilidad de que de 150 intentos de comunicaciones (llamados) sólo en 2 casos el servidor de ocupado.

Ejercicio 3.19: Un técnico en semillas desea inspeccionar el funcionamiento de 20 cámaras de cría. Para esto toma dos cámaras al azar y registra la temperatura de las mismas. Si estas dos cámaras funcionan correctamente, el grupo de 20 será aceptado. Cuáles son las probabilidades que tal grupo de 20 cámaras sea aceptado si contiene:

a) 4 cámaras con registros de temperaturas no adecuadas
b) 8 cámaras con registros de temperatura no adecuadas
c) 12 cámaras con registros de temperaturas no adecuadas

Capítulo 4

Distribución de estadísticos muestrales

Margot Tablada

4. Distribución de estadísticos muestrales

Motivación

En numerosas situaciones deseamos utilizar los resultados del análisis de datos muestrales para elaborar conclusiones que puedan ser extendidas a la población de la que proviene la muestra. A este proceso inductivo se lo denomina **Inferencia Estadística**.

Si la muestra es una *ventana* a través de la cual observamos a la población podemos asegurar que aquello que vemos en la muestra está presente en la población; pero no podemos decir que aquello que no vemos, no está presente. Esto sugiere que si toda muestra contiene una parte de la población, dos muestras de una misma población podrían "mostrar" cosas diferentes e inclusive puede que la diferencia sea muy grande. ¿Cómo decidir en qué muestra confiaremos? ¿Podemos otorgar una medida de confiabilidad al cálculo obtenido en una muestra, para así establecer una medida del error potencial que podríamos tener al concluir sobre la población, de la mano de la muestra?

Vemos que inferir acerca de una población en base a lo observado en solo una de las posibles muestras, implica riesgo: el riesgo de concluir erróneamente por haber seleccionado una muestra que no represente adecuadamente a la población, ya que existe la posibilidad de que la estimación no sea buena por errores aleatorios debidos al muestreo. En este sentido, se hace necesario conocer el comportamiento de los estadísticos obtenidos en las posibles muestras; es decir, conocer su **distribución en el muestreo**.

En este capítulo abordaremos las distribuciones de los estadísticos media muestral y varianza muestral y el Teorema Central del Límite, que da sustento a las conclusiones que se obtienen en los estudios que se realizan con muestras.

Conceptos teóricos y procedimientos

La Inferencia Estadística hace referencia a un conjunto de procedimientos que, mediante el uso de estadísticos muestrales, permiten elaborar conclusiones sobre parámetros poblacionales desconocidos. Conocer o estimar a un parámetro de la distribución de una variable es posible a través de un estadístico. Dado que un estadístico será obtenido a partir de una muestra, es claro imaginar que hay más de una muestra posible de ser elegida y que entonces el valor del estadístico dependerá de la muestra seleccionada. Los valores de los estadísticos cambian de una muestra a otra. Interesa entonces tener una medida de estos cambios para cuantificar la medida del error en el que podría incurrirse al hacer una inferencia.

Distribución de estadísticos

Hemos señalado que el estudio de una muestra se realiza con el fin de concluir sobre la población de la cual ésta proviene. A los fines de presentar conceptos teóricos de distribución en el muestreo, haremos un muestreo cuyos resultados podamos visualizar fácilmente. Para ello, supongamos que contamos con una población finita de valores que puede asumir una variable aleatoria μ y, por razones de simplicidad para el desarrollo y presentación de resultados, supongamos que los valores en la población son: 1; 3; 5; 7 y 9, de modo que N=5. Caractericemos la distribución de la variable y veamos si al trabajar con muestras, podemos aproximarnos a esa distribución. Aproximarnos a la distribución implica poder conocer o **estimar** los parámetros de la distribución de la variable. La idea es utilizar información de la muestra, que pueda representar a los parámetros.

Para caracterizar a la distribución de la variable Y, podemos realizar un gráfico y calcular el valor de la esperanza (μ) y de la varianza (σ^2) de la variable aleatoria (Figura 4.1).

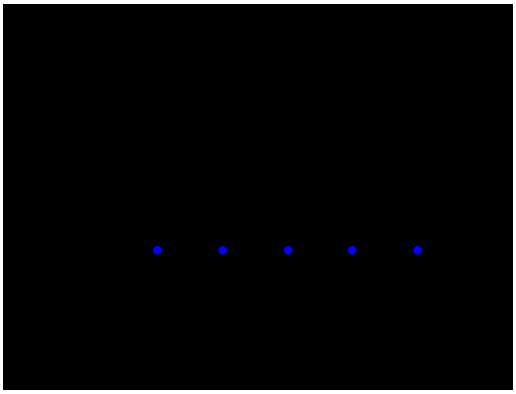


Figura 4.1. Distribución de la variable aleatoria Y, con $\mu = 5$ y $\sigma^2 = 8$

Distribución de la media muestral

Señalamos que mediante la observación de una muestra podemos aproximarnos a lo que ocurre en la población. Entonces, la media calculada con los valores observados en una muestra de tamaño *n*, ¿puede estimar a la media de la población de la cual fue extraida la muestra? Para responder a esto, desde la población propuesta, tomemos muestras de tamaño *n*=2 en un muestreo con reposición y en cada muestra calculemos su media (Tabla 4.1).

Tabla 4.1: Valores que conforman las muestras y medias muestrales, de 10 muestras de tamaño n=2 obtenidas en un muestreo con reposición desde una población finita

Muestra	Valores en la muestra	Media	Muestra	Valores en la muestra	Media
1	9; 1	5	6	5; 7	6
2	3; 5	4	7	1; 3	2
3	7; 1	4	8	3; 1	2
4	7;1	4	9	3; 5	4
5	9;9	9	10	5;9	7

El valor de la media muestral varía entre aquellas muestras que están conformadas por diferentes valores de la variable. Podemos pensar, entonces, que **la media muestral es una variable**. A su vez, vemos que hay muestras cuyas medias son valores más próximos a la media poblacional ($\mu = 5$) que los obtenidos en otras muestras. Además, las 10 muestras presentadas no son todas las posibles muestras de tamaño 2 que se podrían obtener desde la población propuesta. Estas consideraciones nos hacen notar que usar la media de una muestra de tamaño *n* para aproximarnos al valor de μ, involucra la necesidad de conocer el comportamiento de las medias que se obtendrían con las muestras de tamaño *n*, es decir, conocer la distribución del estadístico (variable aleatoria) media muestral.

Para estudiar la distribución de la variable aleatoria media muestral, consideremos todas las muestras posibles de tamaño *n*=2, que se podrían obtener desde la población propuesta haciendo un muestreo con reposición. Hay 25 muestras posibles.

A continuación se listan los valores que conforman cada muestra de tamaño *n*=2, indicando la media de cada muestra ($\bar{y}$).

Valores en la muestra	$\bar{y}$	Valores en la muestra	$\bar{y}$	Valores en la muestra	$\bar{y}$	Valores en la muestra	$\bar{y}$	Valores en la muestra	$\bar{y}$
1;1	1	3;3	3	5;1	3	7;1	4	9;1	5
1;3	2	3;1	2	5;3	4	7;3	5	9;3	6
1;5	3	3;5	4	5;5	5	7;5	6	9;5	7
1;7	4	3;7	5	5;7	6	7;7	7	9;7	8
1;9	5	3;9	6	5;9	7	7;9	8	9;9	9

Dado que tenemos todos los posibles valores de la media muestral, podemos tabular y graficar la distribución de la **variable aleatoria media muestral** ($\bar{Y}$) como se muestra en la Figura 4.2.

$\bar{Y}$	f($\bar{y}$)
1	1/25= 0,04
2	2/25= 0,08
3	3/25= 0,12
4	4/25= 0,16
5	5/25= 0,20
6	4/25= 0,16
7	3/25= 0,12
8	2/25= 0,08
9	1/25= 0,04

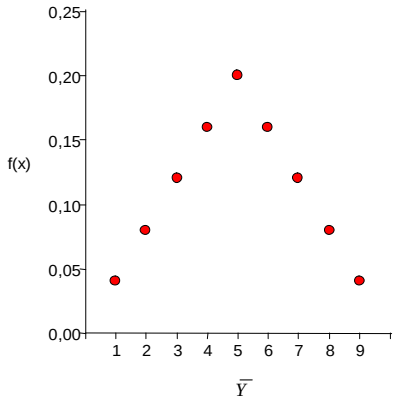


Figura 4.2: Distribución de la variable aleatoria media muestral en muestras de tamaño n=2 con reemplazo

La distribución señala que son más probables (más frecuentes) los valores de media muestral cercanos a 5. Calculemos la esperanza ($\mu_{\bar{y}}$) y la varianza ($\sigma^2_{\bar{y}}$) de la distribución:

$$\mu_{\bar{y}} = 5 = \mu \qquad \text{y} \qquad \sigma^2_{\bar{y}} = 4$$

Vemos que:

a) el promedio de la media muestral tiene igual valor que la media de la población de la que se extrajeron las muestras.

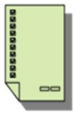
Cuando se señala que "la media muestral es un estimador insesgado de la media poblacional ", se hace referencia a la condición $\mu_{\bar{y}} = \mu$.

b) la varianza de la media muestral no es igual a la varianza de la población muestreada. Sin embargo, si dividimos a la varianza poblacional por el tamaño de la muestra *n*=2, obtenemos el valor de la varianza de la media muestral.

$$\sigma^2_{\bar{y}} = 4 = \frac{\sigma^2}{n} = \frac{8}{2}$$

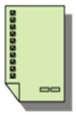
A la raíz cuadrada de $\sigma^2_{\bar{y}}$: $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ se la denomina **error estándar (EE)**.

La igualdad $\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$, se verifica con muestras obtenidas en poblaciones infinitas o desde poblaciones finitas en las que se hace muestreo con reemplazo.
Para el muestreo sin reemplazo en poblaciones finitas al calcular $\sigma^2_{\bar{y}}$ se debe usar un factor de corrección, de modo que $\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$, con N=tamaño de la población.



El error estándar es una medida de confiabilidad de las medias muestrales. A veces se confunde con la desviación estándar, pero ahora sabemos que son estadísticos diferentes. Mientras que la desviación estándar representa los desvíos de los valores de una variable respecto de su media, el error estándar representa los desvíos de los valores de la variable media muestral respecto de la media poblacional.

La distribución de la media muestral caracterizada por los parámetros $\mu_{\bar{y}}$ y $\sigma^2_{\bar{y}}$, se muestra simétrica y está claro que su varianza decrece si aumenta el tamaño de la muestra. Este aspecto es muy importante ya que en una distribución con menor varianza los datos se concentran más alrededor de la media. Esto nos lleva a pensar que con muestras de mayor tamaño, la media muestral sería un estimador más **preciso** de μ.



Si bien el aumento del tamaño muestral produce menor varianza en la distribución de las medias muestrales, puede ocurrir que a partir de cierto valor los cambios en esa varianza no sean relevantes.

Identificando un modelo de distribución para la media muestral

Dado que la media muestral varía de muestra de muestra, sería importante poder identificar un modelo de probabilidad que represente a la distribución de la variable media muestral, ya que con ello podríamos calcular errores en los que se podría incurrir cuando se usan las medias muestrales para realizar inferencia estadística.

Para poder visualizar el ajuste de un modelo de distribución a un conjunto de medias muestrales y las implicancias del **tamaño muestral** en la distribución de las medias muestrales, supongamos una población de pesos de pollos a la faena, con datos suficientes como para obtener una cantidad importante de muestras, ya que utilizaremos un muestreo sin reemplazo. Los datos, para seguir esta ilustración, se encuentran en el archivo [*faena]*.

En primera instancia, visualicemos la distribución de los valores poblacionales y obtengamos medidas resumen (Figura 4.3).

Medidas resumen

Resumen	peso
n	1000,00
Media	3135,68
D.E.	148,30
Var(n)	21970,02
CV	4,73
Mín	2652,25
Máx	3562,90

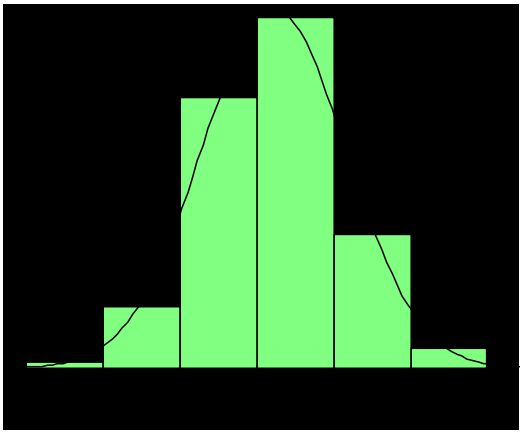


Figura 4.3: Histograma y medidas resumen de la distribución poblacional de pesos de pollos a la faena. Se ha superpuesto el polígono de frecuencias correspondiente al ajuste de un modelo de distribución normal

Observemos que los valores de peso se encuentran entre 2652,25 g y 3562,9 g. Por redondeo a un valor entero, la esperanza de la distribución es μ**= 3136** g y la varianza es σ^2**=21970** g^2; el coeficiente de variación corresponde a un 5%.

La forma de la distribución sugiere que el modelo de distribución Normal sería una buena aproximación. El modelo de la distribución Normal establece que el 95% de los valores de la variable se concentran alrededor de μ a una distancia de 1,96 veces el desvío estándar. Suponiendo este modelo, un 95 % de los pesos concentrados alrededor de μ se encontrarían, por redondeo, entre 2845 g y 3426 g como lo muestra la Figura 4.4.

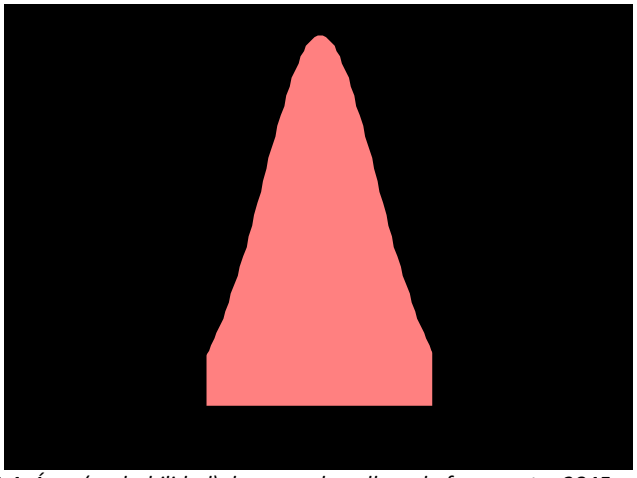


Figura 4.4. Área (probabilidad) de pesos de pollos a la faena entre 2845 g y 3426 g

Los valores 2845 g y 3426 g han sido obtenidos considerando la desviación estándar de la población (148,22 g), de modo que a una distancia de 290,51 g (esto es, 1,96 × 148,22 g) hacia ambos lados de la media μ (o sea, entre 3136 g - 290,51 g=2845 g y 3136 g + 290,51 g= 3426 g), encontramos un 95% de las realizaciones de esta variable aleatoria. Esto indica que valores de peso menores a 2845 g o superiores a 3426 g son poco probables, ya que ocurrirían solo en un 5% del total de pollos.

Vemos que considerando la desviación estándar podemos establecer un intervalo de valores entre los cuales se encuentra el promedio poblacional. De acuerdo a cuántas unidades de DE consideremos, abarcaremos un determinado porcentaje de valores de la variable, que están próximos a μ. De modo similar al planteado, podríamos obtener el conjunto de pesos que se concentran en un 99% alrededor de μ, en cuyo caso los valores se encontrarían a 2,576 veces la DE.

Siguiendo un análisis similar al que hemos presentado para los datos de la población de pollos, y dado que no alimentaríamos a todos los pollos con el suplemento sino a una muestra de ellos, a través de lo que obtengamos en una muestra elegida al azar:

- *¿cómo podemos aproximarnos al valor de μ utilizando la media muestral?, ¿lo que observamos en la muestra elegida ocurrirá en cualquiera de las posibles muestras?*
- *dado que el **error estándar** (**EE**) indica la variabilidad de la media muestral y que su valor depende del tamaño de la muestra ¿por qué decimos que es una medida de **confiabilidad**?*

Visualicemos la distribución en el muestreo y respondamos estos interrogantes. Para ello, realicemos sucesivos muestreos tomando 100 muestras de tamaños *n*=5, *n*=10, *n*=15 y *n*=30. En el programa InfoStat, seleccionamos en el menú *Aplicaciones*, la opción *Didácticas y, luego,* la opción *Remuestreo* (Figura 4.5).

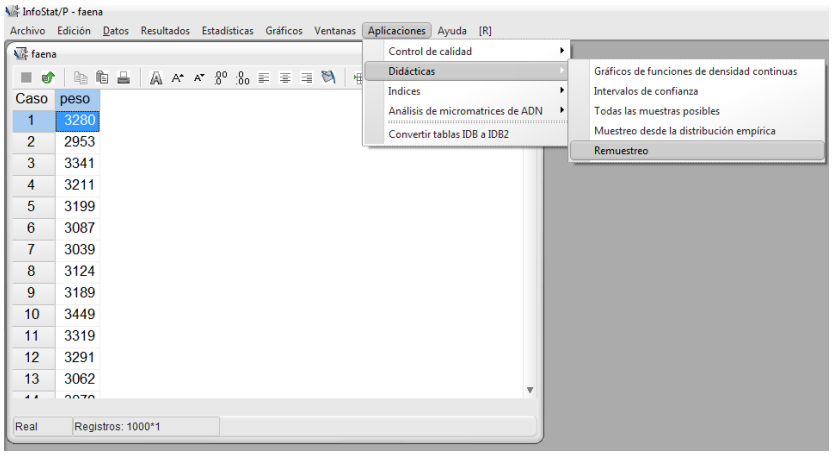


Figura 4.5. Ventana de diálogo con el archivo faena y el acceso a la aplicación Remuestreo

En la ventana de diálogo de Remuestreo se debe indicar a la columna "peso" como la que contiene los datos de la población a muestrear. Al *Aceptar*, aparece una ventana en la cual indicaremos el número de muestras a extraer y el *tamaño muestral* para hacer el muestreo. Para que se ejecute el remuestreo se presiona *Aceptar* (Figura 4.6).

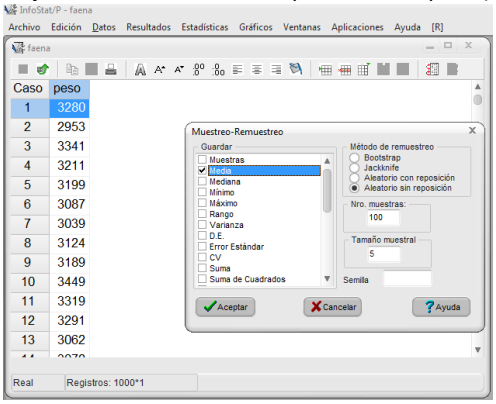


Figura 4.6. Ventana de diálogo de la opción Remuestreo. Se ejemplifica la obtención de las medias de 100 muestras de tamaño n=5, a partir de una población con N=1000

Como resultado del muestreo se generará una tabla que contendrá la identificación, el tamaño y la media, de cada muestra (Figura 4.7).

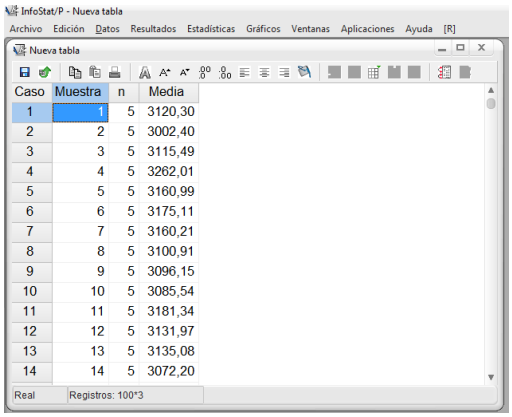


Figura 4.7. Tabla generada con las medias de 100 muestras de tamaño n=5

Para hacer los muestreos con los diferentes tamaños de muestra debemos repetir el procedimiento tantas veces como tamaños muestrales vayamos a utilizar. Obtendremos tantas tablas nuevas, como diferentes tamaños muestrales usemos.

Con los datos de cada muestreo, podemos graficar las diferentes distribuciones empíricas mediante histogramas. Al construir un histograma tenemos disponible una opción que permite ajustar la distribución a distintos modelos de probabilidad. Obtenido un histograma, pediremos un *ajuste Normal* (Figura 4.8).

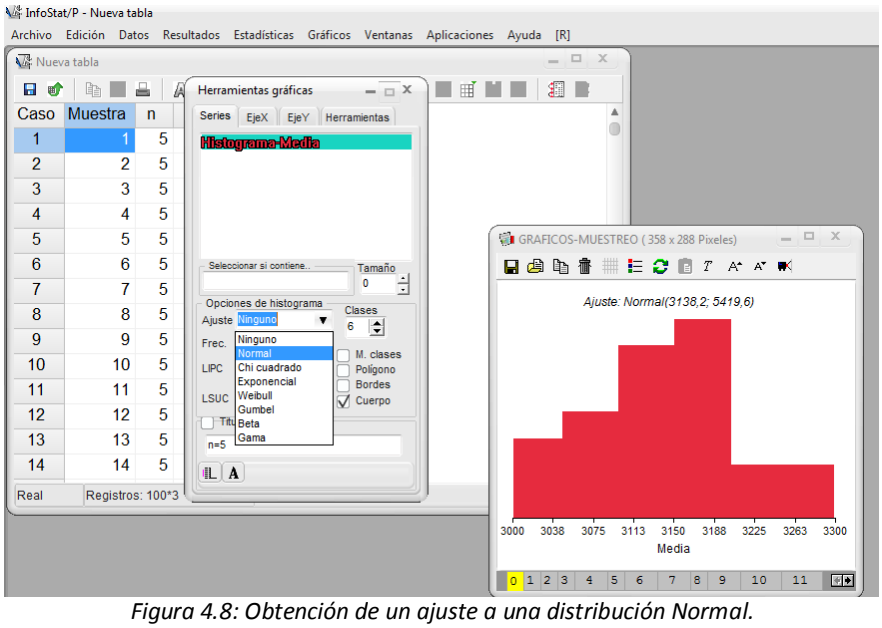


Figura 4.8: Obtención de un ajuste a una distribución Normal.

Al hacer un ajuste en el gráfico se informará, en un cuadro de texto, sobre el tipo de ajuste y los valores estimados para los parámetros de la distribución ajustada.
En la Figura 4.9 se muestran las distribuciones de la variable media muestral y las estimaciones de los parámetros para el ajuste solicitado. A los fines de mejorar la presentación se han modificado atributos de los gráficos (como la omisión del eje Y, entre otros), usando opciones de la ventana de *Herramientas gráficas* que acompaña a cada gráfico.

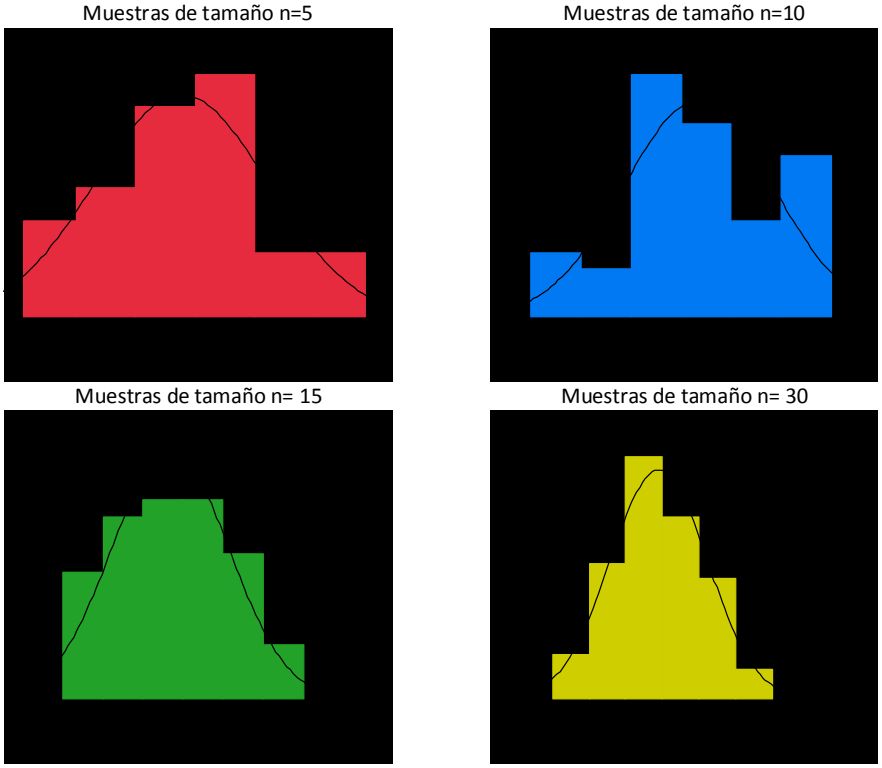


Figura 4.9: Histogramas de frecuencias relativas de la variable media muestral (correspondientes a pesos en gramos) de muestras extraídas desde una misma población utilizando diferentes tamaños muestrales. A cada histograma se le superpone el polígono de frecuencias relativas acumuladas que correspondería si los datos siguieran una distribución Normal.

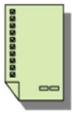
¿Qué podemos observar en los histogramas?
Retomemos uno de los interrogantes que planteamos anteriormente:

- *¿Cómo podemos aproximarnos al valor de μ utilizando la media muestral?, ¿lo que observamos en la muestra elegida ocurrirá en cualquiera de las posibles muestras?*

El valor de la media en cada histograma es prácticamente el mismo, sin importar el tamaño muestral, y es muy próximo al de la media de la población original (3136 g), tal como esperaríamos, puesto que $\mu_{\bar{y}} = \mu$. Esto ocurre porque las frecuencias de aquellas medias muestrales que son menores que μ están "en equilibrio" con las frecuencias de las medias que son mayores que μ. Sin embargo las distribuciones no son iguales.

Las distribuciones tienen diferentes rangos de variación y a medida que aumenta el tamaño muestral, dicho rango disminuye. Por esto, la distribución se vuelve menos aplanada a medida que el tamaño de la muestra aumenta. Con *n*=5 los pesos promedios varían entre 3000 g y 3300 g; con *n*=30 el rango de variación es entre 3075 g y 3200 g.

Es claro que al aumentar el tamaño de la muestra la varianza de la distribución de las medias muestrales es menor. No todas las medias muestrales tienen un valor próximo a μ, pero al tomar muestras de tamaño grande se observa que mayor cantidad de valores son cercanos a μ. Este ejemplo lleva a pensar que para estimar a μ, la media de una muestra de tamaño 30 sería más **confiable** que la obtenida con una muestra de tamaño 5. Por otro lado, a mayor tamaño muestral, mejora el ajuste a la distribución normal.



Podríamos preguntarnos ¿qué tamaño muestral es lo suficientemente grande para garantizar que la media muestral tendrá distribución aproximada a la normal? No hay un tamaño determinado; éste depende de la distribución original desde la que se obtienen las muestras. Cuanto más se aproxime la distribución original a una normal, menor será el tamaño muestral necesario para que la distribución de la media muestral sea normal, pero independientemente de la forma de la distribución original de los datos, la distribución de las medias muestrales tiende al modelo Gaussiano conforme aumenta el tamaño muestral.

Si partimos de una población cuya distribución no es normal, al tomar muestras de tamaño suficientemente grande la media muestral tiende a distribuirse normalmente con esperanza igual a la esperanza de la población original y varianza igual a la varianza de la población original, dividida por el tamaño de muestra considerado. Por lo tanto si queremos calcular probabilidades para eventos de la distribución de $\bar{Y}$, podríamos utilizar el procedimiento de estandarización y calcular el área, que corresponde a la probabilidad en cuestión, bajo una curva N(0;1).

El hecho de relacionar la distribución de la media muestral con una distribución Normal (0;1) cuando el tamaño muestral aumenta, se ha enunciado en un teorema conocido como **Teorema Central del Límite** (TCL).

El TCL se refiere a la distribución de la variable $Z = \frac{\left(\bar{Y} - \mu\right)}{\sigma / \sqrt{n}}$. Cuando n tiende a infinito, la variable Z tiende a una distribución N(0;1). Tomando el ajuste al modelo normal para la distribución de las medias de muestras de tamaño n=30, la media poblacional es estimada en 3137 g y el EE es de 27,4 g. El 95% de las medias muestrales

se encontrarán a 53,7 g (esto es, 1,96 × 27,4 g), tanto por debajo como por encima del valor de μ, o sea entre 3083 g y 3191 g. Valores fuera de este rango pueden ocurrir pero ello es poco probable (solo en un 5% de las muestras). La Figura 4.10 muestra este comportamiento y el que fuera obtenido para la distribución original de la que se extrajeron las muestras.

Figura 4.10: Distribución de la variable peso a faena (izquierda) y peso promedio a faena en muestras con n=30(derecha). El área sombreada en cada distribución corresponde a valores (en gramos) entre los cuantiles 0,05 y 0,95

A diferencia de lo observado en la población original de pesos a faena, en la cual la **DE** (variación de la variable peso respecto a su μ) era de 148,22 g y el 95% de los pesos se concentraba alrededor de μ entre 2845 g y 3426 g, en la distribución de las medias de muestras con n=30, el **EE** (variación de la variable media respecto a su esperanza) es de 27,4 g y el 95% de las medias se concentran alrededor de μ, entre 3083 g y 3191 g. La Figura 4.11, superpone ambas distribuciones e ilustra la concentración de valores alrededor de la media de cada distribución.

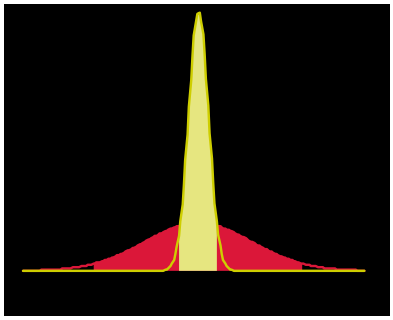
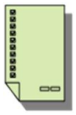


Figura 4.11: Distribuciones de las variables peso a faena (color oscuro) y peso promedio a faena, en muestras con n=30 (color claro). Las áreas sombreadas corresponden a valores (en gramos) entre los cuantiles 0,05 y 0,95 de cada distribución

Vemos que utilizando la media de una muestra podemos aproximarnos o estimar a la media de la población pero que la aproximación no será la misma con cualquier muestra, puesto que en la distribución de las medias éstas se ubican a diferentes distancias respecto de μ. No obstante, sabiendo que el modelo probabilístico de la

distribución de las medias muestrales corresponde al modelo normal podemos conocer la probabilidad de ocurrencia del valor de la media de la muestra elegida.



Tanto para calcular la probabilidad de ocurrencia de determinados valores como para obtener cuantiles en una distribución de la variable aleatoria media muestral, recordemos utilizar en InfoStat el menú Estadísticas, opción Probabilidades y cuantiles, indicando el modelo de la distribución y los valores de sus parámetros.

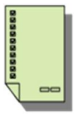
Nos queda pendiente un interrogante: ¿por qué decimos que el **error estándar** es una medida de **confiabilidad**?
La desviación estándar es una medida del error del muestreo (de la variación en la muestra); el error estándar (EE) es una medida de la variación del estimador (en este caso, la media muestral) que permite cuantificar el error de estimación (variación entre las estimaciones).

El **EE** permite obtener una medida de confiabilidad de la estimación o aproximación al verdadero valor de μ. Por ejemplo, si estimamos a μ con una muestra de 30 pollos, con el 95% de las muestras tendríamos un **error de estimación** de a lo sumo 1,96 × 27,4 g = **53,7 g** (por defecto o por exceso) ya que la estimación (es decir la media de la muestra) será un valor entre 3083 g y 3191 g. Dicho de otra manera, si deseamos estimar al verdadero valor de μ eligiendo una muestra de pollos cuyo peso promedio esté a lo sumo a una distancia de 53,7 g de la media verdadera, y sabemos que en la población el peso tiene una **desviación estándar** de **148,22 g**, deberíamos extraer una muestra de 30 pollos. Esto es:

error de estimación= $53,7 = 1,96 * EE = 1,96 * \frac{148,22}{\sqrt{n}}$

$$\text{luego: } n = \left(\frac{1,96 \times 148,22}{53,7} \right)^2 = (5,41)^2 = 29,3 \cong 30 \text{ pollos}$$

El EE puede ser disminuido eligiendo un tamaño muestral lo suficientemente grande como para que la media de la muestra elegida pertenezca al rango de medias muestrales que se encuentran a una distancia deseada de μ.



La varianza de las medias muestrales es inversamente proporcional al tamaño de la muestra. Luego, a través del tamaño de la muestra se puede controlar la variabilidad de distribución del estadístico media muestral y por tanto la confiabilidad que se puede tener de la media de una muestra particular. Si la muestra es de un tamaño n grande, es menos probable obtener una media muestral muy alejada de la media poblacional.

Distribución de una función de la varianza muestral

De manera similar a lo planteado para estudiar a distribución de las medias de todas las muestras posibles de tamaño n=2 con reposición, que obtuvimos de la población conformada por los valores: 1; 3; 5; 7 y 9, calculemos la varianza de cada muestra. Obtenemos los siguientes resultados:

Valores en la muestra	S^2	Valores en la muestra	S^2	Valores en la muestra	S^2	Valores en la muestra	S^2	Valores en la muestra	S^2
1;1	0	3;3	0	5;1	8	7;1	18	9;1	32
1;3	2	3;1	2	5;3	2	7;3	8	9;3	18
1;5	8	3;5	2	5;5	0	7;5	2	9;5	8
1;7	18	3;7	8	5;7	2	7;7	0	9;7	2
1;9	32	3;9	18	5;9	8	7;9	2	9;9	0

Vemos que la varianza cambia según la muestra**; la varianza muestral es una variable aleatoria**. ¿Cómo se distribuyen los valores de la varianza muestral? Dado que tenemos todos los posibles valores de la varianza muestral (S^2) para las muestras de tamaño n=2, podemos tabular y graficar la distribución de la **variable aleatoria varianza muestral**.

S^2	f(S^2)
0	5/25= 0,20
2	8/25= 0,32
8	6/25= 0,24
18	4/25= 0,16
32	2/25= 0,08

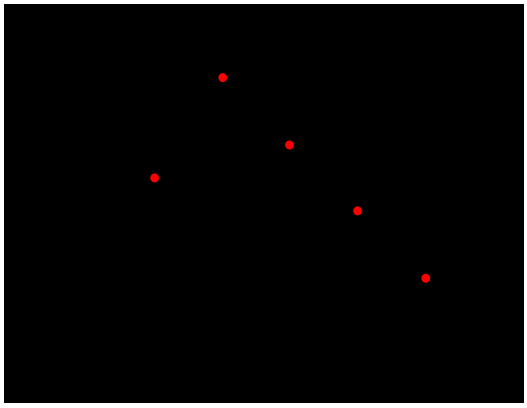


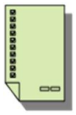
Figura 4.12. Distribución de la variable aleatoria varianza de muestras de tamaño n=2 con reemplazo

Calculemos la esperanza (μ_{S^2}) de la distribución: $\mu_{S^2} = 8$. Los valores de la variable son más frecuentes a la izquierda de la media de la distribución.

Recordando los parámetros de la distribución de la variable aleatoria Y en la población finita con N= 5, la esperanza era 5 y la varianza 8. Podemos ver entonces que la esperanza de la variable aleatoria varianza muestral es igual a la varianza de la población de la que se extrajeron las muestras:

$$\mu_{S^2} = \sigma^2 = 8$$

Este resultado indica que la varianza muestral puede utilizarse para estimar la varianza poblacional.



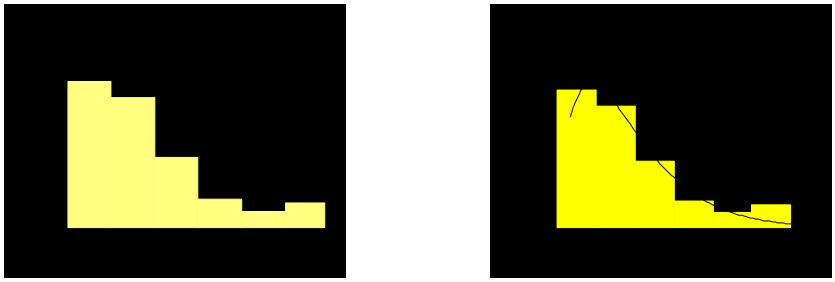
La condición $\mu_{S^2} = \sigma^2$ señala que la varianza muestral es un estimador insesgado de la varianza poblacional.

Repitiendo, en forma análoga a lo presentado con las medias muestrales, veamos qué ocurre con las varianzas de los pesos de pollos a la faena cuando se toman muestras de tamaño 5; 10; 15 y 30. Usaremos la opción Remuestro de las Aplicaciones Didácticas de InfoStat, pero ahora obtendremos las varianzas muestrales. Al igual que en el caso de las medias muestrales, la idea es visualizar la distribución de las varianzas muestrales y poder identificar un modelo de probabilidad que ajuste la distribución.

En el caso de las varianzas muestrales el ajuste a un modelo no se realiza sobre la distribución de los valores de S^2, sino sobre el estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$, de modo que obtenidas las varianzas para cada tamaño de muestra, es necesario calcular los valores de este estadístico. Esto puede realizarse utilizando la opción *Fórmulas* del menú *Datos*, del programa InfoStat, cuando se conoce un valor para σ^2.

La Figura 4.13 muestra las distribuciones de los valores de S^2 y del estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$, para cada tamaño de muestra utilizado, con el ajuste de la correspondiente distribución Chi-cuadrado.

Muestras de tamaño n=5



Muestras de tamaño n=10

Muestras de tamaño n=15

Muestras de tamaño n=30

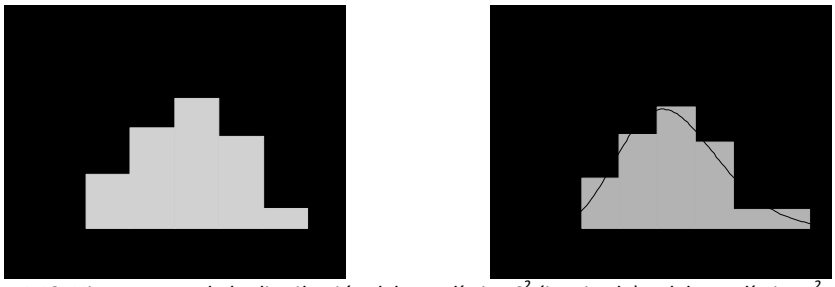


Figura 4.13. Histogramas de la distribución del estadístico S^2 (izquierda) y del estadístico χ^2, con el correspondiente ajuste (derecha).

¿Qué podemos observar en los histogramas anteriores?
La distribución de la varianza muestral es **asimétrica derecha** y se vuelve más simétrica a medida que *n* crece. La distribución de la varianza muestral de muestras obtenidas desde una distribución Normal y escalada por $\frac{(n-1)}{\sigma^2}$ se aproxima a la distribución Chi-cuadrado con n-1 grados de libertad. Esto indica que si deseamos calcular probabilidades referidas a valores de la varianza muestral, debemos utilizar una distribución χ^2 con grados de libertad que dependerán del tamaño muestral con el que se obtuvo la varianza.

Uso de la tabla de la Distribución Chi-cuadrado
Para calcular la probabilidad de que una variable distribuida como una Chi-cuadrado con ν grados de libertad sea menor o igual a un cierto valor, se utiliza la tabla de la distribución acumulada. Cada fila de la tabla corresponde a una distribución Chi-cuadrado para n-1 grados de libertad, de modo que según sea el tamaño muestral nos ubicaremos en una de las filas. En dicha fila buscaremos el valor de x (o el valor aproximado) y leeremos la probabilidad acumulada hasta x, en la cabecera de la columna en la que se encuentra x. Por ejemplo si X se distribuye como una χ^2 con 5 grados de libertad entonces: P ($X \leq 3,99$) = F (3,99) = 0,45.

Comentarios

En este Capítulo hemos experimentado dos ideas centrales: la media muestral y la varianza muestral son variables aleatorias, vale decir no podemos predecir con exactitud su valor y este varía de muestra a muestra. La media muestral es un estimador insesgado de la esperanza de la distribución de la que se extraen las muestras y la varianza muestral lo es de la varianza de dicha distribución poblacional. Las medias de muestras de tamaño *n* siguen una distribución que se aproxima al modelo Normal al aumentar el tamaño muestral, aún cuando los datos originales provinienen de poblaciones no normales.
El error estándar de la media muestral es una medida de confiabilidad las medias muestrales de tamaño *n* y permite conocer el máximo error que podría tener una estimación basada en la media muestral. Se puede calcular el tamaño muestral necesario para estimar a μ con una precisión deseada. Es decir, determinando un valor de distancia entre la estimación y el verdadero valor del parámetro. Una función de las varianzas muestrales, de muestras de tamaño *n*, tiene una distribución teórica denominada Chi-cuadrado con *n*-1 grados de libertad y puede ser usada para calcular probabilidades relativas a varianzas muestrales

Notación

Media de la distribución de las medias de muestras de tamaño *n*: $\mu_{\bar{y}}$

Varianza de la distribución de las medias de muestras de tamaño *n*: $\sigma^2_{\bar{y}}$

Error estándar de la distribución de las medias de de muestras de tamaño *n*: $EE = \sigma_{\bar{y}}$

Distribución de la variable aleatoria media muestral $\bar{Y}$, para muestras aleatorias de tamaño *n* extraídas de una población con esperanza μ y varianza σ^2: $\bar{Y} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

Estadístico Chi-cuadrado: $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$

Distribución del estadístico χ^2: $\chi^2 \sim \chi^2_{n-1}$

Definiciones

Definición 4.1: Error Estándar

La desviación estándar (raíz cuadrada de la varianza) de la variable aleatoria media muestral de muestras de tamaño n, recibe el nombre de ***Error Estándar*** *y es expresado como:* $EE = \sigma_{\bar{Y}} = \sqrt{\sigma^2_{\bar{Y}}} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$

Definición 4.2: Estadístico Chi-cuadrado

Cuando las varianzas muestrales son obtenidas de muestras provenientes de una población con esperanza μ y varianza σ^2, el estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ *, sigue una distribución Chi-cuadrado con n-1 grados de libertad.*

Definición 4.3: Teorema Central del Límite

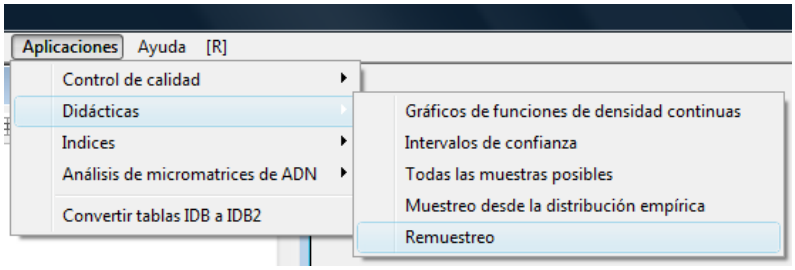
El teorema, hace referencia a la distribución del estadístico Z, proveniente de la estandarización de la variable aleatoria media muestral, postulando que aunque X no se distribuya como una variable aleatoria normal, si tiene varianza finita, entonces para n suficientemente grande, la distribución de: Z= $\left(\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$ *converge en distribución a una N(0,1). Se dice entonces que Z posee una distribución asintóticamente normal.* ***Nota:*** *Cuando se dice que una variable con distribución $F_n(.)$ converge en distribución a una distribución G(.), cuando n tiende a infinito, se quiere indicar que* $\forall \varepsilon > 0 \ \exists \ n_0$ *tal que* $|F_n(yx) - G(yx)| < \varepsilon \ \forall yx \in \Re$ *si* $n > n_0$.

Ejercicios

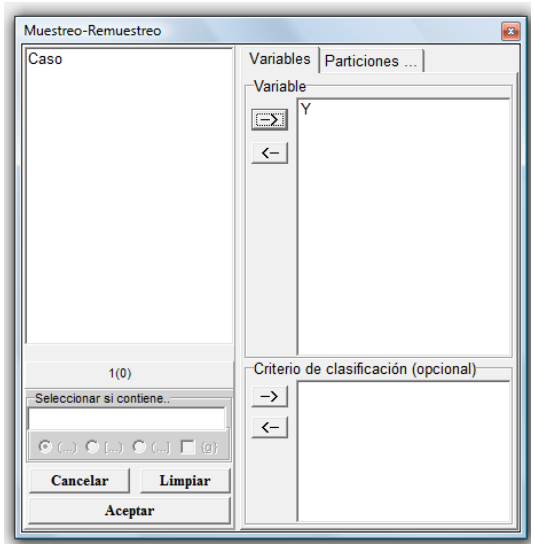
Ejercicio 4.1: Para estudiar empíricamente la distribución de la medias muestrales, utilice un procedimiento de simulación. Suponga que los datos de la variable Y (archivo Ejercicio-1CapituloDEM), representan a una población con μ=27.96 y σ2=27.77. La simulación consiste en generar un número grande de experimentos (200) en los cuales se obtengan muestras con n=3, n=10 y n=25, a partir de un muestreo sin reposición.

Para obtener los resultados de la simulación siga los siguientes pasos:

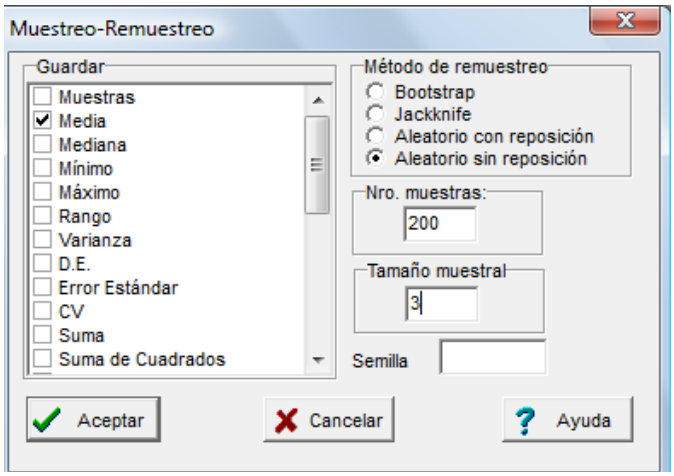
a) En el programa InfoStat, abra el archivo que contiene los datos poblacionales y seleccione ***Aplicaciones*** $\Rightarrow$ ***Didácticas*** $\Rightarrow$ ***Remuestreo,*** como se muestra en la siguiente ventana.



b) A continuación se mostrará la siguiente ventana de diálogo donde deberá indicar que Y es la columna con los datos a utilizar.



c) En la siguiente ventana de diálogo elija el Método de remuestreo: Aleatorio sin reposición, e ingrese el Nro. de muestras y el Tamaño muestral.



d) Al aceptar esta configuración del remuestreo, se generará una nueva tabla con los 200 valores generados.

e) Con los resultados construya un histograma de frecuencias relativas que incluya el ajuste de un modelo normal.

f) Repita el procedimiento del remuestreo usando los tamaños muestrales n=10 y n=25. Recuerde utilizar la tabla de datos con la variable Y. Construya los correspondientes histogramas. En todos los gráficos mantenga la misma escala (minimos y maximos) en el eje X y en el Eje Y, así como también la cantidad de clases.

g) ¿Cuál es el promedio de las medias muestrales para los tres escenarios? ¿Cómo es este promedio respecto del promedio de la población?

h) ¿Cómo es la varianza de las medias obtenidas en cada muestreo respecto de la varianza de la población? Justifique.

i) Comparando los resultados, si Ud. tuviera que estimar a la media de la población: ¿qué estrategia utilizaría? Justifique.

Ejercicio 4.2: En una población de plantas de una especie ornamental la variable aleatoria altura se distribuye en forma aproximada a una normal con media 30 cm y desviación estándar 6 cm.

De acuerdo al enunciado, en cada afirmación indique si es verdadera o falsa. Justifique sus respuestas.

a) Para que las medias de muestras extraídas de la población tengan distribución normal el tamaño muestral deberá ser superior a 100.

b) En la distribución de 200 medias muestrales obtenidas en muestras de tamaño n=10 los valores se concentrarán más alrededor de μ que en una distribución en base a las medias de 100 muestras de tamaño n=20.

c) El error estándar es una estimación de la variabilidad de la altura promedio de muestras de n plantas tomadas de la población.

d) La probabilidad de que en una muestra aleatoria de plantas la altura promedio sea menor a 30 cm, es mayor al tomar una muestra de tamaño 100 que al tomar una muestra de tamaño 10.

e) La variabilidad de la altura promedio en muestras de tamaño n será menor que la variabilidad de la altura de las plantas en la población.

f) La variabilidad de la altura promedio en muestras de 10 plantas es menor que la variabilidad en muestras de 100 plantas.

g) Tomando una muestra de tamaño 100 se obtendrá una estimación más precisa del verdadero promedio de la altura de las plantas de la población, que tomando una muestra de tamaño 10.

Ejercicio 4.3: Si la distribución de la variable aleatoria producción de leche/vaca/lactancia de un establecimiento lácteo se aproxima a una distribución normal con media μ=7000 litros y desvío estándar σ=800 litros.

a) ¿Cuál es la probabilidad de que la media de la producción por lactancia en una muestra de 5 vacas exceda el valor de 7500 litros?

b) En muestras de 5 vacas ¿Cuál es la producción promedio sólo superada por un 5% de las producciones promedio?

Ejercicio 4.4: Uso De la tabla de la Distribución Chi-cuadrado

En la tabla de Distribución Chi-cuadrado acumulada se pueden encontrar algunos cuantiles de la distribución para diferentes grados de libertad. Para calcular la probabilidad de que una variable distribuida como una chi-cuadrado con ν grados de libertad sea menor o igual a un cierto valor se procede de la siguiente forma:

Se busca en la tabla la fila que corresponde a los grados de libertad de la distribución y dentro de esa fila se localiza (de manera exacta o aproximada) el valor x. Luego se lee la probabilidad buscada mirando el encabezamiento de la columna correspondiente.

Por ejemplo, si X se distribuye como una χ2 con 5 grados de libertad entonces:

$P(X \leq 6,1) = F(6,1) = 0,70$

Como ejercicio de uso de la tabla encontrar:

a) $P(X \leq 20,5)$ si X se distribuye como una χ2 con 15 grados de libertad.

b) $P(S2(n-1)/\sigma 2 \leq 10)$ si S2 fue obtenido a partir de una muestra de tamaño 10.

Ejercicio 4.5: En un criadero de semillas se está probando una nueva variedad de maíz que saldrá a la venta si en una muestra de 50 parcelas experimentales el desvío estándar de su rendimiento no supera los 23 kg/ha.

a) ¿Cuál es la probabilidad de que la variedad salga a la venta si la verdadera desviación estándar es 20?

b) ¿Cuál es el valor por debajo del cual está el 99% de los valores posibles de desviaciones estándar muestrales basadas en muestras de tamaño 30, si la verdadera desviación estándar es 20?

Capítulo 5

Estimación de parámetros y contraste de hipótesis

Julio A. Di Rienzo

Análisis de regresión

5. Estimación de parámetros y contraste de hipótesis

Motivación

La toma de decisiones basada en criterios estadísticos se fundamenta en el conocimiento de la forma en que se distribuyen las variables aleatorias. Por ejemplo, para establecer la aptitud de una localidad-región para un cultivo se consideran, entre otras cosas, el régimen de lluvias y de temperaturas. Estas consideraciones contemplan explícita o implícitamente el cálculo de probabilidad de la ocurrencia de eventos que, ya sea por exceso y/o por defecto, hacen fracasar una cosecha. Cuando esta probabilidad es grande se concluye que, para las demandas del cultivo en cuestión, la localidad-región no es apta o lo es marginalmente. El cálculo de esas probabilidades implica conocer la función de distribución de la variable (aleatoria) objeto de estudio. Esta función está caracterizada por *parámetros* que en la práctica son desconocidos. El propósito de este capítulo es discutir la problemática de la *estimación* de parámetros relativos a etas distribuciones, su confiabilidad y contrastar hipótesis sobre ellos.

Conceptos teóricos y procedimientos

Recordemos que la distribución de una variable aleatoria se simboliza usualmente como F(x). Su argumento (x) representa valores particulares de la variable aleatoria y su resultado es un valor comprendido entre 0 y 1. La función de distribución devuelve la probabilidad de que la variable aleatoria se realice con valores menores o iguales al argumento dado (probabilidad acumulada). Por ejemplo, si F(.) fuera la función de distribución de la variable *milímetros de precipitación anual* de una localidad, entonces podríamos evaluarla para un milimetraje particular: por ejemplo F(700). Si F(700)=0,30, diremos que la probabilidad de que en un año cualquiera el milimetraje de precipitación anual sea igual o menor a 700 mm es 0,30. Luego, en promedio, 3 de cada 10 años, tendrán precipitaciones iguales o inferiores a 700 mm. Recíprocamente, la probabilidad de que llueva más de 700 mm será 0,70.

Esta función se puede visualizar utilizando un gráfico de dispersión con los valores de milimetraje en el eje X y la probabilidad acumulada correspondiente en el eje Y (Figura 5.1). En esta figura puede leerse la probabilidad antes mencionada. También se lee que por debajo de 1200 mm ocurrirán casi todas las precipitaciones que puedan registrarse anualmente y por lo tanto será muy poco probable la ocurrencia de precipitaciones mayores a 1200 mm.
En la mayoría de las aplicaciones prácticas no se cuenta con estas funciones de distribución. Sin embargo, podemos tener datos para construirlas. Por ejemplo, si tuviéramos 150 registros de precipitación anual para la localidad en cuestión podríamos obtener los que se llama la **función de distribución empírica** cuya gráfica, para un ejemplo particular hipotético, se muestra en la Figura 5.2.

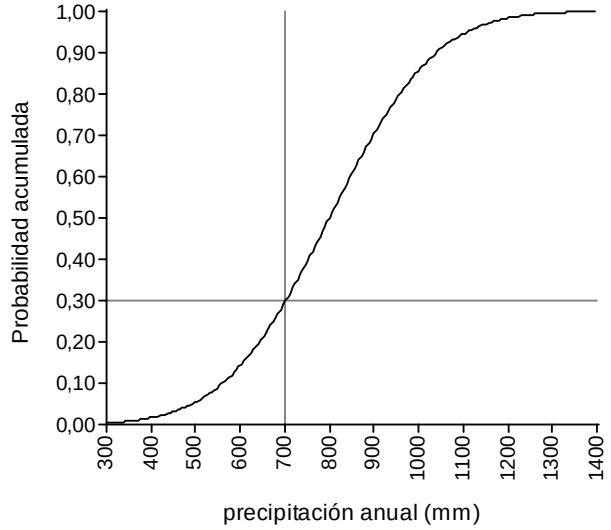


Figura 5.1: Función de distribución de la variable precipitación anual (mm).

Esta función aproxima bastante bien al modelo teórico y puede ser adecuada para muchas aplicaciones prácticas. Sin embargo, uno de sus problemas es que la lectura de las probabilidades de eventos muy extremos es difícil de realizar, ya sea porque no hay datos para esos eventos o porque la información es muy incompleta. Esta situación se agrava cuando la disponibilidad de datos es más reducida. Por ejemplo, si se tuviera una serie de 30 registros de precipitaciones anuales para nuestra localidad hipotética, podríamos encontrar la distribución empírica que se ilustra en la Figura 5.3.

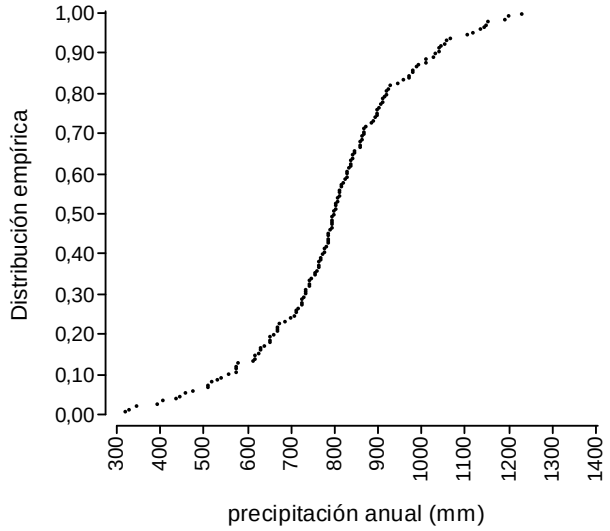


Figura 5.2: Función de distribución empírica de la variable precipitación anual (mm) obtenida a partir de 150 observaciones.

A medida que disminuye la disponibilidad de observaciones, más imprecisa es la forma de la distribución empírica, y más difícil el cálculo de probabilidad de ocurrencia de eventos extremos. En este punto hay dos caminos posibles: conseguir más datos o, suponer que la variable en estudio sigue una función de distribución **teórica conocida** y utilizar los datos disponibles para **estimar** los parámetros que la caracterizan. La ventaja de la última aproximación es que al tener una función de distribución conocida, ya no dependemos de la disponibilidad de datos en las regiones extremas del rango de variación de la variable aleatoria para poder calcular la probabilidad de los eventos extremos. La desventaja es que la pertinencia de la función teórica escogida es una suposición del cálculo, y si la variable en estudio sigue una distribución diferente, el cálculo de probabilidades será inapropiado, especialmente, cuando estamos interesados en asignar probabilidades a eventos extremos.

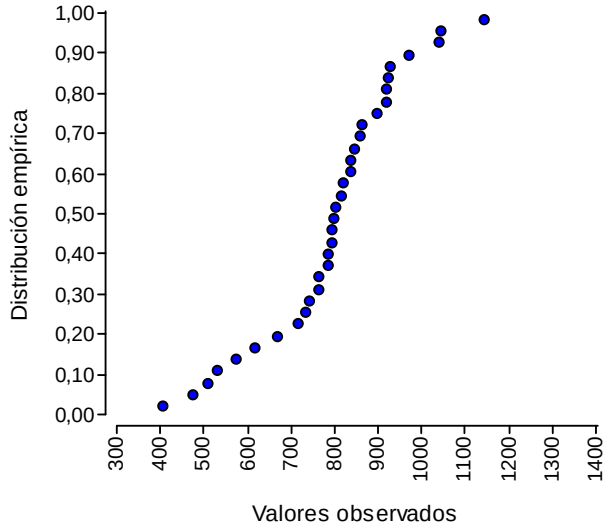


Figura 5.3: Función de distribución empírica de la variable precipitación anual (mm) obtenida a partir de 30 observaciones.

Modelo estadístico

Parece oportuno introducir aquí el concepto de modelo estadístico. Este concepto permite vincular la función de distribución de una variable aleatoria con la práctica común de la experimentación, que consiste en la comparación del comportamientode una variable (aleatoria) bajo diferentes escenarios o condiciones experimentales.

Los estadísticos tratan a las observaciones de un experimento (o muestreo) como las realizaciones de un conjunto de variables aleatorias. Aún en presencia de variabilidad aleatoria es posible encontrar patrones en los datos y la identificación, y caracterización de los mismos es el propósito del análisis estadístico. Para ello **las_observaciones se idealizan mediante un modelo estadístico**. Vamos a restringir nuestra discusión al caso de los modelos lineales que constituyen la base de la estadística aplicada a la experimentación agropecuaria.

Un modelo estadístico incluye una **parte fija** y otra **aleatoria**. La parte aleatoria nos recuerda el carácter variable de las observaciones, mientras que la fija describe la tendencia, lo repetible, lo esperable en promedio. Las partes fija y aleatoria caracterizan a los **parámetros de posición** y **dispersión** de la variable en estudio, respectivamente. Por ejemplo, un modelo para las precipitaciones anuales en tres localidades podría ser el siguiente:

$$Y_{ij} = \mu + \lambda_i + \varepsilon_{ij}$$

Este modelo dice que Y_{ij}, que podría denotar el valor observado de precipitación en la j-ésima localidad y en el i-ésimo año es la resultante de sumar el nivel medio de precipitaciones anuales μ, común a todas las localidades, más λ_i, el efecto de la i-ésima localidad sobre el promedio de las precipitaciones anuales. La discrepancia entre la suma $(\mu+\lambda_i)$ y el valor observado en la i-ésima localidad, j-ésimo año, está representada por ε_{ij}. Este último término se considera aleatorio y se conoce como el **término del error**. Si $\mu=800$ y los efectos de las localidades sobre la media son $\lambda_1=-180$, $\lambda_2=120$ y $\lambda_3=60$ y, además, suponemos que la función de distribución de los errores es normal con media 0 y varianza 30000, el gráfico de las funciones de distribución se puede visualizar en la Figura 5.4. El número 30000 se propuso sólo a los efectos de la ejemplificación.

En la Figura 5.4 puede leerse que precipitaciones anuales menores a 700 mm ocurren frecuentemente en la Localidad 1 y son algo menos frecuentes en la Localidad 2 (la probabilidad aproximada de este evento es 0,50 y 0,30 para las localidades 1 y 2 respectivamente). Mientras tanto, para la Localidad 3 esa probabilidad es pequeña: cercana a 0,10.

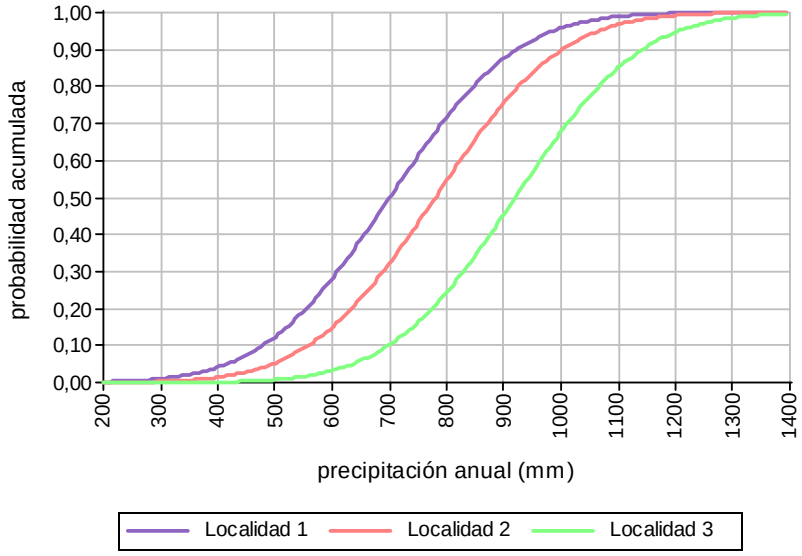


Figura 5.4: Funciones de distribución para el modelo $Y_{ij}=\mu+\lambda_i+\varepsilon_{ij}$ *con* $\mu=800$, $\lambda_1=-100$, $\lambda_2=-20$ *y* $\lambda_3=120$ *y* ε_{ij} *~N(0;30000).*

La Figura 5.5 muestra un caso similar al anterior excepto que las tres localidades tienen efecto nulo sobre el valor medio de precipitaciones anuales. En este caso las funciones de distribución de las precipitaciones anuales de las tres localidades son indistinguibles

por sus parámetros de posición. Supondremos, en cambio, diferencias en sus parámetros de dispersión. Para la ilustración: $\mu = 800$, los efectos de las localidades son nulos y los errores se supondrán normales con media 0 y varianzas diferentes: 30000, 10000 y 80000 para las localidades 1, 2 y 3 respectivamente.
Aunque el milimetraje que acumula la probabilidad 0,5 es el mismo en todas las localidades (800 mm), precipitaciones anuales menores a 650 mm constituyen un evento raro en la Localidad 2, tienen una probabilidad aproximada de 0,20 en la Localidad 1 y ocurren en 3 de cada 10 años en la Localidad 3.

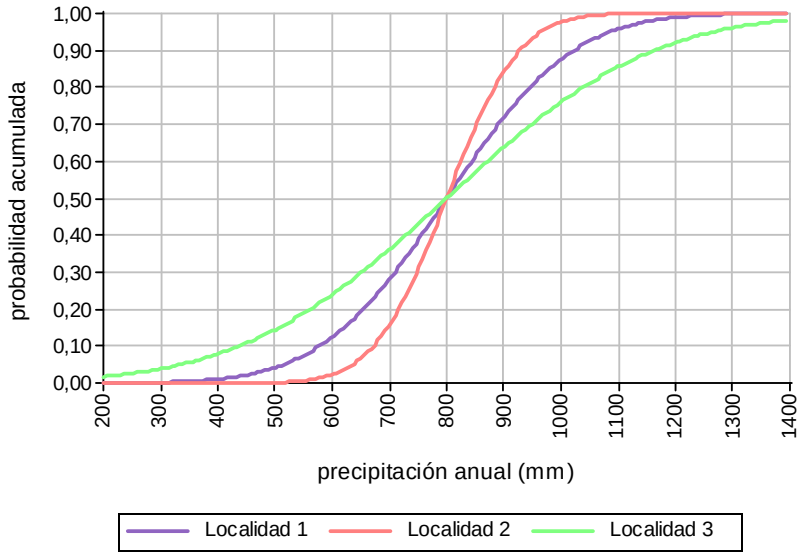
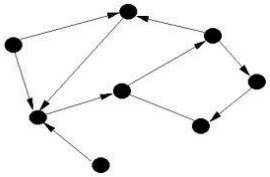


Figura 5.5: Funciones de distribución para el modelo $Y_{ij} = \mu + \lambda_i + \varepsilon_{ij}$ *con* $\mu = 800$, $\lambda_1 = \lambda_2 = \lambda_3 = 0$, *y* ε_{i1} *~N(0;30000),* ε_{i2} *~N(0;10000),* ε_{i3} *~N(0;80000).*



Los modelos estadísticos constituyen una forma sintética y eficiente de representar el proceso aleatorio que genera las observaciones. Cambios en los parámetros de posición y dispersión permiten contemplar una gran variedad de situaciones.

A continuación nos concentraremos en el problema de la estimación de los parámetros que caracterizan a los modelos estadísticos, en particular, a los modelos estadísticos lineales.

Estimación puntual

Cuando se aproxima el parámetro de una distribución a través de un valor calculado a partir de una muestra decimos que se está haciendo una **estimación puntual** del parámetro. Supongamos que tenemos una muestra aleatoria $\{y_1,y_2,...,y_n\}$ de la variable Y, cuya función de distribución acumulada es $F(y;\theta)$. En esta notación estamos indicando que F depende del parámetro θ. Por otra parte, θ es desconocido y no podremos utilizar F(.) a menos que asignemos un valor a θ. Para **estimar** este parámetro usaremos los valores observados en la muestra. Con este objetivo propondremos una función $\hat{\theta}(.)$ que, partiendo de la muestra disponible, produce un valor razonable para el parámetro objeto de estimación. Hemos escogido como símbolo de la función el mismo símbolo del parámetro, y para distinguirlos, marcamos a este último con un acento circunflejo.
No daremos, en lo que sigue, definiciones matemáticas. Aunque ello implica una pérdida de precisión en las definiciones, esperamos, sin embargo, que esto ayude al lector no especializado a lograr la conceptualización deseada.
Toda función basada en una muestra se conoce como **estadístico muestral**. Los estimadores son estadísticos muestrales y en consecuencia son **variables aleatorias**, ya que son funciones de variables aleatorias. Para que un estadístico muestral sirva como estimador, debemos evaluar algunas propiedades que caracterizan a los estimadores.
La elección de un buen estimador, entre un conjunto de posibles estimadores, se realiza teniendo en cuenta 4 propiedades:

- Consistencia
- Insesgamiento
- Eficiencia
- Cerramiento

Consistencia

Diremos que un estimador es ***consistente*** si éste se "aproxima" al parámetro cuanto mayor es el tamaño muestral. Un ejemplo clásico de estimador consistente es la media muestral $\bar{Y}$. La consistencia es la propiedad más importante de un estimador e implica que la estimación mejora (en términos de proximidad entre la estimación y el parámetro estimado) con el incremento en el número de observaciones disponibles. Si un estimador no es consistente, no sirve.

Insesgamiento

Esta propiedad pide a un estimador que, para cualquier tamaño muestra, su valor esperado sea el valor de parámetro. En términos prácticos, esta propiedad implica que si se tomaran muchas muestras de tamaño *n* y se calcula con cada una de ellas el **estimador insesgado**, entonces el promedio de todas estas estimaciones será el valor del parámetro. Cuando esta propiedad no se cumple se dice que el estimador es **sesgado**. El **sesgo** puede ser positivo o negativo. Esta propiedad no es contradictoria de la propiedad de consistencia, pero si un estimador es consistente pero sesgado esto implica que el sesgo se achica con el incremento del tamaño muestral. Se puede probar que la media muestral (promedio) es un **estimador insesgado** de la media poblacional.

Eficiencia

Cuando un estimador es *eficiente* no existe otro, dentro de su categoría, que tenga menor varianza. Esta propiedad es deseable porque implica mayor estabilidad de las estimaciones (estabilidad en el sentido de que si se tomara otra muestra la estimación resultaría "parecida"). La media y la mediana muestrales son, ambos, estimadores *consistentes* e *insesgados* de la media de una variable aleatoria. Si la variable cuya media se quiere estimar tuviera distribución normal, la media muestral es el *estimador de mínima varianza* dentro de los estimadores insesgados, y por lo tanto: el *estimador eficiente*. Cuando la distribución admite valores extremos, propios de las distribuciones asimétricas, como puede ser la distribución exponencial, esta propiedad la tiene la mediana.

Cerramiento

Esta propiedad indica que el estimador siempre produce valores admisibles para el parámetro. Por ejemplo, la varianza es una medida de variabilidad y su cota inferior es 0. Si un estimador de la varianza produce, eventualmente, resultados negativos, entonces no cumple con la propiedad de cerramiento.

Confiabilidad de una estimación

Como se indicó anteriormente los estimadores son variables aleatorias ya que se construyen a partir de una colección de ellas (muestra). Es necesario entonces dar una medida de su confiabilidad. Esto puede hacerse calculando su **error estándar**.

Error estándar

El error estándar de un estimador es la raíz cuadrada de su varianza y la expresión para calcularlo es propia de cada estimador. Por ejemplo, el **error estándar de la media muestral** se calcula como la desviación estándar dividida por la raíz cuadrada del tamaño muestral. Su fórmula es:

$$EE_{\bar{Y}} = S/\sqrt{n}$$

Es útil expresar el error estándar en términos relativos. Si EE representa el error estándar de un estimador $\hat{\theta}$, el error estándar relativo es $EE / \hat{\theta}$. Un error estándar relativo de hasta 0,20 podría ser admisible, pero un error estándar relativo de 0,80 implicaría que la discrepancia promedio del estimador respecto del valor que está estimando, representa aproximadamente un 80% del mismo.

Intervalo de confianza

Otra forma de reportar la incertidumbre de una estimación es dando un ***intervalo de confianza*** para el parámetro que se quiere estimar. Estos intervalos tienen una **probabilidad diseñada** de contener al verdadero valor del parámetro. Esta probabilidad se fija usualmente en 0,95 o superior. Intervalos de menor confianza, como por ejemplo 0,90 o 0,80 son admisibles, aunque en estos casos es conveniente dar alguna explicación que justifique su utilización. La probabilidad de un intervalo de confianza corresponde a la probabilidad de que el intervalo contenga al verdadero valor del parámetro. Sin embargo, para una muestra particular, una vez que los límites se han calculado, asignar una probabilidad al intervalo obtenido no es más aplicable (ya que no es más un intervalo de **límites aleatorios**) y por ello se dice que el intervalo tiene una ***confianza*** del p%, donde p es la probabilidad diseñada.

Un ejemplo típico es la construcción del intervalo de confianza para la media de una población. Este intervalo se calcula partiendo del hecho que:

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

Esta expresión indica que la **diferencia estandarizada de la media muestral respecto de la media poblacional** sigue una distribución de tipo T. Esta distribución es simétrica, acampanada, centrada en cero y está caracterizada por un parámetro conocido como *grados de libertad*. En este caso, el parámetro grados de libertad vale n-1 (el tamaño de la muestra menos uno). La distribución T es una distribución similar a una distribución normal estándar, aunque más achatada. Cuando los grados de libertad de la T son grandes, ésta es indistinguible de una normal estándar.

Mediante manipulación algebraica es posible derivar los límites inferior (LI) y superior (LS) del intervalo de confianza (bilateral) para la media, dado un nivel de confianza $(1-\alpha)\times 100\%$. Si el intervalo tiene una confianza del 95%, entonces $(1-\alpha) = 0.95 \Leftrightarrow \alpha = 0.05$. A continuación se dan las expresiones para obtener los límites del intervalo de confianza:

$$LI = \bar{Y} - T_{1-\alpha/2;n-1}\, S/\sqrt{n}; \quad LS = \bar{Y} + T_{1-\alpha/2;n-1}\, S/\sqrt{n}$$

En dicha expresión, $\bar{Y}$ representa la media muestral y $S/\sqrt{n}$ el estimador de su error estándar. Luego, dada una muestra, la construcción del intervalo de confianza **bilateral**

(tiene límite inferior y superior) para la media poblacional se obtiene sumando y restando de la media muestral, $T_{1-\alpha/2;n-1}$ veces su error estándar.

El coeficiente $T_{1-\alpha/2;n-1}$ corresponde al percentil $(1-\alpha/2)$ de una distribución T con *n*-1 grados de libertad. Si deseamos un intervalo de confianza al 95% entonces $1-\alpha=0.95$ de donde $\alpha=0.05$ y por lo tanto $1-\alpha/2=0.975$. Luego, si tuviésemos una muestra de tamaño *n*=20, el coeficiente por el que habría que multiplicar al error estándar de la media (para restar y sumar, a fin de obtener los límites inferior y superior respectivamente), sería el percentil 0,975 de una T con 19 grados de libertad.

El coeficiente es fácil de obtener con la calculadora de probabilidades y cuantiles de InfoStat (Figura 5.6) seleccionando *T Student (v)* y completando los campos marcados con los grados de libertad apropiados y la probabilidad acumulada. El *[Valor de x]* para la probabilidad ingresada es el cuantil 0,975 de la distribución.

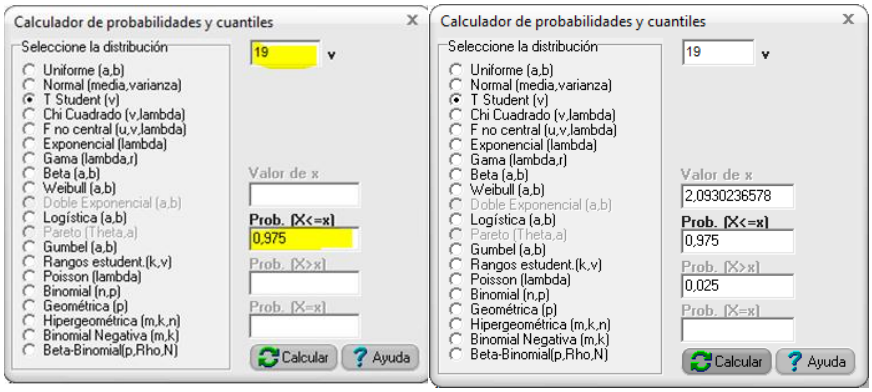


Figura 5.6: Ventana de diálogo de la calculadora de probabilidades y cuantiles. En el ejemplo se muestran resaltados los campos que deben llenarse para calcular el percentil 0,975 de una T con 19 grados de libertad (izquierda) y el resultado al accionar el botón calcular (Derecha).

El coeficiente calculado es 2,093. Cuanto mayor es el tamaño de la muestra menor es el coeficiente T utilizado, pero éste tiene una cota inferior de **1,96**; es por ello que, como un procedimiento aproximado, basado en la suposición de normalidad para la variable, se puede obtener un intervalo de confianza al 95% partiendo del valor estimado, sumándole y restándole **2** veces su error estándar. Los percentiles de una T con los grados de libertad apropiados se pueden consultar también en una tabla de cuantiles de esta distribución, como la se encuentra en el Anexo Tablas Estadísticas.

Aplicación

Residuos de insecticida en apio

Los siguientes datos corresponden a los residuos de un insecticida (en ppm) en plantas de un lote de apio:

0,40	0,77	0,28	0,41	0,74	0,74	0,34	0,22	0,33	0,34
0,42	0,17	0,22	0,23	0,35	0,48	0,42	0,59	0,21	0,48
0,67	0,66	0,34	0,37	0,34	0,52	0,32	0,33	0,27	0,32

Las normas de comercialización establecen que si el residuo de insecticida es igual o mayor que 0,50 ppm, es peligroso para el consumo humano. El contenido de residuos promedio obtenido del lote es: $\overline{Y} = 0,41\text{ppm}$ y la desviación estándar estimada S=0,1686 ppm.

Estrategia de análisis

Estimaremos el intervalo de confianza para el residuo promedio trabajando con α=0,001, de manera tal que sólo 1 de cada mil procedimientos de muestreo basados en un tamaño muestral de 30 unidades muestrales, tengan un nivel medio de residuos fuera del intervalo calculado. Vamos a utilizar lo que se llama un intervalo de confianza **unilateral derecho**, estos intervalos tienen límite inferior en el –infinito y un límite superior dado por $LS = \overline{Y} + T_{1-\alpha;n-1}\, S/\sqrt{n}$. La razón de utilizar el límite unilateral derecho es que no estamos interesados en establecer si la verdadera media está por encima de un valor pequeño sino si está por debajo de una cantidad crítica: 0,50 ppm. La diferencia al construir un intervalo unilateral derecho, respecto de uno bilateral, es que el cuantil de la T que debemos utilizar no es cuantil $1-\alpha/2$ sino el $1-\alpha$.

Para el problema que estamos resolviendo $T_{1-\alpha;n-1} = T_{0.999;29} = 3,3962$. En consecuencia con una media muestral $\overline{Y} = 0,41$ y un error estándar $EE = 0.1686/\sqrt{30} = 0.03078201$ el límite superior del intervalo de confianza unilateral derecho será ≈0,514.

¿Por qué utilizamos un nivel de confianza del 99,9% y no del 95%? La razón es que queremos proteger al consumidor. Cuanto mayor es la confianza más amplio es el intervalo de confianza y esto implica que serán rechazados más lotes que si usáramos un intervalo de confianza al 95%.

Conclusión

Esta muestra es compatible con una media de la concentración de insecticida_superior al límite tolerado y deberá rechazarse.

Contraste de hipótesis

Como se indicó anteriormente los modelos estadísticos tienen una parte fija y otra aleatoria que caracterizan, respectivamente, los parámetros de posición y dispersión de la variable aleatoria bajo estudio. Vamos a centrar nuestra discusión sobre el contraste de hipótesis en el contexto de los modelos lineales. Estos modelos son la base teórica y conceptual del análisis de la varianza y del análisis de regresión (que se discutirán más adelante) y que constituyen el cuerpo principal de métodos estadísticos aplicados a la experimentación agropecuaria.

En los modelos lineales la parte aleatoria puede estar representada por un único término (modelo lineal clásico) o por un conjunto de componentes (modelo lineal mixto). En estos modelos se supone que los componentes aleatorios siguen una distribución normal con esperanza cero. Cada componente aleatorio tiene una varianza determinada y cuando hay más de uno se suponen mutuamente independientes. La parte fija, en tanto, modela la esperanza de la variable aleatoria.

El **contraste de hipótesis** consiste en establecer el valor de verdad (verdadero-falso) de una o más **proposiciones** enunciadas sobre los parámetros de la parte fija o sobre los parámetros de la parte aleatoria de un modelo estadístico. Por ello, antes de proceder con un contraste de hipótesis, debemos proponer un modelo para los datos y estimar sus parámetros.

El modelo verdadero es desconocido para el investigador, por lo que, el que se propone, es sólo un modelo plausible para los datos. En el contraste de hipótesis siempre hay dos modelos competidores: el **modelo nulo** y el **alternativo**, este último, con un número mayor de parámetros. Usualmente el modelo propuesto por el investigador es el modelo alternativo**.** El contraste de hipótesis sirve para establecer si el modelo alternativo es necesario para explicar los datos que se observan o si un modelo más simple (modelo nulo), con un número menor de parámetros, es suficiente.

En el lenguaje del contraste de hipótesis se contrastan una **hipótesis nula** *vs*. una **hipótesis alternativa**. La hipótesis nula que se simboliza con $\mathbf{H}_0$ sostiene que el **modelo nulo** es el **correcto**, mientras que la hipótesis alternativa, que se simboliza con $\mathbf{H}_1$, establece que el **modelo alternativo** es el **correcto**.

Para establecer si la hipótesis nula es consistente o no con los datos (verdadera o falsa) se realiza una **prueba estadística** (test) que asigna una medida de **confiabilidad** a la hipótesis nula. La prueba se basa en un estadístico muestral (calculado a partir de los datos observados) y la medida de confiabilidad se calcula teniendo en cuenta la distribución muestral de ese estadístico cuando la hipótesis nula es cierta. La **confiabilidad** se expresa en términos de **probabilidad** y se la conoce como ***valor p*** (en inglés p-value). Cuanto menor es el *valor p* menos confianza tenemos en la hipótesis

nula. Para decidir cuándo dejamos de "creer" en la hipótesis nula se fija un umbral. Si el ***valor p*** está por debajo del umbral decimos que la hipótesis nula no es consistente con los datos observados (la hipótesis nula se **rechaza**) y se acepta la hipótesis alternativa.
El umbral utilizado para decidir cuándo rechazamos la hipótesis nula se conoce como **nivel de significación** de la prueba y se simboliza con α. Cuando la hipótesis nula se rechaza se dice que la prueba fue **significativa**. En caso contrario diremos que no hay evidencia suficiente para rechazar la hipótesis nula (o que la prueba no fue significativa). Un nivel de significación estándar es 0,05, pero niveles de significación como 0,01 y 0,001 son también convencionales.

Nivel de significación

¿Cuál es la racionalidad detrás del nivel de significación? Cuando una hipótesis nula se somete a prueba es posible concluir que ésta es falsa aún cuando sea verdadera. Este error se conoce como **error de tipo I**. Puede ocurrir debido a que los datos disponibles sean, por azar, muy desfavorables para la hipótesis nula. Está claro que si la hipótesis nula fuera cierta la frecuencia con que aparecerán "datos desfavorables" será pequeña. El nivel de significación **es la probabilidad máxima y admisible de cometer el error de tipo I**. Luego el nivel de significación es el instrumento que tiene el investigador para controlar la tasa con que puede ocurrir este tipo de error. Obviamente que todos quisiéramos que la tasa de error de tipo I fuera cero o muy pequeña, el problema es que cuando disminuimos la tasa de error de tipo I aumenta la probabilidad de ocurrencia de otro tipo de error: el **error de tipo II**. Este error corresponde a la **aceptación** de la hipótesis nula cuando es **falsa**. Su probabilidad de ocurrencia se simboliza con β.
Para ejemplicar el contraste de hipótesis, consideremos un caso simple donde tenemos una muestra de 20 observaciones (n=20): $\{Y_1,Y_2,...,Y_n\}$ que corresponden al *peso seco de plantines* de *Melillotus* recolectados a los 30 días desde la germinación. *Melillotus* es un género de leguminosas forrajeras que se asocian a bacterias para fijar simbióticamente nitrógeno. La eficiencia de fijación de nitrógeno depende, entre otras cosas, de la cepa bacteriana con la que interactúa la planta. En el experimento que examinamos los datos se obtuvieron utilizando una cepa experimental de *Rhizobium* (género de bacterias fijadoras de nitrógeno) como inoculante. Se quiere establecer si esa cepa es mejor que la utilizada en un inoculante comercial (tradicional).
Supongamos que existe suficiente experiencia con el inoculante tradicional para **saber** que el promedio del peso seco de los plantines a los 30 días de edad es μ_0. Además, supondremos que el investigador tiene gran control de las condiciones bajo las cuales se realiza el experimento, de manera tal que cualquier diferencia en el promedio de peso seco debe atribuirse a la nueva cepa.

Estas suposiciones las hacemos para simplificar el problema. En la práctica son difícilmente aceptables. Por ello se hacen experimentos comparativos en los que se evalúan simultáneamente ambos inoculantes. Un ejemplo de este tipo se presenta en el próximo capítulo.

El modelo nulo para este experimento es:

$$Y_i = \mu_0 + \varepsilon_i$$

Este modelo sugiere que todas las observaciones comparten la media μ_0 y que toda la variación observada se debe a variaciones aleatorias atribuibles a variabilidad biológica y errores de medición.
El modelo alternativo, a continuación, es una extensión del modelo nulo al que se le agrega el parámetro τ .

$$Y_i = \mu_0 + \tau + \varepsilon_i$$

Los términos de los dos modelos anteriores se interpretan de la siguiente manera:

Y_i: simboliza una observación (el índice "i" indica que se trata de la i-ésima observación, *i* varía de 1 a 20)

μ_0: es una **constante conocida** que representa el peso promedio de plantines cuando se utiliza el inoculante comercial.

τ: corresponde al efecto del nuevo inoculante. Se espera que este parámetro sea positivo. En tal caso el nuevo inoculante será mejor que el comercial.

ε_i: es la diferencia entre la i-ésima observación y su valor esperado. En el caso del modelo nulo el valor esperado es μ_0 y en el caso del modelo alternativo es $\mu_0 + \tau$. Este término es la discrepancia de cada observación respecto a su valor esperado y se supone que es una variable aleatoria normal con media cero y varianza σ^2 . Supondremos además que los errores son mutuamente independientes. Esta última suposición es necesaria para derivar la distribución del estadístico utilizado para contrastar los modelos nulo y alternativo.

La hipótesis nula se puede enunciar como: $H_0 : \mu = \mu_0$ mientras que la hipótesis alternativa postula que $H_1 : \mu = \mu_0 + \tau;\ \tau \neq 0$ o, equivalentemente: $H_0 : \tau = 0$ vs $H_1 : \tau \neq 0$.

Para establecer si la hipótesis nula es aceptada o no, debemos construir un estadístico cuya distribución sea conocida cuando la hipótesis nula es cierta, y que cambie de manera previsible cuando la hipótesis nula falla. Consideremos el siguiente estadístico:

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

En el numerador del estadístico Z encontramos la diferencia entre la media del peso de los plantines estimada con la muestra y el valor esperado de la media bajo la hipótesis nula (modelo nulo). En el denominador encontramos el **error estándar** de la media de peso de los plantines (obsérvese que en el denominador aparece σ^2, la varianza del término de error). Se puede demostrar que si la hipótesis nula es cierta, el estadístico Z se distribuye como una Normal estándar. La gráfica de la función de densidad Normal se muestra en la Figura 5.7.

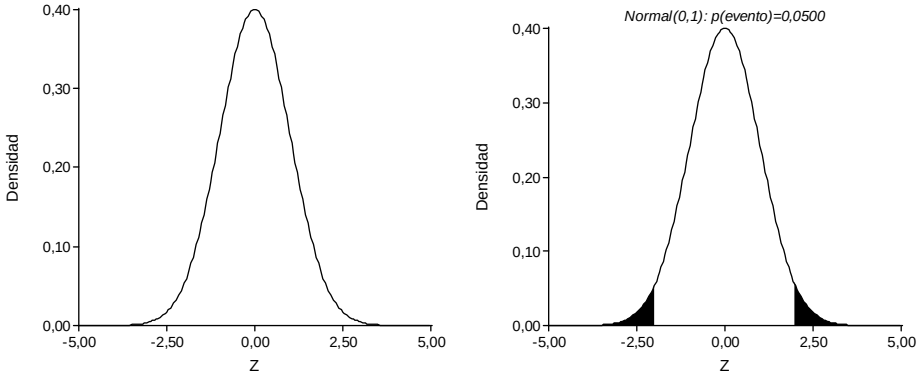


Figura 5.7: Función de densidad de una Normal estándar (gráfico de la izquierda). Función de densidad donde se ha marcado la probabilidad de la región de rechazo bajo H_0 en una prueba bilateral (gráfico de la derecha).

En la imagen de la derecha de la Figura 5.7 se han marcado dos áreas, por debajo de la curva, cuya superficie total (suma), es 0,05. Por tratarse de un área bajo la curva de densidad, el **valor 0,05** es una probabilidad que corresponde a la **probabilidad** de obtener una realización de una Normal estándar **fuera de la región** delimitada por dos puntos que corresponden a: **- 1,96 y 1,96**. La región delimitada por estos puntos se conoce como **región de aceptación** de la hipótesis nula y fuera de esta región está la **región de rechazo**. Si el estadístico Z, calculado a partir de la muestra, "cae" en la región de aceptación la hipótesis nula se acepta, sino se rechaza. Por lo tanto **0,05** es la **probabilidad de que Z se realice en la región de rechazo cuando la hipótesis nula es cierta**. Esta es otra forma de conceptualizar el **nivel de significación**: *probabilidad de que el estadístico utilizado para contrastar las hipótesis se realice en la región de rechazo cuando la hipótesis nula es cierta*. Por lo tanto, el contraste tiene un nivel de significación del 5%.

Contrastes bilateral y unilateral

En el punto anterior ejemplificamos un contraste de hipótesis bilateral. La naturaleza bilateral se origina en la forma en que la hipótesis alternativa está planteada, y tiene como consecuencia que la región de rechazo se dividida en dos partes.

Una de las formas de plantear las hipótesis del ejemplo de *Melillotus* fue: $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_0 + \tau$. En esta forma de enunciar las hipótesis τ puede asumir cualquier valor, ya sea positivo o negativo. De esta manera el investigador está indicando implícitamente que no sabe qué esperar del nuevo inoculante: puede ser tanto mejor como peor que el inoculante comercial. Si por el contrario, el investigador supusiera que el nuevo inoculante es mejor o a lo sumo igual que el comercial, entonces sus hipótesis podrían aprovechar esta información adicional y enunciarse como $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_0 + \tau; \quad \tau > 0$. Obsérvese que hemos agregado la condición de que τ es mayor que cero. Esta condición implica que el investigador espera que la media del peso de los plantines con el nuevo inoculante sea mayor que con el inoculante comercial de referencia, si la hipótesis nula falla. Volvamos al estadístico de la prueba:

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Cuando la hipótesis nula es cierta, el promedio del estadístico Z es cero. Cuando la hipótesis nula falla y la hipótesis alternativa no indica en qué sentido puede hacerlo (contraste bilateral), el promedio de Z puede ser positivo o negativo. Por ello, en ese caso el investigador debe dividir la región de rechazo en dos, poniendo una parte a la derecha y otra a la izquierda, de la región de aceptación.

Cuando la hipótesis alternativa explicita el sentido en que la hipótesis nula puede fallar, el investigador pude ubicar la región de rechazo a uno u otro lado de la región de aceptación, según corresponda. Si el promedio esperado cuando la H_0 falla es positivo, la ubicación será a la derecha; caso contrario, a la izquierda.

La anticipación del sentido en que la hipótesis nula puede fallar agrega información que puede utilizarse para construir un contraste más efectivo. Decimos más efectivo en el sentido que será capaz de **rechazar una hipótesis nula falsa** con un tamaño de muestra menor que si se aplicara un contraste bilateral. Es por ello que se dice que los contrates (pruebas) bilaterales son más conservadores.

La Figura 5.8 muestra la probabilidad de la región de rechazo para un contraste de hipótesis **unilateral derecho**, utilizando un nivel de significación del 5%. La región de aceptación queda a la izquierda del valor 1,645, que corresponde al cuantil 0,95 de una Normal estándar.

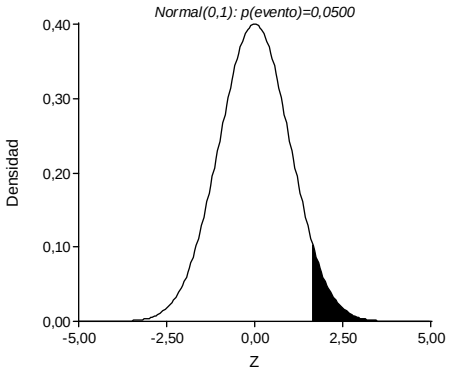


Figura 5.8: Función de densidad de una Normal estándar donde se ha marcado la probabilidad de la región de rechazo bajo H_0 en una prueba unilateral derecha.

Valor p

Supongamos que el estadístico de la prueba se llama E y que E se distribuye, cuando la hipótesis nula es cierta, con una distribución que podemos llamar D. Además supongamos que el valor del estadístico obtenido con la muestra dada es $\check{E}$. Entonces el *valor p* se calcula como $P\left(\left(E > abs(\check{E})\right) | H_0\right)$ o $2P\left(\left(E > abs(\check{E})\right) | H_0\right)$ según que la prueba sea unilateral o bilateral, respectivamente. P(.) hace referencia a la probabilidad de un evento formado por aquellos valores de E que en valor absoluto sean mayores al valor de $\check{E}$ observado en la muestra. Si el ***valor p*** es **menor que el nivel de significación** esto implica que el estadístico de la prueba se realizó en la **región de rechazo**. De allí que en la práctica moderna sólo se examina el *valor p* como criterio para decidir si la hipótesis nula es aceptada o no.

El estadístico calculado en un contraste de hipótesis se obtiene a partir de los datos de una muestra. De allí que el valor de un estadístico varía aún si tomaramos otra muestra de igual tamaño. Por lo tanto, con los datos disponibles en una muestra dada, calculamos sólo uno de todos los valores posibles. El valor p mide cuan probable es obtener, en muestreos repetido.s valores del estadístico iguales o más extremos (más pequeños o más grandes) que el calculado con la muestra dada suponiendo que la hipótesis nula fuera cierta. Si esa probabilidad es pequeña quiere decir que el estadístico calculado no está dentro de un conjunto de resultados frecuentes (región de aceptación) bajo la distribución propuesta en H_0, por lo cual concluiremos que la hipotesis nula debe rechazarser.

La Figura 5.9 muestra 3 funciones de densidad de una Normal estándar. En la primera se ha sombreado la probabilidad de la región de rechazo (nivel de significación) para una prueba unilateral derecha con un nivel de significación del 5% (Figura 5.9a). La segunda y tercera muestran dos casos de *valores p* (áreas sombreadas): uno en el que se rechaza H_0 (Figura 5.9b) y otro en el que no se rechaza (Figura 5.9c).

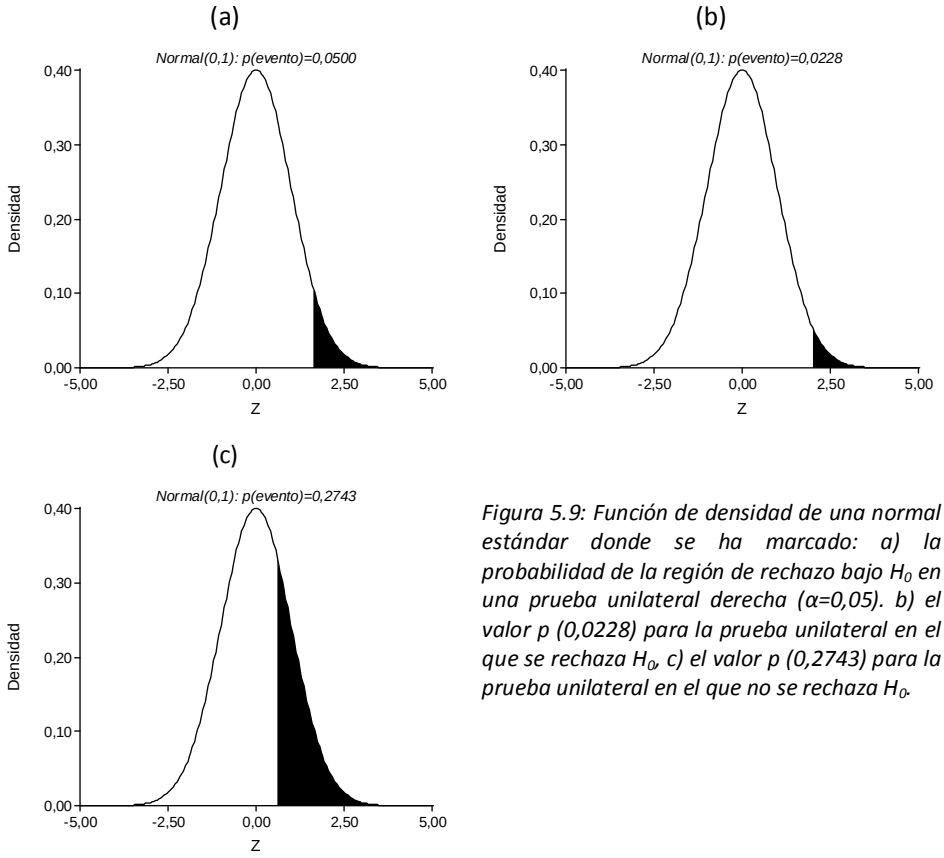


Figura 5.9: Función de densidad de una normal estándar donde se ha marcado: a) la probabilidad de la región de rechazo bajo H_0 en una prueba unilateral derecha (α=0,05). b) el valor p (0,0228) para la prueba unilateral en el que se rechaza H_0, c) el valor p (0,2743) para la prueba unilateral en el que no se rechaza H_0.

Intervalo de confianza y contraste de hipótesis

Existe una correspondencia entre los resultados del contraste de hipótesis y el intervalo de confianza para el parámetro sobre el cual se han formulado las hipótesis. Para contrastes de hipótesis simples esa correspondencia es simple y permite predecir el resultado de un contraste a partir del intervalo de confianza correspondiente. En el caso que ejemplificamos sobre el peso de plantines de *Melillotus*, si el intervalo de confianza bilateral al 95% para la media **incluyera** a μ_0 entonces esto implicaría que el contraste de hipótesis bilateral con un nivel de significación del 5% **no rechazaría** la hipótesis nula:

$H_0 : \mu = \mu_0$. De igual forma si un contraste bilateral al 5% condujera al rechazo de H_0, entonces μ_0 no quedaría incluido en el intervalo de confianza bilateral al 95%.

Potencia

Las pruebas estadísticas para el contraste de hipótesis están afectadas por el ruido o nivel de incertidumbre en el experimento. La **incertidumbre** es modelada y cuantificada por los parámetros de dispersión del modelo. Éstos capturan la variabilidad de los componentes aleatorios. Llamaremos a la incertidumbre de un modelo, en un sentido amplio: **error experimental**. Un modelo con mayor error experimental es un modelo con mayor incertidumbre y por lo tanto con menor **precisión** en sus estimaciones.

La incertidumbre es indeseable. A veces, puede controlarse desde el diseño del experimento: aumentando las repeticiones del mismo, teniendo en cuenta la heterogeneidad previsible de las unidades experimentales (bloqueo) o examinando los protocolos utilizados en busca de causas de variabilidad que puedan controlarse, capacitando a los investigadores-técnicos, utilizando nuevos instrumentos de medición, entre otras acciones.

Cuando la hipótesis nula no se rechaza puede deberse a dos causas: la hipótesis nula es cierta o el experimento no tuvo la **potencia suficiente** para detectar que la hipótesis nula es falsa. Esto último ocurre cuando el **modelo verdadero** es diferente del modelo nulo (y por lo tanto la hipótesis nula es falsa), pero la discrepancia entre ambos es pequeña y/o el tamaño del experimento es insuficiente para detectarla dada la magnitud del error experimental. La probabilidad de que un experimento de tamaño y error experimental determinados pueda detectar una discrepancia específica entre modelos se conoce como **potencia**. Esta probabilidad se representa usualmente con la letra griega π. Luego, un aspecto importante del diseño de un experimento debe contemplar el número de repeticiones necesarias para que, dado un nivel de error experimental, la prueba estadística tenga una potencia razonable para detectar una discrepancia dada (por ejemplo una potencia igual o mayor que 0,80).

Para ejemplificar, volvamos al experimento con la nueva cepa de Rhyzobium. Recordaremos que las hipótesis eran $H_0 : \mu = \mu_0$ vs $H_1 : \mu_0 + \tau;\ \tau > 0$. Con estas hipótesis asumimos que la nueva cepa, sólo puede ser igual o mejor que la cepa tradicional. Si $\tau = 2$ mg, entonces H_0 es falsa. ¿Podríamos detectar que esta hipótesis es falsa si nuestro tamaño muestral fuera de 20 plantas y la varianza del error experimental fuera de 10 mg^2? Para poder responder a esta pregunta tenemos que calcular la probabilidad de que el estadístico del contraste "se realice" en la región de rechazo, cuando $\tau = 2$ mg. Éste es el cálculo de la potencia.

Observar que no sólo decimos que la hipótesis nula es falsa, sino que estamos explicitado cuánto es el efecto de la nueva cepa del inoculante sobre la media del peso seco de los plantines. Si no realizamos esta explicitación no podemos calcular la potencia.

Hasta ahora sabemos que el estadístico de la prueba con la que estamos haciendo la ejemplificación se distribuye como una Normal estándar, cuando la hipótesis nula es cierta. Eso se explicita incluyendo un H_0 sobre el símbolo ~.

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \overset{H_0}{\sim} N(0,1)$$

Cuando la hipótesis nula falla, Z no sigue más una distribución Normal estándar sino una distribución Normal, también con varianza 1, pero desplazada en el sentido que indicado por el signo del valor esperado del numerador. Si la esperanza del numerador es positiva entonces Z es una Normal desplazada hacia la derecha (con media mayor que cero), sino estará desplazada a la izquierda (con media negativa). Para generalizar, podemos decir que: $Z = \frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2 / n}} \sim N\left(\frac{\mu - \mu_0}{\sqrt{\sigma^2 / n}}, 1\right)$

La expresión anterior indica que Z tiene distribución Normal con media igual a la diferencia estandarizada de la verdadera media de Y (μ) respecto de su media hipotética bajo hipótesis nula (μ_o) y con varianza que sigue siendo 1.

Esta distribución no depende de la hipótesis nula, pero cuando la hipótesis nula es cierta entonces la media de Z se hace cero y decimos que tiene distribución normal estándar. Esta es la forma más general de plantear la distribución del estadístico de este contraste.

Volviendo a la pregunta: ¿con qué probabilidad podríamos detectar que la hipótesis es falsa si $\tau = (\mu - \mu_0) = 2$ mg, el tamaño muestral fuera de 20 plantas y la varianza del error experimental fuera de 10 mg^2? Por el planteo del problema el contraste es unilateral derecho, por lo que si trabajamos con un nivel de significación del 5% el punto que delimita la región de aceptación y rechazo es el cuantil 0,95 de una Normal estándar. Este valor es 1,645. Luego la probabilidad de "caer" en la región de rechazo cuando la hipótesis nula falla es:

$$P\left(Z \geq 1,645 \,|\, Z \sim N\left(\frac{2}{\sqrt{10/20}}, 1 \right) \right)$$

La probabilidad que tenemos que calcular se basa entonces en una *N*(2,83;1). Esta probabilidad se muestra gráficamente en la Figura 5.10. En esta figura se observan dos curvas de densidad Normal. A la izquierda: una normal estándar. A la derecha: una N(2,83;1) correspondiente a la distribución de Z cuando τ=2 mg. El área sombreada corresponde a la probabilidad de que Z se realice en la zona de rechazo cuando Z~N(2,83;1). Esta probabilidad es la potencia de rechazar la hipótesis nula. En el ejemplo la potencia vale 0,8820. Para todo fin práctico esta es una potencia razonable.

La mayor parte de la veces no es posible anticipar el valor de τ y entonces no puede calcularse la potencia. Sin embargo, podemos proponer un conjunto plausible de valores para τ y calcular la potencia para cada uno de ellos. Luego podemos hacer un gráfico de dispersión con los valores posibles de τ en el eje X y las potencias calculadas en el eje Y. Este gráfico se conoce como **curva de potencia** y es muy útil para que el investigador pueda evaluar, bajo sus condiciones experimentales, qué sensibilidad tendrá su experimento.

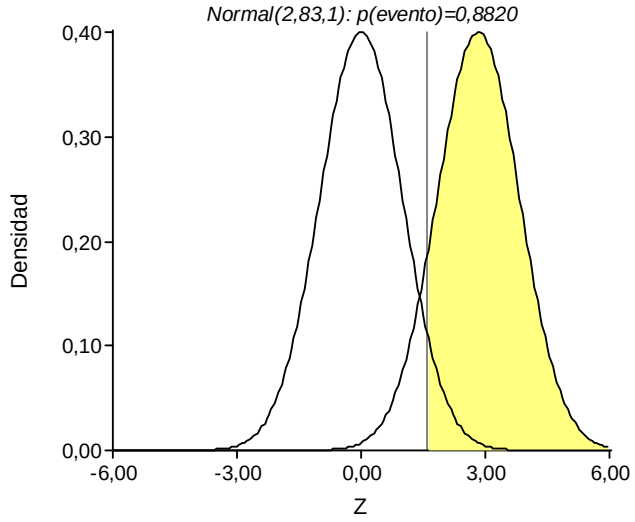


Figura 5.10: Dos curvas de densidad Normal. La que se encuentra a la izquierda del gráfico corresponde a una normal estándar. La que se encuentra a la derecha es una N(2,83;1) correspondiente a la distribución de Z cuando τ=2 mg. El área sombreada corresponde a la probabilidad de que Z se realice en la zona de rechazo cuando la distribución de Z es una N(2,83;1). Esta probabilidad es la potencia de rechazar la hipótesis nula. En el ejemplo la potencia vale 0,8820. Para todo fin práctico esta es una potencia razonable.

Estimación de parámetros y contraste de hipótesis

Para hacer la curva anterior utilizando InfoStat:

1. Abrir una nueva tabla

2. Agregar 99 nuevas filas de manera tener un total de 100 filas en la tabla. Menú Datos>>Acciones sobre filas>>Insertar nueva fila

3. Cambiar el nombre de la primera columna. La llamaremos Thau.

4. Llenar la columna Thau con una secuencia comenzando en 0 y saltando de a 0,03. Ver menú Datos>>Acciones sobre filas>>Llenar con…>> otros>>Secuencia.

5. Renombrar a la segunda columna como potencia.

6. Seleccionar del menú Datos>>Formulas.

7. En el campo de edición poner la siguiente expresión y accionar el botón calcular **potencia=1-distnormal(1,645;thau/raiz(10/20);1)**

8. Ahora hay dos columnas en el archivo de datos: la primera Thau, la segunda potencia. En el menú Gráficos seleccionar el ítem Diagrama de dispersión

El gráfico resultante se muestra en la Figura 5.11. Para valores de τ superiores a 1,75 mg, un experimento basado en 20 plantas y con una varianza del error experimental de aproximadamente 10 mg^2, tendrá una potencia 0,80 o superior.

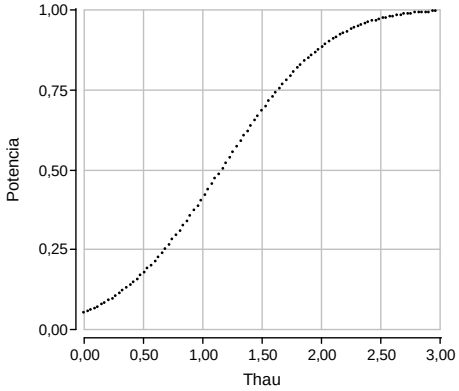


Figura 5.11: Curva de potencia en función de τ para un experimento con 20 plantas y una variabilidad experimental cuantificada por una varianza de 10 mg^2.

π *La curva anterior es útil para saber qué potencia tiene un experimento de un tamaño dado. A veces, sin embargo, se quiere saber: ¿qué tamaño debería tener el experimento (en términos del número de repeticiones) para obtener una potencia apropiada para detectar un determinado efecto de tratamiento?*

De manera similar a la curva anterior, se puede construir una curva de potencia en función de "n", dado un τ. Supongamos por ejemplo que queremos detectar valores de τ a partir de 1 mg. En la Figura 5.11 se observa que para *n*=20, la potencia para un τ =1 es menor que 0,50, así que para alcanzar una potencia de 0,80 o más tendremos que utilizar un número de repeticiones mayor. Calcularemos la potencia con tamaños muestrales crecientes a partir de n=20.
Para hacer la curva de potencia, en función de *n* , en InfoStat:

9. Abrir una nueva tabla

10. Agregar 99 nuevas filas de manera tener un total de 100 filas en la tabla. Menú *Datos>>Acciones sobre filas>>Insertar nueva fila*

11. Cambiar el nombre de la primera columna. La llamaremos "n".

12. Llenar la columna **n** con una secuencia comenzando en 20 y saltando de a 1. Ver menú *Datos>>Acciones sobre columnas>>Llenar con…>> Otros>>Secuencia.*

13. Renombrar a la segunda columna como **potencia**.

14. Seleccionar del menú *Datos>>Formulas*.

15. En el campo de edición poner la siguiente expresión y accionar el botón calcular **potencia=1-distnormal(1,645;1/raiz(10/n);1)** (Observar que ahora el lugar de Thau hay un 1 y el lugar donde ahora aparece la "n" antes había un 20).

16. Ahora hay dos columnas en el archivo de datos: la primera **n**, la segunda **potencia.** En el menú *Gráficos* seleccionar el ítem *Diagrama de dispersión*.

La curva indica que se requerirían 60 plantas para poder detectar con una probabilidad de 0,80 un $\tau = 1$mg o mayor. Si logísticamente no es posible este tamaño en un único experimento, entonces podríamos realizar varios experimentos más pequeños hasta completar el número requerido.

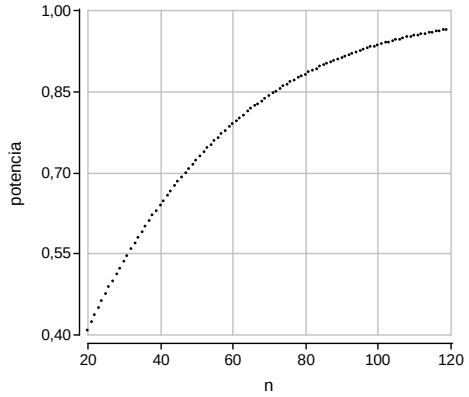


Figura 5.12: Curva de potencia en función de n para un experimento donde $\tau = 1$ mg y una variabilidad experimental cuantificada por una varianza de 10 mg^2.

Definiciones

Definición 5.1: Estimador puntual

Estadístico muestral que asigna un valor al parámetro que está estimando.

Definición 5.2: Consistencia

Propiedad de un estimador que cuando se cumple implica que la varianza y el sesgo de un estimador tienden a cero para n que tiende a infinito. Esta propiedad es una de las propiedades más importantes e implica que a mayor esfuerzo muestral, mejor es nuestra estimación.

Definición 5.3: Insesgamiento

Es una propiedad de los estimadores que, cuando se cumple, implica que dado un tamaño muestral "n" el promedio sobre todas las muestras posibles de tamaño "n" es igual al valor del parámetro estimado.

Definición 5.4: Intervalo de confianza

Región que contiene con una confianza dada al verdadero valor del parámetro estimado. La confianza se expresa en una escala porcentual y usualmente es mayor que 90%. Sus valores usuales son 95% y 99%.

Definición 5.5: Contraste de hipótesis

Comparación de una hipótesis llamada nula vs. una llamada alternativa.

Definición 5.6: Nivel de significación

Se aplica al contraste de hipótesis y es la máxima probabilidad de cometer el Error de Tipo I. O sea en el contraste de hipótesis considerado el error de Tipo I ocurre con una probabilidad igual o menor que el nivel de significación. El nivel de significación lo establece el investigador, se simboliza con la legra griega α y sus valores usuales son 0,05 y 0, 01. Cuando este nivel no se explicita se sobreentiende que es 0,05.

Definición 5.7: Hipótesis nula

En un contraste de hipótesis la hipótesis nula establece que el modelo nulo es el modelo verdadero. Esto se realiza a través de proposiciones sobre los parámetros del modelo cuyo valor de verdad debe establecerse mediante una prueba estadística apropiada.

Definición 5.8: Hipótesis alternativa

En un contraste de hipótesis la hipótesis alternativa especifica la forma en que puede fallar la hipótesis nula. Representa al modelo alternativo.

Definición 5.9: Error tipo I

Rechazar la hipótesis nula cuando es cierta.

Definición 5.10: Error tipo II

Aceptar la hipótesis nula cuando es falsa.

Definición 5.11: Valor p

Medida probabilista de confiabilidad de la hipótesis nula. Cuanto menor es el valor p menos confianza tenemos en la sustentabilidad de la hipótesis nula. Cuando el valor p es menor que el nivel de significación, el estadístico de la prueba se está realizando en la región de rechazo y por lo tanto debemos rechazar la hipótesis nula.

Definición 5.12: Potencia

Probabilidad de rechazar una hipótesis nula falsa.

Definición 5.13: Curva de potencia

Grafico de la potencia de una prueba como función del número de repeticiones en un experimento o como función de la mínima alteración de la hipótesis nula que se quiere detectar.

Análisis de regresión

Ejercicios

Ejercicio 5.1: Supongamos que se conoce que la distribución del perímetro de cabezas de ajo blanco cosechados en un establecimiento hortícola en la última campaña, sigue una distribución aproximada a una Normal con media de 18 cm y varianza de 10 cm^2 y se ha obtenido una muestra de 25 cabezas en la cual la media del perímetro es de 19 cm:

a) Si con el valor de la media muestral se desea estimar el verdadero valor del perímetro promedio de la población de ajos cosechados ¿Qué valores de la distribución de las medias de muestras de tamaño 25 conforman los límites de un intervalo de confianza al 95%?

b) Si con la muestra obtenida se desea realizar un contraste bilateral para la $H_0 : \mu = 18\ cm$ con un nivel de significación del 5% ¿Qué valores de la distribución de las medias de muestras de tamaño 25 conforman los límites de la zona de aceptación de la hipótesis nula?

c) ¿Qué concluiría con los resultados obtenidos, aumentó o no la media del perímetro de ajo?

Ejercicio 5.2: Considerar la variable rendimiento de maíz, cuya distribución es normal con media μ y desviación estándar σ. Para estimar el rendimiento promedio del maíz bajo el efecto de un herbicida, se toma una muestra de tamaño 40 y se obtiene un promedio de 60 qq/ha. Se sabe por experiencias anteriores que la varianza poblacional σ^2 es 25 $(qq/ha)^2$.

a) Construir los intervalos de confianza del 95% y 99% para μ.

b) ¿Cómo cambia el intervalo anterior (95%) si el tamaño de la muestra fuese 100 y se obtiene el mismo promedio?

c) ¿Cómo se modifica el intervalo del 95% calculado en a) si la desviación estándar fuese de 7 qq/ha?

Ejercicio 5.3: Una empresa dedicada a la comercialización de semillas desea estimar la altura promedio de un sorgo forrajero que ha desarrollado. Para ello toma una muestra de 50 plantas y se calcula la media de la altura, la que resulta ser 130 cm. Se sabe por experiencias anteriores que la desviación estándar es 22 cm.

a) Construir los intervalos de confianza para μ con una confianza del 95% y 99% respectivamente. Comparar la amplitud de ambos intervalos y concluir el eecto del nivel de confianza sobre la amplitud.

Ejercicio 5.4: Uso de la tabla de la Distribución "T" de Student.

La tabla de la distribución T de Student del anexo contiene los cuantiles $t_{p,\nu}$ para algunos valores de p, con p $\in$ [0.55, 0.995] (encabezamiento de la tabla) y gl: ν, con ν=1, 2,...,50. Suponga que se quiere calcular la P(T $\leq$ 4.3) donde T es una variable aleatoria que tiene distribución T de Student con 2 gl.

Se busca en el cuerpo de la tabla el valor 4.3 dentro de la fila que corresponde a ν=2, y en el encabezamiento de la columna se lee 0.975 que es la probabilidad buscada. El valor 4.3 es el cuantil 0.975 de la distribución T de Student con 2 gl.

Si por el contrario la probabilidad requerida hubiera sido P(T$\leq$4.3) entonces se procede de igual manera que en el párrafo anterior, pero la lectura de la probabilidad se hace en el pie de la columna. Luego P(T $\leq$-4.3) = 0.025.

Obtener las siguientes probabilidades:

a) n=50, P (T $\leq$ 2)
b) n=50, P(T > 2)
c) n=5, P(T $\leq$ -1.5)
d) ¿Cuál es el valor del cuantil 0.975 para una distribución T de Student con 5 gl? ¿Qué significa este valor?
e) ¿Cuál es el cuantil 0.30 para una distribución T de Student con 42 gl? ¿Qué significa este valor?

Ejercicio 5.5: Se desea establecer el contenido vitamínico de un alimento balanceado para pollos. Se toma una muestra de 49 bolsas y se encuentra que el contenido promedio de vitaminas por cada 100 g es $\overline{X}$ =12 mg. y que la desviación estándar S =2 mg.

f) Encontrar el intervalo de confianza del 95%, para el verdadero promedio del contenido de vitaminas.

Ejercicio 5.6: El espárrago es una planta perenne cuyo cultivo comercial puede tener una duración de 15 años y su implantación es costosa. Dada la extensión del sistema radicular, la profundidad del suelo es fundamental, considerándose indispensable contar con un promedio mínimo de 80 centímetros de sustrato permeable. Se realizan 14 determinaciones de la profundidad del sustrato permeable (en cm) en puntos tomados al azar en dos campos (A y B). Los valores registrados fueron los siguientes:

A:	72	78	86	78	90	104	76	70	83	75	90	81	85	72
B:	86	90	76	76	82	89	93	81	83	97	108	98	90	83

Los resultados del análisis estadístico fueron:

Intervalos de confianza
Bilateral- Estimación paramétrica

Campo	Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
A	Prof(cm)	Media	81.43	2.45	14	76.13	86.73
B	Prof(cm)	Media	88.07	2.42	14	82.85	93.29

a) A partir de los intervalos de confianza al 95% determinar si estos campos son aptos para el cultivo.

b) ¿Hay diferencias en la profundidad del sustrato permeable entre ambos campos? Ayuda: observar si los valores de LI y LS de ambos intervalos, se superponen.

Ejercicio 5.7: Un productor decide probar el funcionamiento de su máquina y para ello, luego de cosechar una parcela, cuenta en 10 unidades de 1 m^2 la cantidad de semillas que quedan en el suelo. Las normas técnicas indican que la media del número de semillas caídas por m^2 no debería ser superior a 80. Los resultados, en semillas/m^2, fueron:

77 73 82 82 79 81 78 76 76 75

a) Construir un intervalo de confianza para μ con una confianza del 90%.

b) Concluir sobre el funcionamiento de la máquina.

Ejercicio 5.8: Se quiere calcular el tamaño de una muestra para estimar μ en una población normal con desviación estándar igual a 13.

a) ¿Cuál debería ser el tamaño mínimo de la muestra para asegurar una amplitud de 9 unidades para el intervalo de confianza al 95%?

b) ¿Qué sucede si la confianza cambia al 99%?

Ejercicio 5.9: Para estimar el rendimiento promedio del trigo en un departamento del sur cordobés se relevan los campos de distintos productores mediante un esquema de muestreo aleatorio simple. Se conoce por experiencias anteriores que σ es igual a 0.7 qq/ha y que el promedio histórico es 26 qq/ha.

a) ¿Qué número de campos se deben evaluar para estimar la media de rendimiento con una confianza del 95% si la amplitud del intervalo no debe ser mayor que el 2.5% del promedio histórico?

b) Si la varianza de la distribución aumenta (proponga σ=1.4), ¿aumenta o disminuye el tamaño muestral necesario para mantener la misma amplitud? Justificar la respuesta.

Ejercicio 5.10: Una variable aleatoria sigue una distribución N(μ, 144) con μ desconocido.

a) ¿Se descartaría la hipótesis μ=15 en favor de la alternativa $\mu \neq 15$, para α=0.05, si de una muestra aleatoria de n=64 observaciones se obtiene una media igual a 20?

b) Construir un intervalo de confianza del 95% para μ.

c) Considerando la misma hipótesis del punto a), ¿qué sucedería con un nivel de significación del 1%?

d) Construir un intervalo de confianza del 99% para μ.

e) Probar H_0: $\mu=15$ versus H_1: $\mu>15$ para $\alpha=0.05$ y $\alpha=0.01$. Comparar con los resultados obtenidos en los puntos a) y c).

Ejercicio 5.11: Los siguientes datos corresponden a rendimientos de maíz (en kg/ha) bajo distintas densidades de siembra: baja= 50.000 plantas/ha, media= 70.000 plantas/ha y alta= 90.000 plantas/ha en dos ambientes: alta y baja productividad.

Ambiente	Baja	Media	Alta
Alto	12818	12490	11780
Alto	11869	12506	10881
Alto	12819	12502	11774
Alto	12189	12419	10578
Alto	13275	14197	13037
Alto	9405	10363	11046
Alto	10687	10144	10940
Bajo	8063	8284	7625
Bajo	8832	9703	9938
Bajo	10302	10489	10779
Bajo	9239	9525	9122
Bajo	8672	9180	9135
Bajo	10149	10442	9786
Bajo	7605	7426	7399

a) Construir intervalos de confianza bilaterales al 95% para la media poblacional de rendimientos para cada una de las densidades de siembra en los ambientes de alto y bajo rendimiento.

b) Realizar una representación gráfica de los intervalos de confianza obtenidos.

Ejercicio 5.12: Los siguientes son datos de incidencias relativas de Esclerotinia (podredumbre del capítulo). Cada dato es el cociente entre la incidencia de una línea comercial respecto de una nueva línea que se espera sea resistente. Los datos se recolectaron en 20 localidades que cubren un amplio número de condiciones ambientales. En cada localidad se obtuvieron datos de incidencia de ambas líneas comparadas.

1,91	1,60	0,83	1,44	1,78
1,75	0,68	2,24	0,81	1,50
0,94	1,45	1,14	0,13	0,53
1,44	1,60	1,58	0,92	0,73

a) ¿Es la nueva línea mejor? Observe que: bajo la hipótesis nula de igualdad de medias de incidencia, el valor esperado de la incidencia relativa es 1, pero si la línea experimental es mejor, el cociente debería aumentar (por la forma en que se propuso el índice, la nueva línea está en el denominador).
Por otra parte no contamos con un conocimiento previo de la varianza de error experimental. De este modo tendremos que estimarla a partir de los datos disponibles. En tal caso la prueba Z es aproximada. La prueba correcta es la prueba T para un parámetro. Su estadístico se muestra a continuación y la región crítica para un nivel de significación del 5% en una prueba unilateral derecha es el cuantil 0,95 de una T con 19 grados de libertad. Este cuantil, que se puede obtener de la calculadora de probabilidades y cuantiles de InfoStat es: 1,729.

$$T = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \overset{H_0}{\sim} T_{(n-1)}$$

b) Construya el intervalo de confianza (unilateral ¿izquierdo?) al 95%

c) Verifique que llegaría a la misma conclusión usando un intervalo de confianza o realizando un contraste de hipótesis.

Ejercicio 5.13:Se acepta que después de 3 años de almacenamiento el vigor de un arbusto forrajero medido como peso seco alcanzado a los 20 días de la germinación es de 45 miligramos promedio. Se propone un nuevo método de almacenamiento para aumentar el vigor. Se evalúan para ello 20 lotes de 10 semillas cada uno y al cabo de 3 años se las hace germinar, obteniéndose los siguientes resultados de peso seco promedio a los 20 días:

49	43	56	57	59	65	52	51	50	55
60	65	53	57	67	56	53	37	45	42

a) Plantear las hipótesis nula y alternativa asociadas al problema.

b) Realizar un contraste de hipótesis con un nivel de significación α=0.01.

c) De acuerdo a la conclusión que se obtuvo en el punto anterior, ¿se justifica realizar un cálculo de potencia?; ¿por qué?

Ayuda: si se tuviera que calcular la potencia con la que se realizó el contraste, acepte la varianza muestral calculada como si se tratara de la varianza poblacional y tomar a la media muestral como estimador de la verdadera media poblacional.

Ejercicio 5.14: Un tipo de ratón de laboratorio muestra una ganancia media de peso de 65 gr. durante los primeros tres meses de vida. Doce ratones fueron alimentados con una nueva dieta desde su nacimiento hasta los primeros tres meses de vida, observándose las siguientes ganancias de peso (en gr):

65	62	64	68	65	64	60	62	69	67	62	71

a) ¿Hay razón para creer que la dieta produce una variación significativa en la cantidad de peso ganado? Trabajar con α=0.05.

Ejercicio 5.15: Cuando la cantidad de semillas de soja que quedan en el suelo luego de pasar la cosechadora es igual o mayor a 80 semillas/m^2, la pérdida de producción, en qq/ha, es grande. Un productor decide probar el funcionamiento de su máquina y para ello, luego de cosechar una parcela, cuenta en 10 unidades de 1 m^2 la cantidad de semillas quedan en el suelo. Los resultados fueron, en semillas/m^2:

77 73 82 82 79 81 78 76 76 75

a) ¿Se puede concluir, trabajando con un nivel de significación del 10%, que la cosechadora está funcionando bien?, es decir, ¿está la perdida dentro de los límites admisibles?
b) Construir un intervalo de confianza para μ apropiado para el problema.

Ejercicio 5.16: Un experimentador avícola considera que al suministrar una ración especial a pollitos de la raza Cornich, ha de lograr un peso medio superior a 700 gr. por animal luego de cuatro semanas de alimentación. Para verificarlo alimenta con la ración a un lote de 50 pollitos y a los 28 días obtiene un peso promedio de 730 gr. con una desviación estándar de 40.21 gr.

c) Establecer las hipótesis nula y alternativa y realizar el contraste correspondiente utilizando α=0.05.
d) Construir un intervalo de confianza para μ.

Ejercicio 5.17: Los siguientes resultados se obtuvieron al analizar los registros de las precipitaciones ocurridas en dos zonas: A y B. Para conocer la precipitación promedio de cada zona se construyeron los correspondientes intervalos de confianza al 95%.

Zona	n	Media	DE	LI(95%)	LS(95%)
A	39	547.29	154.07	497.35	597.24
B	45	614.35	113.96	598.61	630.09

Teniendo en cuenta la información anterior responder las siguientes cuestiones, justificando la respuesta.

a) ¿Cuál sería la decisión en cada zona, al realizar un contraste de hipótesis bilateral para μ=500?

b) ¿Esperaría encontrar diferencias estadísticamente significativas entre las medias de las precipitaciones observadas en cada zona?

Ejercicio 5.18: Para evaluar la homogeneidad de la fertilidad de un suelo se tomaron alícuotas de 20 extracciones de suelo y se midió su contenido de nitrógeno. Los resultados, en ppm, fueron:

0.50 0.48 0.39 0.41 0.43 0.49 0.54 0.48 0.52 0.51
0.49 0.47 0.44 0.45 0.40 0.38 0.50 0.51 0.52 0.45

Se acepta que un suelo es homogéneo en fertilidad, si el contenido de nitrógeno presenta una varianza de a lo sumo 0.005.

a) Con los datos de la muestra, construir un intervalo de confianza apropiado (unilateral o bilateral) al 90% y evaluar a partir de él si el suelo es homogéneo o no en su fertilidad.

Capítulo 6

Comparación de dos poblaciones

Laura A. Gonzalez

6. Comparación de dos poblaciones

Motivación

En muchas situaciones de toma de decisiones, se necesita determinar si los parámetros de dos poblaciones son iguales o diferentes. Una empresa, por ejemplo, puede querer probar si sus empleadas reciben un salario menor que sus empleados por realizar el mismo trabajo. Un laboratorio puede necesitar indagar el efecto de una droga en un determinado grupo de animales frente a otro grupo. También para comparar el efecto de dos virus sobre plantas de tabaco, el aumento de peso en animales alimentados con dos pasturas diferentes. En cada uno caso se busca, más que el valor real de los parámetros, la relación entre sus valores, es decir, cuáles son las diferencias. ¿Las empleadas ganan, en promedio, menos que los empleados por hacer el mismo trabajo? ¿Un grupo de animales reacciona, en promedio, de manera diferente que otro grupo frente a un tratamiento? ¿Hay diferencias en el aumento de peso promedio de novillos alimentados con diferentes pasturas? ¿El efecto de un fungicida es mayor que otro? En este capítulo presentamos métodos estadísticos para responder preguntas referidas a la comparación (a nivel de medias) de dos poblaciones.

Conceptos teóricos y procedimientos

Distribución en el muestreo para la diferencia entre dos medias

Cuando se desea comparar dos poblaciones se usan dos muestras m_1= {Y_{11}, Y_{21},…, Y_{n1}} y m_2= {Y_{12}, Y_{22},…, Y_{n2}}, provenientes de las poblaciones 1 y 2 respectivamente.
Para el caso de medias poblacionales, nos interesa la distribución muestral de la **diferencia entre medias muestrales**. Tenemos la población 1 y la población 2 cuyos parámetros son las medias μ_1 y μ_2 y las desviaciones estándar σ_1 y σ_2 respectivamente.

Supongamos que se toma una muestra aleatoria de la distribución de la población 1, y otra muestra aleatoria de la distribución de la población 2. Si luego restamos las dos medias de las muestras, obtenemos: $\overline{Y}_1 - \overline{Y}_2$ que es la diferencia entre las dos medias muestrales.

La diferencia será positiva si $\overline{Y}_1$ es mayor que $\overline{Y}_2$, y negativa si $\overline{Y}_2$ es mayor que $\overline{Y}_1$. Al construir la distribución de todas las diferencias posibles de las muestras $\overline{Y}_1 - \overline{Y}_2$, se tiene la distribución muestral de la diferencia entre las medias muestrales. La desviación estándar de la distribución de las diferencias entre las medias de las muestras se conoce como **error estándar de la diferencia entre dos medias** y, si se conocen las varianzas poblacionales, se calcula usando la siguiente expresión:

$$\sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}$$

donde:

σ_1^2 es la varianza de la población 1

n_1 es el tamaño de la muestra de la población 1

σ_2^2 es la varianza de la población 2

n_2 es el tamaño de la muestra de la población 2

En esta comparación el valor esperado es $\mu_1 - \mu_2$, bajo la creencia de que no hay diferencias entre grupos o que la misma se supone cero o nula.

Contraste de hipótesis para la diferencia entre dos medias

Estos contrastes sirven por ejemplo para:

a) Comparar el contenido de ácidos grasos en semillas de dos variedades distintas.
b) Comparar la presión arterial de individuos antes y después de suministrarles un medicamento.
c) Comparar el efecto de dos dosis de un fungicida.
d) Comparar los porcentajes de preñez bajo dos protocolos de inseminación artificial.
e) Comparar los porcentajes de lecturas positivas para una virosis en distintas pruebas Elisa.

Los objetivos de la inferencia pueden ser:

f) Estimar la diferencia entre las medias $\mu_1 - \mu_2$ de las poblaciones de las cuales proceden.
g) Contrastar hipótesis sobre un valor postulado para la diferencia de medias poblacionales.

Por ejemplo, supongamos que un ingeniero agrónomo desea estudiar el aumento de peso en animales alimentados con dos pasturas diferentes analizando si las medias son

o no iguales, se puede utilizar una prueba de dos colas o bilateral. En este caso las hipótesis serían:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

También pueden ser reescritas como:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Si existe conocimiento sobre la relación de las medias y se quiere saber, por ejemplo, si alguna de las medias es menor o mayor que la otra, entonces se puede recurrir a pruebas de una cola o unilaterales.

Si se quiere saber si $\mu_1 < \mu_2$, el contraste será unilateral izquierdo y las hipótesis:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 < \mu_2$$

Si lo que se quiere probar es que $\mu_1 > \mu_2$, el contraste será unilateral derecho y las hipótesis:

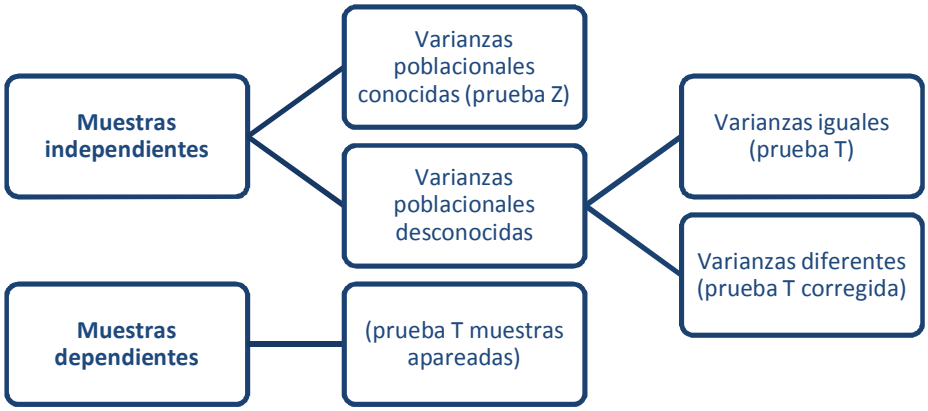
$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 > \mu_2$$

Lo que el investigador está interesado en probar va en la hipótesis alternativa, mientras que la igualdad de medias poblacionales va en la hipótesis nula.

El estadístico a usar en el contraste de medias depende de:

a) La naturaleza del muestreo (muestras independientes o apareadas)
b) Si se conocen las varianzas poblacionales
c) Si las varianzas poblacionales son iguales o diferentes

Los diferentes casos se pueden sintetizar en el siguiente esquema:



Cuando en las parcelas o unidades experimentales no se esperan respuestas diferenciales, es decir son homogéneas, se tendrán **muestras independientes**. Por ejemplo si se busca comparar el contenido de ácidos grasos en semillas de dos variedades distintas, o comparar los porcentajes de preñez bajo dos protocolos de inseminación artificial.

Si las muestras están relacionadas, esto es: los resultados del primer grupo no son independientes de los del segundo, se tendrán lo que se llaman **observaciones apareadas**. Este es el caso de la comparación de la presión arterial de individuos antes y después de suministrarles un medicamento, o si se comparan dos variedades de soja sembradas cada una en cinco localidades diferentes.

En estos últimos ejemplos, el análisis de los datos considerándolos apareados permite controlar factores externos, y así realizar un análisis más preciso. Si las muestras son independientes, los estadísticos para comparar dos poblaciones necesitan, no sólo de la diferencia de medias $\overline{Y}_1 - \overline{Y}_2$ sino también de la variabilidad de la variable estudiada en cada población. Las varianzas σ_1^2 y σ_2^2 pueden ser conocidas o no y a su vez iguales o diferentes. Analicemos ahora las diferentes situaciones.

Muestras independientes y varianzas conocidas

El estadístico será:

$$Z = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} \sim N(0,1)$$

Los límites del intervalo de confianza bilateral, con confianza 1-α, para la diferencia de medias están dados por:

$$\left(\overline{Y}_1 - \overline{Y}_2\right) \pm z_{(1-\alpha/2)} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

Por ejemplo, se montó un ensayo para comparar dos especies forrajeras en función de la producción de materia seca. El ensayo consistió en tomar 12 lotes de semillas de cada especie y hacerlas germinar, obteniéndose los siguientes valores de peso seco promedio a los 20 días (mg), archivo *[EspecieAyB]*:

Especie A	60	65	63	67	56	53	77	55	52	61	61	59
Especie B	49	45	56	57	59	65	52	51	50	62	45	48

Supongamos que se sabe que la desviación estándar poblacional es, para ambas especies, de 5 mg. La pregunta de interés es: ¿hay diferencias entre las forrajeras, a nivel del peso seco promedio? Trabajaremos con α= 0,10.

La hipótesis a plantear serían:

$$H_0 : \mu_A - \mu_B = 0 \quad \text{versus} \quad H_1 : \mu_A - \mu_B \neq 0$$

Para tener una primera descripción de los datos se obtienen los siguientes resultados, usando InfoStat:

Cuadro 6.1. Medidas resumen.

Especie	Variable	n	Media	D.E.	Mín	Máx
A	Peso seco	12	60,75	6,89	52,00	77,00
B	Peso seco	12	53,25	6,52	45,00	65,00

Como puede verse, a partir de los datos se puede calcular la desviación estándar de la variable peso seco para cada especie, sin embargo como tenemos la información de su valor poblacional, lo usamos. El estadístico para este problema se calcula de la siguiente manera:

$$Z = \frac{\left(\bar{Y}_A - \bar{Y}_B\right) - \left(\mu_A - \mu_B\right)}{\sqrt{\left(\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)}} = \frac{\left(60,75 - 53,25\right) - \left(0\right)}{\sqrt{\left(\frac{25}{12} + \frac{25}{12}\right)}} = 3,67$$

Las zonas de aceptación y rechazo de la hipótesis nula se muestran en la Figura 6.1.

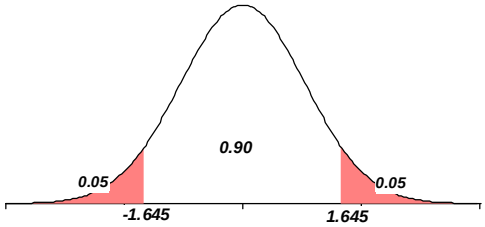


Figura 6.1: Zonas de aceptación y rechazo de la hipótesis nula, para el ejemplo de la comparación de dos forrajeras.

La región de aceptación para un nivel de significación del 10% está delimitada por los valores -1,645 y 1,645, correspondientes a los cuantiles $\alpha/2$ y $1-\alpha/2$ respectivamente, de una distribución Normal Estándar. Como Z= 3,67 es mayor que el punto crítico Z_2^* = 1,645, se rechaza la hipótesis nula de igualdad de medias poblacionales, o sea que la diferencia entre los pesos secos de las forrajeras en estudio es diferente de cero.

Esta prueba no se encuentra en el menú Estadísticas>Inferencia basada en dos muestras de InfoStat, porque no es habitual que se conozcan σ_1^2 y σ_2^2. Si se desea obtener el valor p para esta prueba, se deberá recurrir al calculador de probabilidades y cuantiles del menú Estadísticas>Probabilidades y cuantiles para obtener la P (Z>3,67)= 0,00012. Como este valor p es menor que α= 0,10 se rechaza la hipótesis nula.

Ahora bien, si la diferencia en producción de materia seca de dos especies forrajeras, transcurridos 20 días de la germinación, es superior a 10 mg, la producción de semillas esperada al final de la cosecha, será diferente. ¿Qué se puede decir con respecto a esta afirmación?
Para contestar esta pregunta recurriremos al intervalo de confianza:

$$\left(\bar{Y}_1 - \bar{Y}_2\right) \pm z_{(1-\alpha/2)} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = 7,5 \pm 1,645\left(2,0412\right) = \left[4,14;10,85\right]$$

Vemos que el valor 10 mg está incluido en el intervalo de confianza, con lo cual se puede afirmar que la diferencia en producción de materia seca, entre ambas forrajeras, no es superior a 10 mg. Entonces se concluirá que, si bien a los 20 días de germinación de las semillas hay diferencias en la producción de materia seca entre las especies, la diferencia no es superior a 10 mg, con lo cual la producción de semillas esperada al final de la cosecha, no será diferente.

Muestras independientes y varianzas poblacionales desconocidas e iguales

En el caso que σ_1^2 y σ_2^2 sean desconocidas, se podrán estimar usando las varianzas muestrales S_1^2 y S_2^2. Hay dos estadísticos diferentes para este caso, es por ello que debemos averiguar si las varianzas son iguales o diferentes. Para saberlo deberemos plantear las siguientes hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Suponiendo normalidad para las observaciones de ambas muestras, la prueba de homogeneidad de varianzas se basa en el siguiente estadístico:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1,n_2-1)}$$

Bajo H_0 este estadístico se distribuye como una F con n_1-1 y n_2-1 grados de libertad.

La conclusión la obtendremos con el valor p para el contraste de homogeneidad de varianzas, que hallaremos con el nombre "pHomVar", en la salida de InfoStat.

Si con la prueba anterior se concluye que las varianzas son iguales, para la inferencia de las medias usaremos el siguiente estadístico:

$$T = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim T_{n_1+n_2-2}$$

donde: $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

Los límites del intervalo de confianza bilateral, con confianza 1-α, para la **diferencia de medias** están dados por:

$$(\overline{Y}_1 - \overline{Y}_2) \pm t_{(1-\alpha/2);n_1+n_2-2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Por ejemplo, tenemos el caso donde se busca comparar dos formulaciones de un mismo insecticida sobre el porcentaje de mortalidad de chinche verde evaluada como número de insectos muertos de un total de 100 iniciales. El ensayo se realizó tomando 20 lotes de 100 insectos cada uno y asignando al azar 10 lotes para la formulación A y el resto para la formulación B. Los valores obtenidos fueron los siguientes y se encuentran en el archivo *[FormulaciónAyB]*:

Formulación A	85	86	92	87	92	90	95	90	92	91
Formulación B	87	86	84	80	89	85	92	89	86	90

¿Existen diferencias estadísticamente significativas entre formulaciones considerando la mortalidad promedio de los insectos? Trabajando con α= 0,05 y postulando las hipótesis como:

$$H_0 : \mu_A = \mu_B \quad \text{versus} \quad H_1 : \mu_A \neq \mu_B$$

Realizaremos una prueba **T para observaciones independientes** usando InfoStat (menú Estadísticas > Inferencia basada en dos muestras > Prueba T).

Cuadro 6.2. Prueba T para muestras Independientes (varianzas iguales)

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media (2)
Form	Mort	{A}	{B}	10	10	90,00	86,80

LI(95)	LS(95)	Var(1)	Var(2)	pHomVar	T	gl	p-valor	Prueba
0,12	6,28	9,78	11,73	0,7904	2,18	18	0,0426	Bilateral

Observando los resultados, para el contraste de hipótesis de igualdad de varianzas el valor p indica que las varianzas son homogéneas (pHomVar= 0,7904 es mayor que α= 0,05). El estadístico T= 2,18 que figura en la salida fue calculado con la expresión llamada T y los grados de libertad (gl) fueron calculados como: n_1+n_2-2.

Para la prueba de medias el **valor p** (en la salida se encuentra como valor p), es igual a 0,0426 resulta menor que α= 0,05 indica el rechazo de la hipótesis de igualdad de

medias. Es decir, hay diferencias estadísticamente significativas entre ambas formulaciones considerando la mortalidad de los insectos.
¿Cuál es la diferencia promedio en mortalidad entre las dos formulaciones? Para responder a esta pregunta se utiliza el intervalo de confianza para la diferencia de medias: LI(95)= 0,12 y LS(95)= 6,28.
Observemos que los límites de intervalo de confianza para la diferencia son positivos, esto indicaría que una diferencia positiva entre ambas formulaciones, es decir, la formulación A presenta mayor mortalidad promedio. Analicemos ahora el intervalo de confianza para la mortalidad de formulación A (menú Estadísticas > Inferencia basada en una muestra > Intervalos de confianza).

Cuadro 6.3. Intervalos de confianza.

Bilateral - Estimación paramétrica

Form	Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
A	Mort	Media	90,00	0,99	10	87,76	92,24

Para la formulación A, los valores de mortalidad estarán entre 87,76 y 92,24.

Muestras independientes y varianzas poblacionales desconocidas y diferentes

El estadístico que usaremos es: $T' = \dfrac{\left(\overline{Y}_1 - \overline{Y}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)}} \sim t_v$

donde: $\nu = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$ representa los grados de libertad.

Los límites del intervalo de confianza bilateral, con confianza 1-α, para la diferencia de medias están dados por:

$$\left(\overline{Y}_1 - \overline{Y}_2\right) \pm t_{(1-\alpha/2);\nu} \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

Por ejemplo, un laboratorio está interesado en estudiar la disminución de la actividad enzimática (medida en unidades internacionales) de una reacción con calor respecto a

la misma reacción en frío. La actividad enzimática se observa en 10 tubos con calor y 10 con frío. Los datos se encuentran en el archivo *[FríoCalor]*. Los resultados fueron:

Temp.	Activ.Enz.
Calor	7,61
Calor	7,64
Calor	7,57
Calor	7,60
Calor	7,76

Temp.	Activ.Enz.
Calor	7,51
Calor	7,66
Calor	7,54
Calor	7,46
Calor	7,66

Temp.	Activ.Enz.
Frío	7,00
Frío	7,16
Frío	6,99
Frío	6,87
Frío	7,61

Temp.	Activ.Enz.
Frío	6,80
Frío	7,19
Frío	6,98
Frío	7,27
Frío	6,87

¿Existen diferencias estadísticamente significativas entre ambas condiciones de temperatura analizando la actividad enzimática? (α= 0,05).
Las hipótesis que plantearemos son:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

Realizando una prueba T para observaciones independientes con InfoStat (menú Estadísticas > Inferencia basada en dos muestras > Prueba T), obtenemos:

Cuadro 6.4. Prueba T para muestras Independientes (varianzas diferentes).

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)
Temp.	Activ.Enz.	{Calor}	{Frío}	10	10	7,60	7,08

LI(95)	LS(95)	Var(1)	Var(2)	pHomVar	T	gl	p-valor	Prueba
0,35	0,70	0,01	0,06	0,0053	6,48	11	<0,0001	Bilateral

Analizando los resultados, para el contraste de **hipótesis de igualdad de varianzas** el valor p indica que las varianzas no son homogéneas (pHomVar= 0,0053 es menor que α= 0,05). El estadístico T= 6,48 que figura en la salida fue calculado con la expresión llamada T' y los grados de libertad fueron calculados con la expresión llamada "ν", que corresponde al ajuste de los grados de libertad, necesario en este caso. Note que si las varianzas hubieran sido homogéneas, esta prueba tendría 18 grados de libertad, pero sólo se usaron 11 (gl= 11). La diferencia (7 grados de libertad) es el costo que se pagó por tener varianzas heterogéneas.
Para la prueba de medias el *valor p*< 0,0001 es menor que α= 0,05, lo que indica el rechazo de la hipótesis nula de igualdad de medias. Es decir, hay diferencias estadísticamente significativas entre ambas condiciones de la reacción enzimática utilizando la medida de unidades internacionales.
¿Cuál es la magnitud de la diferencia entre las dos condiciones de reacción?
Para responder a esta pregunta se utiliza el intervalo de confianza para la diferencia de medias: LI(95)= 0,35 y LS(95)= 0,70.
Dado que los límites de intervalo de confianza para la diferencia son positivos, se infiere que la reacción con calor produce mayor actividad enzimática que con frío.

Como se hallaron diferencias entre las reacciones, sería de interés analizar el intervalo de confianza para la media, en la condición de temperatura que produce mayor actividad. Para hallar el intervalo requerido, se recurre a InfoStat (menú Estadísticas > Inferencia basada en una muestra > Intervalos de confianza). El resultado es:

Cuadro 6.5. Intervalos de confianza.

Bilateral - Estimación paramétrica

Temp.	Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
Calor	Activ.Enz.	Media	7,60	0,03	10	7,54	7,66

Observemos entonces que si la reacción es llevada a cabo con calor, los valores de reacción estarán entre 7,54 y 7,66 unidades internacionales.

Muestras dependientes

En este caso, los datos se obtienen de muestras que están relacionadas, es decir, los resultados del primer grupo no son independientes de los del segundo. Dadas las muestras m_1 y m_2 consideremos una muestra de las diferencias entre los datos de cada muestra:

$m_d = \{Y_{11} - Y_{12}, Y_{21} - Y_{22}, \ldots, Y_{n1} - Y_{n2}\} = \{D_1, D_2, \ldots, D_n\}$ (observemos que $n_1 = n_2 = n$)

La prueba T para muestras apareadas es aplicable en el caso que las observaciones de m_1 y m_2 se obtengan de a pares, como por ejemplo mediciones de monóxido a la mañana y tarde de un mismo día. También cuando se mide la presión arterial en cada uno de los individuos de un grupo experimental antes y después de la administración de una droga. Estas observaciones no son independientes ya que la presión arterial posterior a la administración de la droga depende de la presión arterial inicial.

La inferencia se basa en un estadístico que se conoce como **prueba T para muestras apareadas** y que depende de la media y la varianza de las diferencias y del valor hipotetizado para el promedio poblacional de las diferencias (δ). Las hipótesis que podríamos plantear son:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

o bien:

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

donde δ se refiere al promedio poblacional de las diferencias entre los valores de la variable de ambos grupos, bajo la hipótesis nula. El estadístico usado es:

$$T = \frac{\left(\bar{D} - \delta\right)}{\sqrt{\left(\frac{S_D^2}{n}\right)}} \sim t_{n-1}$$

donde $\bar{D}$ es la media muestral de las diferencias y S_D^2 la varianza muestral de las diferencias.

Los límites del intervalo de confianza bilateral, con confianza 1-α, para la diferencia de medias están dados por:

$$\left(\bar{D}\right) \pm t_{(1-\alpha/2);n-1}\sqrt{\left(\frac{S_D^2}{n}\right)}$$

Por ejemplo, para evaluar el crecimiento (medido en peso seco) de plantines de repollo sometidos a dos condiciones hídricas, una con riego no restringido (a capacidad de campo) y la otra con riego restringido (una vez cada 15 días), ocho equipos de trabajo obtuvieron datos para ambas condiciones. Cada dato, aportado por un equipo de trabajo corresponde al peso seco promedio de 50 plantas. Archivo *[RepolloRiegoRyNR].*

Se muestra a continuación los datos y las diferencias de peso seco entre los valores de Riego NR y Riego R, para cada equipo.

Equipo	1	2	3	4	5	6	7	8
Riego NR	0,487	0,408	0,360	0,431	0,576	0,660	0,400	0,540
Riego R	0,387	0,820	0,788	0,889	0,578	0,680	0,410	0,550
Diferencias	0,1	-0,412	-0,428	-0,458	-0,002	-0,02	-0,01	-0,01

¿Es la diferencia de peso seco entre condiciones de riego estadísticamente significativa, para un nivel de significación del 5%?

Las hipótesis:

$$H_0 : \mu_R - \mu_{NR} = 0 \quad \text{versus} \quad H_1 : \mu_R - \mu_{NR} \neq 0$$

Realizando una prueba T para observaciones apareadas con InfoStat (menú Estadísticas > Inferencia basada en dos muestras > Prueba T apareada), obtenemos:

Cuadro 6.6. Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	LI(95%))	LS(95%))	T	Bilateral
Riego R	Riego NR	8	0,16	0,23	-0,04	0,35	1,88	0,1023

Para la prueba de medias el valor p= 0,1023 es mayor que α= 0,05, indicando el no rechazo de la hipótesis de igualdad de medias. Es decir, no hay diferencias estadísticamente significativas entre ambas situaciones de riego. Los límites del intervalo de confianza (con 95% de confianza) para la diferencia de medias son LI(95%)= -0,04 y LS(95%)= 0,35, como el intervalo incluye el cero concluimos que no existe diferencia entre ambas condiciones.

Aplicación

Rendimiento según época de cosecha

En un estudio para analizar la evolución de tubérculos almacenados, se deseaba comparar dos épocas de cosecha: abril y agosto, las que determinan diferentes periodos de almacenamiento. La variable en estudio fue la pérdida de peso por deshidratación (en gramos). El archivo *[Epoca]* contiene las observaciones del estudio.

Época	Peso
Abril	35,56
Abril	36,89
Abril	47,05
Abril	44,36
Abril	42,05

Época	Peso
Abril	43,58
Abril	37,63
Abril	40,21
Abril	39,98
Abril	41,54

Época	Peso
Agosto	33,25
Agosto	27,75
Agosto	32,15
Agosto	21,16
Agosto	25,19

Época	Peso
Agosto	23,42
Agosto	26,87
Agosto	22,36
Agosto	24,13
Agosto	30,22

Estrategia de análisis

Lo primero que se debe decidir es el tipo de observaciones que se tienen, para este problema la naturaleza del estudio indica que son datos independientes dado que hay dos épocas de almacenamiento de los tubérculos. Las hipótesis podrían ser:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 \neq 0$$

Con InfoStat (menú Estadísticas > Inferencia basada en dos muestras > Prueba T), obtenemos los siguientes resultados:

Cuadro 6.7. Prueba T para muestras Independientes

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)
Epoca	Peso	{Abril}	{Agosto}	10	10	40,89	26,65

LI(95)	LS(95)	Var(1)	Var(2)	pHomVar	T	gl	p-valor	Prueba
10,59	17,88	12,81	17,25	0,6648	8,21	18	<0,0001	Bilateral

Si analizamos la salida para el contraste de hipótesis de igualdad de varianzas, el valor p indica que las varianzas son homogéneas (pHomVar= 0,6648 es mayor que α= 0,05). El estadístico T= 8,21 para la prueba de medias arroja un valor p= 0,0426 es menor que α= 0,05, lo que indica el rechazo de la hipótesis de igualdad de medias. Por lo tanto, podemos afirmar que hay diferencias estadísticamente significativas entre ambas épocas de almacenamiento cuando se analiza el peso de los tubérculos.

Para encontrar la diferencia de peso promedio entre ambas épocas utilicemos el intervalo de confianza para la diferencia de medias. Así se puede ver que la diferencia de peso estará entre 10,59 y 17,88 gramos con una confianza del 95%. Como los límites

de intervalo de confianza para la diferencia son positivos se observa que en abril se presentan tubérculos con mayor peso promedio. Analicemos ahora los intervalos de confianza (menú Estadísticas > Inferencia basada en una muestra > Intervalos de confianza), para el peso de los tubérculos en cada época:

Cuadro 6.8. Intervalos de confianza.

Bilateral- Estimación paramétrica

Epoca	Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
Abril	Peso	Media	40,89	1,13	10	38,32	43,45

En abril, los valores de peso promedio estarán entre 38,32 y 43,45 g.

Cuadro 6.9. Intervalos de confianza

Bilateral- Estimación paramétrica

Epoca	Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
Agosto	Peso	Media	26,65	1,31	10	23,68	29,62

En agosto, los valores de peso promedio estarán entre 23,68 y 29,62 g.

Conclusión

Se puede decir que considerando las épocas de abril y agosto, es recomendable hacer el almacenamiento de tubérculos en abril, ya que se obtiene menor pérdida por deshidratación. Los pesos promedios estarán entre 38,32 y 43,45 g para tubérculos almacenados en abril mientras que en agosto los valores estarán entre 23,68 y 29,62 g.

Calidad de semilla bajo dos sistemas de polinización

Se efectuó un experimento en plantas de lúpulo, para estudiar el efecto de la polinización sobre el peso promedio de las semillas obtenidas. Se usaron 10 plantas, la mitad de cada planta fue polinizada y la otra mitad no. Se pesaron las semillas (en gramos) de cada mitad por separado, registrándose de cada planta un par de observaciones. El archivo *[Poliniza]* contiene los valores registrados.

Polinizado	0,79	0,77	0,44	0,61	0,47	0,56	0,93	0,87	0,58	0,69
No polinizado	0,22	0,13	0,33	0,26	0,14	0,16	0,28	0,31	0,21	0,15

Estrategia de análisis

Este problema registra observaciones apareadas ya que se pesaron semillas de la parte sin polinizar y polinizadas en cada planta.
Las hipótesis que podemos plantear son:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 \neq 0$$

Con InfoStat (menú Estadísticas > Inferencia basada en dos muestras > Prueba T apareada), obtenemos:

Cuadro 6.10. Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	Media(1)	Media(2)	DE(dif)
Poliniz.	NoPoliniz.	10	0,45	0,67	0,22	0,17

LI(95%))	LS(95%))	T	Bilateral
0,33	0,57	8,42	<0,0001

Para la prueba de medias el valor p< 0,0001 es menor que α= 0,05, lo que indica el rechazo de la igualdad de medias. Es decir, hay diferencias estadísticamente significativas entre ambas condiciones de polinización.

El intervalo de confianza para la diferencia de medias: LI(95%)= 0,33 y LS(95%)= 0,57 indica que la diferencia entre ambas condiciones con una confianza del 95%. Como los límites de intervalo de confianza para la diferencia son positivos se puede afirmar que las plantas polinizadas producen un mayor peso promedio de semillas.

Para hallar el intervalo sólo para las plantas polinizadas, se recurre a InfoStat (menú Estadísticas > Inferencia basada en una muestra > Intervalos de confianza). El resultado es:

Cuadro 6.11. Intervalos de confianza

Bilateral- Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(95%))	LS(95%))
Poliniz.	Media	0,67	0,05	10	0,55	0,79

Conclusión

Para el lúpulo es recomendable usar la técnica de polinización ya que la misma produce mayor cantidad de semillas. Los pesos promedios esperados de las semillas estarán entre 0,55 y 0,79 gramos.

Ejercicios

Ejercicio 6.1: En un ensayo de biotecnología reproductiva se compararon dos productos, A y B, que se utilizan para el control de la dinámica folicular y cuya finalidad es sincronizar el día, del ciclo ovulatorio de las vacas, en el que cesa el crecimiento del folículo y comienza la regresión. La medición se realiza por ultrasonografía. Un producto se considera mejor que otro si la varianza de la variable "día en que se produce la regresión" es menor. Así, si la varianza es igual a cero implicaría sincronización total, es decir en todas las vacas se produce el evento en el mismo día.

Producto A	3	5	6	2	5	3	2	5	4	6	4	5
Producto B	3	3	2	3	3	3	3	2	3	2	3	3

a) Contrastar la hipótesis que establece que la varianza de la variable en la población que recibe el producto B es menor que la varianza de la variable en la población que recibe producto A. Utilizar un nivel de significación del 5% y el menú Probabilidades y Cuantiles de InfoStat para encontrar los puntos críticos.

Ejercicio 6.2: Dos lotes de pollos de la misma raza y edad fueron alimentados durante 30 días con dos tipos diferentes de alimento balanceado. Los aumentos de peso, en gramos, fueron:

Balanceado A	329	363	298	243	391	333	369	432	440	397	409	350
Balanceado B	353	405	372	345	377	409	428	421	357	372	409	367

a) Pruebar si existen diferencias estadísticamente significativas entre los aumentos de peso promedio de los dos lotes. Trabaje con un nivel de significación de 5%.

b) Estimar la diferencia entre las medias de los tratamientos, con una confianza del 95%. ¿Recomendaría algún balanceado?

Ejercicio 6.3: Una empresa semillera quiere comparar el desempeño de dos variedades de maíz en una amplia región para la cual ambas variedades están recomendadas. Para realizar el ensayo se dispone que en cada una de las 6 estaciones experimentales que la empresa tiene en la zona se siembren dos parcelas, una para cada variedad. Al final del ciclo del cultivo se obtuvieron los siguientes rendimientos (qq/ha):

Estación experimental	1	2	3	4	5	6
Variedad A	50	60	55	40	48	52
Variedad B	52	61	57	42	48	54

a) Para hacer el contraste ¿utilizaría una prueba T para muestras independientes o una prueba T apareada?

b) ¿Qué supuestos se deben cumplir para que la prueba sea válida?

c) ¿Es la diferencia de rendimientos entre variedades estadísticamente significativa, para un nivel de significación del 1%?

d) Construir un intervalo de confianza al 99% para la diferencia de medias.

Ejercicio 6.4: Se está experimentando con un herbicida en maíz, y para ponerlo a prueba se evalúan los rendimientos de 12 parcelas experimentales. En 6 de ellas se utilizó el nuevo herbicida y en las restantes un herbicida tradicional como control. Los resultados del ensayo, expresados en quintales por hectárea, son los siguientes:

Nuevo herbicida	66.02	70.62	64.37	65.17	64.58	61.33	62.11	62.75	58.41	69.63
Tradicional	62.34	67.18	67.10	55.74	59.00	57.78	64.25	60.31	63.05	60.07

a) Para hacer el contraste ¿utilizaría una prueba T para muestras independientes o una prueba T apareada?

b) ¿Qué supuestos se deben cumplir para que la prueba sea válida?

c) ¿Qué se puede decir del desempeño del nuevo herbicida en relación al control, trabajando con un nivel de significación $\alpha = 0.10$?

d) Construir un intervalo de confianza para la diferencia de medias poblacionales.

e) Si después de analizar los datos, encuentra que el estadístico usado pertenece a la región de no rechazo de la hipótesis nula, ¿cuál de las siguientes opciones representa mejor el resultado obtenido? Justificar la respuesta.

d) Ambos herbicidas producen el mismo efecto sobre el rendimiento.

e) Los herbicidas producen distinto efecto sobre el rendimiento.

f) Los herbicidas no producen efectos sobre el rendimiento.

g) Ninguna de las anteriores.

f) ¿Cuál sería la potencia que se alcanzaría con 10 repeticiones por tratamiento y si se busca detectar una diferencia entre herbicidas de 5 qq/ha?

Ejercicio 6.5: Un grupo de conejos fue sometido a una serie de situaciones de tensión que producían una respuesta de temor. Después de un período de tiempo bajo estas condiciones, los conejos fueron comparados con los de un grupo control, que no había sido sometido a tensión. La variable de respuesta fue el peso (en mg) de la glándula suprarrenal. Los resultados fueron:

Experimental	3.8	6.8	8.0	3.6	3.9	5.9	6.0	5.7	5.6	4.5	3.9	4.5
Control	4.2	4.8	2.3	6.5	4.9	3.6	2.4	3.2	4.9	4.8		

a) Comparar el peso de la glándula suprarrenal entre el grupo control y el experimental con un nivel de significación del 1%.

b) Construir un intervalo de confianza para la diferencia de medias poblacionales.

Ejercicio 6.6: Para probar la eficacia de un tratamiento de poda en un bosque de Raulí, un investigador decide comparar el incremento del diámetro de los fustes de los árboles podados, con el incremento en árboles sin poda. Para ello se localizan 20 lotes de los cuales a 10 se los poda y al resto no. Al cabo de 3 años se obtienen los incrementos promedio para cada lote siendo los resultados los siguientes (en cm):

Con poda	0.290	0.305	0.280	0.320	0.350	0.297	0.300	0.298	0.315	0.324
Sin poda	0.300	0.303	0.270	0.300	0.320	0.310	0.280	0.302	0.298	0.301

a) ¿Cuál es el efecto de la poda? Trabaje con un nivel de significación del 5%.

Ejercicio 6.7: La siguiente tabla presenta los resultados de una experiencia conducida para probar la hipótesis de que una dieta rica en lecitina favorece la producción de leche, en vacas de la raza Holando-Argentino. En este experimento se seleccionaron 18 tambos homogéneos en cuanto al manejo, de los cuales 9 fueron asignados aleatoriamente para recibir un suplemento de lecitina y los restantes actuaron como control. Debido a fallas en el seguimiento de uno de los tambos que no recibía el suplemento de lecitina, sus datos fueron descartados. Los resultados, expresados en lts/día promedio por vaca son los siguientes:

Sin lecitina	13.0	14.5	16.0	15.0	14.5	15.2	14.1	13.3	
Con lecitina	17.0	16.5	18.0	17.3	18.1	16.7	19.0	18.3	18.5

Sean μ_{SL} la media de producción diaria de leche para animales de la raza Holando Argentino alimentados normalmente y μ_{CL} la media de producción de los animales alimentados con una dieta rica en lecitina.

a) En base a los datos experimentales verificar la hipótesis: H_0: $\mu_{CL}=\mu_{SL}$ vs. H_1: $\mu_{CL}>\mu_{SL}$ (utilizar $\alpha= 0.05$).

Ejercicio 6.8: Un investigador supone que el estrés que se produce en vacas fistuladas puede disminuir los niveles de fósforo en sangre. Para probar su hipótesis selecciona 8 vacas y a cada una de ellas le extrae una muestra de sangre antes de la fistulación y otra muestra después. Los resultados son:

Vaca	1	2	3	4	5	6	7	8
Antes de la fistulación.	8.69	7.13	7.79	7.93	7.59	7.86	9.06	9.59
Después de la fistulación	7.24	7.10	7.80	7.95	7.50	7.79	9.00	9.48

a) ¿Qué conclusión se puede extraer acerca de la fistulación? Utilizar $\alpha= 0.01$.

Estimación de parámetros y contraste de hipótesis

Ejercicio 6.9: Un criadero de semillas interesado en evaluar el comportamiento bajo riego de 2 híbridos de maíz realizó el siguiente ensayo: se tomaron 2 surcos de 50 m. y se delimitaron 10 sectores de 5 m. cada uno. Se sabe que el perfil de infiltración del agua es distinto a lo largo del surco de riego. Para evitar que este factor afecte la evaluación del rendimiento de los híbridos, en cada uno de los sectores de 5 metros de surco se asignaron aleatoriamente cada uno de ellos. Los datos obtenidos en qq/ha fueron:

Sector	1	2	3	4	5	6	7	8	9	10
Híbrido 1	123	121	119	115	111	105	106	114	120	127
Híbrido 2	127	130	118	117	114	110	115	120	125	133

a) Concluir acerca del comportamiento de los híbridos bajo riego. Utilizar $\alpha = 0.05$.

Ejercicio 6.10: En un experimento se estudió el efecto de dos métodos (A y B) de escarificación del tegumento, sobre la viabilidad de las semillas. De un conjunto de 100 semillas se eligieron al azar 50 que fueron tratadas con uno de los métodos y las restantes se trataron con el otro método. En cada tratamiento se determinó el porcentaje de semillas no viables. En el análisis de los datos con Infostat se reportaron los resultados que se detallan (α=0.10). En función de éstos asignar el valor de Verdadero (V) o Falso (F) a cada una de las consignas del cuadro.

Grupo (1)	Grupo (2)	Media (1)	Media (2)	LI (90%)	LS (90%)	P (Var.Hom.)	T	P(prueba Bilateral)
Mét. A	Mét. B	8.00	8.87	1.58	-0.17	0.0151	2.22	0.046

I. De acuerdo al experimento, los datos deben analizarse con una prueba T para observaciones apareadas	
II. Para la prueba T no fue necesario ajustar los grados de libertad	
III. El valor 8.00 (en la salida se presenta como media(1)), es una estimación puntual del porcentaje de semillas no viables obtenido con el método A	
IV. Los resultados muestran que la varianza del porcentaje de semillas no viables bajo el método A es diferente a la varianza obtenida usando el método B	
V. Con un nivel de confianza de 90% se puede esperar que la diferencia entre las medias del porcentaje de semillas no viables sea superior a 1.58%	
VI. Para comparar los porcentajes de semillas no viables de ambos métodos, la hipótesis nula del contraste establece que los promedios poblacionales son iguales a cero	
VII. Para el contraste de medias el valor p= 0.0467, sugiere que la probabilidad de que las diferencias observadas sean por azar es menor que 0.10	
VIII. Los límites del intervalo de confianza son los puntos críticos del contraste realizado, para un nivel de significación de 0.10	
IX. El contraste realizado indica que la diferencia entre las medias es significativamente mayor a 1.58%	
X. Como hay diferencias entre las medias y los límites del intervalo de confianza son negativos se infiere que el promedio de semillas no viables con el método A es mayor	

Capítulo 7

Análisis de regresión

Julio A. Di Rienzo

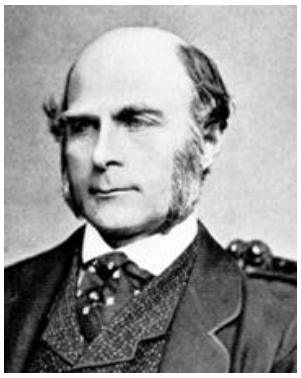
7. Análisis de regresión

Motivación

Muchas veces estamos interesados en describir cómo cambia una variable (que llamaremos dependiente) en función de una (o varias) llamada/s independiente/s. Por ejemplo: ¿cómo afecta al rendimiento del maíz la densidad de siembra en distintos ambientes?, ¿qué dosis de insecticida es requerido para eliminar el 50 de una población de insectos?, ¿cómo responden los rendimientos del trigo a diversas dosis de fertilización nitrogenada?, ¿cuánto más fertilización es siempre mejor?, ¿el efecto de la fertilización es el mismo en cualquier ambiente?, ¿bajo qué condiciones se produce el máximo número de bacterias por cm^3 de cultivo de bacterias? Para responder estas preguntas los investigadores ajustan modelos de regresión a experimentos diseñados o a estudios observacionales. Primeramente abordaremos el modelo de regresión lineal simple, luego introduciremos el modelo de regresión lineal múltiple.

Conceptos teóricos y procedimientos

El análisis de regresión involucra un conjunto de técnicas estadísticas cuyo propósito es la construcción de un modelo para la estimación de la media de una **variable dependiente** a partir de una variable o varias **variables independientes** o también llamadas **regresoras**. Por ejemplo si el propósito fuera establecer la forma en que el rendimiento del maíz es afectado por la densidad de siembra, el rendimiento correspondería a la variable dependiente y la densidad de siembra a la regresora. La variable dependiente se simboliza, usualmente, con la letra "Y" y las variables independientes con la letra x (si hay más de una se enumera x_1, x_2,…).
Genéricamente diremos que las observaciones de la variable dependiente varían según una función $f(.)$ que depende de la/s variable/s independiente/s. Esta función está caracterizada por un conjunto de parámetros (desconocidos) representados por el vector de parámetros $\mathbf{b}$.



Francis Galton

El término regresión fue introducido por Francis Galton en su libro Natural inheritance (1889) y fue confirmado por su amigo Karl Pearson. Su trabajo se centró en la descripción de los rasgos físicos de los descendientes (variable Y) a partir de los de sus padres (variable X). Estudiando la altura de padres e hijos a partir de más de mil registros de grupos familiares, se llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que revelaban también una tendencia a regresar a la media. Fuente: Wikipedia

La dependencia de $f(.)$ de las regresoras $\mathbf{x}$'s y del conjunto de parámetros $\mathfrak{b}$ se indica escribiendo $f(\mathbf{x},\mathfrak{b})$. Para tener en cuenta que las observaciones de Y no son idénticas cuando los valores de x sí lo son, se suma a $f(\mathbf{x},\mathfrak{b})$ un término, conocido como **error** y que se simboliza con ε. Los errores son perturbaciones aleatorias propias de cada observación Y. Luego la i-ésima observación de la variable dependiente se puede representar de acuerdo al siguiente modelo estadístico.

$$Y_i = f(\mathbf{x}_i,\mathfrak{b}) + \varepsilon_i$$

Supondremos además que:

$$\varepsilon_i \sim N(0,\sigma^2);\ \mathrm{cov}(\varepsilon_i,\varepsilon_{i'}) = 0\ \forall i \neq i'$$

La expresión anterior especifica que los errores son variables aleatorias normales con media cero y varianza σ^2 común a todas las observaciones y que los errores son independientes $(\mathrm{cov}(\varepsilon_i,\varepsilon_{i'}) = 0;\ \text{para toda i e i´ diferentes})$.

Dependiendo de la forma de la función $f(.)$ se tiene un modelo de **regresión lineal** o un modelo de **regresión no lineal**. Dependiendo del número de regresoras se tiene un modelo de **regresión simple** (una regresora) o un modelo de **regresión múltiple** (más de una regresora). Un tratamiento más extenso de los modelos de regresión se puede encontrar en Draper y Smith (1988).

Regresión lineal simple

El modelo de regresión lineal simple se define por la forma particular de la función $f(.)$. Ésta se muestra en la siguiente expresión:

$$f(x_i,\mathfrak{b}) = \beta_0 + \beta_1 x_i$$

EL primer coeficiente (β_0) corresponde a la **ordenada al origen** y el segundo (β_1) a la **pendiente**. La Figura 7.1 ilustra un ejemplo sobre el cambio del peso de un animal "promedio" en función del tiempo desde el comienzo de un experimento (fijado arbitrariamente como tiempo cero). En esta recta la ordenada al origen vale 10 g y la pendiente 5 g. Estos datos indican que al comienzo del experimento los animales pesaban en promedio 10 g y que su peso promedio se incrementó en 5 g por día.

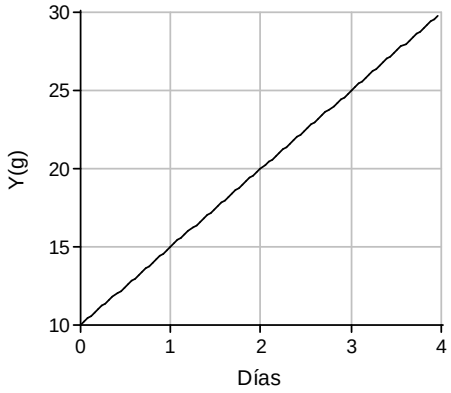


Figura 7.1: Recta que ilustra un modelo de regresión lineal simple donde la ordena al origen vale 10 g y la pendiente 5 g

Estimación

Llamaremos estimación del modelo de regresión a la asignación de valores a β_0 y β_1. A las estimaciones las simbolizaremos con $\tilde{\beta}_0$ y $\tilde{\beta}_1$ respectivamente. Para estimar el modelo hacen falta pares de datos (Y,X). Las estimaciones van a depender de estos datos y cambiarán si utilizamos un conjunto de datos diferentes, aún, cuando los nuevos datos se obtuvieran bajo las mismas condiciones experimentales. Esto implica que si repitiéramos un experimento y analizáramos sus resultados mediante análisis de regresión, las rectas ajustadas no serían exactamente las mismas. Esta situación parece paradójica ya que sugiere que el fenómeno que queremos modelar no puede ser modelado. El origen de estas variaciones está en lo que conocemos como error experimental. El error experimental se conceptualiza como una variable aleatoria que introduce perturbaciones sobre los valores que deberíamos observar de la variable dependiente. Además se asume que los errores son perturbaciones no sistemáticas y que por lo tanto su promedio es cero. Esto quiere decir que si tomáramos medidas repetidas de Y para un mismo valor de la regresora, en promedio, los errores se cancelarían. Luego la magnitud de la diferencia entre estimaciones obtenidas con conjuntos diferentes de datos depende de la magnitud del error experimental y del número de pares de datos (Y,x) utilizados. La magnitud del error experimental se ha

representado por σ^2 en las suposiciones del modelo de regresión y el número de pares por **n**.

Cuanto mayor es el error experimental mayor es la discrepancia entre estimaciones basadas en conjuntos diferentes de datos pero estas discrepancias puede controlarse aumentando el número de pares (Y,x) y hacerlas tan pequeñas como queramos. En la práctica no se toman distintos conjuntos de datos para ajustar un modelo, sin embargo podemos calcular la confiabilidad de las estimaciones mediante su error estándar y/o sus intervalos de confianza.

Aplicación

Lámina de agua en los perfiles del suelo de un cultivo

El archivo [Agua] contiene datos de disponibilidad de agua en un cultivo de soja en los distintos perfiles del suelo hasta una profundidad de 60 cm, obtenidos a los 100 días desde la emergencia. La disponibilidad de agua se expresa en milímetro de lámina de agua. Los valores de profundidad corresponden a 10, 20, 30, 40, 50 y 60 cm, pero el contenido de agua corresponde a los perfiles que van de [0-10) cm, [10-20) cm, etc. El propósito de este estudio es cuantificar cómo cambia la disponibilidad de agua con la profundidad del perfil analizado en un cultivo de soja. Los datos son parte de un estudio es más ambicioso que pretende comparar el efecto de distintos cultivares sobre el perfil de agua en el suelo. En esta aplicación sólo consideramos un cultivar. Para cada perfil hay tres repeticiones correspondientes a tres puntos de muestreo dentro de la parcela experimental.

Estrategia de análisis

El diagrama de dispersión del agua disponible vs la profundidad del perfil muestra un decaimiento sostenido de la disponibilidad y que este decaimiento parece seguir una relación lineal (Figura 7.2).

Análisis de regresión

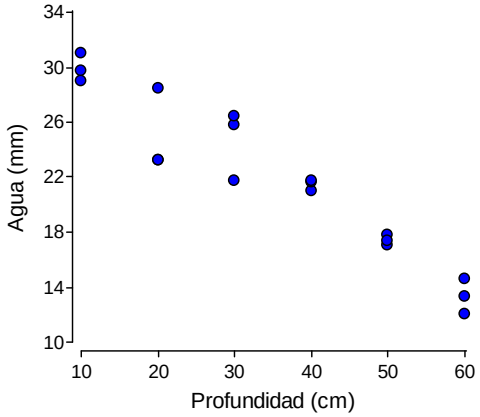


Figura 7.2: Disponibilidad de agua según la profundidad del perfil analizado en un cultivo de soja a los 100 días desde la emergencia.

Utilizando el software estadístico InfoStat ajustaremos un modelo de regresión lineal simple.

Para ajustar un modelo de regresión lineal simple, bajo los supuestos del modelo lineal clásico abrir el archivo [Agua]. En el menú *Estadísticas* seleccione el submenú *Regresión lineal*. Aparecerá la pantalla que se muestra a la izquierda de la Figura 7.3. Seleccione Profundidad (cm) en el panel izquierdo de la ventana y "muévalo" al panel *Regresoras*. De la misma forma seleccione Agua (mm) y "muévalo" al panel *Variable dependiente.* La imagen de la ventana resultante se muestra a la derecha de la Figura 7.3.

Las determinaciones del contenido de agua en los distintos perfiles del suelo dentro de cada punto de muestreo están correlacionadas. Esto viola el supuesto de independencia y, si bien se puede seguir tratando como un problema de regresión, la estructura de correlación debería incluirse en el análisis. La forma habitual de realizar esto es ajustando un modelo lineal mixto.

Para continuar, accione el botón *Aceptar*. Esta acción abrirá la siguiente pantalla (Figura 7.4 -izquierda). Por el momento, no modificaremos nada en esta pantalla. Sólo accionaremos el botón *Aceptar*. Esta acción generará dos salidas. Una gráfica con el diagrama de dispersión y la superposición de la recta ajustada y otra correspondiente al modelo estimado (Cuadro 7.1).

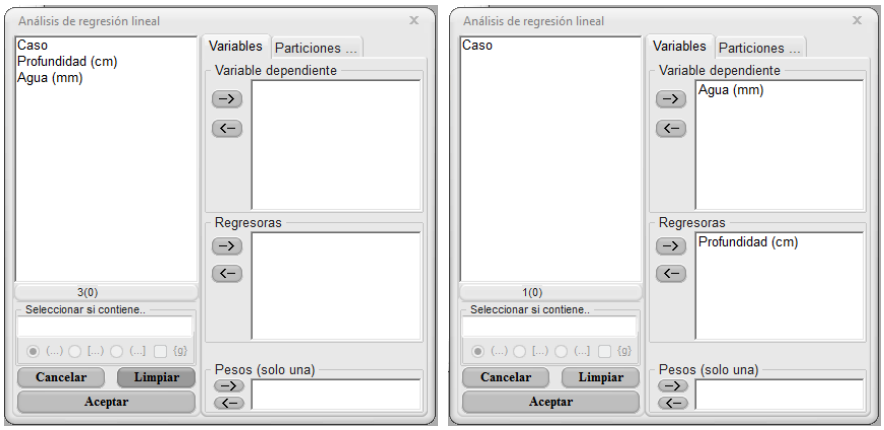


Figura 7.3: Diálogo inicial del análisis de regresión lineal en InfoStat.

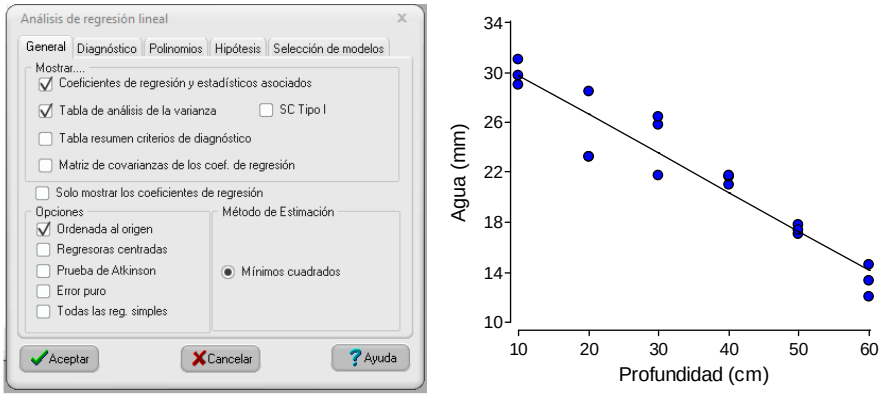


Figura 7.4: Diálogo de opciones del análisis de regresión lineal en InfoStat y salida gráfica del modelo de regresión lineal simple.

Análisis de regresión

Cuadro 7.1: Análisis de regresión lineal aplicada a los datos del archivo [Agua].

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Agua (mm)	18	0,90	0,90	4,18	77,04	79,71

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	32,83	0,99	30,72	34,93	33,08	<0,0001	
Profundidad (cm)	-0,31	0,03	-0,37	-0,26	-12,20	<0,0001	141,25

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	*valor p*
Modelo	507,84	1	507,84	148,95	<0,0001
Profundidad (cm)	507,84	1	507,84	148,95	<0,0001
Error	54,55	16	3,41		
Total	562,39	17			

El

Cuadro 7.1 tiene 3 tablas. El encabezamiento indica que la variable dependiente es *Agua (mm)*, que el número de pares de datos utilizados es 18, que el coeficiente de determinación (R^2) es 0,90, que su versión ajustada (R^2 Aj) también da un valor de 0,90, que el error cuadrático medio de predicción (ECMP) es 4,18 y que los criterios AIC y BIC producen valores de 77,04 y 79,71. Más adelante volveremos sobre este encabezamiento.

La segunda tabla contiene la estimación del modelo. Si pudiéramos tener un perfil a profundidad 10 cm por encima del suelo su contenido de agua estimado equivaldría a una lámina de 32,83 milímetros (const - $\hat{\beta}_0$). Muchas veces la interpretación física de la ordenada al origen puede no tener sentido, pero la presencia de la ordenada en el modelo es comúnmente necesaria a pesar de lo paradójica que resulte su interpretación. El parámetro de mayor interés en este ejemplo es la pendiente de la recta ajustada. La pendiente estimada $\left(\hat{\beta}_1\right)$ aparece en la línea correspondiente a la variable regresora (*Profundidad (cm*)). Su valor es -0,31. Es un punto importante del análisis de regresión establecer si la pendiente verdadera $\left(\beta_1\right)$ es distinta o no de cero. La hipótesis nula es $H_0 : \beta_1 = 0$. Si $\hat{\beta}_1$ fuera cero entonces diríamos que no importa cuál sea la profundidad del perfil analizado el contenido de agua permanece constante. En la columna de valores p, el *valor p* correspondiente a la pendiente es <0,0001. Esto se interpreta diciendo que la probabilidad de obtener una estimación de 0,31 unidades o más en cualquier sentido (+ o -) es, para los datos examinados, menor que 1 en 10000 si el verdadero valor de la pendiente fuera cero. Esto implica, bajo los criterios clásicos de la inferencia estadística, que la pendiente de -0,31 es estadísticamente distinta de cero y por lo tanto a mayor profundidad en el suelo el contenido de agua decae (coeficiente negativo) y ese decaimiento es de 0,31 mm de lámina de agua por cada centímetro de profundización. Luego a los 50 centímetros tendremos un decaimiento de 15,5 mm en la lámina de agua respecto del valor inicial (el correspondiente a la profundidad 0) que se estimó en 32,82 mm.

Luego el contenido promedio de agua en un perfil que se toma entre los 40 y los 50 centímetros de profundidad será 32,82-0,31*50 = 17,32.

El error estándar (EE) es una medida de confiabilidad de las estimaciones. Para la constante $\left(\hat{\beta}_0\right)$ el error estándar es 0,99 y para la pendiente $\left(\hat{\beta}_1\right)$ 0,03. Estos errores representan un error relativo del 3% y 10% aproximadamente para cada uno de sus respectivos parámetros. No existen reglas escritas sobre la valoración de estos errores relativos pero en general un error relativo de hasta un 10% es aceptable y hasta un 20% admisible, aunque esto necesariamente depende de las aplicaciones. El error estándar de una estimación está directamente vinculado con la construcción de los intervalos de confianza. Cuanto mayor sea el error estándar mayor será el intervalos de confianza y por lo tanto mayor la incertidumbre de la estimación. Por ejemplo para la pendiente del modelo estimado, el intervalo [-0,37;-0,26] contiene a la verdadera pendiente con una confianza del 95%. De igual manera el intervalo [30,72; 34,93] hace lo propio con la

ordenada al origen. Una forma de ver simultáneamente el efecto que introduce la incertidumbre de las estimaciones $\check{\beta}_0$ y $\check{\beta}_1$ sobre el modelo estimado es obteniendo la **banda de confianza** para los promedios de contenido de agua en función de la profundidad del perfil. Para ello invocaremos nuevamente el análisis de regresión lineal y en la ventana de diálogo de opciones, solapa *Diagnóstico* marcaremos *Graficar* > *Bandas de Confianza* como se ilustra a continuación Figura 7.5. El gráfico resultante se muestra en la Figura 7.6.

Figura 7.5: Diálogo de opciones del análisis de regresión lineal en InfoStat y salida gráfica del modelo de regresión lineal simple.

No debe sorprendernos que haya puntos del diagrama de dispersión que caen fuera de la banda de confianza ya que se trata de una banda de confianza para la media no para los datos. Si quisiéramos construir una **banda de predicción** para los valores observables de Y entonces deberíamos tildar la opción correspondiente (tarea para el lector). En tal caso la banda de predicción estará por fuera de la de confianza.

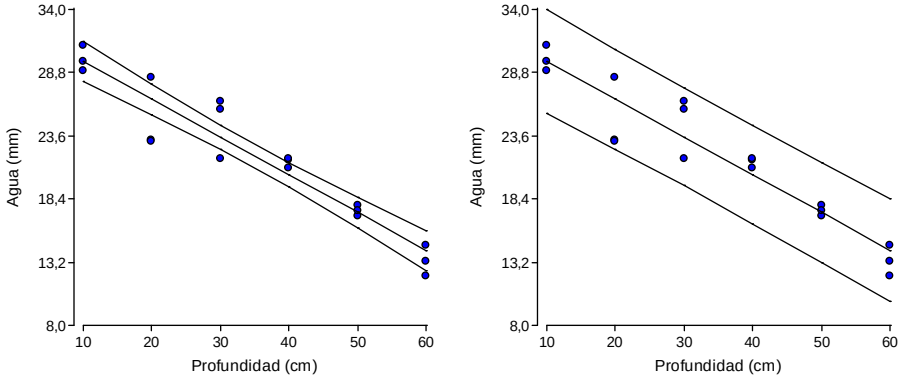


Figura 7.6: Diálogo de opciones del análisis de regresión lineal en InfoStat y salida gráfica del modelo de regresión lineal simple.

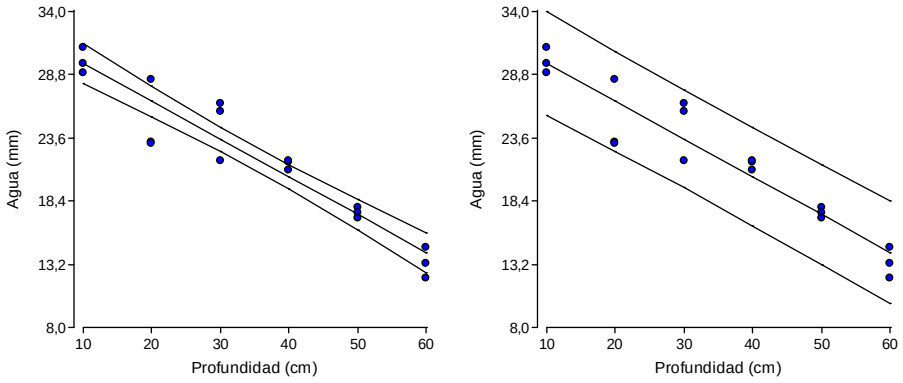


Figura 7.7. Gráfico mostrando la recta ajustada y las bandas de confianza (izquierda) y bandas de predicción (derecha) para el contenido de agua en los distintos perfiles del suelo.

La tercera parte de la salida del análisis de regresión corresponde a una tabla de análisis de la varianza para el modelo de regresión. De ella se desprenden dos cantidades que hemos nombrado anteriormente. El **coeficiente de determinación** y el coeficiente de determinación ajustado. El primero es el cociente entre la suma de cuadrados (sc) correspondiente a la pendiente (fila rotulada con el nombre de la variable independiente) dividida por la suma de cuadrados total. En el ejemplo R^2= 507,84/562,39. El coeficiente R^2 se interpreta como la fracción de variación observada en la variable de respuesta explicada por las variaciones observadas en la variable regresora. Luego con un R^2=0,90, diremos que la profundidad del suelo explica el 90% de la variabilidad observada en el contenido de agua del experimento analizado. El **coeficiente de determinación ajustado** se calcula como

$$R^2_{aj} = 1-(1-R^2)\left(\frac{n-1}{n-p}\right);\ p = \#\ parametros$$

En este caso p=2 (la ordenada al origen y la pendiente). La interpretación es similar a la realizada para R^2 pero este coeficientes es más conservador y es siempre menor o igual a R^2. Un R^2 ajustado mucho menor que R^2, es una indicación de que el modelo incluye covariables que no son necesarias y en tal caso es recomendable una reducción del modelo eliminando regresoras innecesarias.
Más allá del cálculo de los coeficientes de determinación, la tabla de análisis de la varianza es útil en algunos casos especiales como el que ilustraremos más adelante.

Conclusión

La función ajustada para el **valor predicho** de rendimientos será entonces la que se presenta a continuación, donde $\hat{Y}$ representa el espesor de la lámina de agua según la profundidad (P) del perfil examinado:

$$\hat{Y} = 32{,}83\text{-}0{,}31P$$

Esta ecuación sugiere que la lámina de agua decae a 0,31 mm por cada centímetro de profundidad.

Residuos vs. Predichos

Una herramienta diagnóstico esencial para revisar la adecuación del modelo ajustado es revisar el gráfico de residuos vs los valores predichos. Los **residuos** de un modelo se obtienen restando a cada valor observado de la variable dependiente su valor predicho. Los **residuos estudentizados** son un tipo especial de residuos obtenidos al dividir los residuos por sus errores estándares. La ventaja de utilizar **residuos estudentizados** es que el analista puede rápidamente saber cuando un residuo es grande (ya sea positivo o negativo).

Si el modelo está bien ajustado y los supuesto del modelo (normalidad, homoscedasticidad e independencia se cumplen), el 95% de los residuos estudentizados estarán entre -2 y 2.

Luego un residuo menor a -4 implica que el dato correspondiente es extremadamente pequeño para el modelo ajustado, recíprocamente un residuo mayor +4 implicará que el valor observado es muy grande en relación a lo que predice el modelo. Por lo tanto la presencia de residuos estudentizados muy grandes o muy pequeños implica que hay datos que están siendo mal modelados. Esto puedo querer decir dos cosas: los datos son errados (mal transcriptos, mal medidos, la unidad experimental sobre la que se tomó el dato es aberrante – animal o planta enferma por ejemplo) y por lo tanto es mejor eliminarlos de la base de datos, o el modelo que estamos tratando de ajustar a

los datos es inapropiado. No se puede dar un consejo general en este caso, el investigador tendrá que evaluar la situación y decidir el curso de acción.

Una palabra de advertencia. Cuanto mayor es el número de datos, más probable es encontrar residuos estudentizados grandes en valor absoluto, esto no debe sorprender porque estos residuos son poco probables y por esa misma razón aparecen cuando se tienen muchos datos. Un valor cuya probabilidad es 1/1000 difícilmente aparezca en una base de datos de 20 observaciones, pero seguramente aparecerá en una base de 5000 datos.

El gráfico de residuos estudentizados vs valores predichos es una salida estándar de InfoStat. Para los modelos de regresión lineal simple o polinómicos antecede al gráfico que muestra el ajuste. En el caso de regresión múltiple, este es el único grafico que InfoStat da por defecto. La Figura 7.9 muestra un gráfico de residuos vs predicho para el ejemplo de la lámina de agua.

¿Qué esperamos ver en un gráfico de residuos estudentizados vs predichos? Lo ideal es observar una nube de puntos alrededor del cero, confinada en el 95% de los casos a la banda -2, 2 y sin que aparezca ninguna "estructura llamativa".

Si observáramos que los datos con valores predichos bajos tienen residuos estudentizados negativos y viceversa, los que tienen valores predichos altos tuvieran residuos positivos, entonces estaríamos ante una anomalía. Igualmente si pudiéramos identificar con colores las observaciones que realizaron distintos colaboradores un experimento y las observaciones de los distintos colaboradores aparecieran sistemáticamente con residuos estudentizados positivos o negativos, esto debería llamarnos la atención. Igualmente si la variabilidad (rango de variación vertical de los puntos) es mayor para predichos altos que para predichos bajos, entonces estaremos frente a un problema de falta de homogeneidad de varianzas. La interpretación de gráficos de residuos es una destreza que se adquiere mirando estos gráficos.

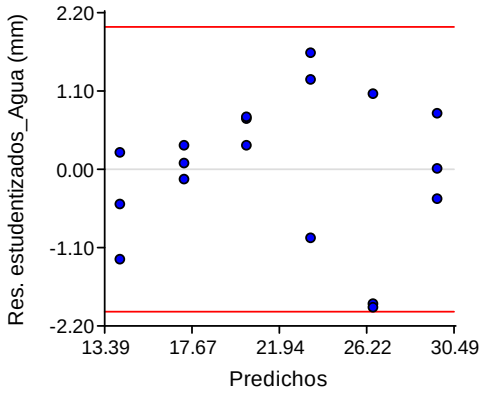


Figura 7.8. Residuos estudentizados vs predichos.

El gráfico mostrado en la Figura 7.9 se insinúa una curvatura que pudiera sugerir la necesidad de ajustar un modelo polinómico de segundo grado. No obstante esta insinuación, la evidencia no es fuerte en este sentido. Afortunadamente para este caso, disponemos de varias observaciones de Y para los distintos valores de X y podemos hacer un contraste formal de hipótesis para la falta de ajuste.

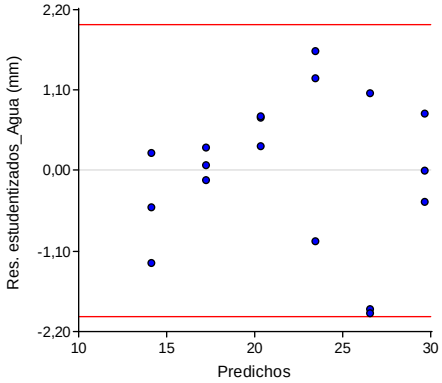


Figura 7.9. Residuos estudentizados vs predichos.

Falta de ajuste

Cuando se dispone de repeticiones de lecturas de Y para todos o al menos algún valor de la regresora es posible hacer una prueba estadística que se conoce como prueba de falta de ajuste. En el ejemplo que estamos examinando hay tres repeticiones para cada valor de x, así que el procedimiento puede ser aplicado. La hipótesis nula de esta prueba es que el modelo ajusta y la alternativa es que hay **falta de ajuste** (lack of fit). Si el **valor p** de la prueba es menor que el nivel de significación la hipótesis nula se rechaza y en consecuencia el modelo lineal no es enteramente apropiado para modelar los datos observados. Para aplicar esta prueba a los datos del ejemplo del agua invoquemos el análisis de regresión lineal y en la ventana de opciones (solapa General)

seleccionemos **Error puro** como se muestra en la Figura 7.10. El resultado de aplicar esta opción se visualiza en la parte correspondiente a análisis de la varianza de la salida (Cuadro 7.2). La prueba aparece con el título Lack of Fit. Tiene asociada un valor p de 0,2780 por lo que no hay evidencia en contra de que el ajuste lineal sea el apropiado para este conjunto de datos.

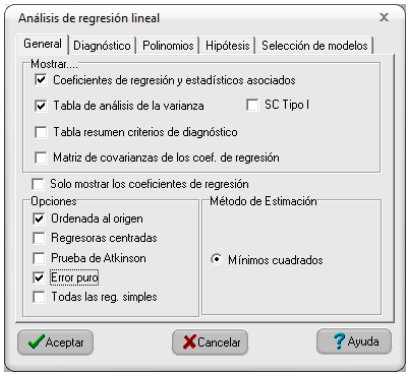


Figura 7.10. Ventana de opciones mostrando la selección Error puro. Con esta opción tildada se obtiene la prueba de falta de ajuste para el modelo lineal planteado (lack of fit test).

Cuadro 7.2: Análisis de regresión lineal de los datos del archivo [Agua] con prueba de bondad de ajuste

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Agua (mm)	18	0,90	0,90	4,18	77,04	79,71

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	32,83	0,99	30,72	34,93	33,08	<0,0001	
Profundidad (cm)	-0,31	0,03	-0,37	-0,26	-12,20	<0,0001	141,25

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	507,84	1	507,84	148,95	<0,0001
Profundidad (cm)	507,84	1	507,84	148,95	<0,0001
Error	54,55	16	3,41		
Lack of Fit	17,76	4	4,44	1,45	0,2780
Error Puro	36,79	12	3,07		
Total	562,39	17			

Regresión lineal múltiple

El modelo de regresión múltiple es una generalización del modelo lineal simple. Aparece en distintos contextos, todos caracterizados por la presencia de más de una regresora. El modelo de **regresión lineal múltiple** puede sintetizarse de la siguiente manera.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{(p-1)} x_{i(p-1)} + \varepsilon_i$$

Supondremos también que:

$$\varepsilon_i \sim N\left(0, \sigma^2\right);\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$$

Comenzaremos desarrollando un caso particular de regresión múltiple: la regresión polinómica y luego nos concentraremos en el caso general.

Regresión polinómica

La regresión polinómica puede basarse en una o más variables regresoras. Abordaremos su presentación con el caso de una regresora. El modelo de regresión polinómica requiere la especificación del grado del polinomio que se quiere ajustar. Así, si el polinomio es de grado 2, y la variable regresora la representamos por x, el modelo lineal que ajustaremos mediante regresión múltiple será:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Supondremos también que:

$$\varepsilon_i \sim N\left(0, \sigma^2\right);\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$$

Como puede observarse el modelo de regresión polinómica a una regresora es un modelo de regresión múltiple donde las regresoras son potencias de la regresora original. Aplicaciones típicas se encuentra en modelos de crecimiento, pero muchos modelos de regresión pueden incluir términos polinómicos para lograr ajustes más apropiados de los datos observados.

El problema principal con la regresión polinómica es la especificación del grado del polinomio ya que suele no haber una justificación teórica que permita sugerirlo independientemente de los datos y en consecuencia la selección del grado se realiza empíricamente. Como el ajuste del modelo polinómico mejora con el grado, el desafío es encontrar un ajuste razonable con el menor grado.

Aunque no pude tomarse como regla, lo usual es no superar el grado 3 ya que de otra forma el modelo resultante no estará capturando lo esencial de la relación entre variable dependiente y regresora sino también el error experimental. Luego un modelo

sobre ajustado a los datos de un experimento particular carece de la generalidad y aplicabilidad que el investigador trata de encontrar.

Aplicación

Respuesta del cultivo a la fertilización nitrogenada

En este ejemplo se estudia el rendimiento de trigo en el oeste de la provincia de Buenos Aires, según el nivel de fertilización nitrogenada. El propósito es encontrar una dosis óptima [datos: fertilización en trigo]. Los datos contienen dos columnas: la dosis de nitrógeno en kg de nitrógeno por ha y el rendimiento en kg/ha.

Estrategia de análisis

Lo primero es mirar la relación empírica que hay entre el rendimiento y el aporte de nitrógeno al suelo. Para ello realizaremos un diagrama de dispersión entre rendimiento (eje Y) y aporte de nitrógeno (eje X) como se muestra en la Figura 7.11. En ella podemos ver que a mayor aporte de nitrógeno mayor es el rendimiento. Sin embargo, parece que el crecimiento del rendimiento empezara a decaer con las dosis mayores. El ajuste de una regresión lineal simple y sus residuos estudentizados se muestran en la Figura 7.12. El gráfico de residuos estudentizados pone claramente de relieve que el ajuste de una recta es insuficiente para estos datos. Cuando los residuos estudentizados muestra una curvatura, como la que se observa en el ejemplo, es un buen indicio de la necesidad de incorporar al modelo un término cuadrático de la regresora: en este caso el nitrógeno.

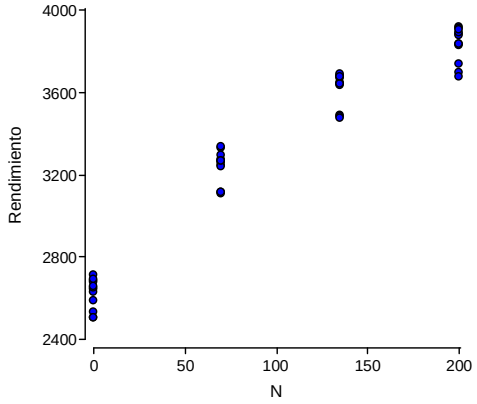


Figura 7.11. Diagrama de dispersión entre rendimiento de trigo (kg/ha) y aporte de nitrógeno al suelo (kg/ha).

Análisis de regresión

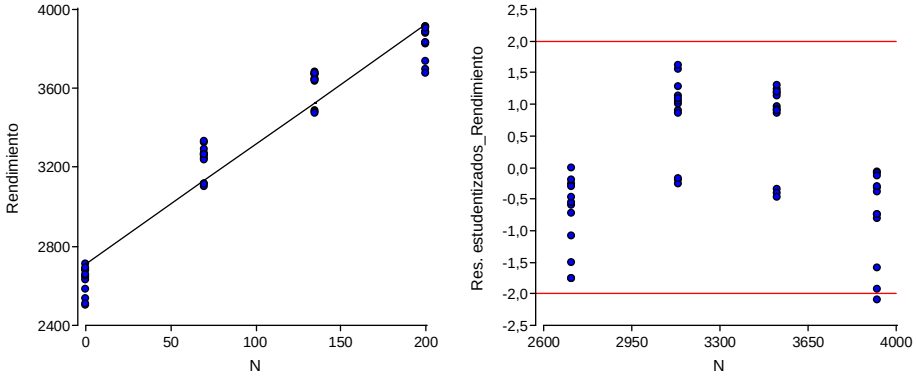


Figura 7.12. Recta ajusta a la relación entre rendimiento de trigo (kg/ha) y aporte de nitrógeno al suelo (kg/ha) (izquierda) y residuos estudentizados vs predicho (derecha).

Para ajustar un modelo polinómico de segundo grado invocaremos el procedimiento de *Análisis de regresión* con *Rendimiento* como variable dependiente y *N* (nitrógeno) como independiente. En la ventana de diálogo del análisis de regresión, seleccionar la solapa *Polinomios* y especificar que nitrógeno (N) entra al modelo como un polinomio de segundo grado (Figura 7.13). La representación gráfica del ajuste obtenido se muestra en la Figura 7.14. Puede apreciarse que los residuos estudentizados han cambiado sustancialmente y ahora no se observa la curvatura mostrada en la Figura 7.12. La salida en la ventana de resultados se presenta en el

Cuadro 7.3.

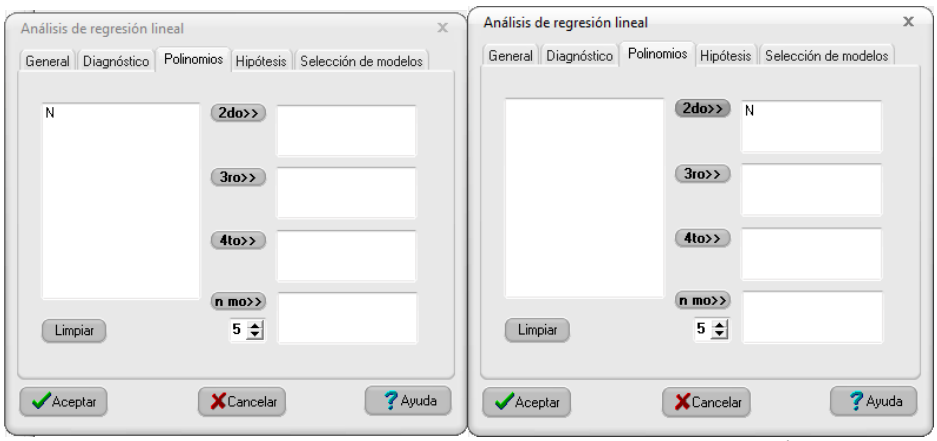


Figura 7.13. Recta ajusta a la relación entre rendimiento de trigo (kg/ha) y aporte de nitrógeno al suelo (kg/ha) (izquierda) u residuos estudentizados vs predicho (derecha).

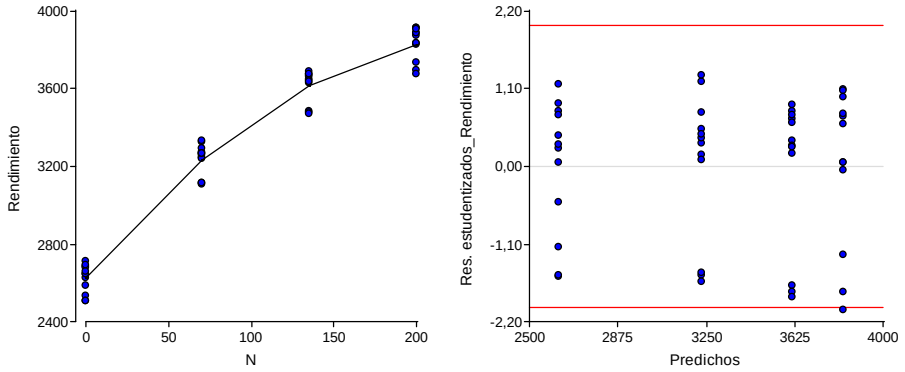


Figura 7.14. Polinomio de orden 2 ajustado a la relación entre rendimiento de trigo (kg/ha) y aporte de nitrógeno al suelo (kg/ha) (izquierda) y residuos estudentizados vs predicho (derecha).

Los resultados presentados en el

Cuadro 7.3 se agrupan en tablas. La primera indica que el número total de datos analizados fue 48 y que la determinación del modelo fue 0,97 (muy alta). Los estadísticos ECMP, AIC y BIC son discutidos en cursos de estadística más avanzados. La segunda tabla, la más importante, contiene las estimaciones de los parámetros del modelo, sus errores estándares, los intervalos de confianza y las pruebas T para la hipótesis nula de que dice que el valor poblacional del parámetro es cero. El *valor p* para esta hipótesis se calculó de acuerdo a un contraste bilateral. El estadístico **Cp-Mallows** es un indicador de la importancia relativa de las variables incluidas en el modelo. Su valor es mayor mientras más importante es la variable para explicar las variaciones de Y.

De acuerdo a esta tabla la ordenada al origen estimada es de 2622,947 kg. Éste valor es perfectamente interpretable en este experimento y corresponde al nivel medio de rendimiento sin agregado de nitrógeno. La pendiente de la parte lineal $\left(\hat{\beta}_1\right)$ se estimó en 10,143kg y la pendiente de la componente cuadrática $\left(\hat{\beta}_2\right)$ se estimó en -0,021kg. Estos coeficientes no pueden interpretarse independientemente ya que están asociados a la misma regresora y actúan de manera simultánea sobre la variable de respuesta.

La función ajustada para el valor esperado de rendimientos será entonces la que se presenta a continuación, donde $\hat{Y}$ representa el rendimiento promedio esperado de acuerdo al aporte de nitrógeno (N).

$$\hat{Y} = 2622,947 + 10,143 * N - 0,021 * N^2$$

Cuadro 7.3: Análisis de regresión lineal aplicada a los datos del archivo [Agua].

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Rendimiento	48	0,97	0,97	7189,41	561,20	568,68

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
Const	2622,947	22,456	2577,719	2668,175	116,806	<0,0001	
N	10,143	0,535	9,066	11,220	18,966	<0,0001	353,894
N^2	-0,021	0,003	-0,026	-0,015	-7,995	<0,0001	64,558

Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	p-valor
Modelo	10115326,97	2	5057663,48	800,18	<0,0001
N	9711271,34	1	9711271,34	1536,44	<0,0001
N^2	404055,63	1	404055,63	63,93	<0,0001
Error	284429,03	45	6320,65		
Total	10399756,00	47			

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	10115326,97	2	5057663,48	800,18	<0,0001
N	10115326,97	2	5057663,48	800,18	<0,0001
Error	284429,03	45	6320,65		
Total	10399756,00	47			

Conclusión

Si el modelo ajustado fuera una recta con pendiente positiva, la mejor dosis sería la máxima. Pero en un modelo cuadrático la dosis que maximiza (o minimiza) la respuesta se calcula derivando la función e igualando la derivada a cero. Si $\hat{\beta}_2$ es negativo entonces en ese punto se alcanza un máximo (sino un mínimo). Luego la dosis que maximiza los rendimientos en nuestro ejemplo será.

$$\frac{-\hat{\beta}_1}{2\hat{\beta}_2} = \frac{-10,143}{2(-0,021)} = 241,5$$

El rendimiento predicho de máximo rendimiento en función del nitrógeno será:

$$\hat{Y} = 2622,947 + 10,143 * 241,5 - 0,021 * 241,5^2 = 3847,7$$

Es interesante observar el la dosis máxima ensayada estuvo por debajo del punto donde se alcanza el máximo. Un nuevo ensayo debería incluir valores superiores de aporte de nitrógeno para verificar esta predicción.

Regresión con múltiples regresoras

El modelo de regresión lineal con múltiples regresoras o simplemente modelo de regresión múltiple es una extensión natural de la regresión lineal simple. La variable de respuesta cambia según una tasa constante (llamada pendiente parcial o coeficiente de regresión parcial) a los cambios de cada una de las regresoras. El procedimiento para ajustar un modelo de regresión múltiple es usualmente por mínimos cuadrados y esto conduce a la solución de un sistema de ecuaciones lineales. Desde el punto de vista operativo el ajuste de estos modelos, utilizando software estadístico, es similar al utilizado para regresión simple, excepto que se agregan más regresoras al modelo y que la interpretación de los coeficientes, ahora llamados **coeficientes de regresión parcial**, es diferente.

La ventaja de utilizar modelos de regresión múltiple es consisten en la posibilidad de estudiar el efectos de varias regresoras simultáneamente.

El modelo de regresión múltiple permite asimismo incluir factores de clasificación mediante la utilización de variables auxiliares (dummy variables) extendiéndolos para ajustar una amplia variedad de datos experimentales u observacionales. La forma general de estos modelos es:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{(p-1)} x_{i(p-1)} + \varepsilon_i$$

Supondremos además que:

$$\varepsilon_i \sim N\left(0, \sigma^2\right);\ \mathrm{cov}(\varepsilon_i, \varepsilon_{i'}) = 0\ \forall i \neq i'$$

Aplicación

Condiciones óptimas de cultivo de bacteria

Se quiere poner a punto el cultivo de una cepa de la bacteria Rhizobium que es usada en inoculaciones de semillas para favorecer la fijación de nitrógeno. Para ello se prueban 2 tiempos de cultivo (24 y 48 hs), 3 temperaturas (27, 35 ,43) y 5 concentraciones de nutrientes expresadas como proporciones (0,6, 0,8, 1,0, 1,2, 1,4) respecto de una solución testigo. Para cada combinación de los factores: tiempo, temperatura y concentración de nutrientes se obtuvo el número de bacterias por cm^3 que representa

la variable dependiente (Y). El archivo que contiene los resultados de este ensayo es *[Rhizobium]*.

Estrategia de análisis

A diferencia de lo que ocurre en el modelo de regresión lineal simple, la visualización de la variable dependiente en función de cada una de las regresoras suele no ser informativa. La forma equivalente de hacer esto es graficando lo que se llaman residuos parciales. Esta técnica la discutiremos más adelante. El ajuste de modelo lineal múltiple se muestra en el Cuadro 7.4.

Como se puede observar en la tabla de *Coeficientes de regresión y estadísticos asociados* (Cuadro 7.4) todos los coeficientes tienen un *valor p* pequeño, menor que el nivel usual de significación de 0,05, y por lo tanto diremos que los coeficientes que están siendo estimados son estadísticamente distintos de cero (esta es la hipótesis nula que este procedimiento pone a prueba). Que los coeficientes de regresión parcial sean estadísticamente distintos de cero implica que cuando se producen cambios en las regresoras, estos cambios se traducen en modificaciones en el número medio de bacterias por cm^3. ¿Cómo deben interpretarse esos coeficientes? Vamos a dejar para después una discusión sobre la ordenada al origen. Como el tiempo está medido en horas, por cada hora adicional de cultivo, y manteniendo las otras regresoras fijas en algún valor, dentro del rango en que se ajustó el modelo, se ganan en promedio 2,79 bacterias por cm^3. Es decir, si mantenemos un cultivo a temperatura de 30 grados y a una concentración de nutrientes 0,9, entonces el incremento promedio en el número de bacterias por cm^3 que se observará entre las 24 y 25 horas de cultivo o entre 28 y 29 horas, será 2,79. Los valores 30 y 0,9 fueron escogidos arbitrariamente y la interpretación sigue siendo válida con cualquier combinación de ellos siempre y cuando sus valores se encuentren dentro del rango de variación de los mismos en el experimento. Por ejemplo no sería válido suponer que el cambio en el número promedio de bacterias por cada hora de cultivo es 2,79 cuando fijamos la concentración en 3.

Los otros coeficientes también son positivos así que en cada caso valdrá una interpretación equivalente, caso contrario, si los coeficientes de regresión parcial fueran negativos, lo único que cambia es que a cambios positivos en las regresoras se observarán decrecimientos en la variable dependiente. La tabla de Análisis de la Varianza en la salida, no nos ofrece información adicional, excepto que el coeficiente de determinación R^2 se obtiene dividiendo la suma de cuadrados atribuible al modelo (78113,27) por la suma de cuadrados total (141432,24).

Cuadro 7.4. Modelo de regresión múltiple para el número de bacterias por cm3 en función del tiempo de cultivo, la temperatura de cultivo y la concentración de nutrientes expresados en términos relativos a una solución estándar.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	33	0,55	0,51	2784,48	353,11	360,59

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	-159,58	55,41	-272,90	-46,25	-2,88	0,0074	
Tiempo	2,79	0,69	1,39	4,19	4,07	0,0003	19,04
Temp	2,55	1,23	0,03	5,08	2,07	0,0476	7,17
Nut	93,82	29,15	34,19	153,45	3,22	0,0032	13,04

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	78113,27	3	26037,76	11,93	<0,0001
Tiempo	36143,43	1	36143,43	16,55	0,0003
Temp	9340,47	1	9340,47	4,28	0,0476
Nut	22612,17	1	22612,17	10,36	0,0032
Error	63318,97	29	2183,41		
Total	141432,24	32			

La ordenada al origen tiene un coeficiente negativo, esto implica que el modelo no ajusta bien cerca del origen. Cuando la temperatura de cultivo, el tiempo de cultivo y la concentración de nutrientes es cero, el valor natural para el número de bacteria por cm^3 debería corresponderse con la concentración por cm^3 del inoculo original. Aún cuando sabemos que el modelo no ajusta bien cerca del origen, en general, no nos preocupamos tanto por eso en la medida que el ajuste del modelo, en la región de las regresoras donde nos interesa investigar, sea bueno. ¿Cómo decidimos si el ajuste es bueno? Una medida habitual para tomar esta decisión es mirar el R^2. En este caso vale 0,55. ¿Qué dice este valor? El mínimo es 0 y el máximo 1 y cuando más cercano a uno "mejor". Si R^2 fuera 1 entonces los valores de la variable dependiente observados coincidirían, todos, con los valores predichos por el modelo. Por lo tanto parece que el R^2 de 0,55 nos deja a mitad de camino.

Sin embargo, tenemos que decir que a pesar de la tradición de utilizar R^2 como un criterio de bondad de ajuste, el R^2 no es una medida de la calidad del modelo ajustado sino sólo una medida aproximada de cuan predictivo es el modelo para valores individuales observables en el futuro de la variable dependiente. Esta medida de la habilidad predictiva del modelo es sólo válida si el modelo ha sido bien ajustado. Entonces, ¿cómo verificamos que el modelo fue bien ajustado? La calidad del ajuste se

juzga por distintos criterios diagnósticos, casi todos ellos basados en la observación de los residuos. Los residuos son las diferencias entre los valores observados y los valores predichos, pero hay muchas formas de residuos dependiendo de cómo calculemos el valor predicho y si el residuo es transformado por algún factor de estala (dividiendo por su error estándar, por ejemplo). La discusión sobre métodos y medidas de diagnóstico puede ser muy extensa, para aquellos que quieran tener una introducción más detallada de este tópico consultar el libro de Daper & Smith (1988). En este material sólo abordaremos algunos métodos de diagnóstico que, a juicio del autor, son los más efectivos para identificar anomalías en el ajuste de un modelo de regresión lineal. A continuación revisaremos las herramientas de diagnóstico y su aplicación al ejemplo que estamos tratando.

Residuos parciales

El análisis de los residuos parciales es una técnica destinada a observar cómo se comporta la variable dependiente en relación a una regresora cuando las otras están fijadas. Estos gráficos permiten visualizar la *forma* de la relación entre la variable dependiente y una regresora particular, una vez que el efecto de las otras regresoras ha sido removido. La Figura 7.15 muestra la manera de pedir los residuos parciales en InfoStat.

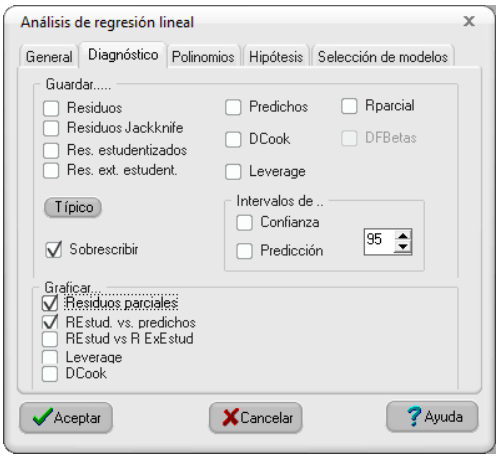


Figura 7.15. Ventana de diálogo indicando la forma de pedir la obtención de gráficos de residuos parciales

La Figura 7.16 muestra los residuos parciales obtenidos para tiempo, temperatura (Temp) y Nutrientes (Nut). Hay dos casos que merecen atención: los residuos parciales en función de la temperatura y los residuos parciales en función de la concentración de nutrientes. No es posible encontrar patrones llamativos en el caso de los residuos parciales con el tiempo ya que este factor sólo se evaluó para dos valores diferentes. Sin embargo, queda claro que a mayor tiempo mayor el número de células por cm^3.

Análisis de regresión

El gráfico de residuos parciales en función de la temperatura muestra que después de la temperatura 35 hay un decaimiento de la producción de bacterias, esto sugiere que la forma en que el número de bacterias se relaciona con la temperatura sigue una curva con un máximo próximo a 35 grados. La forma más sencilla de incorporar esta información al modelo de regresión es agregando una nueva regresora que es el cuadrado de la temperatura, así estaremos ajustando un modelo de regresión lineal múltiple que incluye un polinomio de segundo grado para la temperatura. Para el caso de los nutrientes pasan dos cosas distintas, una es que también, parece haber un máximo cerca de 1,22 y además que la variabilidad en el número de bacterias, entre repeticiones, aumenta con el incremento en la disponibilidad de nutrientes. El primer punto puede aproximarse también incluyendo un término cuadrático para los nutrientes, con lo cual el modelo de regresión múltiple incluiría también un polinomio de grado dos para la concentración de nutrientes.

El problema de la mayor variabilidad, asociada a mayores concentraciones de nutrientes, es un problema que puede abordarse incluyendo en el modelo una función de varianza. En este material no trataremos este caso, pero el lector interesado puede revisar el Tutorial de Modelos Mixtos con InfoStat (Di Rienzo, et. al 2009) que se distribuye conjuntamente con InfoStat y puede accederse desde el menú *Estadística>>Modelos lineales generales y mixtos>> Tutorial*. La no inclusión de la función de varianza tiene como consecuencia que los estimadores de los parámetros tengan un mayor error estándar pero los estimadores son aún, consistentes e insesgados.

En el archivo correspondiente a este ejemplo están calculados los cuadrados de Tiempo y Nut, pero están ocultos. Con la tabla de Rhizobium abierta y aplicando la combinación de teclas [Ctrl] [E] se abrirá un ventana de diálogo. En ella encontrará la lista de columnas en la tabla de datos. Las que no se encuentran tildadas son la que están ocultas. Tíldelas para que se hagan visibles y cierre la ventana de dialogo apretando el botón *Aceptar.* Luego invoque nuevamente al análisis de regresión lineal y en la ventana de diálogo de especificación de variables incluya a los términos cuadráticos de temperatura y concentración de nutrientes. El resultado del ajuste de este modelo se presenta en el Cuadro 7.5.

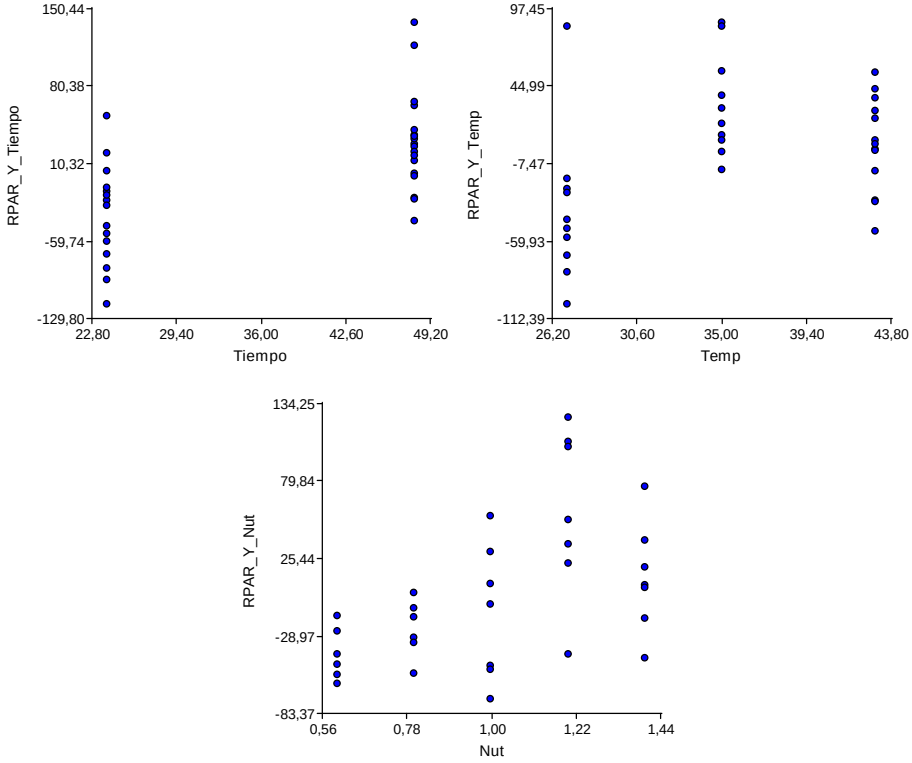


Figura 7.16. Ventana de diálogo indicando la forma de pedir la obtención de gráficos de residuos parciales.

Mirando la sección de *Coeficientes de regresión y estadísticos asociados* se puede observar que la inclusión de Temp2 (temperatura al cuadrado) está respaldada por un valor p significativo. Paradójicamente la inclusión de Nut2 no sólo no parece estar justificada sino que en este nuevo modelo ni siquiera aparece Nut con un efecto significativo. Este comportamiento singular del modelo obedece a que Nut y Nut2 están correlacionadas y están aportando información muy parecida respecto a la variable dependiente y por lo tanto están enmascarando mutuamente sus efectos. La solución es sacar una de ellas y por su puesto eliminaremos Nut2.

Análisis de regresión

Cuadro 7.5. Modelo de regresión múltiple para el número de bacterias por cm3 en función del tiempo de cultivo, la temperatura de cultivo y la concentración de nutrientes.

Análisis de regresión lineal

Variable	N	R²	R² Aj	ECMP	AIC	BIC
Y	33	0,69	0,63	2365,86	344,99	355,46

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	-1211,73	307,82	-1843,32	-580,15	-3,94	0,0005	
Tiempo	2,87	0,59	1,66	4,09	4,86	<0,0001	27,78
Temp	57,14	16,88	22,51	91,76	3,39	0,0022	16,09
Temp2	-0,78	0,24	-1,27	-0,29	-3,24	0,0031	15,17
Nut	359,84	211,81	-74,77	794,44	1,70	0,1008	7,82
Nut2	-130,91	104,62	-345,57	83,75	-1,25	0,2216	6,55

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	97582,58	5	19516,52	12,02	<0,0001
Tiempo	38306,10	1	38306,10	23,59	<0,0001
Temp	18613,25	1	18613,25	11,46	0,0022
Temp2	17072,21	1	17072,21	10,51	0,0031
Nut	4687,14	1	4687,14	2,89	0,1008
Nut2	2542,78	1	2542,78	1,57	0,2216
Error	43849,66	27	1624,06		
Total	141432,24	32			

En la nueva salida (Cuadro 7.6) se observa nuevamente que Nut tiene un efecto altamente significativo. Vemos además que el R^2 es ahora de 0,67, lo que implica que hemos mejorado la capacidad predictiva del modelo, siempre y cuando el modelo sea correcto.

Cuadro 7.6. Modelo de regresión múltiple para el número de bacterias por cm3 en función del tiempo de cultivo, la temperatura de cultivo y la concentración de nutrientes con términos cuadráticos solo para la temperatura.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	33	0,67	0,63	2286,76	344,85	353,83

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	-1087,20	294,21	-1689,86	-484,54	-3,70	0,0009	
Tiempo	2,89	0,60	1,66	4,11	4,83	<0,0001	26,55
Temp	56,93	17,05	22,01	91,84	3,34	0,0024	14,80
Temp2	-0,77	0,24	-1,27	-0,28	-3,20	0,0034	13,90
Nut	96,68	25,41	44,62	148,73	3,80	0,0007	18,01

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	95039,80	4	23759,95	14,34	<0,0001
Tiempo	38636,47	1	38636,47	23,32	<0,0001
Temp	18479,12	1	18479,12	11,15	0,0024
Temp2	16926,53	1	16926,53	10,22	0,0034
Nut	23979,78	1	23979,78	14,47	0,0007
Error	46392,44	28	1656,87		
Total	141432,24	32			

El gráfico de residuos estudentizados vs valores predichos es una salida estándar de InfoStat, en el caso de modelos de regresión lineal simple o polinómicos, antecede al gráfico que muestra el ajuste. En el caso de regresión múltiple, este es el único grafico que InfoStat da por defecto. El gráfico resultante del ajuste anterior se muestra en la Figura 7.9 .

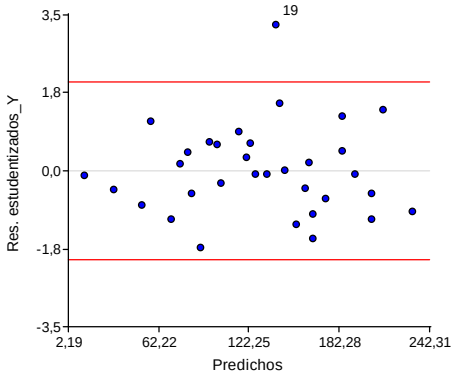


Figura 7.17. Residuos estudentizados vs predichos para el modelo ajustado en el Cuadro 7.6.

El gráfico mostrado en la Figura 7.9 no muestra ninguna anomalía que haga sospechar problemas en el modelo. Por supuesto que hay un dato que está por fuera de la banda [-2, 2], pero deberíamos esperar que 1 de cada 20 datos (bandas de predicción al 95%) produzca un residuo estudentizados por fuera de esta banda y tenemos 30 datos. No obstante revisaremos otra medida diagnóstico que es la Distancia de Cook. Ésta mide el cambio en el vector de parámetros estimados si eliminamos una a una las observaciones que utilizamos para ajustar el modelo. Luego habrá una distancia de Cook para cada dato: la distancia que se obtiene cuando se elimina ese dato. Cuando esta distancia supera el valor 1, entonces decimos que la observación en cuestión es influyente y un criterio a seguir es ver si nuestras conclusiones persisten aún eliminando esa observación influyente. Si las conclusiones cambian entonces el modelo no es confiable ya que conduce a conclusiones diferentes por el efecto de una única observación. InfoStat permite graficar las distancias de Cook. Estas se muestran en el eje Y y el número de observación en el eje X de un gráfico de dispersión. Para el modelo ajustado en el Cuadro 7.6 el gráfico de las distancias de Cook se muestra en la Figura 7.18. Aunque hay una observación que se destaca del resto (#19), su distancia de Cook es menor que 1 y por lo tanto no debe preocupar.

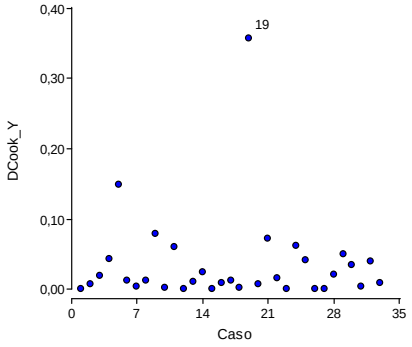


Figura 7.18. Distancias de Cook vs número de caso para el modelo ajustado en el Cuadro 7.6.

Conclusión

El modelo finalmente ajustado para el número de bacterias por cm^3 es el siguiente:

$$\begin{aligned}\hat{Y} &= -1087,2 + 2,89 Tiempo + \\ &+ 56,93 Temperatura - 0,77 Temperatura^2 + \\ &+ 96,68 Nutrientes\end{aligned}$$

El máximo número de bacterias se obtendrá a las 48 horas (máximo tiempo de cultivo evaluado) con una concentración relativa de nutrientes de 1,4 (máximo evaluado) y a una temperatura de 36,97 grados, que se obtiene derivando la ecuación con respecto a la temperatura e igualando a cero.

Ejercicios

Ejercicio 7.1: En este capítulo se introdujo un conjunto de términos que se listan a continuación. ¿Puede recordar su significado?

a) Regresión lineal simple
b) Regresión polinómica
c) Regresión múltiple
d) Residuo
e) Residuo estudentizado
f) Predicho
g) Banda de confianza
h) Banda de predicción
i) Coeficiente de determinación
j) Ordenada al origen
k) Pendiente
l) Prueba de falta de ajuste
m) Coeficiente de determinación ajustado
n) Coeficientes de regresión parcial
o) Residuo parcial

Ejercicio 7.2: Los datos en el archivo [proteinasentrigo] contienen los resultados de la calibración de un instrumento de reflectancia infrarroja para la medición del contenido de proteínas en 24 muestras de trigo. Las variables son: Y = contenido porcentual de proteína y L3L4=índice que combina las reflectancias de radiación infrarroja en las longitudes de onda L3 y L4 (los nombres L3 y L4 no tienen un significado especial). Como la medición infrarroja es más económica que la medición estándar, el objetivo es hallar una expresión matemática para determinar el contenido de proteínas usando sólo el índice L3L4.

a) ¿Describa y estime el modelo propuesto?
b) De una medida de la capacidad predictiva del modelo
c) Construya una banda de confianza para los valores medios estimados
d) Construya un intervalo de confianza (utilizando el InfoStat para el valor de L3L4=8,00

Ejercicio 7.3: ¿A qué temperatura hace ebullición el agua en la cima del Aconcagua? El archivo [Ebullición del agua] contiene datos observados de temperatura de ebullición del agua a distintas altitudes.

a) Estime que temperatura hace ebullición el agua en la cima del Aconcagua, Mendoza (6962 msnm).
b) De un intervalo de predicción para la temperatura de ebullición calculada en el punto anterior.
c) ¿Cómo cambia la temperatura de ebullición cuando se asciende de 0 a 500 m? De acuerdo al modelo propuesto para describir la relación entre estas magnitudes, ¿es este cambio constante, no importado de qué altitud se parta?

Ejercicio 7.4: En un experimento para evaluar la efectividad de un insecticida sobre la sobrevida de dos especies de insectos (A y B) se obtiene que, en ambos casos, es posible ajustar un modelo lineal para la sobrevida (Y) versus la concentración (en ppm) del insecticida utilizado (X), siendo los modelos ajustados los siguientes:

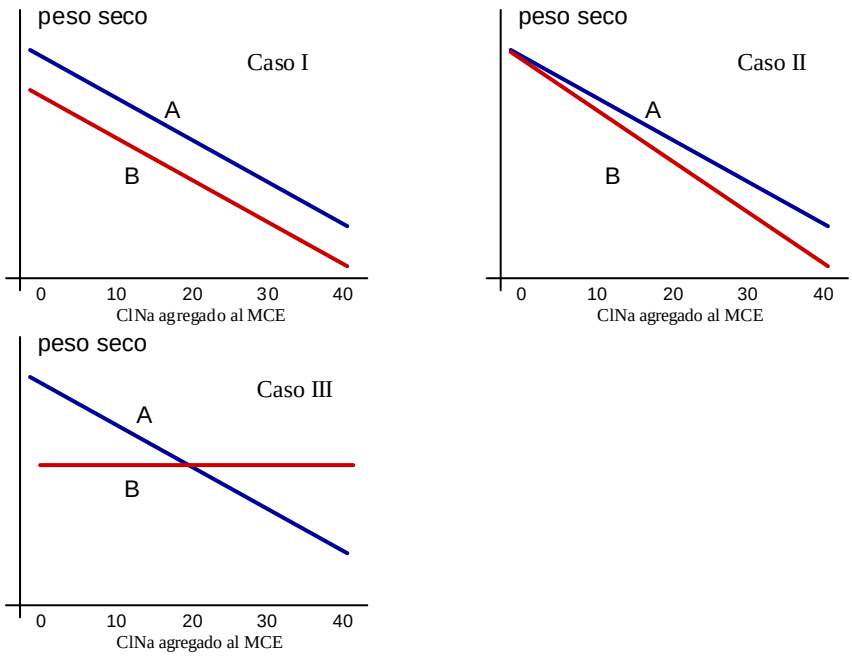
$$\textit{Especie A: } Y = 80 - 15X; \qquad \textit{Especie B: } Y = 60 - 15X.$$

De acuerdo a estos resultados:

a) ¿Es el insecticida igualmente efectivo en ambas especies?
b) ¿Qué interpretación se puede hacer de cada una de estas ecuaciones?
c) ¿Cómo se modifica la sobrevida por cada incremento unitario en la concentración del insecticida agregado?
d) Si se quisiera que ambas especies tengan una sobrevida de a lo sumo 20, ¿cuántas ppm se debería agregar del insecticida?

Ejercicio 7.5: En un ensayo de resistencia a la sequía, dos especies de leguminosas (A y B) fueron comparadas. El experimento consistió en registrar el peso seco total de 10 plantas al cabo de 30 días desde la siembra. Las condiciones comparadas fueron las siguientes: medio de cultivo estándar (MCE), MCE+10 g/l de ClNa, MCE+20 g/l de ClNa, MCE+30 g/l de ClNa, MCE+40 g/l de ClNa. Los tres gráficos que se presentan después de las consignas, muestran tres resultados posibles para esta experiencia. Los gráficos representan las rectas que modelan la esperanza del peso seco en relación al agregado de ClNa en cada caso.

a) ¿Qué conclusión se obtendría, en cada una de estas situaciones acerca de la resistencia a la sequía de ambas especies, asumiendo que si la especie soporta mayor contenido de ClNa será más resistente?
b) ¿Qué significan (o que interpretación tienen) la diferencia y la similitud de las ordenadas al origen de las rectas ajustadas en los casos I, II, y III?
c) ¿Qué significan (o que interpretación tienen) la diferencia y la similitud de las pendientes de las rectas ajustadas en los casos I, II, y III?



Ejercicio 7.6: Se desea probar la efectividad de un nuevo fungicida para el control de roya en trigo. Se probaron distintas dosis en gramos de principio activo por ha (gr.p.a./ha) en 10 parcelas de 100 plantas cada una. A los 15 días de la aplicación se realizó una evaluación del daño, como el tamaño promedio de las machas en hoja bandera. Los datos son los siguientes:

Dosis(X)	100	125	200	250	275	300	325	350	375	400
Daño (Y)	50	48	39	35	30	25	20	12	10	5

a) Ajustar un modelo de regresión lineal para el daño en función de la dosis y construir las bandas de predicción y de confianza.

b) Predecir el daño (tamaño promedio de las manchas) que se hallará si se aplican 260 gr.p.a./ha

Anlaisis de regresión

Ejercicio 7.7: Para estudiar el efecto de la temperatura sobre el vigor durante la germinación, se dispusieron semillas de alfalfa en germinadores a distintas temperaturas. A los 6 días se midió la longitud de las plántulas, obteniéndose los siguientes datos:

T (^{O}C)	Longitud de Plantas (mm)					
10	13	18	15	19	11	17
15	20	24	15	17		
20	22	27	31	21	26	
25	24	25	28	23		

a) ¿Qué diferencia hay en los datos de este ejercicio con respecto a los anteriores?
b) Construir el diagrama de dispersión entre longitud de plántula y temperatura y verificar si existe una tendencia lineal.
c) Realizar un análisis de regresión lineal ¿En cuanto se incrementa la longitud de plantas por cada incremento de un grado en la temperatura?
d) ¿Cuál es el intervalo de confianza para la tasa de cambio de la longitud de plantas?
e) De acuerdo al modelo ajustado, ¿qué temperatura permite obtener mayor vigor?

Ejercicio 7.8: En el archivo [intercepcionderadiacionenmaiz] se encuentran datos de intercepción solar desde los 15 a los 65 días desde la emergencia en un cultivo de maíz de un hibrido comercial. Los datos fueron obtenidos para dos densidades del cultivo Alta (140 kplantas/ha) y Baja (80 kplantas/ha) que se obtuvieron variando la distancia entre líneas. La barra de intercepción de radiación fotosintética activa (RAFA) fue medida cada 10 días. Para cada momento de medición se realizaron determinaciones en 8 puntos del cultivo elegidos al azar. En cada punto se realizaron 4 determinaciones de la RAFA y lo que se reporta en el archivo de datos es el promedio de estas 4 determinaciones. Por lo tanto el archivo de datos tiene 6 determinaciones x 8 puntos de muestreo x 2 densidades de siembra=96 registros y tres columnas: Densidad (Alta, Baja), Días (días desde la emergencia, 15, 25, …) y RAFA. El propósito del estudio es establecer que densidad de siembra es más efectiva para la intercepción de la radiación solar. Como una forma de medir esta eficiencia se quiere calcular el tiempo necesario desde la emergencia para captar el 50% de la RAFA en ambas densidades.

a) Ajuste el modelo de regresión apropiado.
b) En base al modelo ajustado calcule a los cuantos días se alcanza, en cada densidad, la captura del 50% de la RFA.

Capítulo 8

Estudios de correlación y asociación

Julio A. Di Rienzo

8. Estudios de correlación y asociación

Motivación

Es común en las Ciencias Biológicas buscar relaciones entre variables y cuantificar la magnitud de estas asociaciones. Cuando las variables que queremos relacionar son cuantitativas el método estadístico más usado es el **análisis de correlación**. Cuando las variables son cualitativas o categorizadas, el análisis de **tablas de contingencia** y las **pruebas de bondad de ajuste** son estrategias usuales a seguir. En este Capítulo se desarrollan estas estrategias de análisis.

Conceptos teóricos y procedimientos

Presentaremos tres medidas frecuentemente usadas para medir la correlación entre pares de variables cuantitativas: el **Coeficiente de Correlación de Pearson**, el **Coeficiente de Correlación de Spearman** y el **Coeficiente de Concordancia**.

Coeficiente de correlación de Pearson

Es un estadístico cuyos valores varían entre -1 y 1. En cualquiera de los extremos de este rango la correlación es máxima pero en sentidos opuestos. Mientras que una correlación cercana a 1 indica una asociación positiva (ambas variables crecen y decrecen conjuntamente), una correlación cercana a -1 indica lo contrario, es decir, que si una variable crece la otra disminuye y viceversa. La correlación de Pearson (ρ) - se lee rho- entre las variables X e Y se define como:

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

En la expresión del coeficiente, el término cov(X,Y) se refiere a la covarianza entre X e Y, y Var(X) y Var(Y) son las varianzas de X e Y respectivamente. La covarianza es una medida que va entre –infinito y +infinito y cuanto más grande en valor absoluto es esta cantidad más asociación hay entre las variables. Al dividir la covarianza por la raíz cuadrada del producto de las varianzas, se confina el valor del cociente al intervalo [-1,1]. Entonces, este cociente permite tener una escala acotada para medir la covariación. Es estimador de (ρ), que se simboliza usualmente con la letra latina equivalente "r", se calcula según la expresión (1). El número "n" en esta expresión se refiere al número de pares (X,Y).

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right)}} \qquad (1)$$

Un caso especial ocurre cuando $\rho = 0$. En tal caso no hay asociación entre X e Y y diremos que X e Y no están correlacionadas. Cuando X e Y siguen una distribución normal bivariada, es posible construir un contraste de hipótesis para H_0:$\rho = 0$ vs. H_1: $\rho \neq 0$. El estadístico utilizado para realizar este contraste es:

$$T = r\sqrt{\frac{n-2}{1-r^2}} \overset{H_0}{\sim} T_{n-2}$$

Este estadístico sigue una distribución T de Student con n-2 grados de libertad cuando la hipótesis nula es cierta.

Aplicación

Ácidos grasos en semillas

El ácido oleico es un ácido graso mono insaturado de la serie omega 9, típico de los aceites vegetales como el aceite de oliva, del aguacate (palta), etc. El ácido linoleico es un ácido graso poli insaturado esencial para el organismo humano (el organismo no puede sintetizarlo) y tiene que ser ingerido con los alimentos. Al ácido linoleico y a sus derivados se les conoce como ácidos grasos omega 6. El ácido linolénico es también un ácido graso esencial de la familia omega-3. Los datos en el archivo [Aceites] tienen determinaciones de los tres ácidos grasos y contenido de proteínas en diversas

muestras de semillas de un híbrido comercial de girasol. Se quiere estudiar cómo se relaciona el contenido de estos ácidos grasos y el contenido proteico.

Estrategia de análisis

Es útil para estudiar las relaciones entre variables cuantitativas graficarlas unas versus las otras mediante diagramas de dispersión. Las matrices de diagramas de dispersión permiten tener una imagen simultánea de todas estas relaciones. Aunque los gráficos sirven para anticipar los resultados del análisis, la cuantificación de la asociación es un paso esencial y para ello se debe calcular alguna de las medidas de correlación.

La imagen de la matriz de diagramas de dispersión para los datos del archivo *[Aceite]* se muestra en la Figura 8.1.

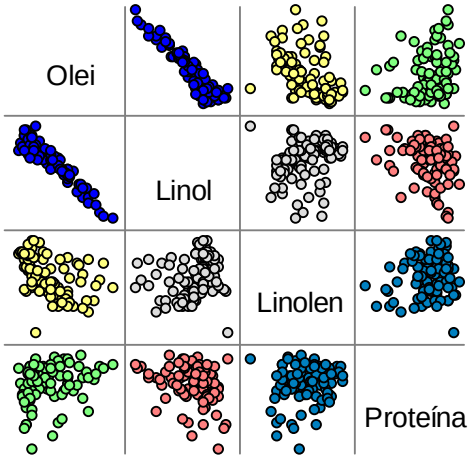


Figura 8.1. Matriz de diagramas de dispersión para el contenido de distintos ácidos grasos y proteínas.

Es fácil ver que los ácidos: oleico y linoleico están fuertemente correlacionados y que esta correlación es negativa. La cuantificación de estas relaciones se observan en el Cuadro 8.1. En este cuadro se presenta una matriz que contiene los coeficientes de correlación de Pearson (triangular inferior) y sus pruebas de hipótesis respectivas (triangular superior). Para obtener la matriz del Cuadro 8.1, en el software InfoStat seleccione el menú *Estadísticas >>Análisis de correlación.* A continuación aparecerá el diálogo de selección de variables que debe llenarse como se muestra en la Figura 8.2 (izquierda) y a continuación el diálogo que permite especificar qué medida de correlación utilizar Figura 8.2 (derecha). Seleccionar la opción *Pearson.* En la diagonal principal se observan las correlaciones de cada variable con sí misma. Este coeficiente es siempre 1 y no tiene ningún valor interpretativo. Por debajo de la diagonal principal

(triangular inferior) están los coeficientes de correlación calculados. Por encima de la diagonal principal (triangular superior) los valores p correspondientes para las hipótesis

H_0:ρ =0 vs. H_1: $\rho \neq 0$.

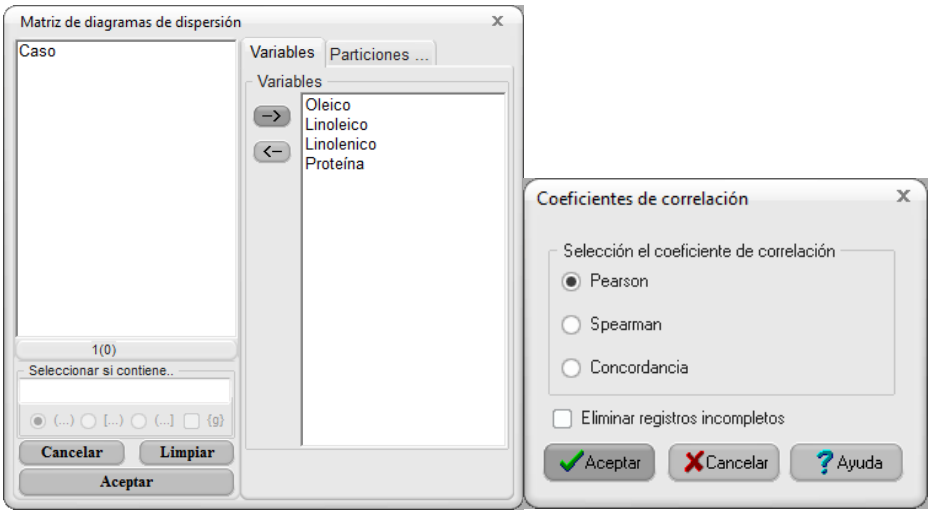


Figura 8.2. InfoStat. Ventanas de diálogo para el cálculo de la correlación de Pearson.

Se observa que la correlación entre oleico y linoleico es fuerte, negativa (-0,93) y significativa (p< 0,000001). Se correlaciona negativamente con el ácido linolénico y aunque esta correlación es débil (-0,47) es significativa (p= 0,000002). Por otra parte el ácido oleico se correlaciona positivamente con el contenido de proteínas (0,29) e igualmente aunque esta correlación es pequeña, es significativa (p=0,004365). La interpretación de los otros coeficientes es similar. Por último se quiere observar que la correlación entre ácido linolénico y el contenido de proteínas es positiva (0,16) pero no significativa (p= 0,119157).

Cuadro 8.1. Coeficiente de correlación de Pearson. En la diagonal principal se observan las correlaciones de cada variable con sí misma. Este coeficiente es siempre 1 y no tiene ningún valor interpretativo. Por debajo de la diagonal principal están los coeficientes de correlación calculados. Por encima de la diagonal principal los p-valores para las hipótesis H0:$\rho = 0$ vs. H1: $\rho \neq 0$.

Correlación de Pearson: Coeficientes\probabilidades

	Oleico	Linoleico	Linolenico	Proteína
Oleico	1,000000	0,000000	0,000002	0,004365
Linoleico	-0,934921	1,000000	0,017301	0,006484
Linolenico	-0,467880	0,245027	1,000000	0,119157
Proteína	0,291491	-0,278916	0,161833	1,000000

Conclusión

Se halló una fuerte correlación negativa entre el contenido de ácido oleico y linoleico. Ambos ácidos grasos se correlacionan positiva y negativamente con el contenido de proteínas respectivamente, aunque estas correlaciones son débiles. El ácido linolénico no se correlaciona con el contenido de proteínas y se correlaciona negativamente con el ácido oleico y positivamente con el linolénico, aunque estas correlaciones son también débiles.

Coeficiente de correlación de Spearman

El **coeficiente de correlación de Spearman** (también conocido como coeficiente de correlación no paramétrico de Spearman) es una medida de correlación que mide la monotonía con que se mueven dos variables aleatorias (X e Y). Para calcular el coeficiente se substituyen los valores observados X e Y por sus posiciones en una lista ordenada de menor a mayor. Esta transformación se conoce como **transformación rango** (del inglés *rank transformation*). En la siguiente tabla se muestra la aplicación de esta transformación a los datos X e Y. La columna "d" se explicará más adelante.

X	**Y**	R(X)	R(Y)	d
10,2	20,2	7	7	0
8,0	6,3	3	3	0
14,1	15,8	4	4	0
15,0	19,1	4	6	-1
15,9	18,7	6	4	1
11,3	10,2	3	3	0
6,0	8,8	1	2	-1

Si X_i^r e Y_i^r son los valores transformados del par $\left(X_i, Y_i\right)$ a partir de los rangos de X y de Y , R(X) y R(Y), definimos $d_i = X_i^r - Y_i^r$ entonces el coeficiente de correlación de Spearman se calcula como:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Para los datos de la tabla donde se ejemplifica la transformación rango el coeficiente sería:

$$r_s = 1 - \frac{6\left((-1)^2 + (1)^2 + (-1)^2\right)}{7\left(7^2 - 1\right)} = 0{,}9464$$

Cuando existen valores repetidos (empates), ya sea en X o en Y, no hay un orden natural para esas observaciones. Por ejemplo si se tuviera la secuencia ordenada: {5, 3, 7, 5, 6, 12, 5, 12}, ¿cuál es el número de orden del primer 5? Por convención la transformación rango se realiza en dos etapas. En la primera se ordena la secuencia numérica {3, 5, 5, 5, 6, 7, 12, 12} y luego se asignan número correlativos: {1, 2, 3, 4, 5, 6, 7, 8}. No está claro porque a uno de los cincos le tocó un 2 y otro un 4 o porque uno de los 12 tiene un 7 y el otro un 8. Solución: promediar los órdenes de los datos repetidos. La transformación rango para estos datos sería: {1, 3, 3, 4, 5, 6, 7,5, 7,5}. Luego los datos originales fueron asignados de la siguiente forma {5(3), 3(1), 7(6), 5(3), 6(5), 12(7,5), 5(3), 12 (7,5)},

Cuando ocurren empates se recomienda utilizar, como algoritmo de cálculo de r_s, la fórmula de cálculo del coeficiente de correlación de Pearson pero aplicada a los pares transformados $\left(X_i^r, Y_i^r\right)$. El coeficiente de Spearman también varía entre -1 y 1 y se interpreta de manera similar a los descripto para el coeficiente de correlación de Pearson: Valores cercanos a 1 o -1 implica alta correlación positiva o negativa respectivamente y 0 falta de correlación.

Un contraste de hipótesis para H_0:$\rho_s = 0$ vs. H_1: $\rho_s \neq 0$, se puede realizar utilizando el hecho de que el estadístico tiene distribución T de Student con n-2 grados de libertad cuando la hipótesis nula es cierta.

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

Mientras que el coeficiente de correlación de Pearson mide el grado de proporcionalidad de las cambios entre los pares (X,Y), el coeficiente de Spearman mide monotonía de cambio sin importar la proporcionalidad. En este sentido es un coeficiente que mide una forma más genérica de asociación. Esto tiene sus ventajas y desventajas. La ventaja es que se puede tener una alta asociación aún cuando se X e Y

se midan en escalas no lineales. Por esta misma razón, tener una alta correlación de Spearman implica que los valores de una de las variables sean predecibles por los valores de la otra. Esto podría ser indeseable cuando se trata de utilizar una variable fácil de medir como subrogante (substituta) de otra difícil de medir. Para este caso nos interesaría que la correlación midiera proporcionalidad de los cambios. Debe decirse por otra parte que cuando el coeficiente de correlación de Pearson es alto (en valor absoluto), el coeficiente de Spearman también lo es.

Aplicación

Ácidos grasos en girasol

Aplicaremos el cálculo del coeficiente de correlación de Spearman a los mismos datos que se utilizaron en la sección anterior para ejemplificar el cálculo del coeficiente de correlación de Pearson: archivo *[Aceites].*

Estrategia de análisis

La estrategia de análisis es similar a la planteada para el caso del coeficiente de Pearson. Para invocar el cálculo del coeficiente de Spearman se debe proceder de manera similar a lo hecho anteriormente, eligiendo el menú *Estadísticas> Análisis de correlación* y completando las ventanas como se muestra en la Figura 8.3. Obsérvese que en el diálogo derecho de la imagen se seleccionó Spearman.

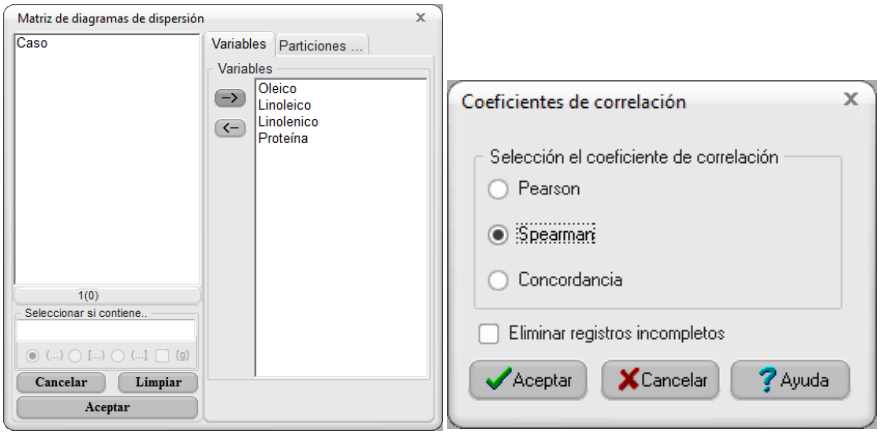


Figura 8.3. InfoStat. Ventanas de diálogo para el cálculo del coeficiente de correlación de Spearman.

La matriz coeficientes de correlación y *valores p* se muestra en el Cuadro 8.2. No hay diferencias con los resultados presentados anteriormente (Cuadro 8.1).

Conclusión

Se concluye de idéntica manera que para el caso del coeficiente de correlación de Pearson.

Cuadro 8.2. Correlación de Spearman. En la diagonal principal se observan las correlaciones de cada variable con sí misma. Este coeficiente es siempre 1 y no tiene ningún valor interpretativo. Por debajo de la diagonal principal están los coeficientes de correlación y por encima de ella se encuentran los valores p para las hipótesis H_0:$\rho = 0$ vs. H_1: $\rho \neq 0$.

Correlación de Spearman: Coeficientes\probabilidades

	Oleico	Linoleico	Linolenico	Proteína
Oleico	1,000000	0,000000	2,46E-09	0,000822
Linoleico	-0,881292	1,000000	0,000271	0,004209
Linolenico	-0,567491	0,367253	1,000000	0,202271
Proteína	0,339291	-0,292626	0,132711	1,000000

Coeficiente de concordancia

Es una medida de la **concordancia** de dos variables aleatorias. Va más allá de medir proporcionalidad como lo hace Pearson, este coeficiente mide el grado de igualdad de mediciones. Tiene la siguiente expresión.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

En la expresión el factor ρ hace referencia al coeficiente de correlación de Pearson, σ_x, σ_y a las desviaciones estándares poblacionales de X e Y, σ_x^2, σ_y^2 a las correspondientes varianzas y μ_x, μ_y a las respectivas medias poblacionales. El estimador del coeficiente de concordancia modificado tiene la siguiente expresión:

$$\rho_c = \frac{\frac{2}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}\right)}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2}$$

Aplicación

Condición corporal de animales

Una herramienta de gran utilidad para el manejo nutricional del rodeo, es la determinación de la "condición corporal" de los vientres. Una de las escalas va del 1 al

9, siendo 1 el valor correspondiente a una vaca extremadamente delgada y 9 el correspondiente a una vaca muy gorda.
¿Es la condición corporal un criterio reproducible entre distintos observadores que pueda utilizarse como estándar y para la valoración del estado de los vientres? El coeficiente de concordancia es el coeficiente ideal para medir la reproducibilidad de una medida.

Estrategia de análisis

Para evaluar la calidad de la condición corporal con escala 1-9, se utilizó un rodeo de 120 animales y cada animal fue valorado en su condición corporal independientemente por 4 técnicos calificados. Los 120 animales se seleccionaron para reflejar condiciones corporales que cubrieran el rango completo de la escala de medición. Los datos están disponibles en el archivo [Condicion corporal]. Se solicitó a los técnicos que se abstengan de introducir valores fraccionarios manteniéndose en la escala de los números enteros.

Siguiendo el mismo procedimiento que con los otros dos coeficientes pero eligiendo la opción Concordancia en la ventana de diálogo correspondiente se obtienen los resultados que se presentan en el Cuadro 8.3. Se observa que las concordancias son todas positivas, cercanas a 0,85.

Cuadro 8.3. Coeficiente de Concordancia. En la diagonal principal el coeficiente es siempre 1. Por debajo de la diagonal principal están los coeficientes de concordancia. Por encima de la diagonal principal se observa el código "sd" (sin dato) ya que no existe una prueba para la hipótesis de coeficiente H_0:$\rho = 0$ vs. H_1: $\rho \neq 0$ implementada en InfoStat.

Concordancia: Coeficientes\probabilidades

	T1	T2	T3	T4
T1	1,00	sd	sd	sd
T2	0,84	1,00	sd	sd
T3	0,84	0,86	1,00	sd
T4	0,87	0,88	0,85	1,00

Conclusión

Técnicos bien entrenados pueden reproducir índice de condición corporal con una correspondencia promedio no inferior al 80%.

Análisis de tablas de contingencia

Abordaremos la problemática del estudio y cuantificación del grado y sentido de la asociación entre variables de naturaleza categórica mediante el análisis de tablas de contingencia. Este material es introductorio y no pretende cubrir el amplio espectro de

métodos estadísticos disponibles para el estudio de variables categóricas. Un material de referencia sobre este tema es Agresti (1990).

Una tabla de contingencia es una tabla de doble entrada que contiene en el encabezado de filas y columnas las modalidades de dos variables categóricas asignadas a filas y columnas respectivamente. El cuerpo de la tabla contiene las frecuencias observadas para las combinaciones de las modalidades correspondientes a las filas y columnas. Además, una fila y una columna adicionales contienen los totales de filas y columnas respectivamente. La Figura 8.4 ilustra la forma general de una tabla de contingencia en la que dos variables categóricas llamadas A y B son asignadas a filas y columnas respectivamente. La variable A tiene tres modalidades: A1, A2 y A3, mientras que B sólo tiene dos: B1 y B2. Dada una muestra de tamaño "n" en la que se registra para cada unidad muestral la modalidad observada de A y de B, el contenido de cada celda corresponde al número de casos que comparten simultáneamente las correspondientes modalidades de A y B. Por lo tanto e una tabla de contingencia contiene son números enteros mayores o iguales que cero. Las tablas de contingencia tienen una fila adicional que totaliza el contenido de las columnas (marginales columna) y una columna adición que totaliza el contenido de las filas (marginales fila). Además, hay una celda adicional que contiene el total de la tabla.

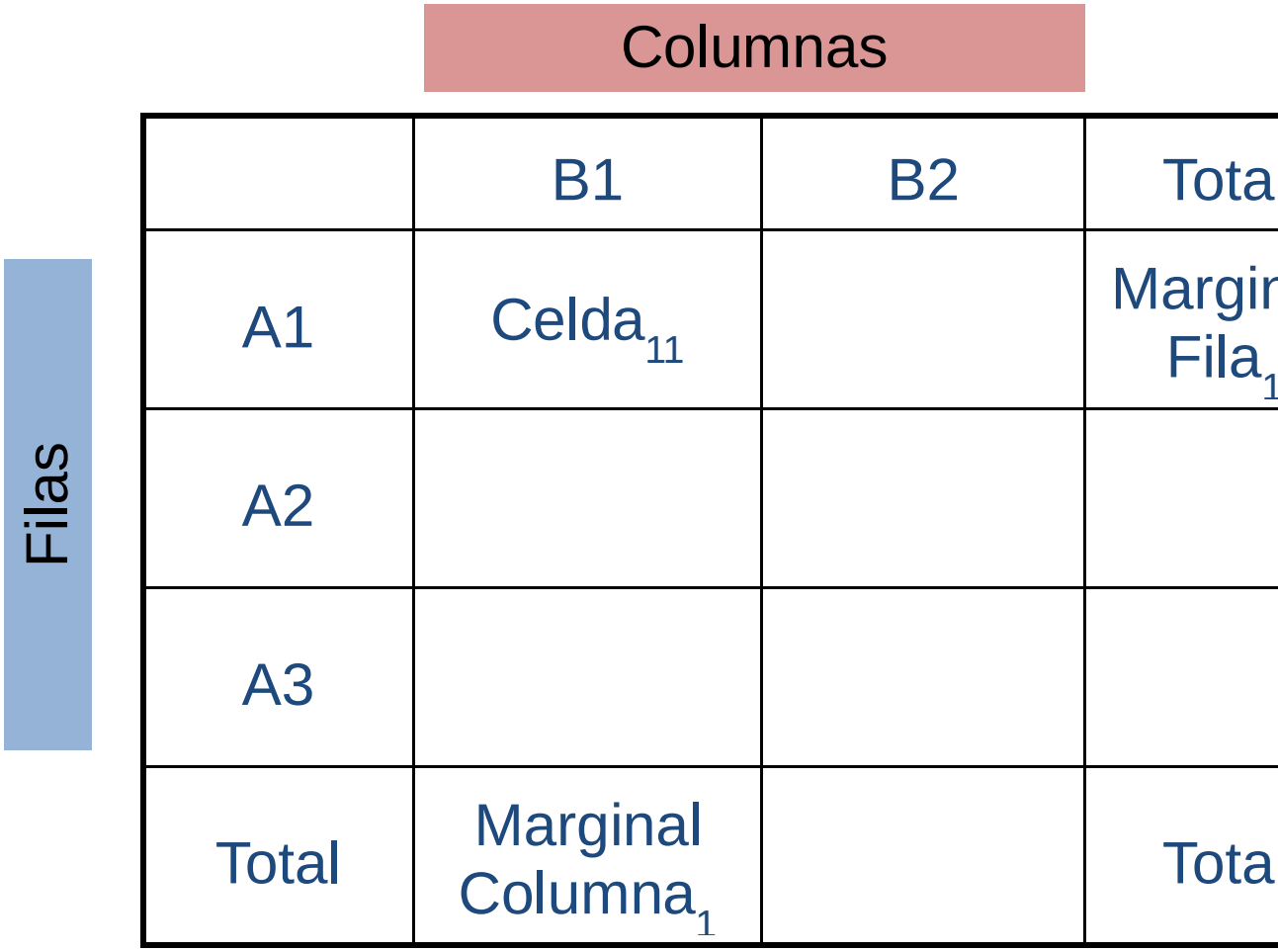


Figura 8.4. Esquema general de una tabla de contingencia para dos variables A y B, la primera con 3 modalidades: A1, A2 y A3 y la segunda con 2: B1 y B2.

Un ejemplo típico es el siguiente: Se quiere evaluar si la germinación o no de semillas está asociada a la condición de haber sido tratadas con un fungicida. En la siguiente tabla, aproximadamente 3000 semillas, divididas en dos lotes de tamaño similar, fueron tratadas con fungicida o dejadas como control no tratadas. Luego las semillas se hicieron germinar y se registró el número de germinadas y no germinadas en cada uno de los grupos: control y tratadas con fungicida. El resultado de este conteo se presenta en la Tabla 8.1.

Tabla 8.1: Tabla de contingencia donde se resume el conteo de semillas germinadas y no germinadas según que fueran tratadas o no (control) con fungicida.

Condición	no germinó	germinó	Total
Control	245	1190	1435
Fungicida	123	1358	1481
Total	368	2548	2916

La pregunta que el investigador quiere responder es si la aplicación del fungicida brinda una protección que finalmente se traduce en un mayor poder germinativo. Los porcentajes de germinación en uno y otro grupo parecen favorecer esa conclusión (Tabla 8.2).

¿Cómo se prueba que la mayor germinación observada en las semillas tratadas es evidencia estadísticamente significativa de que el uso de un fungicida mejora el poder germinativo? Hay algunas alternativas para probar este postulado pero utilizaremos un basada en la hipótesis (nula) de que la germinación una semilla es un evento independiente de la semilla haya sido "curada" con fungicida.

Tabla 8.2: Tabla de contingencia donde se resume el porcentaje de semillas germinadas y no germinadas según que fueran tratadas o no (control) con fungicida.

Condición	no germinó (%)	germinó (%)	Total
Control	17,07	82,93	100,00
Fungicida	8,31	91,69	100,00
Total	12,62	87,38	100,00

*La clave para probar si la hipótesis es sustentada por los datos es calcular las **frecuencias esperadas (E)** (suponiendo cierta la hipótesis de independencia) y compararlas con las **frecuencias observadas (O)**. La*

Tabla 8.3 contiene tales frecuencias esperadas**.** Estas frecuencias se comparan con las observadas mediante el estadístico chi-cuadrado cuya expresión es la siguiente:

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\left(\frac{\left(O_{ij} - E_{ij}\right)^2}{O_{ij}}\right)$$

En la expresión anterior O_{ij} hace referencia a la frecuencia observada en la i-ésima fila, j-ésima columna de la tabla de contingencia, E_{ij} a la correspondiente frecuencia esperada y los argumentos f y c, de los términos de sumatoria, al número de filas y columnas de la tabla de contingencia respectivamente. En el ejemplo $O_{21} = 123$ y $E_{21} = 186,9$, mientras que $f = 2$ y $c = 2$. Por la forma en que se calculan, las

frecuencias esperadas no son necesariamente números enteros y no deben redondearse.

Si la hipótesis nula es cierta, el estadístico presentado se distribuye como una **Chi-cuadrado** con $(f-1)(c-1)$ grados de libertad (en este ejemplo sería 1). Esta prueba es siempre unilateral derecha por lo que para un nivel de significación del 5% la región de aceptación estará delimitada a la derecha por el cuantiles 0,95 de una chi-cuadrado con 1 grado de libertad.

Si utilizamos la calculadora de *Probabilidades y cuantiles* del menú *Estadísticas* de InfoStat obtendremos un valor aproximado 3,84 para este cuantil (en los parámetros de la chi-cuadrado que muestra InfoStat aparece, además de los grados de libertad, un segundo parámetro, el parámetro de no centralidad, este debe dejarse en cero que es su valor por defecto). Luego si el valor observado del estadístico -para los datos de la Tabla 8.1- supera este límite diremos que la hipótesis de independencia es falsa y por lo tanto la insinuación de que el fungicida ejerce un efecto protector que beneficia la germinación debe aceptarse.

El valor calculado de chi-cuadrado es 50,81, muy por encima de 3,84. Asimismo, si calculáramos su p-valor éste sería <0,0001 con lo que, para un nivel de significación del 5%, concluiríamos de idéntica manera rechazando la hipótesis nula. Más adelante la se discutirá como utilizar el software InfoStat para obtener este estadístico.

Tabla 8.3: Tabla es frecuencias esperadas de semillas germinadas y no germinadas según que fueran tratadas o no (control) con fungicida.

Condición	no germinó	germinó	Total
Control	181,1	1253,9	1435
Fungicida	186,9	1294,1	1481
Total	368	2548	2916

¿Cómo se calcularon las frecuencias esperadas de la Tabla 8.3?

*Si no hubiera efecto fungicida, entonces la mejor estimación de la probabilidad de germinación sería dividir el número total de semillas germinadas (2548) por el total de semilla utilizadas (2916). Esta probabilidad estimada es 0,8738. Luego usando esa probabilidad podemos calcular el número esperado de semillas germinadas para el total de semilla control (1435) y para el total de semillas tratadas (1481). El cálculo es muy sencillo. El número esperado de semillas germinadas en el control (si no hubiera efecto fungicida) debería estimarse multiplicando la probabilidad (marginal) de germinación por el total de semillas en el control, esto es: 1435*0,8738=1253,9 y de idéntica manera el número esperado de semillas germinadas en el grupo de semillas tratadas (siguiendo con la suposición de que no existe efecto fungicida) sería 1481*0,8738=1294,1. Los números 1253,9 y 1294,1 son los que aparecen en la columna "germinó" de la*

Tabla 8.3.

Luego los número que aparecen en la columna "no germinó" se obtienen por diferencia (181,1 es lo que le falta a 1253,9 para sumar 1435).Como regla práctica las frecuencias esperadas se calculan según la expresión y los grados de liberta como ($(f-1)(c-1)$.

$$celda_{ij} = \frac{total\ fila_i * total\ columna_j}{total\ general}$$

Razón o cociente de chances

Es bastante intuitivo comparar la probabilidad de que ocurra un evento bajo dos condiciones diferentes si $\pi_A^{(1)}$ representa la probabilidad de que ocurra el evento A en la condición 1 y $\pi_A^{(2)}$ su probabilidad en la condición 2, entonces $RR = \pi_A^{(1)} / \pi_A^{(2)}$ es conoce como **riesgo relativo**. Este estadístico es útil para comparar probabilidades, es simple de interpretar y mide cuantas veces un evento es más probable en una condición que en otra. Sin embargo bajo cierto plantes de muestro el riesgo relativo no puede calcularse. Una forma diferente de comparar probabilidades es utilizar el **cociente de chances** (*odds ratio* en inglés).

Si un suceso A tiene probabilidad π_A, su **chance** se define como: $chance(A) = \pi_A / (1-\pi_A)$. Esta es una forma diferente de representar una probabilidad y su resultado se interpreta como las veces que ocurre un éxito por cada ocurrencia de un fracaso. Por ejemplo, si $\pi_A = 0,50$ la $chance = 1$ e indica que por cada fracaso ocurre un éxito. Éste es el ejemplo de la tirada de una moneda donde se dice que 1 de cada 2 tiradas sale cara (o cruz). Si $\pi_A = 0,95$ la $chance = 19$ y su resultado se interpreta diciendo que 19 de cada 20 veces son éxitos.

Este cociente mide cuanto mayor (o menor) es la chance de que ocurra un éxito bajo una condición respecto de la otra. Cuando la probabilidad de éxito es pequeña en ambas condiciones (inferiores a 0,20), el cociente de chances se aproxima bastante al riesgo relativo y se considera una buena aproximación de éste.

Para el ejemplo del fungicida, la probabilidad estimada de que una semilla germine cuando pertenece al grupo Control es 1190/1435=0,8292683. La probabilidad de esto

ocurra en el grupo al que se le aplica fungicida es 1358/1481=0,916948. La chance en el control es 0,8292683/(1-0,8292683)=4,857143 y la chance en el grupo con fungicida es 0,916948/(1-0,916948)= 11,04065. Así que, en el control, la relación éxitos-fracasos es 5 a 1 (por cada 5 éxitos ocurre un fracaso – 5 de cada 6 semillas germinan) mientras que esta relación es 11 a 1 en las semillas tratadas. La razón de chances de que una semilla germine bajo el tratamiento con fungicida respecto del control es 11,04065/4,857143=2,27 y diremos que la chance de que una semilla germine en el grupo tratado con fungicida es aproximadamente 2 veces la chance de que eso ocurra en el grupo control. Es útil mirar el intervalo de confianza para la razón de chances. El intervalo bilateral se obtiene según la expresión dada abajo, donde OR representa la razón de chances estimada, n_{ij} son las frecuencias observadas en cada celda de la tabla 2x2 y $z_{1-\alpha/2}$ es el cuantil $(1-\alpha/2)$ de una distribución Normal estándar:

$$\exp\left(\ln(OR) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right)$$

El software InfoStat lo calcula y para este ejemplo el intervalo de confianza al 95% es [1,81, 2,86]. La amplitud de este intervalo refleja la calidad de la estimación. En este caso el intervalo indica que la razón de chances está aproximadamente entre 2 y 3 e indica una buena estimación.

Aplicación

Condición corporal y éxito de inseminación

En un establecimiento ganadero se quiere establecer si la condición corporal de las vacas (medida en la escala del 1 al 5) afecta y de qué manera el éxito de la inseminación.

Estrategia de análisis

Para evaluar la relación entre CC y éxito de la inseminación, 160 vacas fueron inseminadas y se registró su CC. Sólo se consideraron vacas con CC 2, 3 y 4. Posteriormente se estableció si las vacas habían quedado preñadas o no. Los datos generados por este ensayo se muestran en la Tabla 8.4.

Tabla 8.4: Tabla es frecuencias observadas de vacas preñadas y no preñadas inseminadas artificialmente y clasificadas según su condición corporal.

CC	Preñadas	No preñadas	Total
2	23	7	30
3	76	4	80
4	46	4	50
Total	147	13	160

Si asumimos como hipótesis nula que la condición corporal no se vincula con el éxito de la inseminación, los valores esperados pueden calcularse. Utilizaremos InfoStat para calcular las frecuencias esperadas y calcular el estadístico chi-cuadrado. Para ello debemos reorganizar los datos en una tabla conteniendo tres columnas como se muestra a en la Tabla 8.5. Estos datos se encuentran cargados en el archivo [PreñezyCCorporal]. Una vez abierto el archivo debe invocarse el análisis de una tabla de contingencia. Para ello selecciones el menú *Estadísticas*, ítem *Datos categorizados*, sub-ítem *Tablas de contingencias*.

Tabla 8.5: Tabla que muestra la forma en que deben organizarse los datos para ser procesados por InfoStat

CC	Preñada	Conteo
2	SI	23
3	SI	76
4	SI	46
2	NO	7
3	NO	4
4	NO	4

Una vez que se acepta este diálogo aparece la ventana de selección de variables. En ella la condición corporal (CC) y la Preñez deben asignarse a la lista de *Criterios de clasificación*. La variable conteniendo los conteos debe asignarse a la lista de *Frecuencias*. La Figura 8.5 ilustra estas asignaciones.

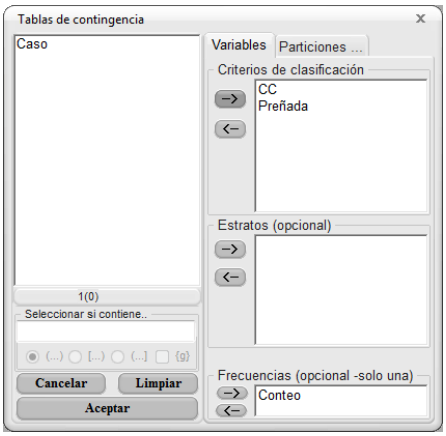


Figura 8.5. InfoStat. Ventana de diálogo que muestra InfoStat para la selección de variables del análisis de tablas de contingencias.

Una vez que se aceptan las especificaciones del diálogo de selección de variables, aparecerá la ventana de opciones del análisis de tablas de contingencia. Esta ventana

tiene dos solapas: *Selección de filas y columnas* y *Opciones*. El contenido de ambas solapas se muestra en la Figura 8.6. Obsérvese que la columna que tiene la información sobre el éxito de la inseminación (Preñada) e ubicó en la lista "Columnas" y la que contiene la información sobre la CC en la lista "Filas". Esta forma de asignación reproduce el arreglo de datos de la Tabla 8.4. La ubicación de Preñada y CC como columnas o filas es indiferente a los fines de probar la independencia de estos criterios de clasificación pero la elección de su posición en filas o columnas puede facilitar la presentación e interpretación de los resultados. En la solapa *Opciones* se han tildado tres opciones que no se encuentran tildadas por defecto: *Frecuencias relativas por filas*, *Frecuencias esperadas bajo independencia*, *Desviaciones de lo esperado bajo indep., estandarizadas* y *Frecuencias relativas como porcentajes*. Estas opciones tienen su correlato en los resultados que se presentan en la próxima sección.

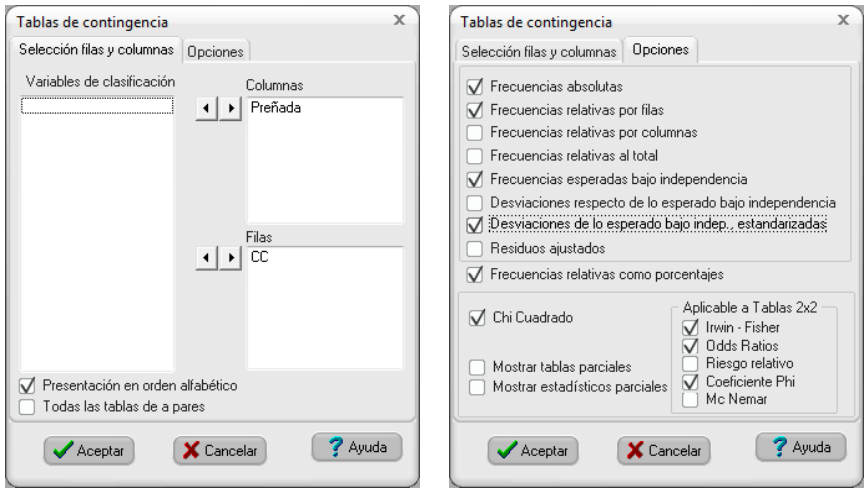


Figura 8.6. InfoStat. Ventana de diálogo para la selección de filas y columnas y opciones de resultados en el análisis de tablas de contingencia con InfoStat.

El Cuadro 8.4 presenta el resultado del análisis de los resultados mostrados en la Tabla 8.4. En esta salida se muestra la tabla de frecuencias absolutas (los datos observados), la tabla de frecuencias relativas por filas, expresadas como porcentajes, la tabla de frecuencias esperadas bajo la hipótesis de independencia y la tabla de desviaciones estandarizadas respecto de lo esperado bajo la hipótesis de independencia.

El estadístico chi-cuadrado de para la tabla examinada fue 8,79. Para una chi-cuadrado con 2 grados de libertad, su *valor p* es 0,0123. Con un nivel de significación del 5% este *valor p* indica que la hipótesis nula debe rechazarse o como usualmente se dice el resultado del a prueba chi-cuadrado fue significativo. A veces es útil saber porqué la hipótesis nula falla. La tabla de desvíos estandarizados respecto de lo esperado permite individualizar las partes de la tabla de frecuencias que más contribuyen al chi-cuadrado. Si se eleva al cuadrado cada una de las entradas de esta tabla, su suma reproduce el

estadístico chi-cuadrado (8,79). Por lo tanto cuanto mayor en valor absoluto es una entrada mayor es su contribución al chi-cuadrado. Como regla práctica, si una entrada tiene valor absoluto mayor que 2 esto es indicativo que está haciendo una contribución significativa al chi-cuadrado. En el ejemplo sólo la celda correspondiente a la condición corporal 2, columna “No preñada” tiene un desvió estandarizado mayor que 2 (2,81), indicando que, cuando la condición corporal es 2, hay más fracasos de la inseminación de lo esperado si la condición corporal no estuviera relacionada con el éxito de esta técnica de manejo reproductivo.

Cuadro 8.4. Tabla de contingencias en el que se presenta una tabla de frecuencias absolutas (los datos observados), una tabla de frecuencias relativas por filas, expresadas como porcentajes, la tabla de frecuencias esperadas bajo la hipótesis de independencia y una tabla de desviaciones estandarizadas respecto de lo esperado bajo la hipótesis de independencia.

Tablas de contingencia

Frecuencias: Conteo
Frecuencias absolutas
En columnas:Preñada

CC	NO	SI	Total
2	7	23	30
3	4	76	80
4	4	46	50
Total	15	145	160

Frecuencias relativas por filas(expresadas como porcentajes)
En columnas:Preñada

CC	NO	SI	Total
2	23,33	76,67	100,00
3	5,00	95,00	100,00
4	8,00	92,00	100,00
Total	9,38	90,63	100,00

Frecuencias esperadas bajo independencia
En columnas:Preñada

CC	NO	SI	Total
2	2,81	27,19	30,00
3	7,50	72,50	80,00
4	4,69	45,31	50,00
Total	15,00	145,00	160,00

Desviaciones de lo esperado bajo indep., estandarizadas
En columnas:Preñada

CC	NO	SI	Total
2	2,50	-0,80	sd
3	-1,28	0,41	sd
4	-0,32	0,10	sd
Total	sd	sd	sd

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	8,79	2	0,0123
Chi Cuadrado MV-G2	7,33	2	0,0257
Coef.Conting.Cramer	0,17		
Coef.Conting.Pearson	0,23		

Conclusión

En el rodeo evaluado, la condición corporal afecta significativamente el éxito de la inseminación y el análisis sugiere que la condición corporal 2 está relacionada con una mayor frecuencia de fracasos. No hay evidencia que sugiera diferencias en los resultados de la inseminación entre las condiciones 3 y 4.

Pruebas de bondad de ajuste

Un caso de tabla de contingencia diferente al presentado anteriormente es aquel en el que las frecuencias esperadas son deducidas desde un modelo teórico cuyos parámetros se estiman independientemente de los datos disponibles. Un ejemplo clásico de esta situación esta relacionado con un experimento de Gregor Mendel.

*Mendel (1822-1884) fue un monje naturalista nacido en Heinzendorf, Austria, considerado como padre de la genética moderna, trabajando con arvejas (*Pisum sativum*) se interesó, entre otras cosas, por la herencia de dos características del tegumento de las semillas: la textura, que podía se lisa o rugosa y el color que podía ser amarillo o verde.*

El monje investigador imaginó que tanto el color como la textura del tegumento se debían a la contribución que hacían los padres, mediante sus "alelos", a la composición de una partícula que regulaba la expresión del carácter: "el gen". En los organismos diploides como las arvejas de Mendel o los humanos, los cromosomas se encuentran apareados, proviniendo un miembro del par de parte del padre y el otro de la madre. Los alelos paterno y materno de un gen se encuentran en los respectivos cromosomas. Mendel idealizaba que si un progenitor era puro, en el sentido de que portaba, por ejemplo, los dos alelos que producían semillas de color amarillo (homocigota para color amarillo) y el otro progenitor era también homocigota pero para el color verde, su cruza (F1) produciría semilla de color amarillo o verde según cuál de los colores fuera el carácter dominante. El esquema siguiente asume que los padres (P) son homocigotas y que el color amarillo es el color dominante. Los individuos portadores de ambos alelos dominantes son identificados como AA y los individuos portadores de los alelos para el verde con aa. El carácter verde es, en este ejemplo, el carácter recesivo.

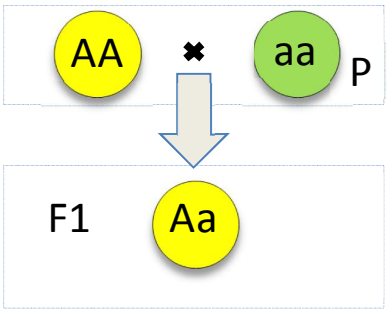


Figura 8.7. Cruzamiento de dos parentales homocigotas dominante y recesivo respectivamente para el color de tegumento

El resultado de cruzar individuos F1 produce la generación F2 como se ilustra en la Figura 8.8. Desde el punto de vista genotípico hay, en promedio, ¼ de genotipos homocigotas dominantes, ¼ de homocigotas recesivos y ½ de heterocigotas.

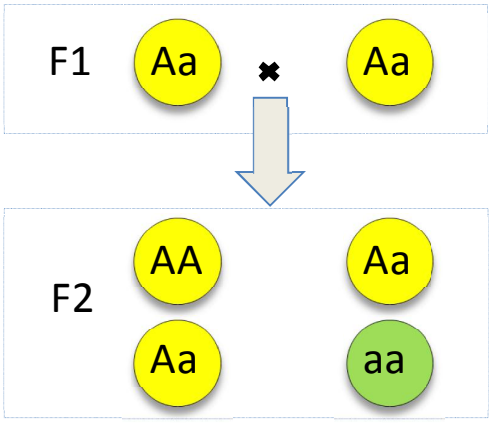


Figura 8.8. Cruzamiento de dos parentales heterocigotas para el color de tegumento

Si cruzamos individuos heterocigotas para dos caracteres como el color de tegumento con alelos A (amarillo dominante) y a (verde) y la textura del tegumento B (lisa dominante) b (rugosa) y ambos caracteres heredan independientemente los resultados teóricos del cruzamiento se presenta en la Figura 8.9. Fenotípicamente se debe esperar que 9/16 semillas sean amarillas lisas, 3/16 amarillas rugosas, 3/16 lisas verdes y 1/16 semillas verdes rugosas.

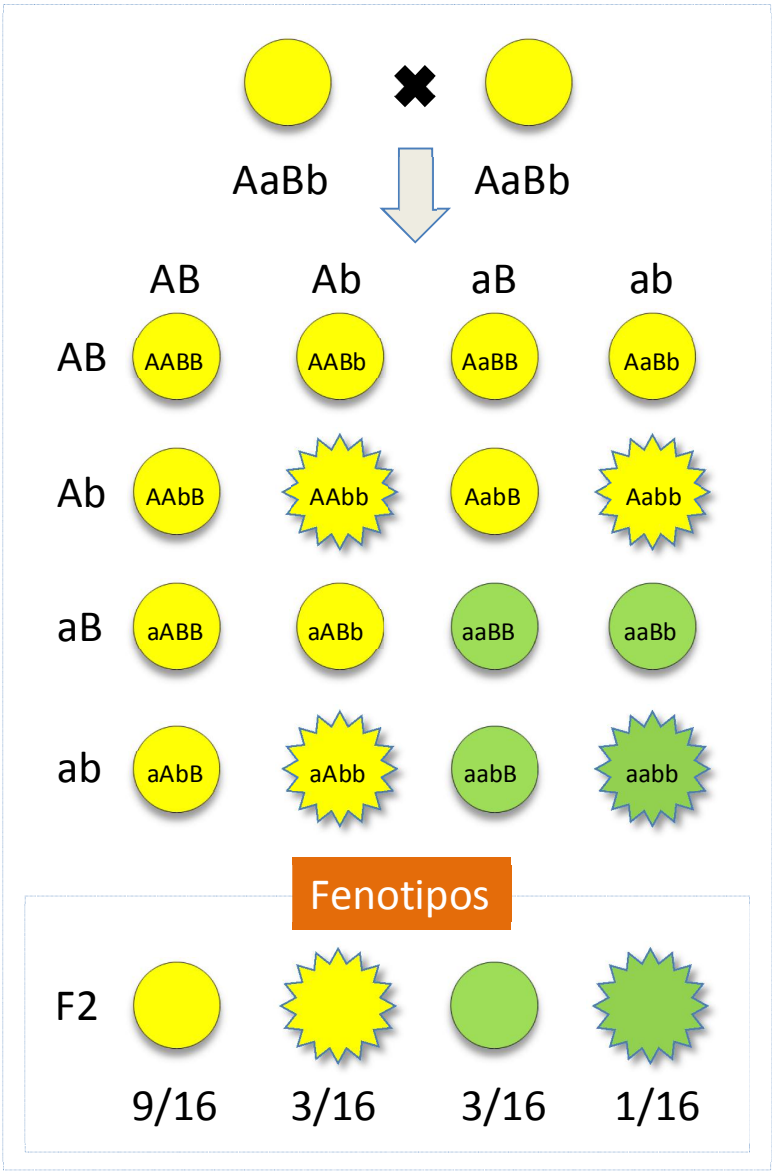


Figura 8.9. Esquema de segregación de dos parentales heterocigotas para el color y textura de tegumento de semillas de arvejas. El tegumento liso y amarillo son las expresiones dominantes.

La Tabla 8.6 muestra el resultado del experimento realizado por Mendel en 1866 sobre este cruzamiento. Las frecuencias presentadas corresponden a la clasificación de 539 semillas de arvejas, según color y textura del tegumento.

La pregunta es sobre la independencia del carácter textura y el carácter color. Éstos se heredan independientemente? Como en el ejemplo anterior tendremos que encontrar las frecuencias esperadas y compararlas con las observadas. La comparación también lo haremos mediante el estadístico chi-cuadrado. A diferencia del ejemplo del fungicida, las frecuencias esperadas se deducen del modelo teórico ilustrado en Figura 8.9 y no son necesarios datos experimentales observados para estimarlas, excepto conocer el total de semillas observadas. También tendremos que encontrar una forma general para el cálculo de los grados de libertad. El estudio de frecuencias observadas respecto de esperadas por un modelo cuyos parámetros no dependen de los datos observados, es lo que se conoce como un análisis de bondad de ajuste.

Tabla 8.6: Tabla es frecuencias de semillas clasificadas según el color (Amarillo o Verde) y textura del tegumento (Lisas, Rugosas) obtenidas del cruzamiento de parentales heterocigotas para ambos caracteres.

		Tegumento		
		L	R	Total
Color	A	301	96	397
	V	112	30	132
	Total	403	126	539

La Tabla 8.7 presenta las frecuencias esperadas para el número de semillas derivadas del modelo de segregación independiente de dos caracteres mendelianos: color y textura del tegumento.

El estadístico chi-cuadrado para este ejemplo será:

$$\chi^2 = \frac{(301-303,2)^2}{303,2} + \frac{(96-101,1)^2}{101,1} + \frac{(112-101,1)^2}{101,1} + \frac{(30-33,7)^2}{33,7} = 1,856731$$

Tabla 8.7: Tabla es frecuencias esperadas según el color (Amarillo o Verde) y textura del tegumento (Lisas, Rugosas) deducidas de un modelo de segregación independiente de dos caracteres mendelianos (color y textura)

		Tegumento	
		L	R
Color	A	539 x 9/16=303,2	539 x 3/16=101,1
	V	539 x 3/16=101,1	539 x 1/16=33,7

Lo que debemos establecer son los grados de libertad de la distribución del estadístico chi-cuadrado cuando la hipótesis nula es cierta. La forma general de calcularlo es por la diferencia de la dimensión del espacio de parámetros para calcular las frecuencias esperadas cuando no se aplican las restricciones impuestas por la hipótesis nula y la dimensión del espacio de parámetros necesarios para estimar las frecuencias esperadas

bajo las restricciones implicadas en la hipótesis nula. En una tabla 2 x 2 hay que rellenar 4 celdas, pero como que el total general de semillas observadas está dado, sólo hay tres celdas que pueden moverse independientemente. Luego la dimensión del espació de parámetros es 3. Por otra parte la hipótesis nula establece que las frecuencias esperadas se obtienen multiplicando el total general por las probabilidades esperadas por el modelo genético. Estas cuatro probabilidades definen un punto en un espacio de dimensión 4. Laa matemáticas nos dice que la dimensión de un punto es cero, de allí que los grados de libertad del chi-cuadrado del experimento de Mendel será 3-0=3. Usando la calculadora de probabilidades y cuantiles de InfoStat, podemos calcular el valor p de 1,856731 como la probabilidad de estar por encima de ese valor en una chi-cuadrado con 3 grados de libertad. El valor p es 0,60267. Para un nivel de significación del 5%, este valor p sugiere que la hipótesis de herencia independiente es consistente con los datos observados.

Aplicación

Color de las flores, espinas y porte de un arbusto

Una planta ornamental puede tener flores Rojas o Blancas, tener porte Arbustivo o rastrero y tener o no Espinas. Cada uno de estos caracteres está regulado por un gen, siendo los caracteres dominantes: flores rojas, porte arbustivo y con espinas (*RAE*). Se cruzaron parentales homocigotos dominantes (RRAAEE) con parentales homocigotas recesivos (rraaee) para obtener la F1 y luego se cruzaron F1xF1. La siguiente tabla contiene los resultados de este último cruzamiento, del que se dispone de 200 plantas. Se quiere saber si los tres caracteres se heredan independientemente.

Tabla 8.8: Tabla es frecuencias fenotípicas observadas según el color de las flores, porte de la planta y presencia de espinas en plantas obtenidas del cruzamiento de heterocitas para los tres caracteres de una planta ornamental.

Flores	Porte	Espinas	Frecuencias fenotípicas observadas en 200 plantas
Rojas	Arbustivo	Si	86
Rojas	Arbustivo	No	28
Rojas	Rastrero	Si	30
Rojas	Rastrero	No	7
Blancas	Arbustivo	Si	26
Blancas	Arbustivo	No	9
Blancas	Rastrero	Si	11
Blancas	Rastrero	No	3

Estrategia de análisis

Para analizar estos datos debemos establecer las frecuencias esperadas bajo la hipótesis de herencia independiente. Una tabla de clasificación con todas las combinaciones genotípicas ayudará a este fin. La primera columna y la primera fila de la siguiente tabla contienen los posibles genotipos de los progenitores. El cuerpo de la tabla contiene una codificación de los fenotipos resultantes.

Tabla 8.9: Tabla es cruzamientos posibles: La primera columna y la primera fila de la siguiente tabla contienen los posibles genotipos de los progenitores. El cuerpo de la tabla contiene una codificación de los fenotipos resultantes.

	RAE	**RAe**	**RaE**	**Rae**	**rAE**	**rAe**	**raE**	**rae**
RAE	*RAE*	*RAE*	*RAE*	*RAE*	*RAE*	*RAE*	*RAE*	*RAE*
RAe	*RAE*	*RAe*	*RAE*	*RAe*	*RAE*	*Rae*	*RAE*	*RAe*
RaE	*RAE*	*RAE*	*RaE*	*RaE*	*RAE*	*RAE*	*RaE*	*RaE*
Rae	*RAE*	*RAe*	*RaE*	*Rae*	*RAE*	*Rae*	*RaE*	*Rae*
rAE	*RAE*	*RAE*	*RAE*	*RAE*	*rAE*	*rAE*	*rAE*	*rAE*
rAe	*RAE*	*RAe*	*RAE*	*RAe*	*rAE*	*rAe*	*rAE*	*rAe*
raE	*RAE*	*RAE*	*RaE*	*RaE*	*rAE*	*rAE*	*raE*	*raE*
rae	*RAE*	*RAe*	*RaE*	*Rae*	*rAE*	*rAe*	*raE*	*rae*

De las 64 celdas de la tabla muchas contribuirán a un único fenotipo. Por ejemplo la fila 1 produce plantas de flores arbustivas con espinas y flores rojas. Si se resumen las frecuencias fenotípicas obtenemos la siguiente tabla de frecuencias relativas esperadas. Éstas resultan de dividir las frecuencias fenotípicas por 64 que es el número total de genotipos posibles.

Tabla 8.10: Tabla es frecuencias fenotípicas observadas y esperadas según el color de las flores, porte de la planta y presencia de espinas en plantas obtenidas del cruzamiento de heterocigotas para los tres caracteres de una planta ornamental.

Flores	**Porte**	**Espinas**	**Frecuencias fenotípicas teóricas**	**Frec. relativas esperadas**	**Frec. esperadas en 200 pts**	**Frec. observadas en 200 pts**
Rojas	Arbustivo	Si	27	27/64	84,38	86
Rojas	Arbustivo	No	9	9/64	28,12	24
Rojas	Rastrero	Si	9	9/64	28,12	30
Rojas	Rastrero	No	3	3/64	9,38	4
Blancas	Arbustivo	Si	9	9/64	28,12	26
Blancas	Arbustivo	No	3	3/64	9,38	9
Blancas	Rastrero	Si	3	3/64	9,38	14
Blancas	Rastrero	No	1	1/64	3,12	0

Una vez que se dispone de las frecuencias esperadas podemos compararlas con las frecuencias observadas mediante el estadístico chi-cuadrado. Los grados de libertad de esta prueba son 7-0=7.

Para realizar esta prueba con InfoStat, seleccionaremos del menú Estadísticas, el ítem *Inferencia basada en una muestra*, sub-ítem *Prueba de bondad de ajuste (multinomial),* como se muestra en la Figura 8.10. Al invocar este procedimiento se abre una ventana

específica para la carga de las frecuencias observadas y ya sean las proporciones o las frecuencias esperadas como se muestra en la Figura 8.11. En esta ventana al accionar el botón aceptar, aparece el valor del estadístico chi-cuadrado, sus grados de libertad y el *valor p*. Como podrá observarse, existe un dispositivo para cuando hay que corregir los grados de libertad. Por defecto la corrección es cero.

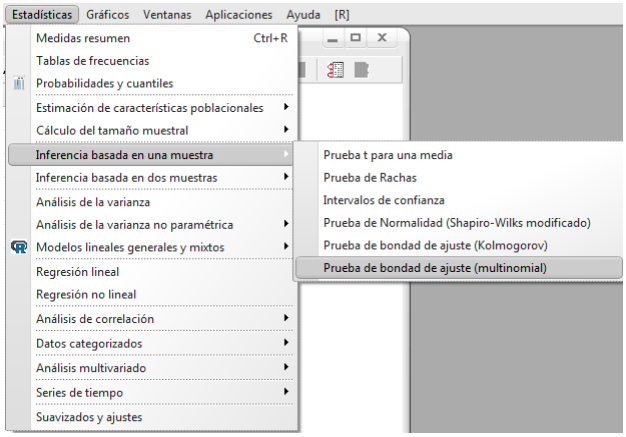


Figura 8.10. InfoStat. Secuencia de ítems de menú para realizar un contraste de hipótesis para bondad de ajuste.

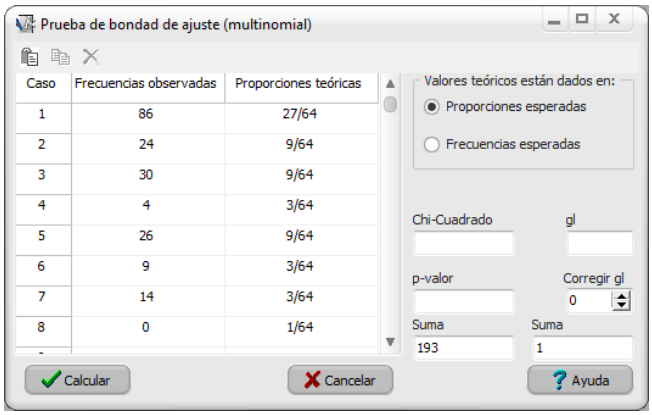


Figura 8.11. InfoStat. Ventana de diálogo para la carga de frecuencias observas y frecuencias o proporciones esperadas.

Conclusión

No se puede rechazar la hipótesis que sostiene que los caracteres color de flor, presencia de espinas y porte son caracteres que "segregan" independientemente.

Ejercicios

Ejercicio 8.1: Para establecer que sistema de monitoreo de insectos es más efectivo se realizó un estudio donde el número total de un insecto plaga fue estimado en 20 parcelas de ¼ de hectárea que cubrían desde bajas al bajas a altas densidades poblacionales. Las parcelas estaba sembradas 60000 plantas por hectárea. Se tomó una muestra sistemática de 300 pantas por parcela y se contó el número total de los insectos de interés. El número total de plantas evaluadas fue de 6000 plantas. Este es un esfuerzo de muestreo impráctico para monitoreo rutinario. Al mismo tiempo se utilizaron 2 métodos de monitoreo: a) Recorrer la parcela en forma de W. El recorrido total es de 103 m aproximadamente y tomando una planta por cada 4 metros produce una muestra de aproximadamente 25 plantas. b) Usar 10 trampas para captura de insectos por parcela ubicadas equidistantemente dentro de la parcela. Los resultados se encuentran en el archivo [Densidadesdeinsectos]. El archivo contiene 3 columnas: Sistemático 300p, Muestreo W y Trampas. Los datos que se consignan es esta tabla son el promedio de insectos por planta en los dos primeros casos y el promedio de insectos por trampa en el tercero.

a) Esquematice, mediante matrices de diagramas de dispersión, las relaciones entre estas determinaciones de densidad.
b) ¿Qué coeficiente de asociación entre variables cuantitativas utilizaría en este caso?, ¿porqué?
c) ¿Es la medida de asociación escogida, entre el muestreo sistemático y los dos métodos de monitoreo significativas?
d) ¿Cuál de los dos sistemas propuestos para monitoreo correlaciona mejor con la densidad estimada por el muestreo sistemático?

Ejercicio 8.2: En un estudio se hicieron mediciones de perímetro y peso de cabezas de ajo. Los datos que se obtuvieron fueron los siguientes:

Perímetro (cm)	12.39	12.39	12.71	9.8	12.3	10.12	11.81	11.41	9.4	11.49
Peso (grs.)	32.27	29.39	30.8	15.6	29.8	16.87	28.11	23.29	14.11	25.37

a) ¿Cómo se espera que sea la correlación entre peso y perímetro? ¿Positiva? ¿negativa?, ¿sin correlación?
b) Calcular el coeficiente correlación de Pearson entre peso y perímetro
c) ¿Es significativo el coeficiente encontrado?

Ejercicio 8.3: Si quiere establecer si ¿el uso de suplementos en las raciones de vacas aumenta éxito de la inseminación? Los datos que se presentan a continuación son un resumen del archivo [Suplementos].

Suplemento	No preñada	Preñada	Total
NO	31	219	250
SI	13	237	250
Total	44	456	500

a) Establecer si hay asociación o no con el uso de suplementos alimentarios y la obtención de una preñez

b) En caso afirmativo calcular la razón de chances.

Ejercicio 8.4: La siguiente tabla contiene la distribución de 18223 hogares argentinos clasificados según régimen de tenencia de la vivienda y región [datos EPH2007]. Estos datos son un extracto de la Encuesta Permanente de Hogares, realizada por INDEC en 2007.

	Régimen de tenencia de la vivienda			
REGION	Propietario	Inquilino	Otro	Total
Cuyo	1138	244	225	1607
Buenos Aires	2095	434	322	2851
NEA	1557	263	190	2010
NOA	2446	395	429	3270
Pampeana	4164	1155	633	5952
Patagonia	1685	571	277	2533
Total	13085	3062	2076	18223

a) ¿Existe asociación estadísticamente significativa entre el régimen de tenencia de la vivienda y la región del país que se considere?

b) ¿Hay alguna región donde la propiedad de la vivienda sea más prevalente que en otras regiones?

Ejercicio 8.5: Se quiere corroboran si las siguientes frecuencias fenotípicas de una planta ornamental se corresponden las proporciones fenotípicas 9:3:3:1, utilizando un nivel de significación del 5%.

Fenotipos	Frecuencias Observadas
Hojas verdosas, bordes lisos	926
Hojas verdosas, brotes dentados	288
Hojas rojizas, sin lisos	293
Hojas rojizas, con dentados	104

Capítulo 9

Diseño y análisis de experimentos a un criterio de clasificación

Carlos Walter Robledo

9. Diseño y análisis de experimentos a un criterio de clasificación

Motivación

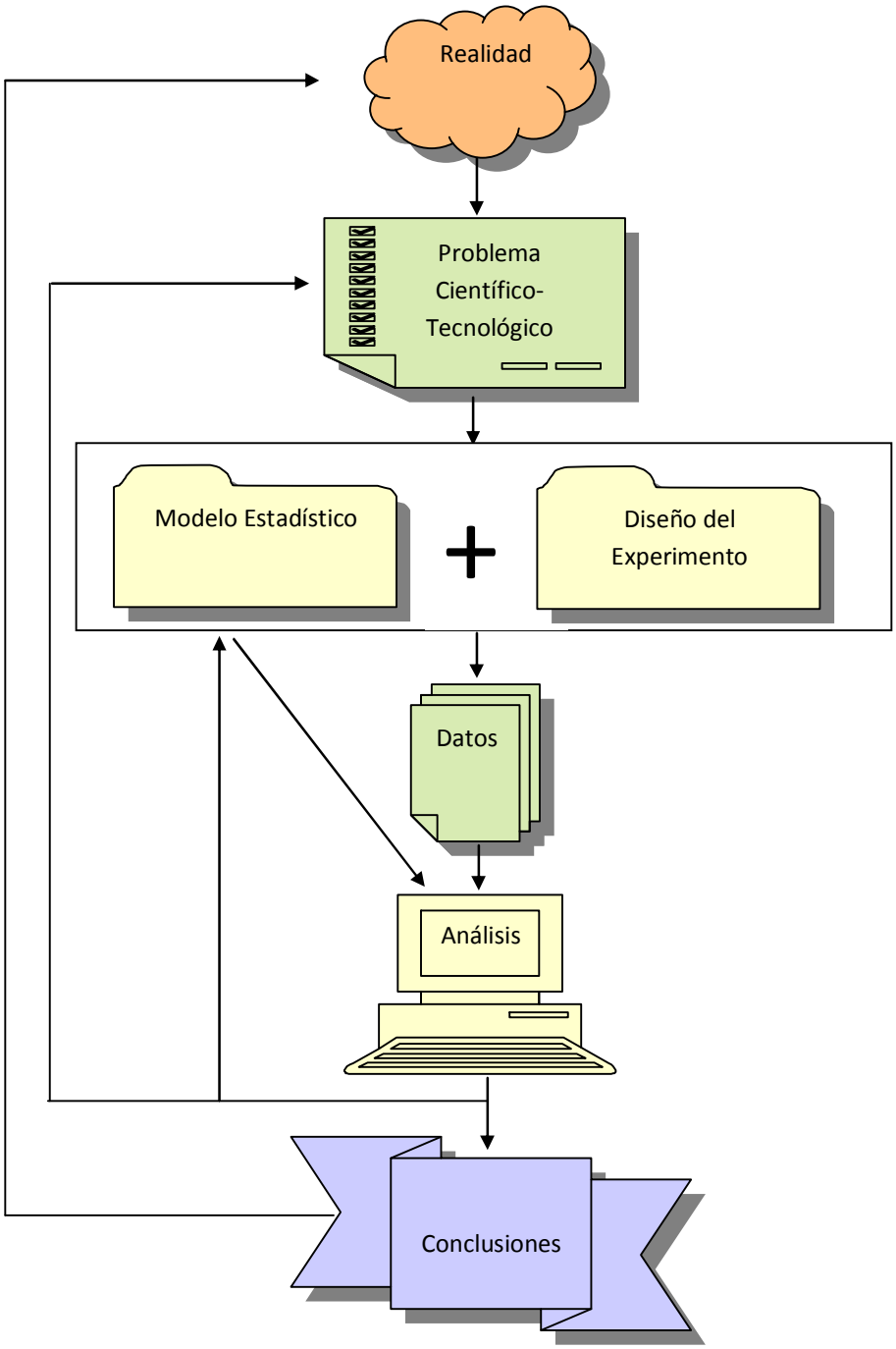
En las Ciencias Agronómicas es frecuente conducir ensayos con fines de evaluar comparativamente dos o más poblaciones, identificadas por algún criterio que las distinga o separe como es la aplicación de distintos tratamientos (criterio de clasificación). Para analizar estos experimentos es común recurrir a la técnica del Análisis de la Varianza (ANAVA). Más formalmente, el ANAVA es un método estadístico cuya finalidad es contrastar hipótesis referidas a las medias dos o más poblaciones, generalmente definidas por la asignación de dos o más tratamientos a un conjunto de unidades experimentales. En este capítulo se introducen dos temáticas relacionadas: (a) la generación de datos experimentales, siguiendo conceptos básicos del diseño de experimentos y (b) técnicas de análisis de datos en experimentos comparativos utilizando la técnica estadística del ANAVA.

Conceptos teóricos y procedimientos

Un primera idea a considerar, es que el problema que se esté estudiando, a fin de elaborar conclusiones que permitan luego realizar recomendaciones de tipo tecnológico-productivas, es el que permite identificar qué metodología/s estadística/s debiera/n utilizarse (ya que es factible usar uno o más en la práctica) y de qué forma se debieran conducir los ensayos con la finalidad de registrar datos que posibiliten realizar las evaluaciones o comparaciones que fueren de interés en el problema bajo estudio (contrastes de hipótesis). El esquema siguiente representa esta idea.

En el esquema se expone que los problemas a investigar surgen de la realidad, es decir de la capacidad del investigador de observar y percibir las necesidades de investigación que plantea el medio. Para abordar ese problema puede diseñarse un estudio experimental donde se busca comparar y analizar diferencias entre distintos trataientos o condiciones experimentales para inferir sobre posibles **efectos de tratamientos**. El diseño de experimentos y el análisis de los datos relevados en el experimento son de crucial importancia para garantizar cierta confiabilidad en las conclusiones que se deriven del estudio.

El esquema también representa un hecho que tiene que ver con este "motor" de investigación que es el análisis estadístico. Así, el análisis estadístico surge como una herramienta para generar conocimiento a partir de los datos. El análisis de los datos de un experimento particular permite sugerir modificaciones a modo de *feed-back* o retroalimentación del sistema para generar nuevos datos e incluso para modificar el modelo estadístico adoptado para analizar los datos. El análisis estadístico también permite enriquecer la identificación y caracterización del problema científico-tecnológico y así reformular las hipótesis que se desean evaluar.


Realidad
Problema
Científico-
Tecnológico
Modelo Estadístico
+
Diseño del
Experimento
Datos
Análisis
Conclusiones

Criterios de clasificación e hipótesis del ANAVA

Supongamos que se desea evaluar si un conjunto (dos o más) de medias poblacionales son iguales y en caso que no lo sean, identificar cuál o cuáles son diferentes y cuál o cuáles no lo son, desde un punto de vista estadístico y a partir de la información muestral o experimental que se tiene sobre esas poblaciones.
El problema puede formularse en términos de una hipótesis nula y una alternativa, las que se escriben de la siguiente forma:

H_0 : $\mu_1 = \cdots = \mu_a$

H_1 : Al menos una de las a medias poblacionales es distinta

donde *a* representa la cantidad de medias poblacionales a comparar. Estas *a* poblaciones que están involucradas en el estudio, deben distinguirse o estar separadas en base a algún criterio que el investigador establezca claramente. De esta manera, si se encuentran diferencias entre los valores esperados de todas o de al menos un par de ellas, se podrá inferir sobre la causa de los efectos que generan las diferencias.
A modo de ejemplo de estos criterios de clasificación de datos podemos citar el factor "variedades". Supongamos que se tiene un ensayo comparativo de rendimientos, donde se registran datos de rendimiento de grano para varias parcelas y que estas paracelas han sido sembradas con distintas variedades. En este caso las *a* poblaciones a evaluar serian las correspondientes a datos de rendimiento de las *a* variedades, μ_1 podría representar el rendimiento medio poblacional de parcelas donde se siembra la variedad que se identifique con el número 1. Es decir que $\mu_1 = E(Y_1)$, esto es la esperanza de la variable aleatoria Y_1 (el rendimiento de la variedad 1), μ_2 es la media poblacional de la variable aleatoria Y_2 (rendimiento de la variedad 2) y así sucesivamente.
Otro ejemplo de criterio de clasificación podría ser la dosis de fertilizante que se usa para lograr un cultivo. Si un técnico estuviera interesado en evaluar comparativamente los rendimientos medios de un híbrido cuando no se lo fertiliza respecto a fertilizar con 100, 200, 300 o 400 kg/ha de urea como fuente de nitrogeno, se podría diseñar un experimento con cinco poblaciones (*a*= 5) a evaluar, una correspondiente a un tratamiento control o no fertilizado y otras respondiendo a las cuatro dosis distintas de fertilización que se pretenden evaluar. Si en el diseño del estudio experimental fijamos o controlamos la mayoría de los factores que pueden impactar la respuesta, al observar diferencias entre poblaciones, éstas podrán ser asignadas con mayor confianza a los tratamientos. Por ejemplo, la diferencia entre la media de la población de rendimientos sin fertilizar y la media de la población de rendimientos con 200 kg/ha de urea permite inferir sobre el efecto de fertilizar con 200 kg/ha del producto.

El proceso generador de datos (PGD)

El origen de los datos necesarios para probar la hipótesis estadística de igualdad de *a* medias poblacionales puede ser observacional o experimental.
En las Ciencias Sociales, como por ejemplo en las Ciencias Económicas, no es factible realizar experimentos –sea por cuestiones básicamente prácticas o por cuestiones éticas, pero sí es posible observar y registrar o tomar datos directamente de la realidad, sin modificaciones o manipulaciones introducidas por el investigador o técnico en la génesis o proceso que da origen a los datos. Cuando el estudio es de esta naturaleza, igualmente puede ser de interés realizar comparaciones de las observaciones realizadas bajo distintas condiciones con ANAVA. No obstante es más difícil, cuando no imposible, concluir sobre relaciones causales ya que factores no controlados que actúan en la realidad pueden enmascarar las diferencias entre condiciones debidas al factor de clasificación considerado como factor "tratamiento".
En otras ciencias sí es factible conducir experimentos. En estos casos, es posible generar datos experimentalmente, bajo condiciones controladas por el investigador, por lo que en numerosos casos se sustituye la palabra *población* por la de *tratamiento* y se realizan conclusiones del tipo causa-efecto.

En los estudios observacionales como experimentales hay un denominador común conceptual que ayuda a explicar el origen de los datos desde un punto de vista estadístico y que genéricamente llamaremos proceso generador de los datos (PGD). En Estadística existen muchos modelos que han sido propuestos como PGD, uno de los más usados es el modelo lineal aditivo.

El *modelo lineal* que se puede utilizar para contrastar la hipótesis de igualdad de *a* medias poblacionales se puede escribir de la siguiente forma:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{con} \quad i = 1,...,a \quad \text{y} \quad j = 1,...,n$$

donde:

- a denota la cantidad de poblaciones o tratamientos en evaluación
- n indica la cantidad de unidades experimentales que se evaluarán de cada población o tratamiento
- Y_{ij} *es la j-ésima* observación de la i-ésima población o tratamiento
- μ es la media general
- τ_i *es el efecto de l*a i-ésima población o tratamiento
- ε_{ij} es una variable aleatoria normal independientemente distribuida con esperanza 0 y varianza σ^2 $\forall i,j$.

Este modelo lineal nos ayuda a explicar que cada magnitud que registramos como dato en nuestro estudio proviene la suma de la acción de varios componentes: una cantidad fija desconocida, denotada por μ , más una componente τ_i, también desconocida, y que es usada para explicar cómo cambia la observación Y_{ij} debido al hecho de pertenecer a la población o tratamiento *i*, más un término aleatorio ε_{ij} (componente aleatoria sobre la cual el investigador no tiene control) que ayuda a explicar la variabilidad "natural o propia" que existe entre dato y dato dentro de una misma población o tratamiento. Si dos unidades de análisis son tratadas de igual manera, es decir pertenecen a la misma población, sería de esperar que su respuesta (el dato recolectado desde la unidad) sea el mismo. No obstante, en la práctica se observan diferencias entre las respuestas de unidades experimentales tratadas de igual manera. La variabilidad de las respuestas de unidades experimentales tratadas con el mismo tratamiento o pertenecientes a la misma población es la cantidad que en el modelo se denota por σ^2 y se conoce como **variabilidad residual**.

Una representación gráfica del modelo lineal presentado es la siguiente:

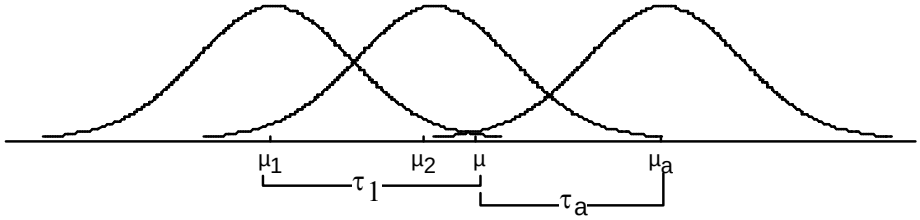


Figura 9.1: Representación del modelo lineal del ANAVA

En la Figura anterior se grafican las funciones de densidad normales de la variable aleatoria de interés bajo cada población, mostrando el punto de equilibrio de todas ellas (μ), las esperanzas de cada una de ellas (μ_i) y los corrimientos de las esperanzas respecto del punto de equilibrio representando los efectos de tratamiento (τ_i). Consideremos nuevamente el ejemplo relacionado a evaluar comparativamente los rendimientos medios de un cierto híbrido cuando no se lo fertiliza respecto a situaciones donde se fertiliza con 100, 200, 300 o 400 Kg/ha de urea. El modelo lineal nos ayudaría a explicar como se produjo el rendimiento de, por ejemplo, la parcela *j*, o unidad experimental *j*, donde se cultiva experimentalmente el híbrido con 100 Kg/ha del fertilizante. Este valor de rendimiento es representado simbólicamente como $Y_{100,j}$ y según el modelo esta cantidad es producida por la suma de tres componentes. La primera, es una cantidad fija desconocida μ que representa el valor esperado del rendimiento del híbrido independientemente del tratamiento que reciba, este valor se estima con la media general de todos los renimientos, se supone que el rendimiento que estamos tratando de explicar tendrá que asumir un valor cercano a esa media general. La segunda, es el efecto τ_{100} que representa el cambio en el rendimiento (que puede ser positivo o negativo) por el hecho de haber utilizado 100 Kg/ha en esa parcela. La tercera es la cantidad $\varepsilon_{100,j}$ también desconocida por el investigador que es debida

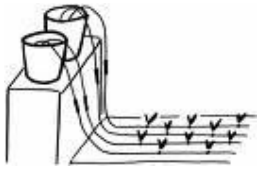
exclusivamente a las condiciones y característica propias de la parcela que utilizó y a condiciones no controladas como podrían ser condiciones climáticas, presencia/ausencia de plagas, malezas, que se presentaron en la parcela durante el cultivo de la misma pero para las cuales no hay suficiente información como para tratarlas separadamente.

¿para qué nos sirve un modelo en el que cada uno de los tres términos que lo componen son todos desconocidos? Es posible calcular "aproximaciones" a los verdaderos valores de las componentes no aleatorias y a la varianza de la componente aleatoria?

Obtenida las aproximaciones, es decir habiendo estimado los parámetros del modelo, podremos obtener un **valor predicho** por el modelo para cada una de las unidades de análisis. La diferencia entre el valor observado de la variable en una unidad de análisis y el valor predicho por el modelo para esa misma unidad se denomina **residuo** y es un predictor del término de error aleatorio.

Conceptos del diseño de experimentos

El primero, es el concepto de **unidad experimental (UE)**, que hace referencia a la mínima unidad de análisis sobre la que se realizará una medición.



En las Ciencias Agropecuarias se suele usar el término "parcela experimental" para referirise a la unidad experimental ya que comúnmente se trabaja efectivamente con parcelas de tierra como unidad experimental. No obstante, las UE también podrían ser macetas, árboles, animales, ratones de laboratorio, …

En los estudios experimentales la UE se define como la mínima porción del material experimental sobre el cual un tratamiento puede ser realizado o aplicado. Para un buen diseño siempre es conveniente tener repeticiones de UE, es decir un número mayor que uno de UE que reciben un tratamiento particular.

El concepto de **tratamiento** se refiere a la acción o acciones que se aplican a las unidades experimentales con la finalidad de observar cómo responden y así "simular experimentalmente bajo condiciones controladas" las poblacionales que interesam comparar.

En estudios observacionales, las UE a veces son llamadas unidades observacionales. Por ejemplo, en estudios socio-económicos podemos citar como unidades observacionales de un estudio comparativos a las empresas, las personas o los productores.

La importancia de pensar en las UE antes de realizar el estudio, es decir durante la etapa de diseño del mismo, radica en la necesidad de reconocer cualquier estructura (no aleatoria) que éstas pudieran tener.

El reconocimiento a priori de la heterogeneidad que exista entre ellas previo a la asignación de tratamientos o a su clasificación es importante para diseñar el estudio. Si es posible elegir las unidades necesarias para conducir el estudio de forma tal que sean lo más similares posibles entre sí (concepto de homogeneidad de unidades experimentales) diremos que el diseño de experimento más conveniente desde un punto de vista estadístico es el conocido como **diseño completamente aleatorizado** (**DCA**). Aquí, ya que no se distingue ninguna estructura de UE, los tratamientos serán aplicados a las mismas de forma totalmente aleatoria, es decir cualquier UE puede recibir cualquier tratamiento. Mientras que, si no es posible disponer de UE homogéneas, pero es posible agruparlas de forma tal que cada grupo de unidades sea internamente homogéneo, y dentro de cada grupo hay suficientes UE como para comparar los tratamientos diremos que un diseño recomendado desde el punto de vista estadístico es el conocido como **diseño en bloques aleatorizados***,* aquí los tratamientos son aleatorizados dentro de cada bloque de UE.
La asignación de los tratamientos a las unidades experimentales, y su conducción a lo largo del estudio, puede contribuir a que uno de los supuestos importantes en el modelo lineal, el supuesto de independencia, se cumpla.

La elección aleatoria de las unidades de observación y la asignación aleatoria de tratamientos a las unidades experimentales son mecanismos recomendados para evitar falta de independencia.

En la experimentación agronómica a campo, también se toman otros cuidados para evitar la presencia de datos correlacionados experimentalmente. Por ejemplo, para que el rendimiento de una parcela sea independiente del rendimiento de las parcelas vecinas, se puede recurrir a distintas variantes como dejar espacio suficiente entre una parcela y otra. Otra variante es no dejar espacios libres, con el fin de simular mejor las condiciones reales de cultivo, y luego evaluar sólo el sector central de cada parcela. La superficie de la parcela que no producirá datos para el análisis se suele denominar **bordura**.
La **aleatorización** es otro concepto fundamental del diseño de experimentos**,** que centra su atención en minimizar efectos sistemáticos. En un diseño experimental, la aleatorización hace referencia al proceso mediante el que se asigna cual tratamiento recibirá cada una de las unidades experimentales.
En un DCA un mecanismo de aleatorización puede ser el siguiente: por ejemplo se pueden elegir al azar, desde las $a \times n$ unidades experimentales disponibles*,* un grupo de *n* unidades experimentales y luego se elige al azar un tratamiento para asignar a esas unidades. Debe registrarse claramente que tratamiento recibió cada unidad experimental para que cuando se registre la medición de la variable aleatoria que se esté estudiando, se asocie el dato a la unidad experimental.

Existen distintas técnicas de asignación al azar de los tratamientos a las unidades experimentales. Se pueden colocar papelitos con números para representar a cada unidad experimental, mezclarlos en una bolsa, y luego sacar un papelito que identificará una unidad experimental, papelito que no será repuesto en la bolsa. Luego, desde otra bolsa con papelitos que identifican a cada tratamiento, sacar un papel identificatorio del tratamiento que recibirá la unidad experimental recién elegida, papelito que si será repuesto en la bolsa. Así se establece la asociación "unidad experimental–tratamiento que recibirá" de forma aleatoria, procedimiento que se repetirá para cada una de las unidades experimentales.

Finalmente, otro concepto fundamental del diseño de experimentos es el de **repetición.** Cada una de las *n* unidades experimentales que reciben un mismo tratamiento y que permiten generar *n* datos independientes ofician de repetición.

Tabla 9.1: Estructura de una tabla de datos de un experimento unifactorial o a una vía de clasificación

Tratamientos					Media	Varianza
1	y_{11}	y_{12}	...	y_{1n}	$\overline{y}_1$	S_1^2
2	y_{21}	y_{22}	...	y_{2n}	$\overline{y}_2$	S_2^2
:	:	:	:::	:	:	:
a	y_{a1}	y_{a2}	...	y_{an}	$\overline{y}_a$	S_a^2

Las repeticiones juegan un rol importante ya que permiten evaluar la variabilidad de los datos registrados dentro de cada tratamiento. Esta variabilidad se estima por medio de la varianza muestral de las repeticiones. A la varianza muestral como medida de dispersión la denotamos como S^2. Ahora, como tenemos varias poblaciones a la notación de la varianza muestral le agregamos como subíndice la letra *i* , según lo hemos introducido en el modelo lineal, para distinguir las varianzas muestrales de las muestras correspondientes a distintas poblaciones o tratamientos que estamos interesados en evaluar, esto es S_i^2.

Bajo el supuesto de que los $a \times n$ términos de error aleatorio del modelo lineal tienen todos la misma varianza σ^2 (supuesto de varianza constante u homogeniedad de varianzas), cada una de las *a* varianzas muestrales S_i^2 nos ofrecen buenos estimadores del parámetro poblacional σ^2. Este supuesto de varianzas homogéneas nos habilita a promediar las S_i^2 para obtener un estimador de σ^2. El promedio de las *a* varianzas muestrales S_i^2 es un nuevo estadístico que recibe el nombre de **cuadrado medio dentro** o **cuadrado medio del error experimental** (**CME**).

EL CME representa una medida de la variabilidad dentro de los tratamientos, o dicho de otra manera, entre las repeticiones. Si es bajo, relativo a otras medidas de variabilidad en el estudio, implica que la varibilida experimental es baja, esto es que las respuestas de unidades experimentales que recibieron el mismo tratamiento varía relativamente poco (como es de esperar en estudios bien diseñados).

Análisis de la varianza de un DCA

El ANAVA para contrastar la hipótesis de igualdad de medias poblacionalas entre los distintos tratamientos, respecto a la hipótesis de que al menos un par de tratamientos difiere estadísticamente, se basa en la comparación de dos "varianzas muestrales", una es la varianza dentro de tratamientos o CME y otra es la varianza entre tratamientos o entre medias de tratamientos. Esta comparación de dos varianzas se realiza por medio de la **prueba F** basada en el estadístico F igual al cociente de dos varianzas. Por ello la técnica se denomina Análisis de Varianza (ANAVA).

La primera varianza, introducida en la sección anterior, es denominada **cuadrado medio dentro** (**CMD**) o **cuadrado medio del error** y representa la variabilidad observada de unidad a unidad que reciben el mismo tratamiento y no asignable a ninguna causa particular; es la denominada varianza debida al **error experimental**. El cuadrado medio dentro, como toda varianza puede ser escrito también como el cociente de una suma de cuadrados y sus grados de libertad, que en este diseño con N= *axn* unidades experimentales, son N-a. El CMD es un estimador de la varianza residual, es decir de la variabilidad entre observaciones que no se debe a las fuentes de variación que se reconocen a priori; en el DCA sería equivalente a la variabilidad entre observaciones que no tiene que ver con diferencias entre tratamientos, sino con diferencias observadas dentro de los tratamientos.

La segunda varianza muestral que forma parte del estadístico F, surge de la idea de que es posible plantear otro estimador de σ^2. Bajo normalidad, si la hipótesis nula de igualdad de medias y las suposiciones de homogeneidad de varianzas fuesen verdaderas, las *a* poblacionales serían iguales. Las *a* medias muestrales que se pueden calcular con los *n* datos de cada tratamiento, tienen varianza $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ según lo observado en el estudio de distribuciones en el muestreo.

Bajo el supuesto de homogeneidad de varianzas y de igualdad de medias poblacionales, entonces se puede obtener un segundo buen estimador de σ^2 si pensamos a $\hat{\sigma}^2 = n \times S^2_{\bar{X}}$. El nuevo estimador es conocido como **cuadrado medio entre**

tratamientos, cuadrado medio tratamientos o simplemente como cuadrado medio entre (**CME**).

El cuadrado medio, como toda varianza, puede ser escrito también como el cociente de una suma de cuadrados y sus grados de libertad, que en este diseño con a *tratamientos es* a-1*.*

Bajo la hipótesis nula, es decir cuando no hay diferencias significativas entre las medias de los tratamientos, $S_{\bar{X}}^2$ tenderá a ser baja, ya que las medias muestrales de los *a* tratamientos serán muy parecidas.

En el caso que la hipótesis nula de igualdad de medias poblacionales no fuera verdadera, ocurrirá que $S_{\bar{X}}^2$ tenderá a crecer a medida que las *a* medias poblacionales sean cada vez más distintas.

Si denotamos como σ_E^2 y σ_D^2 a las varianzas estimadas respectivamente por el CME y el CMD, luego bajo la hipótesis nula de igualdad de medias poblacionales o de tratamiento, ocurrirá que $\sigma_E^2 = \sigma_D^2$, en caso contrario (hipótesis nula falsa) ocurrirá que $\sigma_E^2 > \sigma_D^2$, por lo que podemos reescribir las hipótesis clásicas del ANAVA (referidas a medias poblacionales o esperanzas) como la siguiente hipótesis unilaterial que compara dos varianzas poblacionales:

$$H_0 : \sigma_E^2 = \sigma_D^2 \qquad vs \qquad H_1 : \sigma_E^2 > \sigma_D^2$$

La prueba del ANAVA consiste en calcular el estadístico F utilizando los estimadores de σ_E^2 y σ_D^2 (es decir los cuadrados medios) de la siguiente forma:

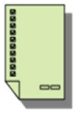
$$F = \frac{CME}{CMD}$$

Este estadístico tiene, bajo H_0, una distribución $F_{(a-1),(N-a))}$ con N igual al número total de unidades experimentales.

Luego, para un nivel de significación α, si F es mayor que el cuantil $(1-\alpha)$ de la distribución $F_{(a-1),(N-a)}$ se rechaza H_0, implicando que H_1 es verdadera. El rechazo de H_0 implica que las medias poblacionales (expresadas como a media poblacional más un efecto de tratamiento o población) no son iguales y por lo tanto, que algún $\tau_i \neq 0$; así se concluye que no todas las medias de tratamiento son iguales.

El ANAVA se basa en dos estimadores independientes de la varianza común del conjunto de tratamientos: uno basado en la variabilidad dentro de los tratamientos, y otro basado en la variabilidad entre los tratamientos. Si no hay diferencias entre las medias de los tratamientos, estos dos estimadores estiman al mismo parámetro, de lo contrario el segundo tiende a ser mayor cuanto mayor es la diferencia entre medias de tratamientos.

Luego, a pesar de que la hipótesis de interés del ANAVA se refiera a la igualdad de las esperanzas de dos o más distribuciones, la técnica del ANAVA se basa en la comparación de varianzas para inferir acerca de la igualdad de las esperanzas.
El análisis de la varianza se suele resumir en una tabla conocida como Tabla de Análisis de la Varianza en la que se resumen los estadísticos y cálculos básicos para obtener el CME y el CMD, estadísticos claves para la prueba de hipótesis. En la columna titulada "**Fuentes de Variación**" se destacan tres celdas con sus correspondientes títulos. En ellas se indican los contenidos de las celdas dentro de la fila respectiva. En la fila titulada "Entre Tratamientos" existen cuatro celdas, en las que se presentan las siguientes cantidades: **Suma de Cuadrados Entre Tratamientos** (SCE), **Grados de Libertad** de la suma de cuadrados entre tratamientos (gle), **Cuadrados Medios** Entre Tratamientos (CME) y el estadístico F correspondiente al cociente del CME/CMD. La fila titulada "Dentro (**Error Experimental**)" se completa con las siguientes cantidades: Suma de Cuadrados Dentro de Tratamientos (SCD), Grados de Libertad de la suma de cuadrados dentro de tratamientos (gld) y Cuadrado Medio Dentro de Tratamientos(CMD). En la titulada "Total" se completa con la Suma de Cuadrados Total (SCT) y Grados de Libertad Total (glt).



Esta presentación tan tradicional de las salidas de un ANAVA, permite ordenar los cálculos cuando estos se realizan sin un software estadístico. No obstante, el valor más importante de la salida del ANAVA cuando éste se realiza con software es el valor p asociado al estadístico F.

Como en otras pruebas estadísticas, el **valor p** de la prueba se compara con el nivel de significación fijado y si el valor p es menor que α, se concluye rechazando la hipótesis nula. En una ANAVA siempre que el valor F sea grande, se pone en evidencia que las diferencias entre tratamientos son mayores a las diferencias observadas dentro de tratamientos es decir a aquellas que podrían darse por azar o por la variabilidad natural de la respuesta.Consecuentemente valores altos de F se asocian con valores p bajos y llevan al rechazo de la hipótesis de igualdad de medias de tratamientos.
Cuando el ANAVA se realiza con InfoStat, se obtiene además de la Suma de Cuadrados Total y las Sumas de Cuadrados de cada componente, una Suma de Cuadrados del Modelo. Esta última es proporcional a la variabilidad en la respuesta explicada por el

modelo lineal completo que se propone. El cociente entre la Suma de Cuadrados del Modelo y la Suma de Cuadrados Total, se denomina **coeficiente de determinación** o R^2. Este coeficiente, al ser una proporción, verifica que $0 \leq R^2 \leq 1$, siendo deseable valores superiores, digamos que en la práctica, a 0.60 y mientras mayores, mejor. El coeficiente de determinación suele expresarse en porcentaje y se interpreta como el porcentaje de la variabilidad total en Y que es explicada o contabilizada en el modelo de ANAVA propuesto. El complemento a 100% es una medida de la variabilidad no explicada por el modelo.

Aplicación

Ensayo comparativo de rendimiento

Para comparar los rendimientos medios de 4 cultivares híbridos de un cultivo (tratamientos) en un ambiente, se realiza un experimento bajo un diseño a campo con 10 repeticiones o parcelas por tratamiento. Cada parcela tiene una superficie total de 5 surcos por 25 metros de largo cada uno. No obstante, la parcela útil es de 3 surcos por 15 metros cada uno. El resto es considerado bordura y no se registran los pesos de cosecha en esa porción de la parcela. Los resultados se encuentran en el archivo [Híbridos]. Los datos de rendimientos parcelarios se registran en qq/ha a humedad constante (14% de humedad).

Estrategia de análisis

En primer lugar, planteamos la hipótesis estadística a contrastar:

H_0: $\mu_1 = \cdots = \mu_4$

H_1: Al menos uno de las 4 cultivares tiene media poblacional distinta a las demás

En segundo lugar, asumimos un modelo lineal para un diseño completamente aleatorizado a un criterio de clasificación. Esto es, suponemos que las unidades experimentales pudieron ser elegidas de forma tal que son homogéneas en suelo, pendiente, humedad, topografía, sombreados y otros factores que podrían impactar los rendimientos y que las variedades se asignaron aleatoriamente a las unidades experimentales. Cada rendimiento observado en el experimento se puede explicar de la siguiente manera:

Análisis de experimentos a un criterio de clasificación

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

donde:

i=1,...,$a = 4$ variedades,

j=1,..,$n = 10$ repeticiones

Y_{ij}	*representa el rendimiento de la j-ésima parcela del i-ésimo cultivar*
μ	*representa la media general de los rendimientos*
τ_i	*es el efecto sobre el rendimiento del i-ésimo cultivar*
ε_{ij}	*es una variable aleatoria normal independientemente distribuida con esperanza 0 y varianza* σ^2 $\forall i,j$

Luego, podremos proceder a conducir el ANAVA para probar la hipótesis planteada. Para ello, abrir el archivo *[Híbridos]* de InfoStat. Luego en el menú *Estadísticas* seleccionar el submenú *Análisis de la Varianza*. Seleccionar *Cultivar* en el panel izquierdo de la ventana y "agregarlo" al panel *Variables de clasificación*. De la misma forma seleccionar *Rend.* y agregrarlo al panel *Variables dependientes.* La imagen de la ventana resultante se muestra a la derecha de la Figura 9.2

Figura 9.2: InfoStat. Diálogo inicial del análisis de la varianza

Para continuar, accione el botón *Aceptar*. Esta acción abrirá la siguiente pantalla Figura 9.3. Por el momento, no modificaremos nada en esta pantalla. Sólo accionaremos el botón *Aceptar.* Esta acción generará la salida correspondiente al modelo estimado.

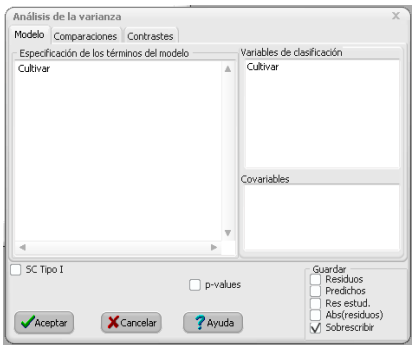


Figura 9.3: InfoStat. Diálogo de opciones del Análisis de la Varianza.

Cuadro 9.1: Análisis de la varianza aplicado a los datos del archivo [Híbridos].

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rend.	40	0,32	0,26	23,73

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	10026,83	3	3342,28	5,68	0,0027
Cultivar	10026,83	3	3342,28	5,68	0,0027
Error	21194,85	36	588,75		
Total	31221,68	39			

El coeficiente de variación (CV) de la salida anterior se calcula así:

$$CV = \frac{\sqrt{CMError}}{Media\ general} \times 100 = \frac{\sqrt{588,75}}{102,27} \times 100 = 23,73$$

La primer tabla presenta la información complementaria al ANAVA: (1) Se destaca la variable dependiente en análisis: en nuestro ejemplo *Rend*; (2) Se informa que en total se han utilizado *N*= 40 datos para conducir el ANAVA; (3) Se reporta un coeficiente de determinación R^2 = 0,32 por tanto el modelo lineal adoptado para conducir el ANAVA explica el 32% de la variabilidad total en los datos. Este coeficiente, representa sólo una porción de la variabilidad total por tanto deducimos que otros factores distinto a la genética (híbrido usado) estarán impactando la variabilidad de los rendimientos; (4) El coeficiente de variación, CV, de la variable respuesta rendimiento que es igual a 23,7%. El CV brinda información acerca de la relación porcentual entre la variabilidad residual (no explicada por el modelo) y la media de los datos. A menor CV, mejor calidad de información disponible en el estudio. La segunda tabla es la del ANAVA propiamente dicha, en el formato que hemos presentado. En la línea identificada como "Error" y en

la columna titulada como CM podemos leer el valor del Cuadrado Medio Dentro, y en la línea identificada como Cultivar el valor del Cuadrado Medio Entre Tratamientos (es importante destacar que en un modelo lineal a un criterio de clasificación, el Cuadrado Medio Entre es igual al Cuadrado Medio de Modelo). Así, en la columna titulada como F, se puede leer el cociente CME/CMD que es igual a 5,68, con un valor-*p* igual a 0,0027, lo que sugiere el rechazo de la hipótesis nula de igualdad de medias de tratamientos si se trabaja con un nivel de significación del 5% o α= 0,05.

Conclusión

Si bien el coeficiente de determinación R^2 es bajo (0,32) el modelo lineal adoptado para conducir el ANAVA permite rechazar la hipótesis nula (*P*<0,05). El coeficiente de variación es bajo y sugiere un experimento informativo por lo que podría concluirse que la variabilidad residual (no explicada por el modelo) en proporción a la media de los datos, fue mantenida bajo control en el experimento. Estos resultados indican que el factor híbrido es estadísticamente significativo para explicar diferencias de rendimientos medios entre estos 4 materiales; no obstante existe un porcentaje alto de variabilidad que es explicado por algún o algunos otro(s) factores no tenidos en cuenta en el análisis. Hay al menos un híbrido que rinde diferente a los demás.

Pruebas 'a Posteriori': Comparaciones múltiples de medias

Cuando se rechaza la hipótesis nula del ANAVA podemos concluir que existen diferencias significativas (p<0.05) entre al menos dos de las medias poblacionales de en evaluación.

Se plantea ahora el problema de detectar cuál o cuáles son los tratamientos que tienen medias poblacionales diferentes y cuáles son iguales, si es que hay algunos tratamientos que no se diferencian estadísticamente. Este problema se resolverá en base a **pruebas de comparaciones múltiples de medias** conocidas en general y más técnicamente como **comparaciones 'a posteriori'** del ANAVA.

En el ANAVA del problema en el que se evalúan 4 híbridos, utilizando los datos en el archivo [Híbridos], concluimos (ver sección anterior) que se rechazaba la hipótesis nula de igualdad de medias poblacionales de estos híbridos. El problema que abordaremos ahora es el detectar cuál o cuáles medias de híbridos son las distintas. Existen un conjunto importante de pruebas 'a posteri' disponibles que pueden realizarse tras haberse rechazado (exclusivamente) la hipótesis nula del ANAVA en base al test F.

Si el número de tratamientos es suficientemente grande, es probable que la diferencia entre la media mayor y la menor sea declarada como significativa por una prueba T de comparación de medias de dos poblaciones, aún cuando la H_0 no fue rechazada en el ANAVA. Así, realizando comparaciones de a pares usando la prueba T, cada una con un nivel α, la probabilidad de rechazar incorrectamente H_0, al menos una vez, incrementa con el número de tratamientos. Luego, teniendo como objetivo controlar α, y en algunos casos contralar β, existen varios procedimientos de comparaciones múltiples 'a posteriori' .

Existe una gama muy amplia de alternativas para llevar adelante este tipo de pruebas, las que por su naturaleza, pueden clasificarse en **pruebas tradicionales** y **pruebas basadas en conglomerados**.

Los *procedimientos tradicionales* generalmente presentan una menor tasa de error tipo I que los *procedimientos basados en conglomerados* cuando se trabaja en experimentos que no tienen un buen control de los niveles de precisión usados para la comparación de medias. No obstante, con un número alto de medias de tratamiento, los procedimientos tradicionales pueden producir salidas de difícil interpretación ya que una misma media puede pertenecer a más de un grupo de medias. Por el contrario, los métodos jerárquicos para comparaciones de medias producen agrupamientos mutuamente excluyentes (partición del conjunto de medias de tratamientos) y por tanto cada media solo clasificará en un grupo de la partición.

Se presentarán aquí solo dos pruebas tradicionales: las pruebas de Fisher y de Tukey y, de los procedimientos que no generan superposiciones entre grupos de medias estadísticamente indistinguibles, solo se presentará la prueba de Di Rienzo, Guzman y Casanoves (DGC), sugiriéndose al lector que revise la presentación más amplia hecha en esta temática en el Manual de InfoStat.

Prueba de Fisher

La **prueba de Fisher** es similar a la prueba de Tukey, en el sentido de comparar todos los pares de media muestrales con un estadístico y decidir en función de tal comparación si las medias poblaciones correspondientes son estadísticamente diferentes o no. No obstante, el estadístico de la prueba es diferente. En vez de usar los cuantiles de la distribución de rangos estudentizados utiliza los cuantiles de una de una distribución t de Student con los grados de libertad del cuadrado medio dentro de tratamientos y es particular para cada comparación de medias ya que depende del número de repeticiones por tratamiento. Luego, la diferencia mínima significativa entre el tratamiento i-ésimo y el tratamiento j-ésimo, **DMSf**, está dada por:

$$DMSf_{ij} = t_{gld;(1-\alpha/2)} \sqrt{CMD \; \frac{n_i + n_j}{n_i \; n_j}}$$

Con la prueba de Fisher es más fácil rechazar la hipótesis de igualdad de medias que con la prueba de Tukey, por esta razón se dice que este último es más conservador y el primero más potente.

Prueba de Tukey

El **prueba de Tukey**, al igual que cualquier procedimiento tradicional para la comparación de medias, examina con un mismo estadístico todas las diferencias de

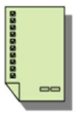
medias muestrales en estudio. Si hay *a* medias, luego habrá $\binom{a}{2} = \frac{a!}{(a-2)!\,2!}$ diferencias de medias posibles.

El estadístico propuesto por Tukey para este tipo de comparación es el siguiente:

$$DMSt = q_{a,gld;(1-\alpha)}\sqrt{\frac{CMD}{n}}$$

donde $q_{a,gld;\ (1-\alpha)}$ es el cuantil (1-α) que se obtiene de la distribución de Rangos Studentizados para *a* tratamientos y los grados de libertad dentro; α es el nivel de significación en base al cual se rechazó la H_0 del ANAVA y *n* es el número de repeticiones en base a las que se calculan las medias muestrales. Si el tamaño de muestra no fuera el mismo para cada tratamiento, deberá reemplazarse *n* por la media armónica de los $\{n_i\}$, esto es:

$$n_0 = \frac{a}{\sum_{i=1}^{a}\frac{1}{n_i}}$$



Si el valor absoluto de la diferencia entre un par de medias supera a DMSt, se dice que esta diferencia es estadísticamente significativa.

Se concluirá en consecuencia que las esperanzas asociadas a esa diferencia son distintas con un nivel de significación α.
Cabe destacar que cuando los tamaños muestrales son muy diferentes, esta prueba de Tukey puede dejar de ser confiable, caso en el cual podría utilizarse algún procedimiento de contraste múltiple que considere tal situación, como el de Scheffé (1953).

Prueba de Di Rienzo, Guzmán y Casanoves (DGC)

Este procedimiento de comparación de medias (Di Rienzo, *et al.*, 2002), utiliza la técnica multivariada del análisis de conglomerados (encadenamiento promedio o UPGMA), sobre una matriz de distancia entre medias muestrales de tratamiento.
Como consecuencia del análisis de conglomerado se obtiene un dendrograma en el cual puede observarse la secuencia jerárquica de formación de conglomerados. Si se designa como Q a la distancia entre el origen y el nodo raíz del árbol (aquel en el cual se unen todas las medias), la prueba utiliza la distribución de Q bajo la hipótesis: $H_0 : \mu_1 = \cdots = \mu_a$ para construir una prueba con nivel de significación α. Las medias (o grupos de medias) unidas en nodos que están por encima de Q, se pueden considerar estadísticamente diferentes para el nivel de significación α. El método presupone igual

número de repeticiones por tratamiento, en caso contrario el algoritmo implementado utiliza la media armónica del número de repeticiones.

Aplicación

Comparación de redimientos promedios

En InfoStat para realizar una Prueba 'a posteriori', cualquiera sea ella, debe invocarse el Menú *Estadísticas* seleccione el submenú *Análisis de la Varianza*. Aparecerá la pantalla que ya hemos presentado anteriormente. Tras seleccionar *Cultivar* en el panel izquierdo de la ventana y agregarlo al panel *Variables de clasificación* y seleccionar *Rend* para luego agregarlo al panel *Variables dependiente,* al pulsar el botón Aceptar, aparecerá una nueva ventana, como la que presentáramos en la Figura 9.3. Al activar la solapa "Comparaciones" de esta ventana, se presentará un nuevo diálogo como el que se presenta a continuación:

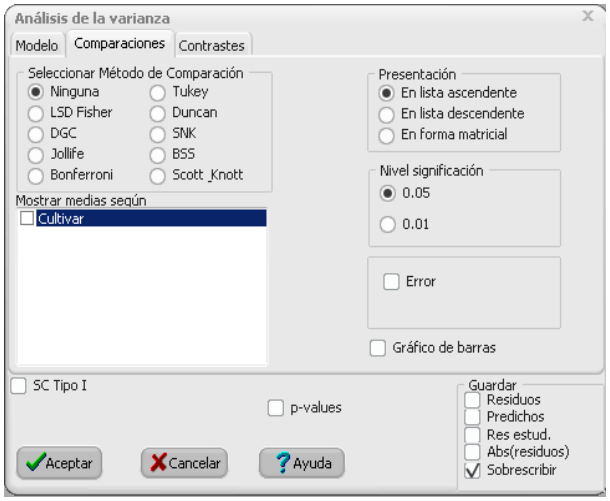


Figura 9.4: Diálogo de Comparaciones Múltiples de a pares de medias o Pruebas 'a Posteriori' del ANAVA en InfoStat

Para cualquier procedimiento que se elija, InfoStat permite definir el nivel de significación nominal usado para la prueba seleccionada (0,05 o 0,01 son los valores usuales). Además, se puede optar por el tipo de presentación de los resultados de las comparaciones múltiples (en forma de lista ascendente, descendente o en forma matricial). Si solicita presentación en lista, las comparaciones se muestran en una lista en la cual letras distintas indican diferencias significativas entre las medias que se comparan. Si seleccionamos la Prueba de Tukey y pulsamos el botón *Aceptar*, obtendremos la siguiente salida en la ventana de Resultados de InfoStat.

Cuadro 9.2: Análisis de la varianza y el test 'a posteriori' de Tukey aplicado a los datos del archivo [Híbridos].

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rend.	40	0,32	0,26	23,73

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	10026,83	3	3342,28	5,68	0,0027
Cultivar	10026,83	3	3342,28	5,68	0,0027
Error	21194,85	36	588,75		
Total	31221,68	39			

Test:Tukey Alfa=0,05 DMS=27,72246
Error: 588,7457 gl: 36

Cultivar	Medias	n	E.E.		
2,00	76,68	10	7,67	A	
4,00	105,44	10	7,67		B
1,00	106,90	10	7,67		B
3,00	120,06	10	7,67		B

Medias con una letra común no son significativamente diferentes(p<= 0,05)

Si se solicita presentación matricial, InfoStat presenta las comparaciones en una matriz cuya diagonal inferior tendrá como elementos las diferencias entre las medias y en la diagonal superior se presenta el símbolo "*" indicando los pares de medias que difieren estadísticamente al nivel de significación elegido. Si en la ventana de diálogo de la solapa Comparaciones de InfoStat seleccionamos la Prueba LSD de Fisher, los resultados serán los del Cuadro 9.3.

Cuadro 9.3: Análisis de la varianza y el test 'a posteriori' LSD de Fisher aplicado a los datos del archivo Híbridos

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rend.	40	0,32	0,26	23,73

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	10026,83	3	3342,28	5,68	0,0027
Cultivar	10026,83	3	3342,28	5,68	0,0027
Error	21194,85	36	588,75		
Total	31221,68	39			

Test:LSD Fisher Alfa=0,05 DMS=22,00731
Error: 588,7457 gl: 36

Cultivar	Medias	n	E.E.		
2.00	76,68	10	7,67	A	
4.00	105,44	10	7,67		B
1.00	106,90	10	7,67		B
3.00	120,06	10	7,67		B

Medias con una letra común no son significativamente diferentes(p<= 0,05)

Conclusión

Las medias muestrales, ordenadas en forma ascendente, muestran que el cultivar 2 tiene el menor de los rendimientos (76,68 qq/ha), le sigue el cultivar 4 (105,44 qq/ha), el cultivar 1 (106,90 qq/ha) y el cultivar 3 es el de mayor rendimiento de los cultivares comparados (120,06 qq/ha).

Las tres pruebas presentadas (Tukey y LSD de Fisher), nos muestran idénticos resultados, asignando la letra A al cultivar 2 y la letra B a los cultivares 4, 1, 3. Tratamientos que comparten una misma letra no se pueden declarar como estadísticamente diferentes, es decir las diferencias muestrales observadas pueden haberse dado por azar y por tanto no ser repetibles. Por ello, los investigadores sólo concluyen sobre diferencias que resultan estadísticamente significativas. Así los resultados del experimento particular pueden extenderse a la población ya que se espera estabilidad de las relaciones halladas.

Los resultados de las pruebas a posteriori en el ejemplo nos permite concluir que:

(1) El cultivar 2 posee una media significativamente diferente (y menor) a las medias poblacionales de los otros tres cultivares; y

(2) Las medias poblacionales no difieren significativamente entre los cultivares 4, 1 y 3.

Es probable plantearse porque no es significativa la diferencia entre el cultivar 4 y 3, ya que sus medias muestrales difieren en 120,06-105,44=14.52 qq/ha, diferencia que

agronómicamente puede ser de relevancia económica en grandes superficies de cultivo. La respuesta pasa por considerar la magnitud del Cuadrado Medio del Error del ANAVA, que es parte del cálculo del estadístico *Diferencia Mínima Significativa (DMS)*, parece que las diferencias entre estas medias son de la magnitud de las diferencias dentro de tratamiento. Las DMS que declara a dos medias poblacionales como significativamente diferentes si la diferencias entre las medias muestrales en la Prueba de Tukey (DMS=27,72246 qq/ha) es diferente a la obtenida en la prueba LSD de Fisher donde la DMS es menor (DMS= 22,00731 qq/ha).

Verificación de supuestos del ANAVA

El modelo lineal del ANAVA plantea **supuestos** que deben cumplirse para que el estadístico F=CME/CMD tenga la distribución F con (*a*-1) y *a*(*n*-1) grados de libertad y por tanto los valores p reportados sean válidos.

Estos supuestos plantean exigencias acerca de los términos de error aleatorios ε_{ij} y se pueden establecer como: (a) **independencia** entre términos de error aleatorio, (b) **distribución normal** de los términos de error aleatorio, con esperanza cero, y (c) que la varianza de los términos de error se mantenga constante para todo *i*, *j* ; este último supuesto puede entenderse también como **homogeneidad de varianzas** dentro de cada tratamiento, o que la variabilidad de las observaciones bajo los distintos tratamientos es la misma o no difiere significativamente.

En caso que alguno de estos supuestos (normalidad, homogeneidad de varianzas o independencia) no se cumplan, impactarán sobre la distribución del estadístico F y con ello el verdadero nivel de significancia de la prueba de hipótesis del ANAVA, afectando así la calidad de las conclusiones que finalmente buscamos obtener, con probabilidades de los Errores Tipo I y II que no son las esperadas.

Existen distintas técnicas de validación de supuestos, pero las que se presentan aquí se basan en los predictores de los errores, es decir los residuos.

El **residuo** e_{ij} de la observación j-ésima del tratamiento i-ésimo fue definido como el predictor de ε_{ij}, y puede ser calculado como la diferencia entre el valor observado y el valor predicho por el modelo lineal dado. Para un DCA a un criterio de clasificación, el residuo asoaciado a una UE particular se calcula como:

$$e_{ij} = y_{ij} - \overline{y}_i$$

Para calcular todos los residuos con InfoStat, es necesario entrar al submenú *Análisis de la Varianza* y especificar la variable de clasificación y la respuesta, tal cual lo hemos aprendido a hacer para conducir el ANAVA propiamente dicho. Cuando se llega a la ventana de opciones del ANAVA deben tildarse las celdas de *Guardar Residuos, Predichos, Residuales Estudentizados (Res.Estud.) y Absolutos de los Residuos (Abs(residuos))* como se muestra en la siguiente Figura, para que se agreguen las columnas respectivas en la tabla de datos con que estemos trabajando.

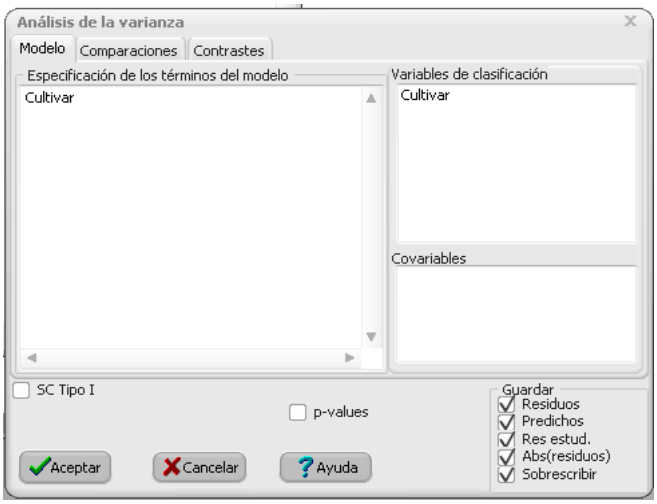


Figura 9.5: InfoStat. Diálogo de opciones del ANAVA, para la generación de residuos, predichos y otros estadísticos necesarios para la verificación de supuestos, en InfoStat

Una vez generadas estas columnas con los residuos, los predichos, los residuos estudentizados (una forma de residuos que estandariza de manera tal que la variación de los mismos quede comprendida entre -4 y 4 y así se puedan identificar fácilmente residuos "altos" o "bajos") y los valores absolutos de los residuos, procederemos a verificar el cumplimiento de los supuestos de normalidad, independencia y homogeneidad de varianzas de los ε_{ij}, mediante las siguientes pruebas de hipótesis e interpretaciones gráficas.

Normalidad

Tomando los residuos como dato de análisis, una de las técnicas más usadas es construir un **Q-Q plot normal**. Mediante esta técnica se obtiene un diagrama de dispersión en el que, si los residuales son normales y no hay otros defectos del modelo, los residuos observados se alinean sobre una recta a 45° como se muestra en la siguiente figura ya que correlacionan bien con los residuos esperados bajo el supuesto que la muestra de datos realmente sigue una distribución normal. El gráfico compara los cuantiles observados con los cuantiles esperados bajo normalidad.
La presencia de ligeras violaciones de este supuesto no es muy grave para el ANAVA, no afectándose de forma importante la probabilidad de cometer Error de Tipo I. La Figura 9.7 ilustra el Q-Q plot de residuos del problema de los Híbridos que venimos estudiando a lo largo de este Capítulo. En las siguientes figuras se presentan los diálogos de InfoStat para generar el Q-Q Plot mostrado.

Para acceder a la ventana de diálogo que permite seleccionar la variable para hacer el QQ-Plot de interés, acceder al Menú *Gráficos*, submenú *Q-Q Plot*. Tras elegir la variable

RDUO-Rend. y pulsar el botón *Aceptar*, se presentará una segunda ventana de diálogo, que permite elegir el modelo de distribución a validar como se muestra a continuación.

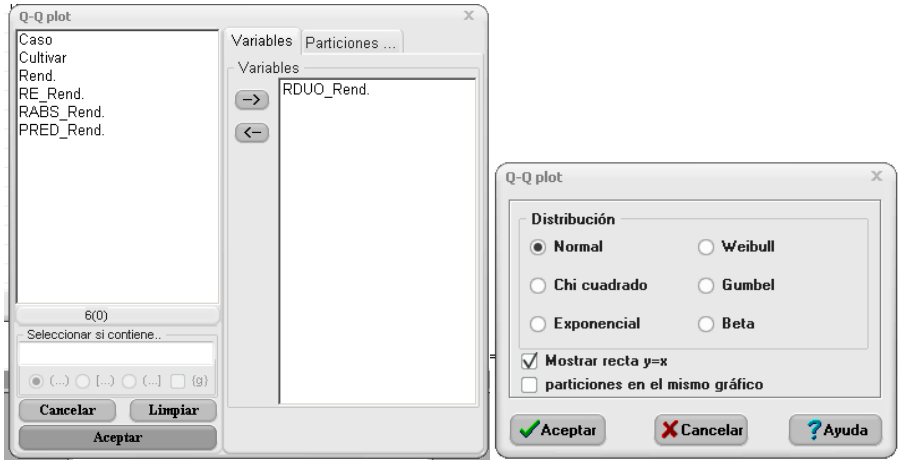


Figura 9.6: InfoStat. Diálogos para generar un Q-Q plot para prueba de distribución normal.

Tras accionar el botón *Aceptar*, se construirá el gráfico como el que se muestra en la la siguiente figura:

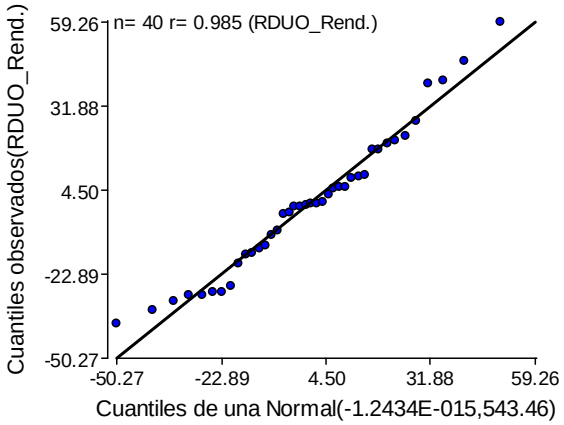


Figura 9.7: Q-Q Plot de los residuos del ANAVA en InfoStat

Homogeneidad de varianzas

Cuando los términos de error tienen **varianzas homogéneas** y el modelo explica bien a los datos (es decir no queda ninguna fuente de variación sistemática que aún se pueda remover), el gráfico de dispersión de residuos vs. predichos presentará una nube de puntos sin patrón alguno. Por ello, los investigadores usan los gráficos de dispersión de

residuos con patrones aleatorios como indicador de un buen ajuste del modelo a sus datos.

Un patrón en este tipo de gráficos que indica falta de homogeneidad en las varianzas se muestra en la Figura 9.8. La heterogeneidad de varianzas de pone de manifiesto ya que a medida que crecen los valores predichos por el modelo, aumentan las dispersiones de los residuos; así los tratamientos con mayores valores predichos tienen más variabilidad entre sus repeticiones que los tratamientos con menor valor predicho. Este tipo de patrón es indeseable ya que puede llevarnos a cometer errores en las conclusiones; frecuentemente se asocia con una mayor probabilidad de cometer Error Tipo II, es decir no detectar diferencias entre tratamientos cuando éstas realmente existen.

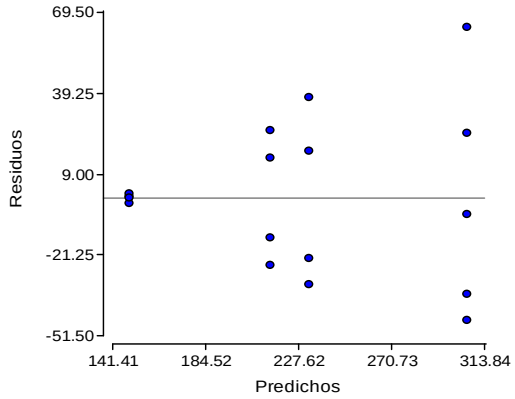


Figura 9.8: Gráfico de Residuos en función de Predichos en un ejemplo con falta de homogeneidad de varianzas.

En el ejemplo de aplicación, para generar esta gráfica, se debe entrar al menú *Gráficos* submenú *Diagrama de Dispersión* y asociar *RE-Rend* al Eje Y y *PRED-Rend* al Eje X. Se obtendrá así el diagrama a la derecha del diálogo del Diagrama de dispersión de la siguiente Figura, que sugiere que la variabilidad de los rendimientos en el híbrido de menor rinde pareciera diferente a la variabilidad del rendimiento en los otros híbridos. Para estas situaciones donde se observan diferencias o algún patrón particular, existen pruebas formales para detectar la significancia de las mismas como es la Prueba de Levene que se construye como un ANAVA del valor absoluto de los residuos. Si ese ANAVA presenta un valor p pequeño se concluye que la heterogeneidad de varianzas es importante y, como podría afectar la potencia de nuestras conclusiones, se recurre otro tipo de ANAVA donde no es necesario suponer varianzas homogéneas como es el caso del ANAVA bajo un modelo lineal mixto.

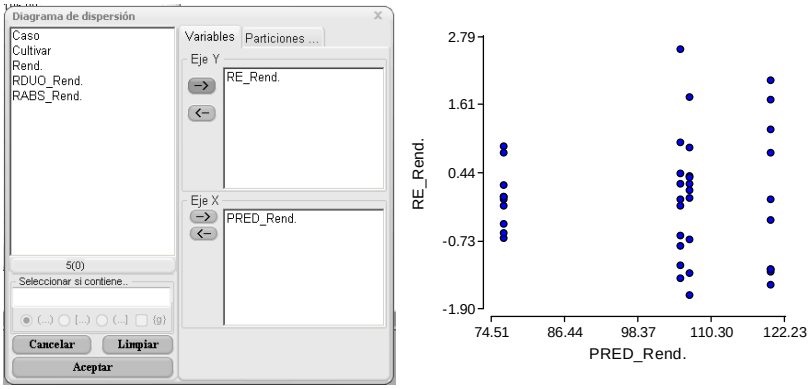
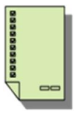


Figura 9.9: nfoStat. Gráfico de Residuos vs. Predichos

Independencia

Una ayuda valiosa para estudiar la posible falta de independencia entre los errores es realizar un gráfico de los residuos según la secuencia en el tiempo o espacio físico en que han sido colectados los datos; por supuesto que para tal prueba debe conocerse cómo ha sido el mecanismo de recolección de datos. Si los residuos aparecen en secuencias de varios valores positivos seguidos de varios valores negativos puede ser un indicio claro de la falta de independencia. Otro posible patrón indicativo de falta de independencia es una sucesión alternante de residuales positivos y negativos. Siempre que se detecte cualquier patrón distinto al aleatorio (falta de patrón), se debe sospechar del incumplimiento del supuesto de independencia.



La falta de independencia es un problema potencialmente peligroso y difícil de corregir, por lo que es importante prevenirlo. La aleatorización en la asignación de los tratamientos a las unidades experimentales, en la secuencia de medición de los resultados del ensayo, o en cualquier otra etapa experimental que pueda introducir una fuente sustancial de error, es uno de los métodos más eficaces de controlar la falta de independencia.

En el ejemplo de los híbridos, esta gráfica no se puede realizar porque no se registró la secuencia de tiempo en que se realizaron las mediciones de las parcelas, ni tampoco las ubicaciones de las parcelas en el campo, como para poder realizar una gráfica que permita evaluar la posible falta de independencia (temporal o espacial), que pueda haber ocurrido en este experimento. De la inspección del gráfico Q-Q Plot de normalidad de los residuos del modelo lineal del ANAVA adoptado, se puede informar que no se observa una alejamiento importante del modelo normal. Algo similar ocurre con el gráfico de dispersión de los residuos versus los predichos, en el sentido que no se observa un patrón de heterogeneidad de varianzas de relevancia (excepto por el cultivar

de menor rendimiento). Por lo que podría asumirse que los términos de error verifican los supuestos y tomar como válidas las conclusiones realizadas tanto para el ANAVA como para las pruebas 'a posteriori' conducidas. Cuando los supuestos de Normalidad y Homocedasticidad (homogeneidad de varianzas) no se cumplen, algunos investigadores recurren a la transformación de los datos a otras escalas, como la logarítmica, raíz cuadrada o arco seno, donde los supuestos puede ser que se cumplan. Por ende las comparaciones de realizan en la escala donde el ANAVA es válido.

Ejercicios

Ejercicio 9.1: En la Provincia de Córdoba se produce aproximadamente el 95% del maní tipo confitería destinado a exportación. En el año 2006 se realizó un estudio en el que se indagaron estrategias tecnológicas productivas y características socio-económicas de los productores de maní de la Provincia de Córdoba. A partir de este estudio, se pudo clasificar a los productores como pequeños a medianos productores independientes (Tipo de Productor I), grandes productores (Tipo de Productor II) y pequeños a medianos productores no independientes asociados a grandes productores (Tipo de Productor III). Luego, otros investigadores estudiaron si los rendimientos medios logrados por esta tipología de productores diferían entre sí, con la hipótesis científica de que los Productores Tipo II y III lograban rendimientos medios superiores a lo alcanzados por los Tipo I. En el archivo [Mani] (disponible por gentileza de la Lic. Mara LLop) se encuentran los rendimientos de 27 productores entrevistados (9 de cada Tipo) a los que se les solicitó información veraz (cartas de porte del grano entregados para su venta) sobre los volúmenes cosechados, los que permitieron calcular rendimientos promedios por hectárea logrado por cada productor.

Se solicita:

a) Plantear las hipótesis estadísticas que se podrían contrastar en este problema y reflexionar sobre la naturaleza del estudio (observacional vs experimental)

b) Realizar el Análisis de la Varianza ($\alpha = 0.05$)

c) Valide los supuestos de homogeneidad de varianzas y de normalidad de los términos de error aleatorio

d) Si corresponde, realizar la prueba LSD de Fisher.

e) Redactar conclusiones.

Ejercicio 9.2: Una empresa agrícola necesita establecer si le conviene, desde el punto de vista económico, fertilizar sus cultivos de soja. Para este propósito se realizó un ensayo en un lote de 20 has, dividido en parcelas de una hectárea cada una, en el que se evaluaron cuatro estrategias de fertilización: (a) No fertilizar, (b) usar el Fertilizante A, (c) usar el Fertilizante B y (d) usar el Fertilizante C, asignando los tratamientos en forma aleatoria. Cada parcela fue laboreada culturalmente con la misma tecnología de siembra directa en cuanto al manejo de plagas, malezas, densidades de siembra, variedades, fecha de siembra y control de humedad en el suelo. La única diferencia entre ellas fue el fertilizante utilizado.

Considere ahora que el precio de la tonelada de soja es de $1200, los costos de producción de cada parcela son del orden de los 15 qq/ha (sin incluir el costo del Fertilizante), el costo por hectárea de usar el Fertilizante A es de 5 qq/ha, del utilizar el Fertilizante B de 3,5 qq/ha, de usar el Fertilizante C de 2 qq/ha, y que los rendimientos obtenidos (qq/ha) fueron:

Sin fertilizar	Fertilizante A	Fertilizante B	Fertilizante C
19	**33**	**33**	**28**
20	**35**	**31**	**24**
22	**29**	**35**	**25**
23	**31**	**34**	**26**
21	**30**	**32**	**27**

a) Trabajar con la variable Y=Beneficio Económico($/ha), la que se calcula en este caso como Rendimiento (qq/ha) × Precio de la Producción($/qq) ï Costos de Producción ($/ha). Realizar previamente una representación gráfica comparativa de los Beneficios Económicos($/ha) logrados en las parcelas de este estudio experimental.

b) Conduzca un ANAVA con la variable Y=Beneficio Económico($/ha), verifique los supuestos de homogeneidad de varianzas y normalidad, y de ser necesario una prueba de comparaciones múltiples.

c) ¿Cuál de los fertilizantes recomendaría?

Ejercicio 9.3: Se desea evaluar la calidad de plantas de olivos producidas por esqueje o estaca, cuando éstas son sometidas a un tratamiento promotor del enraizamiento (lavado durante 48 horas antes de ser plantadas en el almázigo). Para ello, se toman 10 estacas de una cierta Variedad (Arbequina) y se las planta directamente (Tratamiento A) en macetitas de enraizamiento, dándosele luego el manejo convencional para que enraícen (humedad ambiente, temperatura, fertiriego, fungicidas, bactericidas) y a otras 10 estacas de la misma Variedad se las somete previamente al lavado con agua corriente durante 48 horas (Tratamiento B), para luego seguir con el manejo convencional para que enraícen. Se presenta a continuación la altura de las plantas (cms) lograda a partir de esos esquejes, al cabo de 90 días de haber sido plantadas:

Sin lavar	8	12	15	16	9	16	14	15	11	14
Con lavado	9	9	8	12	10	11	13	14	9	10

a) Realizar la prueba del test F del análisis de varianza, previa verificación de los supuestos de normalidad y homogeneidad de varianzas, usando un nivel de significación del 5%.

b) Comprobar que el valor del estadístico T para comparar dos poblaciones con varianzas homogéneas, cuando es elevado al cuadrado, reproduce el valor del estadístico F del ANAVA.

c) ¿Qué se concluye sobre las diferencias en altura de las plantas logradas al cabo de 90 días de haber sido plantadas?

Ejercicio 9.4. Se desea conocer el efecto de las cepas de inoculantes de Rhizobium, fijadoras de nitrógeno atmosférico, sobre el contenido de nitrógeno de plantas de trébol rojo. Para ello se dispone de 30 macetas de trébol rojo en un invernadero. Se asignan al azar 5 macetas para cada una de las cepas y se procede a inocularlas. Los resultados son los siguientes (en mg. de nitrógeno/Kg de Materia Seca):

Cepa I	Cepa II	Cepa III	Cepa IV	Cepa V	Cepa VI
29.4	27.7	19.1	18.6	11.6	16.9
29.0	24.3	16.9	18.8	11.8	17.3
32.1	24.8	15.8	20.5	14.2	19.1
32.6	25.2	17.0	20.7	14.3	19.4
33.0	27.9	19.4	21.0	14.4	20.8

a) ¿Cuales son las unidades experimentales?, ¿Cuántas repeticiones hay?
b) Plantear las hipótesis científicas y estadísticas del experimento
c) Realizar el Análisis de la Varianza ($\alpha = 0.05$) y concluir sobre si las distintas cepas producen el mismo nivel de fijación de nitrógeno o no.
d) Si corresponde, realizar una prueba ña posterioriò e indique que cepa o cepas recomendaría.

Ejercicio 9.5 Se desea estudiar el efecto de la carga animal sobre la producción de materia seca en una pastura implantada. Para ello se divide un lote en 28 potreros y se asignan aleatoriamente 7 potreros a cada una de las 4 cargas animales en estudio (2 nov./ha., 4 nov./ha, 6 nov./ha. y 8 nov./ha.). Los resultados fueron los siguientes expresados en toneladas de materia seca por hectárea.

								Media
carga 2	2.6	1.9	3.1	2.8	2.2	2.0	2.7	2.47
carga 4	3.3	3.6	3.0	3.5	3.2	3.9	3.4	3.41
carga 6	3.1	2.0	2.5	3.1	2.3	3.0	2.2	2.60
carga 8	2.5	2.3	2.8	1.8	2.7	2.6	2.0	2.39

a) Plantear un modelo lineal que permita recomendar alguna carga en especial.
b) ¿Qué supuestos se requieren para el análisis de este ensayo?
c) Realizar el análisis y concluya. Trabajar con un nivel de significación de 0.05.

Análisis de experimentos a un criterio de clasificación

Ejercicio 9.6 Una empresa de agroquímicos ha producido un nuevo inoculante para soja, que saldrá a la venta si con su aplicación se obtienen mayores rendimientos que sin su utilización. Para evaluar al inoculante se realiza un experimento inoculando 14 lotes de semillas. La mitad de los 14 lotes se inoculan con una dosis baja (Dosis 1) y la otra mitad con una dosis más alta (Dosis 2). Además se incluyen en el ensayo 6 lotes de semillas sin inocular (testigo o control). El experimento se realiza en un mismo ambiente y se implementa usando la variedad y la forma de manejo de cultivo más difundida para ese ambiente. Cada lote de semillas se asigna al azar a una de las parcelas del ensayo que se consideran homogéneas desde un punto de vista práctico. Se midió el rinde en gr/m2 por cada parcela y luego se lo llevó a qq/ha. Se trabajó con un nivel de significación del 0.05, usando el siguiente modelo:

$$y_i = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1,...,a \qquad j = 1,...,n \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Del análisis estadístico se obtuvieron los siguientes resultados:

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rinde	20	0.48	0.42	6.92

Cuadro de Análisis de la Varianza

F.V.	SC	gl	CM	F	p-valor
Modelo	74.07	2	37.04	8.00	0.0036
Trat	74.07	2	37.04	8.00	0.0036
Error	78.66	17	4.63		
Total	152.73	19			

Test:LSD Fisher Alfa:=0.05 DMS:=2.48901
Error: 4.6272 gl: 17

Trat	Medias	n		
Sin Inocul.	28.17	6	A	
Inoc. Dosis 2	32.05	7		B
Inoc. Dosis 1	32.62	7		B

Letras distintas indican diferencias significativas(p<= 0.05)

De acuerdo con estos resultados asignar la condición de Verdadero (V) o Falso (F) a cada una de las siguientes afirmaciones:

En el experimento se utilizó igual cantidad de repeticiones para cada tratamiento	
El diseño experimental utilizado en este ensayo fue el diseño completamente aleatorizado	
En el modelo lineal, una de las tres componentes, representa el rendimiento promedio bajo el tratamiento i-esimo	
La hipótesis nula del ANAVA establece que los promedios de los rendimientos obtenidos con cualquiera de las dos dosis de inoculante y con el tratamiento testigo, son estadísticamente iguales	
La fuente de variación "Trat" tiene 2 grados de libertad porque en el experimento hay dos tratamientos en evaluación	
Como el valor p=0.0036 es menor que el nivel de significación, se puede conlcuir que la variabilidad de los rendimientos entre tratamientos es menor a la variabilidad dentro de los tratamientos	
El valor p= 0.0036 permite concluir que los rendimientos obtenidos en parcelas sembradas con semillas con la misma condición de inoculación fueron menos variables que los obtenidos en parcelas sembradas con semillas con diferentes condiciones de inoculación	
La diferencia mínima significativa de Fisher indica una cota mínima para diferencia que debe existir entre las medias muestrales de dos tratamientos para declarar a las medias poblaciones de estos tratamientos como estadistícticamente diferentes	
El uso de inoculante permite obtener un mayor rendimiento	
Convendría usar la dosis más alta del inoculante ya que al aumentarla se obtuvo mayor rendimiento	
Dado que las diferencias muestrales o experimentales observadas, entre no inocular e inocular, son estadísticamente significativas se podría recomendar la inoculación ya que la probabilidad de azar en estas diferencias es baja. Se considera que estas diferencias no se dieron por azar y que es probable que se vuelvan a repetir en otra situación donde se comparen cultivos de soja sin inoular e inoculados como los analizados en este experimento.	

Capítulo 10

Análisis de experimentos con varios criterios de clasificación

Mónica Balzarini

10. Análisis de experimentos con varios criterios de clasificación

Motivación

Hemos presentado el ANAVA como un método estadístico cuya finalidad es contrastar hipótesis referidas a la comparación de medias de dos o más poblaciones. Supusimos que esas poblaciones están conformadas por unidades de análisis expuestas a distintas condiciones, que hemos llamado "tratamientos". Así, el factor tratamiento es entendido como un criterio de clasificación, ya que luego de su aplicación a las unidades experimentales, éstas quedan clasificadas según los distintos niveles de tratamiento. No obstante, existen situaciones donde los criterios de clasificación de las unidades son muchos y el modelo lineal de ANAVA debe extenderse para contemplarlos en el análisis.

Conceptos teóricos y procedimientos

Más de un criterio de clasificación

En algunas ocasiones los tratamientos se definen por la combinación de dos o más factores, por ejemplo combinaciones del factor "principio activo" del producto terapéutico en uso y el factor "dosis" de aplicación del producto. Si los principios activos son 2 y las dosis son 2, entonces decimos que existe una estructura factorial de tratamientos que produce 4=2x2 tratamientos. Ahora, existen dos criterios de clasificación de los datos y ambos están relacionados a cuestiones que interesan evaluar (tratamientos). En experimentos con estructura factorial de tratamientos, surge una

nueva pregunta referida a la existencia o no de interacción entre ambos factores tratamientos.

Además de estos experimentos con dos criterios de clasificación en la estructura de tratamientos, existen otros donde la multiplicidad de criterios de clasificación se da a nivel de las unidades experimentales (UE). Por último, otro caso frecuente, se da cuando las UE son clasificadas por dos criterios, pero uno se refiere al factor tratamiento (factor de interés) y otro a un factor que genera variabilidad entre las UE, tal es el caso del Diseño en Bloques Aleatorizados. Aún cuando el factor de bloqueo de UE, no es el factor sobre el que se quiere concluir, interesa tenerlo en cuenta durante el análisis ya que puede ocasionar variaciones sistemáticas importantes sobre la variable respuesta y, de ser ignorado, podría conducirnos a sobreestimar la variabilidad esperada entre repeticiones y por tanto afectar las comparaciones entre medias de tratamiento. Estos factores de la estructura de las UE suelen ser denominados factores de control, y al contemplarlos en el análisis es posible disminuir el impacto negativo que algunos "ruidos" experimentales podrían tener sobre las conclusiones. En cualquiera de las situaciones, la principal pregunta de los modelos de ANAVA que discutiremos es: ¿cómo afectan los tratamientos a la respuesta?, ¿Hay diferencias, a nivel medio, entre tratamientos?

Cuando los datos son explicados por un modelo de clasificación en términos de factores, ya sean estos de tipo factores tratamientos o factores de control, la pregunta que siempre está presente es ¿cómo afectan los distintos niveles de los factores a la variable respuesta? La estimación de un modelo lineal de ANAVA, expresado en término de constantes desconocidas relacionadas a los efectos de los **factores**, permitirá responder esta pregunta.

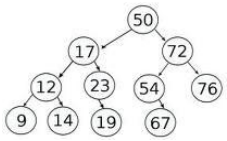
Supongamos que se tienen datos de una variable respuesta Y para *a* niveles de un factor A y *b* niveles de un factor B. Los niveles han sido fijados o determinados por el experimentador ya que son precisamente los efectos de esos niveles de los factores que interesan comparar. Luego un modelo lineal para el valor esperado bajo el i-ésimo nivel del factor A (i=1,...,a) y el j-ésimo nivel del factor B (j=1,...,b) podría ser

$$\mu_{ij} = E(Y_{ij}) = \mu + \alpha_i + \beta_j$$

con μ, α y β constantes desconocidas que representan la media general de las observaciones, el efecto del factor A y el efecto del factor B. El modelo lineal anterior se denomina modelo de **ANAVA de efectos fijos a dos vías de clasificación**; este modelo asume que los efectos de ambos factores son aditivos, es decir no existe interacción o dependencia entre estos efectos. Algunos modelos a dos criterios de clasificación permiten adicionar otros términos compuestos formados a partir de los efectos de los factores principales. Un ejemplo de término compuesto es el efecto de interacción entre los factores, que describiremos más adelante.

Estructuras en los datos

El **modelo estadístico** es una simplificación de la realidad. No obstante, si proporciona un buen ajuste para los datos permitirá comprender mejor esta realidad y posiblemente predecir futuros valores de la variable de interés. El modelo es una abstracción del proceso generador de datos (PGD) que captura aquellas características del proceso que permiten responder alguna pregunta particular.
En todo estudio experimental deben reconocerse dos estructuras: 1) la estructura de las unidades experimentales (UE) y 2) la estructura de los tratamientos. El **diseño del experimento** es el mecanismo usado para vincular estas dos estructuras.



Las estructuras presentes en los datos son partes del proceso generatriz de datos que debemos reconocer para poder postular un buen modelo para su análisis

La **estructura de unidades experimentales** sale a luz cuando nos preguntamos sobre el material experimental: Son las UE homogéneas?. Si la respuesta es afirmativa, diremos que no existe estructura en las UE y usaremos un diseño completamente aleatorizado (DCA), ya que si todas las UE son iguales, cualquiera podría recibir un tratamiento particular.

La homogeneidad de las UE es clave para decidir el diseño experimental a usar ya que siempre es de interes comparar los resultados obtenidos con distintos tratamientos pero en condiciones homogéneas de comparación.

Si la respuesta a la pregunta sobre la homogeneidad del material no es afirmativa, estaremos frente a un estudio donde existe la posibilidad de confundir efectos y esto no es deseado. Por tanto, intentaremos controlar este ruido extra que impone variabilidad entre las UE desde el principio del experimento (aún cuando no recibieran tratamientos distintos).

*Una forma de controlar variabilidad entre UE (no debida a efectos de tratamientos) es a través del "bloqueo o **estratificación de UE**". Cuando existe este tipo de estructura en las UE, el diseño experimental más difundido es el diseño en bloques completos al azar (DBCA).*

Independientemente de cuál fuera la condición de la estructura de las UE (digamos sin estructura o estratificadas), tendremos que pensar sobre la estructura de los tratamientos: Los tratamientos se encuentra definidos por un único factor, es decir existe sólo una vía o criterio de clasificación? Si la respuesta es afirmativa entonces

diremos que no hay estructura de tratamientos. Si para conformar un tratamiento debemos combinar dos o más factores, diremos que hay **estructura de tratamientos**. En este último caso puede ser que los factores se encuentren "cruzados" o "anidados".

Se habla de **factores cruzados** cuando cada nivel de un factor se combina con cada uno de los niveles del otro factor para formar un tratamiento. Ejemplo: En un ensayo comparativo de rendimiento de girasol, se evalúan una serie de cultivares en distintas localidades. Por ejemplo, se evalúan 10 cultivares de girasol en 25 localidades pertenecientes a la región girasolera argentina. Si todos los cultivares son evaluados en todas las localidades, se tendrán 10×25=250 tratamientos producto de la combinación de los distintos niveles de los dos factores.

Se habla de **factores anidados** cuando los niveles de un factor son distintos para cada nivel del otro factor. Ejemplo: En un rodeo lechero se evalúa la capacidad del toro a través de sus hijas, para ello, se inseminan 16 madres, 8 madres tendrán hijas del toro A y 8 madres tendrán hijas del toro B, en este caso, tenemos dos factores, uno dado por los toros, con dos niveles porque hay dos toros y el otro factor dado por las madres, el cual tiene 16 niveles. Pero las madres que son inseminadas con el semen del toro A, no son las mismas que las madres inseminadas con el toro B, por ello se dice que el factor madre está anidado en el factor toro.

Para citar otro ejemplo de anidamiento de factores, supongamos que se evalúa el daño provocado por un virus en diferentes hospederos vegetales en distintas zonas pertenecientes a una región. Se evalúo el daño en 5 hospederos: maíz, trigo, cebada, centeno, avena. Las localidades evaluadas fueron 9. Tenemos dos factores o fuentes de variación reconocidas a priori y sobre las que nos interesa inferir: el factor localidad y el factor hospederos. El primero tiene 9 niveles y el segundo 5 niveles. Los hospederos de una localidad son diferentes a los hospederos que se encuentran en otra localidad, por ello decimos que el factor hospedero se encuentra anidado en el factor localidad.

Cuando los factores tratamiento están cruzados se dice que se tiene una estructura **factorial** de tratamientos y el diseño suele denominarse bifactorial, trifactorial o multifactorial según se crucen los niveles de dos, tres o más factores, respectivamente.

Finalmente, la estructura de la variable respuesta también debe ser contemplada. Por ejemplo, cuando la respuesta se mide repetidamente en el tiempo sobre una misma UE, los datos podrían estar clasificados por el factor tirmpo de medición o por el factor sujeto. Este tipo de estructuras sobre la respuesta son objeto de estudio en cursos de estadística avanzada.

En este capítulo, se introducen dos modelos de ANAVA particulares: (a) el modelo del ANAVA para un diseño en bloques completos al azar que responde a una estructura particular de UE, y (b) el modelo del ANAVA para un diseño bifactorial que responde a una estructura particular de tratamientos.

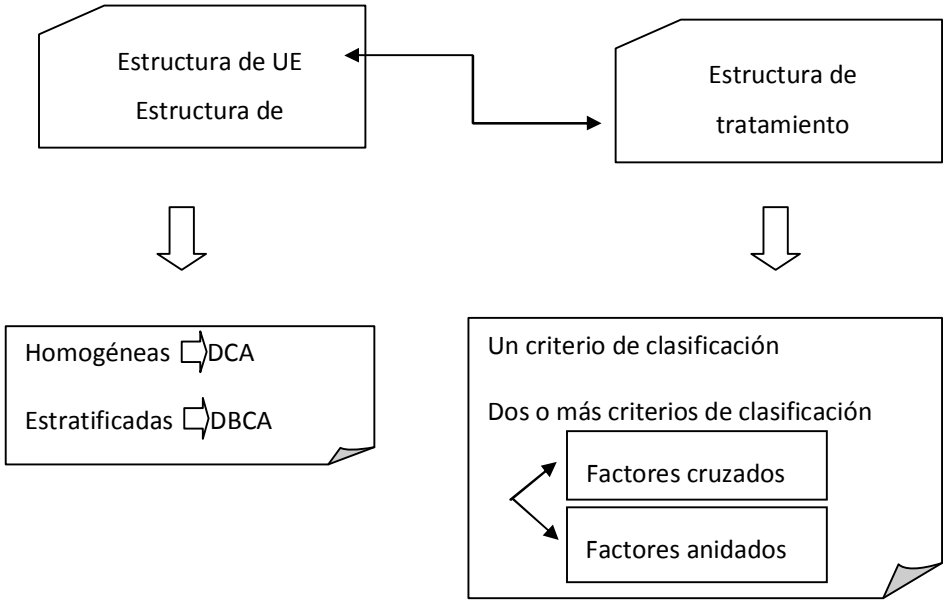


Figura 10.1: Estructuras presentes en un Diseño Experimental

Diseño en Bloques Completos al Azar

Si la UE disponibles para realizar un experimento no son homogéneas, se debe reconocer el o los factores que las hacen heterogéneas de manera que la variabilidad en la respuesta inducida por tal heterogeneidad no se confunda con la variabilidad experimental. Cuando las UE no son homogéneas, pueden no reaccionar o responder a los tratamientos de la misma manera o con la misma capacidad debido a sus diferencias intrínsecas.

Estas fuentes de variación sistemática, que se reconocen en el momento de planificar el estudio, deben ser contempladas en el diseño del experimento y en el análisis de los datos para disminuir el error experimental. Este hecho implica que se debe reconocer *a priori* la estructura presente en las UE.

La forma tradicional de controlar la variación del material experimental en experiencias planificadas es formando grupos o **bloques de UE homogéneas**. Los bloques de UE se construyen de manera tal que las unidades experimentales dentro de un bloque, varíen menos entre sí que UE en distintos bloques. El principio que subyace un bloqueo eficiente es **homogeneidad dentro del bloque y heterogeneidad entre bloques.** Por ejemplo: en el siguiente esquema se observa que las UE (parcelas del lote) podría variar debido a un efecto 'sombra' sobre el

terreno que ocasiona la cortina forestal; el criterio de bloqueo será entonces el nivel de sombra que recibe la parcela y los bloques se dispondrán de manera tal que las parcelas en un mismo bloque sean "homogéneas" respecto al criterio de bloqueo, es decir tengan un nivel de sombreo similar. Cada bloque en el esquema siguiente es un conjunto de tres parcelas con niveles de sombreo similar. Así si se quieren comparar tres tratamientos, estos se asignarán a las parcelas de un mismo bloque de manera aleatoria. En cada bloque se repetirá el proceso de aleatorización.

<luminosidad (+ sombra) >luminosidad (- sombra)

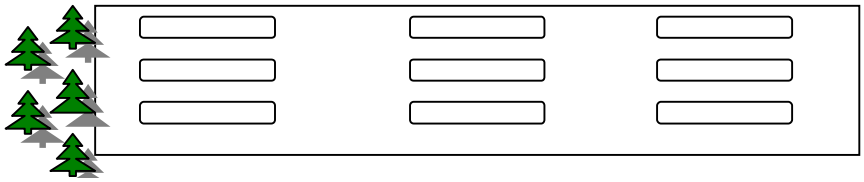


Figura 10.2: Esquema de localización de parcelas en un diseño en bloques con tres repeticiones, ubicadas de iquierda a derecha en el terreno experimental

En síntesis, reconocidos los grupos de UE homogéneas, los tratamientos, de ser posible, se comparan dentro de cada bloque. Si todos los tratamientos se disponen en un bloque, es decir si el bloque tiene tantas UE como tratamientos, el diseño será en bloques completos. Si la asignación de los tratamientos a las UE del bloque se hace al azar, entonces un diseño que reúne todas las características expuestas se denomina **Diseño en Bloques Completos al Azar (DBCA).**

Con el DBCA se pretende eliminar del error experimental de la variabilidad debida al factor de estratificación o bloqueo, esto disminuye los errores de estimación y aumenta la precisión de las comparaciones de las medias de tratamientos.

Los criterios de bloqueo pueden deberse no sólo a las características relacionadas con las unidades experimentales sino también, en algunas circunstancias, a aspectos ligados con la colecta de información o la realización de los tratamientos. A las características relacionadas con las UE se las denomina naturales mientras que al resto se las llama inducidas. Por ejemplo, si tenemos un conjunto de UE homogéneas pero algunos subgrupos de este conjunto son manejados por distintos operarios, o a distintos tiempos, el factor operario y el factor tiempo pueden introducir una fuente de variación en la respuesta (inducida). En este caso sería apropiado que cada operario trabaje con todos los tratamientos a comparar, o que si el experimento se lleva a cabo en varios días o momentos de tiempo, que en cada día se releve el dato de una repetición por tratamiento. Entonces, si contamos con 5 días para evaluar un ensayo donde hay 15

parcelas que han sido tratadas con 3 fertilizantes foliares, sería más recomendable en cada día evaluar tres parcelas, una para de cada tratamiento de fertilización, que evaluar repeticiones de un mismo tratamiento en un día y repeticiones de otro en otro día. Si hacemos esto último, y hay algún efecto del día de medición (supongamos un día de mucha más temperatura que otro), el efecto día quedará confundido con el efecto tratamiento. El bloqueo de UE pretende disminuir el **confundimiento** de factores.

DBCA: los tratamientos son asignados según la estructura de parcelas de manera tal que cada tratamiento aparezca una vez en cada bloque, todos los tratamientos estén en todos los bloques y la aleatorización de los tratamientos a las UE se realice dentro de cada bloque.

Las unidades experimentales que conforman un bloque no necesariamente deben ser adyacentes. Por ejemplo, cuando se comparar cultivares y se dispone de parcelas en la loma de un terreno, otras a una altimetría media y otras en un bajo. Las diferencias del suelo debidas a la topografía podrían afectar la respuesta. Entonces sembraremos todos los cultivares en la loma, todos en el medio y todos en el bajo. Habrá tres bloques o repeticiones definidas por el factor topografía, y en cada bloque estarán todas los tratamientos (cultivares). En caso contrario (algunos cultivares solo están en la loma y otros sólo en el bajo), el efecto cultivar se podría confundir con el efecto topografía. A continuación se muestran dos diseños que se condujeron siguendo un arreglo de bloques completos al azar (DBCA), con tres repeticiones para evaluar tres tratamientos, es decir un total de nueve UE (Figura 10.3). Previo a la aplicación de los tratamientos, el suelo del lote de ensayo fue monitoreado intensivamente a través de determinaciones de conductividad eléctrica y elevación, obtenidas con maquinaria de precisión, con las que se logró un mapa de variabilidad espacial. En la Figura de la derecha los bloques se dispusieron mejor que la de la izquierda ya que se observa mayor homogeneidad de las parcelas dentro de cada bloque, respecto al mapa de variabilidad de suelo.

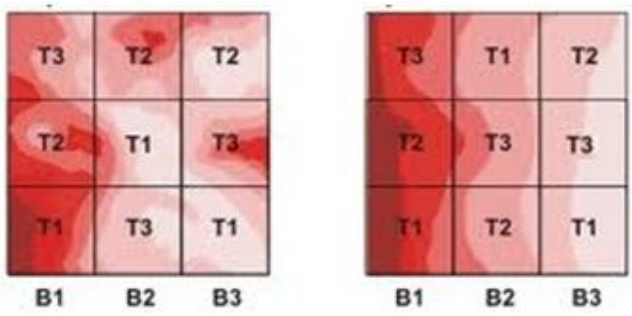


Figura 10.3: Esquema de localización de parcelas en dos diseños en bloques con tres repeticiones o bloques (B1, B2 y B3)

El control experimental debe ser realizado apropiadamente: 1) tratamientos asignados al azar a las unidades experimentales para neutralizar los efectos de factores no controlados, 2) tratamientos repetidos para poder estimar el error experimental y 3) estructura de unidades experimentales controlada (bloqueo si es necesario).

Cuando el número de tratamientos es dos, el DBCA es análogo al diseño de muestras apareadas para comparar la media de dos poblaciones ya que en cada caso de análisis o repetición se aplican y comparar los dos tratamientos.

Analisis de la varianza para un DBCA

El modelo para analizar un diseño en bloques completamente aleatorizados, es:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

donde: Y_{ij} es la respuesta del i-ésimo tratamiento en el j-ésimo bloque

μ es la media general

τ_i es el efecto del i-ésimo tratamiento i = 1, ...,a

β_j es el efecto del j-ésimo bloque j = 1, ...,b

ε_{ij} es el término de error aleatorio.

Si se puede suponer que existe **aditividad bloque-tratamiento** que significa NO interacción entre los bloques y los tratamientos y que los ε_{ij} son independientes e

idénticamente distribuidos N(0, σ^2) puede obtenerse una prueba F para la hipótesis de igualdad de medias de tratamientos como se hizo en el DCA.
Las hipótesis que se somete a prueba en un ANAVA para un DBCA, como en el DCA a una vía de clasificación, y está establecida sobre la medias de las poblaciones relacionadas a cada tratamiento ($\mu_i = \mu + \tau_i$ con $i = 1, \ldots, a$):

$H_0 : \mu_1 = \mu_2 = , \ldots , = \mu_a$

H_1 : Al menos un par de medias poblacionales difiere

Algebraicamente, en el contexto del ANAVA, existe una forma conveniente de expresar la magnitud de la variabilidad debida a los bloques en el contexto de las otras fuentes de variación intervinientes:

SCTotal = SCtratamiento + SCbloque + SCerror

Es decir que la suma de los desvíos cuadrados de cada observación con respecto a la media general puede ser particionada en tres sumas de cuadrados, una indicadora de las diferencias entre tratamientos: **Suma de Cuadrados de Tratamientos** (SCtratamiento), otra de la diferencia entre bloques: **Suma de Cuadrados de Bloques** (SCbloque) y otra que expresa la variación aleatoria de unidades experimentales que recibieron el mismo tratamiento después de descontar las variaciones debidas a las diferencias entre bloques, es decir el error experimental: **Suma de Cuadrados del Error** (SCerror). Si las diferencias entre unidades experimentales debidas al factor de bloqueo no son considerada, es decir si omitimos el efecto bloque en el modelo, la Suma de Cuadrados de Bloques se adiciona a la Suma de Cuadrados del Error. Así, el error experimental aumenta y como consecuencia se pierde eficiencia en la prueba de la hipótesis de interés. Los resultados del ANAVA también se presentan en una tabla igual al DCA, excepto que debido al bloqueo de las UE habrá una fila de la tabla indicando la variabilidad de la respuesta entre bloques.

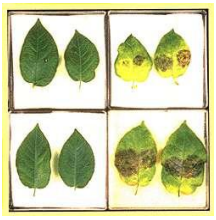
La comparación entre las medias de bloques, en general, no es de interés:

1- porque por construcción se espera que sean diferentes

2- porque en general no se asocian con cuestiones de interés, sólo responden a un factor que se debe controlar, es decir a una estrategia para evaluar los tratamientos en forma más precisa. Pero el principal interés recae siempre en la comparación de tratamientos.

3- porque la aleatorización fue realizada solo dentro de los bloques. Tal restricción de aleatorización hace que el estadístico construido entre CMBloque y CMError no siga una distribución F teórica. No obstante, el cociente puede ser usado para realizar sugerencias sobre la necesidad de bloqueo en experiencias futuras similares a la realizada.

Como se presentó para el modelo de ANAVA correspondiente a un DCA, los valores ajustados o predichos por el modelo permiten calcular los residuos que se usarán para evaluar el cumplimiento de los supuestos que sustentan al ANAVA clásico.



Aparte de los supuestos que aprendimos a evaluar en el contexto de un DCA, en el DBCA hay otro supuesto: la estructura de parcelas no debe interactuar con la estructura de tratamientos, es decir el efecto de los bloques debe ser aditivo al de los tratamientos.

El supuesto de no interacción bloque-tratamiento, implica decir que si un tratamiento es mejor que otro, esta relación entre ellos debe estar presente en todos los bloques. De no ser así, sería engañoso hacer recomendaciones acerca de los tratamientos en forma independiente a los bloques. Podemos recurrir a métodos de control del supuesto de **aditividad bloque-tratamiento** usando gráficos de líneas para representar la respuesta para cada nivel del factor tratamiento para cada uno de los bloques separadamente. Si existe aditividad las líneas dibujadas serán paralelas, en caso contrario habrá cruzamientos de las líneas (interacción o falta de aditividad bloque-tratamiento.

Aplicación

DBCA en ensayo comparativo de variedades de trigo

Para evaluar la adpatación y potenciales de rendimientos de un conjunto de variedades bajo las condiciones de clima y suelo de una región, es común que se implementen ensayos comparativos de rendimiento. En el ensayo usado en esta ilustración se compararon 10 variedades de trigo en un DBCA con 3 repeticiones, una de las variedades es la variedad comercial (testigo) de mayor difusión en la región y las otras 9 son variedades que se pretenden introducir comercialmente porque se supone superan a la variedad testigo. Los datos se encuentran en el archivo [trigo].

A continuación se presentan los resultados obtenidos luego de seleccionar a la variable “Rendimiento” como dependiente, al factor bloque (factor de control) y al factor variedad (factor tratamiento) como criterios de clasificación en el Menú de ANAVA de InfoStat.

Cuadro 10.1: ANAVA para un DBCA donde el factor "Bloque" representa el factor de control experimental y el factor "Variedad" el tratamiento

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	30	0.92	0.87	5.33

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	6557027.13	11	596093.38	19.36	<0.0001
Bloque	259665.00	2	129832.50	4.22	0.0315
Variedad	6297362.13	9	699706.90	22.72	<0.0001
Error	554237.67	18	30790.98		
Total	7111264.80	29			

Test:LSD Fisher Alfa=0.05 DMS=301.00661

Error: 30790.9815 gl: 18

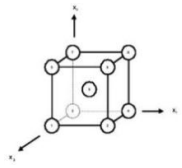
Variedad	Medias	n	E.E.			
V2	2504.00	3	101.31	A		
V6	2504.33	3	101.31	A		
Testigo	3066.33	3	101.31		B	
V1	3066.67	3	101.31		B	
V3	3473.00	3	101.31			C
V7	3474.33	3	101.31			C
V4	3645.00	3	101.31			C
V8	3646.33	3	101.31			C
V9	3760.67	3	101.31			C
V5	3761.33	3	101.31			C

Medias con una letra común no son significativamente diferente(p<= 0.05)

Se observa que los criterios de ajuste del modelo son buenos, que existe poca variabilidad residual, que el modelo explica alto porcentaje de la variabilidad en los datos de rendimiento (92%). Al menos una variedad muestra diferencias estadísticamente significativas (P<0,0001) respecto a las otras en lo que se refiere al promedio de sus rendimientos. La prueba LSD muestra que l rendimiento logrado con las variedades V2 y V6, fueron estadísticamente inferior al obtenido con el testigo comercial, que la variedad V1 no se diferenció estadísticamente del testigo y que las restantes variedades sí superan estadísticamente el rendimiento del testigo comercial bajo las condiciones ambientales del ensayo. El valor p en la fila en la que se encuentra el efecto de bloque sugiere que fue oportuna la decisión de usar un DBCA ya que las diferencias de rendimientos de distintos bloques no fueron menor.

Diseño con estructura factorial de tratamientos (Bifactorial)

El uso de experimentos factoriales se realiza cuando se reconoce la existencia de una estructura de tratamientos. Cuando se cruzan dos factores para definir un tratamiento (diseño bifactorial) las diferencias de la respuesta en relación a los niveles de cada uno de los factores se denominan efectos principales y las diferencias de los efectos de un factor entre distintos niveles del otro se denominan efectos de interacción entre factores. La presencia de interacción significativa señala cambios en las diferencias observadas bajo los niveles de un factor entre distintos niveles del otro factor. Cuando se cruzan niveles de varios factores para conformar un tratamiento, el experimentador se pregunta si es posible identificar los efectos de cada uno de los factores por separado (efectos principales) y eventualmente probar hipótesis también sobre la interacción entre los factores.



Entonces, los experimentos con arreglo factorial de tratamiento permiten responder a la siguiente pregunta: Las variaciones en la respuesta debidas a los efectos de un factor son independientes de los niveles del otro factor? Hay interacción entre factores o no?

Los modelos factoriales se conocen como **modelos de efectos aditivos** si los términos que modelan la interacción están ausentes y como **modelo con efectos multiplicativos de interacción** si además de los efectos principales de cada uno de los dos factores se adiciona un término que se refiere al efecto que surge del producto de los dos (interacción).

Modelo aditivo para un diseño bifactorial bajo un DCA

El modelo para un experimento con estructura factorial de tratamientos definida por dos factores cruzados, sin estructura de parcelas, es decir siguiendo un diseño completamente aleatorizado para asignar los tratamientos a las UE, y suponiendo falta de interacción (modelo aditivo) es el siguiente:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \text{con i=1,...,a; j=1,...,b}$$

donde Y_{ij} representa la respuesta al i-ésimo nivel del factor A y j-ésimo nivel de factor B, μ representa una media general, α_i el efecto que produce el i-ésimo nivel del factor A (con a niveles), β_j corresponde al efecto del j-ésimo nivel del factor B (con b niveles) y ε_{ij} es el término de error aleatorio asociado a la observación ij-ésima que como siempre se supone es una variable aleatoria normal, con esperanza cero y varianza σ^2.

Si el supuesto de aditividad (no interacción) no se cumple entonces el experimento está deficientemente diseñado ya que harían falta repeticiones de los tratamientos (combinación de los niveles de ambos factores) para inferir sobre efectos de interacción.

La tabla del ANAVA para un bifactorial tiene dos filas en lugar de una (como en el DCA a un criterio de clasificación) para evaluar los tratamientos. Cada fila se asocia a un factor tratamiento. Si el modelo es aditivo, la interacción no está presente. No obstante lo más frecuente es que también haya un termino en el modelo (y por tanto una fila en la tabla de ANAVA) para el factor interacción.

Aplicación

Diseño bifactorial sin repeticiones

Para ejemplificar una situación donde hay dos factores de interés y no existen repeticiones para cada tratamiento definido por la combinación de éstos se presenta un experimento factorial en el que es de interés estudiar los factores cepa usada en la inoculación de alfalfa con tres niveles y el factor cultivar de alfalfa con cinco niveles en la producción de forraje.

Supongamos que los 3×5=15 tratamientos resultantes se asignan a las UE (parcelas) según un diseño completamente aleatorizado. Se conoce por experiencias previas (o se supone) que no hay interacción entre los efectos de cepa y cultivar y por tanto el efecto de interacción no se incluirá en el modelo de análisis. Los factores se han designado como C (cepa) y CV (cultivar). Los 15 tratamientos de interés surgen del cruzamiento de ambos factores, es decir cada nivel de un factor se asocia con cada uno de los niveles del otro. En este experimento, cada uno de los tratamientos se evaluó una sola vez, es decir los tratamientos combinatoriales no están repetidos. No obstante esto, existen repeticiones para cada nivel de un factor si éste se observa a través de los niveles del otro. La variable observada es el rendimiento. Los datos están en el archivo [Alfalfa]. Se presenta a continuación los resultados obtenidos mediante el ANAVA de InfoStat, luego de haber seleccionado al Rendimiento como variable respuesta o dependiente, y a los factores “Cepa” y “Cultivar” como criterios de clasificación.

Cuadro 10.2: ANAVA de un experimento con DCA y dos factores sin interacción.

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	18	0.90	0.83	3.77

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	3563310.32	7	509044.33	12.96	0.0003
Cepa	291483.78	2	145741.89	3.71	0.0623
Cultivar	3271826.55	5	654365.31	16.66	0.0001
Error	392669.66	10	39266.97		
Total	3955979.99	17			

Test:LSD Fisher Alfa=0.05 DMS=360.50411

Error: 39266.9665 gl: 10

Cultivar	Medias	n	E.E.			
CV2	4503.83	3	114.41	A		
CV6	5065.76	3	114.41		B	
CV1	5068.32	3	114.41		B	
CV3	5472.53	3	114.41			C
CV4	5644.88	3	114.41			C
CV5	5761.36	3	114.41			C

Medias con una letra común no son significativamente diferentes(p<= 0.05)

Se concluye que hay efecto de cepa solo marginalmente (p=0,06); este efecto es significativo si se trabaja con un alfa del 10% pero no si se trabaja con un alfa del 5%. Por el contrario, si existen claras evidencias de efecto de cultivar o genotipo (p=0,0001). En el caso del factor cepa, al no ser significativo para el nivel de significancia que fijamos *a priori*, no se realizan pruebas de comparaciones múltiples. Para el factor cultivar, por tener cinco niveles y un valor p que sugiere que al menos un cultivar difiere estadísticamente de los otros, se necesita indagar más. Esto se puede realizar haciendo comparación múltiples de medias a posteriori del ANAVA. Se solicitó una prueba LSD de Fisher para conocer cuál o cuáles de las medias de cultivar son diferentes. En el siguiente gráfico se visualiza la diferencia promedio entre CV, como así también la posible interacción entre los efectos de cepa y cultivar. No obstante, por la falta de repeticiones en el ensayo, este efecto de interacción no puede evaluarse estadísticamente, es decir no podemos decir si la interacción que se observa en la figura es azarosa o se puede atribuí a un patrón real de diferencias entre cepas que cambian con los cultivares.

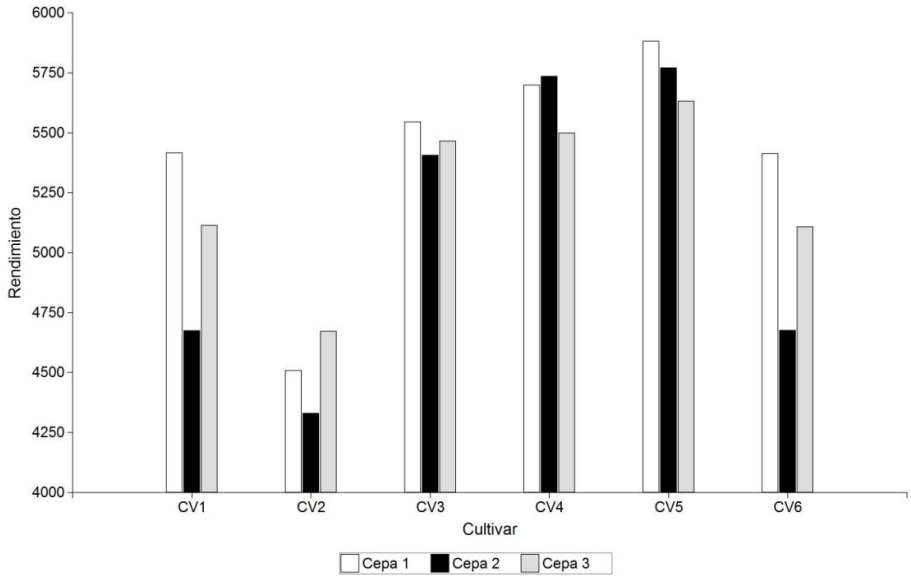


Figura 10.4. Rendimiento según tratamientos definidos por la combinación del cultivar usado y la cepara de la inoculación recibida.

Arreglos factoriales con interacción

Si el experimentador supone o sospecha que la respuesta a dos o más factores además de involucrar la suma de los efectos individuales de esos factores depende de la combinación específica de los niveles de éstos, entonces el modelo para el experimento factorial deberá incluir términos de **interacción** que den cuenta de este hecho.
Por ejemplo, en la evaluación del fenotipo o expresión de un ser vivo (persona, animal, planta) se supone que existen dos factores con efecto principal: el Genotipo (es decir el conjunto de sus genes) y el Ambiente. No obstante, los modelos utilizados para explicar variaciones fenotípicas no se encuentran completos sino se adiciona el término de interacción Genotipo*Ambiente.

Existen numerosos ejemplos que dos individuos con igual genotipo pueden mostrar expresiones fenotípicas bien diferentes si se desarrollan en ambientes distintos. Es la combinación específica del factor Genotipo y del factor Ambiente, la que define la expresión del carácter observado.

La inclusión de términos de interacción en el modelo conlleva la necesidad de tener repeticiones para cada tratamiento porque de otra forma no es posible estimar los parámetros adicionales y evaluar desde un ANAVA la significación estadística de la interacción. Cuando el experimento tiene dos factores, existen solo interacciones de primer orden, cuando tiene tres factores, existen interacciones de primer y de segundo orden y así los órdenes de la interacción siguen creciendo para arreglos factoriales con mayor número de factores.
El modelo lineal para un **experimento bifactorial** con interacciones es una ampliación del modelo para el experimento bifactorial de **efectos aditivos**, bajo un DCA, se expresa como:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

con i=1,...,a; j=1,...,b; k=1,..,nij

donde Yijk representa la respuesta en la k-ésima repetición del i-ésimo nivel del factor A y j-ésimo nivel de factor B, μ representa la media general, αi el efecto que produce el i-ésimo nivel del factor A, βj corresponde al efecto del j-ésimo nivel del factor B y los términos δij representan los efectos adicionales (interacciones) de las combinaciones de los niveles de los factores. Los términos de error εijk asociados a cada observación se suponen como es usual, normal e independientemente distribuidos con esperanza cero y varianza común σ^2. La tabla de ANAVA tendrá una fila extra, para evaluar la significancia de la interacción. En general, si esta resulta significativa se estudia la interacción y no los efectos principales de los factores. Mientras que si la interacción no es significativa se analiza el efectos de cada factor separadamente y en término de las medias de sus niveles.

Aplicación

DCA con estructura bifactorial de tratamientos y repeticiones

Las investigaciones en agricultura deben orientarse al desarrollo y aplicación de tecnologías que incrementen las fuentes primarias de alimento pero de manera social, económica y ambientalmente sustentable. La alimentación de la población mundial requiere cada vez más de un sistema de agricultura sostenible que pueda mantener el ritmo de crecimiento de la población. Los pronosticados aumentos de temperaturas y de lluvia hacen pensar que, en Argentina, seguirá avanzando la frontera agrícola, incrementándose la necesidad de cambios tecnológicos rápidos para no perder sostenibilidad. Las mayores escalas de producción agrícola, así como el incremento en el costo de la tierra y la necesidad de bajar el nivel de insumos destinados a la producción plantean fuertes motivaciones para la adaptación a la innovación tecnológica. La agricultura de precisión que habilita el manejo sitio específico de los lotes constituye un enfoque prometedor para favorecer una agricultura sostenible.

Análisis de experimentos con varios criterios de clasificación

Las nuevas tecnologías asociadas a la agricultura de precisión proporcionan la oportunidad de medir con mayor precisión la variabilidad espacial no sólo en el rendimiento sino también en propiedades de suelo. Para el manejo sitio-específico en los lotes, se reliza una delimitación de zonas dentro de los mismos que expresan una combinación relativamente homogénea de factores de rendimiento, y que consecuentemente pueden ser tratados diferencialmente, por ejemplo, algunos sitios podrían recibir dosis reducida de fertilizantes. Para poder hacer recomendaciones de dosis de fertilización según sitio en un cultivo, se realizaron las siguientes actividades: 1) delimitación de tres zonas homogéneas en base a variabilidad espacial de variables de suelo, 2) selección aleatoria de seis areas del lote de cada una de las tres zonas, 3) de las 6 áreas seleccionadas para cada ZM, dos seleccionadas al azar recibieron una dosis alta de nitrógeno, otras dos una dosis reducida a la mitad en su contenido de nitrógeno y otras dos se dejaron sin fertilización, 4) en cada una de las 18 áreas se obtuvo el rendimiento del cultivo. Los datos se encuentran en el archivo [Fertlizantes]. El ANAVA arrojó los resultados que se muestran en la tabla de salida de InfoStat.

La interacción entre los factores Zona de Manejo y Nivel de Fertilización resultó significativa ($p<0,0001$) razón por la cual no se estudian los efectos principales de los factores a través de las medias de todos los datos. Es necesario estudiar o "abrir" la interacción, esto es estudiar los efectos de un factor dentro de cada uno de los niveles del otro. En este ejemplo se analizaron las respuestas del cultivo bajo las distintas dosis dentro de cada Zona de Manejo con el objetivo de planificar el futuro manejo por sitio del lote. Los resultados sugieren que en las zonas clasificadas como BUENAS desde el mapa de variabilidad de suelo, es posible reducir la dosis de fertilizante a la mitad sin ocasionar cambios significativos en los niveles productivos.

Cuadro 10.3: ANAVA de un experimento con DCA y dos factores con interacción

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	18	0.98	0.96	0.91

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	3804100.78	8	475512.60	55.48	<0.0001
Zona	1730038.11	2	865019.06	100.93	<0.0001
Dosis	1205680.11	2	602840.06	70.34	<0.0001
Zona*Dosis	868382.56	4	217095.64	25.33	0.0001
Error	77131.50	9	8570.17		
Total	3881232.28	17			

Test:LSD Fisher Alfa=0.05 DMS=209.41959 Error: 8570.1667 gl: 9

Zona	Dosis	Medias	n	E.E.						
POBRE	Sin F	9374.50	2	65.46	A					
POBRE	Reducida	9538.00	2	65.46	A		B			
MEDIA	Sin F	9738.00	2	65.46			B			
MEDIA	Reducida	10111.50	2	65.46				C		
MEDIA	ALTA	10112.50	2	65.46				C		
BUENA	Sin F	10438.50	2	65.46					D	
BUENA	Reducida	10549.50	2	65.46					D	E
POBRE	ALTA	10616.50	2	65.46					D	E
BUENA	ALTA	10694.50	2	65.46						E

Medias con una letra común no son significativamente diferentes(p<= 0.05)

Una vez calculados los residuos se puede verificar el cumplimiento de los supuestos de normalidad, independencia y homogeneidad de varianzas de los términos de error mediante pruebas de hipótesis e interpretaciones gráficas como se ha explicado anteriormente. Estas pruebas usualmente se construyen reparametrizando el modelo factorial como un modelo a una vía de clasificación considerando el factor tratamiento que surge de la combinación de los factores originales. Aunque en los dos ejemplos anteriores se han presentado experimentos con **estructura factorial de tratamientos** donde los tratamientos se han dispuestos sobre las parcelas según un DCA, otras combinación de estructuras de tratamientos y estructuras de parcela son posible. Este hecho hace que existan una amplia variedad de arreglos o diseños experimentales. En el ejemplo que sigue se usará un modelo bifactorial pero donde los tratamientos se asignaron a las UE siguiendo un DBCA.

Aplicación

Ensayo para comparar calidad de embalaje

En un establecimiento agropecuario que embala productos perecederos es de particular importancia la resistencia de los embalajes. El material de embalaje es plástico termocontraible y los productos envasados deben pasar por un horno a cierta temperatura para lograr que el envoltorio plástico se contraiga. La empresa ha estado embalando los productos con un método tradicional que no le ha dado los resultados esperados.

Decide entonces evaluar nuevos materiales de embalaje. En el mercado le ofrecen 2 nuevos materiales (N1 y N2) que, a diferencia del tradicional, requieren circulación de aire al entrar al horno. La velocidad de circulación del aire depende del tamaño de los productos a embalar, por lo que se decide probar 3 velocidades distintas para el ventilador (1000, 2000 y 3000 rpm). De la combinación de los factores: material, con 2 niveles, y velocidad del ventilador, con 3 niveles, surge una estructura factorial con 6 tratamientos.

Se decide hacer 3 repeticiones para la experiencia, pero como no se puede realizar todo el ensayo en un solo turno de trabajo, se hace una corrida del experimento en cada uno de tres turnos, mañana, tarde y noche (M, T y N respectivamente). Si bien no interesa evaluar el factor turno, este se modela para descontar las posibles diferencias en la respuesta para cada uno de ellos, es decir se lo usa como factor de bloqueo. La variable que se mide para evaluar los tratamientos es la resistencia del embalaje, medida en una escala de 0 a 100. Los datos están en el archivo [Embalaje].

Estrategia de análisis

Se ajustará un ANAVA para un DBCA con estructura factorial de tratamientos, es decir una combinación de los modelos discutidos en este Capítulo. El modelo de análisis es:

$$Yijk = \mu + Materiali + Velocidad + Material*Velocidadij + Turnok + \varepsilon ijk$$

La forma de solicitar este modelo en InfoStat es seleccionando "resistencia" como *Variable dependiente*, Velocidad, Material y Bloque como *Variables de clasificación* y presionando *Aceptar*. En la ventana de diálogo del modelo, especificar la ecuación del modelo de la siguiente manera:

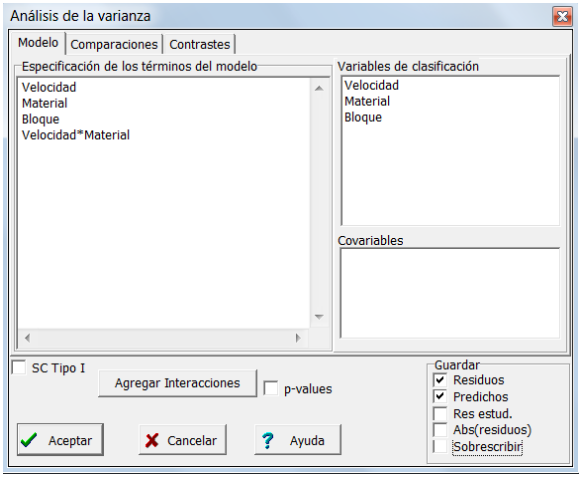


Figura 10.5. InfoStat. Ventana de Diálogo para especificar un modelo bifactorial-DBCA.

Luego del ajuste, una vez corrobarando el cumplimiento de los supuestos estadísticos del modelo a través del análisis de los residuos, se procederá a comparar las medias de los factores, es decir estudiar los efectos principales si no hay interacción significativa. Si la interacción Material*Velocidad resultase significativa se abrirá la interacción limitando las comparaciones de los efectos de un factor dentro de cada uno de los niveles del otro factor.

Cuadro 10.4: Resultados de un ANAVA para un diseño bifactorial en BCA Análisis de la varianza.

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Resistencia	18	0,96	0,93	13,60

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	13239,56	7	1891,37	31,23	<0,0001
Velocidad	1515,11	2	757,56	12,51	0,0019
Material	11150,22	1	11150,22	184,13	<0,0001
Bloque	19,11	2	9,56	0,16	0,8561
Velocidad*Material	555,11	2	277,56	4,58	0,0387
Error	605,56	10	60,56		
Total	13845,11	17			

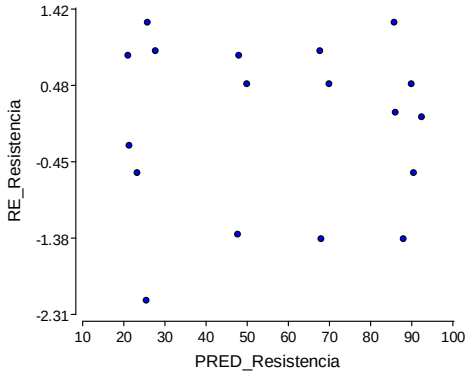


Figura 10.6. Residuos versus Predichos

El modelo representa un buen ajuste, tal lo muestra el gráfico de dispersión de residuos vs predichos, el valor relativamente grande del coeficiente de determinación y el valor pequeño de CV. El coeficiente de determinación R^2 es mayor al 90%, sugiriendo que el modelo ajustado explica un importante porcentual de la variabilidad total en los datos. No se observan diferencias entre bloques, por lo que se supone que no existen diferencias sistemáticas entre los turnos de trabajo.

La salida resultante del ANAVA sugiere la presencia de interacción estadísticamente significativa entre los factores Velocidad y Material (P=0,0387). Para estudiar la interacción se solicita en la solapa de comparación de medias, una prueba a posteriori (por ejemplo, LSD de Fisher) y se pide que se muestren las medias de la interacción y no las media de los efectos principales.

Cuadro 10.5: Comparación de medias de tratamientos definidos por la combinación del factor Velocidad y el factor Material. Prueba LSD de Fisher para la resistenciadel embalaje como variable dependiente

Test:LSD Fisher Alfa=0.05 DMS=14.15708
Error: 60.5556 gl: 10

Velocidad	Material	Medias	n	E.E.				
1000	N2	22.00	3	4.49	A			
2000	N2	26.33	3	4.49	A			
3000	N2	48.67	3	4.49		B		
1000	N1	68.67	3	4.49			C	
3000	N1	86.67	3	4.49				D
2000	N1	91.00	3	4.49				D

Letras distintas indican diferencias significativas(p<= 0.05)

Para visualizar la interacción es común realizar gráficos de barras de la respuesta en función de un factor como eje X distintas particiones de los datos producidas por el segundo factor de interés. En este ejemplo, mostramos la resistencia de los distintos

materiales para las distintas velocidades. Se observa que el material N1 es el de mayor resistencia promedio y su dependencia respecto a la velocidad no es lineal; no existen diferencias estadísticamente significativas (P>0,05) entre 2000 y 3000 rpm pero sí con 1000 donde se observa una menor resistencia para este material. La relación con la velocidad no es la misma para el material N2, donde no se encuentran diferencias entre 1000 y 2000 rpn y recién con 3000 rpm se incrementa la resistencia. Más allá de la presencia de interacción, el gráfico muestra que el nivel medio de la resistencia es diferente entre materiales.

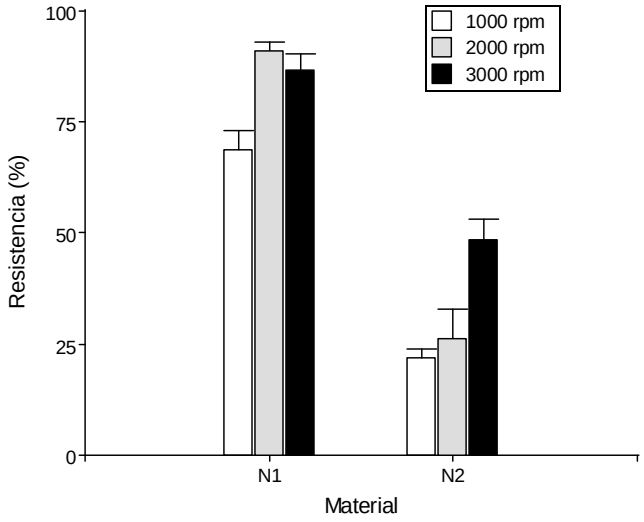


Figura 10.7. Residuos versus Predichos

Conclusión

Si bien la hipótesis sobre efecto turno no es de interés y por las restricciones a la aleatorización que implica el hecho de que los tratamientos se asignaron al azar dentro de cada turno la prueba F para turno no es válida. Se recomienda el uso del material N1 con la velocidad 2000 ya que esta velocidad (con este material) permite obtener la mejor de las resistencias, siendo este valor no diferente al obtenido con más rpm y por tanto más trabajo.

Otros caminos por recorrer en la modelación estadística

Los Agrónomos estamos acostumbrados a trabajar con modelos estadísticos para variables continuas y estudios experimentales, como son los modelos de **ANAVA** y **regresión** presentados en este libro. No obstante, es importante resaltar que el modelo estadístico refleja un proceso generador que no puede generar datos con distintas

características que los datos relevados. Si esto sucediera, las inferencias basadas en un modelo alejado de los datos no resultarán confiables.

La idea es construir modelos a partir de una clase de modelos que representen apropiadamente el proceso generador de datos y la naturaleza de los datos disponibles. Debido a la complejidad de los fenómenos aleatorios de origen biológico, la Bioestadística se expande continuamente en lo que se refiere a tipos o clases de modelos que se podrían ajustar a un conjunto de datos biológicos. También crece la disciplina a nivel de métodos de estimación de los parámetros del modelo para tales clases.

Por ejemplo, hemos aprendido que en los modelos de efectos fijos existe una única **componente aleatoria**, que denominamos el término de error, que permite ajustar las diferencias entre los valores observados y aquellos predichos por el modelo. Para esa componente aleatoria es necesario especificar las características de la distribución de probabilidad asociada. Los efectos de los **parámetros** son constantes fijas y atribuibles a un conjunto finito de niveles de un factor, que ocurren en los datos y sobre los cuales se desea hacer inferencia. Bajo los supuestos del modelo de muestreo ideal, las tablas de ANAVA basadas en mínimos cuadrados ordinarios proveen el método natural para las estimaciones de interés en el marco de los modelos de efectos fijos como los presentados.

Pero, este tipo de modelos ¿es suficiente para atender una adecuada representación de la realidad en todo momento? ¿Porqué siempre considerar a los efectos de los factores como constantes fijas?

La respuesta a ambas preguntas es: los modelos que hemos aprendido en este curso introductorio son sólo algunos de los que conforman el cuerpo conceptual de la Bioestadística actual.

Por ejemplo, a veces es necesario o conveniente considerar a un factor como aleatorio. Supongamos que 15 operarios que están trabajando en una plantación frutal son seleccionados al azar desde cada una de tres lotes de un establecimiento agropecuario los cuales pueden ser diferentes en cuanto a la dureza del suelo. Se registra la variable profundidad del hollado que realizan para la plantación sobre 5 hoyos producidos por la misma persona. Uno de los objetivos del estudio es comparar los tres lotes de plantación en estudio, vale decir se desea estimar y comparar los efectos de estos lotes. El factor lote se incorporará al modelo como un factor de efectos fijos. Sin embargo, también existe interés en conocer cuál es la variación de la profundidad del hoyado debida al operario que interviene en la producción del mismo. No se desea estimar y comparar los efectos de las personas que casualmente intervinieron en esta muestra. Sino que, suponiendo que ellos podrían proveer una estimación de la variabilidad debida al factor mano de obra, se desea estimar la magnitud de dicha fuente de variación. El factor operario se incorporará al modelo como un **factor de efectos aleatorios**.

Si se trabaja con un modelo de ANAVA con ambos tipos de efectos en el modelo, efectos fijos y aleatorios, entonces el modelo se llama **Modelo Mixto**. Asumiendo los efectos de operario como aleatorios, el interés del análisis también recaerá en la

estimación de la varianza de esos efectos. Luego, para modelar los datos de este ejemplo, consideramos que existen 2 criterios de clasificación, uno fijo y otro aleatorio y que por tanto el modelo contiene 2 fuentes aleatorias de variación: varianza entre operarios y varianza residual. Ambas explican la variación en la respuesta y por ello se conocen como componentes de varianza.
Bajo el **Modelo Lineal Mixto** (MLM), la varianza de la variable en estudio es la suma de estas las distintas componentes de varianza. En los MLM sólo es necesario sostener el supuesto de normalidad, pudiendo lograr estimaciones en casos de datos que no son independientes y/o en casos donde las varianzas no son homogéneas. La mayor flexibilidad del modelo mixto de ANAVA ha expandido, de manera importante, la selección de ésta técnica con respecto al ANAVA del modelo lineal general.

*El modelo de muestreo ideal conduce al ML clásico que tiene como supuestos la distribución normal, la heterogeneidad de varianzas (heterocedasticidad) y la independencia de los términos de error aleatorios. Bajo linealidad, cuando el supuesto de normalidad se puede sostener pero hay falta de homogeneidad de varianzas y/o independencia, cobran importancia los Modelos Lineales Mixtos (MLM). Debido al advenimiento de las técnicas computacionales y de cálculo numérico, actualmente se pueden también ajustar modelos lineales sin necesidad de asumir distribución normal (**Modelos Lineales Generalizados**, MLG). Por ejemplo, datos de una respuesta discreta es mejor usar en un MLG que un ML clásico. Si la tendencia a modelar es no linear, serpa más conveniente un modelos no lineales (MNL).*

La técnica de ANAVA y los métodos de estimación asociados (basados en Sumas de Cuadrados) han sido usados ampliamente para modelos lineales de efectos fijos con distribuciones normales. En muchas situaciones que se alejan de los supuestos del modelo de muestreo ideal, las tablas de ANAVA representan una sobresimplificación y una pérdida de información y eficiencia ya que no contienen los estadísticos suficientes. Otros procedimientos de estimación, como son aquellos basados en la función de máxima verosimilitud (MV o ML de sus siglas en Inglés), son preferibles en contextos donde no pueden sostenerse los supuestos de independencia y homogeneidad de varianza del modelo de muestreo ideal. Bajo normalidad, en la mayoría de los modelos de interés práctico, los estimadores MV proveen resultados analíticos. Bajo no normalidad, si bien es difícil obtener resultados analíticos, se obtienen estimadores por maximización numérica de la función de MV. El procedimiento de MV tiene la particularidad de ser un procedimiento general y eficiente (al menos cuando el tamaño muestral es grande). Una ventaja adicional de la estimación MV es que se puede trabajar tanto con datos balanceados como desbalanceados, ya sea con distinto número de repeticiones por celda o aún con celdas faltantes.
Estos comentarios se presentan para indicar que la Bioestadística es una disciplina en continuo desarrollo. Desde los protocolos que incluyen el diseño de un estudio

experimental u observacional hasta la elaboración de conclusiones se transitan numerosos caminos. Tanto en la etapa del análisis exploratorio de datos, que generalmente coincide con las primeras etapas descriptivas o cuantitativas de los estudios, como en la etapa de modelación estadística, frecuentemente reservada para estados más avanzados de las investigaciones, las posibilidades de análisis de datos son numerosas. La naturaleza de la variable y, más internamente, del proceso generador de los datos, define en gran medida la tecnología de información más apropiada para resolver un problema particular. Esta obra ha presentado métodos estadísticos clásicos, no obstante las posibilidades del análisis de datos en la práctica se extiende más allá de lo explora.

Ejercicios

Ejercicio 10.1: Los datos siguientes corresponden a un experimento realizado por Charles Darwin en 1876. En cada maceta se plantan dos brotes de maíz, uno producido por fertilización cruzada, y el otro por auto-fertilización. El objetivo era mostrar las ventajas de la fertilización cruzada. Los datos son las alturas finales de las plantas después de un período de tiempo, se encuentran en el archivo [Cruzamientos].

a) ¿Alguno de los dos tipos de maíz es demostrablemente mejor?

b) Si es así, ¿cómo se puede describir la diferencia?

Ejercicio 110.2: Se dan los tiempos de sobrevida (en unidades de 10 horas) de animales, sometidos a 3 tipos de veneno, y 4 tratamientos antitóxicos. Los datos se encuentran en el archivo [Veneno].

a) Describir la influencia de los dos factores en la sobrevida, analizando primero la existencia o no de interacción entre ambos.

Ejercicio 10.3: El siguiente conjunto de datos corresponde a proteína bruta en leche obtenida con dos suplementos (A y B) en dos dosis (1 y 2). Cada observación corresponde al contenido de proteína bruta en leche de una muestra compuesta obtenida por tambo.

Tambo	Control	A1	A2	B1	B2
I	3.19	3.03	3.06	3.22	3.33
II	3.16	3.07	3.08	3.28	3.20
III	3.25	3.23	3.24	3.45	3.45
IV	3.48	3.30	3.33	3.44	3.39
V	3.25	3.25	3.24	3.35	3.54
VI	3.10	3.05	2.93	3.28	3.35

a) Calcular la estadística descriptiva básica.

b) Identificar el modelo lineal para los datos anteriores.

c) Calcular la tabla de análisis de la varianza y, si corresponde, utilizar alguna técnica de comparaciones múltiples.

d) ¿Qué suplementación se recomendaría si el objetivo es maximizar la concentración de proteína bruta en la leche?

Análisis de experimentos con varios criterios de clasificación

Ejercicio 10.4: En la siguiente tabla se muestran los resultados de un experimento montado según un diseño completamente aleatorizado con cuatro repeticiones, en el que nemátodos de género Pratylenchus fueron criados en cuatro condiciones de temperatura y discriminados según sexo para evaluar el efecto del sexo y la temperatura sobre la expresión fenotípica de diversos caracteres morfométricos. Los resultados presentados corresponden al largo promedio de la cola en unidades experimentales conformadas por 5 individuos.

	Hembras				Machos			
Temp. (C)	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4
16	29.2	32.5	34.6	32.6	27.2	24.7	27.3	26.2
21	30.1	30.4	31.4	35.8	26.7	26.5	27.2	27.2
25	31.6	30.2	29.5	30.0	26.2	26.3	28.2	26.2
28	29.6	28.4	28.4	28.1	24.8	25.4	25.6	26.2

a) Identificar el modelo lineal para este experimento.
b) Representar gráficamente los valores medios según sexo y temperatura.
c) Construir la tabla de análisis de la varianza correspondiente.
d) Concluir sobre el efecto de la temperatura y el sexo sobre la expresión del largo de la cola y relacione sus conclusiones con la representación gráfica obtenida en ´b´.

Ejercicio 10.5: Considere el Ejercicio 10.4 suponga que debido al tamaño del experimento las repeticiones se realizaron en laboratorios diferentes. Considere que las repeticiones como bloques.

a) Identificar el modelo lineal para las observaciones de este experimento.
b) Construir una tabla de análisis de la varianza.
c) Concluir sobre la acción del sexo, la temperatura y su eventual interacción.

Ejercicio 10.6: Se realizó un experimento para estudiar el efecto de la cepa y del sustrato en la producción de un hongo comestible conocido como Gírgola (Pleorotus ostratus). Para la realización del ensayo se utilizaron bolsas del mismo material y en cada bolsa se colocó un tipo de sustrato en el que se sembró un tipo de cepa. Se evaluaron 3 cepas colocando cada una de ellas en cada tipo de sustrato. Los sustratos fueron: Paja de trigo + aserrín de álamo (PT-A), Paja de alfalfa + aserrín de álamo (PA-A) y Paja de trigo (PT). Se emplearon 4 bolsas por tratamiento evaluándose, al final del periodo de cultivo, el rendimiento en kg por bolsa. A continuación se presentan los resultados obtenidos con el análisis de la varianza y un gráfico construido para el problema:

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rend	36	0.72	0.64	11.16

Cuadro de Análisis de la Varianza (SC tipo III)

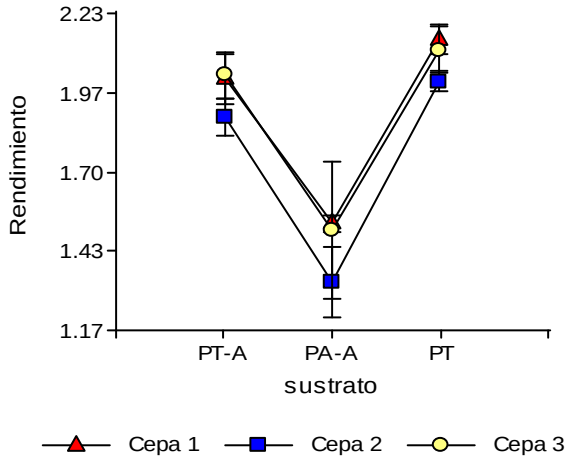
F.V.	SC	gl	CM	F	p-valor
Modelo	2.95	8	0.37	8.76	<0.0001
cepas	0.18	2	0.09	2.13	0.1381
sustrato	2.76	2	1.38	32.81	<0.0001
cepas*sust	0.01	4	2.2E-03	0.05	0.9944
Error	1.14	27	0.04		
Total	4.08	35			

Test:LSD Fisher Alfa=0.05 DMS=0.17180

Error: 0.0421 gl: 27

sustrato	Medias	n	E.E.		
PA-A	1.45	12	0.06	A	
PT-A	1.97	12	0.06		B
PT	2.09	12	0.06		B

Medias con una letra común no son significativamente diferentes(p<= 0.05)



Análisis de experimentos con varios criterios de clasificación

Asignar a cada una de las siguientes afirmaciones una V o una F según sea Verdadera o Falsa

Según el ANAVA se usó un modelo para un diseño completamente aleatorizado con arreglo factorial de tratamientos	
El gráfico indica una interacción significativa entre sustrato y cepa	
Los resultados muestran que no habría efecto del factor cepa	
Los resultados del ANAVA indican una interacción estadísticamente significativa entre los dos factores	
Con el sustrato paja de alfalfa + aserrín de álamo, se obtuvo el menor rendimiento promedio	
Para comparar los resultados de los distintos sustratos es necesario hacerlo dentro de cada cepa	
El efecto de cepa no se puede evaluar por presencia de interacción	
La cepa 2 produce un decrecimiento estadísticamente significativo del rendimiento respecto de al menos alguna de las otras cepas, independientemente del sustrato	

Capítulo 11

Ensayos multiambientales comparativos de rendimientos

Mónica Balzarini

11. Ensayos multiambientales comparativos de rendimientos

Motivación

Los datos provenientes de redes de ensayos comparativos, conducidos a campo en numerosos ambientes (ensayos multiambientales) son importantes en agricultura porque proveen conocimientos específicos del material vegetal disponible para cultivo y sus relaciones con los ambientes donde pueden producirse dentro de una región de interés. El término genotipo se refiere a un cultivar o a un híbrido. El término ambiente se relaciona al conjunto de climas, suelos, factores bióticos (plagas y enfermedades) y condiciones de manejo de un ensayo individual en una localidad determinada en un año. La exploración de patrones de interacción Genotipo*Ambiente, ofrece posibilidades, especialmente en la selección y adopción de genotipos que muestren interacción positiva con algunas localidades y sus condiciones ambientales prevalecientes (exploración de adaptación específica) o de genotipos con baja frecuencia de rendimientos pobres o fracaso del cultivo (exploración de estabilidad de rendimientos, adaptación en sentido amplio).

En este Capítulo se ejemplifica el análisis de una red de ensayos a partir de técnicas y métodos estadísticos que hemos aprendido en este curso. El objetivo de este Capítulo es ilustrar cómo se integra el uso de herramientas de análisis estadístico en un problema particular. Se ha seleccionado el análisis de redes de ensayos porque incluye conceptos de diseño de experimentos, particularmente diseño en bloques completos al azar y diseño factorial e ilustra el uso de gráficos presentados en el Capítulo 1, como los biplots y los diagramas de dispersión, a modo de herramientas complementarias. El problema agronómico que se aborda tiene que ver con la respuesta de una pregunta

importante para la producción vegetal: ¿qué material genético sembrar en un determinado ambiente?

Contexto del problema

Los cultivos de trigo, soja, girasol y maíz son los más importantes en el aporte a la sustentabilidad económica y biológica de los sistemas de producción agrícola en numerosos ambientes de la región centro de Argentina. Por ello, existe una oferta continua de nuevos cultivares y tecnologías de manejo para el área.

Las asociaciones de productores de la región, las empresas agropecuarias que cultivan una superficie importante del área de cultivo, los semilleros y otras empresas que proveen de material para la siembra y para la protección de los cultivos, así como las Universidades y el INTA en su rol de instituciones de investigación agropecuaria, se enfrentan continuamente al desafío de tener que recomendar tecnologías de producción de estos cultivos (cultivares o híbridos, esquemas de fertilización, manejo del agua, manejo del suelo, entre otras).

Las respuestas que se dan a cada productor se sustentan principalmente en la experimentación a campo de las nuevas tecnologías. En esta región, como en otras del país, se establecen anualmente numerosas redes de ensayos comparativos de rendimiento que permiten evaluar las distintas alternativas de producción en los ambientes explorados por los productores. Uno de los principales objetivos de las redes de ensayos multiambientales comparativos de rendimientos, es generar información que permita mejorar la toma de decisiones y evaluar el comportamiento de distintos materiales comerciales y precomerciales por su potencial y estabilidad de rendimiento.

Los efectos de la **interacción Genotipo*Ambiente** sugieren que las diferencias entre genotipos no son consistentes a través de los ambientes. La respuesta diferencial de los genotipos según el ambiente no deben ser ignorada, sino por el contrario analizada, usando las técnicas apropiadas, para explorar las ventajas y desventajas potenciales de la adaptación de los distintos genotipos en los ambientes de interés. La información provista por las redes de ensayos multiambientales permiten ganar conocimiento sobre el tipo y magnitud de la interacción Genotipo*Ambiente que se debe esperar en una región dada y así constituye una herramienta para establecer estrategias de manejo sitio-específcicas si fuere necesario.

La variable respuesta más común en redes de ensayos comparativos es el rendimiento, aunque en la práctica también se registran numerosas covariables para complementar los análisis de rendimientos. El diseño experimental más común en redes de ensayos comparativos es el DBCA dentro de cada ambiente. El término “Ambiente” suele estar asociado a distintas localidades y sitios de ensayos, a distintas fechas de siembra, a distintos años o campañas agrícolas o a la combinación de éstos. Las redes de ensayos comparativos son de distintos “tamaños”, no obstante es común disponer de 5 a 10 ambientes con 5 a 10 genotipos evaluados en cada ambiente, según un diseño con 2 o 3

repeticiones de cada genotipo en cada ambiente, comúnmente 2 o 3 bloques completos por ambiente.
La evaluación de redes de ensayo para el rendimiento del cultivo de interés comienza, como todo análisis estadístico, con gráficos descriptivos. Usualmente, se realizan gráficos de barras para indicar los rendimientos promedios de los distintos genotipos en cada ambiente. Cuando los genotipos y/o ambientes son numerosos, estos gráficos se realizan particionando la información por ambiente.

ANAVA a dos criterios de clasificación y BIPLOT

Seguido del análisis gráfico, suelen realizarse ANAVAs para cada ambiente independientemente. El objetivo de estos ANAVAs por ambiente es evaluar la calidad de los ensayos en los distintos ambientes. Los coeficientes de variación (CV) de los ensayos en cada ambiente son buenos indicadores de la calidad del mismo. Ensayos con CV mayores a 30-40% suelen ser descartados de la base de datos de la red. En redes de ensayos, es común que no todos los ensayos sean conducidos con igual precisión; muchas veces las personas involucradas con los ensayos no son las mismas.
Los ANAVA por ambiente también sirven para considerar si la precisión de los ensayos es similar, es decir si hay homogeneidad de varianzas residuales a través de los ambientes. Si esto ocurriese tiene más sentido realizar un análisis conjunto bajo el modelo clásico que cuando hay heterogeneidad de varianzas residuales. Generalmente, para que un ensayo se considere con menor precisión que otro su varianza residual (Cuadrado Medio del Error) debe ser tres o más veces mayor a la del ensayo considerado más preciso. Diferencias de varianzas residual de menor magnitud usualmente no invalidan las conclusiones obtenidas a partir del análisis conjunto de los datos bajo el supuesto de homogeniedad de varianzas.
La comparación de gentipos en redes de ensayo suele realizarse mediante modelos de ANAVA bifactorial (Genotipo y Ambiente son los factores) con interacción. Si el diseño experiental ha sido un DBCA dentro de cada ambiente, entonces el modelo de ANAVA debe incluir también el efecto de bloque anidado en el ambiente, ya que los bloques de un ensayo no son los mismos que los bloques de otro ensayo.
Por el rol principal que juega la interacción Genotipo*Ambiente, el término de interacción es de particular interés en los análisis de redes de ensayo. No sólo importa saber si es estadísticamente significativo o no, sino que también interesa saber (cuando resulta estadísticamente significativo) cuáles fueron los Genotipos y los Ambientes más responsables de la significancia estadística de la interacción. Es decir cuáles Genotipos y cuáles ambientes son los de mayor contribución a la componente de interacción.
Para estudiar la interacción, el efecto global de interacción suele descomponerse en uno, dos o más términos multiplicativos. Estos términos ponderan mediante scores de genotipo y scores de ambiente la contribución relativa de éstos en la explicación de la interacción. La descomposición del efecto de interacción se realiza vía Análisis de

Componentes Principales y por ello los resultados pueden viasualizarse en gráficos del tipo Biplot.

Estos modelos con efectos de Genotipo, Ambiente e interacción modelada vía ACP, suelen denominarse modelos lineales-bilineales. El nombre se debe a que el modelo para la respuesta del genotipo *i* en el ambiente *j* comprende una parte sistemática que involucra los efectos aditivos principales de genotipo y ambiente (componentes lineales) como así también uno o más términos multiplicativos para explicar patrones en el término de interacción Genotipo*Ambiente (componentes bilineales).

Comúnmente la parte aleatoria del modelo involucra al término de error y a la varianza residual del término de interacción, *i.e.* la parte de la interacción GE no explicada por el modelo multiplicativo. Proceduralmente, la estimación de los parámetros de interacción Genotipo*Ambiente en un modelo lineal-bilineal y para tablas de datos balanceadas (es decir cuando se tienen todos los Genotipos en todos los Ambientes) se hace por medio del Análisis de Componentes Principales de una matriz Z que contiene los residuos del modelo de ANAVA bifactorial aditivo, es decir luego de ajustar por el modelo de efectos principales. El análisis de esta matriz de residuos provee los scores de genotipos y ambientes respectivamente. Generalmente los dos primeros términos multiplicativos o componentes principales (CP1 y CP2) son suficientes para explicar los principales patrones de interacción; la variabilidad remanente en la matriz de efectos de interacción se interpreta como ruido o variabilidad no asociada a patrones significativos y por tanto repetibles de interacción.

Los primeros modelos lineales-bilineales usados en redes de ensayos agrícolas multiambientales fueron llamados modelos de efectos aditivos e interacción multiplicativa o modelos AMMI (del inglés, Additive Maineffects and Multiplicative Interaction) por Gauch (1988). Realizado el análisis de componentes principales, el biplot de la CP1 y CP2 es usado para identificar asociaciones entre genotipos y ambientes. Marcadores de genotipo con valores altos de CP1 sugieren que los rendimientos de estos genotipos se correlacionan positivamente con los ambientes que también tienen scores altos de CP1. Vale decir, el genotipo muestra alguna ventaja, relativa a los otros genotipos y a lo sucedido en otros ambientes, en ese ambiente. Los genotipos con valores altos de CP1 se correlacionan negativamente con ambientes con valores bajos de CP1.

Genotipos con valores cercanos a cero en la CP1 son interpretados como adaptados a los ambientes de prueba o de menor contribución en la interacción Genptipo*Ambiente, es decir más estables. Mientras más alta es la CP1, más interacción. Por ello, es común que luego del Biplot, también se presente una gráfica relacionando producción (medias de rendimiento por genotipo) y estabilidad (valores de CP1 promedio para cada Genotipo). Generalmente esta medida de estabilidad se expresa en escala estandarizada y al cuadrado, así es posible asignar valores umbrales para decidier si la interacción, medida a través de esta función de la CP1, sugiere que la inestabilidad es significativa o no.

Aplicación

Red de ensayos de Trigo

Se analizarán a modo ilustrativo ensayos que fueron conducidos en 5 ambientes correspondientes a distintas localidades del área de cultivo de trigo en el Sur de la Región Triguera. En cada ambiente se usaron dos repeticiones para cada una de 7 variedades de trigo usando un diseño de parcelas de bloques completos al azar para controlar el efecto de diferencias de altitud ("loma" y "bajo") que se observaron en cada sitio. Cada unidad experimental (parcela) tenía 6 metros de ancho y 200 mts de largo. Por las dimensiones de las unidades experimentales, se suele usar el nombre de **macroparcelas**. Este tipo de parcelas se usa comúnmente en ensayos a campo donde se evalúan materiales precomerciales con materiales comerciales usados como testigos y se desa cultivar a los genotipos en las condiciones habituales de trabajo del productor ya que el objetivo principal del ensayo es la recomendación de cultivares para el productor en su ambiente específico.
En el ejemplo que se presenta, se sembraron variedades de trigo de ciclo intermedio a largo. Las fechas de siembra y las prácticas culturales fueron las recomendadas en cada ambiente. Todos los lotes usados en esta red de ensayo habían sido cultivados con soja de primera como antecesor. De esta manera hay menos posibilidad de que el efecto del cultivo antecesor se confunda con efectos de cultivar. Todos los ensayos contaron con buena cantidad de agua útil para el cultivar al momento de la siembra. La macroparcelas se cosecharon con la maquinaria que usa el productor y se pesaron en monotolbas con balanza. Los datos de rendimiento de las distintas parcelas se corrigieron re-expresándolos a todos a un mismo valor de humedad (14 % = humedad comercial). Los datos se encuentran en el archivo *[Red]*.

Estrategia de análisis

Primero se realizaron gráficos de barras indicando el comportamiento promedio (a través de las repeticiones) de cada material en cada ambiente. Luego se realizó un ANAVA bajo un modelo que incluyó los efectos de Genotipo, Ambiente, Genotipo*Ambiente y el efecto de Bloque anidado dentro de cada ambiente. Este último término se indica en InfoStat con la sintaxis Ambiente>Bloque.
Posteriormente se ajustó un ANAVA sin interacción (modelo aditivo) y se guardaron los residuos. Se suponen que estos residuos miden no sólo el error experimental como en cualquier otro modelo estadístico sino también la interacción ya que ésta no se consideró al ajustar el modelo. Los residuos fueron primero promediados para tener sólo un valor por combinación de Genotipo y Ambiente y luego dispuestos en una matriz Z de tantas filas como genotipos y tantas columnas como ambientes. La matriz Z fue sometida a un ACP y se construyó un gráfico Biplot para visualizar los resultados del análisis de la interacción.

Finalmente, con la CP1 generada a partir del ACP de la matriz de residuos del modelo aditivo y las medias de Genotipos se realizó un gráfico de dispersión para analizar simultánamente estabilidad y producción de cada material evaluado. A este gráfico se le trazaron dos líneas de referencia: (1) a nivel de las ordenadas para indicar el rendimiento promedio y (2) a nivel de las abscisas para indicar la significancia estadística de la estabilidad o inestabilidad. Esta última se juzgó según el valor de una variable aleatoria Chi-cuadrado con 1 grado de libertad ya que los valores del eje corresponden al valor de la CP1 al cuadrado que teóricamente se distribuye como una Chi-Cuadrado con un grado de libertad (Chi-cuadrado=3,84). Valores superiores sugieren inestabilidad y valores menores estabilidad del genotipo a través de los ambientes. Por ende, si se buscan genotipos de altos rendimientos y baja inestabilidad ambiental, hay que observar cuáles son los genotipos situados más arriba y más hacia la izquierda de la gráfica.

Resultados y discusión

Las gráficas descriptivas anteriores muestran que se registraron diferencias entre cultivares en todos los ambientes, pero que estas diferencias cambian con los ambientes. Por ejemplo, la variedad IV con un desempeño relativamente bueno en los ambientes A, B, C y D resultó una variedad de pobre rendimiento en los ambientes E y F, que además fueron los ambientes en promedio mas pobres o de menor rendimiento. El ANAVA para el análisis conjunto de los ensayos de la red sugiere que la interacción Genotipo*Ambiente es estadísticamente distinta de cero (P=0,0002). Por tanto el análisis de los efectos principales de genotipo debiera postergarse hasta comprender mejor el fenómeno de interacción.

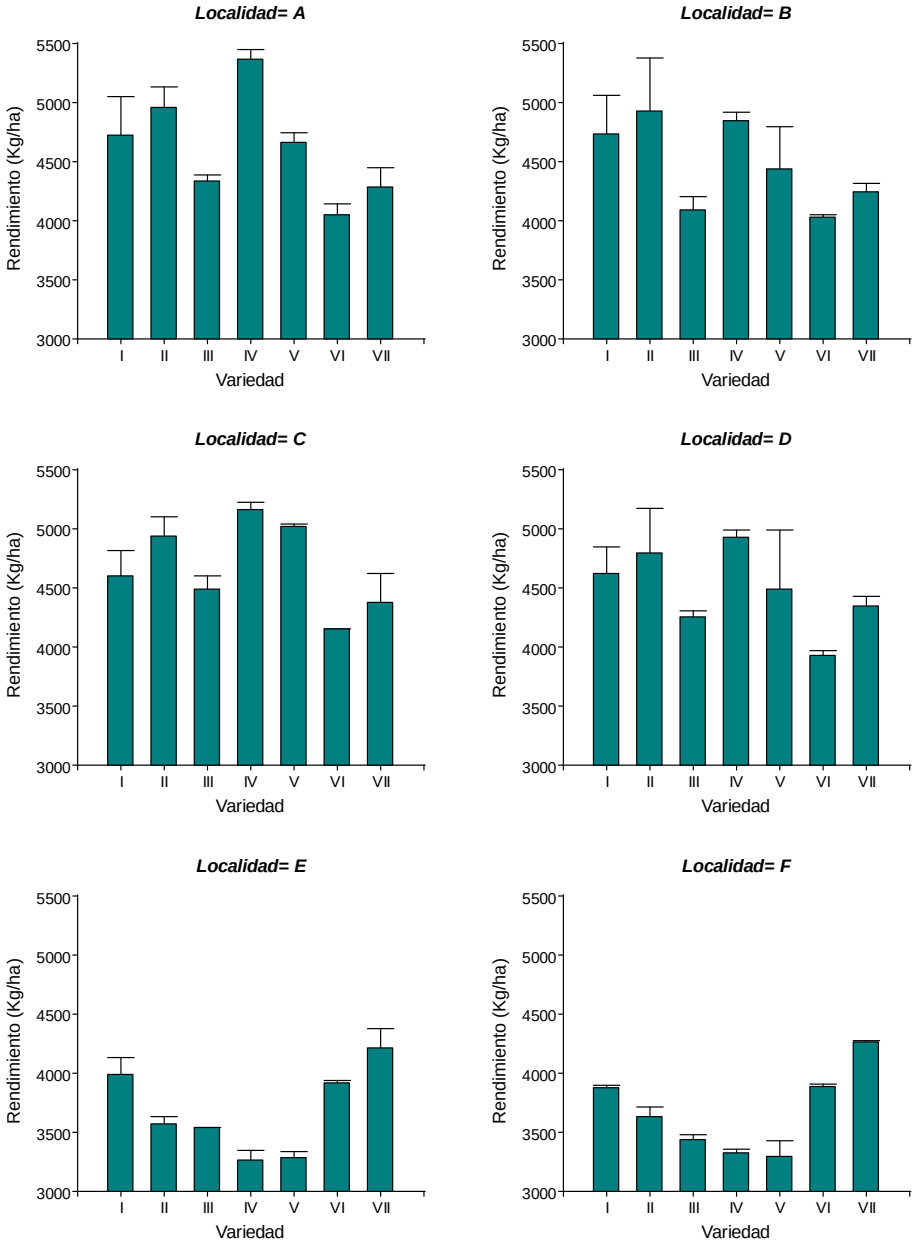


Figura 11.1. Medias de rendimiento (más E.E.) de 7 genotipos (Variedades I,II,III,IV,V,VI y VII) en 6 ambientes de la región de cultivo (A,B,C,D,E y F).

Cuadro 11.1. ANAVA para una red de ensayos comparativos de variedades de trigo conducidos bajo un DBCA en cada ambiente

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	84	0.91	0.80	6.05

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	25479618.95	47	542119.55	8.13	<0.0001
Localidad>Bl	249467.26	6	41577.88	0.62	0.7103
Localidad	15098297.85	5	3019659.57	45.28	<0.0001
Variedad	3002571.83	6	500428.64	7.50	<0.0001
Localidad*Variedad	7129282.02	30	237642.73	3.56	0.0002
Error	2400983.40	36	66693.98		
Total	27880602.36	83			

Test:LSD Fisher Alfa=0.05 DMS=213.82367
Error: 66693.9834 gl: 36

Variedad	Medias	n	E.E.			
VI	3995.28	12	74.55	A		
III	4026.42	12	74.55	A		
V	4199.98	12	74.55	A	B	
VII	4285.95	12	74.55		B	C
I	4424.99	12	74.55			C
II	4472.19	12	74.55			C
IV	4482.98	12	74.55			C

Letras distintas indican diferencias significativas(p<= 0.05)

La probabilidad de que las diferencias observadas en el compartamiento relativo de los genotipos en los distintos ambientes sean sólo por azar es baja (P=0,0002). Por tanto, la interacción se presupone que es un efecto repetible e interesa indagar sobre cuáles genotipos son los que más contribuyeron a la significancia de la interacción. La Figura siguiente es el Biplot de los efectos de interacción. Se observa que el cultivar IV en mayor medida, y luego el V y el II, se desempeñaron relativo a los otros mejor en los ambientes A, B, C y D que en los ambientes F y E. En estos dos ambientes los genotipos de mejor desempeño relativo respecto al rendimiento fueron los genotipos VI y VII.

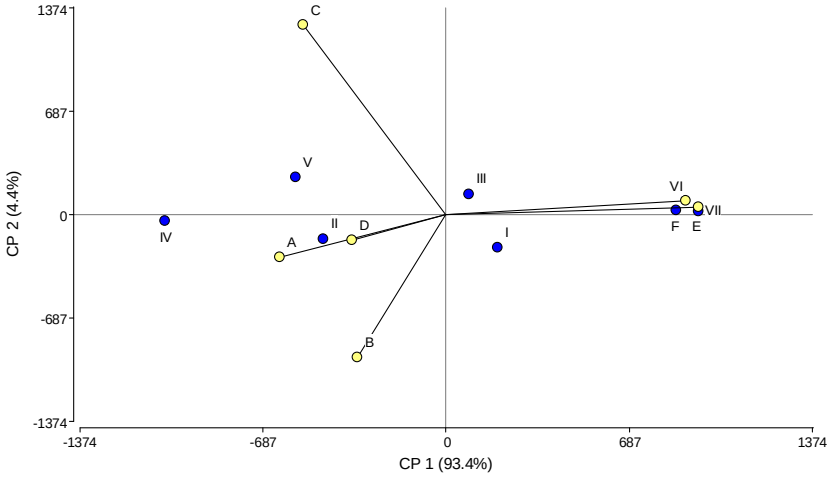


Figura 11.2. Biplot del ACP de los efectos de interacción entre 7 genotipos (I,II,III,IV,V,VI y VII) y 6 ambientes (A,B,C,D,E y F).

La Figura siguiente combina información sobre producción y estabilidad. Teniendo en cuanta ambas medidas el Genotipo I es el mejor posicionado, *i.e.* con un rendimiento alto relativo a la media de los rendimientos y un indicador de inestabilidad de valor bajo, es decir de mayor estabilidad de rendimientos a través de los ambientes. A nivel de rendimiento medio, la variedad I es similar a las variedades II y IV. No obstante esta última alcanza ese valor promedio con fuertes cambios a través de los ambientes y por tanto existen ambientes (como E y F) donde su cultivo puede resultar riesgoso.

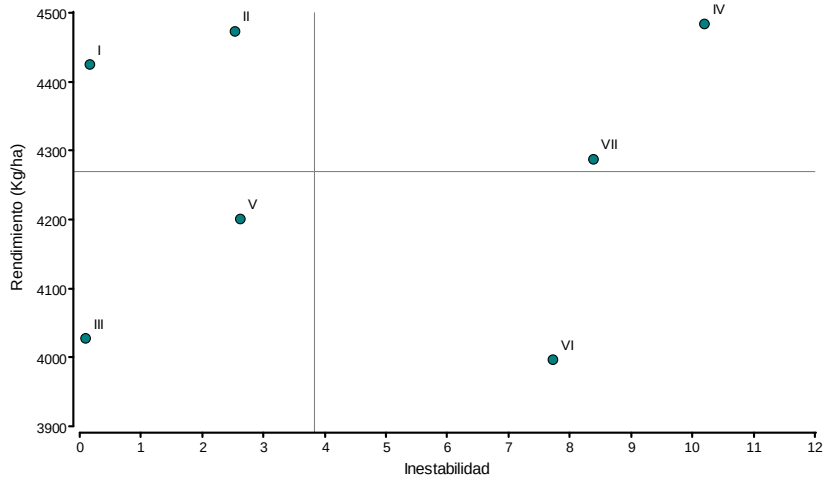


Figura 11.3. Rendimiento promedio e indicador de inestabilidad (menores valores indica estabilidad) de rendimientos a través de los ambientes de ensayo para 7 genotipos.

12. Referencias

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons.

Balzarini, M. (2008). *Análisis Multivariado. Curso de posgrado. FCA-UNC.* Córdoba, Argentina.

Draper, N. R., & Smith, H. (1988). *Applied Regression Analysis* (Third ed.). New York: John Wiley & Sons.

Di Rienzo, J. A., Casanoves, F., Gonzalez, L. A., Tablada, E. M., Díaz, M. d., Robledo, C. W., y otros. (2007). *Estadística para las Ciencias Agropecuarias.* Córdoba: Brujas.

Di Rienzo, J. A., Macchiavelli, R., & Casanoves, F. (2010). Modelos Mixtos en InfoStat. Córdoba, Córdoba, Argentina.

Hacking. (1991). *La domesticación del azar: La erosión del determinismo y el nacimiento de las ciencias del caos.* Barcelona: Editorial Gedisa.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6 ed.). Prentice Hall.

Levin, R. I., & Rubin, D. S. (2004). *Estadística para administración y economía* (Séptima ed.). Méjico: Pearson Educación.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility". *Biometrics (International Biometric Society) 45 (1)* , 255–268. doi:10.2307/2532051. PMID 2720055. http://jstor.org/stable/2532051.

Nickerson, C. A. (1997). A Note on A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics (International Biometric Society) 53 (4)*, 1503–1507. doi:10.2307/2533516. http://www.jstor.org/stable/2533516.

Peña, D. (2002). *Análisis Multivariado.* España: Mc Graw Hill.

13. Tablas Estadísticas

Tabla de Números Aleatorios

81	4	37	23	59	51	32	71	89	37	66	28	38	49	59	49	33	77	42
82	24	34	34	71	62	74	66	32	26	75	20	47	68	86	92	81	19	9
73	34	62	51	22	38	24	28	45	44	25	68	74	68	26	64	44	79	94
76	27	21	30	62	52	44	30	84	6	44	60	31	31	39	4	18	33	59
4	48	54	8	86	7	43	52	86	63	84	74	72	91	29	96	73	5	60
18	10	75	64	40	44	2	66	24	45	58	44	73	79	66	95	25	49	80
34	100	36	14	79	51	49	35	93	97	28	4	78	2	34	58	40	9	48
53	46	39	11	61	33	12	8	70	28	2	7	87	58	7	59	2	68	48
79	25	52	36	53	64	29	57	84	26	56	11	15	69	52	42	20	12	99
66	10	24	92	19	74	100	85	39	5	39	39	58	8	49	34	41	77	70
99	84	99	91	41	88	9	33	24	99	96	98	18	89	44	93	12	17	92
50	28	33	52	84	40	21	5	49	92	21	31	2	62	53	13	96	69	85
76	55	77	53	13	39	64	43	58	64	31	78	56	95	49	57	2	64	56
93	35	75	28	48	100	98	48	27	12	94	27	84	43	32	18	19	13	77
7	17	21	49	100	15	59	83	10	67	99	4	26	88	33	27	80	63	73
72	38	80	72	69	22	19	17	65	68	66	84	83	97	86	8	55	74	93
7	5	58	68	42	70	2	16	23	35	60	45	35	60	43	62	69	7	58
19	34	58	54	20	91	95	72	16	37	46	57	93	31	97	2	96	81	6
40	72	65	99	49	40	10	68	88	14	11	84	22	91	55	44	79	85	84
99	37	83	34	31	43	86	58	30	67	21	2	54	27	46	11	32	43	10
2	16	91	60	88	6	26	5	58	44	97	90	90	28	12	78	67	45	5
80	7	47	41	67	64	96	49	84	42	87	33	15	28	58	64	42	49	74
53	20	35	44	18	26	47	6	1	55	6	74	62	56	23	51	78	15	19
73	88	60	42	74	2	31	32	85	40	21	42	68	35	51	58	87	5	10
32	13	59	78	14	50	89	18	41	63	35	49	67	72	31	66	79	22	14
67	51	56	9	52	98	83	41	16	43	50	27	94	48	66	6	20	43	23
95	52	3	87	98	43	17	72	50	58	31	27	92	46	31	69	72	67	27
45	67	22	41	55	27	32	44	80	34	57	10	37	30	5	65	59	27	99
82	63	70	7	59	37	61	58	99	31	33	69	10	79	32	50	56	48	78
97	50	13	19	83	27	23	55	88	57	67	8	58	76	56	62	15	76	56
46	37	31	68	62	89	98	57	60	70	24	76	44	57	86	62	83	26	59
76	22	34	79	33	45	32	43	76	7	45	12	61	24	29	20	24	45	65
44	94	14	84	72	5	19	19	61	47	18	21	41	96	17	45	63	5	6
20	65	87	43	77	46	73	38	74	18	73	62	25	18	24	68	27	64	51
34	14	3	89	68	56	33	33	67	14	9	38	58	95	32	14	54	34	65
13	80	93	61	53	61	95	63	35	52	80	83	84	61	25	76	20	13	73
35	98	76	30	2	7	1	88	19	9	39	44	39	38	40	42	60	15	10
81	33	39	20	88	46	73	62	41	93	49	53	48	40	17	40	83	12	53
19	26	69	65	72	64	9	28	14	75	57	35	25	90	49	23	83	71	30
63	36	77	14	9	94	59	3	16	100	89	93	93	97	4	69	90	97	40
53	44	47	62	82	41	77	18	59	65	31	86	41	39	78	77	24	65	79
15	63	14	64	93	89	55	27	46	27	67	38	38	26	94	24	82	86	63
85	13	32	99	4	4	46	40	95	10	33	30	98	3	53	17	86	63	93
5	83	68	8	51	95	7	37	42	38	57	99	58	74	53	42	67	1	68
49	19	61	29	69	26	39	58	4	42	22	11	99	2	53	17	13	76	5
83	76	63	26	32	66	42	55	85	15	72	78	27	51	25	82	71	38	13
58	24	35	54	45	36	69	36	41	92	85	16	59	99	99	12	58	19	51
29	45	5	17	94	51	56	13	55	79	39	18	62	58	9	59	36	46	45
87	4	54	61	45	75	31	68	92	96	51	76	20	41	28	80	69	88	84
95	4	25	62	86	89	90	88	21	66	33	32	6	59	82	3	67	41	44
4	44	99	80	20	29	89	21	44	33	85	77	25	26	40	50	25	47	77
34	78	11	64	83	68	5	56	53	34	32	14	90	31	57	47	82	84	31
33	23	22	97	13	28	2	91	85	67	49	41	81	74	94	28	49	82	25
56	14	92	52	25	15	60	46	29	5	54	91	58	19	88	15	29	86	36
43	77	74	77	84	66	49	38	72	84	86	77	9	4	26	69	38	65	31

Probabilidades bioniales

Tamaño de muestra (N), número de eventos (n) y probabilidad de ocurrencia del vento (p)

N	n	p=0.01	p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
2	0	0.9801	0.9025	0.8100	0.6400	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100	0.0025	0.0001
	1	0.0198	0.0950	0.1800	0.3200	0.4200	0.4800	0.5000	0.4800	0.4200	0.3200	0.1800	0.0950	0.0198
	2	0.0001	0.0025	0.0100	0.0400	0.0900	0.1600	0.2500	0.3600	0.4900	0.6400	0.8100	0.9025	0.9801
3	0	0.9703	0.8574	0.7290	0.5120	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010	0.0001	0.0000
	1	0.0294	0.1354	0.2430	0.3840	0.4410	0.4320	0.3750	0.2880	0.1890	0.0960	0.0270	0.0071	0.0003
	2	0.0003	0.0071	0.0270	0.0960	0.1890	0.2880	0.3750	0.4320	0.4410	0.3840	0.2430	0.1354	0.0294
	3	0.0000	0.0001	0.0010	0.0080	0.0270	0.0640	0.1250	0.2160	0.3430	0.5120	0.7290	0.8574	0.9703
4	0	0.9606	0.8145	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001	0.0000	0.0000
	1	0.0388	0.1715	0.2916	0.4096	0.4116	0.3456	0.2500	0.1536	0.0756	0.0256	0.0036	0.0005	0.0000
	2	0.0006	0.0135	0.0486	0.1536	0.2646	0.3456	0.3750	0.3456	0.2646	0.1536	0.0486	0.0135	0.0006
	3	0.0000	0.0005	0.0036	0.0256	0.0756	0.1536	0.2500	0.3456	0.4116	0.4096	0.2916	0.1715	0.0388
	4	0.0000	0.0000	0.0001	0.0016	0.0081	0.0256	0.0625	0.1296	0.2401	0.4096	0.6561	0.8145	0.9606
5	0	0.9510	0.7738	0.5905	0.3277	0.1681	0.0778	0.0312	0.0102	0.0024	0.0003	0.0000	0.0000	0.0000
	1	0.0480	0.2036	0.3280	0.4096	0.3602	0.2592	0.1562	0.0768	0.0284	0.0064	0.0004	0.0000	0.0000
	2	0.0010	0.0214	0.0729	0.2048	0.3087	0.3456	0.3125	0.2304	0.1323	0.0512	0.0081	0.0011	0.0000
	3	0.0000	0.0011	0.0081	0.0512	0.1323	0.2304	0.3125	0.3456	0.3087	0.2048	0.0729	0.0214	0.0010
	4	0.0000	0.0000	0.0005	0.0064	0.0284	0.0768	0.1562	0.2592	0.3602	0.4096	0.3280	0.2036	0.0480
	5	0.0000	0.0000	0.0000	0.0003	0.0024	0.0102	0.0312	0.0778	0.1681	0.3277	0.5905	0.7738	0.9510
6	0	0.9415	0.7351	0.5314	0.2621	0.1176	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000	0.0000	0.0000
	1	0.0571	0.2321	0.3543	0.3932	0.3025	0.1866	0.0937	0.0369	0.0102	0.0015	0.0001	0.0000	0.0000
	2	0.0014	0.0305	0.0984	0.2458	0.3241	0.3110	0.2344	0.1382	0.0595	0.0154	0.0012	0.0001	0.0000
	3	0.0000	0.0021	0.0146	0.0819	0.1852	0.2765	0.3125	0.2765	0.1852	0.0819	0.0146	0.0021	0.0000
	4	0.0000	0.0001	0.0012	0.0154	0.0595	0.1382	0.2344	0.3110	0.3241	0.2458	0.0984	0.0305	0.0014
	5	0.0000	0.0000	0.0001	0.0015	0.0102	0.0369	0.0938	0.1866	0.3025	0.3932	0.3543	0.2321	0.0571
	6	0.0000	0.0000	0.0000	0.0001	0.0007	0.0041	0.0156	0.0467	0.1176	0.2621	0.5314	0.7351	0.9415
7	0	0.9321	0.6983	0.4783	0.2097	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	0.0000	0.0000	0.0000
	1	0.0659	0.2573	0.3720	0.3670	0.2471	0.1306	0.0547	0.0172	0.0036	0.0004	0.0000	0.0000	0.0000
	2	0.0020	0.0406	0.1240	0.2753	0.3177	0.2613	0.1641	0.0774	0.0250	0.0043	0.0002	0.0000	0.0000
	3	0.0000	0.0036	0.0230	0.1147	0.2269	0.2903	0.2734	0.1935	0.0972	0.0287	0.0026	0.0002	0.0000
	4	0.0000	0.0002	0.0026	0.0287	0.0972	0.1935	0.2734	0.2903	0.2269	0.1147	0.0230	0.0036	0.0000
	5	0.0000	0.0000	0.0002	0.0043	0.0250	0.0774	0.1641	0.2613	0.3177	0.2753	0.1240	0.0406	0.0020
	6	0.0000	0.0000	0.0000	0.0004	0.0036	0.0172	0.0547	0.1306	0.2471	0.3670	0.3720	0.2573	0.0659
	7	0.0000	0.0000	0.0000	0.0000	0.0002	0.0016	0.0078	0.0280	0.0824	0.2097	0.4783	0.6983	0.9321
8	0	0.9227	0.6634	0.4305	0.1678	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.0746	0.2793	0.3826	0.3355	0.1977	0.0896	0.0313	0.0079	0.0012	0.0001	0.0000	0.0000	0.0000
	2	0.0026	0.0515	0.1488	0.2936	0.2965	0.2090	0.1094	0.0413	0.0100	0.0011	0.0000	0.0000	0.0000
	3	0.0001	0.0054	0.0331	0.1468	0.2541	0.2787	0.2187	0.1239	0.0467	0.0092	0.0004	0.0000	0.0000
	4	0.0000	0.0004	0.0046	0.0459	0.1361	0.2322	0.2734	0.2322	0.1361	0.0459	0.0046	0.0004	0.0000
	5	0.0000	0.0000	0.0004	0.0092	0.0467	0.1239	0.2187	0.2787	0.2541	0.1468	0.0331	0.0054	0.0001
	6	0.0000	0.0000	0.0000	0.0011	0.0100	0.0413	0.1094	0.2090	0.2965	0.2936	0.1488	0.0515	0.0026
	7	0.0000	0.0000	0.0000	0.0001	0.0012	0.0079	0.0313	0.0896	0.1977	0.3355	0.3826	0.2793	0.0746
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0039	0.0168	0.0576	0.1678	0.4305	0.6634	0.9227
9	0	0.9135	0.6302	0.3874	0.1342	0.0404	0.0101	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0830	0.2985	0.3874	0.3020	0.1556	0.0605	0.0176	0.0035	0.0004	0.0000	0.0000	0.0000	0.0000
	2	0.0034	0.0629	0.1722	0.3020	0.2668	0.1612	0.0703	0.0212	0.0039	0.0003	0.0000	0.0000	0.0000
	3	0.0001	0.0077	0.0446	0.1762	0.2668	0.2508	0.1641	0.0743	0.0210	0.0028	0.0001	0.0000	0.0000
	4	0.0000	0.0006	0.0074	0.0661	0.1715	0.2508	0.2461	0.1672	0.0735	0.0165	0.0008	0.0000	0.0000
	5	0.0000	0.0000	0.0008	0.0165	0.0735	0.1672	0.2461	0.2508	0.1715	0.0661	0.0074	0.0006	0.0000
	6	0.0000	0.0000	0.0001	0.0028	0.0210	0.0743	0.1641	0.2508	0.2668	0.1762	0.0446	0.0077	0.0001
	7	0.0000	0.0000	0.0000	0.0003	0.0039	0.0212	0.0703	0.1612	0.2668	0.3020	0.1722	0.0629	0.0034
	8	0.0000	0.0000	0.0000	0.0000	0.0004	0.0035	0.0176	0.0605	0.1556	0.3020	0.3874	0.2985	0.0830
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0020	0.0101	0.0404	0.1342	0.3874	0.6302	0.9135
10	0	0.9044	0.5987	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0914	0.3151	0.3874	0.2684	0.1211	0.0403	0.0098	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.0042	0.0746	0.1937	0.3020	0.2335	0.1209	0.0439	0.0106	0.0014	0.0001	0.0000	0.0000	0.0000
	3	0.0001	0.0105	0.0574	0.2013	0.2668	0.2150	0.1172	0.0425	0.0090	0.0008	0.0000	0.0000	0.0000
	4	0.0000	0.0010	0.0112	0.0881	0.2001	0.2508	0.2051	0.1115	0.0368	0.0055	0.0001	0.0000	0.0000
	5	0.0000	0.0001	0.0015	0.0264	0.1029	0.2007	0.2461	0.2007	0.1029	0.0264	0.0015	0.0001	0.0000
	6	0.0000	0.0000	0.0001	0.0055	0.0368	0.1115	0.2051	0.2508	0.2001	0.0881	0.0112	0.0010	0.0000
	7	0.0000	0.0000	0.0000	0.0008	0.0090	0.0425	0.1172	0.2150	0.2668	0.2013	0.0574	0.0105	0.0001
	8	0.0000	0.0000	0.0000	0.0001	0.0014	0.0106	0.0439	0.1209	0.2335	0.3020	0.1937	0.0746	0.0042
	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0098	0.0403	0.1211	0.2684	0.3874	0.3151	0.0914
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0060	0.0282	0.1074	0.3487	0.5987	0.9044

Probabilidades bioniales

Tamaño de muestra (N), número de eventos (n) y probabilidad de ocurrencia del vento (p)

N	n	p=0.01	p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
11	0	0.8953	0.5688	0.3138	0.0859	0.0198	0.0036	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0995	0.3293	0.3835	0.2362	0.0932	0.0266	0.0054	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0050	0.0867	0.2131	0.2953	0.1998	0.0887	0.0269	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000
	3	0.0002	0.0137	0.0710	0.2215	0.2568	0.1774	0.0806	0.0234	0.0037	0.0002	0.0000	0.0000	0.0000
	4	0.0000	0.0014	0.0158	0.1107	0.2201	0.2365	0.1611	0.0701	0.0173	0.0017	0.0000	0.0000	0.0000
	5	0.0000	0.0001	0.0025	0.0388	0.1321	0.2207	0.2256	0.1471	0.0566	0.0097	0.0003	0.0000	0.0000
	6	0.0000	0.0000	0.0003	0.0097	0.0566	0.1471	0.2256	0.2207	0.1321	0.0388	0.0025	0.0001	0.0000
	7	0.0000	0.0000	0.0000	0.0017	0.0173	0.0701	0.1611	0.2365	0.2201	0.1107	0.0158	0.0014	0.0000
	8	0.0000	0.0000	0.0000	0.0002	0.0037	0.0234	0.0806	0.1774	0.2568	0.2215	0.0710	0.0137	0.0002
	9	0.0000	0.0000	0.0000	0.0000	0.0005	0.0052	0.0269	0.0887	0.1998	0.2953	0.2131	0.0867	0.0050
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0054	0.0266	0.0932	0.2362	0.3835	0.3293	0.0995
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0036	0.0198	0.0859	0.3138	0.5688	0.8953
12	0	0.8864	0.5404	0.2824	0.0687	0.0138	0.0022	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1074	0.3413	0.3766	0.2062	0.0712	0.0174	0.0029	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0060	0.0988	0.2301	0.2835	0.1678	0.0639	0.0161	0.0025	0.0002	0.0000	0.0000	0.0000	0.0000
	3	0.0002	0.0173	0.0852	0.2362	0.2397	0.1419	0.0537	0.0125	0.0015	0.0001	0.0000	0.0000	0.0000
	4	0.0000	0.0021	0.0213	0.1329	0.2311	0.2128	0.1208	0.0420	0.0078	0.0005	0.0000	0.0000	0.0000
	5	0.0000	0.0002	0.0038	0.0532	0.1585	0.2270	0.1934	0.1009	0.0291	0.0033	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0005	0.0155	0.0792	0.1766	0.2256	0.1766	0.0792	0.0155	0.0005	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0033	0.0291	0.1009	0.1934	0.2270	0.1585	0.0532	0.0038	0.0002	0.0000
	8	0.0000	0.0000	0.0000	0.0005	0.0078	0.0420	0.1208	0.2128	0.2311	0.1329	0.0213	0.0021	0.0000
	9	0.0000	0.0000	0.0000	0.0001	0.0015	0.0125	0.0537	0.1419	0.2397	0.2362	0.0852	0.0173	0.0002
	10	0.0000	0.0000	0.0000	0.0000	0.0002	0.0025	0.0161	0.0639	0.1678	0.2835	0.2301	0.0988	0.0060
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0029	0.0174	0.0712	0.2062	0.3766	0.3413	0.1074
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0022	0.0138	0.0687	0.2824	0.5404	0.8864
13	0	0.8775	0.5133	0.2542	0.0550	0.0097	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1152	0.3512	0.3672	0.1787	0.0540	0.0113	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0070	0.1109	0.2448	0.2680	0.1388	0.0453	0.0095	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.0003	0.0214	0.0997	0.2457	0.2181	0.1107	0.0349	0.0065	0.0006	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0028	0.0277	0.1535	0.2337	0.1845	0.0873	0.0243	0.0034	0.0001	0.0000	0.0000	0.0000
	5	0.0000	0.0003	0.0055	0.0691	0.1803	0.2214	0.1571	0.0656	0.0142	0.0011	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0008	0.0230	0.1030	0.1968	0.2095	0.1312	0.0442	0.0058	0.0001	0.0000	0.0000
	7	0.0000	0.0000	0.0001	0.0058	0.0442	0.1312	0.2095	0.1968	0.1030	0.0230	0.0008	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0011	0.0142	0.0656	0.1571	0.2214	0.1803	0.0691	0.0055	0.0003	0.0000
	9	0.0000	0.0000	0.0000	0.0001	0.0034	0.0243	0.0873	0.1845	0.2337	0.1535	0.0277	0.0028	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0006	0.0065	0.0349	0.1107	0.2181	0.2457	0.0997	0.0214	0.0003
	11	0.0000	0.0000	0.0000	0.0000	0.0001	0.0012	0.0095	0.0453	0.1388	0.2680	0.2448	0.1109	0.0070
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0113	0.0540	0.1787	0.3672	0.3512	0.1152
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0013	0.0097	0.0550	0.2542	0.5133	0.8775
14	0	0.8687	0.4877	0.2288	0.0440	0.0068	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1229	0.3593	0.3559	0.1539	0.0407	0.0073	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0081	0.1229	0.2570	0.2501	0.1134	0.0317	0.0056	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0003	0.0259	0.1142	0.2501	0.1943	0.0845	0.0222	0.0033	0.0002	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0037	0.0349	0.1720	0.2290	0.1549	0.0611	0.0136	0.0014	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0004	0.0078	0.0860	0.1963	0.2066	0.1222	0.0408	0.0066	0.0003	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0013	0.0322	0.1262	0.2066	0.1833	0.0918	0.0232	0.0020	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0002	0.0092	0.0618	0.1574	0.2095	0.1574	0.0618	0.0092	0.0002	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0020	0.0232	0.0918	0.1833	0.2066	0.1262	0.0322	0.0013	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0003	0.0066	0.0408	0.1222	0.2066	0.1963	0.0860	0.0078	0.0004	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0014	0.0136	0.0611	0.1549	0.2290	0.1720	0.0349	0.0037	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0002	0.0033	0.0222	0.0845	0.1943	0.2501	0.1142	0.0259	0.0003
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0056	0.0317	0.1134	0.2501	0.2570	0.1229	0.0081
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0009	0.0073	0.0407	0.1539	0.3559	0.3593	0.1229
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008	0.0068	0.0440	0.2288	0.4877	0.8687
15	0	0.8601	0.4633	0.2059	0.0352	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1303	0.3658	0.3432	0.1319	0.0305	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0092	0.1348	0.2669	0.2309	0.0916	0.0219	0.0032	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0004	0.0307	0.1285	0.2501	0.1700	0.0634	0.0139	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0049	0.0428	0.1876	0.2186	0.1268	0.0417	0.0074	0.0006	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0006	0.0105	0.1032	0.2061	0.1859	0.0916	0.0245	0.0030	0.0001	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0019	0.0430	0.1472	0.2066	0.1527	0.0612	0.0116	0.0007	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0003	0.0138	0.0811	0.1771	0.1964	0.1181	0.0348	0.0035	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0035	0.0348	0.1181	0.1964	0.1771	0.0811	0.0138	0.0003	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0007	0.0116	0.0612	0.1527	0.2066	0.1472	0.0430	0.0019	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0001	0.0030	0.0245	0.0916	0.1859	0.2061	0.1032	0.0105	0.0006	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0006	0.0074	0.0417	0.1268	0.2186	0.1876	0.0428	0.0049	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0139	0.0634	0.1700	0.2501	0.1285	0.0307	0.0004
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0032	0.0219	0.0916	0.2309	0.2669	0.1348	0.0092
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0047	0.0305	0.1319	0.3432	0.3658	0.1303
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0047	0.0352	0.2059	0.4633	0.8601

Probabilidades bioniales

Tamaño de muestra (N), número de eventos (n) y probabilidad de ocurrencia del vento (p)

N	n	p=0.01	p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
16	0	0.8515	0.4401	0.1853	0.0281	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1376	0.3706	0.3294	0.1126	0.0228	0.0030	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0104	0.1463	0.2745	0.2111	0.0732	0.0150	0.0018	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0005	0.0359	0.1423	0.2463	0.1465	0.0468	0.0085	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0061	0.0514	0.2001	0.2040	0.1014	0.0278	0.0040	0.0002	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0008	0.0137	0.1201	0.2099	0.1623	0.0667	0.0142	0.0013	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0001	0.0028	0.0550	0.1649	0.1983	0.1222	0.0392	0.0056	0.0002	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0004	0.0197	0.1010	0.1889	0.1746	0.0840	0.0185	0.0012	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0001	0.0055	0.0487	0.1417	0.1964	0.1417	0.0487	0.0055	0.0001	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0012	0.0185	0.0840	0.1746	0.1889	0.1010	0.0197	0.0004	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0002	0.0056	0.0392	0.1222	0.1983	0.1649	0.0550	0.0028	0.0001	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0013	0.0142	0.0667	0.1623	0.2099	0.1201	0.0137	0.0008	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0002	0.0040	0.0278	0.1014	0.2040	0.2001	0.0514	0.0061	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0085	0.0468	0.1465	0.2463	0.1423	0.0359	0.0005
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0018	0.0150	0.0732	0.2111	0.2745	0.1463	0.0104
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0030	0.0228	0.1126	0.3294	0.3706	0.1376
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0033	0.0281	0.1853	0.4401	0.8515
17	0	0.8429	0.4181	0.1668	0.0225	0.0023	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1447	0.3741	0.3150	0.0957	0.0169	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0117	0.1575	0.2800	0.1914	0.0581	0.0102	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0006	0.0415	0.1556	0.2393	0.1245	0.0341	0.0052	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0076	0.0605	0.2093	0.1868	0.0796	0.0182	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0010	0.0175	0.1361	0.2081	0.1379	0.0472	0.0081	0.0006	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0001	0.0039	0.0680	0.1784	0.1839	0.0944	0.0242	0.0026	0.0001	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0007	0.0267	0.1201	0.1927	0.1484	0.0571	0.0095	0.0004	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0001	0.0084	0.0644	0.1606	0.1855	0.1070	0.0276	0.0021	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0021	0.0276	0.1070	0.1855	0.1606	0.0644	0.0084	0.0001	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0004	0.0095	0.0571	0.1484	0.1927	0.1201	0.0267	0.0007	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0001	0.0026	0.0242	0.0944	0.1839	0.1784	0.0680	0.0039	0.0001	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0006	0.0081	0.0472	0.1379	0.2081	0.1361	0.0175	0.0010	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0021	0.0182	0.0796	0.1868	0.2093	0.0605	0.0076	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0052	0.0341	0.1245	0.2393	0.1556	0.0415	0.0006
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0102	0.0581	0.1914	0.2800	0.1575	0.0117
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0019	0.0169	0.0957	0.3150	0.3741	0.1447
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0023	0.0225	0.1668	0.4181	0.8429
18	0	0.8345	0.3972	0.1501	0.0180	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1517	0.3763	0.3002	0.0811	0.0126	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0130	0.1683	0.2835	0.1723	0.0458	0.0069	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0007	0.0473	0.1680	0.2297	0.1046	0.0246	0.0031	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0093	0.0700	0.2153	0.1681	0.0614	0.0117	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0014	0.0218	0.1507	0.2017	0.1146	0.0327	0.0045	0.0002	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0002	0.0052	0.0816	0.1873	0.1655	0.0708	0.0145	0.0012	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0010	0.0350	0.1376	0.1892	0.1214	0.0374	0.0046	0.0001	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0002	0.0120	0.0811	0.1734	0.1669	0.0771	0.0149	0.0008	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0033	0.0386	0.1284	0.1855	0.1284	0.0386	0.0033	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0008	0.0149	0.0771	0.1669	0.1734	0.0811	0.0120	0.0002	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0001	0.0046	0.0374	0.1214	0.1892	0.1376	0.0350	0.0010	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0012	0.0145	0.0708	0.1655	0.1873	0.0816	0.0052	0.0002	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0002	0.0045	0.0327	0.1146	0.2017	0.1507	0.0218	0.0014	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0011	0.0117	0.0614	0.1681	0.2153	0.0700	0.0093	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0031	0.0246	0.1046	0.2297	0.1680	0.0473	0.0007
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0006	0.0069	0.0458	0.1723	0.2835	0.1683	0.0130
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0012	0.0126	0.0811	0.3002	0.3763	0.1517
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0016	0.0180	0.1501	0.3972	0.8345

Probabilidades bioniales

Tamaño de muestra (N), número de eventos (n) y probabilidad de ocurrencia del vento (p)

N	n	p=0.01	p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
19	0	0.8262	0.3774	0.1351	0.0144	0.0011	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1586	0.3774	0.2852	0.0685	0.0093	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0144	0.1787	0.2852	0.1540	0.0358	0.0046	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0008	0.0533	0.1796	0.2182	0.0869	0.0175	0.0018	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0112	0.0798	0.2182	0.1491	0.0467	0.0074	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0018	0.0266	0.1636	0.1916	0.0933	0.0222	0.0024	0.0001	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0002	0.0069	0.0955	0.1916	0.1451	0.0518	0.0085	0.0005	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0014	0.0443	0.1525	0.1797	0.0961	0.0237	0.0022	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0002	0.0166	0.0981	0.1797	0.1442	0.0532	0.0077	0.0003	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0051	0.0514	0.1464	0.1762	0.0976	0.0220	0.0013	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0013	0.0220	0.0976	0.1762	0.1464	0.0514	0.0051	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0003	0.0077	0.0532	0.1442	0.1797	0.0981	0.0166	0.0002	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0022	0.0237	0.0961	0.1797	0.1525	0.0443	0.0014	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0005	0.0085	0.0518	0.1451	0.1916	0.0955	0.0069	0.0002	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0001	0.0024	0.0222	0.0933	0.1916	0.1636	0.0266	0.0018	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0074	0.0467	0.1491	0.2182	0.0798	0.0112	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0018	0.0175	0.0869	0.2182	0.1796	0.0533	0.0008
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0046	0.0358	0.1540	0.2852	0.1787	0.0144
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0093	0.0685	0.2852	0.3774	0.1586
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0144	0.1351	0.3774	0.8262
20	0	0.8179	0.3585	0.1216	0.0115	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1652	0.3774	0.2702	0.0576	0.0068	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0159	0.1887	0.2852	0.1369	0.0278	0.0031	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0010	0.0596	0.1901	0.2054	0.0716	0.0123	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0000	0.0133	0.0898	0.2182	0.1304	0.0350	0.0046	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0022	0.0319	0.1746	0.1789	0.0746	0.0148	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0003	0.0089	0.1091	0.1916	0.1244	0.0370	0.0049	0.0002	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0020	0.0545	0.1643	0.1659	0.0739	0.0146	0.0010	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0004	0.0222	0.1144	0.1797	0.1201	0.0355	0.0039	0.0001	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0001	0.0074	0.0654	0.1597	0.1602	0.0710	0.0120	0.0005	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0020	0.0308	0.1171	0.1762	0.1171	0.0308	0.0020	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0005	0.0120	0.0710	0.1602	0.1597	0.0654	0.0074	0.0001	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0001	0.0039	0.0355	0.1201	0.1797	0.1144	0.0222	0.0004	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0010	0.0146	0.0739	0.1659	0.1643	0.0545	0.0020	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0002	0.0049	0.0370	0.1244	0.1916	0.1091	0.0089	0.0003	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013	0.0148	0.0746	0.1789	0.1746	0.0319	0.0022	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0046	0.0350	0.1304	0.2182	0.0898	0.0133	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0011	0.0123	0.0716	0.2054	0.1901	0.0596	0.0010
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0031	0.0278	0.1369	0.2852	0.1887	0.0159
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0068	0.0576	0.2702	0.3774	0.1652
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0115	0.1216	0.3585	0.8179

Probabilidades Poisson

Número de eventos en filas, parámetro lambda en columnas

n	l=0.1	l=0.2	l=0.4	l=0.8	l=2	l=5	l=10	l=20	l=30	l=40	l=50
0	0.9048	0.8187	0.6703	0.4493	0.1353	0.0067	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0905	0.1637	0.2681	0.3595	0.2707	0.0337	0.0005	0.0000	0.0000	0.0000	0.0000
2	0.0045	0.0164	0.0536	0.1438	0.2707	0.0842	0.0023	0.0000	0.0000	0.0000	0.0000
3	0.0002	0.0011	0.0072	0.0383	0.1804	0.1404	0.0076	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0001	0.0007	0.0077	0.0902	0.1755	0.0189	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0001	0.0012	0.0361	0.1755	0.0378	0.0001	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0002	0.0120	0.1462	0.0631	0.0002	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0034	0.1044	0.0901	0.0005	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0009	0.0653	0.1126	0.0013	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0002	0.0363	0.1251	0.0029	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0181	0.1251	0.0058	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0082	0.1137	0.0106	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034	0.0948	0.0176	0.0001	0.0000	0.0000
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013	0.0729	0.0271	0.0002	0.0000	0.0000
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0521	0.0387	0.0005	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0347	0.0516	0.0010	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0217	0.0646	0.0019	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0128	0.0760	0.0034	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0071	0.0844	0.0057	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0037	0.0888	0.0089	0.0001	0.0000
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019	0.0888	0.0134	0.0002	0.0000
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0846	0.0192	0.0004	0.0000
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0769	0.0261	0.0007	0.0000
23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0669	0.0341	0.0012	0.0000
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0557	0.0426	0.0019	0.0000
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0446	0.0511	0.0031	0.0000
26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0343	0.0590	0.0047	0.0001
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0254	0.0655	0.0070	0.0001
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0181	0.0702	0.0100	0.0002
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0125	0.0726	0.0138	0.0004
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0083	0.0726	0.0185	0.0007
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0054	0.0703	0.0238	0.0011
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034	0.0659	0.0298	0.0017
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0020	0.0599	0.0361	0.0026
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0529	0.0425	0.0038
35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0453	0.0485	0.0054
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0378	0.0539	0.0075
37	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0306	0.0583	0.0102
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0242	0.0614	0.0134
39	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0186	0.0629	0.0172
40	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0139	0.0629	0.0215
41	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0102	0.0614	0.0262
42	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0073	0.0585	0.0312
43	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0051	0.0544	0.0363
44	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0035	0.0495	0.0412
45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0023	0.0440	0.0458
46	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0382	0.0498
47	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0325	0.0530
48	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0006	0.0271	0.0552
49	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0221	0.0563
50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0177	0.0563
51	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0139	0.0552
52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0107	0.0531
53	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0081	0.0501
54	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0060	0.0464
55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0043	0.0422
56	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0031	0.0376
57	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0022	0.0330
58	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0285
59	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0241
60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0201

Tabla de Cuantiles de la una distribución normal estándar

z	P(Z ≤ z)	z	P(Z ≤ z)	z	P(Z ≤ z)	quantil	z
-3.25	0.00058	-1.00	0.15866	1.25	0.89435	0.00001	-4.265
-3.20	0.00069	-0.95	0.17106	1.30	0.90320	0.0001	-3.719
-3.15	0.00082	-0.90	0.18406	1.35	0.91149	0.001	-3.090
-3.10	0.00097	-0.85	0.19766	1.40	0.91924	0.005	-2.576
-3.05	0.00114	-0.80	0.21186	1.45	0.92647	0.01	-2.326
-3.00	0.00135	-0.75	0.22663	1.50	0.93319	0.02	-2.054
-2.95	0.00159	-0.70	0.24196	1.55	0.93943	0.025	-1.960
-2.90	0.00187	-0.65	0.25785	1.60	0.94520	0.03	-1.881
-2.85	0.00219	-0.60	0.27425	1.65	0.95053	0.04	-1.751
-2.80	0.00256	-0.55	0.29116	1.70	0.95543	0.05	-1.645
-2.75	0.00298	-0.50	0.30854	1.75	0.95994	0.06	-1.555
-2.70	0.00347	-0.45	0.32636	1.80	0.96407	0.07	-1.476
-2.65	0.00402	-0.40	0.34458	1.85	0.96784	0.08	-1.405
-2.60	0.00466	-0.35	0.36317	1.90	0.97128	0.09	-1.341
-2.55	0.00539	-0.30	0.38209	1.95	0.97441	0.10	-1.282
-2.50	0.00621	-0.25	0.40129	2.00	0.97725	0.15	-1.036
-2.45	0.00714	-0.20	0.42074	2.05	0.97982	0.20	-0.842
-2.40	0.00820	-0.15	0.44038	2.10	0.98214	0.25	-0.674
-2.35	0.00939	-0.10	0.46017	2.15	0.98422	0.30	-0.524
-2.30	0.01072	-0.05	0.48006	2.20	0.98610	0.35	-0.385
-2.25	0.01222	0.00	0.50000	2.25	0.98778	0.40	-0.253
-2.20	0.01390	0.05	0.51994	2.30	0.98928	0.45	-0.126
-2.15	0.01578	0.10	0.53983	2.35	0.99061	0.50	0.000
-2.10	0.01786	0.15	0.55962	2.40	0.99180	0.55	0.126
-2.05	0.02018	0.20	0.57926	2.45	0.99286	0.60	0.253
-2.00	0.02275	0.25	0.59871	2.50	0.99379	0.65	0.385
-1.95	0.02559	0.30	0.61791	2.55	0.99461	0.70	0.524
-1.90	0.02872	0.35	0.63683	2.60	0.99534	0.75	0.674
-1.85	0.03216	0.40	0.65542	2.65	0.99598	0.80	0.842
-1.80	0.03593	0.45	0.67364	2.70	0.99653	0.85	1.036
-1.75	0.04006	0.50	0.69146	2.75	0.99702	0.90	1.282
-1.70	0.04457	0.55	0.70884	2.80	0.99744	0.91	1.341
-1.65	0.04947	0.60	0.72575	2.85	0.99781	0.92	1.405
-1.60	0.05480	0.65	0.74215	2.90	0.99813	0.93	1.476
-1.55	0.06057	0.70	0.75804	2.95	0.99841	0.94	1.555
-1.50	0.06681	0.75	0.77337	3.00	0.99865	0.95	1.645
-1.45	0.07353	0.80	0.78814	3.05	0.99886	0.96	1.751
-1.40	0.08076	0.85	0.80234	3.10	0.99903	0.97	1.881
-1.35	0.08851	0.90	0.81594	3.15	0.99918	0.975	1.960
-1.30	0.09680	0.95	0.82894	3.20	0.99931	0.98	2.054
-1.25	0.10565	1.00	0.84134	3.25	0.99942	0.99	2.326
-1.20	0.11507	1.05	0.85314	3.30	0.99952	0.995	2.576
-1.15	0.12507	1.10	0.86433	3.35	0.99960	0.999	3.090
-1.10	0.13567	1.15	0.87493	3.40	0.99966	0.9999	3.719
-1.05	0.14686	1.20	0.88493	3.45	0.99972	0.99999	4.265

Tabla de Cuantiles de la Distribución T de Student

En el margen superior se leen los cuatiles y en el margen izquierdo los grados de libertad (ν). Esta tabla tabula valores $P(T \leq t)$ para $t>0$. Si se buscan valores de $t<0$ los cuantiles se leen en el margen inferior.

ν	0.700	0.725	0.750	0.775	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975	0.990	0.995
1	0.727	0.854	1.000	1.171	1.376	1.632	1.963	2.414	3.078	4.165	6.314	12.71	31.82	63.66
2	0.617	0.713	0.816	0.931	1.061	1.210	1.386	1.604	1.886	2.282	2.920	4.303	6.965	9.925
3	0.584	0.671	0.765	0.866	0.978	1.105	1.250	1.423	1.638	1.924	2.353	3.182	4.541	5.841
4	0.569	0.652	0.741	0.836	0.941	1.057	1.190	1.344	1.533	1.778	2.132	2.776	3.747	4.604
5	0.559	0.641	0.727	0.819	0.920	1.031	1.156	1.301	1.476	1.699	2.015	2.571	3.365	4.032
6	0.553	0.633	0.718	0.808	0.906	1.013	1.134	1.273	1.440	1.650	1.943	2.447	3.143	3.707
7	0.549	0.628	0.711	0.800	0.896	1.001	1.119	1.254	1.415	1.617	1.895	2.365	2.998	3.499
8	0.546	0.624	0.706	0.794	0.889	0.993	1.108	1.240	1.397	1.592	1.860	2.306	2.896	3.355
9	0.543	0.621	0.703	0.790	0.883	0.986	1.100	1.230	1.383	1.574	1.833	2.262	2.821	3.250
10	0.542	0.619	0.700	0.786	0.879	0.980	1.093	1.221	1.372	1.559	1.812	2.228	2.764	3.169
11	0.540	0.617	0.697	0.783	0.876	0.976	1.088	1.214	1.363	1.548	1.796	2.201	2.718	3.106
12	0.539	0.615	0.695	0.781	0.873	0.972	1.083	1.209	1.356	1.538	1.782	2.179	2.681	3.055
13	0.538	0.614	0.694	0.779	0.870	0.969	1.079	1.204	1.350	1.530	1.771	2.160	2.650	3.012
14	0.537	0.613	0.692	0.777	0.868	0.967	1.076	1.200	1.345	1.523	1.761	2.145	2.624	2.977
15	0.536	0.612	0.691	0.776	0.866	0.965	1.074	1.197	1.341	1.517	1.753	2.131	2.602	2.947
16	0.535	0.611	0.690	0.774	0.865	0.963	1.071	1.194	1.337	1.512	1.746	2.120	2.583	2.921
17	0.534	0.610	0.689	0.773	0.863	0.961	1.069	1.191	1.333	1.508	1.740	2.110	2.567	2.898
18	0.534	0.609	0.688	0.772	0.862	0.960	1.067	1.189	1.330	1.504	1.734	2.101	2.552	2.878
19	0.533	0.609	0.688	0.771	0.861	0.958	1.066	1.187	1.328	1.500	1.729	2.093	2.539	2.861
20	0.533	0.608	0.687	0.771	0.860	0.957	1.064	1.185	1.325	1.497	1.725	2.086	2.528	2.845
21	0.532	0.608	0.686	0.770	0.859	0.956	1.063	1.183	1.323	1.494	1.721	2.080	2.518	2.831
22	0.532	0.607	0.686	0.769	0.858	0.955	1.061	1.182	1.321	1.492	1.717	2.074	2.508	2.819
23	0.532	0.607	0.685	0.769	0.858	0.954	1.060	1.180	1.319	1.489	1.714	2.069	2.500	2.807
24	0.531	0.606	0.685	0.768	0.857	0.953	1.059	1.179	1.318	1.487	1.711	2.064	2.492	2.797
25	0.531	0.606	0.684	0.767	0.856	0.952	1.058	1.178	1.316	1.485	1.708	2.060	2.485	2.787
26	0.531	0.606	0.684	0.767	0.856	0.952	1.058	1.177	1.315	1.483	1.706	2.056	2.479	2.779
27	0.531	0.605	0.684	0.767	0.855	0.951	1.057	1.176	1.314	1.482	1.703	2.052	2.473	2.771
28	0.530	0.605	0.683	0.766	0.855	0.950	1.056	1.175	1.313	1.480	1.701	2.048	2.467	2.763
29	0.530	0.605	0.683	0.766	0.854	0.950	1.055	1.174	1.311	1.479	1.699	2.045	2.462	2.756
30	0.530	0.605	0.683	0.765	0.854	0.949	1.055	1.173	1.310	1.477	1.697	2.042	2.457	2.750
31	0.530	0.604	0.682	0.765	0.853	0.949	1.054	1.172	1.309	1.476	1.696	2.040	2.453	2.744
32	0.530	0.604	0.682	0.765	0.853	0.948	1.054	1.172	1.309	1.475	1.694	2.037	2.449	2.738
33	0.530	0.604	0.682	0.765	0.853	0.948	1.053	1.171	1.308	1.474	1.692	2.035	2.445	2.733
34	0.529	0.604	0.682	0.764	0.852	0.948	1.052	1.170	1.307	1.473	1.691	2.032	2.441	2.728
35	0.529	0.604	0.682	0.764	0.852	0.947	1.052	1.170	1.306	1.472	1.690	2.030	2.438	2.724
36	0.529	0.603	0.681	0.764	0.852	0.947	1.052	1.169	1.306	1.471	1.688	2.028	2.434	2.719
37	0.529	0.603	0.681	0.764	0.851	0.947	1.051	1.169	1.305	1.470	1.687	2.026	2.431	2.715
38	0.529	0.603	0.681	0.763	0.851	0.946	1.051	1.168	1.304	1.469	1.686	2.024	2.429	2.712
39	0.529	0.603	0.681	0.763	0.851	0.946	1.050	1.168	1.304	1.468	1.685	2.023	2.426	2.708
40	0.529	0.603	0.681	0.763	0.851	0.946	1.050	1.167	1.303	1.468	1.684	2.021	2.423	2.704
41	0.529	0.603	0.681	0.763	0.850	0.945	1.050	1.167	1.303	1.467	1.683	2.020	2.421	2.701
42	0.528	0.603	0.680	0.763	0.850	0.945	1.049	1.166	1.302	1.466	1.682	2.018	2.418	2.698
43	0.528	0.603	0.680	0.762	0.850	0.945	1.049	1.166	1.302	1.466	1.681	2.017	2.416	2.695
44	0.528	0.602	0.680	0.762	0.850	0.945	1.049	1.166	1.301	1.465	1.680	2.015	2.414	2.692
45	0.528	0.602	0.680	0.762	0.850	0.944	1.049	1.165	1.301	1.465	1.679	2.014	2.412	2.690
46	0.528	0.602	0.680	0.762	0.850	0.944	1.048	1.165	1.300	1.464	1.679	2.013	2.410	2.687
47	0.528	0.602	0.680	0.762	0.849	0.944	1.048	1.165	1.300	1.463	1.678	2.012	2.408	2.685
48	0.528	0.602	0.680	0.762	0.849	0.944	1.048	1.164	1.299	1.463	1.677	2.011	2.407	2.682
49	0.528	0.602	0.680	0.762	0.849	0.944	1.048	1.164	1.299	1.462	1.677	2.010	2.405	2.680
50	0.528	0.602	0.679	0.761	0.849	0.943	1.047	1.164	1.299	1.462	1.676	2.009	2.403	2.678
	0.300	0.275	0.250	0.225	0.200	0.175	0.150	0.125	0.100	0.075	0.050	0.025	0.010	0.005

Tabla de Cuantiles de la Distribución Chi-Cuadrado

En el margen superior se lee $P(\chi^2 \leq x)$ para los valores de x que figuran en el cuerpo de la tabla y en el margen izquierdo los grados de libertad (ν).

ν	0.010	0.025	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0.0002	0.0010	0.0039	0.0158	0.0358	0.0642	0.1015	0.1485	0.2059	0.2750	0.3573	0.4549
2	0.0201	0.0506	0.1026	0.2107	0.3250	0.4463	0.5754	0.7133	0.8616	1.0217	1.1957	1.3863
3	0.1148	0.2158	0.3518	0.5844	0.7978	1.0052	1.2125	1.4237	1.6416	1.8692	2.1095	2.3660
4	0.2971	0.4844	0.7107	1.0636	1.3665	1.6488	1.9226	2.1947	2.4701	2.7528	3.0469	3.3567
5	0.5543	0.8312	1.1455	1.6103	1.9938	2.3425	2.6746	2.9999	3.3251	3.6555	3.9959	4.3515
6	0.8721	1.2373	1.6354	2.2041	2.6613	3.0701	3.4546	3.8276	4.1973	4.5702	4.9519	5.3481
7	1.2390	1.6899	2.1674	2.8331	3.3583	3.8223	4.2549	4.6713	5.0816	5.4932	5.9125	6.3458
8	1.6465	2.1797	2.7326	3.4895	4.0782	4.5936	5.0706	5.5274	5.9753	6.4226	6.8766	7.3441
9	2.0879	2.7004	3.3251	4.1682	4.8165	5.3801	5.8988	6.3933	6.8763	7.3570	7.8434	8.3428
10	2.5582	3.2470	3.9403	4.8652	5.5701	6.1791	6.7372	7.2672	7.7832	8.2955	8.8123	9.3418
11	3.0535	3.8157	4.5748	5.5778	6.3364	6.9887	7.5841	8.1479	8.6952	9.2373	9.7831	10.3410
12	3.5706	4.4038	5.2260	6.3038	7.1138	7.8073	8.4384	9.0343	9.6115	10.1820	10.7553	11.3403
13	4.1069	5.0088	5.8919	7.0415	7.9008	8.6339	9.2991	9.9257	10.5315	11.1291	11.7288	12.3398
14	4.6604	5.6287	6.5706	7.7895	8.6963	9.4673	10.1653	10.8215	11.4548	12.0785	12.7034	13.3393
15	5.2294	6.2621	7.2610	8.5468	9.4993	10.3070	11.0365	11.7212	12.3809	13.0297	13.6790	14.3389
16	5.8122	6.9076	7.9616	9.3122	10.3090	11.1521	11.9122	12.6244	13.3096	13.9827	14.6555	15.3385
17	6.4078	7.5642	8.6718	10.0852	11.1249	12.0023	12.7919	13.5307	14.2406	14.9373	15.6328	16.3382
18	7.0149	8.2307	9.3905	10.8649	11.9462	12.8570	13.6753	14.4399	15.1738	15.8932	16.6108	17.3379
19	7.6327	8.9065	10.1170	11.6509	12.7727	13.7158	14.5620	15.3517	16.1089	16.8504	17.5894	18.3377
20	8.2604	9.5908	10.8508	12.4426	13.6039	14.5784	15.4518	16.2659	17.0458	17.8088	18.5687	19.3374
21	8.8972	10.2829	11.5913	13.2396	14.4393	15.4446	16.3444	17.1823	17.9843	18.7683	19.5485	20.3372
22	9.5425	10.9823	12.3380	14.0415	15.2788	16.3140	17.2396	18.1007	18.9243	19.7288	20.5288	21.3370
23	10.1957	11.6885	13.0905	14.8480	16.1219	17.1865	18.1373	19.0211	19.8657	20.6902	21.5095	22.3369
24	10.8564	12.4011	13.8484	15.6587	16.9686	18.0618	19.0373	19.9432	20.8084	21.6525	22.4908	23.3367
25	11.5240	13.1197	14.6114	16.4734	17.8184	18.9398	19.9393	20.8670	21.7524	22.6156	23.4724	24.3366
26	12.1981	13.8439	15.3792	17.2919	18.6714	19.8202	20.8434	21.7924	22.6975	23.5794	24.4544	25.3365
27	12.8785	14.5734	16.1514	18.1139	19.5272	20.7030	21.7494	22.7192	23.6437	24.5440	25.4367	26.3363
28	13.5647	15.3079	16.9279	18.9392	20.3857	21.5880	22.6572	23.6475	24.5909	25.5093	26.4195	27.3362
29	14.2564	16.0471	17.7084	19.7677	21.2468	22.4751	23.5666	24.5770	25.5391	26.4751	27.4025	28.3361
30	14.9534	16.7908	18.4926	20.5992	22.1103	23.3641	24.4776	25.5078	26.4881	27.4416	28.3858	29.3360
31	15.6555	17.5387	19.2806	21.4336	22.9762	24.2551	25.3901	26.4397	27.4381	28.4087	29.3694	30.3359
32	16.3622	18.2907	20.0719	22.2706	23.8442	25.1478	26.3041	27.3728	28.3889	29.3763	30.3533	31.3359
33	17.0735	19.0466	20.8665	23.1102	24.7143	26.0422	27.2194	28.3069	29.3405	30.3444	31.3375	32.3358
34	17.7891	19.8062	21.6643	23.9523	25.5864	26.9383	28.1361	29.2421	30.2928	31.3130	32.3219	33.3357
35	18.5089	20.5694	22.4650	24.7966	26.4604	27.8359	29.0540	30.1782	31.2458	32.2821	33.3065	34.3356
36	19.2327	21.3359	23.2686	25.6433	27.3362	28.7350	29.9730	31.1152	32.1995	33.2517	34.2913	35.3356
37	19.9603	22.1056	24.0749	26.4921	28.2138	29.6355	30.8933	32.0532	33.1539	34.2216	35.2764	36.3355
38	20.6914	22.8785	24.8839	27.3429	29.0931	30.5373	31.8146	32.9919	34.1089	35.1920	36.2617	37.3354
39	21.4262	23.6544	25.6954	28.1958	29.9739	31.4405	32.7369	33.9316	35.0645	36.1628	37.2472	38.3354
40	22.1643	24.4330	26.5093	29.0505	30.8563	32.3450	33.6603	34.8719	36.0207	37.1340	38.2328	39.3353
41	22.9056	25.2145	27.3256	29.9071	31.7402	33.2506	34.5846	35.8131	36.9774	38.1055	39.2187	40.3353
42	23.6501	25.9987	28.1441	30.7654	32.6255	34.1574	35.5099	36.7550	37.9347	39.0774	40.2047	41.3352
43	24.3976	26.7853	28.9647	31.6255	33.5122	35.0653	36.4361	37.6975	38.8924	40.0496	41.1909	42.3352
44	25.1480	27.5746	29.7875	32.4871	34.4002	35.9744	37.3631	38.6408	39.8507	41.0222	42.1773	43.3352
45	25.9012	28.3661	30.6122	33.3504	35.2896	36.8844	38.2910	39.5847	40.8095	41.9950	43.1638	44.3351
46	26.6572	29.1601	31.4390	34.2152	36.1801	37.7955	39.2197	40.5292	41.7687	42.9682	44.1505	45.3351
47	27.4158	29.9562	32.2676	35.0814	37.0718	38.7075	40.1492	41.4744	42.7284	43.9417	45.1373	46.3350
48	28.1770	30.7545	33.0981	35.9491	37.9648	39.6205	41.0794	42.4201	43.6885	44.9154	46.1243	47.3350
49	28.9407	31.5549	33.9303	36.8182	38.8588	40.5344	42.0104	43.3664	44.6491	45.8895	47.1114	48.3350

Tabla de Cuantiles de la Distribución Chi-Cuadrado

En el margen superior se lee $P(\chi^2 \leq x)$ para los valores de x que figuran en el cuerpo de la tabla y en el margen izquierdo los grados de libertad (ν).

ν	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.999
1	0.5707	0.7083	0.8735	1.0742	1.3233	1.6424	2.0723	2.7055	3.8415	5.0239	6.6349	10.8278
2	1.5970	1.8326	2.0996	2.4079	2.7726	3.2189	3.7942	4.6052	5.9915	7.3777	9.2103	13.8150
3	2.6430	2.9462	3.2831	3.6649	4.1083	4.6416	5.3171	6.2514	7.8147	9.3484	11.3448	16.2667
4	3.6871	4.0446	4.4377	4.8784	5.3853	5.9886	6.7449	7.7794	9.4877	11.1433	13.2767	18.4670
5	4.7278	5.1319	5.5731	6.0644	6.6257	7.2893	8.1152	9.2364	11.0705	12.8325	15.0863	20.5147
6	5.7652	6.2108	6.6948	7.2311	7.8408	8.5581	9.4461	10.6446	12.5916	14.4494	16.8118	22.4577
7	6.8000	7.2832	7.8061	8.3834	9.0371	9.8033	10.7479	12.0170	14.0672	16.0128	18.4753	24.3215
8	7.8325	8.3505	8.9094	9.5245	10.2189	11.0301	12.0271	13.3616	15.5073	17.5345	20.0902	26.1248
9	8.8632	9.4136	10.0060	10.6564	11.3887	12.2421	13.2880	14.6837	16.9190	19.0228	21.6661	27.8768
10	9.8922	10.4732	11.0971	11.7807	12.5489	13.4420	14.5339	15.9872	18.3070	20.4832	23.2093	29.5881
11	10.9199	11.5298	12.1836	12.8987	13.7007	14.6314	15.7671	17.2750	19.6751	21.9201	24.7250	31.2645
12	11.9463	12.5838	13.2661	14.0111	14.8454	15.8120	16.9893	18.5493	21.0261	23.3367	26.2170	32.9094
13	12.9717	13.6356	14.3451	15.1187	15.9839	16.9848	18.2020	19.8119	22.3620	24.7356	27.6882	34.5288
14	13.9961	14.6853	15.4209	16.2221	17.1169	18.1508	19.4062	21.0642	23.6848	26.1189	29.1412	36.1237
15	15.0197	15.7332	16.4940	17.3217	18.2451	19.3107	20.6030	22.3071	24.9958	27.4884	30.5779	37.6976
16	16.0425	16.7795	17.5646	18.4179	19.3689	20.4651	21.7931	23.5418	26.2962	28.8454	32.0000	39.2529
17	17.0646	17.8244	18.6330	19.5110	20.4887	21.6146	22.9770	24.7690	27.5871	30.1910	33.4086	40.7896
18	18.0860	18.8679	19.6993	20.6014	21.6049	22.7595	24.1555	25.9894	28.8693	31.5264	34.8053	42.3123
19	19.1069	19.9102	20.7638	21.6891	22.7178	23.9004	25.3288	27.2036	30.1435	32.8523	36.1909	43.8211
20	20.1272	20.9514	21.8265	22.7745	23.8277	25.0375	26.4976	28.4120	31.4105	34.1696	37.5662	45.3147
21	21.1470	21.9915	22.8876	23.8578	24.9348	26.1711	27.6620	29.6151	32.6706	35.4789	38.9322	46.7966
22	22.1663	23.0307	23.9473	24.9390	26.0393	27.3014	28.8225	30.8133	33.9244	36.7807	40.2893	48.2681
23	23.1852	24.0689	25.0055	26.0184	27.1413	28.4288	29.9792	32.0069	35.1725	38.0757	41.6384	49.7280
24	24.2037	25.1063	26.0625	27.0960	28.2412	29.5533	31.1325	33.1962	36.4150	39.3641	42.9798	51.1785
25	25.2218	26.1430	27.1183	28.1719	29.3388	30.6752	32.2825	34.3816	37.6525	40.6465	44.3141	52.6197
26	26.2395	27.1789	28.1730	29.2463	30.4346	31.7946	33.4295	35.5632	38.8851	41.9232	45.6418	54.0516
27	27.2569	28.2141	29.2266	30.3193	31.5284	32.9117	34.5736	36.7412	40.1133	43.1945	46.9630	55.4766
28	28.2740	29.2486	30.2791	31.3909	32.6205	34.0266	35.7150	37.9159	41.3371	44.4608	48.2783	56.8922
29	29.2908	30.2825	31.3308	32.4612	33.7109	35.1394	36.8538	39.0875	42.5570	45.7223	49.5880	58.3008
30	30.3073	31.3159	32.3815	33.5302	34.7997	36.2502	37.9902	40.2560	43.7730	46.9793	50.8921	59.7024
31	31.3235	32.3486	33.4314	34.5981	35.8871	37.3591	39.1244	41.4217	44.9854	48.2319	52.1913	61.0983
32	32.3394	33.3809	34.4804	35.6649	36.9730	38.4663	40.2563	42.5848	46.1943	49.4804	53.4859	62.4871
33	33.3551	34.4126	35.5287	36.7307	38.0575	39.5718	41.3861	43.7452	47.3999	50.7251	54.7754	63.8701
34	34.3706	35.4438	36.5763	37.7954	39.1408	40.6757	42.5140	44.9032	48.6024	51.9660	56.0610	65.2461
35	35.3858	36.4746	37.6231	38.8591	40.2228	41.7780	43.6399	46.0588	49.8018	53.2034	57.3421	66.6198
36	36.4008	37.5049	38.6693	39.9220	41.3036	42.8788	44.7641	47.2122	50.9985	54.4373	58.6192	67.9842
37	37.4156	38.5349	39.7148	40.9839	42.3833	43.9782	45.8864	48.3634	52.1923	55.6680	59.8925	69.3463
38	38.4302	39.5643	40.7597	42.0450	43.4619	45.0763	47.0072	49.5126	53.3836	56.8955	61.1620	70.7037
39	39.4446	40.5935	41.8040	43.1054	44.5395	46.1730	48.1263	50.6598	54.5722	58.1201	62.4280	72.0541
40	40.4589	41.6222	42.8477	44.1649	45.6160	47.2685	49.2439	51.8051	55.7585	59.3417	63.6908	73.4022
41	41.4729	42.6506	43.8909	45.2236	46.6916	48.3628	50.3599	52.9485	56.9424	60.5606	64.9501	74.7456
42	42.4868	43.6786	44.9335	46.2817	47.7662	49.4560	51.4746	54.0902	58.1241	61.7768	66.2063	76.0844
43	43.5005	44.7063	45.9757	47.3390	48.8400	50.5480	52.5879	55.2302	59.3035	62.9904	67.4595	77.4185
44	44.5141	45.7336	47.0173	48.3957	49.9129	51.6389	53.6998	56.3686	60.4809	64.2014	68.7095	78.7503
45	45.5274	46.7607	48.0584	49.4517	50.9849	52.7288	54.8105	57.5053	61.6562	65.4101	69.9569	80.0774
46	46.5407	47.7874	49.0991	50.5071	52.0562	53.8177	55.9199	58.6405	62.8296	66.6165	71.2014	81.3999
47	47.5538	48.8139	50.1394	51.5619	53.1267	54.9056	57.0281	59.7743	64.0011	67.8207	72.4432	82.7201
48	48.5668	49.8401	51.1792	52.6161	54.1964	55.9926	58.1352	60.9066	65.1708	69.0226	73.6827	84.0379
49	49.5796	50.8659	52.2186	53.6697	55.2653	57.0786	59.2411	62.0375	66.3386	70.2224	74.9194	85.3511

14. Soluciones de ejercicios

Capítulo 1

Ejercicio 1

a) Experimental.

b) Severidad (cualitativa ordinal). Rendimiento (cuantitativa continua).

c) Tratamiento (Variable cualitativa nominal), con tres niveles: Sin pulverizar, F1 y F2. Destino (variable cualitativa dicotómica o binaria), con dos niveles: comercial y semilla.

d) Población de tubérculos-semillas que no fueron pulverizados, población de tubérculos semillas al que se les aplicó el fungicida 1 (F1) y población de tubérculos semillas al que se les aplicó el fungicida 2 (F2).

e) n=3.

f) La asociación entre severidad y rendimiento.

g) Medidas resumen, tablas y gráficos.

Ejercicio 2

a) Uno de los técnicos (Técnico 1) propone seleccionar al azar 100 productores y clasificarlos según lo especificado para cada variable. Otro técnico (Técnico 2) piensa que primero deberían separar las planillas según el tipo de manejo y luego elegir al azar 25 productores de cada tipo de manejo clasificándolos según la producción de leche, teniendo también un total de 100 productores.Observacional.

b) Tabla de contingencia

Producción promedio

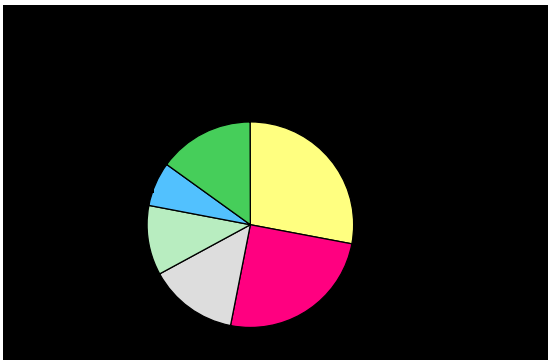
Tratamiento	Alta	Media	Baja	Total
Verdeo	7	11	8	26
Suplemento	14	10	7	31
Verdeo y Suplemento	12	8	5	25
Ninguno	4	6	8	18
Total	37	35	28	100

Ejercicio 3

a) Cuantitativa discreta.
b) Cualitativa nominal o binaria.
c) Cuantitativa discreta.
d) Cuantitativa continua.
e) Cuantitativa continua.
f) Cualitativa ordinal.
g) Cuantitativa continua.

Soluciones de ejercicios

Ejercicio 4



Ejercicio 5

a) 4º. b) 1º. c) 3º. d) 2º.

Ejercicio 6

Producción (t/ha)	FA	FR	FAA	FRA
(17 - 23]	5	0,06	5	0,06
(23 - 28]	21	0,26	26	0,33
(28 - 34]	25	0.31	51	0,64
(34 - 39]	17	0,21	68	0,85
(39 - 45]	9	0,11	77	0,96
(45 - 50]	3	0,04	80	1,00

a) 6%.

b) 36%.

c) 12 productores.

d) (28 - 34]. Es el valor (marca de clase=31) de la variable producción en t/ha cuya ubicación en la distribución, deja por debajo una proporción del 0.5 del total de los datos.

e) (34 - 39]. Es el valor (marca de clase=36.5) de la variable producción en t/ha cuya ubicación en la distribución, deja por debajo una proporción del 0.85 del total de los datos.

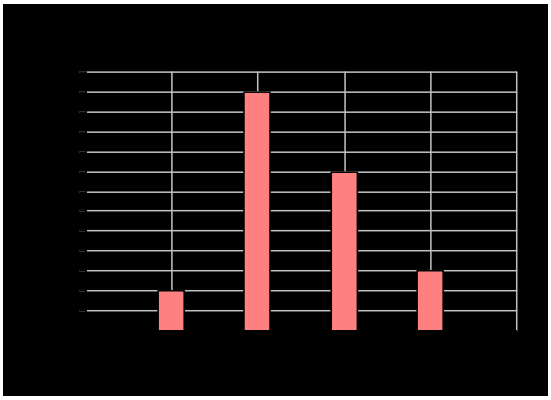
f) Polígono de frecuencias acumuladas.

Ejercicio 7

a) Distribución de frecuencias de la variable número de dientes por hoja

Clase	MC	FA	FR	FAA	FRA
1	1	2	0,08	2	0,08
2	2	12	0,48	14	0,56
3	3	8	0,32	22	0,88
4	4	3	0,12	25	1,00

b)



c) 8%

d) 44%.

Ejercicio 8

a) Medidas resumen

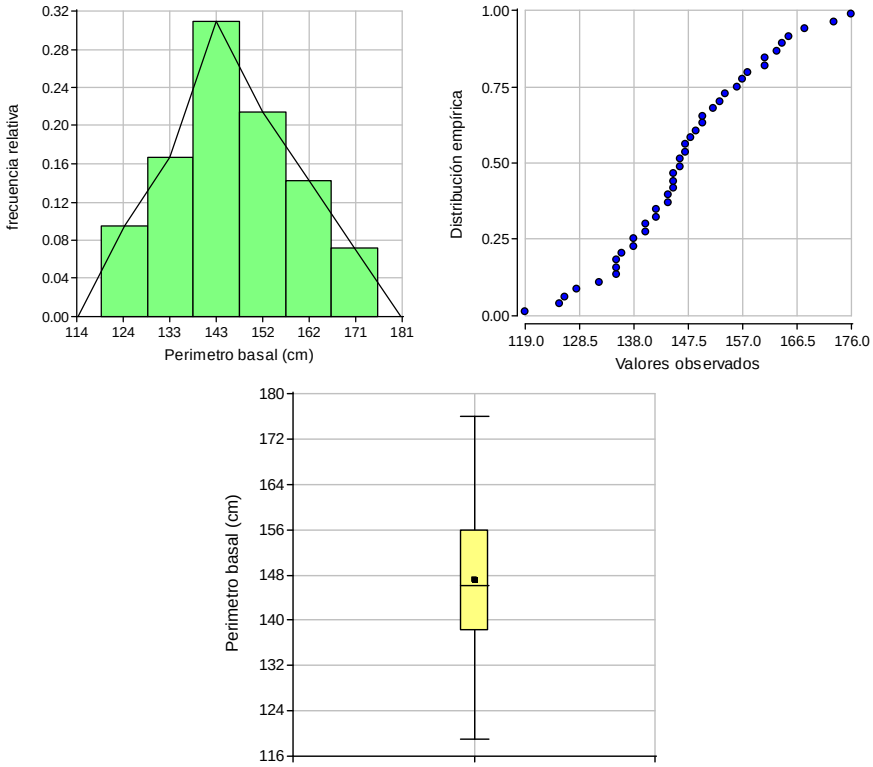
Media	*807,2*
Mediana	*805*
Max.	*995*
Min.	*606*
Rango	*389*
Varianza (n-1)	*10595.3*
D.E.	*102,9*
CV	*12,7*

b)

I. F
II. F
III. V
IV. V
V. F
VI. V
VII. V

Ejercicio 9

a)



Histograma de frecuencias relativas con polígono de frecuencias (arriba izquierda), grafico de distribución empírica (arriba derecha) y grafico de cajas (Box-Plot) (Abajo).

b) El gráfico de distribución empírica permite una lectura directa de los cuantiles.

c)

Medidas resumen	
n	42
Media	147.1
D.E.	12.9
Var(n-1)	166.9
CV	8.8
Mín	119
Máx	176
Mediana	146
P(25)	138
P(75)	156

d) Si.

Soluciones de ejercicios

Ejercicio 10

a) Se recomendaría el híbrido B.

I. Se recomendaría el híbrido A.
II. V
III. F
IV. V
V. F
VI. V
VII. F
VIII. V
IX. V
X. V
XI.
*XII.*F

Capítulo 2

Ejercicio 1

a) Clásico o basado en el espacio probabilístico.
b) No.
c) 1
d) 4/9
e)

y	F(y)
2	1/9
3	3/9
4	6/9
5	8/9
6	1

Ejercicio 2

a) Evento A= “obtener un nivel de producción alto”
b) Frecuencial
c) P(A)=80/320=0,25
d) Evento B=”obtener un nivel bajo de producción y ser productor del grupoA”. P(B)=75/320= 0,234375
e) Evento C=”obtener un nivel bajo de producción dado que el productor pertenece al grupo A”. P(C)=75/120=0,625. Probabilidad condicional.

Ejercicio 3

a) X=Cantidad de tractores vendidos por día
b) La variable tiene 5 posibles resultados. La variable es de tipo discreta
c) P(A)=110/260
d) P(A)=P(x=3)+P(x=4 o más)=25/260+10/260=35/260= 0,1346
e) P(A=vender 3 tractores mañana y vender 3 tractores pasado mañana)

Ejercicio 4

a) Si son mutuamente excluyentes
b) Si son estadísticamente dependientes

Ejercicio 5

a) $P(S)+P(T)+P(U)+P(PG)=\frac{210}{1596}+\frac{35}{1596}+\frac{36}{1596}+\frac{5}{1596}=\frac{286}{1596}=0{,}1792$

b) P(menor de 25 años)=271/1596= 0.1698

c) Si, son mutuamente excluyentes. No son independientes

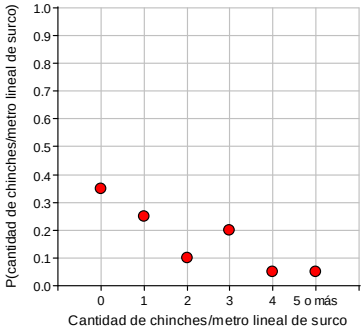
d) $P(T)+P(U)=\frac{5}{715}+\frac{10}{715}=\frac{15}{715}=0.021$

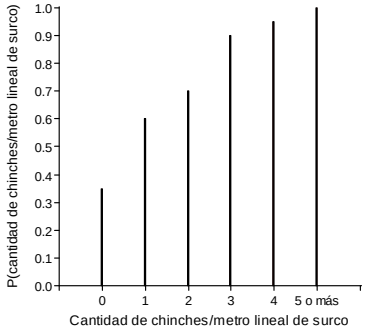
Ejercicio 6

La probabilidad de que un productor sea pequeño o mediano es 0,79. Son eventos mutuamente excluyentes.

Ejercicio 7

a) Función de probabilidad y distribución acumulada de la variable.





b) P(X=3)+ P(X=4)+ P(X=5 o más)=0,2+0,05+0,05=0,3

c) E(X)=0×0,35+1×0,25+2×0,10+3×0,2+4×0,05+5×0,05=1,5

d) La varianza de la variables es

$$V(X)=(0-1{,}5)^2 0{,}35+(1-1{,}5)^2 0{,}25+(2-1{,}5)^2 0{,}1+(3-1{,}5)^2 0{,}2+(4-1{,}5)^2 0{,}05+(5-1{,}5)^2 0{,}05=2{,}25$$

Capítulo 3

Ejercicio 1

a) 0.9032 ; b) 1 ; c) 0.0968 ; d) 0.68268 ; e) 0.14988, f) 0

Ejercicio 2

a) 0.3085 ; b) 0.383

Ejercicio 3

a) 0.0227; b) 0.6827

Ejercicio 4

a) x = 17.022 micrones ; b) el 75% de la distribución de la variable diámetro de un sedimento, comprende valores menores o iguales a 17 micrones.

Ejercicio 5

a) 0.2266 ; b) 0.2902

Ejercicio 6

a) Consumo en fresco: 0.3618×300000=108540 l ; Consumo de queso:0.3984×300000=119520 l y Consumo de leche en polvo: 0.2397×300000=71910 l

Ejercicio 7

Proporción de huevos con espesor de cáscara menor a 10 cmm= 0.0062. Cantidad de huevos con espesor de cáscara menor a 10 cmm= 5000×0.0062=31. Se romen 15.5 huevos con espesor de cáscara menor a 10 cmm

Proporción de huevos con espesor de cáscara comprendido entre 10 y 30 cmm=0.9876. Cantidad de huevos con espesor de cáscara entre 10 y 30 cmm=4938. Se rompen el 10%=493.8 huevos

a) Se rompen: 15 huevos + 494 huevos=509 huevos. Llegan sanos al consumidor=4491 huevos.

Ejercicio 8

a) Categoría I=0.17898×10000=1790 cajones, Categoría II=0.5107×10000=5107 cajones y Categoría III=0.3103×10000=3103 cajones

Ejercicio 9

a) La estrategia A produce un 52% de los frutos de la Categoría II y la B un 55%. Se elige la estrategia B.

Ejercicio 10

a) Proporción de granos que serán retenidos por el tamiz= 0.7977

b) Proporción de granos no retenidos por el tamiz de 8mm que serán retenidos por un tamiz de diámetro de malla igual a 7.5 mm=0.0967

c) Proporción de granos que pasará a través de los dos tamices= 0.1056

Ejercicio 11

a) E(cantidad de callos enraizados en cajas de petri)=1, V(cantidad de callos enraizados en cajas de petri)=0.8

b) P(X<4)=P(X=0)+P(X=1)+P(X=2)+P(X=3)=0.32768+0.4096+0.2048+0.0512= 0.99328

c) P(2 <X< 5)=P(X=3)+P(X=4)=0.0512+0.0064=0.0576

Ejercicio 14

a) P(X<6)= 0.1301414209

b) P(X<3)= 0.01033605068

c) P(X<10)=0.9863047314 (λ=5)

d) P(X=0)= 0.08208499862 (λ=2.5)

Ejercicio 15

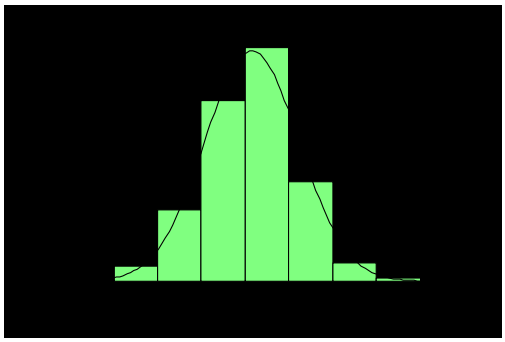
a) Binomial (p=0.70; n=10)

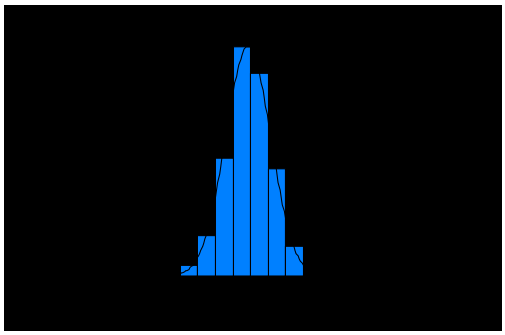
b) 7

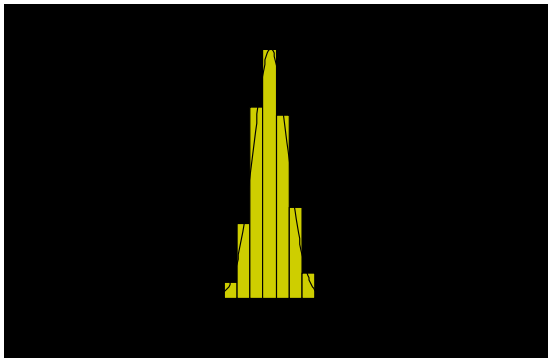
c) P(X=10)= 0.0282475249

Capítulo 4

Ejercicio 1







En los tres muestreos el promedio de las medias muestrales es similar al valor de de la media de la población a partir de la cual se obtienen las muestras y la aproximación es mayor cuando se usan muestras de mayor tamaño.
La varianza de las medias muestrales siempre resultó menor que la varianza poblacional. Esto ocurre porque en la distribución de las medias muestrales la varianza es afectada por el tamaño muestral, siendo cada vez menor a medida que crece el tamaño de la muestra.
Para estimar a la media poblacional de la variable Y es conveniente usar el mayor de los tamaños muestrales. El mayor tamaño muestral conduce a mayor confiabilidad porque produce que en la distribución de las medias, obtenidas con muestras de dicho tamaño, los valores se encuentren más cercanos al valor de su media poblacional el cual coincide con la media de la población de la que se extrajeron las muestras.

Ejercicio 2

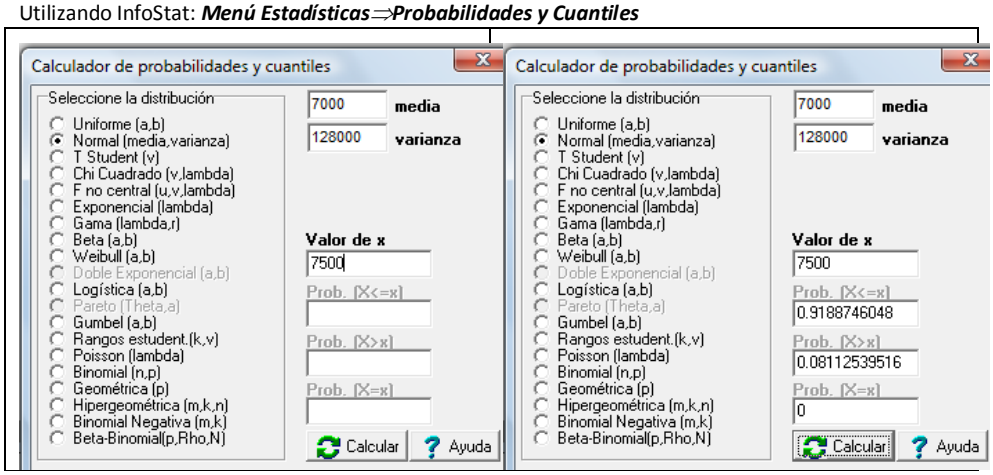
a)Falsa; b)Falsa;c)Verdadera;d)Falsa;e)Verdadera;f)Falsa; g)Verdadera

Ejercicio 3

a) $P\left(Z > \dfrac{7500-7000}{800/\sqrt{5}}\right) = 1 - P\left(Z \le 1,3975\right) \cong 1 - 0,91924 \cong 0,0876$

Soluciones de ejercicios

Utilizando InfoStat: ***Menú Estadísticas⇒Probabilidades y Cuantiles***

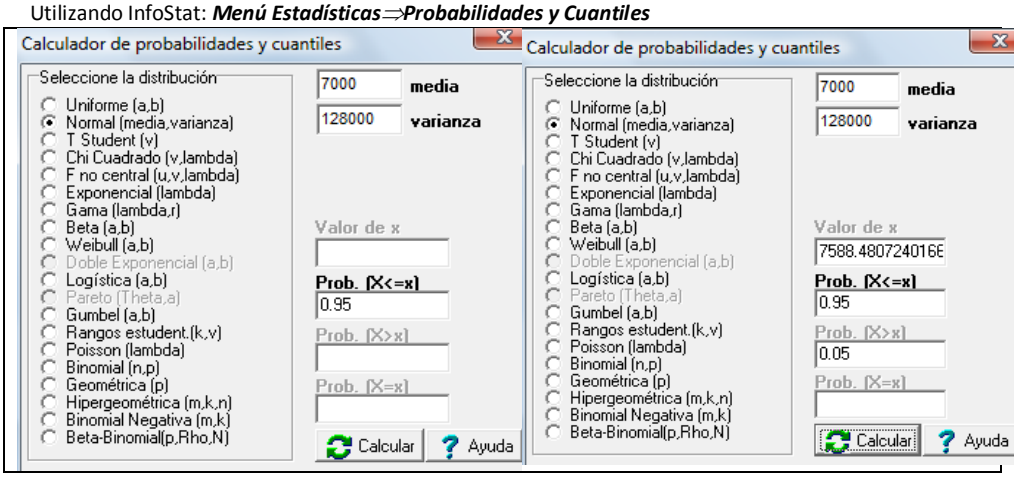


b)

$$P\left(Z > \frac{\bar{y}-7000}{800/\sqrt{5}}\right) = 0,05 \Rightarrow P\left(Z \leq \frac{\bar{y}-7000}{800/\sqrt{5}}\right) = 0,95 \Rightarrow z = 1,645$$

$$\frac{\bar{y}-7000}{800/\sqrt{5}} = 1,645 \Rightarrow \bar{y} = 1,645.\frac{800}{\sqrt{5}} + 7000 = 7588,53$$

Utilizando InfoStat: ***Menú Estadísticas⇒Probabilidades y Cuantiles***



Ejercicio 4

a) 0,85

b)0,65

Ejercicio 5

a) $P\left(\frac{S^2(n-1)}{\sigma^2} \leq \frac{23^2(50-1)}{20^2}\right) = P\left(\frac{S^2(n-1)}{\sigma^2} \leq 64,8\right) \cong 0,95$

b) $P\left(\frac{S^2(n-1)}{\sigma^2} \leq \frac{S^2(30-1)}{20^2}\right) = 0,99 \Rightarrow \frac{S^2(30-1)}{20^2} = 49,5880 \;\; ; \;\; S^2 = 683,97 \Rightarrow S = 26,15$

El 99% de los valores posibles para la desviación estándar en muestras de 30 parcelas son rendimientos menores o iguales a 26,15 kg/ha.

Capítulo 5 .

2)

a) Si α=0.05, [58.45 ; 61.55], amplitud=3.1.

Si α=0.01, [57.96; 62.04] amplitud=4.08;

b) Si α=0.05 y n=100 [59.02 ; 60.98] amplitud=1.96;

c) Si σ= 7, [57.83 ; 62.17], amplitud=4.34.

5)

Con $q_1 = T_{(48;0.025)} = -2.011$ y $q_2 = T_{(48;0.975)} = 2.011$, el intervalo será: [11.43 ; 12.57].

9)

a) n≅18

b) n≅71. El tamaño muestral aumenta porque se requiere un *n* mayor para mantener la misma amplitud de intervalo de confianza.

10)

a) Descartar H_0, Z=3.33;

b) LI=17.06; LS=22.94;

c) Se rechaza H_0

d) LI=16.14, LS=23.86;

e) Se rechaza H_0. La media es mayor que 15.

13)

a) H_0: μ = 45 H_1: μ >45.

b) T= 4.86. Valor de tabla $T_{(19;0.99)}$ = 2.539. Se rechaza H_0.

15)

Prueba T para un parámetro

Valor del parámetro probado: 80

Variable	n	Media	DE	LS(90%)	T	p(Unilateral I)
sem/m2	10	77.90	3.07	79.24	-2.16	0.0294

a) Para H_0: $\mu \geq 80$ versus H_1: $\mu < 80$, p=0.0294 es menor que α=0.10 se rechaza la hipótesis nula. La pérdida está dentro de los límites admisibles.

b) La pérdida es como máximo 79.24 sem/m^2 con una confianza del 90%.

17)

H_0: μ=500 vs. H_0: $\mu \neq 500$

Zona	n	Media	DE	LI(95%)	LS(95%)	
A	39	547.29	154.07	497.35		597.24
B	45	614.35	113.96	598.61		630.09

a) Los intervalos para la zona A contienen el valor μ=500, por lo que se aceptaría la hipótesis nula.

No sucede lo mismo en la zona B.

b) Los intervalos no se superponen, con lo cual si se esperaría encontrar diferencias estadísticamente significativas entre las medias de las precipitaciones observadas en cada zona.

Capítulo 6

1)

Prueba F para igualdad de varianzas

Variable	Grupo(1)	Grupo(2)	n(1)	n(2)	Var(1)	Var(2)	F	p	prueba
Día	{A}	{B}	12	12	1.97	0.20	9.63	0.0004	Unilateral

2)

Prueba T para muestras Independientes

Variable:Peso (g) - Clasific: Balanceado - prueba:Bilateral

	Grupo 1	Grupo 2
	A	B
n	12	12
Media	362.83	384.58
Media(1)-Media(2)	-21.75	
LI(95)	-60.47	
LS(95)	16.97	
pHomVar	0.0292	
T	-1.19	
p-valor	0.2523	

3)

a) Prueba T para muestras apareadas.

b) Normalidad e independencia.

c) y d)

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	LI(99%))	LS(99%))	T	Bilateral
Var A	Var B	6	-1.50	0.84	-2.88	-0.12	-4.39	0.0071

4)

a) Prueba T para muestras independientes.

b) Normalidad, homogeneidad de varianzas, independencia.

c) y d)

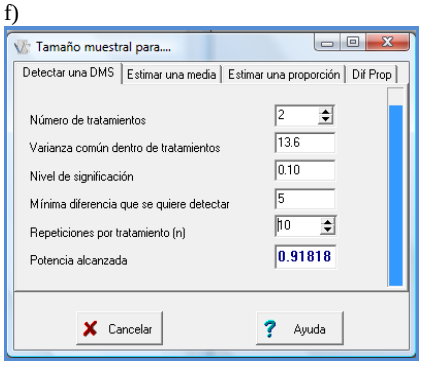
Prueba T para muestras Independientes

Variable:Rend (qq/ha) - Clasific:Herbicida - prueba:Bilateral

	Grupo 1	Grupo 2
	Nuevo	Tradicional
n	10	10
Media	64.50	61.68
Varianza	13.60	13.60
Media(1)-Media(2)	2.82	
LI(95)	-0.71	
LS(95)	6.34	
pHomVar	0.9227	
T	1.68	
p-valor	0.1104	

e) Opción 1.

f)



4)

Prueba T para muestras Independientes

Variable:Peso - Clasific:Grupo - prueba:Bilateral

	Grupo 1	Grupo 2
	Control	Experimental
n	10	12
Media	4.16	5.18
Media(1)-Media(2)	-1.02	
LI(95)	-2.22	
LS(95)	0.17	
pHomVar	0.8773	
T	-1.78	
p-valor	0.0900	

6)

Prueba T para muestras Independientes

Variable:Increm. - Clasific:Tratamiento - prueba:Bilateral

	Grupo 1	Grupo 2
	con poda	sin poda
n	10	10
Media	0.31	0.30
Media(1)-Media(2)	0.01	
LI(95)	-0.01	
LS(95)	0.03	
pHomVar	0.3108	
T	1.23	
p-valor	0.2361	

7)

Prueba T para muestras Independientes

Variable:Prod.Leche - Clasific:Lecitina - prueba:Unilateral

	Grupo 1	Grupo 2
	con	sin
n	9	8
Media	17.71	14.45
Media(1)-Media(2)	3.26	
pHomVar	0.7215	
T	7.25	
p-valor	<0.0001	

8)

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	Bilateral
Antes fist.	Despues fist.	8	0.22	0.50	1.26	0.2469

9)

Prueba T (muestras apareadas)

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	Bilateral
H1	H2	10	-4.80	3.05	-4.98	0.0008

10)

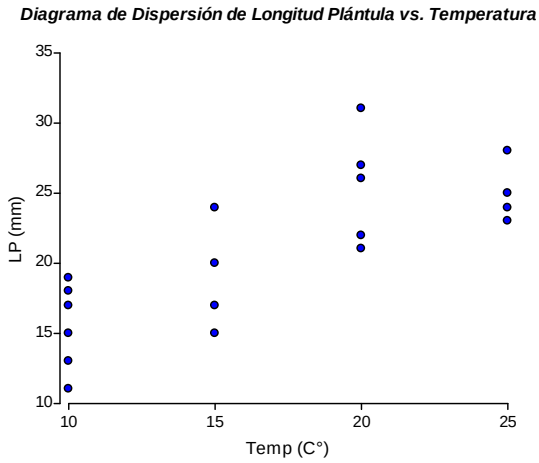
I. F
II. F
III. V
IV. V
V. F
VI. V
VII. V
VIII. F
IX. F
X. F

Capítulo 7

7)

a) En el experimento del ejemplo anterior se registra un solo valor de Y para cada X, en este ejemplo se tomaron varios valores de Y (longitud plántula) para cada valor de X (temperatura). Luego este conjunto de datos también podría analizarse con ANAVA para un modelo de efectos de tratamientos (temperatura)

b) El diagrama de dispersión sugiere que existe una tendencia lineal de la longitud de plántulas en el rango de temperaturas usadas en el experimento.



c) El modelo lineal es: $LP_{ij} = \alpha + \beta Temperatura_i + \varepsilon_{ij}$ con el supuesto de que los términos de error ε_{ij} son variables aleatorias independientes con distribución normal de media cero y varianza σ^2. Los estimadores de los parámetros (coeficientes) del modelo son a=8,69 (estimador de α, ordenada al origen) y b=0,72 (estimador de β, pendiente).

Análisis de regresión lineal

Variable	N	R^2
LP (mm)	19	0,60

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor
const	8,69	2,54	3,32	14,06	3,42	0,0033
Temp (C°)	0,72	0,14	0,42	1,02	5,04	0,0001

Cuadro de Análisis de la Varianza

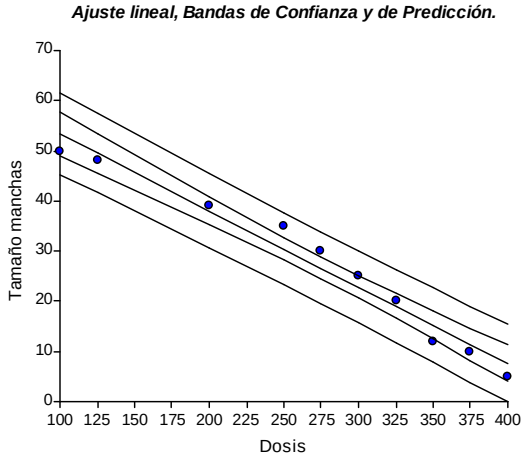
F.V.	SC	gl	CM	F	p-valor
Modelo	317,86	1	317,86	25,41	0,0001
Temp (C°)	317,86	1	317,86	25,41	0,0001
Error	212,66	17	12,51		
Total	530,53	18			

d) Desde el cuadro de ANAVA se desprende que el Modelo explica una parte significativa de la variación en los datos, dado que el valor-p asociado a la hipótesis nula que postula que las variaciones en LP no son explicadas por la relación lineal con la temperatura, es menor que el nivel de significación propuesto. La recta ajustada expresa el valor esperado de LP para cada temperatura. Como tiene pendiente positiva, a mayor temperatura se debe esperar mayor longitud, i.e. a 25°C deberíamos esperar que las plantas germinadas muestren mayor vigor.

6)

a) El diagrama de dispersión sugiere que existe una tendencia lineal de pendiente negativa que modela el tamaño de las manchas en función de la dosis de fungicida usada en el experimento (mayor dosis, menor tamaño de mancha). Los estimadores de los parámetros (coeficientes) del modelo son a=68,49 (estimador de α, ordenada al origen) y b=-0,15 (estimador de β, pendiente). Desde el cuadro de ANAVA se desprende que el Modelo explica una parte significativa de la variación en el tamaño de las

manchas (P<0,0001). En la siguiente figura, se presenta el ajuste (recta central), las bandas de confianza (alrededor de la recta de ajuste) y las bandas de predicción (bandas exteriores).



b) Desde la recta ajustada se predice que el tamaño de la mancha para 260 gr.p.a/ha seria Y=68,49-0,15*260=29,49.

Análisis de regresión lineal

Variable	N	R^2	
Daño	10	0,97	

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor
const	68,49	2,79	62,06	74,92	24,56	<0,0001
Dosis	-0,15	0,01	-0,17	-0,13	-15,65	<0,0001

Cuadro de Análisis de la Varianza

F.V.	SC	gl	CM	F	p-valor
Modelo	2165,70	1	2165,70	245,06	<0,0001
Dosis	2165,70	1	2165,70	245,06	<0,0001
Error	70,70	8	8,84		
Total	2236,40	9			

Capítulo 9

1)

a) H_0: $\mu_1 = \mu_2 = \mu_3$ versus

H_1: Al menos un tipo de productor se diferencia de los otros en los rendimientos medios logrados,

donde μ_1 representa el rendimiento medio logrado por los productores independientes (Tipo de Productor I), μ_2 representa el rendimiento medio logrado por los productores grandes (Tipo de Productor II) y μ_3 representa el rendimiento medio logrado por los productores asociados a grandes productores (Tipo de Productor III).

Finalmente, el estudio es de tipo observacional, con fines comparativos.

b) **Análisis de la varianza**

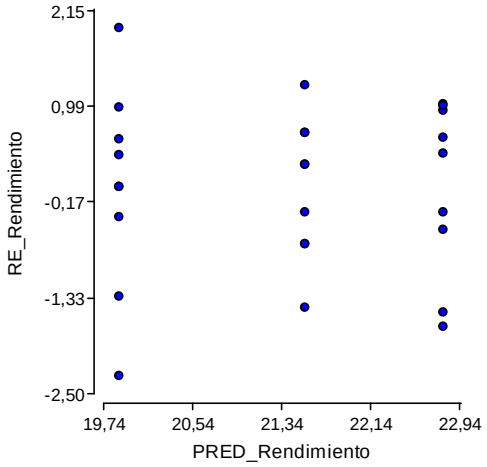
Variable	N	R^2	R^2 Aj	CV	
Rendimiento		27	0,05	0,00	25,78

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	38,24	2	19,12	0,63	0,5425
TipoProd1	38,24	2	19,12	0,63	0,5425
Error	731,35	24	30,47		
Total	769,59	26			

Fijando el nivel de significación en 0,05, como el valor p asociado a la hipótesis de nula acerca de la igualdad de media lograda por los distintos tipos de productores es mayor a 0.05 no se rechaza la hipótesis nula y se concluye que no existen diferencias significativas entre los distintos tipos de productores en cuanto a los rendimientos medios que logran alcanzar en el cultivo del maní.

c) Debemos generar los residuos, residuos estudentizados, valores absolutos de los residuos y los valores predichos –en primer lugar, para poder validar los supuestos solicitados en este punto. Para ello debe reconducirse el ANAVA del punto b) y en la solapa del Modelo en InfoStat tildar las celdas habilitadas a estos fines.
Para validar el supuesto de homogeneidad de varianzas se realiza la insepección visual del siguiente gráfico: el de los residuos estudentizados (RE_Rendimiento) vs. los valores predichos del modelo PRED_Rendimiento):



De la inspección visual de esta gráfica no se observa un fuerte patrón de heterogeneidad. Se conduce a continuación una prueba formal de homogeneidad de varianzas (Levene) basada en los valores absolutos de los residuos.

Análisis de la varianza

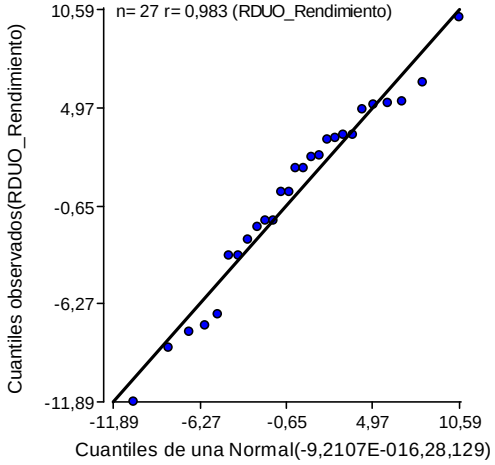
Variable	N	R^2	R^2 Aj	CV
RABS Rendimiento	27	0,03	0,00	72,87

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	6,01	2	3,01	0,31	0,7363
TipoProd1	6,01	2	3,01	0,31	0,7363
Error	232,59	24	9,69		
Total	238,60	26			

Como el valor p es 0,7363, al ser mayor que el nivel de significación, se termina aceptando la hipótesis nula de la Prueba de Levene que postula la homogeneidad de varianzas.

En segundo lugar, para evaluar normalidad, se realiza el gráfico QQ-plot de normalidad de los residuos (RDUO_Rendimiento), que se presenta a continuación:



La gráfica muestra que los residuos observados se alinean sobre una recta a 45°, mostrando que se correlacionan bien con los residuos esperados bajo el supuesto que los residuos tienen distribución Normal.

a) Debido a que no se rechaza la hipótesis nula de igualdad de medias del ANAVA en el punto b) es que no tiene sentido realizar ninguna de las pruebas de comparaciones múltiples conocidas, como la de Fisher sugerida.

b) EL modelo lineal adoptado para probar la hipótesis planteada en a) permite concluir que los distintos tipos de productores no logran producir rendimientos medios que sea significativamente diferentes entres sí. El modelo acredita términos de error aleatorios homogéneos en sus varianzas (p>0,05) y con distribución probablemente normal, lo que permite afirmar que la conclusión a la que se arriba es altamente probable que no sea equivocada.

2)

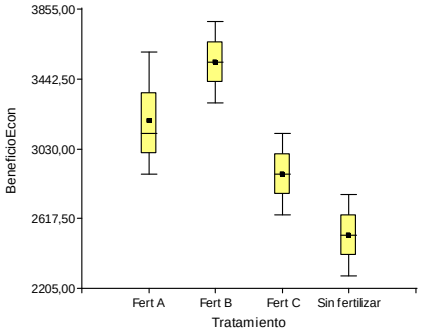
a) En base a la información presentada en este ejercicio, se construyó una tabla InfoStat, la que se presenta a continuación:

Nueva tabla

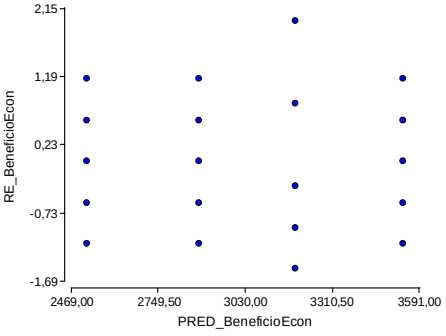
Caso	Tratamiento	Rendimiento	Costo	BeneficioEcon
1	Sin fertilizar	19,00	0,00	2280,00
2	Sin fertilizar	20,00	0,00	2400,00
3	Sin fertilizar	22,00	0,00	2640,00
4	Sin fertilizar	23,00	0,00	2760,00
5	Sin fertilizar	21,00	0,00	2520,00
6	Fert A	33,00	5,00	3360,00
7	Fert A	35,00	5,00	3600,00
8	Fert A	29,00	5,00	2880,00
9	Fert A	31,00	5,00	3120,00
10	Fert A	30,00	5,00	3000,00
11	Fert B	33,00	3,50	3540,00
12	Fert B	31,00	3,50	3300,00
13	Fert B	35,00	3,50	3780,00
14	Fert B	34,00	3,50	3660,00
15	Fert B	32,00	3,50	3420,00
16	Fert C	28,00	2,00	3120,00
17	Fert C	24,00	2,00	2640,00
18	Fert C	25,00	2,00	2760,00
19	Fert C	26,00	2,00	2880,00
20	Fert C	27,00	2,00	3000,00

Real | registros: 20*4

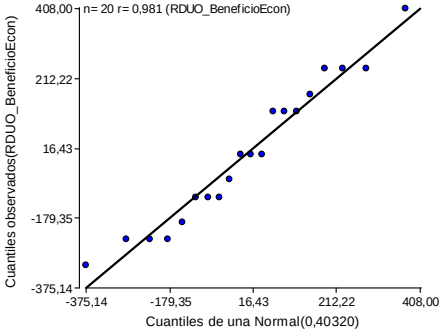
Una posible representación gráfica de interés estadístico es el Box-Plot de la Variable Beneficio Económico:



1) Verificación de Homogeneridad de Varianzas:



2) Verificación de Normalidad de los términos de error:



Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
BeneficioEcon	20	0,79	0,75	7,21

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor		
Modelo	2844540,00	3	948180,00	19,80	<0,0001		
Tratamiento	2844540,00	3	948180,00	19,80	<0,0001		
Error	766080,00	16	47880,00				
Total	3610620,00	19					

Test:LSD Fisher Alfa=0,05 DMS=293,37521

Error: 47880,0000 gl: 16

Tratamiento	Medias	n	E.E.			
Sin fertilizar	2520,00	5	97,86	A		
Fert C	2880,00	5	97,86		B	
Fert A	3192,00	5	97,86			C
Fert B	3540,00	5	97,86			D

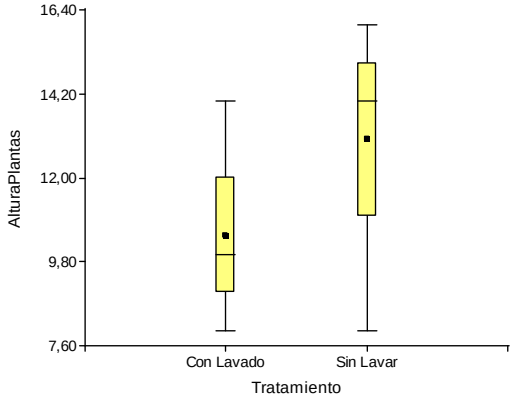
Medias con una letra común no son significativamente diferentes(p<= 0,05)

b) En base a los gráficos se puede concluir que los supuestos de homogeneidad de varianzas y normalidad de los términos de error no se violarían, lo que permite interpretar el valor p del ANAVA sin mayores riesgos a cometer equívocos a la hora de concluir.
Atento a que el valor p del test F de Tratamiento en la tabla del ANAVA es <0,0001, se puede concluir que existen diferencias significativas (p<0.05) en los beneficios económicos medios logrados bajo los distintos tratamientos, rechazando así la hipótesis nula del ANAVA.

Por último, el test d Fisher permite concluir que el Tratamiento con el Fertilizante B genera los beneficios económicos medios más altos respecto de los otros tratamients, con una media de $/ha de 3,540,=. Le sigue el Tratamiento con el Fertilizante A con una media de $/ha de 3192,=, el Fertilizante C con $/ha de 2,.880,=. Finalmente no convendría no fertilizar, ya que muestra lograr beneficios económicos significativamente menores, con una media de $/ha de 2.520,=

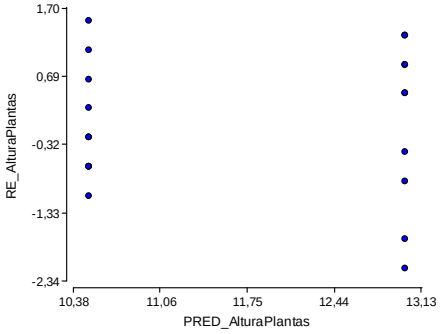
3)

a) Una posible representación gráfica de interés estadístico es el Box-Plot de la Variable Altura de Plantas, en el que se puede observar que no existirían diferencias significativas entre las medias, ya que las variabilidades presentadas por cada tratamiento harían que los intervalos de confianza al 95% se superpongan. Se probara esta afirmación directamente con la Prueba F del ANAVA:

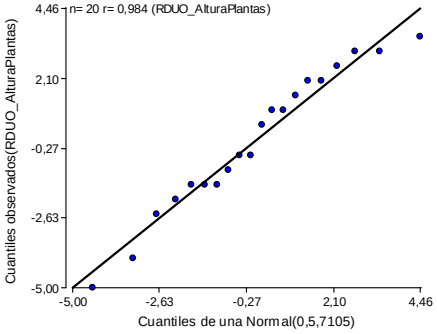


c)

1) Verificación de Homogeneridad de Varianzas



2) Verificación de Normalidad de los términos de error



Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
AlturaPlantas	20	0,22	0,18	20,89

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo		31,25	1	31,25	5,18	0,0352
Tratamiento		31,25	1	31,25	5,18	0,0352
Error	108,50	18	6,03		
Total	139,75	19			

Test:LSD Fisher Alfa=0,05 DMS=2,30677

Error: 6,0278 gl: 18

Tratamiento	Medias	n	E.E.		
Con Lavado	10,50	10	0,78	A	
Sin Lavar	13,00	10	0,78		B

Medias con una letra común no son significativamente diferentes(p<= 0,05)

b)

Prueba T para muestras Independientes

Clasific	Variable	Grupo 1	Grupo 2
Tratamiento	AlturaPlantas	{Con Lavado}	{Sin Lavar}

n(1)	n(2)	Media(1)	Media(2)	pHomVar	T	p-valor	prueba
10	10	10,50	13,00	0,2710	-2,2769	0,0352	Bilateral

Con InfoStat se generó esta tabla trabajando con cuatro decimales, de la que tomando el valor T=-2,2769 al cuadrado se verifica que coincide con el valor F=5,18 de la tabla del ANAVA.

c) En base a los gráficos se puede concluir que los supuestos de homogeneidad de varianzas y normalidad de los términos de error no se violarían, lo que permite interpretar el valor p del ANAVA sin mayores riesgos a cometer equívocos a la hora de concluir.
Atento a que el valor p del test F de Tratamiento en la tabla del ANAVA es 0,0352, se puede concluir que existen diferencias significativas ($p<0,05$) en las alturas de plantas logradas por los dos tratamientos, rechazando así la hipótesis nula del ANAVA. Por último, el test d Fisher permite concluir que el lavado de las estacas genera plantas significativamente más bajas en promedio que el tratamiento sin lavar.

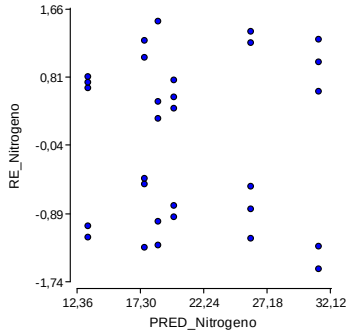
4)

a) Las macetas constituyen la Unidades Experimentales. Hay cinco macetas por Cepa, por lo que hay cinco repeticiones por Tratamiento (esto es, Cepa!).

b) H_0: $\mu_1 = \mu_2 = \ldots = \mu_5$ versus
H_1: Al menos una cepa se diferencia de las otras cepas en la cantidad media de Nitrógeno fijado,

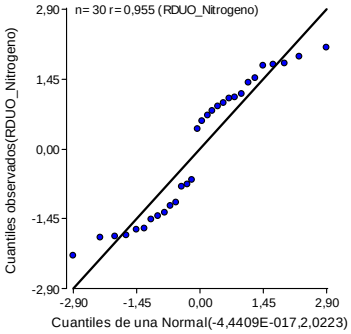
c)

1) Verificación de Homogeneidad de Varianzas:



Este gráfico permite suponer que el supuesto de homogeneidad de varianzas de los términos de error no se violaría.

2) Verificación de Normalidad de los términos de error:



Este gráfico muestra que el supuesto de normalidad podría no cumplirse ya que los residuos observados no se alinean sobre una recta a 45°, mostrando que se correlacionarían muy bien con los residuos esperados bajo el supuesto que los términos de error tienen distribución Normal. Esta situación podría alterar la calidad de la estimación del valor p en el test F del ANAVA.

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Nitrogeno	30	0,95	0,94	7,40

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	1034,08	5	206,82	84,63	<0,0001
Cepa	1034,08	5	206,82	84,63	<0,0001
Error	58,65	24	2,44		
Total	1092,73	29			

Test:LSD Fisher Alfa=0,05 DMS=2,04051

Error: 2,4437 gl: 24

Cepa	Medias	n	E.E.					
V	13,26	5	0,70	A				
III	17,64	5	0,70		B			
VI	18,70	5	0,70		B	C		
IV	19,92	5	0,70			C		
II	25,98	5	0,70				D	
I	31,22	5	0,70					E

Medias con una letra común no son significativamente diferentes(p<= 0,05)

c) Considerando que el valor p del test F de Tratamiento en la tabla del ANAVA para Cepa(Tratamiento) es <0,0001, se puede concluir que existen diferencias significativas (p<0,05) en la cantidad de nitrógeno fijado por las distintas Cepas evaluadas en el experimento, rechazando así la hipótesis nula

del ANAVA. La prueba de Fisher permite concluir que la Cepa que menos fija, significativamente, es la V; que la que más fija es la Cepa I y en segundo lugar la Cepa II; en tanto no se puede concluir entre las Cepa III, VI y IV, ya que comparten letras, destacando que presentan medias significativamente distintas de la media de la Cepa V y de la Cepa II.

5)

a) $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ donde:

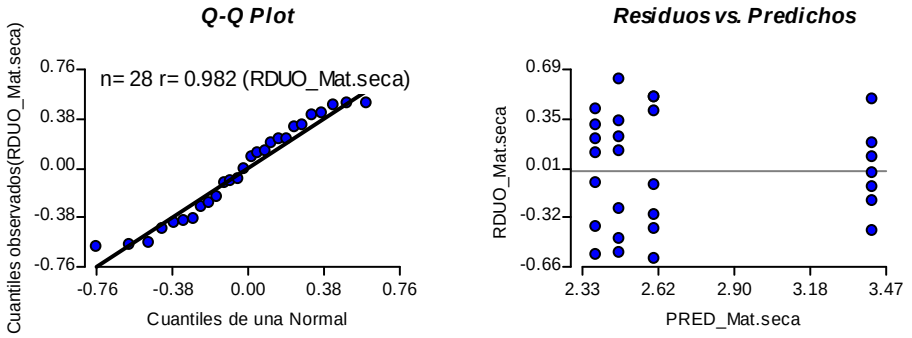
Y_{ij} = es la j-ésima observación de materia seca bajo la i-ésima carga animal, i=2, 4, 6, 8 (esto es, cuatro tratamientos) y j=1,…,7 (n=7)

μ= media general de materia seca.

τ_i = efecto de la i-ésima carga animal,

ε_{ij} = variable aleatoria normal, independientemente distribuida con esperanza cero y varianza σ^2 $\forall$ i j.

b) ε_{ij} están normal e independientemente distribuidos con esperanza cero y varianza σ^2. Para estudiar el cumplimiento de estos supuestos se recurre a métodos gráficos (QQ-plot para normalidad, Residuos vs predichos para homocedasticidad)



El análisis de las figuras precedentes pemitiría asumir que los supuestos normalidad y homogeneidad de varianzas se cumplen.

c) **Cuadro de Análisis de la Varianza (SC tipo III)**

F.V.	SC	gl	CM	F	p-valor	
Modelo		4.69	3	1.56	9.84	0.0002
Tratamiento		4.69	3	1.56	9.84	0.0002
Error	3.81	24	0.16			
Total	8.50	27				

Como p=0.0002 es menor que α=0,05 se rechaza la hipótesis de efectos de tratamientos nulos, es decir al menos un tratamiento (carga animal) produce un efecto diferente. Se realiza la prueba "a posteriori" de Fisher:

Test: LSD Fisher Alfa:=0.05 DMS:=0.43964

Error: 0.1588 gl: 24

Tratamiento	Medias	n	
carga8	2.39	7	A
carga2	2.47	7	A
carga6	2.60	7	A

carga4	3.41	7	B

Letras distintas indican diferencias significativas(p<= 0,05)

Se recomienda la carga animal de 4 novillos/ha, porque es la carga que induce la mayor producción de materia seca, siendo estadísticamente diferente de la producción promedio inducida por resto de las cargas animales.

Capítulo 10

3)

a)

tratamiento	n	Media	E.E.	CV	Mín	Máx
A1	6	3.16	0.05	3.74	3.03	3.30
A2	6	3.15	0.06	4.71	2.93	3.33
B1	6	3.34	0.04	2.80	3.22	3.45
B2	6	3.38	0.05	3.41	3.20	3.54
control	6	3.24	0.05	4.06	3.10	3.48

b) $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $\quad Proteinas_{ij} = \mu + Tratamiento_i + Tambo_j + \varepsilon_{ij}$

c)

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
prot	30	0.86	0.80	2.05

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	0.55	9	0.06	13.76	<0.0001
trat	0.26	4	0.07	14.72	<0.0001
Tambo	0.29	5	0.06	13.00	<0.0001
Error	0.09	20	4.4E-03		
Total	0.64	29			

Test:LSD Fisher Alfa=0.05 DMS=0.08009

Error: 0.0044 gl: 20

trat	Medias	n	E.E.			
A2	3.15	6	0.03	A		
A1	3.16	6	0.03	A		
control	3.24	6	0.03		B	
B1	3.34	6	0.03			C
B2	3.38	6	0.03			C

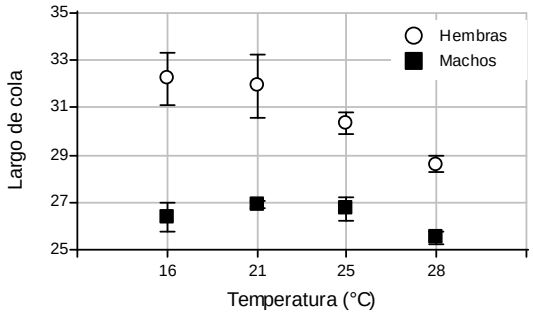
Medias con una letra común no son significativamente diferentes(p<= 0.05)

d) Suplemento B en cualquiera de sus dosis

4)

a) $Y_{ij} = \mu + Sexo_i + Temperatura_j + Sexo*Temperatura_{ij} + \varepsilon_{ij}$

b)



c)

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
largocola	32	0.80	0.74	4.97

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	192.98	7	27.57	13.68	<0.0001
Sexo	155.32	1	155.32	77.07	<0.0001
Temperatura	27.99	3	9.33	4.63	0.0108
Sexo*Temperatura	9.66	3	3.22	1.60	0.2159
Error	48.37	24	2.02		
Total	241.34	31			

d)

Las hembras siempre tienen mayor longitud de cola, independientemente de la temperatura. Es decir, no hay interacción estadísticamente significativa entre los efectos de los factores sexo y temperatura. No obstante, hay efecto estadísticamente significativo de sexo y temperatura diferente de cero.

.

5)

a) $Y_{ijk} = \mu + Sexo_i + Temperatura_j + Sexo*Temperatura_{ij} + Bloque_k + \varepsilon_{ijk}$

b)

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
largocola	32	0.83	0.75	4.92

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	199.79	10	19.98	10.10	<0.0001
Sexo	155.32	1	155.32	78.50	<0.0001
Temperatura	27.99	3	9.33	4.72	0.0114
Bloque 6.82	3	2.27	1.15	0.3528	
Sexo*Temperatura	9.66	3	3.22	1.63	0.2130
Error	41.55	21	1.98		
Total	241.34	31			

c)

El efecto de la temperatura es independiente del sexo para el largo de la cola (No hay interacción sexo*temperatura). Hay un efecto de sexo en la longitud de la cola (las hembras tiene mayor longitud de la cola que los machos) y hay un efecto de la temperatura. No hubo efecto de laboratorios (bloques).

6)

a)

I. V
II. F
III. V
IV. F
V. V
VI. F
VII. F
VIII. F

15. Índice de palabras clave

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

Esta obra se terminó de imprimir en el mes de
Marzo de 2012 en Editorial Brujas.
Córdoba-Argentina