



## Supuestos del Análisis de Varianza

Erika Kaniá Kuhl  
Ing. Agrónomo Dr.



Fisher propuso en la década de los 20 (1920-1930) los clásicos Modelos de Análisis de Varianza (Modelos de clasificación), Modelos de Regresión y Análisis de Covarianza.

Estos métodos fueron desarrollados para variables continuas

Supuestos de estos métodos (Métodos simplistas):

- Datos independientes (no correlacionados).
- Varianzas homogéneas.
- Distribución normal de los errores

Los supuestos son necesarios para la inferencia

(Valor P del ANDEVA no es válido en caso de incumplimiento)



Hoy en día existen técnicas que permiten modelar datos en los cuales las observaciones no son independientes y/o con varianzas heterogéneas y/o que no presentan distribución normal



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

El modelo lineal del ANAVA plantea supuestos que deben cumplirse para que el estadístico F tenga valores p reportados que sean válidos

$Y_{ij}$  es la j-ésima observación del i-ésimo tratamiento  
 $\mu$  es la media general común a todos los tratamientos  
 $\tau_i$  es el efecto fijo del tratamiento i  
 $\varepsilon_{ij}$  es una variable aleatoria normal (error), independientemente distribuida con esperanza 0 y varianza  $\sigma^2$ , que representa la variabilidad



Estos supuestos plantean exigencias acerca de los términos de error

$$\varepsilon_{ij}$$

Se pueden establecer como:

- Independencia entre los errores
- Homogeneidad de varianzas
- Distribución normal de los errores



En caso que alguno de estos supuestos no se cumplan:

Impactarán directamente sobre los valores p reportados, y por lo tanto sobre la calidad de las conclusiones que finalmente buscamos obtener.

La verificación de los supuestos se realiza en la práctica a través de los predictores de los términos de error aleatorio que son los residuos aleatorios asociados a cada observación



$e_{ij}$  = VALOR OBSERVADO – VALOR ESTIMADO

( $e_{ij}$  = residuo asociado a una unidad experimental)

Se calcula como la diferencia entre el valor observado de la variable respuesta y el valor esperado estimado (predicho) bajo el modelo lineal especificado.



### Supuesto Independencia de los errores

Debe existir independencia entre los errores de las unidades experimentales

Evitar la presencia de datos correlacionados experimentalmente

Hay tres casos en los que podemos tener problemas con este supuesto:

- 1) Tener más de un dato por UE (Submuestreo dentro de la UE)
- 2) Falta de aleatorización de los tratamientos a las UE
- 3) Temporal. Cuando a una misma UE se le mide el largo de brote en 5 oportunidades



### Archivo Duraznero

En un huerto de duraznero se condujo un ensayo bajo un Diseño Completamente aleatorizado con cuatro tratamientos (hormonas A, B, C y D) y 5 repeticiones por tratamiento, siendo la unidad experimental una planta. En cada planta se seleccionaron al azar 5 frutos en los cuales se evaluó el diámetro del fruto



### Archivo Duraznero

Submuestreo de frutos dentro de un árbol (UE): esos frutos carecen de independencia, no son independientes entre sí, por lo tanto hay una estructura de correlación dentro del árbol.

¿Cómo solucionamos este problema?

Sacando el promedio de los frutos por Unidad Experimental

Debe existir un solo dato por Unidad experimental para proceder con el análisis.

### Archivo Híbridos

Para comparar 4 variedades de maíz (1, 2, 3 y 4), plantados en el sector de Melipilla en un suelo franco-arcilloso homogéneo y con riego tradicional mediante surcos, se realizó un ensayo con 10 repeticiones por tratamiento. Las variedades fueron sembradas en el mes de noviembre a una densidad de plantación de 90.000 plantas por ha. Los tratamientos se establecieron en parcelas de 3,2 x 7 m (22,4 m<sup>2</sup>), con hileras separadas a una distancia de 0,8 m. La variable respuesta medida fue el rendimiento (toneladas · ha<sup>-1</sup>).



### Supuesto Independencia de los errores

El cumplimiento de este supuesto queda garantizado, pero no asegurado, mediante el procedimiento de **ALEATORIZACIÓN** de los tratamientos a las unidades experimentales, razón por la cual este procedimiento no debe obviarse.

**Supuesto Homogeneidad de varianzas**

Homocedasticidad = Homogeneidad

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_i^2$   
 $H_A : \text{no todas las } \sigma_i^2 \text{ iguales.}$

(al menos dos varianzas distintas)

**Supuesto Homogeneidad de varianzas**

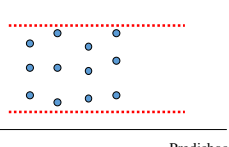
**PRUEBA GRÁFICA**

- Gráfico de dispersión de **residuos vs predichos**

**PRUEBA ESTADÍSTICA**

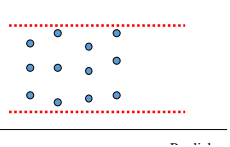
- Test de Levene

**Supuesto Homogeneidad de varianzas**



Realizando un gráfico de dispersión de los **RESIDUOS** v/s los valores **PREDICHOS** por el modelo, se puede observar si existe alguna tendencias sospechosa que sugiera que existe heterogeneidad de varianzas.

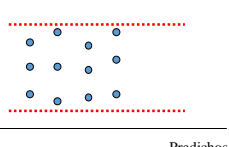
**Supuesto Homogeneidad de varianzas**



Cuando existe **homogeneidad de varianzas**, se debe observar una nube de puntos **sin patrón alguno** (patrón aleatorio).

A medida que aumenta lo predicho por el modelo no existe un aumento de la dispersión de los residuos.

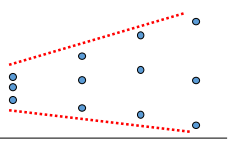
**Supuesto Homogeneidad de varianzas**



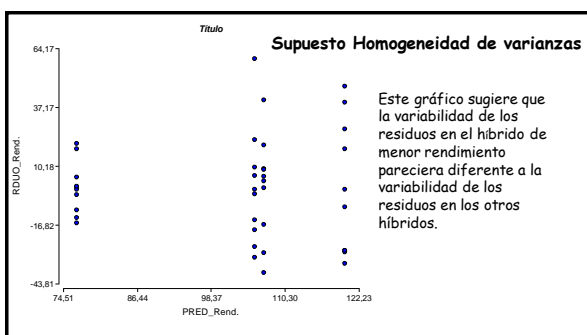
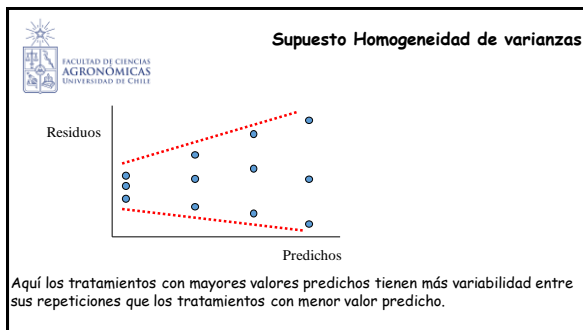
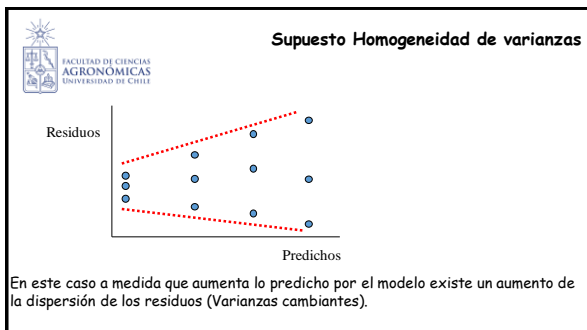
La variación de los residuos debe ser uniforme en todo el rango de los valores pronosticados.

A medida que **aumenta lo predicho** por el modelo **no existe un aumento de la dispersión** de los residuos (Varianzas cambiantes).

**Supuesto Homogeneidad de varianzas**



Cuando existe **heterogeneidad de varianzas**, se observa una estructura como este patrón típico que indica falta de homogeneidad de varianzas



**Supuesto Homogeneidad de varianzas**

- Prueba de Levene (INFOSTAT)

Esta prueba se construye realizando un análisis de la varianza con los valores absolutos de los residuos como variable dependiente.

Es preferible realizar pruebas gráficas para verificar este supuesto, ya que en caso de estar ante un escenario de heterogeneidad de varianzas, este tipo de prueba no nos entrega una estrategia de análisis a seguir.

**Supuesto Homogeneidad de varianzas**

ANÁLISIS DE LA VARIANZA

Variable	N	R <sup>2</sup>	R <sup>2</sup> Ajustado	CV
RABS Rend.	40	0,18	0,11	77,90

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	1528,31	3	509,44	2,64	0,0642
Cultivar	1528,31	3	509,44	2,64	0,0642
Error	6946,69	36	192,96		
Total	8475,00	39			

Para un nivel de significancia del 5 %, se acepta la hipótesis nula en que las varianzas son homogéneas (p-valor 0,064 > 0,05), es decir existe homocedasticidad de varianzas.

**Supuesto Distribución normal de los errores**

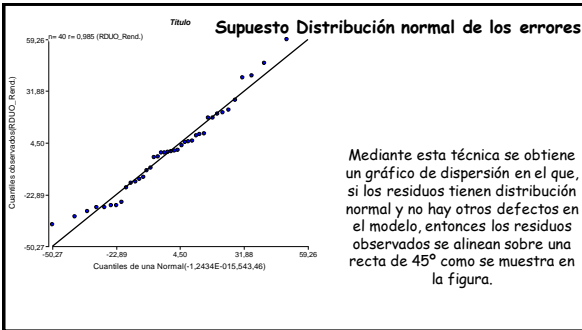
HO: Los errores tiene distribución normal.  
HA: Los errores no tiene distribución normal.

**PRUEBA GRÁFICA**

- Gráfico QQ-PLOT (INFOSTAT)

**PRUEBAS ESTADÍSTICAS**

- Shapiro Wilks (INFOSTAT)
- Kolmogorov-Smirnov
- Anderson Darling
- Ryan-Joiner
- Cramer von Mises
- Lilliefors



**Supuesto Distribución normal de los errores**

FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

Shapiro-Wilks (modificado)

Variable	n	Media	D.E.	W*	p (Unilateral D)
RDUO Rend.	40	0,00	23,31	0,95	0,2975

Como  $p\text{value} = 0,295 > 0,05$ , implica Aceptar  $H_0$ , es decir existe distribución normal de los errores al 5 % de nivel de significancia.

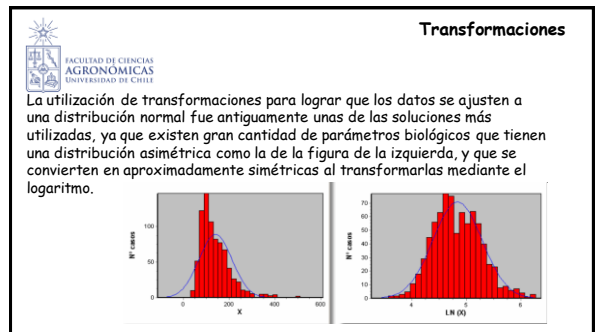
**¿Y si no se cumplen los supuestos?**

FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

**Métodos más contemporáneos:**

- Incluir una función de varianza en el modelo
- Modelos lineales generalizados

**Transformaciones y Estadística no paramétrica (60-70)**



**Transformaciones**

FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

-Se debe realizar una transformación de la variable original.

- Luego, a la variable transformada hay que calcularle sus RESIDUOS, los que deben ser sometidos nuevamente a las pruebas de los supuestos de normalidad y de homogeneidad de varianzas.

Si pasan los supuestos realizar el ANOVA con la variable transformada, luego las comparaciones múltiples se realizan con la escala transformada, pero los resultados finales (promedios de los tratamientos), deben presentarse con los valores de la escala original

**Transformaciones**

FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE


-Se debe realizar una transformación de la variable original.

- Luego, a la variable transformada hay que calcularle sus RESIDUOS, los que deben ser sometidos nuevamente a las pruebas de los supuestos de normalidad y de homogeneidad de varianzas.

Si pasan los supuestos realizar el ANOVA con la variable transformada, luego las comparaciones múltiples se realizan con la escala transformada, pero los resultados finales (promedios de los tratamientos), deben presentarse con los valores de la escala original

Si no pasan los supuestos, probar con otras transformaciones.

De lo contrario habrá que recurrir a la Estadística no Paramétrica.




FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

### Transformaciones

Transformaciones más utilizadas:

1. Transformación raíz cuadrada
2. Transformación logarítmica
3. Transformación de Bliss



FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

### Transformación raíz cuadrada

$\sqrt{y}$

Es recomendada cuando los datos "y" están expresados en números enteros no negativos, los cuales siguen una distribución de Poisson, como por ej:


- n° de arañitas rojas por hoja
- n° de bacterias en un cultivo

$Y^* = \sqrt{y}$

Si los valores observados son cercanos a cero o muy pequeños, se sugiere utilizar:

$Y^* = \sqrt{y + cte}$

Constante = 0.01 ó 0.5, etc.



FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

### Transformación Ln o Log


$\ln(y)$  ó  $\log(y)$

Consiste en tomar los logaritmos de las variables originales.

$Y^* = \log(y)$

Se usa cuando los datos son números enteros positivos que cubren un amplio rango de valores (o que tienen mucha variabilidad) como por ej:

- n° de nemátodos por muestra de suelo
- n° de células somáticas en la leche




FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

### Transformación Ln o Log

$\ln(y)$  ó  $\log(y)$

Hay problemas con la transformación logarítmica si la variable "y" toma algún valor "cero", por lo que en estos casos, o incluso si existen valores muy pequeños, será adecuada emplear la transformación:

$Y^* = \log(y + 1)$



FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE


### Transformación angular (Bliss)

Cuando los datos son proporciones o porcentajes como por ej:

- % Botrytis
- % Palo negro
- % Daño

$Y^* = \arcseno \sqrt{y/100}$

Donde Y es un valor entre 0% y 100%




FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

### Estadística paramétrica

En los **métodos paramétricos** se asume que la población de la cual la muestra es extraída presenta distribución NORMAL.

Esta propiedad es necesaria para que la prueba de hipótesis del ANDEVA sea válida.




**Estadística no paramétrica**

Se denominan pruebas **no paramétricas** aquellas que **no presuponen** una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre (*distribution free*).

La mayor desventaja de la estadística **no paramétrica** es que es mucho menos poderosa que la estadística paramétrica.


Esta basada en rangos, es decir, se considera el *orden relativo que le corresponde* a los datos más que a los valores mismos.



**Estadística no paramétrica**

Al ordenar los datos de  $>$  a  $<$  se obtienen rangos


y	Rango
7	2
20	5
40	6
13	3
2	1
17	4



**Estadística no paramétrica**  
**Prueba de Kruskal-Wallis**

Es equivalente a la prueba del ANOVA para un **Diseño Completamente aleatorizado** y que permite comparar la igualdad de k medias poblacionales (tratamientos) versus que al menos una es distinta.


Las hipótesis son las mismas que para el ANOVA.



**Estadística no paramétrica**  
**Prueba de Friedman**


Es una prueba no paramétrica equivalente a la prueba del ANOVA aplicable a modelos de clasificación en dos sentidos, sin interacción, como es el caso del **Diseño en Bloques completos al azar**.

Las hipótesis son las mismas que para el ANOVA.



**Estadística no paramétrica:**

- Fueron un buen intento para arreglar el problema (60-70)
- Paramétrico: supone distribución
- No paramétrico: de distribución libre
- Esta obsoletos, y perdiendo vigencia a pasos agigantados
- Es una estadística de distribución libre, pero si exige independencia
- Ataca el problema de la no normalidad a través de una transformación de rangos
- Al asignar rangos, se pierde mucha información de las diferencias entre los tratamientos



**Modelación estadística avanzada**



FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

## Modelación estadística avanzada

### ANTES

- La clásica receta de que un experimento hay que analizarlo de una manera en particular esta obsoleta!

### AHORA

- Para un mismo set de datos hay que evaluar diferentes modelos, modelos alternativos.
- Posteriormente hay que aplicar criterios de selección de modelos
- Hay que pensar en toda la fuente de variación que hay en el experimento y colocarla en el modelo, efecto unidad experimental (sujeto), efecto covariable, efecto bloque, etc....y evaluar.
- Según la naturaleza de la variable respuesta se decide cual modelo usar, modelo lineal general, modelo lineal generalizado,



FACULTAD DE CIENCIAS  
AGRONÓMICAS  
UNIVERSIDAD DE CHILE

## Supuestos del Análisis de Varianza

Erika Kania Kuhl  
Ing. Agrónomo Dr.