Tarea 1

IN3242 - Estadística
Departamento de Ingeniería Civil Industrial, Universidad de Chile
Primavera 2023
Profesor Raimundo Undurraga
Auxiliares: Matías Reyes, Leonardo Meneses, Bastián Urzúa, Brandon Galarza, Antonia Villegas, Camila Jáuregui, Francisca Monetta
Puntaje Total: 375 puntos
Fecha límite de entrega: Viernes 8 de septiembre hasta las 23:59 hrs.

I. Conceptos Básicos de Estadística

1. Sea W una variable aleatoria que distribuye normal de parámetros $(\eta, 49m^2)$. Encuentre $\mathbb{P}(W > \eta + m)$ [10 puntos]

Definamos:

$$\mathbb{P}(W > \eta + m) = \mathbb{P}\left(Z > \frac{\eta + m - \eta}{7m}\right)$$

$$= \mathbb{P}(Z > 1/7) = 1 - \mathbb{P}(Z < 1/7) = 1 - 0.5568 = 0.4432$$

2. Las variables aleatorias y y x distribuyen como una normal con medias 4 y 2, varianzas 16 y 4 y covarianza 6.

(i) Calcule la pendiente y el intercepto en la función de esperanza condicional $E[y|x]$ [5 puntos]
Utilizando la siguiente formula

$$\mathbb{E}(Y|X) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(X - \mu_X)$$

Donde $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{3}{4}$

$$\mathbb{E}(Y|X) = 4 + \frac{4}{2} \cdot \frac{3}{4}(X - 2) = 1 + \frac{3}{2}X$$

Luego la pendiente es igual a $\frac{3}{2}$ y el intercepto sucede cuando $X = 0$, es decir, el intercepto es igual a 1.

(ii) Calcule la correlación cuadrada entre y y x [5 puntos]
Como ya se calculó ρ, solo basta elevarla al cuadrado, por lo tanto, $\rho^2 = \frac{9}{16}$

(iii) Calcule la correlación cuadrada entre y y $E[y|x]$ [10 puntos]
En este caso buscamos $\rho = \frac{Cov(Y,\mathbb{E}(Y|X))}{\sigma_Y \sigma_{\mathbb{E}(Y|X)}}$

$$Cov(Y, \mathbb{E}(Y|X) = Cov(Y, 1 + \frac{3}{2}X) = \frac{3}{2}Cov(Y, X) = \frac{3}{2} \cdot 6 = 9$$

$$Var(\mathbb{E}(Y|X)) = Var(1 + \frac{3}{2}X) = (\frac{3}{2})^2 \cdot Var(X) = \frac{9}{4} \cdot 4 = 9$$

$$\rho = \frac{Cov(Y, \mathbb{E}(Y|X))}{\sigma_Y \sigma_{\mathbb{E}(Y|X)}} = \frac{9}{4 \cdot 3} = \frac{3}{4}$$

$$\rho^2 = \frac{9}{16}$$

3. La variable aleatoria X tiene una distribución uniforme continua con $0 < X < 10$.

(i) Cuál es la función de densidad de X? [5 puntos]
Por enunciado se tiene que X∼U(0,10), por lo tanto, $f_X(x) = \frac{1}{10}$

(ii) ¿Cuál es la función de distribución acumulada de X? [5 puntos]
Como se conoce le distribución de X, la cdf es igual a:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{10} & \text{si } 0 \leq x < 10 \\ 1 & \text{si } x \geq 10 \end{cases}$$

(iii) Encuentre la media y varianza de X [5 puntos]
La esperanza de X es igual a $\mathbb{E}(X) = \frac{0+10}{2} = 5$. La varianza de X es igual a $Var(X) = \frac{(10-0)^2}{12} = \frac{100}{12}$

(iv) ¿Cuál es la densidad condicional de $X|X > 5$? [5 puntos]
Notar que lo que se pide es $\mathbb{P}(X = x|X > 5)$, luego por probabilidad condicional esto es

$$\mathbb{P}(X = x|X > 5) = \frac{\mathbb{P}(X = x, X > 5)}{\mathbb{P}(X > 5)}$$

Donde x debe ser mayor que 5 y menor a 10, pues en cualquier otro caso, la probabilidad del numerador se va a 0, bajo estas condiciones:

$$f_{X|X>5}(x) = \mathbb{P}(X = x|X > 5) = \frac{\mathbb{P}(X = x, X > 5)}{\mathbb{P}(X > 5)} = \frac{\frac{1}{10}}{\frac{1}{2}} = \frac{1}{5}$$

(v) ¿Cuál es la esperanza condicional y la varianza condicional de esta variable? [5 puntos]
Dado el resultado anterior, se puede notar que $X|X > 5$ es otra uniforme $Y \sim U(5, 10)$, por lo tanto

$$\mathbb{E}(Y) = \frac{5 + 10}{2} = \frac{15}{2}$$

$$Var(Y) = \frac{(10 - 5)^2}{12} = \frac{25}{12}$$

4. Considere la distribución conjunta de dos variables aleatorias:

y: número de accidentes de tránsito al día en la RM

x: densidad promedio de automóviles en la RM.

Note que x es continua, mientras que y discreta. Suponga que la densidad condicional de y es $f(y|x) = \frac{e^{-\beta x}(\beta x)^y}{y!}$, con $x \geq 0$, y $\beta > 0$, mientras que la distribución marginal de x es $f(x) = \theta e^{-\theta x}$, con $\theta > 0$. Por lo tanto, condicionada en x, y distribuye Poisson de parámetro βx, mientras que x, no condicionada, tiene distribución exponencial.

(i) ¿Cuál es la distribución conjunta, $f(x, y)$ de estas dos variables aleatorias, x e y? [5 puntos]
Usando la fórmula $f(x, y) = f(y|x)f(x)$

$$f(x,y) = \frac{e^{-\beta x}(\beta x)^y}{y!} \cdot \theta e^{-\theta x} = \frac{\theta e^{-(\theta+\beta)x}(\beta x)^y}{y!}$$

(ii) Muestre que el inverso de la densidad marginal de y es $\frac{1}{\delta(1-\delta)^y}$, donde el inverso de δ es $\frac{\beta+\theta}{\theta}$ [5 puntos]
Para calcular la densidad marginal de y, debemos calcular la integral de la distribución conjunta respecto a x, así:

$$f(y) = \int_0^\infty f(x,y)dx = \frac{\theta\beta^y}{y!}\int_0^\infty e^{-(\beta+\theta)x}x^y$$

Luego, para resolver la integral usaremos la función gamma (Γ), la cual es:

$$\Gamma(z) = \int_0^\infty e^{-az}z^t dz = \frac{z!}{a^{z+1}}$$

Así, obtenemos:

$$f(y) = \frac{\theta\beta^y}{y!}\frac{y!}{\beta+\theta^{y+1}} = \frac{\theta\beta^y}{(\beta+\theta)(\beta+\theta^y)} = \left(\frac{\theta}{\beta+\theta}\right)\left(\frac{\beta}{\beta+\theta}\right)^y$$

Recordando que:

$$\delta = \frac{\theta}{\beta+\theta}$$

$$f(y) = \delta(1-\delta)^y$$

(iii) Muestre que $\frac{1}{E[x]} = \theta$ y que $(\frac{1}{Var[x]})^{1/2} = \theta$ [5 puntos]
Usando el resultado de la integral de la función gamma se tiene lo siguiente:

$$\mathbb{E}(X) = \int_0^\infty xf(x)dx = \int_0^\infty x\theta e^{-\theta x}dx = \theta\int_0^\infty xe^{-\theta x}dx = \theta\frac{1!}{\theta^{1+1}} = \theta\frac{1}{\theta^2} = \frac{1}{\theta}$$

Ahora para mostrar lo segundo, se debe recordar que $Var(X) = E(X^2) - E(X)^2$, así tenemos que:

$$\mathbb{E}(X^2) = \int_0^\infty x^2 f(x)dx = \theta\frac{2!}{\theta^{2+1}} = \theta\frac{2}{\theta^3} = \frac{2}{\theta^2}$$

Ya teniendo los valores de $E(X)$ y $E(X^2)$, se calcula la varianza:

$$Var(X) = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}$$

Luego, para mostrar lo pedido:

$$\left(\frac{1}{Var(X)}\right)^{1/2} = (\theta^2)^{1/2} = \theta$$

(iv) Si $E[y|x] = \beta x$, obtenga $E[y], Var[y]$, y $Cov[x, y]$. Exprese en términos de β y θ [5 puntos]

$$E(Y) = E_x(E(Y|X)) = E_x(\beta X) = \beta E(X) = \beta\frac{1}{\theta} = \frac{\beta}{\theta}$$

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = E(\beta X) + Var(\beta X) = \beta E(X) + \beta^2 Var(X) =$$

$$\beta\frac{1}{\theta} + \beta^2\frac{1}{\theta^2} = \frac{\beta}{\theta} + \left(\frac{\beta}{\theta}\right)^2$$

$$Cov(X, Y) = Cov(X, E(Y|X)) = Cov(X, \beta X) = \beta Var(X) = \beta\frac{1}{\theta^2} = \frac{\beta}{\theta^2}$$

5. Defina qué es un estimador insesgado. Defina qué es un estimador consistente. Explique cuidadosamente cuál es la diferencia entre ambos. ¿Todos los estimadores insesgados son consistentes? ¿Todos los estimadores sesgados son inconsistentes? Ilustre con un ejemplo en cada caso [20 puntos]

R: Sea $\hat{\theta}$ un estimador del parámetro poblacional θ. $\hat{\theta}$ es un estimador insesgado de θ si $E(\hat{\theta}) = \theta$. A su vez, $\hat{\theta}$ es un estimador consistente de θ si el límite de $\hat{\theta}$ conforme n tiende a infinito es igual a θ. El insesgamiento es una propiedad de los estimadores en muestras pequeñas, y por tanto no depende del tamaño muestral. En cambio, la consistencia es una propiedad asintótica que aplica a estimadores en muestras grandes, es decir, depende del tamaño muestral y tiene relación con la convergencia en probabilidad del estimador conforme el tamaño muestral crece.

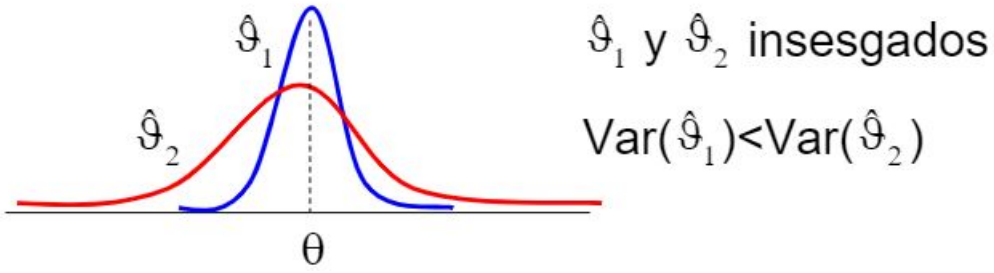
Por otra parte, un estimador insesgado puede ser inconsistente y un estimador sesgado puede ser consistente. Por ejemplo:

- El estimador $\hat{X}+1/n$ es un estimador sesgado del promedio poblacional ya que $E(\hat{X}+1/n) = \mu + 1/n \neq \mu$. No obstante, el límite de $\hat{X} + 1/n$ conforme n tiende a infinito es igual a μ, i.e., $\hat{X} + 1/n$ es un estimador consistente del promedio poblacional.
- El estimador X_1 es un estimador insesgado del promedio poblacional ya que $E(X_1) = E(X_i) = \mu$. Sin embargo, X_1 no depende de n, i.e., no converge en probabilidad a algo distinto de X_1 conforme n tiende al infinito, y por tanto es un estimador inconsistente del promedio poblacional.

6. Dado un estimador del parámetro poblacional θ, se define el error cuadrático medio como $MSE(\theta) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (Sesgo(\hat{\theta}))^2$. Bajo este criterio, un estimador eficiente siempre es mejor que un estimador insesgado. Comente e ilustre gráficamente. [10 puntos]

R: Sea $\hat{\theta}$ un estimador del parámetro poblacional θ. Luego, $\hat{\theta}$ es eficiente si, dentro de todos los estimadores insesgados de θ, $\hat{\theta}$ es aquel con la menor varianza. En efecto, lo primero es reconocer que para que $\hat{\theta}$ sea eficiente es condición necesaria que sea insesgado.

Luego, notar que si $\hat{\theta}$ es insesgado, entonces $MSE(\hat{\theta}) = Var(\hat{\theta})$. Pero como sabemos que $\hat{\theta}$ es eficiente, entonces sabemos que $Var(\hat{\theta})$ es mínima. $MSE(\hat{\theta})$ es un criterio mixto que permite comparar estimadores en base a la importancia relativa de la varianza y el sesgo de los estimadores. En consecuencia, si $\hat{\theta}$ es eficiente, entonces tiene mínimo error cuadrático medio, y por tanto será mejor que cualquier otro estimador insesgado (pues por definición ningún otro estimador insesgado puede tener menor varianza que $Var(\hat{\theta})$).



En la imagen anterior se puede apreciar dos estimadores insesgados, pero uno con menor varianza que el otro.

7. Sean A_i con $i = 1, ..., N$ un conjunto de variables aleatorias independientes e idénticamente distribuidas de media μ y varianza σ^2. Se define el siguiente estimador de la varianza:

$D_*^2 = \frac{1}{N}\sum_{i=1}^{N}(A_i - \bar{A})^2$ donde $\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A_i$ corresponde a la media muestral.

(i) Demuestre que D_*^2 es un estimador sesgado de la varianza poblacional [10 puntos]
A partir de la fórmula de del estimador de la varianza se desarrolla la demostración:

$$D_*^2 = \frac{1}{N}\sum_{i=1}^{N}(A_i - \bar{A})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(A_i^2 - 2A_i\bar{A} + \bar{A}^2)$$

$$= \frac{1}{N}[\sum_{i=1}^{N}A_i^2 - 2\sum_{i=1}^{N}A_i\bar{A} + \sum_{i=1}^{N}\bar{A}^2)]$$

$$= \frac{1}{N}\sum_{i=1}^{N}A_i^2 - 2\bar{A}\frac{1}{N}\sum_{i=1}^{N}A_i + \frac{1}{N} * N\bar{A}^2$$

Por definición del estadístico promedio: $\frac{1}{N}\sum_{i=1}^{N} A_i = \bar{A}$

$$D_*^2 = \frac{1}{N}\sum_{i=1}^{N}(A_i^2 - \bar{A}^2)$$

Dado que se quiere probar que es sesgado, se calcula la esperanza:

$$\mathbb{E}(D_*^2) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}A_i^2 - \bar{A}^2\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(A_i^2) - \mathbb{E}(\bar{A}^2)$$

Recordando que la varianza para una variable aleatoria se puede calcular como $Var(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$ para este caso se tiene:

$$\mathbb{E}(A_i^2) = Var(A_i) + [\mathbb{E}(A_i)]^2 = \sigma^2 + \mu^2 \Rightarrow \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(A_i^2) = \frac{1}{N}\sum_{i=1}^{N}\sigma^2 + \mu^2 = \sigma^2 + \mu^2$$

$$E(\bar{A}^2) = Var(\bar{A}) + [E(\bar{A})]^2 = \frac{\sigma^2}{N} + \mu^2 \Rightarrow E(\bar{A}^2) = \frac{\sigma^2}{N} + \mu^2$$

Finalmente:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(A_i^2) - \mathbb{E}(\bar{A}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2$$

$$= \sigma^2 - \frac{\sigma^2}{N}$$

Se concluye que el estimador es sesgado, ya que en esperanza su valor es distinto al parámetro a estimar, siendo su sesgo $-\frac{\sigma^2}{N}$.

(ii) Proponga un estimador insesgado de la varianza poblacional. [10 puntos]

A partir el resultado anterior, notar que su valor es $D_*^2 = \frac{N-1}{N}\sigma^2$ por lo que al multiplicarlo por $\frac{N}{N-1}$ se obtiene el valor de σ^2 y el estimador sería insesgado respecto a la varianza poblacional.

8. Explique la principal diferencia entre la Ley de los Grandes Números y el Teorema del Límite Central y de un ejemplo en cada caso. [20 puntos]

La Ley de los Grandes Números habla de la consistencia del promedio como estimador de la media poblacional, y establece que conforme n aumenta al infinito el promedio converge a la media poblacional. El Teorema del Límite Central nos dice que a cualquier variable aleatoria X, independiente de la distribución poblacional desde la cual provenga, al aplicarle la transformación estabilizadora $\sqrt{n}(\bar{X} - \mu)$ a su promedio, $\bar{X}$, va a converger en distribución a una Normal con media 0 y varianza σ^2. Luego, conforme n tiende a infinito, $\bar{X}$ converge asintóticamente a una Normal con media μ y varianza σ^2/n. La importancia del TLC es que podemos hacer inferencia estadística sobre el comportamiento de una variable aleatoria sin necesidad de conocer la distribución poblacional desde la cual proviene.

II. Intervalos de Confianza y Test de Hipótesis

9. Suponga que A y B son dos variables aleatorias estadísticamente independientes. Usted posee 50 muestras de A y 50 de B, cuyos valores esperados son μ_A y μ_B, respectivamente, y la varianza es σ^2 para ambas variables.

(a) Construya un IC al $(1-\alpha)\%$ para la diferencia de medias de A y B [10 puntos]

Para construir un IC tenemos que encontrar un estadístico A que contenga al parámetro en cuestión, en este caso, la diferencia de medias, tal que:

$$\mathbb{P}(a_1 < A < a_2) = 1 - \alpha$$

Sean $A_1, ..., A_n$ e $B_1, ..., B_n$ dos muestras aleatorias de distribuciones normales con medias μ_A y μ_B y varianzas σ^2. Como el tamaño de las muestras es bastante grande, tenemos que los estimadores de las medias (o los promedios) distribuyen Normal. Por lo tanto, el estimador de la diferencia de medias es:

$$\bar{A} - \bar{B}$$

El cual distribuye Normal, con media:

$$\mathbb{E}(\bar{A} - \bar{B}) = \mathbb{E}(\bar{A}) - \mathbb{E}(\bar{B})$$

$$= \mu_A - \mu_B$$

Y varianza:

$$Var(\bar{A} - \bar{B}) = Var(\bar{A}) + Var(\bar{B})$$

$$= \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}$$

$$= \sigma^2 \left(\frac{1}{50} + \frac{1}{50} \right)$$

$$= \sigma^2 \left(\frac{2}{50} \right)$$

$$= \sigma^2 \cdot 0.04$$

Es decir, $\bar{A} - \bar{B}$ distribuye $\mathcal{N}\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$. Por lo tanto, podemos definir el estadístico:

$$Z = \frac{(\bar{A} - \bar{B}) - (\mu_A - \mu_B)}{\sqrt{0.04\sigma^2}}$$

Que al ser una normal estandarizada, distribuye $\mathcal{N}(0, 1)$:

$$\mathbb{P}(z_1 < Z < z_2) = 1 - \alpha$$

Sabemos que la normal es simétrica, por lo que:

$$\mathbb{P}(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Definiendo $z_{1-\frac{\alpha}{2}}$ como:

$$\mathbb{P}(Z < z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

Despejando $\mu_A - \mu_B$ en la inecuación obtenemos:

$$\mathbb{P}(\bar{A} - \bar{B} - z_{1-\frac{\alpha}{2}}\sigma\sqrt{0.04} < \mu_A - \mu_B < \bar{A} - \bar{B} + z_{1-\frac{\alpha}{2}}\sigma\sqrt{0.04}) = 1 - \alpha$$

Por lo que el IC al 95% es:

$$(\bar{A} - \bar{B}) \pm z_{1-\frac{\alpha}{2}}\sigma\sqrt{0.04}$$

(b) ¿Cómo cambia el problema si se desconoce el valor de σ^2? Explique [10 puntos]

Si se desconoce el valor de la varianza poblacional σ^2, se puede ocupar algún estimador, como el estimador de la varianza muestral S^2:

$$S^2 = \frac{\sum_{i=1}^{n}(A_i - \bar{A})^2}{n - 1}$$

10. Se observa la muestra 4.2, 6.8, 3.6, 1.8, 5.8, 4.6, 3.4, 6.6, 6.0, 8.0 de tamaño 10 de una distribución $N(5, \sigma^2)$, i.e., tiene media conocida y equivale a 5. Construya un intervalo al 90% de confianza para σ^2. [10 puntos]

Notemos que $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 = \left(\frac{(n-1)S^2}{\sigma^2}\right) \sim \chi^2_{n-1}$

Luego, como $\left(\frac{(n-1)S^2}{\sigma^2}\right) \sim \chi^2_{n-1}$ entonces podemos calcular dos constantes a y b tales que:

$$\mathbb{P}\left(\left(\frac{(n-1)S^2}{\sigma^2}\right) < b\right) = 0.95$$

$$\mathbb{P}\left(a < \left(\frac{(n-1)S^2}{\sigma^2}\right) < b\right) = 0.90$$

Reescribiendo lo anterior se obtiene:

$$0.90 = \mathbb{P}\left(\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}\right)$$

Calculando la varianza muestral se obtiene que:

$$S^2 = \sum_{i=1}^{n}(X_i - \mu)^2/(n-1) = 3.6489$$

Como $\alpha = 0.1$, n=10, $b = \chi^2_{\alpha/2,(n-1)} = \chi^2_{0.05,9} = 16.919$ y $a = \chi^2_{1-\alpha/2,(n-1)} = \chi^2_{0.95,9} = 3.325$

Por lo tanto el intervalo de confianza al 90% es:

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right] = [1.9410; 9.8767]$$

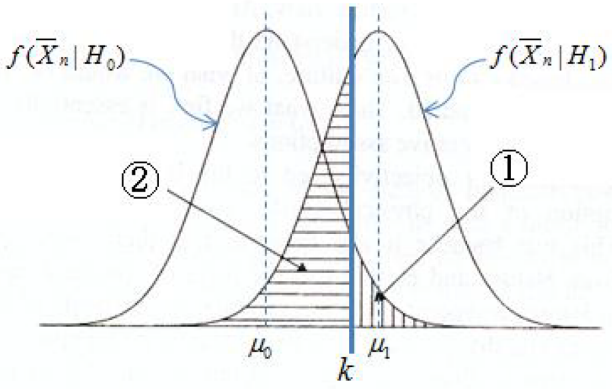
11. Una persona pseudo-fanática del fútbol es invitada al estadio varias veces al mes, pero no siempre quiere ir por miedo a sufrir actos delictivos, siendo la probabilidad que vaya de un 40%. Durante el año será invitada a 100 partidos. Defina una v.a. adecuada para calcular la probabilidad de que esta persona asista a una cierta cantidad de partidos, y calcule el valor esperado y varianza de la distribución de dicha v.a. [10 puntos]

El experimento de "asistir o no a un partido" se puede modelar de forma simple como una v.a $X_i \sim Bernoulli$ que toma valor 1 si decide asistir al Estadio y 0 si no. Tomando en cuenta que las asistencias al Estadio son eventos independientes, se define la siguiente v.a. para calcular la cantidad de asistencias:

$$Y = \sum_{i=1}^{n} X_i$$

Esta sumará la cantidad de veces que la persona decida asistir al partido por lo que puede ser modelada como una variable aleatoria binomial, es decir, $Y_i \sim Binomial$. Se sabe que la esperanza de una Binomial es np y su varianza $np(1-p)$, por lo que reemplazando los parámetros $n = 100$ y $p = 0.40$, se obtiene un valor esperado y varianza de 40 y 24, respectivamente.

12. Suponga dos hipótesis respecto del valor poblacional μ, la hipótesis nula ($H_0 : \mu = \mu_0$) y la hipótesis alternativa ($H_1 : \mu = \mu_1$). Para testear cuál de estas dos hipótesis es la correcta, un aumento del tamaño muestral reduce el error tipo I y aumenta el error tipo II, mejorando así el poder estadístico del test. ¿Verdadero o Falso? Defina y explique detalladamente. [15 puntos]



R: Considerando que el error tipo I es la probabilidad de rechazar H_0 en circunstancias que H_0 es verdadera. También es llamado falso positivo y su probabilidad se denota por α. Y el error tipo II es la probabilidad de no rechazar H_0 en circunstancias que H_1 es verdadera. También se denomina falso negativo y su probabilidad se denota por β.

Además, el tamaño muestral incide en el poder estadístico de la siguiente fomra, al tener una mayor cantidad de datos de la población disminuye la varianza de la distribución de ambas hipótesis (H_0 y H_1), lo cual reduce tanto el error tipo I como el error tipo II, aumentando así el poder estadístico del test.

En base a las dos definiciones anteriores podemos decir que la afirmación es falsa. Un aumento en el tamaño muestral no necesariamente reduce el error tipo I y aumenta el error tipo II al mismo tiempo. Por el contrario, se espera que al aumentar el tamaño muestral ambos errores disminuyan.

13. El Ministerio de Desarrollo Social está a cargo de implementar la Encuesta CASEN, la cual permite medir la incidencia de la pobreza extrema en la población nacional y su evolución en el tiempo. La pobreza extrema es una variable binaria que toma el valor 1 si el hogar está en pobreza extrema (debajo de un valor determinado de ingreso per cápita) y 0 si no. En el gráfico a continuación se muestra la evolución de la pobreza extrema para la población urbana y la población rural entre los años 2006 y 2022. Se denota una caída en la pobreza extrema a lo largo del período, tanto en zonas rurales como en zonas urbanas. Por ejemplo, tal como muestra el cuadro debajo del gráfico, la pobreza extrema en zonas rurales cae de un 18% en 2006 a un 2.6% en 2022. Por otra parte, la pobreza extrema en zonas urbanas cae de un 7.9% en 2006 a un 1.7% en 2022. Lo anterior sugiere que la brecha de pobreza extrema entre zonas rurales y zonas urbanas también ha disminuido. Mientras en 2006 la brecha era de 10.1 puntos porcentuales en favor de zonas rurales, dicha brecha en 2022 era sólo de 0.9 puntos porcentuales.

(a) Asumiendo una muestra de 1500 observaciones por zona por año, construya un intervalo de confianza al 90% ($\alpha = 0.1$) de la tasa de pobreza extrema para zonas rurales en el año 2006 y otro para zonas urbanas en el año 2006. ¿Puede concluir que las tasas de pobreza extrema entre zonas rurales y urbanas son estadísticamente distintas en ese año? [5 puntos]

Construyamos un intervalo de confianza considerando que la fórmula general para el intervalo de confianza de una proporción (en este caso, la tasa de pobreza extrema) es:

$$IC = \bar{X} \pm Z\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

Donde $\bar{X}$ es la tasa de pobreza extrema, Z es el valor crítico de la distribución normal estándar a nivel de confianza deseado, en este caso 90% (1.645), y n es el tamaño de la muestra (1500)

En base a lo anterior realizamos intervalos de confianza para ambas zonas:

Intervalo de confianza Rural

$$0.18 \pm 1.645\sqrt{\frac{0.18(1-0.18)}{1500}} = 0.18 \pm 0.0229 = IC_{RURAL}[0.1571, 0.2029]$$

Intervalo de confianza Urbano

$$0.079 \pm 1.645\sqrt{\frac{0.079(1-0.079)}{1500}} = 0.079 \pm 0.0115 = IC_{URBANO}[0.0905, 0.0675]$$

Ahora, para determinar si las tasas de pobreza extrema entre zonas rurales y urbanas son estadísticamente distintas en el año 2006, podemos comparar los intervalos de confianza. Dado que los intervalos de confianza no se superponen, podemos concluir que existe evidencia para sugerir que las tasas de pobreza extrema en las zonas rurales y urbanas son estadísticamente distintas en ese año. La tasa de pobreza en las zonas rurales (entre 15.71% y 20.29%) es significativamente mayor que en las zonas urbanas (entre 1.58% y 3.62%).

(b) Asumiendo una muestra de 1500 observaciones por zona por año, testee a un 90% de nivel de confianza ($\alpha = 0.1$) si la diferencia en las proporciones de pobreza extrema entre zonas rurales y urbana es mayor a cero. Haga esto para 2006, y luego separadamente para 2022. Explicitar hipótesis nula, hipótesis alternativa, y test a utilizar. ¿Qué concluye? [15 puntos]

Paso 1: Definir Hipótesis nula y alternativa. Este es un test de proporciones. Sea p_R la tasa de pobreza extrema en zona rural y p_U la tasa de pobreza extrema en zona Urbana, el año 2006. Luego,

$$H_0 : p_R = p_U \iff p_R - p_U = 0$$
$$H_1 : p_R > p_U \iff p_R - p_U > 0$$

Hipótesis Nula (H_0): La diferencia en las proporciones de pobreza extrema entre zonas rurales y urbanas en 2006 es igual a cero.
Hipótesis Alternativa (H_1): La diferencia en las proporciones de pobreza extrema entre zonas rurales y urbanas en 2006 es mayor que cero.

Paso 2: Definir un nivel de significancia. Como se especifica en el enunciado el nivel de significancia es 90%.

Paso 3: Definir estadístico de prueba y calcularlo. Como los datos entregados son de la Encuesta CASEN podemos suponer que la cantidad de datos es lo suficientemente grande

para que la muestra distribuya de forma normal. Como se definió un test de proporciones, el estadístico corresponderá a este test:

$$Z = \frac{p_R - p_U}{\sqrt{p(1-p)(\frac{1}{n_R} + \frac{1}{n_U})}}$$

Siendo p

$$p = \frac{n_R p_R + n_U p_U}{n_R + n_U}$$

Paso 4: Identificar el valor crítico. Utilizamos la tabla de distribución normal estándar (tabla Z) y buscamos el valor del nivel de confianza para construir el $Z_{critico}$. Como es un test de una cola, la región de rechazo viene definida por la hipótesis alternativa ($p_R > p_U$). En efecto, rechazamos la hipótesis nula (en favor de la alternativa) si $Z > 1.285$.

Calculando para el año 2006 nos queda lo siguiente:

$$p = \frac{1500 * 0.18 + 1500 * 0.079}{3000} = 0.130$$

$$Z = \frac{0.18 - 0.079}{\sqrt{0.130(1-0.130)(\frac{1}{1500} + \frac{1}{1500})}} = 8.225$$

Ahora, calculamos para el año 2022:

$$p = \frac{1500 * 0.026 + 1500 * 0.017}{3000} = 0.0215$$

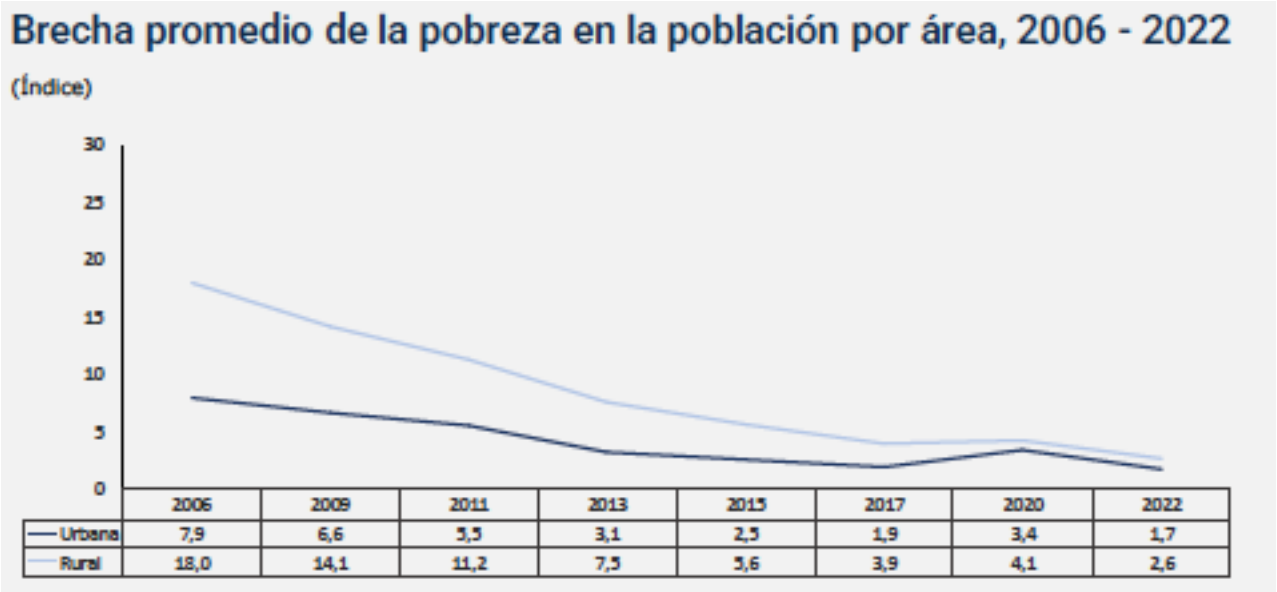
$$Z = \frac{0.026 - 0.017}{\sqrt{0.0215(1-0.0215)(\frac{1}{1500} + \frac{1}{1500})}} = 1.699$$

De acuerdo a lo visto se puede concluir que en ambos casos que la diferencia en las proporciones es mayor a cero por lo que se rechaza la hipótesis nula.

(c) Explique en qué consiste el Teorema del Límite Central. ¿Cuál es la relevancia del Teorema del Límite Central para llevar a cabo este test de hipótesis? ¿Por qué es tan importante? Explique. [15 puntos]

El Teorema del Límite Central nos dice que a cualquier variable aleatoria X, independiente de la distribución poblacional desde la cual provenga, al aplicarle la transformación estabilizadora $\sqrt{n}(\bar{X} - \mu)$ a su promedio, $\bar{X}$, va a converger en distribución a una Normal con media 0 y varianza σ^2. Luego, conforme n tiende a infinito, $\bar{X}$ converge asintóticamente a una Normal con media μ y varianza σ^2/n. La importancia del TLC es que podemos hacer inferencia estadística sobre el comportamiento de una variable aleatoria sin necesidad de conocer la distribución poblacional desde la cual proviene. Nótese que la proporción la podemos interpretar como una media, ya que su definición es casos favorables entre casos totales. Segundo, dado que suponemos que tenemos una muestra lo suficientemente grande, por TCL sabemos que

el estadístico del test de proporciones distribuye normal estándar, con media 0 y varianza 1. Tercero, lo anterior nos permite hacer inferencia estadística para el comportamiento de los elementos que componen el estadístico del test de proporciones, incluido p_m y p_h. En conclusión, sin el TCL no podríamos hacer inferencia estadística de los parámetros de interés de este problema.



	2006	2009	2011	2013	2015	2017	2020	2022
Urbana	7,9	6,6	5,5	3,1	2,5	1,9	3,4	1,7
Rural	18,0	14,1	11,2	7,5	5,6	3,9	4,1	2,6

III. Ejercicios Empíricos

14. Vaya al sitio web del Servicio Nacional de Migraciones:

https://serviciomigraciones.cl/estudios-migratorios/registros-administrativos/

Descargue la base de datos de residencias temporales otorgadas en 2022. Cada fila es una visa otorgada. La base incluye variables demográficas como sexo, nacionalidad, fecha de nacimiento de la persona, nivel educativo, actividad, y profesión, entre otras.

(a) Transforme el formato de la base de datos a R. Crear variable de edad ("edad") actualizada al 31 de Diciembre de 2022. Calcule el promedio, desviación estándar, mediana y moda de la edad. Además, calcule la covarianza entre la variable edad y el mes del año en que la persona ingresó al país. Comente los resultados obtenidos. [10 puntos]

Variable	Promedio	Desviación estándar	Mediana	Moda
Edad	31,9	14,73	31	28

Table 1: Estadísticas descriptivas de la variable edad

Se puede notar que la edad promedio es bastante similar a la mediana, aunque la moda es unos años menor. La desviación hace sentido en base al valor de la edad promedio.
Además al calcular la covarianza entre Edad y el mes en que la persona ingreso al país esta es igual a -2.3635. Se puede notar que están correlacionadas negativamente, es decir, a mayor edad, la gente ingresa al país en los primeros meses del mes.

(b) Crear una variable continua de maximo nivel educacional alcanzado ("educ"), donde "Ninguno" implica 0 anos de educación, "Pre-Basico" 1, "Basica" 6, "Media" 12, "Tecnica" 13, "Superior" 17. Obtenga una gráfico de dispersión con la variable edad y educ. Asegurese de eliminar aquellas observaciones que no contienen información ("en blanco" o "no informa"). ¿Qué concluye? [10 puntos]
En este gráfico no se ve una relación clara entre ambas variables, aunque llama la atención que ambas variables tienen una correlación "fuerte" (si se calcula es de 0.43).

(c) Obtenga un histograma de la frecuencia de la variable *edad*. ¿Qué tipo de distribución puede asociar al resultado obtenido? [10 puntos]
La distribución luce como una normal centrada en los 30 años aproximadamente.

(d) Realice un Test de Normalidad de la variable *edad*. ¿La distribución de la población (de donde proviene la muestra) es Normal? Comente. [15 puntos]
Al realizar múltiples test y contrario a lo que pensaríamos luego de ver el histograma de la variable edad, llegamos a que ésta no proviene de una distribución normal. Los valores delos p-valor de cada test así lo señalan (en todos los casos rechazamos la hipótesis nula de normalidad). Así es como, si bien tiene una apariencia de una normal no significa necesariamente que provenga de este tipo de distribución.

(e) Determine el error estándar del estimador promedio de la variable *edad* y construya un intervalo de confianza al 90% para este. [15 puntos]
De la parte i) tenemos que la edad promedio es de 31,9 años y la desviación estándar es de 14,73 años. El error estándar viene definido como:

$$\frac{S}{\sqrt{n}} = \frac{14,73}{\sqrt{256.038}} = 0,029$$

Ahora, para el IC del promedio del ingreso no laboral, dado que no conocemos la varianza de la muestra, se usa el estadístico:

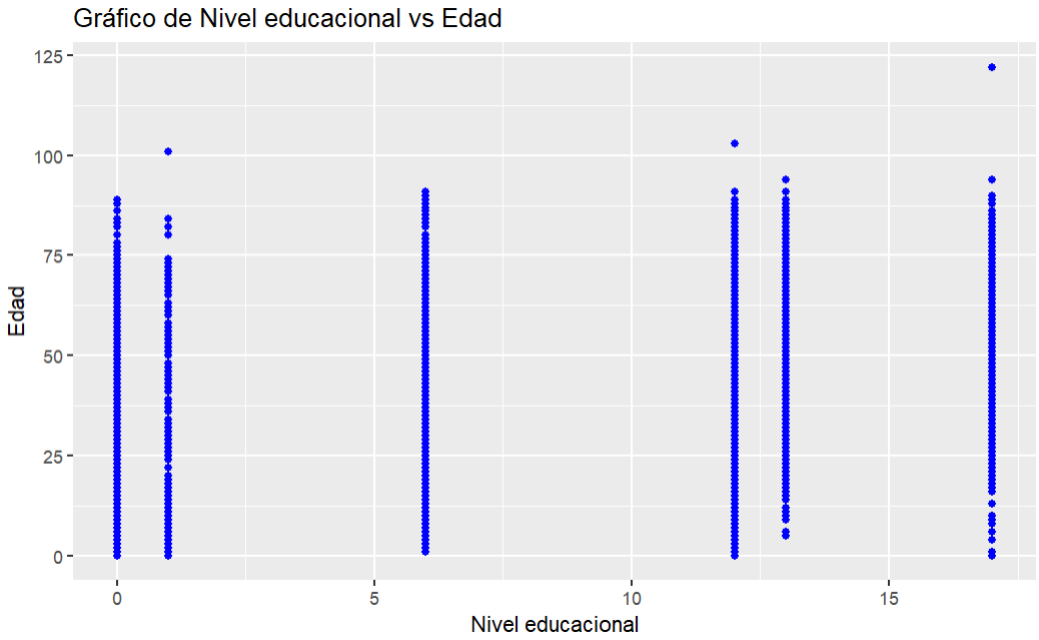


Figure 1: Nivel educacional vs Edad

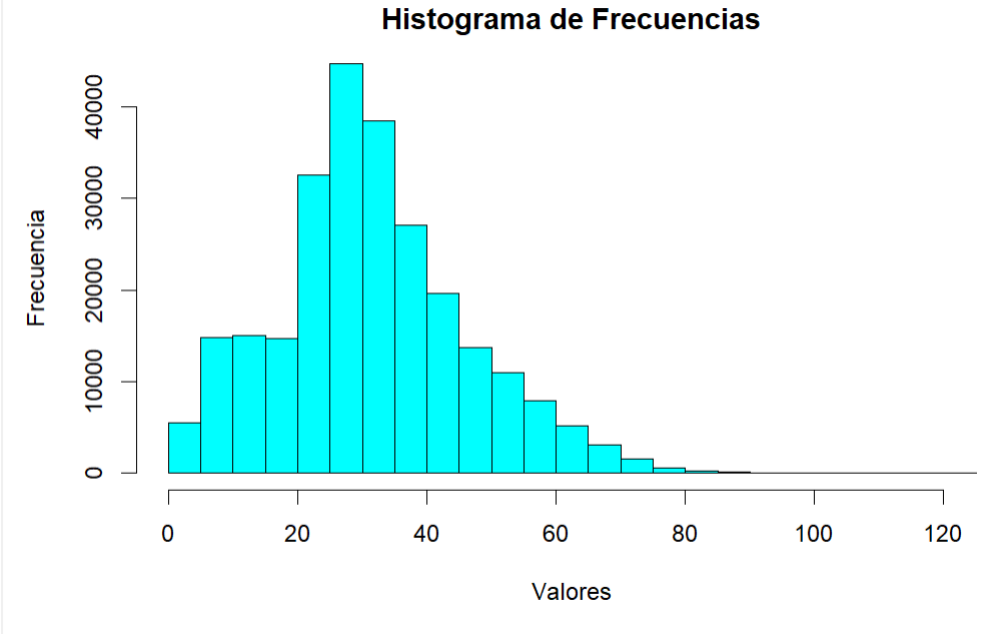


Figure 2: Histograma de edad

```
        Shapiro-Wilk normality test

data:  otorgadas2022_v2$edad[1:5000]
W = 0.984, p-value <2e-16


        Shapiro-Wilk normality test

data:  otorgadas2022_v2$edad[5001:10000]
W = 0.982, p-value <2e-16


        Shapiro-Wilk normality test

data:  otorgadas2022_v2$edad[10001:15000]
W = 0.985, p-value <2e-16
```

Figure 3: Test de normalidad de la distribución

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{31,9 - \mu}{0,029}$$

$$31,9 - t_{\alpha/2} \cdot 0,029 \leq \mu \leq 31,9 + t_{\alpha/2} \cdot 0,029$$

$$31,9 - 1,645 \cdot 0,029 \leq \mu \leq 31,9 + 1,645 \cdot 0,029$$

$$31,852 \leq \mu \leq 31,948$$

Por lo que $\mu \in [31,852; 31,948]$ con un nivel de confianza de un 90%, todo ello bajo el supuesto de normalidad de la población de la cual se obtiene la muestra.

(f) Para cada una de las 10 nacionalidades con mayor numero de visas temporales otorgadas, calcule la edad promedio. Luego testee la hipótesis nula (al 95% de confianza) de que el promedio de edad de aquella nacionalidad con mayor promedio de edad es igual al promedio de edad de aquella nacionalidad con menor promedio de edad. Interprete. [15 puntos]
Se puede notar que la nacionalidad con el menor promedio de edad es Haití y el de mayor promedio Cuba.
Primero se define un test de medias, se establece la hipotesis nula y la alternativa

$$H_0 : \mu_{Haiti} = \mu_{Cuba} \; H_1 : \mu_{Haiti} \neq \mu_{Cuba}$$

El nivel de significancia lo da el enunciado al 95%, $\alpha = 0,05$
Se escoge el estadístico adecuado

$$\frac{(\bar{x_1} - \bar{x_0}) - (\mu_{Haiti} - \mu_{Cuba})}{\sqrt{\frac{\sigma^2_{Haiti}}{n_{Haiti}} + \frac{\sigma^2_{Cuba}}{n_{Cuba}}}}$$

País	Conteo	Edad promedio	Desviación estándar
Venezuela	117.718	32,854	14,965
Colombia	41.018	31,599	14,486
Perú	35,112	32,141	15,032
Bolivia	24.280	27,913	13,099
Ecuador	8.096	29,601	14,473
Haití	6.854	26,722	12,917
Argentina	5.825	35,565	14,641
Brasil	2.139	33,662	13,577
República Dominicana	1.975	32,453	13,718
Cuba	1.654	36,528	14,402

Table 2: 10 países con más visas otorgadas

$$Z = \frac{36,528 - 26,722}{\sqrt{\frac{166,849}{6854} + \frac{207,418}{1654}}}$$

$$Z = \frac{9,806}{0,387}$$

Se construye el IC

$$-9,806 \pm 1,96 \cdot 0,387 = -9,806 \pm 0,759$$

$$\text{IC al } 95\%[-10,565, -9,047]$$

Como el 0 no pertenece al IC, existe evidencia suficiente para rechazar la hipótesis nula, es decir, las edades promedios de los extranjeros que vienen de Haití y Cuba son diferentes estadísticamente hablando al 95%.

15. Diríjase al sitio web del Instituto Nacional de Estadísticas (INE)

https://www.ine.gob.cl/estadisticas/sociales/mercado-laboral/ocupacion-y-desocupacion

Descargue la base de datos de la encuesta de ocupación y desocupación correspondiente al trimestre abril-mayo-junio de 2023 (ENE 2023 05 AMJ). Cada fila es una persona encuestada.

1. Considerando sólo a personas que durante la semana pasada trabajaron al menos una hora, calcule el promedio, desviación estándar, mínimo y máximo de las horas habituales semanales de trabajo. [10 puntos]

Promedio	Dv. estándar	Mínimo	Máximo
42.87	46.66	1	999

Table 3: Estadística descriptiva de la variable C2_1_3: Horas semanales trabajadas habitualmente

En la tabla se puede observar que el máximo es 999 dado que, al ver la descrición de dicha variable, contiene valores como 999 y 888 para decir que no responde y no sabe, respectivamente. Por lo que no entregan infomación pero no referente exactamente a que esa es la cantidad de hora de trabajo semanal.

Promedio	Dv. estándar	Mínimo	Máximo
40.55	13.53	1	168

Table 4: Estadística descriptiva de la variable C2_1_3: Horas semanales trabajadas habitualmente, sin los valors 888 y 999

En la tabla 4 podemos ver la estadística descriptiva de la misma variable pero sin los valores 999 y 888. Se observa que la desviación estándar disminuye en un 33%, lo que tiene sentido dado que el máximo y el mínimo ya no están tan lejanos como antes.

2. Considerando sólo a personas que durante la semana pasada trabajaron al menos una hora, grafique el promedio de horas habituales semanales de trabajo, por edad, donde en el eje Y están las horas promedio trabajadas, y en el eje X la edad. Comente. [10 puntos]

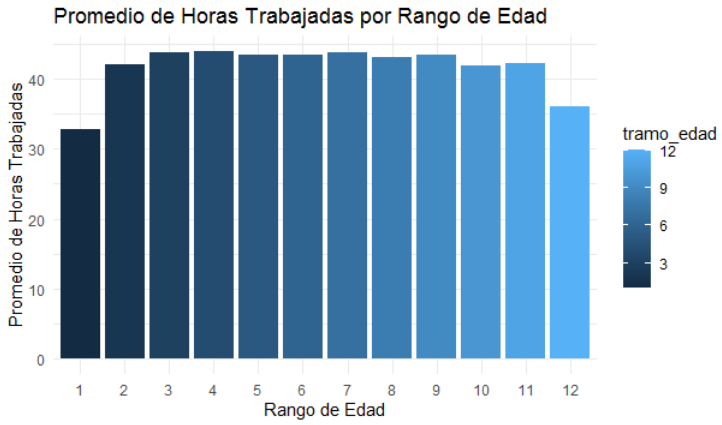


Figure 4: Promedio de horas semanales por grupo etario

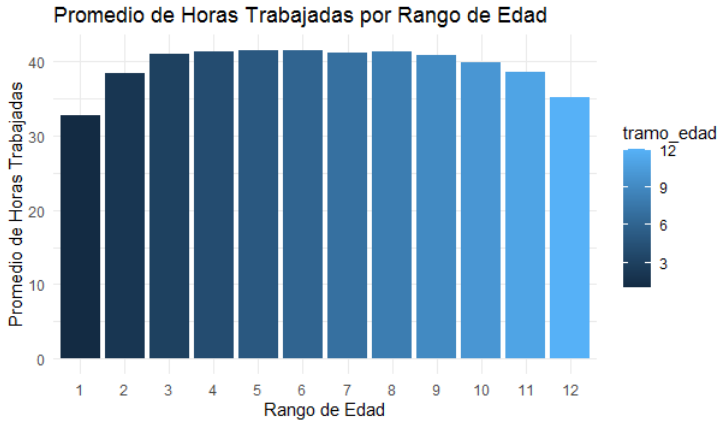


Figure 5: Promedio de horas semanales por grupo etario sin considerar los valores 888 y 999

En este caso también se tienen dos gráficos uno con toda la información y otro que excluyó los valores 888 y 999. Se puede observar que la diferencia no es visualmente notoria.

De ambas figuras se puede observar que hay 12 tramos etarios donde se aprecia que existe una distribución normal, esta se puede ver más claramente en la Figura 5 que en la 4 ya que debemos recordad que la Figura 4 se encuentra sesgadas por los valores 888 y 999 que no hablan exactamente de horas trabajadas sino de no responder.La amplitud máxima se encuentra en los tramos 5 y 6 son los que hacen referencia a rangos de 35-39 y 40-44, respectivamente. La colas de las distribución se encuentran en los primero tres tramos y en los últimos tres. Esto tiene sentido ya que al iniciar en la etapa laboral, entre los 15 y 29 años representados por los tramos 1, 2 y 3, el tiempo que se dedica al trabajo no es completo ya que existen variables como el estudio que limitan las jornadas laborales. Por otro lado, entre los 60 y 70 años representados por los tamos 10, 11 y 12, las jornadas laborales también disminuyen por variables como la jubilación, la salud, la vejez, etc.

3. Considerando sólo a personas que durante la semana pasada trabajaron al menos una hora, grafique un *qqplot* para las variables Edad y horas habituales semanales de trabajo ¿Qué puede concluir para cada variable? [10 puntos]

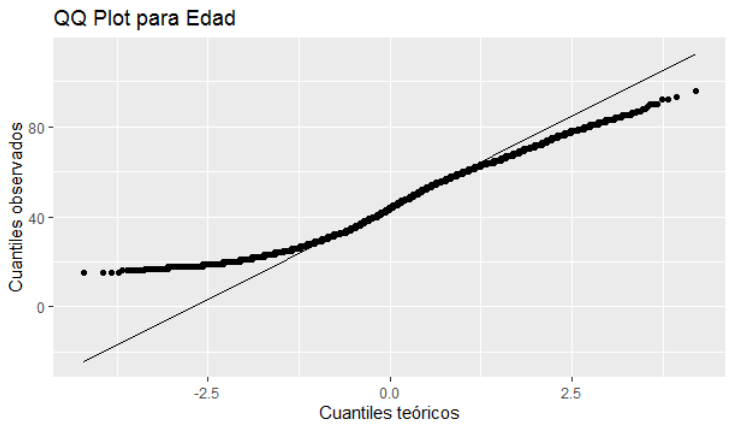


Figure 6: QQPlot para Edad

De la Figura 6 se puede observar que la variable Edad se acerca a la diagonal por lo que su distribución es parecida a la Normal a excepción por la colas que se alejan de la diagonal. Dado esto no podemos asegurar la normalidad de la variable.

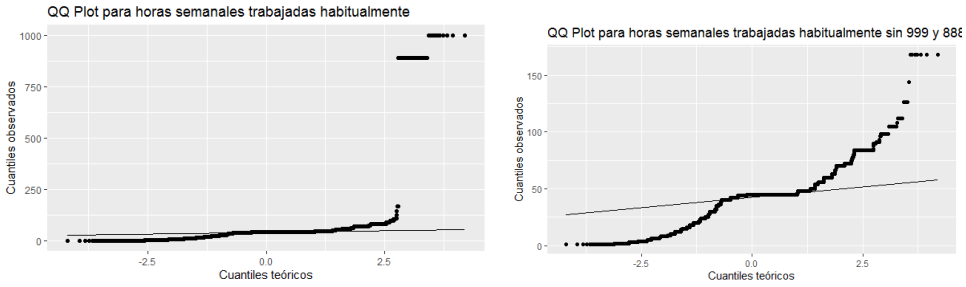


Figure 7: QQPlot para las horas semanales trabajadas habitualmente

En la Figura 7 se encuentran dos gráficos, el de la izquierda muestra las horas semanales trabajadas habitualmente son los datos 888 y 999, en esta se puede observar que esos datos específicamente son los que se alejan de la diagonal considerablemente como outlier por lo

que se decidió realizar un qqPlot sin aquellos datos y se obtuvo el resultado de la derecha. En el qqPlot de las horas semanales trabajadas habitualmente sin los datos 888 y 999 se puede observar que solo el centro se asemeja a la diagonal pero las colas se alejan impidiendo que se pueda considerar esta viable como una distribución normal.

4. Considerando sólo a personas que durante la semana pasada trabajaron al menos una hora, genere un intervalo de confianza para la varianza de la variable Edad. [10 puntos]

   El intervalo de confianza al 95% para la varianza será:

$$\sigma^2_{ingreso} \in [44.1 \ ; 44.4]$$

5. Presente un test de hipótesis para testear si la edad promedio de personas que trabajaron al menos una hora la semana pasada es igual al promedio de edad de personas que no trabajaron al menos una hora la semana pasada. Defina hipótesis nula e hipótesis alternativa, y luego ejecute el test usando el comando correspondiente. Testee empíricamente considerando un alpha=0.1. ¿Qué concluye? [10 puntos]

   Si asuminos normalidad, y α=0.1, para este caso tendremos un test de dos colas:

$$H_0 : \mu_{edad\_si\_trabajo} = \mu_{edad\_no\_trabajo}$$

$$H_a : \mu_{edad\_si\_trabajo} \neq \mu_{edad\_no\_trabajo}$$

   Por lo que así se tiene:

$$p - valor = 0,946 > 0,1$$

$$IC_{\mu_{edad\_si\_trabajo} - \mu_{edad\_no\_trabajo}} = [-\infty; 341280663]$$

   En efecto, utilizando cualquiera de estas evidencias (i.e., que el p-valor sea mayor a 0.1; o que el cero esté incluido en el intervalo), concluimos que NO se rechaza H_0 por lo que no habría diferencias entre las medias de las edades entre los que si trabajaron al menos una hora la semana anterior y los que no.

6. Considerando sólo a personas que durante la semana pasada trabajaron al menos una hora, realice un test de hipótesis para testear si el promedio de horas habituales semanales de trabajo es estadísticamente equivalente entre hombres y mujeres. Defina hipótesis nula e hipótesis alternativa, y luego ejecute el test usando el comando correspondiente. Testee empíricamente considerando un 95% de nivel de confianza estadística. ¿Qué puede concluir al respecto? [15 puntos]

   Si asuminos normalidad, y α=0.05, para este caso tendremos un test de una cola:

$$H_0 : \mu_{horas\_trab\_mujeres} = \mu_{horas\_trab\_hombres}$$

$$H_a : \mu_{horas\_trab\_mujeres} \neq \mu_{horas\_trab\_mujeres}$$

   Por lo que así se tiene:

$$p - valor = 2,2e^{-16} < 0,05$$

$$IC_{\mu_{horas\_trab\_mujeres}-\mu_{horas\_trab\_hombres}} = [-\infty; -3,939]$$

En efecto, utilizando cualquiera de estas evidencias (i.e., que el p-valor sea menor a 0.05; o que el cero no esté incluido en el intervalo), concluimos que SE RECHAZA H_0 por lo que NO es posible afirmar que la media de horas trabajadas por las mujeres es estadísticamente equivalente que la media de horas trabajadas por hombres.