

Maestría en Computo Estadístico
Inferencia Estadística
Tarea 7

23 de noviembre de 2020

Enrique Santibáñez Cortés

Repositorio de Git: Tarea 7, IE.

Escriba de manera concisa y clara sus resultados, justificando los pasos necesarios. Serán descontados puntos de los ejercicios mal escritos y que contenga ecuaciones sin una estructura gramatical adecuada. Las conclusiones deben escribirse en el contexto del problema. Todos los programas y simulaciones tienen que realizarse en R.

1. Las hojas de una planta se examinan buscando insectos. El número de insectos en una hoja sigue una distribución de Poisson con media μ , con la excepción de que muchas de las hojas no tienen insectos pues son inadecuadas para que se alimenten de ellas y esto no es simplemente el resultado de la variación aleatoria de la ley de Poisson.
 - a) Encuentre la probabilidad condicional de que una hoja contenga i insectos, dado que contiene al menos uno.

RESPUESTA

Sea X el número de insectos en una hoja, entonces sabemos que $X \sim \text{Poisson}(\mu)$, entonces ocupando el teorema de Bayes tenemos que la **la probabilidad condicional de que una hoja contenga i insectos, dado que contiene al menos uno es**

$$\mathbb{P}(X = i | X \geq 1) = \frac{\mathbb{P}(X = i, X \geq 1)}{\mathbb{P}(X \geq 1)} = \frac{\mathbb{P}(X = i, X \geq 1)}{1 - \mathbb{P}(X = 0)},$$

para $i = 0$ tenemos que los eventos de tener 0 insectos y tener al menos un insecto son eventos incluyentes, por lo que la probabilidad sería 0. Ahora para $i \geq 1$ observemos que los eventos de obtener i insectos y el evento de tener un insecto este evento está contenido en obtener i insectos por lo que la probabilidad sería calculada arriba se simplifica a

$$\begin{aligned} \mathbb{P}(X = i | X \geq 1) &= \frac{\mathbb{P}(X = i, X \geq 1)}{\mathbb{P}(X \geq 1)} = \frac{\mathbb{P}(X = i, X \geq 1)}{1 - \mathbb{P}(X = 0)} = \frac{\mathbb{P}(X = i)}{1 - \mathbb{P}(X = 0)} = \frac{\frac{\mu^i e^{-\mu}}{i!}}{1 - e^{-\mu}} \\ &= \frac{\mu^i e^{-\mu}}{i!(1 - e^{-\mu})}, \quad i \geq 1. \end{aligned}$$

- b) Supongamos que se observan x_i hojas conteniendo i insectos ($i = 1, 2, 3, \dots$), con $\sum x_i = n$. Muestre que el estimador de máxima verosimilitud de μ satisface la ecuación

$$\hat{\mu} = \bar{x}(1 - e^{-\hat{\mu}}),$$

donde $\bar{x} = \sum x_i / n$.

RESPUESTA

Utilizando el resultado del inciso anterior sabemos que la probabilidad de obtener al menos un insecto dado que ya se observó uno está dado por la función de probabilidad $f(x) = \frac{\mu^x e^{-\mu}}{x!(1 - e^{-\mu})}$, para $x = 1, 2, \dots$. Entonces la función de verosimilitud del proceso anterior es

$$L(\mu) = \left(\frac{\mu^{x_1} e^{-\mu}}{x_1!(1 - e^{-\mu})} \right) \cdot \left(\frac{\mu^{x_2} e^{-\mu}}{x_2!(1 - e^{-\mu})} \right) \cdots \left(\frac{\mu^{x_n} e^{-\mu}}{x_n!(1 - e^{-\mu})} \right) = \frac{\mu^{\sum_{i=1}^n x_i} e^{-n\mu}}{\prod_{i=1}^n x_i!(1 - e^{-\mu})^n}$$

y por lo tanto,

$$\log L(\mu) = \sum_{i=1}^n x_i \log(\mu) - n\mu - \log \left(\prod_{i=1}^n x_i! \right) - \log((1 - e^{-\mu})^n).$$

Derivando e igualando a cero para obtener puntos críticos, obtenemos que

$$\frac{d}{d\mu} \log L(\mu) = \sum_{i=1}^n x_i \frac{1}{\mu} - n - \frac{ne^{-\mu}}{1 - e^{-\mu}}.$$

Así que

$$\begin{aligned} \sum_{i=1}^n x_i \frac{1}{\mu} - n - \frac{ne^{-\mu}}{1 - e^{-\mu}} &= 0 \\ \sum_{i=1}^n x_i \frac{1}{\mu} &= \frac{n - ne^{-\mu} + ne^{-\mu}}{1 - e^{-\mu}} \\ \sum_{i=1}^n x_i \frac{1}{\mu} &= \frac{n}{1 - e^{-\mu}}, \end{aligned}$$

y lo anterior que un punto crítico es

$$\hat{\mu} = \bar{x}(1 - e^{-\hat{\mu}}).$$

Usando el criterio de segunda derivada, verifiquemos que el punto crítico es un máximo. Tenemos que

$$\frac{d^2}{d\mu^2} \log L(\mu) = - \sum_{i=1}^n x_i \frac{1}{\mu^2} - \frac{ne^{-\mu}}{1 - e^{-\mu}}.$$

entonces como la segunda derivada es negativa podemos concluir que el punto crítico es máximo, y **por lo tanto el estimador de máxima verosimilitud para μ es**

$$\hat{\mu} = \bar{x}(1 - e^{-\hat{\mu}}).$$

c) Determine $\hat{\mu}$ numéricamente para el caso $\bar{x} = 3.2$. Utilice R.

RESPUESTA

Ocupemos la librería *rootSolve* para encontrar el valor exacto del estimador de máxima verosimilitud.

```
library(rootSolve)
mle_mu_exp <- function(muh){ # definimos la ecuación
  muh-3.2*(1-exp(-muh))
}

sol_ecuacion <- uniroot(mle_mu_exp, c(.05, 3.5))$root # resolvemos la ecuación
```

Por lo tanto, **podemos concluir que el estimador de máxima verosimilitud para μ es**

$$\hat{\mu} = 3.0481759. \quad \blacksquare.$$

2. Los siguientes son los tiempos (en horas) entre fallas sucesivas del sistema de aire acondicionado en un avión:

97	51	11	4	141	18	142	68	77
80	1	16	106	206	82	54	31	216.
46	111	39	63	18	191	18	163	24

- a) Suponiendo que estas son observaciones independientes de una distribución exponencial con media θ , encuentre $\hat{\theta}_{MLE}$.

RESPUESTA

Sea T los tiempos (en horas) entre fallas sucesivas del sistema de aire acondicionado en un avión, entonces sabemos que $Y \sim \text{Exponencial}(\theta)$ y por lo tanto la función de densidad es

$$f(t) = \theta e^{-t\theta}.$$

Entonces la función de verosimilitud es

$$L(\theta) = \prod_{i=1}^n f(t_i) = \prod_{i=1}^n \theta e^{-t_i\theta} = \theta^n e^{-\theta \sum_{i=1}^n t_i}$$

y por lo tanto,

$$\log(L(\theta)) = n \log(\theta) - \theta \sum_{i=1}^n t_i.$$

Ahora derivamos con respecto a θ e igualamos a cero para obtener puntos críticos, obtenemos que

$$\frac{d}{d\theta} \log(L(\theta)) = \frac{n}{\theta} - \sum_{i=1}^n t_i,$$

así que

$$\frac{n}{\theta} - \sum_{i=1}^n t_i = 0 \Rightarrow \text{el punto crítico es } \hat{\theta} = \frac{n}{\sum_{i=1}^n t_i}.$$

Usando el criterio de segunda derivada, verifiquemos que efectivamente el punto crítico es un máximo, tenemos que

$$\frac{d^2}{d\theta^2} \log(L(\theta)) = -\frac{n}{\theta^2}$$

entonces como la segunda derivada es negativa podemos concluir que el punto crítico es máximo, y **por lo tanto el estimador de máxima verosimilitud para θ es:**

$$\hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^n t_i}.$$

- b) Haga una tabla de frecuencias para estos datos usando las clases $(0, 50]$, $(50, 100]$, $(100, 200]$ y $(200, \infty)$. Calcule el estimador de máxima verosimilitud de las frecuencias esperadas para estas clases bajo el supuesto en el inciso anterior. ¿La distribución exponencial parece ser un modelo adecuado para los datos?

RESPUESTA

Realizamos la tabla de frecuencias con los intervalos solicitados utilizando R. Primero ingresamos los datos

```
t_ejericio2 <-c (97,51,11,4,141,18,142,68,77, # datos
                80,1,16,106,206,82,54,31,216,
                46,111,39,63,18,191,18,163,24)

intervalos <- c(0, 50, 100, 200, Inf) # intervalos
datos_intervalos <- as.factor(cut(t_ejericio2, breaks = intervalos))
frecuencias_observadas <- table(datos_intervalos) # frecuencias.
frecuencias_observadas
```

```
## datos_intervalos
##      (0,50]  (50,100]  (100,200]  (200,Inf]
##           11         8         6         2
thetha <- mean(t_ejercicio2) # estimador mle
fda_intervalos <- pexp(intervalos[2:5], 1/thetha) # probabilidades consi. el mle
frecuencias_esperadas<- c(fda_intervalos[1],diff( fda_intervalos))*27
frecuencias_esperadas # frecuencias de los intervalos

## [1] 12.917671  6.737440  5.346837  1.998051
```

Por lo tanto, tenemos que las frecuencias observadas y las frecuencias considerando el estimador de máxima verosimilitud son

Intervalos	Frecuencias observadas	Frecuencias esperadas
(0, 50]	11	12.9176708
(50, 100]	8	6.7374405
(100, 200]	6	5.3468372
(200, ∞)	2	1.9980515

Por lo que, **podemos concluir que la distribución exponencial parece ser un modelo adecuado para los datos.** ■.

- Se hicieron 27 mediciones de los rendimientos de dos procesos industriales, con los siguientes resultados:

$$\text{Proceso 1: } n_1 = 11 \quad \bar{y}_1 = 6,23 \quad s_1^2 = 3,79$$

$$\text{Proceso 2: } n_2 = 16 \quad \bar{y}_2 = 12,74 \quad s_2^2 = 4,17$$

Suponiendo que los rendimientos se distribuyen normalmente con la misma varianza, encuentre intervalos de confianza para las medias μ_1 y μ_2 y para la diferencia de medias $\mu_1 - \mu_2$.

RESPUESTA

Considerando que los procesos industriales son independientes y como en el problema nos nos dice la varianza de cada proceso podemos decir que es desconocida. Ahora, podemos utilizar el TLC para crear los intervalos de confianza pero como n no son grandes estos intervalos de confianza no serían muy adecuados para este problema. Como nos dan s^2 para cada proceso (s^2 es un estimador de σ^2). Entonces para construir intervalos de confianza del 100%(1 - α) para μ , primero recordemos que para poblaciones normales (sin importar que tan grande sea n) tenemos que

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \quad \text{y} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2, \quad (1)$$

ambas independientes (Vease la pag. 86 para la primera parte, la pag. 96 para segunda parte las demostraciones de lo anterior). Ahora, recordemos el siguiente resultado.

Teorema: 1 (Diapositiva 553, primera parte) Sean X y Y variables aleatorias independientes tales que

$$X \sim N(0, 1), \quad Y \sim \chi_n^2.$$

Definamos $T = X/\sqrt{Y/n}$ entonces $T \sim t_n$.

Entonces ocupando el resultado anterior (1), y podemos crear el cociente de las dos variables (1) para obtener una nueva variable con distribución $t - Student$, es decir, consideremos

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{S}{\sigma}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

Entonces como lo anterior depende de X_1, X_2, \dots, X_n a través de \bar{X} y su distribución no depende de μ , ni de ningún parámetro desconocido podemos considerarlo como nuestro pivote. Ahora considerando una confiabilidad de $(1 - \alpha)$, tenemos que podemos definir constantes a y b (percentiles) tales que

$$\mathbb{P}\left(a < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < b\right) = 1 - \alpha \Leftrightarrow \mathbb{P}\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Y haciendo algunos pasos algebraicos tenemos que

$$\mathbb{P}\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}\right) = 1 - \alpha \Leftrightarrow \mathbb{P}\left(\bar{X} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Y esto implica, que un intervalo de confianza del 100%(1 - α) para μ (adecuado para este problema) sea

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \Leftrightarrow (\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}).$$

Por lo tanto, **un intervalo de confianza del 95% (es decir, considerando $\alpha = 0,05$) para μ_1 es** (considerando que, $n = 11, \bar{y}_1 = 6,23, s_1^2 = 3,79, t_{\alpha/2,10} = 2,228$, utilizando R `qt(0.975, 10)`)

$$\left(6,23 - 2,228 \cdot \sqrt{\frac{3,79}{11}}, 6,23 + 2,228 \cdot \sqrt{\frac{3,79}{11}}\right) = (6,23 - 1,308, 6,23 + 1,308) = \mathbf{(4,922, 7,538)}$$

y **un intervalo de confianza del 95% (es decir, considerando $\alpha = 0,05$) para μ_2 es** (considerando que, $n = 16, \bar{y}_1 = 12,74, s_1^2 = 4,17, t_{\alpha/2,15} = 2,131$, utilizando R `qt(0.975, 16)`)

$$\left(12,74 - 2,131 \cdot \sqrt{\frac{4,17}{16}}, 12,74 + 2,131 \cdot \sqrt{\frac{4,17}{16}}\right) = (12,74 - 1,088, 12,74 + 1,088) = \mathbf{(11,652, 13,828)}.$$

Ahora, para construir un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ no sería factible utilizar el TLC ya que son muestras pequeñas (si fueran muestras grandes el intervalo muy intuitivo). Otro punto a considera es que no conocemos la varianza real de los por lo que no podemos normalizar y encontrar un pivote sencillo. Debido a que en el problema nos dicen que tienen la misma varianza, podemos hacer un razonamiento análogo de como se construyo el intervalo de confianza para las medias por separado. Entonces, primero encontremos un pivote para la estimación de $\mu_1 - \mu_2$. Suponiendo que ambos procesos son independientes entre si, entonces como son normales tenemos que

$$\mathbb{E}(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2, \quad Var(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

donde $\mu_i = \mathbb{E}(\bar{Y}_i)$ y $\sigma_1^2 = \sigma_2^2 = \sigma^2 = Var(\bar{Y}_1)$. Ahora utilizando las propiedades de las distribuciones normales, podemos concluir que

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \Rightarrow \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1).$$

Cómo la varianza de ambos procesos es desconocida tenemos que estimarla. Para ello, sabemos que (por la misma justificación de la primera parte del problema)

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi_{n_1-1}^2 \quad \text{y} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_2-1}^2. \quad (2)$$

Entonces si definimos a

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)},$$

Es sencillo mostrar que $\mathbb{E}(S_p^2) = \sigma^2$ y que $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi_{n_1+n_2-2}^2$ (se demuestra en el ejercicio 3 de las notas de esta tarea). Entonces como $\bar{Y}_1, \bar{Y}_2, S_1^2, S_2^2$ son independientes entre sí (por definición, y por que los procesos son independientes) esto implica que $\bar{Y}_1 - \bar{Y}_2$ y S_p^2 también lo sean. Entonces utilizando el resultado (1) y simplificando tenemos que

$$\begin{aligned} \frac{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\frac{\frac{1}{\sigma} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}}{\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}} &\sim t_{n_1+n_2-2} \Leftrightarrow \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}} \sim t_{n_1+n_2-2} \Leftrightarrow \\ \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} &\sim t_{n_1+n_2-2} \end{aligned}$$

De lo anterior podemos concluir que

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

es un pivote para la estimación de $\mu_1 - \mu_2$. Entonces, el intervalo de confianza del 100%(1 - α) se obtendría como:

$$\mathbb{P} \left(-t_{\alpha/2} < \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{\alpha/2} \right) = 1 - \alpha.$$

Y lo anterior implica que el intervalo de confianza del 100%(1 - α) para $\mu_1 - \mu_2$ (cuando se desconoce la varianza, pero se sabe que tiene la misma varianza y las muestras son pequeñas) sería

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \Leftrightarrow \left(\bar{y}_1 - \bar{y}_2 - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{y}_1 - \bar{y}_2 + t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right).$$

Entonces, con los datos de este ejercicio tenemos que

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 &= 6,23 - 12,74 = -6,51, \quad n_1 = 11, n_2 = 16, s_1^2 = 3,179, s_2^2 = 4,17, \\ t_{\alpha/2, n_1+n_2-2} &= t_{0,025,25} = 2,06, \quad \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{1}{11} + \frac{1}{16}} = 0,392, \\ s_p &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} = \sqrt{\frac{(11 - 1)(3,179) + (16 - 1)(4,17)}{(11 + 16 - 2)}} = \sqrt{\frac{94,34}{25}} = 1,94. \end{aligned}$$

Por lo tanto, el intervalo de confianza del 95 % para $\mu_1 - \mu_2$ es

$$(-6,51 - (2,06)(1,94)(0,392), -6,51 + (2,06)(1,94)(0,392)) \Leftrightarrow (-8,08, -4,943411) \quad \blacksquare.$$

Interpretación: como el intervalo de confianza no contiene el cero, entonces no hay evidencia para decir que los dos procesos tengan el mismo comportamiento medio.

4. Un experimento para determinar el efecto de una medicina en la concentración de glucosa en la sangre de ratas diabéticas dio los siguientes resultados:

Grupo control:	2,05	1,82	2,00	1,94	2,12	
Grupo tratamiento:	1,71	1,37	2,04	1,50	1,69	1,83

Analiza la hipótesis de que el tratamiento no tiene efecto sobre la media de la concentración de la glucosa en la sangre. Mencione las hipótesis bajo las cuales realiza el análisis.

RESPUESTA

Definición: 1 Sea X_1, X_2, \dots, X_{n_1} v.a. independientes con distribución $\text{Normal}(\mu_X, \sigma_X^2)$ (donde σ_X^2 es desconocida, y Y_1, Y_2, \dots, Y_{n_2} v.a. independientes con distribución $\text{Normal}(\mu_Y, \sigma_Y^2)$ (donde σ_Y^2 es desconocida, y X_i es independiente a Y_i . Entonces si queremos verificar que las medias de la población son distintas, entonces se plantean el juego de hipótesis:

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X \neq \mu_Y.$$

Esta prueba tiene el estadístico de prueba

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

donde

$$s_p = \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}.$$

Y la región de rechazar la hipótesis nula H_0 con un nivel α es si $|t| > t_{n_1+n_2-2, \alpha/2}$.

Observemos que el tamaño de las muestras son pequeñas, por lo que no podemos utilizar el TLC para determinar un estadístico de prueba. Ahora, no conocemos ni la distribución de la población ni la varianza real, pero para este caso supondremos que la distribución es normal. Para validar este planteamiento de que las poblaciones son normales, utilizaremos la prueba de normalidad de Shapiro Wilk, la cual consiste en una prueba de bondad de ajuste muy usada para probar la Hipótesis Nula de que la muestra viene de una población normal contra la Hipótesis Alternativa de que proviene de alguna otra población no normal. Su consideración equivale a verificar normalidad mediante un método no gráfico, sino propiamente estadístico. Utilizando la función integrada de R tenemos que

```
grupo_control <- c(2.05, 1.82, 2.00, 1.94, 2.12) # datos.
grupo_tratamiento <- c(1.71, 1.37, 2.04, 1.50, 1.69, 1.82)
```

```
shapiro.test(grupo_control) # test shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  grupo_control
## W = 0.98323, p-value = 0.9511
```

```
shapiro.test(grupo_tratamiento)
```

```
##
```

```
## Shapiro-Wilk normality test
##
## data: grupo_tratamiento
## W = 0.9806, p-value = 0.9545
```

Entonces, como los estadísticos de prueba W para ambas pruebas están cercanos a uno podemos concluir que efectivamente los datos provienen de una distribución normal.

Nos interesa saber si el tratamiento tiene efecto sobre la media de la concentración de la glucosa en la sangre entonces podemos usar la prueba t -student (definición 1) para validar este resultado. Tenemos el siguiente juego de hipótesis

$$H_0 : \mu_T = \mu_C \quad \text{vs.} \quad H_1 : \mu_T \neq \mu_C.$$

donde μ_T es la media del grupo tratamiento y μ_C es la media del grupo control. Nuestro estadístico de prueba es (realizamos los cálculos en R)

```
n_1 <- length(grupo_control) # tamaños
n_2 <- length(grupo_tratamiento)

x_bar <- mean(grupo_control) # medias muestrales
y_bar <- mean(grupo_tratamiento)

std2_x <- var(grupo_control) # varianza muestral
std2_y <- var(grupo_tratamiento)

sp <- sqrt(((n_1-1)*std2_x+(n_2-1)*std2_y)/(n_1+n_2-2)) # sp

t <- (x_bar-y_bar)/(sp*sqrt(1/n_1+1/n_2)) # estadístico prueba
t

## [1] 2.566404
```

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2,566404,$$

```
qt(0.975, 9)
```

```
## [1] 2.262157
```

Entonces con $\alpha = 0,05$ tenemos que el percentil de la distribución t -student es $t_{n_1+n_2-2, \alpha/2} = t_{9, 0,025} = 2,262157$. Por lo tanto, como $|t| = 2,57 > 2,26 = t_{9, 0,025}$ **podemos rechazar la hipótesis nula y concluir que existe un efecto de la medicina en la concentración de glucosa en la rangre de ratas diabéticas.** ■.

5. Sea Y el tiempo hasta que falla cierto componente eléctrico. La distribución de Y es exponencial con media θ/t , donde t es la temperatura a la cual el componente opera. Supongamos que n componentes se prueban de manera independiente a temperaturas t_1, t_2, \dots, t_n , respectivamente, y que sus tiempos de vida observados son y_1, y_2, \dots, y_n . Derive una expresión para el estimador de máxima verosimilitud de θ .

RESPUESTA

Teorema: 2 Sea X una v.a. con función generadora de momentos $M_x(t)$, entonces si $Y = aX + b$, entonces $M_Y(t) = e^{bt} M_X(at)$.

Observemos que si $Y \sim \text{Exponencial}(\theta/t)$ y por lo tanto su generadora de momentos es

$$M_Y(p) = \left(1 - \frac{pt}{\theta}\right)^{-1}.$$

Ahora consideremos la variable aleatoria $W = Y/t$, utilizando (2) tenemos

$$M_W(p) = e^{t0} M_Y\left(\frac{p}{t}\right) = \left(1 - \frac{p}{\theta}\right)^{-1}.$$

Por lo que podemos concluir que $W \sim \text{Exponencial}(\theta)$. Entonces como tenemos observaciones de la variable Y pero con diferentes temperaturas, podemos considerar tenemos que la función de verosimilitud es

$$L(\theta|w_1, \dots, w_n) = L(\theta|y_1/t_1, \dots, y_n/t_n) = \prod_{i=1}^n f_{W_i}(w_i) = \prod_{i=1}^n \theta e^{-\frac{\theta}{t_i} y_i} = \theta^n e^{-\theta \sum_{i=1}^n \frac{y_i}{t_i}},$$

y por lo tanto

$$\log(L(\theta)) = n \log(\theta) - \theta \sum_{i=1}^n \frac{y_i}{t_i}$$

Ahora derivamos con respecto a θ e igualamos a cero para obtener los puntos críticos, obtenemos que

$$\frac{d}{d\theta} \log(L(\theta)) = \frac{n}{\theta} - \sum_{i=1}^n \frac{y_i}{t_i}$$

así que

$$\frac{n}{\theta} - \sum_{i=1}^n \frac{y_i}{t_i} = 0 \Rightarrow \text{el punto crítico es } \hat{\theta} = \frac{n}{\sum_{i=1}^n \frac{y_i}{t_i}}.$$

Usando el criterio de segunda derivada verifiquemos que efectivamente el punto crítico es un máximo, tenemos que

$$\frac{d^2}{d\theta^2} \log L(\theta) = -\frac{n}{\theta^2}.$$

Por lo tanto, como la segunda derivada es negativa podemos concluir que el punto crítico es máximo, y **por lo tanto el estimador de máxima verosimilitud para θ es:**

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \frac{y_i}{t_i}}. \quad \blacksquare.$$

6. Diez componentes electrónicos con tiempos de vida distribuidos exponencialmente fueron probados por períodos de tiempo determinados. Tres de los componentes sobrevivieron sus periodos de prueba y los siete restantes fallaron en los siguientes tiempos.

No. Componente:	1	2	3	4	5	6	7	8	9	10
Período de prueba:	81	72	70	60	41	31	31	30	29	21
Tiempo de falla:	2	—	51	—	33	27	14	24	4	—

Encuentre el estimado de máxima verosimilitud para la media de la distribución exponencial θ .

RESPUESTA

Nota: este problema lo sentí un poco fuera de lugar debido a que se tienen que saber un poco más de conocimiento de temas (análisis de supervivencia) de estadística para resolverlo, o no se si existe una manera más sencilla de plantearlo.

Ocupemos la teoría de análisis de supervivencia tenemos el siguiente teorema.

Teorema: 3 Sea X_1, X_2, \dots, X_n una m.a con función de densidad $f(t)$ y constantes de censuras asociadas C_1, C_2, \dots, C_n respectivamente. Bajo la censura por la derecha con tiempos de censura fijos, la función de verosimilitud, $L(\theta)$, de los datos observados (t_i, δ_i) , $i = 1, 2, \dots, n$ está dada por

$$L(\theta) = \prod_{i=1}^n [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}$$

donde $\delta_i = I\{Y \leq C_i\}$ es su indicador de censura, t_1, t_2, \dots, t_n es una realización con censura presente de T_1, T_2, \dots, T_n , y $S(t_i)$ es la función de supervivencia.

Ocupemos el teorema (3) para el caso exponencial, es decir. Sea X_1, X_2, \dots, X_n una m.a. con distribución $Exp(\theta)$ la cual esta censurada por la derecha, de tamaño n . Entonces la función de densidad es

$$f(x_i) = \theta \exp(-\theta y_i), \quad y > 0.$$

Ahora calculemos la función de supervivencia,

$$S(x_i) = \mathbb{P}(X_i > t) = 1 - F(t) = 1 - (1 - \exp(-\theta y_i)) = \exp(-\theta y_i).$$

Denotemos por

$$(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)$$

el conjunto de datos de tiempos de supervivencia observados. De esa manera, si X_i no está censurado ($\delta_i = 1$), con tiempo de supervivencia observado en x_i , el valor realizado t_i de T_i es igual a x_i , por el contrario, si X_i está censurado en c_i ($\delta_i = 0$), entonces $t_i = c_i$ y $X_i > c_i$. Supongamos ahora que las observaciones han sido reordenadas de manera que t_1, t_2, \dots, t_r denotan las r observaciones no censuradas y $t_{r+1}, t_{r+2}, \dots, t_n$ las $n - r$ observaciones censuradas. Entonces ocupando el teorema (3) tenemos la función de verosimilitud para θ formada en base a \mathbf{x} , es

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n [\theta \exp(-\theta x_i)]^{\delta_i} [\exp(-\theta x_i)]^{1-\delta_i} \\ &= \theta^r \exp\left(-\theta \sum_{i=1}^n t_i\right). \end{aligned}$$

La simplificación de la igualdad anterior se debe a que tenemos r observaciones no censuradas y $n - r$ censuradas. Así, la función de log-verosimilitud es

$$\log L(\theta) = r \log \theta - \theta \sum_{i=1}^n t_i.$$

Ahora derivamos con respecto a θ e igualamos a cero para obtener los puntos críticos, obtenemos que

$$\frac{d}{d\theta} \log L(\theta) = \frac{r}{\theta} - \sum_{i=1}^n t_i$$

así que

$$\frac{r}{\theta} - \sum_{i=1}^n t_i = 0 \Rightarrow \text{el punto crítico es } \hat{\theta} = \sum_{i=1}^n \frac{r}{t_i}.$$

Usando el criterio de segunda derivada verifiquemos que efectivamente el punto crítico es un máximo, tenemos que

$$\frac{d^2}{d\theta^2} \log L(\theta) = -\frac{r}{\theta^2}$$

Por lo tanto, como la segunda derivada es negativa podemos concluir que el punto crítico encontrado es un máximo, y por lo tanto el estimador de máxima verosimilitud para θ es

$$\hat{\theta} = \sum_{i=1}^n \frac{r}{t_i}.$$

Regresando al problema original, observemos que tenemos un m.a. con distribución $Exp(\theta)$ con censurada a la derecha (ya que se pone un período de prueba). Entonces podemos aplicar todo lo descrito anteriormente para determinar el estimado de máxima verosimilitud para θ . Tenemos los datos observados

$$(t_i, \delta_i) = (2, 1), (72, 0), (51, 1), (60, 0), (33, 1), (27, 1), (14, 1), (24, 1), (4, 1), (21, 0).$$

entonces tenemos 7 datos no censurados (es decir, $r = 7$) esto implica que el estimador de máxima verosimilitud es

$$\sum_{i=1}^n \frac{r}{t_i} = \frac{7}{2 + 72 + 51 + 60 + 33 + 27 + 14 + 24 + 4 + 21} = \frac{7}{308} = \mathbf{0,02272727} \quad \blacksquare.$$

Esto se puede interpretar que el tiempo esperado de sobrevivencia para los componentes es de 44 (definición de esperanza de una distribución exponencial $1/0.02272727$) unidades de tiempo.

7. Una máquina de bebidas está diseñada para descargar, cuando opera apropiadamente, al menos 7 onzas de bebida por taza con una desviación estándar de 0.2 onzas. Si un estadístico selecciona una muestra aleatoria de 16 tazas para examinar el servicio al cliente y este está dispuesto a tomar un riesgo $\alpha = 0,05$ de cometer un error de Tipo I, calcule la potencia de la prueba y la probabilidad (β) de tener un error del Tipo II si la media poblacional de la cantidad despachada es:

a) 6.9 onzas por taza.

b) 6.8 onzas por taza.

Puede asumir que los datos son normales.

RESPUESTA

Tenemos que el proceso opera apropiadamente si cada taza tiene al menos 7 onzas. Entonces, nos interesa probar si el funcionamiento de la máquina no opera adecuadamente. Como sabemos que la población es normal con desviación estandar conocida, por lo que nuestro juego de hipótesis es

$$H_0 : \mu \geq 7 \quad \text{vs.} \quad H_1 : \mu < 7.$$

Entonces para este juego de hipótesis considerando un α la probabilidad (β) de tener un error del Tipo II esta definido como (ver pag. 220 Diapositivas)

$$\beta = \mathbb{P} \left(Z > -z_\alpha - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \right) = 1 - \Phi \left(-z_\alpha - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \right)$$

y la potencia de la prueba se define como

$$\text{Potencia de la prueba} = \mathbb{P}\left(Z \leq -z_{\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right) = \Phi\left(-z_{\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right).$$

Por lo que para este problema consideramos $\alpha = 0,05$, $\sigma = 0,2$, $\mu_0 = 7$, $n = 16$ podemos graficar la potencia de la prueba y la probabilidad de obtener un error del Tipo II para cualquier algún μ_1 de la siguiente forma

```
library(tidyverse)
n <- 16 # datos del problema
mu_0 <- 7
std_pobla <- 0.2
alpha <- 0.05

z_alpha <- qnorm(0.95) # cuantil normal estandar

mu_1 <- seq(6.5,7.5,0.01) # rango de mu_1

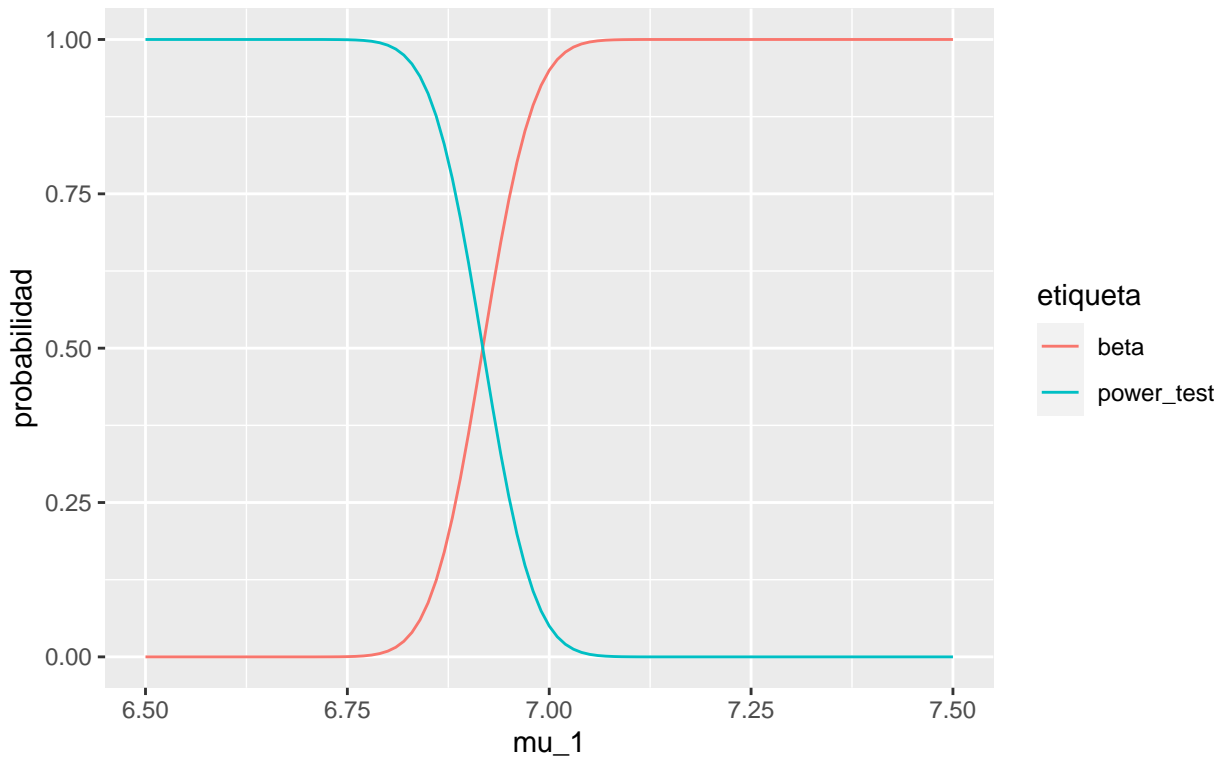
beta <- 1- pnorm(-z_alpha-(sqrt(n)*(mu_1-mu_0))/std_pobla)

power_test <- pnorm(-z_alpha-(sqrt(n)*(mu_1-mu_0))/std_pobla)

datos <- data.frame(mu_1=rep(mu_1,2), probabilidad = c(beta,power_test),
                    etiqueta = c(rep("beta",length(mu_1)), rep("power_test",length(mu_1))) )

ggplot(datos, aes(x= mu_1, y= probabilidad, group=etiqueta, col=etiqueta))+
  geom_line()+
  labs(title="Grafica de beta y la potencia de la prueba, para distintos medias
          alternativas.")
```

Grafica de beta y la potencia de la prueba, para distintos medias alternativas.



casos particulares.

```
power_mu_6_9 <- datos[(datos$mu_1==6.9)&(datos$etiqueta=="power_test"),]$probabilidad
beta_mu_6_9 <- datos[(datos$mu_1==6.9)&(datos$etiqueta=="beta"),]$probabilidad
power_mu_6_8 <- datos[(datos$mu_1==6.8)&(datos$etiqueta=="power_test"),]$probabilidad
beta_mu_6_8 <- datos[(datos$mu_1==6.8)&(datos$etiqueta=="beta"),]$probabilidad
```

Para el caso particular: a) $\mu_1 = 6,9$ la potencia de la prueba es igual a 0.63876, y la probabilidad de obtener un error del Tipo II es igual a 0.36124. Y para a) $\mu_1 = 6,8$ la potencia de la prueba es igual a 0.9907423, y la probabilidad de obtener un error del Tipo II es igual a 0.0092577. Por lo tanto, para el caso 2 tenemos una potencia casi de uno, lo que podemos interpretar que la prueba es muy buena para detectar que nuestra muestra provenga de una población distinta de la especifica la hipótesis nulas, es decir, que nos indica rechazar H_0 cuando sea falsa. ■.

8. Construya un intervalo de confianza aproximado del 90 % para el parámetro λ de una distribución de Poisson. Evalúe su intervalo si una muestra de tamaño 30 produce $\sum x_i = 240$.

RESPUESTA

Sea X_1, X_2, \dots, X_n una muestra aleatoria de tamaño n de una distribución $\text{Poisson}(\lambda)$. Entonces consideremos n grande, lo que implica podemos usar el teorema de límite central y concluir que

$$\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\bar{X} - \lambda}{\frac{\lambda}{\sqrt{n}}} \rightarrow N(0, 1)$$

cuando $n \rightarrow \infty$. Recordemos que el estimador de máxima verosimilitud para la varianza para una distribución poisson por lo que podemos utilizar $\hat{\lambda} = \bar{X}$ y así tener un pivote para λ , es decir,

$$\frac{\bar{X} - \lambda}{\frac{\bar{X}}{\sqrt{n}}} \rightarrow N(0, 1)$$

Ahora considerando una confiabilidad de $(1 - \alpha)$, tenemos que podemos definir constantes a y b (percentiles) tales que

$$\mathbb{P} \left(a < \frac{\bar{X} - \lambda}{\frac{\bar{X}}{\sqrt{n}}} < b \right) = 1 - \alpha \Leftrightarrow \mathbb{P} \left(-z_{\alpha/2} < \frac{\bar{X} - \lambda}{\frac{\bar{X}}{\sqrt{n}}} < z_{\alpha/2} \right) = 1 - \alpha.$$

Y haciendo algunos pasos algebraicos tenemos que

$$\mathbb{P} \left(-z_{\alpha/2} < \frac{\bar{X} - \lambda}{\frac{\bar{X}}{\sqrt{n}}} < z_{\alpha/2} \right) = 1 - \alpha \Leftrightarrow \mathbb{P} \left(\bar{X} - z_{\alpha/2} \cdot \sqrt{\frac{\bar{X}}{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \sqrt{\frac{\bar{X}}{n}} \right) = 1 - \alpha.$$

Y por lo tanto, **que un intervalo de confianza del 100%(1 - α) para λ (utilizando el TLC) sea**

$$\bar{x} \pm z_{\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}} \Leftrightarrow \left(\bar{x} - z_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} \right).$$

Entonces, un intervalo de confianza aproximado del 90% ($\alpha = 10$, $z_{0,05} = 1,64$) para λ considerando $n = 30$ y $\sum x_i = 240$ es

$$\left(\bar{x} - z_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} \right) = \left(8 - (1,64) * \sqrt{\frac{8}{30}}, 8 + (1,64) * \sqrt{\frac{8}{30}} \right) = \mathbf{(7,153, 8,846)}. \blacksquare.$$

9. Sea $X_1, \dots, X_n \sim Uniforme(0, \theta)$ y sea $Y = \max\{X_1, \dots, X_n\}$. Deseamos probar

$$H_0 : \theta = 1/2 \quad \text{vs.} \quad H_1 : \theta > 1/2.$$

Supongamos que decidimos probar esta hipótesis rechazando H_0 cuando $Y > c$.

- Calcule el poder de la prueba.
- ¿Qué elección de c tiene que tomarse para que la prueba tenga una significancia de 0.05?
- En una muestra de tamaño $n = 20$ con $Y = 0,48$, ¿cuál es el p-valor? ¿Qué conclusión sobre H_0 haría?
- En una muestra de tamaño $n = 20$ con $Y = 0,52$, ¿cuál es el p-valor? ¿Qué conclusión sobre H_0 haría?

RESPUESTA

Por como esta definido Y , tenemos que $Y = X_{(n)}$ es decir, es el estadístico de orden n . Como $X \sim Unif(0, \theta)$ implica que la densidad de Y sea

$$f_Y(y) = n \left(\int_0^y \frac{1}{\theta} dy \right)^{n-1} \frac{1}{\theta} = \frac{y^{n-1}}{\theta^n}.$$

Entonces la probabilidad de rechazar H_0 dado que H_0 (es decir, α) es

$$\mathbb{P}(\text{Rechazar } H_0 | H_0) = \mathbb{P}(Y > c | H_0) = \int_c^{\theta_0} \frac{ny^{n-1}}{\theta_0^n} dy = \frac{y^n}{\theta_0^n} \Big|_c^{\theta_0} = 1 - \frac{c^n}{\theta_0^n} = \alpha.$$

Entonces, para un α se rechaza H_0 si Y excede $\theta_0(1-\alpha)^{1/n}$. Ocupando lo anterior, podemos concluir que la potencia del prueba es

$$\begin{aligned}\text{Poder de la prueba} &= \mathbb{P}(\text{Rechazar } H_0 | H_1) = 1 - \mathbb{P}(\text{No rechazar } H_0 | H_1) = 1 - \int_0^{\theta_0(1-\alpha)^{1/n}} \frac{ny^{n-1}}{\theta_1^n} dy \\ &= 1 - \left. \frac{y^n}{\theta_1^n} \right|_0^{\theta_0(1-\alpha)^{1/n}} = 1 - \frac{\theta_0^n(1-\alpha)}{\theta_1^n}.\end{aligned}$$

Y para este problema en específico sería como $\theta_0 = \frac{1}{2}$, **a) el poder de la prueba dado un α y un θ_1 es $1 - \frac{1-\alpha}{2^n\theta_1^n}$.** Ahora para que la prueba tenga una significativa de 0,05 tenemos que

$$0,05 = 1 - \frac{c^n}{\theta_0^n} \Rightarrow c^n = \frac{(1-0,05)}{2^n} \Rightarrow c = \frac{(1-0,05)^{1/n}}{2},$$

es decir, **b) tenemos que elegir a c como $\frac{(1-0,05)^{1/n}}{2}$.**

Ahora, el p -valor para este problema se define como

$$\begin{aligned}p\text{-valor} &= \mathbb{P}(\text{el estadístico } Y \text{ sea mayor del valor observado} | H_0) = \mathbb{P}(Y > y_{obs} | H_0) \\ &= \int_{y_{obs}}^{\theta_0} \frac{ny^{n-1}}{\theta_0^n} dy = \left. \frac{y^n}{\theta_0^n} \right|_{y_{obs}}^{\theta_0} = 1 - \frac{y_{obs}^n}{\theta_0^n}.\end{aligned}$$

Entonces si el tamaño $n = 20$ con $Y = 0,48$ **c) el p -value es**

$$p\text{-valor} = 1 - \frac{0,48^{20}}{0,5^{20}} = 0,5579976.$$

Por lo que no hay evidencia significativa de rechazar H_0 . Ahora si consideramos el tamaño $n = 20$ con $Y = 0,52$ **d) el p -value es 0**, ya que $y_{obs} > \theta_0 = 0,5$. Por lo que en este caso, si existe evidencia significativa de rechazar H_0 y por lo tanto concluir que $\theta > 1/2$. ■.

10. Sea $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$.

a) Sea $\lambda_0 > 0$. Asumiendo que n es suficientemente grande, proponga un estadístico de prueba para examinar

$$H_0 : \lambda = \lambda_0 \quad \text{vs.} \quad H_1 : \lambda \neq \lambda_0.$$

RESPUESTA

Recordemos la definición de la prueba de Wald.

Definición: 2 Considera el juego de hipótesis

$$H_0 : \lambda = \lambda_0 \quad \text{vs.} \quad H_1 : \lambda \neq \lambda_0.$$

Asuma que $\hat{\theta}$ es asintóticamente Normal:

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\text{Var}(\hat{\theta})}} \rightarrow N(0, 1),$$

cuando $n \rightarrow \infty$. La prueba de Wald de considerando α es rechazar H_0 cuando $|W| > z_{\alpha/2}$ donde W es el estadístico de prueba definido como

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}}.$$

Sabemos que el estimador de máxima verosimilitud para λ de una población Poisson es \bar{X} (demostrado en clase), es decir, $\hat{\lambda} = \bar{X}$ y de igual manera sabemos que la varianza de este estimador es $\frac{\bar{X}}{n}$. Entonces, por el TLC podemos decir que

$$\frac{\bar{X} - \lambda_0}{\sqrt{\frac{\bar{X}}{n}}} \sim N(0, 1)$$

cuando $n \rightarrow \infty$. Entonces como tenemos un estimador que es asintoticamente Normal, podemos ocupar la prueba de Wald (definición 2) y **tener el estadístico de prueba**

$$\frac{\bar{X} - \lambda_0}{\sqrt{\frac{\bar{X}}{n}}}.$$

b) Establezca la región de rechazo para un nivel de significancia de $\alpha = 0,05$.

RESPUESTA

Por el inciso anterior, y por la definición de la prueba de Wald (2) tenemos que la región de rechazo con $\alpha = 0,05$ es

$$P_{\theta_0}(|W| > z_{\alpha/2}) = 0,05 \Leftrightarrow P_{\theta_0}(|W| > 1,96) = 0,05,$$

que es equivalente a rechazar H_0 cuando $|W| > 1,96$.

c) Sea $\lambda_0 = 1, n = 20$ y $\alpha = 0,05$. Simule $X_1, \dots, X_n \sim \text{Poisson}(\lambda_0)$ y aplique la prueba que propuso. Repita esto 10^4 veces y cuente que tan a menudo rechaza la hipótesis nula. ¿Qué tan cercana es la tasa de error tipo I de 0.05? ¿Qué hay de la tasa de error tipo II?

RESPUESTA

```
set.seed(19970808)
lambda_0 <- 1 # datos del problema
n <- 20
z_alpha <- qnorm(0.975)
simulacion_104 <- rpois(n*10**4, lambda = lambda_0) # simulaciones
simulacion_20_104 <- matrix(simulacion_104, ncol=n)
media_simulaciones <- rowMeans(simulacion_20_104) # media
W <- abs(media_simulaciones-lambda_0)/sqrt(media_simulaciones/n) # estadístico

table(W > z_alpha)[1] # tabla de frecuencias

## FALSE
## 9488
```

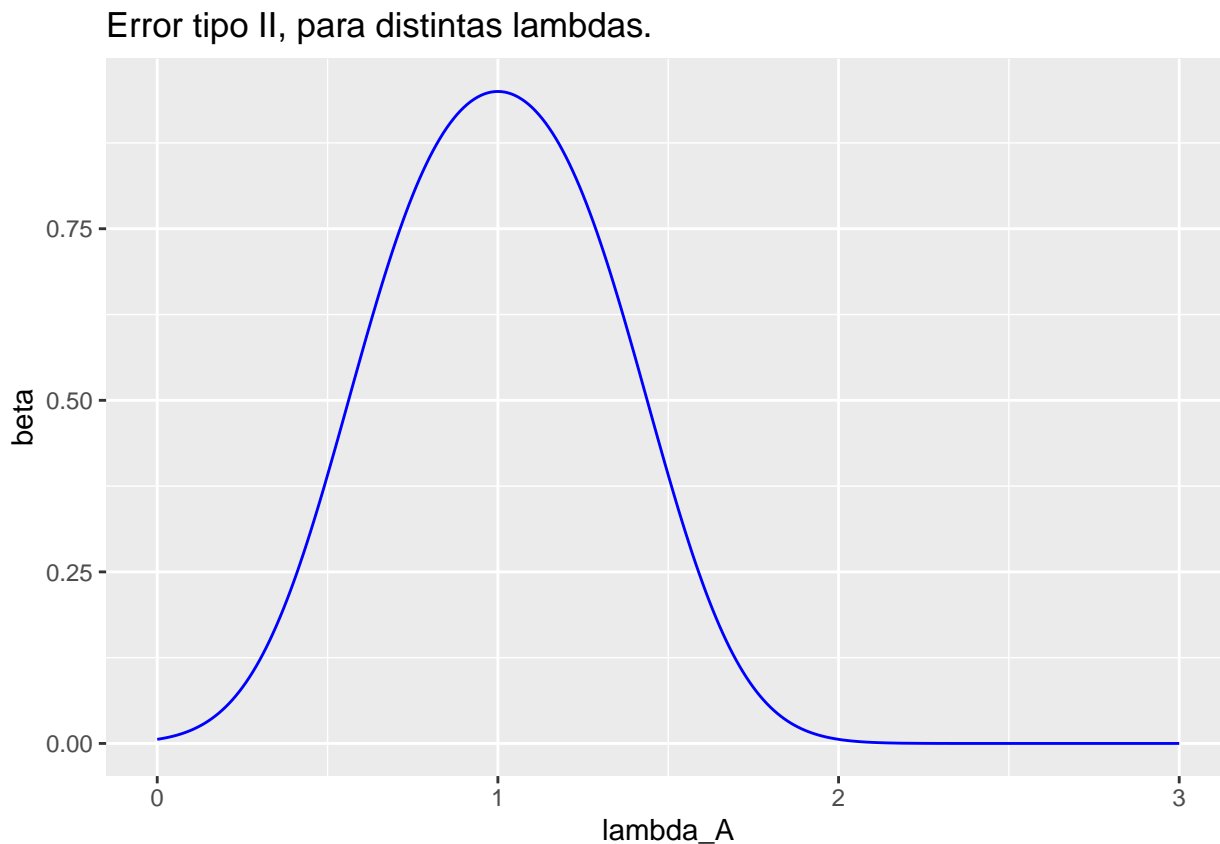
Por lo tanto, en el 0.0512 % de las simulaciones es mayor en el 0.9488 % es menor, lo que coinciden con la probabilidad de rechazar la hipótesis nula. Es decir, **estamos muy cerca del error tipo I de 0.05**. Sobre el error tipo II, tenemos que para este juego de hipótesis considerando $\bar{X} \sim N(\lambda_A, \frac{1}{\sqrt{20}})$ entonces

$$\begin{aligned} \beta &= \mathbb{P}(\text{No rechazar } H_0 | H_A) = \mathbb{P}(\bar{X} > \lambda_0 + z_{\alpha/2} \sqrt{\sigma/n}) + \mathbb{P}(\bar{X} < \lambda_0 - z_{\alpha/2} \sqrt{\sigma/n}) \\ &= \mathbb{P}(\bar{X} > 1,438) + \mathbb{P}(\bar{X} < 0,562) \end{aligned}$$

Entonces, calculemos y gráfiquemos el error tipo II para distintas λ_A .


```
lambda_A <- seq(0, 3, 0.01)
beta <- pnorm(1.43826, mean=lambda_A, sd=(1)/(sqrt(20)) , lower.tail = TRUE) - pnorm(0.5617307 , mean=1.43826, sd=(1)/(sqrt(20)) , lower.tail = TRUE)

ggplot(data= data.frame(lambda_A, beta), aes(x=lambda_A, y=beta))+
  geom_line(col="blue")+
  labs(title="Error tipo II, para distintas lambdas.")
```



11. En la librería boot de R acceder los datos *cd4* los cuales son conteos de células CD4 en pacientes VIH-positivos antes y después de un año de tratamiento con un antiviral.

- a) Construya un intervalo de confianza bootstrap para el coeficiente de correlación entre los conteos base y los conteos después del tratamiento.

RESPUESTA

Consideremos la metodología de Bootstrap, específicamente consideremos los intervalos BC_a los cuales son una mejora con respecto a los intervalos básicos y de percentiles Bootstrap. Entonces tenemos que el intervalo BC_a con cobertura $1 - 2\alpha$ es

$$BC_a = (\hat{\theta}_{\alpha_1}^*, \hat{\theta}_{\alpha_2}^*)$$

donde

$$\alpha_1 = \phi \left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)} \right)$$

$$\alpha_2 = \phi \left(z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})} \right)$$

Y donde z_0 es un cuantil asociado a la proporción de réplicas bootstrap menores que $\hat{\theta}$ (el estimador

original):

$$z_0 = \phi^{-1} \left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B} \right)$$

es decir, z_0 es una medida del sesgo de $\hat{\theta}^*$. Y por último a se define como

$$a = \frac{\sum_{i=1}^n (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6 \left(\sum_{i=1}^n (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2 \right)^{3/2}},$$

donde,

$\hat{\theta}_{(.)}$ = valor de $\hat{\theta}$ eliminando la i – ésima observación

$\hat{\theta}_{(i)}$ = promedio de $\hat{\theta}'_i$ s (sobre los n datos).

Entonces, con lo anterior ya podemos construir los intervalos de confianza BC_a .

```
library(boot) # cargamos la libreria
head(cd4,3) # top 3 registros
```

```
## baseline oneyear
## 1      2.12      2.47
## 2      4.35      4.61
## 3      3.39      5.26
```

Para este problema simularemos 5000 muestras Bootstrap y calcularemos el coeficiente de correlación.

```
n <- nrow(cd4)
corr_obs <- cor(cd4$baseline, cd4$oneyear) # correlación observ.
b <- c()
B <- 5000
set.seed(19970808) # fijamos la semilla
for(i in 1:B) {
  ind_sample <- sample(1:n, size=n, replace=TRUE)
  b[i] <- cor(cd4$baseline[ind_sample], cd4$oneyear[ind_sample])
}
```

Tenemos que la correlación observada en la muestra es igual a 0.7231654. Ahora, calculamos los cuantiles para el intervalo.

```
z0 <- qnorm(mean(b < corr_obs)) # calculamos z0
rjack <- c() # inicializamos theta
for(i in 1:n){ # estimamos los jackknife para theta
  rjack[i] <- cor(cd4$baseline[-i], cd4$oneyear[-i])
}
rm <- mean(rjack) # calculamos la media de los estimadores
a <- sum((rm-rjack)^3)/(6*(sum((rm-rjack)^2))^(1.5)) # calculamos a
aen2 <- 0.05 # definimos alpha
alf1 <- pnorm(z0 + (z0+qnorm(aen2))/(1-a*(z0+qnorm(aen2)))) # calculamos q'
alf2 <- pnorm(z0 + (z0+qnorm(1-aen2))/(1-a*(z0+qnorm(1-aen2))))
```

Por lo tanto, $\alpha_1=0.0510521$ y $\alpha_2=0.9511489$. Por lo que, ahora podemos calcular los cuantiles con la muestra bootstrap simulado.

```
ci_cd4 <- quantile(b,probs=c(alf1,alf2))
```

Por lo tanto, el intervalo BC_a para el coeficiente de correlación entre los conteos VIH-positivos antes y después de un año de tratamiento con un antiviral es

$$BC_a = (\hat{\theta}_{\alpha_1}^*, \hat{\theta}_{\alpha_2}^*) = (0,5520196, 0,8448081).$$

Nota: Otra respuesta sería validar el supuesto de normalidad de los datos para que el intervalo de confianza Bootstrap por percentiles tenga justificación sólida y no existan desventajas de usarlo.

b) Calcule el coeficiente estimado de correlación corregido por sesgo usando Jackknife.

RESPUESTA

El estimador Jackknife para el sesgo se define como

$$\hat{B}_j = (n-1) (\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

así, el estimador Jackknife para θ se define como

$$\tilde{\theta} = \hat{\theta} - \hat{B}_j = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}.$$

Esto implica que

$$\mathbb{E}(\tilde{\theta}) = n\mathbb{E}(\hat{\theta}) - (n-1)\mathbb{E}(\hat{\theta}_{(\cdot)}).$$

Entonces, con lo anterior podemos calcular el coeficiente estimado de correlación corregido por sesgo.

```
set.seed(19970808)
rjack <- c() # inicializamos los estimadores
r_jack_corregido <- c()
for(i in 1:n){ # estimamos los jackknife para theta
  rjack[i] <- cor(cd4$baseline[-i], cd4$oneyear[-i])
  r_jack_corregido[i] <- n*corr_obs - (n-1)*rjack[i]
}
rm <- mean(rjack) # calculamos la media de los estimadores

estimador_sesgo_jack <- (n-1)*(rm-corr_obs)
estimador_jack <- corr_obs - estimador_sesgo_jack

estimadro_jack_error_est <- sqrt((sum( (estimador_jack-mean(estimador_jack))^2)/(n*(n-1))))
```

Entonces, el estimador para el sesgo $\hat{B}_j = -0.0067843$. Y por lo tanto, el **coeficiente estimado de correlación corregido por sesgo es de 0.7299497**. ■.

12. Los siguientes 15 datos forman una muestra aleatoria de una distribución Gamma con parámetro de forma $\alpha = 3$ y parámetro de escala $\beta = 2$ (la media es $\alpha\beta$ y la varianza $\alpha\beta^2$)

14,18	10,99	3,38	6,76	5,56	1,26	4,05	4,61
1,78	3,84	4,69	2,12	2,39	16,75	4,19	

Encuentre un intervalo de confianza para la mediana de la distribución.

RESPUESTA

Recordemos que la mediana de una variable con distribución Gamma es de forma cerrada (es decir, no existe una igualdad). Entonces para construir un intervalo de confianza para la media consideramos la metodología descrita en el ejercicio 11, es decir, utilizaremos bootstrap para construir los intervalos BC_a .

Realizamos un pequeño cambio a la metodología, como conocemos la distribución exacta de la muestra, es decir, sabemos que la distribución de la m.a es de una Gamma(3, 2) podemos realizar bootstrap paramétrico para tener un mejor intervalo de confianza. Con estas consideraciones procedemos a construir el intervalo.

```
set.seed(19970808) # fijamos la semilla

df_gamma <- c(14.18, 10.99, 3.38, 6.76, 5.56, 1.26, 4.05, 4.61, 1.78, 3.84, 4.69, 2.12,
             2.39, 16.75, 4.19) # m.a.

median_obs <- median(df_gamma) # mediana observada
n <- length(df_gamma) # tamaño

B <- 5000 # número de repeticiones
b <- c()
for(i in 1:B){
  m_a_gamma <- rgamma(n, 3, 1/2) # generamos la muestra aleatoria
  b[i]<-median(m_a_gamma) # calculo de la mediana.
}
```

Ya generamos la 5000 muestras considerando la distribución Gamma(3,2), la media de la m.a. original es 4.19. Ahora calculamos los cuantiles para el intervalo.

```
z0 <- qnorm(mean(b < median_obs)) # calculamos z0
rjack <- c() # inicializamos theta
for(i in 1:n){ # estimamos los jackknife para theta
  rjack[i] <- median(df_gamma[-i])
}
rm <- mean(rjack) # calculamos la media de los estimadores
a <- sum((rm-rjack)^3)/(6*(sum((rm-rjack)^2))^(1.5)) # calculamos a
aen2 <- 0.05 # definimos alpha
alf1 <- pnorm(z0 + (z0+qnorm(aen2))/(1-a*(z0+qnorm(aen2)))) # calculamos q'
alf2 <- pnorm(z0 + (z0+qnorm(1-aen2))/(1-a*(z0+qnorm(1-aen2))))
```

Por lo tanto, $\alpha_1=2,1129675 \times 10^{-5}$ y $\alpha_2=0.2011266$. Por lo que, ahora podemos calcular los cuantiles con la muestra bootstrap simulado.

```
ci_cd4 <- quantile(b, probs=c(alf1,alf2))
```

Por lo tanto, el intervalo BC_a para la media de la distribución es

$$BC_a = (\hat{\theta}_{\alpha_1}^*, \hat{\theta}_{\alpha_2}^*) = (2,2976991, 4,5319194). \blacksquare$$

Ejercicios de las Notas:

1.1 (Diapositiva 36). Encuentre en el anterior ejemplo los valores $z_{\alpha/2}$ utilizados.

RESPUESTA

Tenemos que la siguiente igualdad del error de estimación

$$\text{Error de estimación} \equiv E \equiv z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

Y como en el ejemplo anterior (de las notas) tenemos que $\sigma^2 = 1$ y $E = 1$. Utilizando R calculamos los cantiles solicitados,

```
qnorm(0.05)
```

```
## [1] -1.644854
```

```
qnorm(0.025)
```

```
## [1] -1.959964
```

es decir, para $1 - \alpha = 0,9$ tenemos que $z_{10/2} = -1,644854$ y para $1 - \alpha = 0,95$ tenemos que $z_{0,05/2} = -1,959964$.

1.2 Si se quisiera tener un IC del 100 % de confiabilidad, ¿qué pasa con $z_{\alpha/2}$? ¿Tendría este intervalo algún interés práctico?

RESPUESTA

Tener un intervalo de confianza del 100 % eso implicaría que $z_{\alpha/2} \rightarrow \infty$ por lo que el intervalo sería toda la recta real, por lo que no nos diría absolutamente nada de la variabilidad del parámetro de interés. Es decir, si simulamos muchas veces el experimento a estudiar y calculamos el parámetro de interés entonces en 100 % de los experimentos este parámetro estará en la recta real, en términos prácticos no aporta información adicional del parámetro de interés. ■.

2. (Diapositiva 58) En este último intervalo deberíamos notar que la relación entre la media de la variable original y la de su transformación logarítmica está afectada por una constante que aquí no aparece. Recuerda la relación entre medias y varianzas que hemos establecido al final del capítulo 3, para la Normal y Lognormal. Trata de incorporar esta información para entender mejor la diferencia numérica entre estos intervalos.

RESPUESTA

Recordemos la relación que existe entre las medias y varianzas para la Normal y Lognormal.

Teorema: 4 Sea $X \sim N(\mu, \sigma^2)$ y $Y \equiv e^X \sim \text{Lognormal}(\mu, \sigma^2)$ entonces

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = M_x(1) = e^{\mu + \frac{\sigma^2}{2}}$$

y además

$$V(Y) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)$$

En el problema nos dicen que la distribución de la población es normal, y por ende la distribución de los valores de calcio sería lognormal. Además nos dicen que $\bar{x} = 0,6976$ y $s^2 = 0,3595$. Entonces una forma de encontrar el IC para μ , la media de los valores de oxalato de calcio es utilizando la relación de la media y varianza de la Normal y Lognormal (4), es decir, tenemos el IC del 95 % sería

$$\begin{aligned} e^{\hat{x} + \hat{\sigma}^2/2} \pm z_{\alpha/2} e^{2\hat{x} + \hat{\sigma}^2} (e^{\hat{\sigma}^2 - 1}) &= (e^{0,6976 + 0,0552^2/2} \pm z_{0,05/2} e^{2 \cdot 0,6976 + 0,0552^2} (e^{0,0552^2} - 1)) \\ &= 2,011988 \pm 0,02421291 \\ &= (1,987775, 2,036201) \end{aligned}$$

Si comparamos el intervalo de confianza calculado en las notas considerando el pivote considerando una t - student el cual fue (1,802, 2,239), observamos que claramente que este intervalo es más grande que el nosotros calculamos considerando la distribución Lognormal. Es decir, podemos concluir que considerando la relación entre la distribución Normal y Lognormal obtenemos un mejor intervalo de confianza con la misma confiabilidad de 95 %. ■.

3. (Diapositiva 110) Muestra las 3 propiedades mencionadas arriba para S_p^2 , donde

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

a) Mostrar que $\mathbb{E}(S_p^2) = \sigma^2$.

RESPUESTA

Recordemos S_1^2 y S_2^2 son estimadores insesgados de σ^2 , es decir, $\mathbb{E}(S_1^2) = \mathbb{E}(S_2^2) = \sigma^2$. Entonces tenemos que

$$\begin{aligned}\mathbb{E}(S_p^2) &= \mathbb{E}\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}\right) \\ &= \frac{(n_1 - 1)\mathbb{E}(S_1^2) + (n_2 - 1)\mathbb{E}(S_2^2)}{(n_1 + n_2 - 2)} \\ &= \frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{(n_1 + n_2 - 2)} \\ &= \frac{(n_1 + n_2 - 2)\sigma^2}{(n_1 + n_2 - 2)} \\ &= \sigma^2.\end{aligned}$$

Lo anterior, demuestra que $\mathbb{E}(S_p^2) = \sigma^2$ y por lo tanto podemos decir que S_p^2 es un estimador insesgado para σ^2 .

b) Mostrar que $\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi_{n_1 + n_2 - 1}^2$.

RESPUESTA

Utilizando la definición S_p^2 , tenemos

$$\begin{aligned}\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} &= \frac{(n_1 + n_2 - 2)}{\sigma^2} \left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} \right) \\ &= \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}.\end{aligned}$$

Ahora, como en el problema se trabajo con dos m.a. con distribución normal independientes entre sí. Entonces nosotros sabemos que (demostrado en clase pag. 96)

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi_{n_1 - 1}^2, \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_2 - 1}^2.$$

Entonces como también sabemos que suma de χ^2 's independientes es igual a otra χ^2 con grados de libertad la suma de los χ^2 's sumadas (demostrado en la tarea 6, ejercicio 2 de las notas), es decir, es cerrada bajo la suma. Entonces ocupando estos resultado, podemos concluir que

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_1 - 1}^2 + \chi_{n_2 - 1}^2 = \chi_{n_1 + n_2 - 2}^2.$$

c) Mostrar que S_p^2 es de mínima varianza.

RESPUESTA

Para demostrar esta propiedad, recordemos la desigualdad de Cramér-Rao y una propiedad de la matriz de información.

Definición: 3 (Diapositiva 140) Sea $\hat{\theta}$ un estimador insesgado para θ entonces:

$$V(\hat{\theta}) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x)\right)^2\right]},$$

donde $f(x)$ es la función de probabilidad o densidad de la población y n el tamaño de la muestra. Además, si se encuentra por algún método un estimador insesgado cuya varianza sea igual al valor éste será el EIMV, es decir, el estimador de mínima varianza.

Teorema: 5 (Diapositiva 141)

$$n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x)\right)^2\right] = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x)\right].$$

Como sabemos que

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \chi_{n_1+n_2-2}^2$$

entonces esto implica que

$$V\left(\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}\right) = 2(n_1 + n_2 - 2).$$

Entonces podemos calcular la varianza de S_p^2 como

$$\begin{aligned} V\left(\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}\right) &= 2(n_1 + n_2 - 2) \\ \frac{(n_1 + n_2 - 2)^2 V(S_p^2)}{\sigma^4} &= 2(n_1 + n_2 - 2) \\ V(S_p^2) &= \frac{2(n_1 + n_2 - 2)\sigma^4}{(n_1 + n_2 - 2)^2} \\ V(S_p^2) &= \frac{2\sigma^4}{(n_1 + n_2 - 2)}. \end{aligned}$$

Ahora calculemos la cota de Cramér-Rao (definición 3), para ello ocupemos el teorema (5).

$$\begin{aligned}
n\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f(x) \right)^2 \right] &= -n\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f(x) \right] \\
&= -n\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \frac{(x - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right) \right] \\
&= -n\mathbb{E} \left[\left(\frac{-(x - \mu)^2}{2\sigma^6} + \frac{1}{2\sigma^4} \right) \right] \\
&= -n \left[\frac{-2\mathbb{E}[(x - \mu)^2] + \sigma^2}{2\sigma^6} \right] \\
&= -n \frac{-2\sigma^2 + \sigma^2}{2\sigma^6} \\
&= -n \frac{-1}{2\sigma^4} \\
&= \frac{n}{2\sigma^4}.
\end{aligned}$$

Entonces, la cota de Cramér-Rao es $\frac{2\sigma^4}{n}$ donde $n = n_1 + n_2$, es decir, para este caso la cota es $\frac{2\sigma^4}{n_1 + n_2}$. Entonces si comparamos la varianza de S_p^2 (ya que es insesgado, por el inciso a)) calculada podemos observar que es igual a la cota de Cramér-Rao solo que con la única diferencia que el denominador pierde 2 grados de libertad al estimar S_1^2 y S_2^2 . Entonces podemos concluir que S_p^2 es de mínima varianza. ■.

4. (Diapositiva 164) Calcula las probabilidad mencionadas arriba.

RESPUESTA

El problema nos dice que tenemos una población normal. La hipótesis nula nos dice que la media es 80. Entonces la probabilidad de que aparezca un valor 80.3 u otro mayor considerando para el caso 1 cuando la varianza es 0.1 (se uso R : `pnorm(80.3, 80, sqrt(.1/10))`)

$$\mathbb{P}(X > 80,3) = 1 - \mathbb{P}(X < 80,3) = 1 - 0,829 = \mathbf{0,0013}.$$

Y para el caso 2, cuando la varianza es 0.29 (se uso R : `pnorm(80.3, 80, sqrt(.29/10))`)

$$\mathbb{P}(X > 80,3) = 1 - \mathbb{P}(X < 80,3) = 1 - 0,0711 = \mathbf{0,039}.$$

Entonces, en términos del problema para el caso 1 estas probabilidades nos dan una noción de rechazar la hipótesis nula. Pero si consideramos el caso 2, no existe evidencia suficiente para pensar que rechazar la hipótesis nula. ■.

5. (Diapositiva 230) Construir esas gráficas para este caso particular.

RESPUESTA

Nos pide graficar la potencia de una prueba para la media de una población normal con varianza desconocida de una cola para dos niveles de significancia, $\alpha = 0,05, 0,1$ y considerando el hecho que desconocemos el valor de la media bajo la hipótesis alternativa. En el mismo problema nos dicen que la potencia de la prueba para los α 's mencionados es

$$\text{Potencia de la prueba } (\alpha = 0,10) = \mathbb{P} \left(T < -t_{0,10,19} + \frac{\sqrt{20}(35 - \mu_A)}{5,554} \right),$$

$$\text{Potencia de la prueba } (\alpha = 0,05) = \mathbb{P} \left(T < -t_{0,05,19} + \frac{\sqrt{20}(35 - \mu_A)}{5,554} \right),$$

Con ayuda de *R* procedemos a calcular la potencia de la prueba para distintos valores de la media bajo la hipótesis alternativa.

```
n <- 20 # Tamaño de la muestra
mu_0 <- 35 # media de la nula.

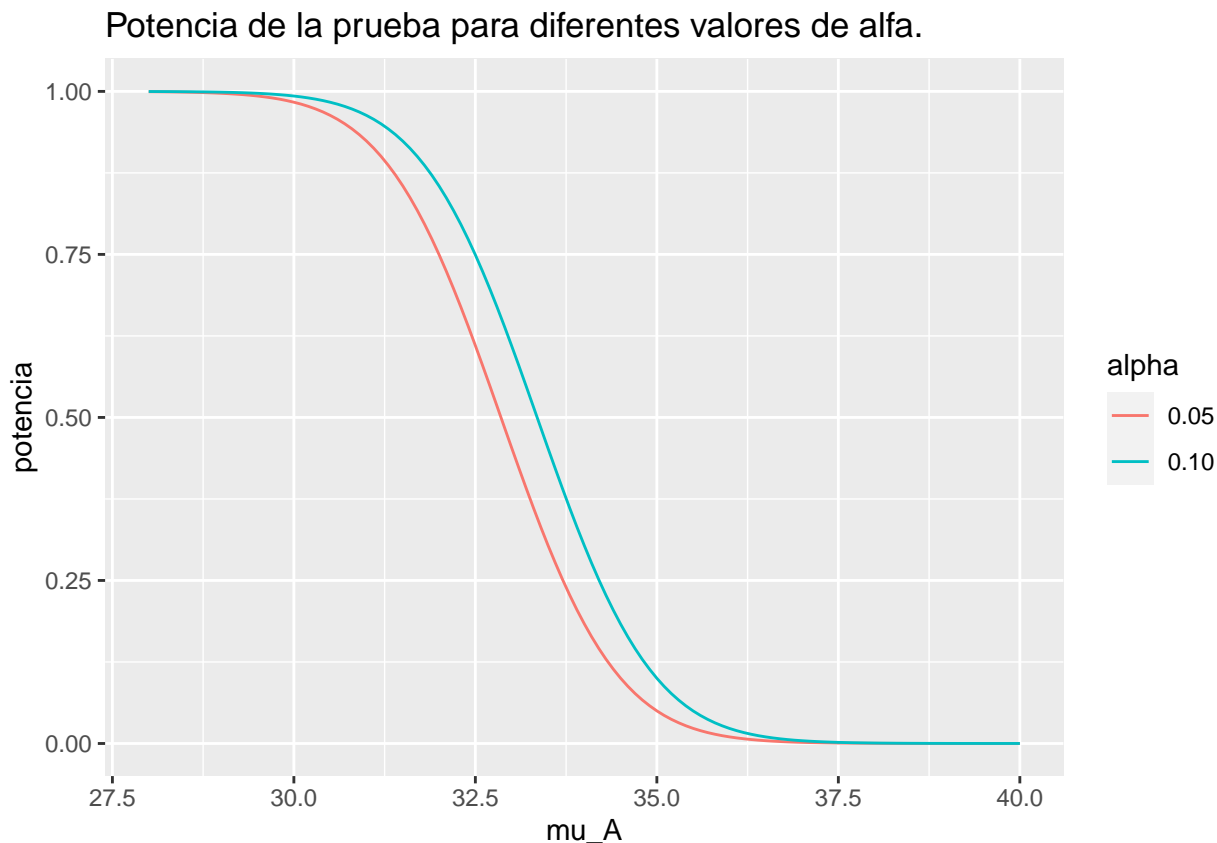
t_10 <- qt(0.90,19) # cuantiles de una dist. t
t_05 <- qt(0.95,19)

mu_A <- seq(28,40,0.1) # media alternativa

potencia_alpha_10 <- pt(-t_10+sqrt(n)*(35-mu_A)/5.554, 19) # potencia
potencia_alpha_05 <- pt(-t_05+sqrt(n)*(35-mu_A)/5.554, 19)

# dataframe aus para grafica.
potencia_T <- data.frame(mu_A= rep(mu_A,2), potencia = c(potencia_alpha_05, potencia_alpha_10),
                        alpha=c(rep("0.05", length(mu_A)), rep("0.10", length(mu_A))))

ggplot(data=potencia_T, aes(x=mu_A, y=potencia, group=alpha, col=alpha))+ # grafica de la potencia
  geom_line()+
  labs(title="Potencia de la prueba para diferentes valores de alfa.")
```



Estas gráficas nos ayudan a entender el efecto de fija un valor específico para α . Ya que implica cambios a la hora de rechazar o no la hipótesis nula. ■.

- (Diapositiva 235) Realiza una prueba de nivel $\alpha = 0,01$ para las siguientes hipótesis: $H_0 : \sigma^2 = 8$ vs $H_A : \sigma^2 < 8$ a partir del archivo datvar.txt.

- Verifica inicialmente normalidad y concluye sobre la plausibilidad de la misma en este conjunto de datos.
- Utiliza el estadístico discutido en este capítulo y el que se presenta además en el capítulo anterior cuando la población no es normal.
- Calcula la potencia de la prueba para valores del cociente 1.5, 2, 2.5, 3.

RESPUESTA

Para hacer inferencia sobre la varianza de la población consideremos la siguiente metodología. Si suponemos normalidad en el proceso, tenemos el estadístico de prueba

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

con región de rechazo

$$\mathbb{P}(U < a) = \alpha$$

donde en la practica se usa $a = \chi_{1-\alpha, n-1}^2$, es decir, rechazar si $u < \chi_{1-\alpha, n-1}^2$. Todo lo anterior es considerando las hipótesis

$$H_0 = \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_A : \sigma^2 < \sigma_0^2.$$

Entonces considerando todo lo anterior, primero comprobemos el supuesto de normalidad para ello grafiquemos los Q–Q plot, un histograma y además utilizaremos la prueba de normalidad de Shapiro Wilk.

```
dat_var <- read.table("datvar.txt") # cargamos los datos
dat_var$V1 <- as.numeric(gsub( ",", ".", dat_var$V1)) # transformamos tipo dato
head(dat_var,3) # top 3
```

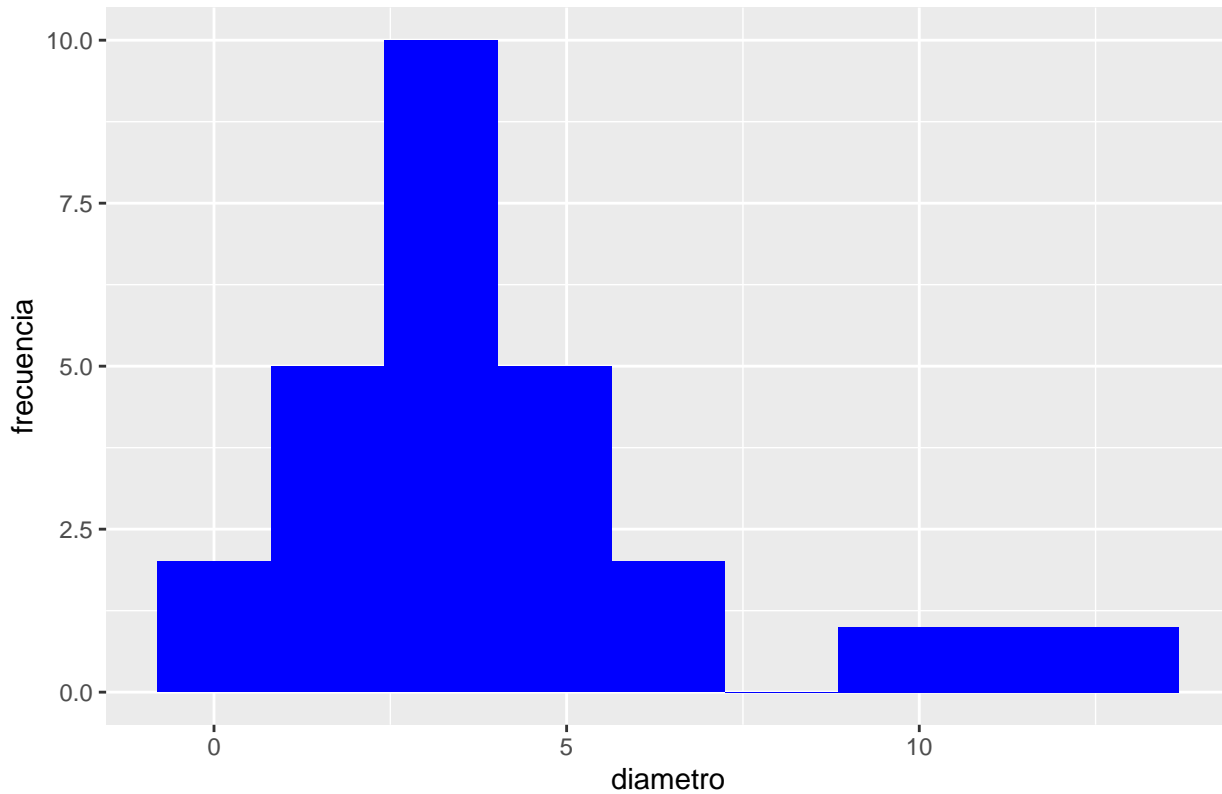
```
##          V1
## 1 3.703027
## 2 3.403226
## 3 1.410564
```

Ahora graficamos el histograma y el Q–Q plot.

```
library(fBasics) # qqplot

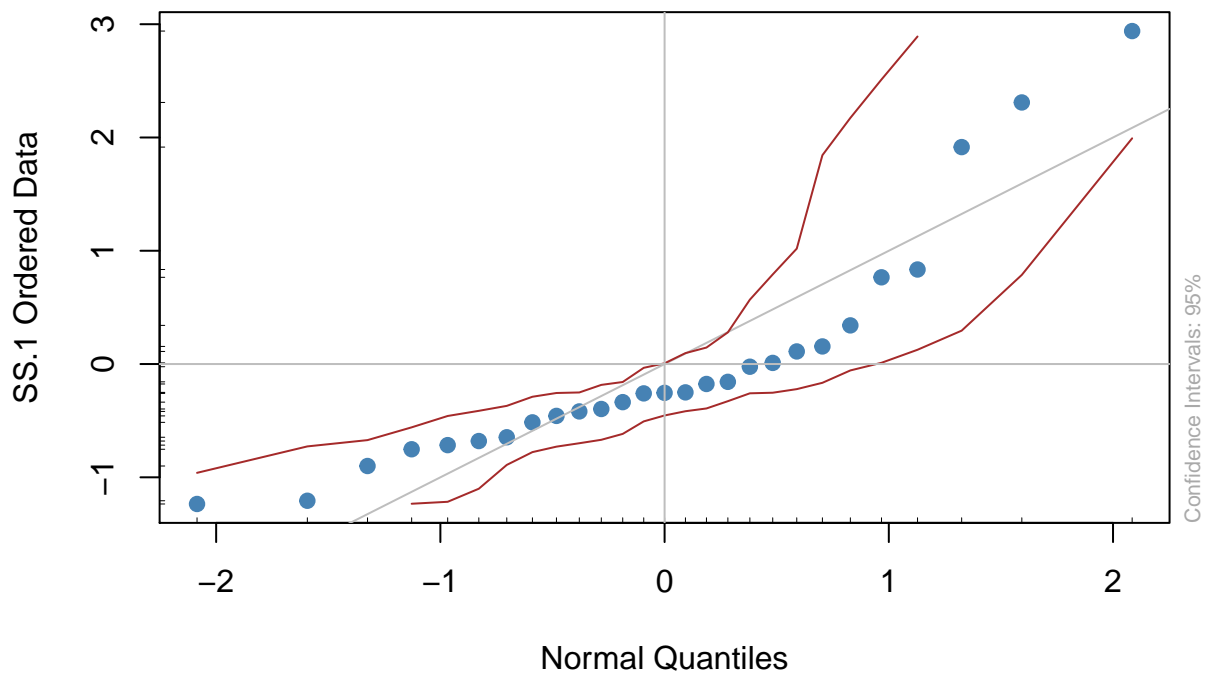
ggplot(dat_var, aes(x=V1)) + # histograma
  geom_histogram(bins=9, fill="blue")+
  labs(title="Histograma del los datos de varianza", y="frecuencia", x="diametro")
```

Histograma del los datos de varianza



```
qqnormPlot(dat_var$V1) # qqplot
```

NORM QQ PLOT



Observamos que existen irregularidades en los extremos del histograma y en el Q–Q plot, por lo que puede deberse a outliers en los datos, pero nos da indicios que los datos. Para reforzar esta conclusión utilicemos la prueba de Shapiro Wilk.

```
shapiro.test(dat_var$V1)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: dat_var$V1  
## W = 0.8285, p-value = 0.0004449
```

Entonces, **como el estadístico de prueba W está cercano a uno podemos concluir que efectivamente los datos provienen de una distribución normal.**

Como ya verificamos el supuesto de normalidad, procedemos a plantear nuestro juego de hipótesis. Nuestro juego de hipótesis son

$$H_0 = \sigma^2 = 8 \quad \text{vs} \quad H_A : \sigma^2 < 8.$$

Entonces, **calculemos el estadístico de prueba**

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

y los cuantiles adecuados (con $\alpha = 0,05$).

```
sigma_ejer_6 <- 8  
n <- nrow(dat_var)  
u <- (n-1)*var(dat_var$V1)/sigma_ejer_6  
  
chi_inf <- qchisq(p=0.95,df=n-1)
```

Entonces como el estadístico de prueba es $U = 30.9362$, y tenemos que la decisión de rechazo para esta prueba es rechazar H_0 si $U < \chi_{1-\alpha, n-1}^2$. Tenemos que $\chi_{1-\alpha, n-1}^2 = 38.8851387$, por lo tanto como $U < \chi_{1-\alpha, n-1}^2$ rechazamos la hipótesis nula. Y por lo tanto, **podemos concluir que hay evidencia significativa para decir que la varianza del proceso se ha alterado.** ■.

Por último calculemos la potencia de la prueba para valores del cociente, la cual para este juego de hipótesis es

$$\text{Potencia de la prueba} = \mathbb{P} \left(U < \left(\frac{\sigma_0^2}{\sigma_A^2} \right) \chi_{1-\alpha, n-1}^2 \right).$$

Entonces ocupamos, procedemos a calcular la potencia con R:

```
cocien_valore <- c(1, 1.5, 2, 2.5, 3)  
  
pchisq(cocien_valore*chi_inf,df=n-1)
```

```
## [1] 0.9500000 0.9997198 0.9999995 1.0000000 1.0000000
```

Es decir, la potencia es muy buena para rechazar H_0 cuando la varianza es igual, el doble, el triple. ■.

7. (Diapositiva 251) La desviación estándar del diámetro interno de tuercas producidas por un proceso en una fábrica es de 0.0004cm. Un día, como parte de acciones correctivas producto de un diagnóstico en los componentes del proceso, son reemplazadas ciertas partes mecánicas y se espera que la variación del proceso cambie, pero no se sabe en qué dirección. Para verificar algún cambio se toma

una muestra aleatoria:

0,501081	0,499666	0,499195	0,500069	0,499803
0,500921	0,499513	0,499984	0,500451	0,500472
0,499908	0,500196	0,499234	0,500193	0,501199
0,499327	0,500036	0,500392	0,499124	0,501195
0,500131	0,499604	0,500323	0,500531	0,500599
0,499044	0,499379	0,499695	0,499600	0,499300
0,500165	0,499721	0,500026	0,498187	0,500645
0,500345	0,499962	0,500556	0,499923	0,499287
0,500179	0,499772	0,499380	0,499102	0,500477
0,499820	0,499415	0,500007	0,500609	0,500209

Realiza un Q-Q Plot y verifica normalidad. Plantea las hipótesis que correspondan, haz la prueba y concluye: ¿Se ha alterado la varianza del proceso? (Los datos se encuentran en el archivo *diam.txt*).

RESPUESTA

El objetivo de este problema es verificar si el reemplazo de ciertas piezas mecánicas implica un cambio en la variación del proceso, este cambio no se sabe si es un cambio para disminuir o ganar variación en el proceso. Para validar lo anterior consideremos la siguiente metodología, si suponemos normalidad en el proceso, tenemos el estadístico de prueba

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

con región de rechazo

$$\mathbb{P}(U < a) + \mathbb{P}(U > b) = \alpha$$

donde en la práctica se usa $a = \chi_{1-\alpha/2, n-1}^2$ y $b = \chi_{\alpha/2, n-1}^2$, es decir, rechazar si $u < \chi_{1-\alpha/2, n-1}^2$ o $u > \chi_{\alpha/2, n-1}^2$. Todo lo anterior es considerando las hipótesis

$$H_0 = \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_A : \sigma^2 \neq \sigma_0^2.$$

Y para esta prueba de hipótesis el p-value se define como (aunque no es universalmente aceptada)

$$= 2 \min \left\{ \mathbb{P} \left(\chi_{n-1}^2 < U_{\text{observado}} \right), \mathbb{P} \left(\chi_{n-1}^2 > U_{\text{observado}} \right) \right\}$$

. Entonces primero comprobemos el supuesto de normalidad para ello grafiquemos los Q-Q plot, un histograma y además utilizaremos la prueba de normalidad de Shapiro Wilk.

Primero cargamos los datos.

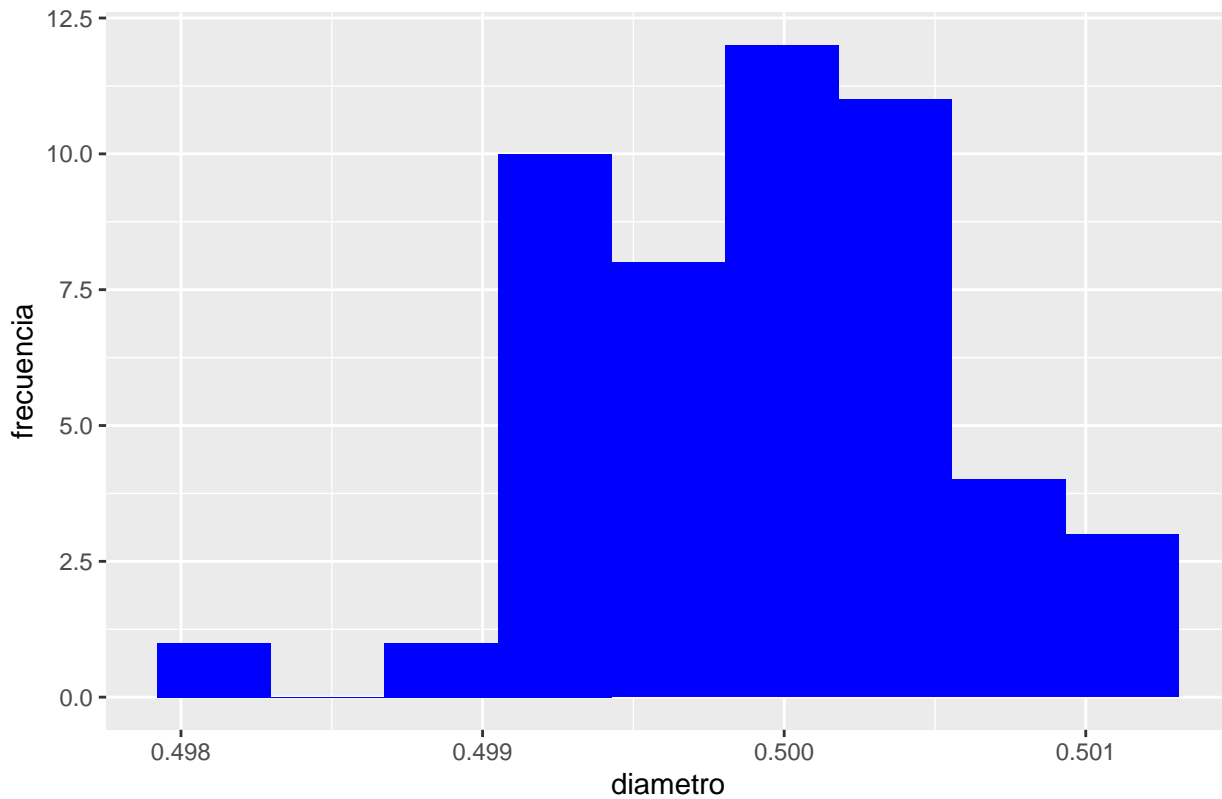
```
diam_tuercas <- read.table(file = "diam.txt", sep = ",", header = T) # load file
head(diam_tuercas, 3) # top 3.
```

```
##      c000001  c000002
## 1          1 0.501081
## 2          2 0.500921
## 3          3 0.499908
```

Ahora graficamos el histograma y el Q-Q plot.

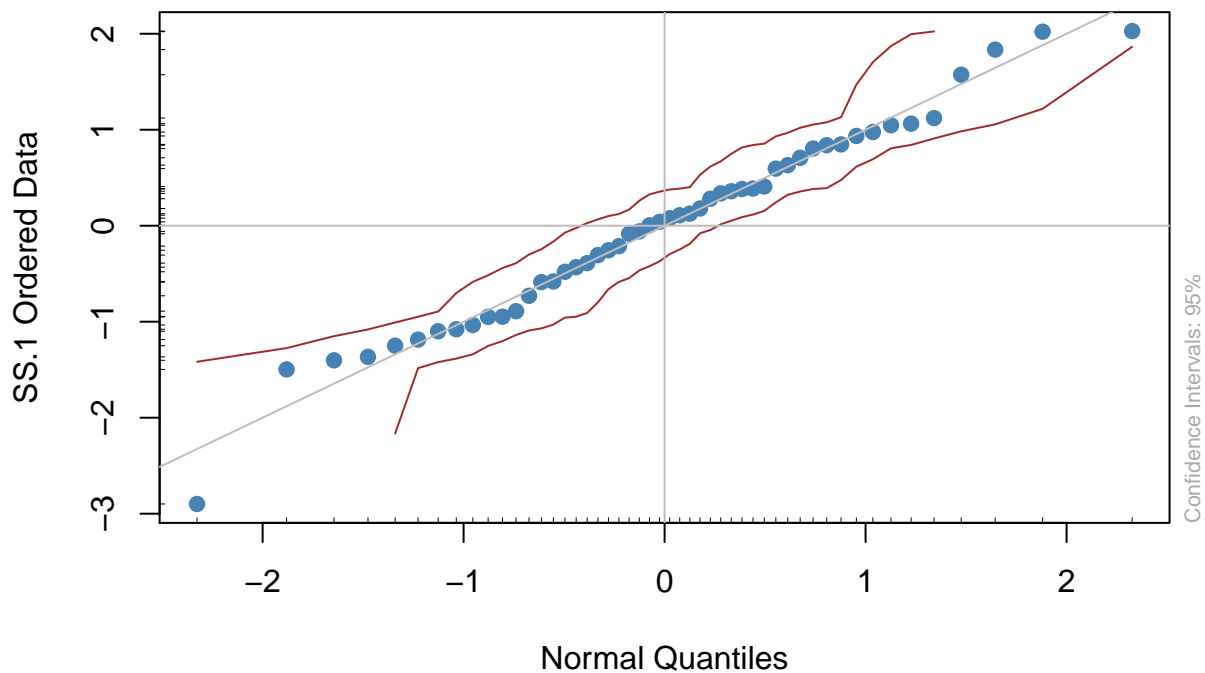
```
ggplot(diam_tuercas, aes(x=c000002)) + # histograma
  geom_histogram(bins=9, fill="blue")+
  labs(title="Histograma del diametro de las tuercas", y="frecuencia", x="diametro")
```

Histograma del diametro de las tuercas



```
qqnormPlot(diam_tuercas$c000002) # qqplot
```

NORM QQ PLOT



Analizando los gráficos podemos observar que los datos provienen de una distribución normal. Para reforzar esta conclusión utilizemos la prueba de Shapiro Wilk.

```
shapiro.test(diam_tuercas$c000002)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: diam_tuercas$c000002  
## W = 0.98259, p-value = 0.6651
```

Entonces, como el estadístico de prueba W está cercano a uno podemos concluir que efectivamente los datos provienen de una distribución normal.

Como ya verificamos el supuesto de normalidad, procedemos a plantear nuestro juego de hipótesis. En este caso sabemos que antes del cambio la desviación estándar era igual a 0.0004 cm, entonces nuestro juego de hipótesis son

$$H_0 = \sigma^2 = 0,0004^2 \quad \text{vs} \quad H_A : \sigma^2 \neq 0,0004^2.$$

Entonces calculemos el estadístico de prueba y los cuantiles adecuados (con $\alpha = 0,05$).

```
sigma_20 <- 0.0004**2  
n <- nrow(diam_tuercas)  
u <- (n-1)*var(diam_tuercas$c000002)/sigma_20  
  
chi_inf <- qchisq(p=0.975,df=n-1)  
chi_sup <- qchisq(p=0.025, df=n-1)
```

Entonces como el estadístico de prueba es $U = 114.4683286$, y tenemos que la decisión de rechazo para esta prueba es rechazar H_0 si $U < \chi^2_{1-\alpha/2, n-1}$ o $U > \chi^2_{\alpha/2, n-1}$. Tenemos que $\chi^2_{1-\alpha/2, n-1} = 70.2224136$ y $\chi^2_{\alpha/2, n-1} = 31.5549165$, por lo tanto como $U > \chi^2_{\alpha/2, n-1}$ rechazamos la hipótesis nula. Y por lo tanto, **podemos concluir que hay evidencia significativa para decir que la varianza del proceso se ha alterado.** ■.

8. (Diapositiva 258) Ya habíamos comentado que cuando se considera la potencia de una prueba de dos colas una de las dos probabilidades que es necesario calcular se vuelve despreciable conforme $\mu_A - \mu_0$ se hace grande o se hace muy pequeña, dependiendo de cuál de las dos direcciones de la Alternativa se estén considerando.

8.1 Calcula la potencia de una prueba bajo normalidad para las hipótesis $H_0 : \mu = \mu_0$ y $H_A : \mu \neq \mu_A$ en valores de $\sqrt{n}\Psi = -3, -2,5, -2, -1,5, -1, -0,5, 0,5, 1, 1,5, 2, 2,5, 3$. Hazlo para valores de $\alpha = 0,05, 0,01$. A partir de la tabla que generes, haz la gráfica correspondiente, (en el eje x: $\sqrt{n}\Psi$, en el eje y: los valores de la potencia, para cada nivel α considerado).

RESPUESTA

La potencia para nuestro juego de hipótesis se define como

$$\text{Potencia de la prueba} = \mathbb{P}\left(Z < -z_{\alpha/2} + \sqrt{n}\Phi\right) + \mathbb{P}\left(Z > z_{\alpha/2} + \sqrt{n}\Phi\right).$$

Entonces con lo anterior, calculemos la potencia con los distintos valores de $\sqrt{n}\Psi$ y α .

```
z_alpha_10 <- qnorm(0.99) # cuantiles alpha  
z_alpha_05 <- qnorm(0.95)  
  
sqr_n_phi <- c(-3, -2.5, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 2.5, 3) # valores sqrt_n_phi  
potencia_alpha_05 <- pnorm(-z_alpha_05+sqr_n_phi)+1-pnorm(z_alpha_05+sqr_n_phi) # potencias
```

```

potencia_alpha_10 <- pnorm(-z_alpha_10+sqr_n_phi)+1-pnorm(z_alpha_10+sqr_n_phi)

potencias_prueba <- data.frame(sqrt_n_phi = rep(sqr_n_phi,2),
                                alphas = c(rep("0.05", length(sqr_n_phi)), rep("0.10", length(sqr_n_phi) )),
                                potencia = c(potencia_alpha_05, potencia_alpha_10)) # data.frame ggplot

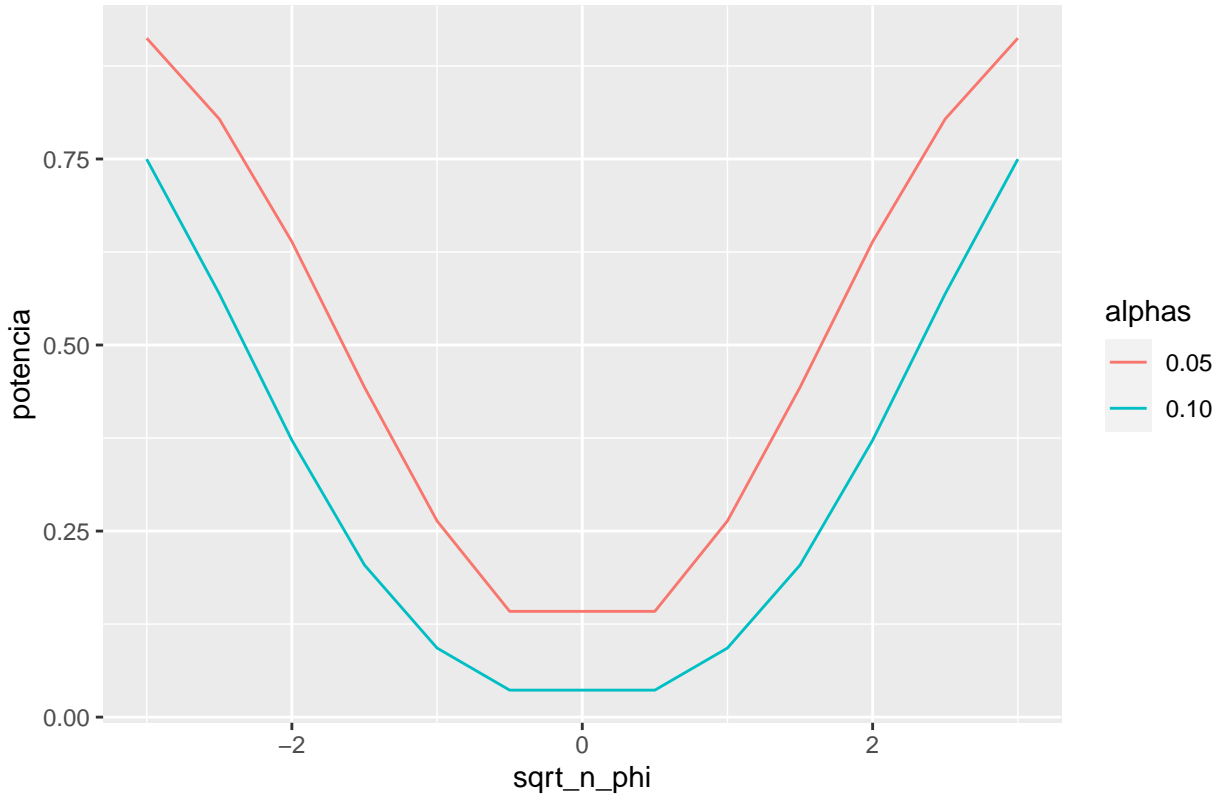
print(potencias_prueba) #imprimimos las potencias

##      sqrt_n_phi alphas  potencia
## 1         -3.0   0.05 0.91231624
## 2         -2.5   0.05 0.80378194
## 3         -2.0   0.05 0.63889380
## 4         -1.5   0.05 0.44324407
## 5         -1.0   0.05 0.26359734
## 6         -0.5   0.05 0.14211717
## 7          0.5   0.05 0.14211717
## 8          1.0   0.05 0.26359734
## 9          1.5   0.05 0.44324407
## 10         2.0   0.05 0.63889380
## 11         2.5   0.05 0.80378194
## 12         3.0   0.05 0.91231624
## 13        -3.0   0.10 0.74973380
## 14        -2.5   0.10 0.56893126
## 15        -2.0   0.10 0.37208817
## 16        -1.5   0.10 0.20436842
## 17        -1.0   0.10 0.09280221
## 18        -0.5   0.10 0.03625304
## 19         0.5   0.10 0.03625304
## 20         1.0   0.10 0.09280221
## 21         1.5   0.10 0.20436842
## 22         2.0   0.10 0.37208817
## 23         2.5   0.10 0.56893126
## 24         3.0   0.10 0.74973380

ggplot(data=potencias_prueba, aes(x=sqrt_n_phi, y=potencia, group=alphas, col=alphas))+
  geom_line() + # Grafica potencia
  labs(title="Potencia de la prueba con distintos valores de phi y alpha.")

```


Potencia de la prueba con distintos valores de phi y alpha.



8.2 Verifica a partir de que valores de $\sqrt{n}\Psi$ una de las dos probabilidades es posible no tomarla en cuenta sin perder precisión. ¿Qué valores de las potencias calculadas desearías usar en una aplicación?

RESPUESTA

Al potencias calculados les concatenamos las dos probabilidades calculadas, donde probabilidad_1 sera $\mathbb{P}(Z < -z_{\alpha/2} + \sqrt{n}\Phi)$ y probabilidad_2 sera $\mathbb{P}(Z > z_{\alpha/2} + \sqrt{n}\Phi)$.

concatenamos las probabilidades.

```
potencias_prueba$probabilidad_1 <-c(pnorm(-z_alpha_05+sqrt_n_phi), pnorm(-z_alpha_10+sqrt_n_phi))
potencias_prueba$probabilidad_2 <-c(1-pnorm(z_alpha_05+sqrt_n_phi), 1-pnorm(z_alpha_10+sqrt_n_phi))
```

```
print(head(potencias_prueba)) # imprimimos los valores grandes y pequeños
```

```
##  sqrt_n_phi alphas  potencia probabilidad_1 probabilidad_2
## 1      -3.0    0.05 0.9123162  1.701588e-06      0.9123145
## 2      -2.5    0.05 0.8037819  1.700154e-05      0.8037649
## 3      -2.0    0.05 0.6388938  1.337720e-04      0.6387600
## 4      -1.5    0.05 0.4432441  8.308497e-04      0.4424132
## 5      -1.0    0.05 0.2635973  4.086313e-03      0.2595110
## 6      -0.5    0.05 0.1421172  1.598228e-02      0.1261349
```

```
print(tail(potencias_prueba))
```

```
##  sqrt_n_phi alphas  potencia probabilidad_1 probabilidad_2
## 19         0.5    0.10 0.03625304  0.03389894  2.354105e-03
## 20         1.0    0.10 0.09280221  0.09236225  4.399602e-04
## 21         1.5    0.10 0.20436842  0.20430339  6.502923e-05
## 22         2.0    0.10 0.37208817  0.37208059  7.580097e-06
```

## 23	2.5	0.10	0.56893126	0.56893057	6.952976e-07
## 24	3.0	0.10	0.74973380	0.74973375	5.010357e-08

Observando las probabilidades observamos que cuando $\sqrt{n}\Phi$ es pequeña la $\mathbb{P}(Z < -z_{\alpha/2} + \sqrt{n}\Phi)$ se vuelve despreciable y cuando $\sqrt{n}\Phi$ es muy grande $\mathbb{P}(Z > z_{\alpha/2} + \sqrt{n}\Phi)$ se hace despreciable. Dependiendo del problema, si es más “costoso” (peligroso, etc) rechazar la hipótesis nula cuando la hipótesis nula es cierta entonces yo escogería una potencia muy alta en este caso mayor a 0.80 sería aceptable.

8.3 Nota la relación entre el tamaño de muestra para una prueba de medias de dos colas con respecto a la potencia que se desearía en una aplicación, cuál usarías?

RESPUESTA

Entonces considerando la potencia adecuada del inciso anterior, observamos que el tamaño de muestras para una prueba de medias de dos colas la que usaría sería $|n\Phi| = 2,5$.

8.4 ¿Qué criterios debes usar para seleccionar el tamaño de la muestra en aplicaciones concretas?

RESPUESTA

En la aplicación el tamaño de muestra depende de que quieres contractar, observemos que el tamaño de la muestra se hace grande para valores del desvío muy pequeños. Recordemos que existen circunstancias en que aumentar el tamaño de la muestra puede resultar caro. No tiene sentido incrementar el tamaño de muestra para la detección de valores de $\mu_A - \mu_0$ muy pequeños que en realidad no interesen; pero si desviaciones pequeñas son importantísimas, es necesario examinar el tamaño de muestra que se requiera para detectarlas. En este sentido, de nuevo, la consideración de la situación será lo que dicte cuál es el desvío mínimo que se quiere detectar con la prueba. ■.

9. (Diapositiva 274) Sea $X \sim \text{Binomial}(n, p)$. Deseamos contrastar las hipótesis $H_0 : p \leq p_0$ contra $H_1 : p > p_0$. Calcule el estadístico $\lambda(X)$ de la prueba de cociente de verosimilitudes. Justifique $\lambda(X) < c$ si $x > c'$, para constantes c, c' . ¿Cómo podemos elegir c para un nivel de significancia α ? ¿Existe c' para cualquier α ?

RESPUESTA

Tenemos que la función de verosimilitud para m.a. con distribución Binomial(n, p) es

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Definamos $\Theta_0 = \{p_0\}$ y $\Theta = \{p\}$ Si $\theta = p$, tenemos que

$$\lambda(X) = \frac{\sup_{\theta \in \Theta_0} f_{\theta}(x_1, \dots, x_n)}{\sup_{\theta \in \Theta} f_{\theta}(x_1, \dots, x_n)} = \frac{\binom{n}{x} p_0^x (1-p_0)^{n-x}}{\binom{n}{x} \hat{p}^x (1-\hat{p})^{n-x}} = \frac{p_0^x (1-p_0)^{n-x}}{\bar{x}^x (1-\bar{x})^{n-x}} = \left(\frac{p_0}{\bar{x}}\right)^x \left(\frac{1-p_0}{1-\bar{x}}\right)^{n-x}.$$

La justificación de lo anterior, es debido a que sabemos que el estimador de máxima verosimilitud para una variable aleatoria con distribución binomial es $\hat{p} = \bar{x}$, por lo que quiere decir que maximiza la función la función de probabilidad. Y por lo tanto, la decisión de rechazo es, rechaza H_0 si $\lambda(X) < c$, donde c es una constante. Ahora calculemos,

$$-\ln \lambda(x) = x(\log \bar{x} - \log p_0) + (n-x)(\log(1-\bar{x}) - \log(1-p_0))$$

Y, considerando las siguientes aproximaciones conocidas

$$\ln(1-\bar{x}) - \ln(1-p_0) \approx \frac{\bar{x} - p_0}{1-p_0}, \quad \ln \bar{x} - \ln p_0 \approx \frac{\bar{x} - p_0}{p_0},$$

tenemos entonces

$$\begin{aligned}
-\ln \lambda(x) &= x(\log \bar{x} - \log p_0) + (n-x)(\log(1-\bar{x}) - \log(1-p_0)) \\
&= x \left(\frac{\bar{x} - p_0}{p_0} \right) - (n-x) \left(\frac{\bar{x} - p_0}{1-p_0} \right) \\
&= n \left(\bar{x} \left(\frac{\bar{x} - p_0}{p_0} \right) - (1-\bar{x}) \left(\frac{\bar{x} - p_0}{1-p_0} \right) \right) \\
&= n(\bar{x} - p_0) \left(-\frac{1-\bar{x}}{1-p_0} + \frac{\bar{x}}{p_0} \right) \\
&= n(\bar{x} - p_0) \left(\frac{\bar{x} - p_0}{p_0(1-p_0)} \right) = \frac{(\bar{x} - p_0)^2}{p_0(1-p_0)/n} \\
&= \left| \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} \right|.
\end{aligned}$$

Por lo anterior, sabemos que se distribuye como una normal estándar. Por lo que nuestra región de rechazo es equivalente a

$$\lambda(x) < c \Leftrightarrow \left| \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} \right| > c' = z_\alpha,$$

Lo anterior es la justificación de $\lambda(x) < c$ si $x > c'$, para constantes c, c' . Además, es claro que existe c' para cualquier α , el cual es $c' = z_\alpha$. Lo que no es claro es como elegir c para algún nivel de significancia, en la practica es tomar c tal que

$$\mathbb{P}(\lambda(x) < c) > \alpha$$

. Pero no es tan claro ya que no conocemos la distribución de $\lambda(x)$. ■.

10.1 (Dispositiva 204, notas pasadas) Hallar los estimadores de máxima verosimilitud de θ y λ .

RESPUESTA

Tenemos siguiente función de probabilidad

$$f(x; \lambda, \theta) = \begin{cases} \lambda e^{-\lambda(x-\theta)}, & x \geq \theta \\ 0 & x < \theta \end{cases}$$

Entonces debido a que $x \geq \theta$ eso implica que $x_{(1)} \geq \theta$. Por lo que la función de máxima verosimilitud es

$$L(\lambda, \theta) = \prod_{i=1}^n \lambda e^{-\lambda(x_i-\theta)} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i + n\lambda\theta}, \quad x_{(i)} > \theta.$$

Y así la función log-verosimilitud es

$$\log L(\lambda, \theta) = n \log \lambda - \lambda \sum_{i=1}^n x_i + n\lambda\theta.$$

Por lo tanto, derivando con respecto a λ e igualando a 0 busquemos el punto crítico de la función de log-verosimilitud.

$$\frac{d}{d\lambda} \log L(\lambda, \theta) = \frac{n}{\lambda} - \sum_{i=1}^n x_i + n\theta.$$

Igualando a cero,

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i + n\theta = 0 \Rightarrow \text{es un punto crítico } \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i - n\hat{\theta}}$$

Ahora probemos que el punto encontrado es un máximo, para ello ocupemos el criterio de segunda derivada.

$$\frac{d^2}{d\lambda^2} \log L(\lambda, \theta) = -\frac{n}{\lambda^2}.$$

Por lo tanto, como la segunda derivada siempre es negativa podemos concluir que el punto crítico encontrado es un máximo. Por lo que, **podemos concluir que el estimador de máxima verosimilitud para λ es $\frac{n}{\sum_{i=1}^n x_i - n\hat{\theta}}$** . Ahora, si realizamos un procedimiento análogo para encontrar el estimador de máxima verosimilitud para θ nos encontramos con la dificultad de que la derivada de la función de verosimilitud no es función de θ por lo que no sería posible encontrarlo de esa forma. Entonces prestemos atención a la condición de que $x_{(1)} > \theta$, si $x \leq \theta$ entonces la verosimilitud se hace cero. Por lo que, **podemos concluir que el estimador de máxima verosimilitud para θ es $\hat{\theta} = \min\{x_1, x_2, \dots, x_n\}$** .

10.2 Al hacer 10 observaciones de tiempos pico, se obtuvieron los siguientes valores (en segundos):

$$3, 11, 0, 64, 2, 55, 2, 20, 5, 44, 3, 42, 10, 39, 8, 93, 17, 82, 1, 30.$$

Calcular los estimadores de θ y λ .

RESPUESTA

Ocupando el inciso anterior, los estimadores de máxima verosimilitud para θ y λ dada la muestra son:

$$\begin{aligned} \hat{\theta} &= \min\{x_1, x_2, \dots, x_n\} = \mathbf{0,64}. \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n x_i - n\hat{\theta}} = \frac{10}{55,8 - 10 * 0,64} = \mathbf{0,2024}. \blacksquare. \end{aligned}$$