Práctica N°3

Ivan Fernando Mujica Mamani

Análisis Estadístico I

Maestría en Ciencia de Datos v.2

Noviembre – 2020

1. Ejercicio 21

Suponga que de una población en la que σ = 10 se toma una muestra aleatoria simple de tamaño 50. Halle el valor del error estándar de la media en cada uno de los casos siguientes (si es necesario use el factor de corrección para una población finita).

A. El tamaño de la población es infinito.

Del problema tenemos que:

$n=50$

$\sigma=10$

usando la definición de error estandar de una muestra :

$$\sigma_x=\frac{\sigma}{\sqrt{(n)}}=\frac{10}{\sqrt{(50)}}=1.4142$$

Aplicacion en R:

```
> # Ejercicio 21
> error_estandar = function (muestra, desviacion) {
+   err =  desviacion / sqrt(muestra)
+   cat(err)
+ }
> error_estandar(50, 10)
1.414214
```

B. El tamaño de la población es N = 50 000.

pide el error estandar para una población finita, por lo tanto usamos:

$$\sigma_X=\sqrt{\left(\frac{N-n}{N-1}\right)}\frac{\sigma}{\sqrt{(n)}}=\sqrt{\left(\frac{50000-50}{50000-1}\right)}\frac{10}{\sqrt{(50)}}=\frac{0.9995\,*\,10}{7.07106}=1.41350$$

C. El tamaño de la población es N = 5000.

pide el error estandar para una población finita, por lo tanto usamos:

$$\sigma_X = \sqrt{\left(\frac{N-n}{N-1}\right)}\frac{\sigma}{\sqrt{(n)}} = \sqrt{\left(\frac{5000-50}{5000-1}\right)}\frac{10}{\sqrt{(50)}} = \frac{0.9950 * 10}{7.07106} = 1.4071$$

D. El tamaño de la población es N = 500.

pide el error estandar para una población finita, por lo tanto usamos:

$$\sigma_X = \sqrt{\left(\frac{N-n}{N-1}\right)}\frac{\sigma}{\sqrt{(n)}} = \sqrt{\left(\frac{500-50}{500-1}\right)}\frac{10}{\sqrt{(50)}} = \frac{0.9496 * 10}{7.07106} = 1.3429$$

2. Ejercicio 22

Regrese al problema de los administradores de EAI. Suponga que se usa una muestra aleatoria simple de 60 administradores.

$n = 60$

$\sigma = 4000$

$\mu = 51800$

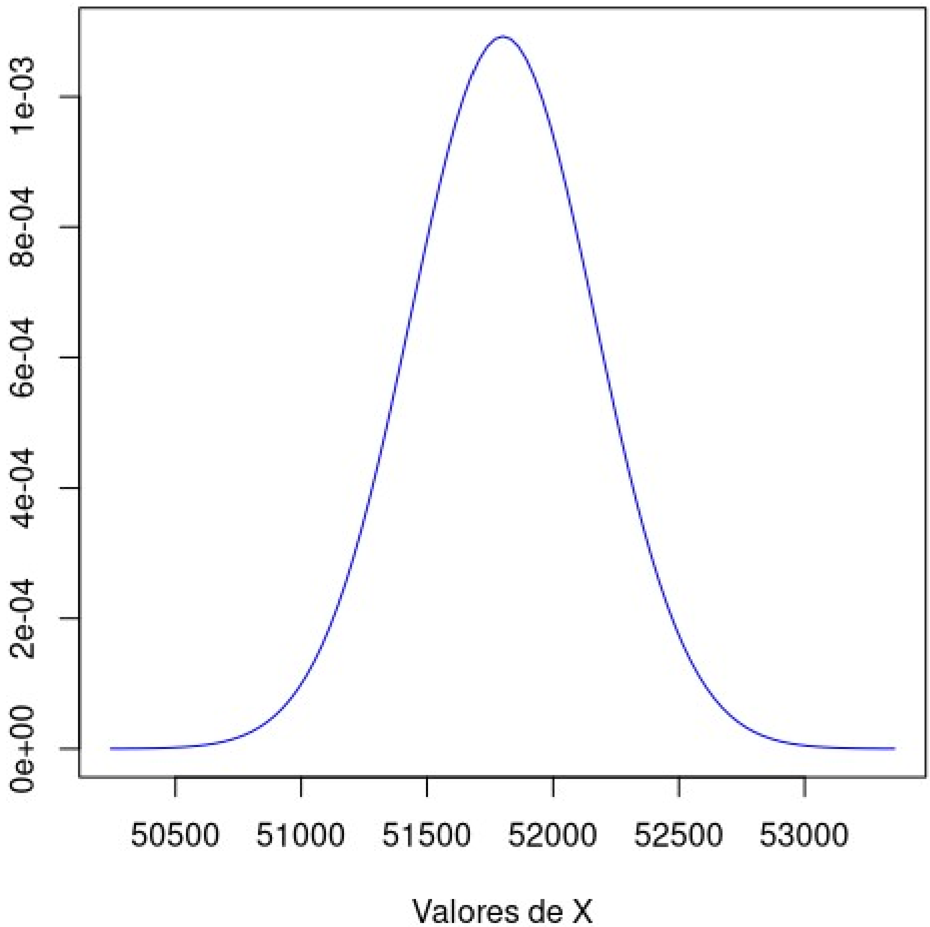
A. Dibuje la distribución muestral de x si se emplean muestras aleatorias simples de tamaño 60.

$$\sigma_x = \frac{\sigma}{\sqrt{(n)}} = \frac{4000}{\sqrt{(60)}} = 516.4$$

Aplicación en R:

```
#A
# Ejercicio 22
# A
miu = 51800
desv = 4000
n = 60
desv_muestral = desv / sqrt(n)
desv_muestral
plot_normal(51800,516.3978)
```

Gráfica:

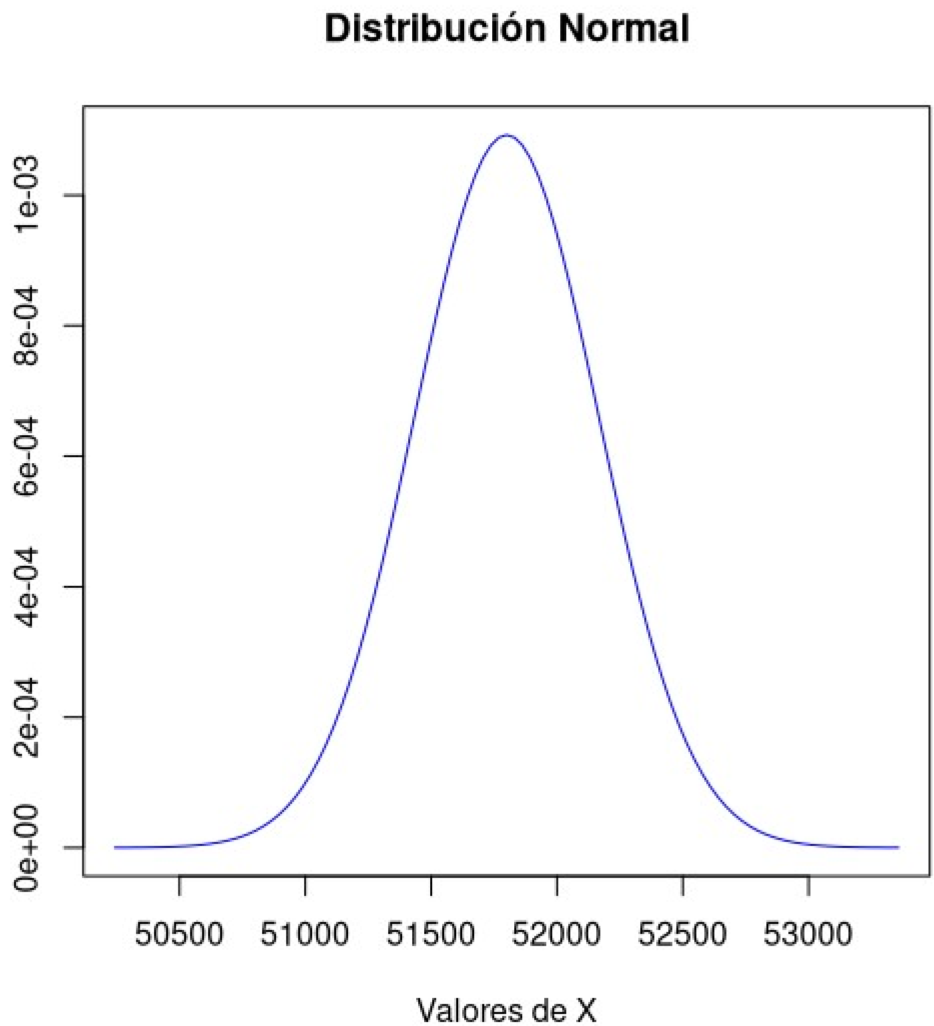


B. ¿Qué pasa con la distribución muestral de si se usan muestras aleatorias simples de tamaño 120?

Con $n = 120$, es decir con una mayor muestra la desviación estandar se decrementa y la distribución muestral tiende a ser mas estrecha.

Aplicación en R:

```
#B
n = 120
desv_muestral = desv / sqrt(n)
desv_muestral
plot_normal(miu,desv_muestral)
```



C. ¿Qué puede decir acerca de lo que le pasa a la distribución muestral de conforme el tamaño de la muestra aumenta? ¿Parece ser lógica esta generalización? Explique.

Con una mayor muestra la desviacion estandar se decrementa y la distribucion muestral tiende a ser mas estrecha. Si es lógica, porque mientras mas grande es la muestra, mas seguros estaremos de media poblacionala porque tenemos más datos de la poblacion en la muestra.

3. Ejercicio por estimación por intervalo de confianza con σ conocido

En una investigación sobre los negocios pequeños que tienen un sitio en la Web se encontró que la cantidad promedio que se gasta en un sitio es $11 500 por año. Dada una muestra de 60 negocios y una desviación estándar σ = $4000, ¿cuál es el margen de error? Use 95% de confianza.

Dado que:

$x=11500$

$\sigma=4000$

$n=60$

$1-\alpha=0.95$

Sabiendo que $\alpha/2=0.025$ y $t_{\alpha/2}=1.984$ (desde la tabla)

$$E=Z_{\alpha/2}\frac{\sigma}{\sqrt{(n)}}=\frac{1.96 * 4000}{\sqrt{(60)}}=1012.14$$

Aplicacion en R:

```
> error = function (muestra, desviacion, nivel) {
+   p=(1+(nivel/100))/2
+   zc<-qnorm(c(p),mean=0,sd=1,lower.tail=TRUE)
+   cat("El Error es,\n")
+   err = zc * desviacion / sqrt(muestra)
+   cat(err)
+ }
> error(60, 4000, 95)
El Error es,
1012.121
```

¿Qué recomendaría si el estudio requiere un margen de error de $500?

500 es menor al error calculado con muestra de 60, para conseguir ese margen se recomienda incrementar n.

4. Ejercicio por estimación por intervalo de confianza con σ desconocido

El número de horas de vuelo de los pilotos del Continental Airlines es 49 horas por mes. Suponga que esta media se basó en las horas de vuelo de una muestra de 100 pilotos de esa empresa y que la desviación estándar muestral haya sido 8.5 horas.

Dado que:

$n=100$

$x=49$

$s=8.5$

$c=1-\alpha=0.95$

Respuestas

A. *¿A 95% de confianza, ¿cuál es el margen de error?*

Sabiendo que $\alpha/2=0.025$ y $t_{\alpha/2}=1.984$ (desde la tabla)

$$E=t_{\alpha/2}\frac{s}{\sqrt{(n)}}=\frac{1.984 * 8.5}{\sqrt{(100)}}=1.6864$$

B. *Dé el intervalo de estimación de 95% para la media poblacional de las horas de vuelo de los pilotos.*

Aplicando la formula $x-E<\mu<x+E$

$49-1.6864<\mu<49+1.6864$

$47.3136<\mu<50.6864$

Aplicacion en R:

```
> # 4
>
> intervalo_media_desv_desconocido=function(muestra,media, desviacion,nivel)
+ {
+   p=(1+(nivel/100))/2
+   tc = qt(c(p),df=muestra)
+   i<-(tc*desviacion)/sqrt(muestra)
+   cat("El intervalo de confianz para la media,\n")
+   cat("al nivel de confianza",nivel,"%es,\n")
+   cat("(",media-i,",",media+i,")\n")
+ }
> intervalo_media_desv_desconocido(100,49,8.5,95)
El intervalo de confianz para la media,
al nivel de confianza 95 %es,
( 47.31362 , 50.68638 )
```

5. Ejercicio determinación del tamaño de la muestra para una estimación por intervalo de la media poblacional con desviacion conocida.

El costo promedio de la gasolina sin plomo en Grater Cincinnati es $2.41 (. En una época de cambios en los precios, un periódico muestrea las gasolineras y presenta un informe sobre los precios de la gasolina. Suponga que en los precios del galón de la gasolina sin plomo la desviación estándar es $0.15; dé el tamaño de muestra n que debe usar este periódico para tener 95% de confianza con cada uno de los márgenes de error siguientes.

Dado que

$c=1-\alpha=0.95$ por lo tanto $Z_{\alpha/2}=1.96$

$\sigma=0.15$

A.Un margen de error de $0.07

E = 0.07

según la fórmula:

$$E = t_{\alpha/2}\frac{s}{\sqrt{(n)}} = \frac{1.96^2 * 0.15^2}{0.07^2} = 18$$

aplicación en R:

```
> #A calculo de n
> tamano_muestra_media=function(error,desviacion,nivel)
+ {
+   p=(1+(nivel/100))/2
+   zc<-qnorm(c(p),mean=0,sd=1,lower.tail=TRUE)
+   n<-((zc*desviacion)/(error))^2
+   cat("El tama?o de la muestra para,\n")
+   cat("un nivel de confianza de ",nivel,"%,\n")
+   cat("con un error y una desviacion","es\n")
+   cat(ceiling(n),"\n")
+ }
> tamano_muestra_media(0.07,0.15,95)
El tama?o de la muestra para,
un nivel de confianza de  95 %,
con un error y una desviacion es
18
```

B.Un margen de error de $0.05

$$n = Z_{\alpha/2}^2\frac{\sigma^2}{E^2} = \frac{1.96^2 * 0.15^2}{0.05^2} = 35$$

```
> tamano_muestra_media(0.05,0.15,95)
El tama?o de la muestra para,
un nivel de confianza de  95 %,
con un error y una desviacion es
35
> 
```

6. Ejercicio determinación del tamaño de la muestra para una estimación por intervalo de la media poblacional con desviacion no conocida.

Los salarios anuales iniciales de estudiantes que acaban de terminar una carrera en administración se espera que estén entre $30 000 y $45 000. Suponga que quiere dar un intervalo de confianza de 95% para estimar la media poblacional de los salarios iniciales. ¿Cuál es el valor planeado de la desviación estándar poblacional?

Dado que

$c=1-\alpha=0.95$ por lo tanto $Z_{\alpha/2}=1.96$

RANGO = 45000 – 30000 = 15000

$\sigma=\frac{15000}{4}=3750$ La desviación es aproximadamente ¼ del rango.

¿Cuán grande deberá ser la muestra si quiere que el margen de error sea

A.$500?

según la fórmula:

$$n=Z_{\alpha/2}^{2}\frac{\sigma^{2}}{E^{2}}=\frac{1.96^{2}*3750^{2}}{500^{2}}=217$$

Aplicación en R:

```
> # 6
> tamano_muestra_media(500,3750,95)
El tama?o de la muestra para,
un nivel de confianza de  95 %,
con un error y una desviacion es
217
>
```

B.$200?

$$n=Z_{\alpha/2}^{2}\frac{\sigma^{2}}{E^{2}}=\frac{1.96^{2}*3750^{2}}{200^{2}}=1351$$

Aplicación en R:

```
> # 6
> tamano_muestra_media(200,3750,95)
El tama?o de la muestra para,
un nivel de confianza de  95 %,
con un error y una desviacion es
1351
>
```

7. **Ejercicio determinación del tamaño de la muestra para una estimación por intervalo para la proporción.**

De acuerdo con estadísticas publicadas por la CNBC, la cantidad de vehículos que no están asegurados es sorprendente. Los resultados muestrales de la CNBC indican que 46 de 200 vehículos no estaban asegurados.

Dado que

$n=200$

$c=1-\alpha=0.95$ por lo tanto $Z_{\alpha/2}=1.96$

$x=46$

A.¿Cuál es la estimación puntual de la proporción de vehículos no asegurados?

Por definicion sabemos que:

$$p=\frac{x}{n}=\frac{46}{200}=0.23$$

B.Dé un intervalo de confianza de 95% para la proporción poblacional.

Por definicion sabemos que:

$$p\pm Z_{\alpha/2}\sqrt{\left(\frac{p(1-p)}{n}\right)}=0.23\pm 1.96\sqrt{\left(\frac{0.23(1-0.23)}{200}\right)}=0.23\pm 0.0552$$

$0.1748<p<0.2883$

8. **Ejercicio determinación del tamaño de la muestra para una estimación por intervalo para la proporción.**

Según Thomson Financial, hasta el 25 de enero de 2006, la mayor parte de las empresas que informaban tener ganancias habían superado las estimaciones. En una muestra de 162 empresas, 104 superaron las estimaciones, 29 coincidieron y 29 se quedaron cortas.

Dado que

$n=162$

$x=29$

A. ¿Cuál es la estimación puntual de la proporción de empresas que se quedaron cortas?

Por definición sabemos que:

$$p=\frac{x}{n}=\frac{29}{162}=0.1790$$

B. Determine el margen de error y dé un intervalo de confianza de 95% para la proporción que superó las estimaciones.

$x=104$

$c=1-\alpha=0.95$ por lo tanto $Z_{\alpha/2}=1.96$

$$p=\frac{x}{n}=\frac{104}{162}=0.6420$$

$$p\pm Z_{\alpha/2}\sqrt{\left(\frac{p(1-p)}{n}\right)}=0.6420\pm 1.96\sqrt{\left(\frac{0.6420(1-0.6420)}{162}\right)}=0.6420\pm 0.0733$$

$0.5687<p<0.7153$

C. ¿De qué tamaño debe de ser la muestra si el margen de error es 0.05?

por definicion sabemos que:

$$n = Z_{\alpha/2}^2 \frac{p * (1-p)}{E^2} = \frac{1.96^2 * 0.6420 * (1-0.6420)}{0.05^2} = 354$$

9. Ejercicios de prueba de hipótesis para sigma al cuadrado conocido

Las declaraciones de impuestos presentadas antes del 31 de marzo obtienen un reembolso que en promedio es de $1056. Considere la población de los declarantes de "última hora" que presentan su declaración los últimos cinco días del periodo para este trámite (normalmente del 10 al 15 de abril)

a) A. Un investigador sugiere que la razón por la que estos declarantes esperan hasta los últimos días se debe a que en promedio obtienen un reembolso menor que los que declaran antes del 31 de marzo. Dé las hipótesis apropiadas de manera que el rechazo de H_0 favorezca la sugerencia de este investigador.

b) B. En una muestra de 400 personas que presentaron su declaración entre el 10 y el 15 de abril, la media de los reembolsos fue $910. Por experiencia se sabe que es posible considerar que la desviación estándar poblacional es $\sigma = \$1600$ ¿Cuál es el valor-p?

C. Con $\propto = 0.5$, ¿Cuál es su conclusión?

D. Repita la prueba de hipótesis anterior usando el método del valor crítico.

Dado que:

$n = 400$

$\bar{x} = 910$

$\sigma = 1600$

$\alpha = 0.05$

Determinamos la hipótesis:

$H_0 : \mu \geq 1056$

$H_a : \mu < 1056$

Se sabe que la distribución muestral de la media muestral tiene media μ y desviación estándar $\frac{\sigma}{\sqrt{n}}$

El valor de z es la media muestral menos la media poblacional, dividida por la desviación estándar.

$$z=\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}=\frac{\overline{910}-1056}{1600/\sqrt{400}}=-1.83$$

```
> n = 400
> mu = 910
> sigma = 1600
> alfa = 0.05
> z = -1.83
> P = 0.0336
> z =(910-1056)/(1600/sqrt(400))
> z
[1] -1.825
>
```

Determinamos la probabilidad utilizando la tabla de distribución normal (valor P).

$$P=P(Z\leftarrow 1.83)=0.0336$$

Como el valor P es menor que el nivel de significancia α: a hipótesis nula se rechaza.

$$P<0.05\rightarrow Se\,rechaza\,H_0$$

El valor crítico es el valor de la tabla de distribución normal correspondiente a una probabilidad de 0.05

$$z=-1.645$$

La región de rechazo, que contiene todos los valores menores que -1.645

Si el valor de la prueba estadística esta dentro de la región de rechazo, entonces se rechaza la hipótesis nula.

$$-1.83\leftarrow 1.645\rightarrow Se\,rechaza\,H_0$$

Entonces, finalmente decimos que:

A) La hipótesis apropiada es $H_0:\mu\geq 1056, H_a:\mu<1056$

B) El valor apropiado para P es $P=0.0336$ **con** $z=-1.83$

C) Con valor $\propto = 0.5$ se rechaza la hipótesis.

D) Utilizando el método del valor crítico también se rechaza la hipótesis

10. Ejercicios de prueba de hipótesis para sigma al cuadrado conocido

Las empresas de seguridad de Wall Street pagaron en 2005 gratificaciones de fin de año de $125 500 por empleado (Fortune, 6 de febrero de 2006). Suponga que se desea tomar una muestra de los empleados de la empresa de seguridad Jones & Ryan para ver si la media de la gratificación de fin de año es diferente de la media reportada para la población.

a. Dé las hipótesis nula y alternativa que usaría para probar si las gratificaciones de fin de año de Jones & Ryan difieren de la media poblacional.

B. Admita que en una muestra de 40 empleados de Jones & Ryan la media muestral de las gratificaciones de fin de año es $118 000. Suponga que la desviación estándar poblacional es $\sigma = \$30000$ y calcule el valor-p.

C. Con $\alpha = 0.05$ como nivel de significancia, ¿cuál es su conclusión?

D. Repita esta prueba de hipótesis usando el método del valor crítico.

Respuestas:

Establecemos los parámetros de entrada:

$n = 40$

$\bar{x} = 118000$

$\sigma = 30000$

$\alpha = 0.05$

a) Determinamos la hipótesis

$H_0 : \mu = 125500$

$H_1 \neq 125500$

La distribución muestral de la media muestral tiene media μ y desviación estándar $\frac{\sigma}{\sqrt{n}}$

b) El valor de z es la media muestral restada por la media poblacional, dividida por la desviación estándar.

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{118000-125500}{30000/\sqrt{40}}=-1.58$$

Aplicación en R:

```
> n = 40
> x = 118000
> sigma = 30333
> alfa = 0.05
>
> z =(118000-125500)/(30000/sqrt(40))
> z
[1] -1.581139
>
```

El valor p es la probabilidad de obtener un valor mas extremo o igual al estadístico de prueba estandarizado z, determinamos la probabilidad utilizando la tabla de distribución normal

$$P=P(Z\leftarrow 1.58\,ó\,Z>1.58)=2\times P(Z\leftarrow 1.58)=2\times 0.0571=0.1142$$

c) Si el valor de P es menor que el nivel de significancia, entonces la hipótesis nula es rechazada.

$$P>0.05\rightarrow Fallo\ para\ rechazar\ la\ hipotesis\ nula\ H_0$$

El valor crítico es el valor en la tabla de la distribución normal correspondiente a la probabilidad de 0.025/0.975:

$$z=\pm 1.96$$

La región de rechazo, luego contiene todos los valores menores que -1.96 y todos los valores mayores que 1.96.

d) **Si la hipótesis de la prueba estadística esta dentro de la región de rechazo. Entonces se rechaza la hipótesis nula**.

$-1.96 \leftarrow 1.58 < 1.6 \rightarrow \textit{Fallo para rechazar la hipotesis nula } H_0$

11. Ejercicio de prueba de hipótesis para sigma al cuadrado desconocido

La Employment and Training Administration informó que la prestación media del seguro de desempleo es $238/semana (The World Almanac, 2003). Un investigador del estado de Virginia anticipó que datos muestrales indicarán que la prestación media semanal del seguro de desempleo en el estado de Virginia es menor que la media de todo el país.

Establecemos los parámetros de entrada:

$n=100$

$\bar{x}=231$

$s=80$

$\alpha=0.05$

***a*. Dé las hipótesis adecuadas de manera que el rechazo de H0 favorezca la afirmación del investigador.**

Determinamos la hipótesis:

$H_0: \mu \geq 238$

$H_a: \mu < 238$

b. En una muestra de 100 individuos la media muestral encontrada fue $231 y la desviación estándar muestral fue $80. ¿Cuál es el valor-p?

$$t=\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}=\frac{231-238}{80/\sqrt{100}}=-0.88$$

Aplicación en R

```
> # 11
> n = 100
> mu = 231
> sigma = 80
> alfa = 0.05
> t = (mu - 238)/(sigma/sqrt(n))
> t
[1] -0.875
>
```

c. Si α =0.05, ¿cuál es su conclusión?

El valor de P es la probabilidad de obtener el valor de las pruebas estadísticas o un valor mas extremo. El valor de P lo encontramos utilizando la tabla t-student en la fila $df=n-1=100-1=99$.

$0.10<P<0.20$

Si el valor p es menor que el nivel de significancia, entonces se rechaza la hipótesis.

$P>0.05 \rightarrow Fallo\,al\,rechazar\,la\,hipotesis\,nula\,H_0$

d. Repita la prueba de hipótesis anterior usando el método del valor crítico.

Determinamos el valor crítico utilizando la tabla t-student con $df=n-1=100-1=99$ y $\alpha=0.05$

$t=1.660$

Entonces la región de rechazo contiene todos los valores menores que -1.660.

Si el valor de la prueba estadística esta dentro de la región de rechazo, entonces se rechaza la hipótesis nula.

$-0.88>-1.660 \rightarrow Fallo\,al\,rechazar\,la\,hipotesis\,nula\,H_0$