PREWORK
SESIÓN 5

Introducción

En este prework, conocerás algunas definiciones y resultados importantes de la estadística inferencial. En el work trabajarás con datos y los cálculos serán muy rápidos y fáciles de hacer con algunas funciones de R.

Objetivos

- Conocer algunas funciones de densidad de variables aleatorias muy útiles y comunes.
- Conocer lo que dice el teorema central del límite
- Conocer procedimientos para llevar a cabo algunos contrastes de hipótesis

Temas

- Teorema central del límite
- Contraste de hipótesis

Teorema central del límite

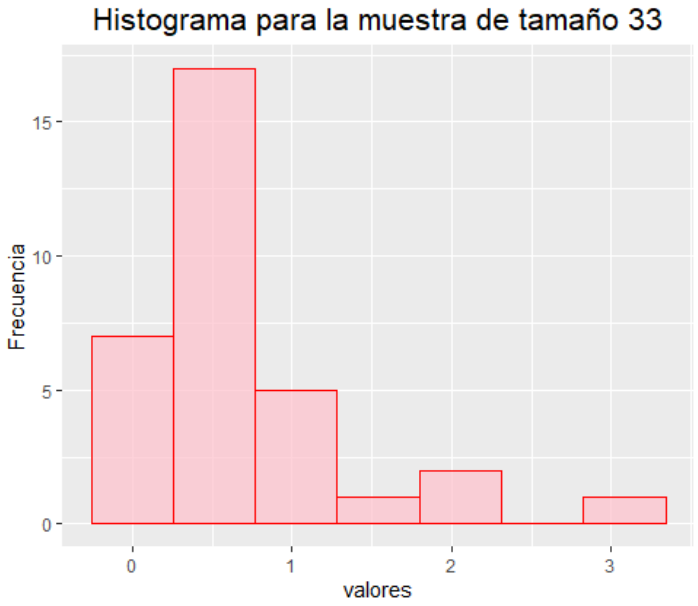
Dada una muestra aleatoria Y_1, Y_2,…, Y_n de *cualquier* población con media μ y varianza σ^2 finita, se cumple que la media muestral (promedio) $\overline{Y}$ está *distribuida normalmente de forma aproximada* con media μ y varianza σ^2/n, siempre y cuando el tamaño muestral sea grande ($n > 30$).

Por ejemplo, si tomamos una muestra aleatoria de tamaño $n = 33$

```
muestra <- c(1.82165160, 1.06824486, 0.38492498, 0.52779737, 0.17989299,
0.38599556, 0.01565589, 0.53166559, 1.08000160, 0.61289266, 0.16050136,
0.35143952, 0.41076615, 1.09468497, 0.53319069, 1.09299258, 0.61343642,
0.15565428, 1.44299912, 0.43475144, 0.60773249, 3.09911364, 0.36185393,
1.00729974, 0.30582083, 0.35948934, 0.20484999, 0.13779880, 0.28064973,
2.03910927, 0.19785169, 0.46706578, 0.30224129)
```

que proviene de una población que está sesgada hacia la derecha, lo cual puede verse al considerar su histograma

```
data <- as.data.frame(muestra)
library(ggplot2)
ggplot(data, aes(muestra)) +
  geom_histogram(colour = 'red', fill = 'pink',
                 alpha = 0.7, bins = 7) + # Intensidad del color fill
  ggtitle('Histograma para la muestra de tamaño 33') +
  labs(x = 'valores', y = 'Frecuencia') +
  theme_get() +
  theme(plot.title = element_text(hjust = 0.5, size = 16))
```



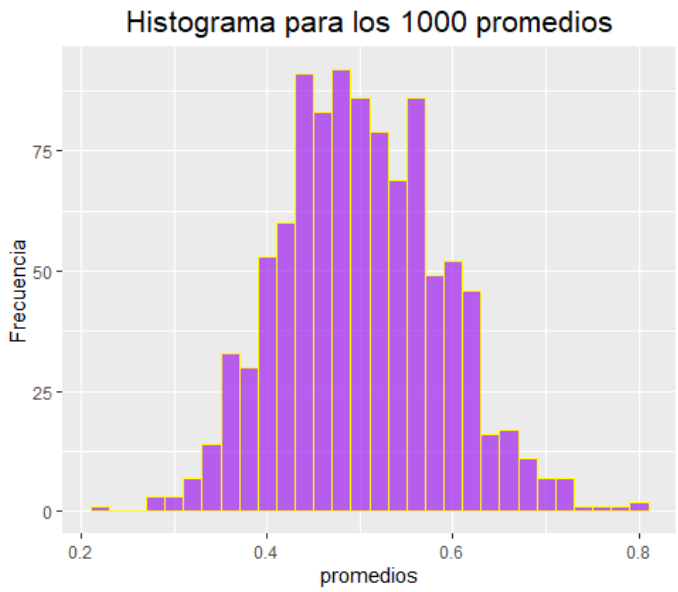
resulta que el promedio de los $n = 33$ valores

```
mean(muestra)
```

```
0.674849
```

es un valor que sí proviene de una población aproximadamente normal. Es decir, si tomáramos 1000 muestras aleatorias diferentes como la anterior (donde cada muestra proviene de la misma población que está sesgada hacia la derecha y por tanto cada uno de los histogramas de las muestras serían similares al histograma anterior), y obtuviéramos los 1000 promedios correspondientes a cada una de las

muestras de tamaño 33, entonces el histograma de estos 1000 promedios, tendría aproximadamente forma de campana (sería simétrico), y luciría como sigue



aún cuando los histogramas de las muestras tomadas inicialmente no tengan forma de campana. Esto siempre ocurre para muestras de cualquier población que satisfacen las condiciones del Teorema central del límite.

Ejemplo. Los tiempos de servicio para los clientes que pasan por la caja en una tienda de venta al menudeo son variables aleatorias independientes con media de 1.5 minutos y varianza de 1.0. Calcule la probabilidad de que 100 clientes puedan ser atendidos en menos de 2 horas de tiempo total de servicio.

Solución.

Denotemos por Y_i el tiempo de servicio para el i-ésimo cliente, entonces queremos calcular

$$P\left(\sum_{i=1}^{100} Y_i \leq 120\right) = P\left(\overline{Y} \leq \frac{120}{100}\right) = P(\overline{Y} \leq 1.20)$$

Como el tamaño muestral es grande, el teorema central del límite nos dice que $\overline{Y}$ está distribuida normalmente en forma aproximada con media $\mu = 1.5$ y varianza $\frac{\sigma^2}{n} = \frac{1}{100}$. Entonces la probabilidad buscada está dada por

pnorm(q = 1.2, mean = 1.5, sd = 1/10)

0.001349898

Entonces, la probabilidad de que 100 clientes puedan ser atendidos en menos de 2 horas es aproximadamente 0.0013. Esta pequeña probabilidad indica que es prácticamente imposible atender a 100 clientes en menos de 2 horas.

Contraste de hipótesis

Los elementos de un contraste de hipótesis

1. Hipótesis nula, H_0
2. Hipótesis alternativa, H_a
3. Estadístico de prueba
4. Región de rechazo

Nota: También llamaremos prueba de hipótesis a un contraste de hipótesis, sin caer en discusiones formales. Buscamos decidirnos por una de las hipótesis y no estamos exentos de cometer errores.

Definición. Se comete un *error tipo I* si H_0 es rechazada cuando H_0 es verdadera.

La *probabilidad de un error tipo I* está denotada por α. El valor de α se denomina *nivel* de la prueba.

Se comente un *error tipo II* si H_0 es aceptada cuando H_a es verdadera. La probabilidad de un *error tipo II* está denotada por β.

Error tipo I y tipo II

H_0: No hay embarazo *vs* H_a: Hay embarazo

Imagen tomada de internet

Contrastes comunes con muestras grandes ($n > 30$)

Suponga que deseamos contrastar dos hipótesis respecto a la media poblacional μ con base en una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ (la población no necesita ser normal). En esta sección presentamos un procedimiento de contrastes de hipótesis que está basado en el estimador $\overline{Y}$ que tiene una distribución muestral normal (aproximadamente) con media μ y error estándar $\frac{\sigma}{\sqrt{n}}$.

Si μ_0 es un valor específico de μ, podemos probar H_0: $\mu = \mu_0$ contra H_a: $\mu > \mu_0$. En este caso, las hipótesis nula y alternativa, el estadístico de prueba y la región de rechazo son como sigue:

H_0: $\mu = \mu_0$.

H_a: $\mu > \mu_0$.

Estadístico de prueba: $Z = \frac{\overline{Y} - \mu_0}{\frac{S}{\sqrt{n}}}$.

Región de rechazo: RR = $\{Z > z_\alpha\}$.

El valor real de z_{α} en la región de rechazo RR se determina al fijar la probabilidad α del error tipo I (el nivel de la prueba), es decir, z_{α} es el número que satisface $P(Z > z_{\alpha}) = \alpha$, o el cuantil de orden $1 - \alpha$.

Ejemplo. Contraste de cola superior

Estamos interesados en contrastar las hipótesis H_0: $\mu = 0.1$ *vs* H_1: $\mu > 0.1$ (contraste de cola superior) con base en una muestra aleatoria de tamaño $n = 40$ de la población.

```
muestra <- c(0.191825830, 0.090832594, 0.078292920, 0.023187365,
0.275329543, 0.120594281, 0.011730131, 0.727012539, 0.108018454,
0.004800318, 0.070778142, 0.539517386, 0.165975518, 0.136258035,
0.216427932, 0.002537893, 0.563361006, 0.027473375, 0.380678788,
0.310481407, 0.142732480, 0.836212104, 0.149678939, 0.288385634,
0.535300943, 0.491167954, 0.429518316, 0.043545325, 0.443696671,
0.078943105, 0.205748181, 0.167813525, 0.017052988, 0.082652468,
0.125213495, 0.166680130, 0.128717925, 0.003860131, 0.045212421,
0.086816614)
```

El valor observado del estadístico de prueba en este caso está dado por

```
(z0 <- (mean(muestra)-0.1)/(sd(muestra)/sqrt(40)))
```

```
3.41015
```

que proviene de una distribución normal estándar aproximadamente.

Supongamos que estamos interesados en encontrar la región de rechazo (de cola superior) con un nivel de significancia $\alpha = 0.05$, debemos encontrar el valor $z_{0.05}$ que satisface $P\left(Z > z_{0.05}\right) = 0.05$, es decir, el cuantil de orden 0.95.

```
(z.05 <- qnorm(p = 0.05, lower.tail = FALSE))
```

```
1.644854
```

A continuación, observamos el valor 1.644854 y la región de rechazo en el eje horizontal

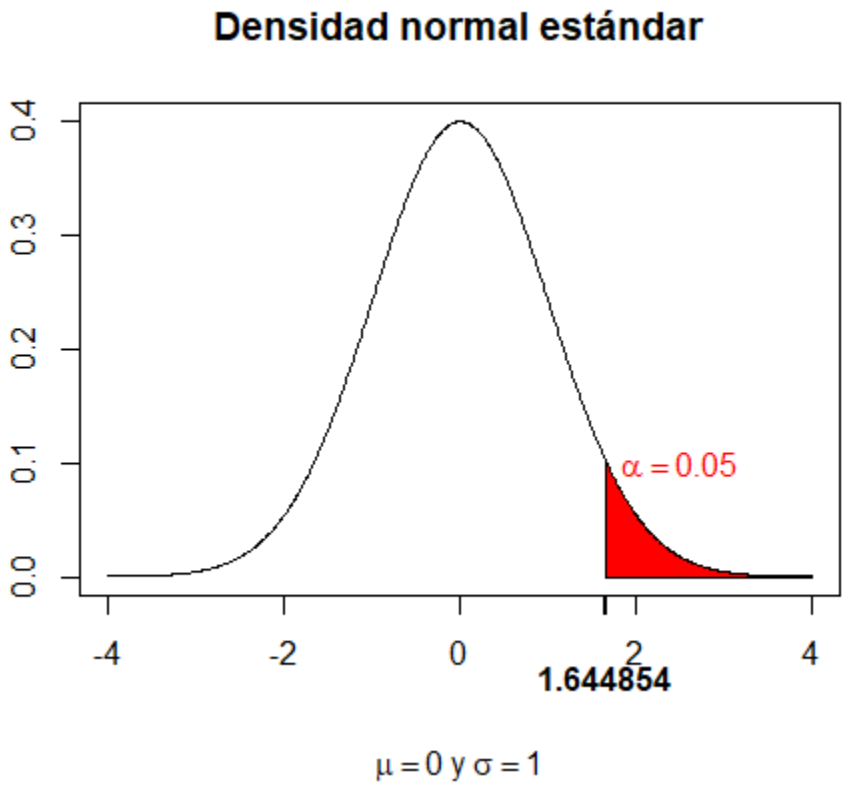
```
x <- seq(-4, 4, 0.01)
y <- dnorm(x)
plot(x, y, type = "l", xlab="", ylab="")
```

```
title(main = "Densidad normal estándar", sub = expression(paste(mu == 0, " y ", sigma == 1)))
```

```
polygon(c(z.05, x[x>=z.05], max(x)), c(0, y[x>=z.05], 0), col="red")
```

```
axis(side = 1, at = z.05, font = 2, padj = 1, lwd = 2)
```

```
text(2.5, 0.1, labels = expression(alpha == 0.05), col = "red")
```



Como

```
3.41015 > 1.644854
```

rechazamos la hipótesis nula.

p-value El p-value lo podemos calcular como

```
(pvalue <- pnorm(z0, lower.tail = FALSE))
```

```
0.0003246356
```

Por lo general, la decisión de rechazar o no la hipótesis nula se toma con base en el p-value. Si el p-value es pequeño (por ejemplo, p-value < 0.05 o p-value < 0.01), se rechaza la hipótesis nula.

Ejemplo. Contraste de cola inferior

Estamos interesados en contrastar las hipótesis H_0: $p = 0.9$ *vs* H_1: $p < 0.9$ (contraste de cola inferior) donde p es la probabilidad de éxito en un experimento que sólo puede resultar en éxito o fracaso (1 o 0) (Ensayo Bernoulli, donde la media es p y la varianza es $p(1 - p)$). Si tomamos una muestra ($n = 45$) aleatoria relacionada con el ensayo Bernoulli

```
muestra <- c(0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0)
```

El valor observado del estadístico de prueba en este caso está dado por

```
(z0 <- (mean(muestra)-0.9)/sqrt((0.9*(1-0.9))/45))
```

```
-7.205108
```

que proviene de una distribución normal estándar aproximadamente.

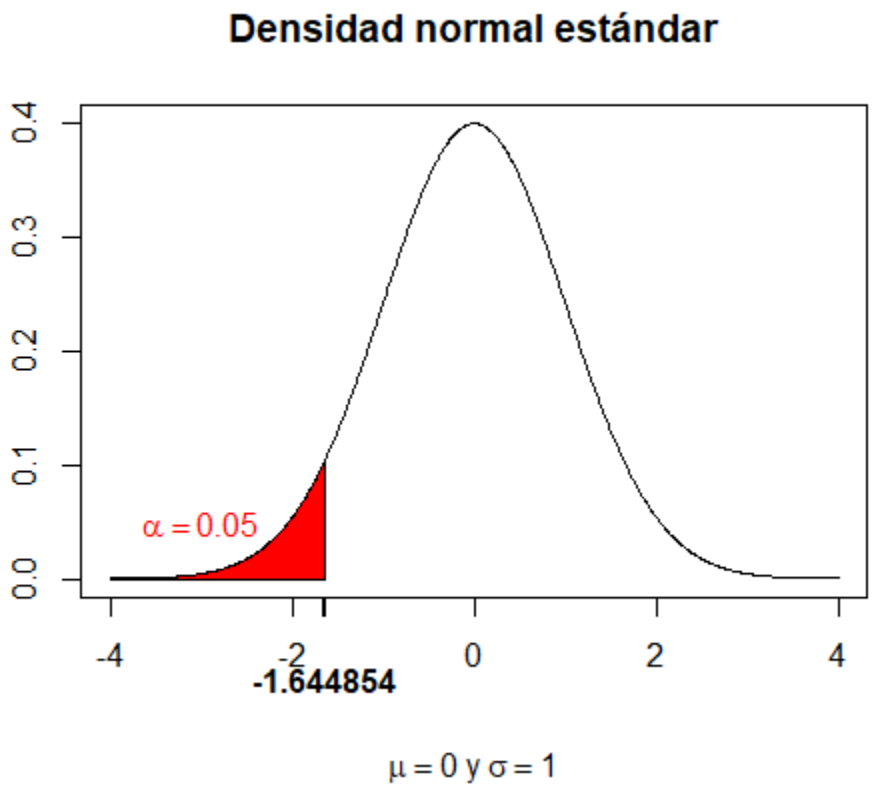
Supongamos que estamos interesados en encontrar la región de rechazo (de cola inferior) con un nivel de significancia $\alpha = 0.05$, debemos encontrar el valor $-z_{0.05}$ que satisface $P\left(Z < -z_{0.05}\right) = 0.05$, es decir, el cuantil de orden 0.05

```
(z.05 <- qnorm(p = 0.05))
```

```
-1.644854
```

A continuación, observamos el valor -1.644854 y la región de rechazo en el eje horizontal

```
x <- seq(-4, 4, 0.01)

y <- dnorm(x)

plot(x, y, type = "l", xlab="", ylab="")

title(main = "Densidad normal estándar", sub = expression(paste(mu == 0, " y ", sigma == 1)))

polygon(c(min(x), x[x<=z.05], z.05), c(0, y[x<=z.05], 0), col="red")

axis(side = 1, at = z.05, font = 2, padj = 1, lwd = 2)

text(-3, 0.05, labels = expression(alpha == 0.05), col = "red")
```



Como

-7.205108 < -1.644854

rechazamos la hipótesis nula.

p-value El p-value lo podemos calcular como

(pvalue <- pnorm(z0))

2.899895e-13

Contraste de hipótesis con muestras pequeñas ($n < 30$) para μ

Supongamos que Y_1, Y_2, ..., Y_n denotan una muestra aleatoria de tamaño *n* de una **distribución normal** con media μ desconocida y varianza σ^2 desconocida. Si $\overline{Y}$ y S denotan la media muestral y la desviación estándar muestral, respectivamente, y si H_0: $\mu = \mu_0$ es verdadera, entonces

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$$

Tiene una distribución *t* con *n-1* grados de libertad.

Contraste de muestra pequeña para μ

Ejemplo. Contraste de cola superior

Estamos interesados en contrastar las hipótesis H_0: $\mu = 170$ *vs* H_1: $\mu > 170$ (contraste de cola superior) con base en una muestra aleatoria de tamaño $n = 15$

```
muestra <- c(166.6896, 175.2299, 170.4218, 176.2738, 183.5532, 179.4669, 179.2014, 173.6239, 176.1826, 182.2429, 176.9100, 166.4572, 172.5695, 180.9723, 180.7529)
```

El valor observado del estadístico de prueba en este caso está dado por

```
(t0 <- (mean(muestra)-170)/(sd(muestra)/sqrt(15)))
```

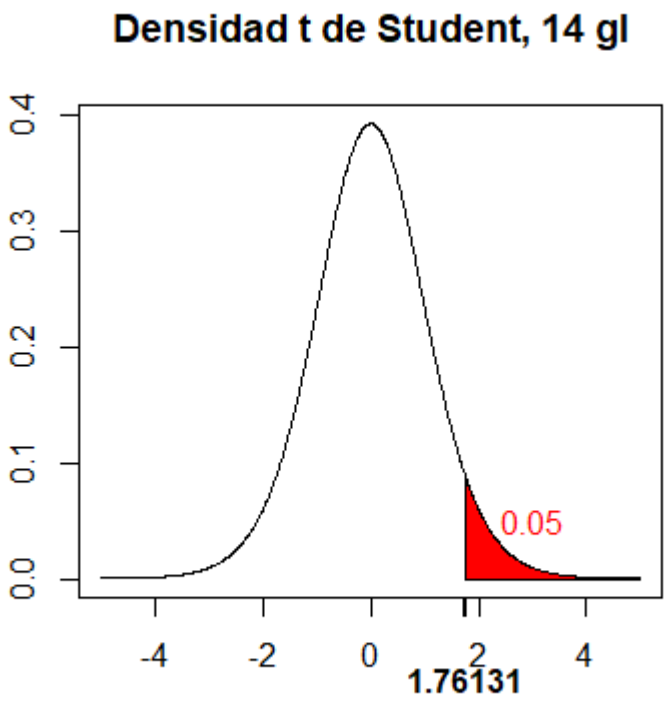
```
4.410437
```

que proviene de una distribución t de Student con $n - 1 = 14$ grados de libertad (gl).

Supongamos que estamos interesados en encontrar la región de rechazo (de cola superior) con un nivel de significancia $\alpha = 0.05$, debemos encontrar el valor $t_{0.05}$ que satisface $P\left(T > t_{0.05}\right) = 0.05$, donde T se distribuye como t de Student con $n - 1 = 14$ gl.

```
(t.05 <- qt(p = 0.05, df = 14, lower.tail = FALSE))
```

```
1.76131
```



Como

4.410437 > 1.76131

rechazamos la hipótesis nula.

p-value El p-value lo podemos calcular como

(pvalue <- pt(t0, df = 14, lower.tail = FALSE))

0.000296395

También podemos usar la función t.test para llevar a cabo el procedimiento de contraste de hipótesis

t.test(x = muestra,

alternative = "greater",

mu = 170)

Ejemplo. Contraste de dos colas Una máquina expendedora de gaseosas fue diseñada para descargar en promedio 7 onzas de líquido por taza. En una prueba de la máquina, diez tazas de líquido se sacaron de la máquina y se midieron. La media y la desviación estándar de las diez mediciones fueron 7.1 onzas y 0.12 onzas respectivamente. ¿Estos datos presentan suficiente evidencia para indicar que la descarga media difiere de 7 onzas? ¿Cuál es la decisión adecuada si $\alpha = 0.10$?

Solución.

Si μ representa la verdadera media (desconocida) de la población de descargas de onzas de líquido por taza, entonces tenemos el siguiente contraste de hipótesis

$$H_0: \mu = 7 \text{ vs } H_1: \mu \neq 7$$

tenemos lo siguiente

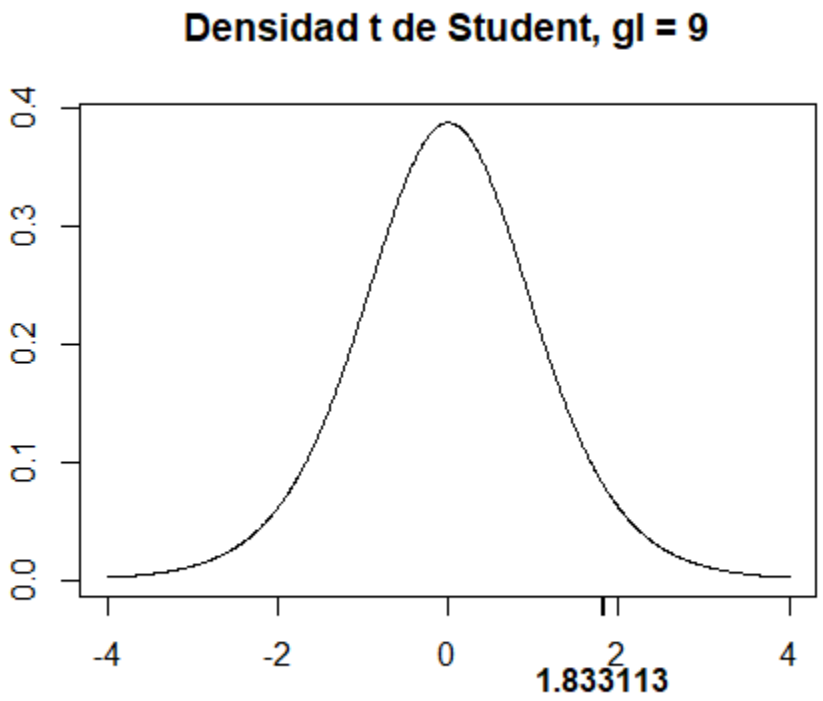
$n = 10, \bar{y} = 7.1, s = 0.12, gl = 10 - 1 = 9$

El valor observado de T es $t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.1-7}{\frac{0.12}{\sqrt{10}}} = 2.635.$

Como $\alpha = 0.10, \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} = 1.833 =$

$$qt(0.05,\ df\ =\ 9,\ lower.tail\ =\ FALSE)$$

Mostramos el cuantil encontrado en el eje de medición (eje horizontal)



La región de rechazo de dos colas está dada por

RR = $\{|t| > t_{0.05}\} = \{|t| > 1.833\} = \{t < -1.833 \text{ o } t > 1.833\}.$

Como el valor observado del estadístico de prueba $t = 2.635$, cae en la región de rechazo, porque $t = 2.635 > 1.833$, se rechaza la hipótesis nula.

Quiz

1. Dada una muestra aleatoria $X_1, X_2, ..., X_n$ $(n > 30)$ de una población (no necesariamente normal) con media μ y varianza σ^2 finitas, ¿Cómo se distribuye aproximadamente la variable aleatoria $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$?

a) Se distribuye binomialmente

b) Se distribuye como una exponencial
c) Se distribuye aproximadamente de forma normal

d) No se puede determinar

2. Consulte en diferentes fuentes cuando consideramos que una muestra es pequeña cuando llevamos a cabo contrastes de hipótesis.

a) n < 10, 000
b) n < 30
c) n < 1000

d) n < 100

3. Dada una muestra aleatoria $X_1, X_2, ..., X_n$ de una población normal con media μ y varianza σ^2 desconocida, ¿Cómo se distribuye la variable aleatoria $\frac{\bar{X}-\mu}{S/\sqrt{n}}$?

a) Se distribuye binomialmente

b) Se distribuye como una exponencial
c) Se distribuye aproximadamente de forma normal

d) Se distribuye como una t de Student con n-1 grados de libertad

BIBLIOGRAFÍA UTILIZADA

- Wackerly, D. et al. (2010). Estadística Matemática con Aplicaciones. Cengage Learning Editores, S.A. de C.V.