Tema 4: Intervalos de Confianza para una Población Normal

Población → Podemos trabajar con estos

↓ ↓

μ ← $\bar{X}$ Estadística

Media Poblacional

Gráficas con R

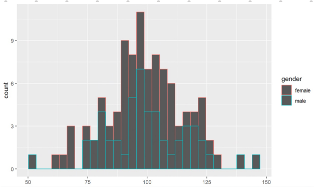
mapeo de columnas a atr estéticos

ggplot(___ , aes (x = ___ , y = ___) + geom_point( )

matriz "aesthetic" → x = iq → cojo matriz $iq : columna

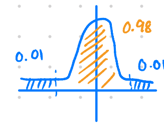
Representar con puntos → , col = ___

col = gender
* pinta matriz$gender




```
#HACEMOS HISTOGRAMA
    ggplot(iq_df, aes( x=iq)) + geom_histogram() #ggplot se encarga: cuentas + histograma
     ggplot(iq_df, aes( x=iq, col=gender)) + geom_histogram()
```

Ejemplo: nivel de confianza del 98%. Para $\bar{X} = 99$



$-z = \text{qnorm}(0.01) = 2.32$

$z = \text{qnorm}(0.99) = -2.32$

Si (curva) o (curva) los nºs ≠

$P(-2'32 < z < 2'32)$ $\left(\sigma^2 = \frac{15^2}{100}\right)$

$$\begin{cases} -2'32 < \frac{\bar{X}-\mu}{\sigma} \to \mu < \bar{X} - 2'32 \cdot \sigma = 99 \cdot 2'32 \cdot \frac{15}{10} = 102.48 \\ 2'32 > \frac{\bar{X}-\mu}{\sigma} \to \mu < 99 + 2'32 \cdot \frac{15}{100} = 92'52 \end{cases}$$

$\mu \in (92.52, 102.48)$

Intervalos de confianza para media si Var[X] conocida

Ejercicio: CIs — Muestra

El fichero "iq.csv" contiene los resultados de un test de cociente intelectual (CI) realizado a una muestra de estudiantes de cierta universidad (con cientos de miles de estudiantes matriculados). ¿Cuál es el cociente intelectual medio de los alumnos de la universidad? Ten en cuenta que los tests de cociente intelectual se diseñan para que la desviación estándar poblacional sea de 15 puntos.

iq.csv = import datset

* copias ↝

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(readr)
iq_df <- read_csv("data_1/iq.csv")
View(iq)
```

1) Busca un estadístico para estimar el parámetro poblacional

Glógico media μ $\xrightarrow{\text{Para}}$ media poblacional

```
# media muestral --> mean
sum(iq_df[, 3])/100
mean(iq_df$iq)
```

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Datos obtenidos → 99
Lo habitual → 100 } ¿los alumnos tienen un cociente bajo? NO!

Estimación puntual

2) Buscar distribución en el muestreo para el estadístico

Asumiremos { ... o bien ⇒ Población infinita
... o bien ⇒ Muestreo aleatorio simple (con reemplazam.)

¿Cuál es la distribución de $\bar{X}$?

- Todos ± en un punto
- Hay pocos q se alejan
- Forma de campana

$\bar{X} \sim N(\mu, \sigma^2)$

Si puedo asumir normalidad...

$$E[X] = E\left[\frac{\sum x_i}{N}\right] = \frac{\sum E[X_i]}{N} = \frac{\sum_{i=1}^{N} \mu}{N} = \frac{N \cdot \mu}{N} = \mu$$

$$Var[X] = Var\left[\frac{\sum x_i}{N}\right] = \frac{1}{N^2} \cdot \sum \underbrace{Var[x_i]}_{\sigma^2} = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N}$$

↑ por indep.

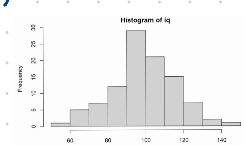
N>>
N<<

Si la población es Normal: $\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)}$

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{15^2}{100}\right)$ • Población ∞ ✓✓

* Para la distrib. Normal utilizamos Histograma

```
#histograma, contar cuantos casos caen para cada numero
iq = iq_df$iq
hist(iq)
#es razonable asumir normalidad
```



3) Escribir matemáticamente la regla del 68-95-99 para el nivel de confianza deseado

68 ⇒ $P(\mu, \frac{15}{10} < \bar{X} < \mu, \frac{15}{10})$

95 ⇒ $P(2\mu, \frac{15}{10} < \bar{X} < 2\mu, \frac{15}{10})$

99 ⇒ $P(3\mu, \frac{15}{10} < \bar{X} < 3\mu, \frac{15}{10})$

$P(\mu - z, \frac{15}{10} < \bar{X} < \mu + z, \frac{15}{10}) = \alpha$

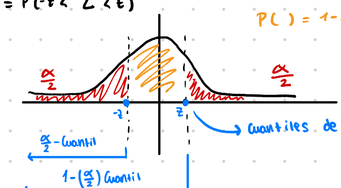
"nivel de confianza" $X \sim N(\mu, \frac{15}{10})$

$= P(-z, \frac{15}{10} < \bar{X} - \mu < z, \frac{15}{10})$ $\bar{X} - \mu \sim N(0, \frac{15}{10})$

$= P(-z < \frac{\bar{X}-\mu}{15/10} < z)$ → $\frac{\bar{X}-\mu}{15/10} \sim N(0,1)$ "Estandarización" "Tipificación"

$= P(-z < Z < z)$

P( ) = 1 - α



cuantiles de la $Z \sim N(0,1)$

$\frac{\alpha}{2}$ - cuantil

$1 - (\frac{\alpha}{2})$ cuantil

qnorm
```

Tema 4: Intervalos de Confianza

*Código → falta

Intervalos de confianza para la varianza Muestral

Estadístico razonable para varianza poblacional σ^2:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \to E[(x-\mu)^2]$$

Este estimador es sesgado → $E[S^2] \neq \sigma^2$

Para ello definimos la Varianza Muestral:

$$\boxed{S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

R: var( )

Perdemos un grado de libertad

Si se toman n muestras indep → población normal : la VA ▸ $\frac{(n-1)\hat{S}^2}{\sigma^2}$

tiene chi-cuadrado n-1 grados de libertad ▸ $\boxed{\frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}}$

Ejercicio: Chips

Una máquina fabrica cierta pieza de un ordenador. El tamaño deseado de la pieza es de 5 cm. En el proceso de fabricación siempre hay circunstancias que no se pueden controlar, por lo que el tamaño de la pieza varía aleatoriamente. El proceso de fabricación está diseñado de forma que el tamaño de cualquier pieza $X_i \sim \mathcal{N}(5, \sigma^2)$, aunque todavía no se conoce σ^2. Para ello, se dispone de los datos almacenados en "pieces.csv". Halla un intervalo de confianza al 99 % para σ^2.

Asumir q las piezas son indep → sino el estadístico no sería chi-cuadrado

1) El estadístico

.csv → dframe   columna

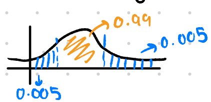
$$\hat{S}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X}) = \text{var(pieces\$pieces\_size.cm)} = 0.00102015$$

2) Hallar la distribución de Muestras

$$\frac{(n-1)\cdot \hat{S}_2}{\sigma^2} \sim \chi^2_{n-1} \qquad \frac{29 \cdot \hat{S}_2}{\sigma^2} \sim \chi^2_{29}$$

n-1 muestras = 29

3) $P\left(a < \frac{29 \cdot \hat{S}^2}{\sigma^2} < b\right)$



a = qchisq(0.005, df = 29) = 13.12

b = qchisq(0.995, df = 29) = 53.33

$$P\left(13.12 < \frac{29 \cdot \hat{S}_2}{\sigma^2} < 53.33\right)$$

$13.12 < \frac{29 \cdot \hat{S}_2}{\sigma^2}$; $\sigma^2 < \frac{29 \cdot \hat{S}_2 = 0.00102015}{13.12}$; $\sigma^2 < 0.002255$

$53.33 > \frac{29 \cdot \hat{S}_2}{\sigma^2}$; $\sigma^2 > \frac{29 \cdot 0.00102015}{53.33}$; $\sigma^2 > 0.00056$

$\sigma^2 \in (0.0022, 0.0005)$

Número de Bins

geom_histogram(bins = 5)

¿Qué nº es conveniente?

↳ nclass.FD(dataframe\$colum)

Intervalos de Confianza para la Media y Varianza desconocida

Si n muestras se toman de una población normal, la VA:

$$\boxed{\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}}$$

Recuerda: $X \sim N(\mu, \frac{\sigma}{n}) \to z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

se distribuye como una "T de Student" con n-1 grados de libertad

$$\boxed{\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim T_{n-1}}$$

Ejercicio:

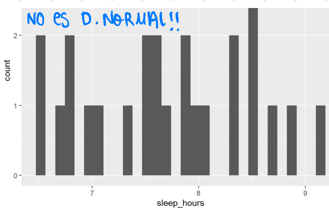
Nueva York es conocida como "la ciudad que nunca duerme". En cierta encuesta se preguntó a una muestra aleatoria de 25 neoyorquinos cuánto tiempo dormían por la noche. Los datos se encuentra en "new_york.csv". ¿Apoyan los datos la afirmación de que los neoyorquinos duermen menos de 8 horas por noche en promedio? Usa un nivel de confianza del 96 %.

X_i: "tiempo q se duerme en Ny"

μ es indep { reemplazamiento no; población ∞ sí ✓

¿Es $\bar{X}$ normal?

↳ ggplot( ) + geom_histogram



Asumir Normalidad:

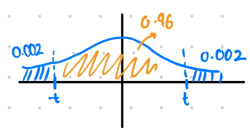
1) Estimador Puntual

```
mean (new_york$sleep_hours)
sd(new_york$sleep_hours)
```
= 7.73
= 0.77

2) $\frac{\bar{X} - \mu}{sd/\sqrt{n}} = \frac{\bar{X} - 7.73}{0.77/\sqrt{25}} \sim t_{24}$

$0.96 = P\left(-t < \frac{\bar{X} - \mu}{sd/\sqrt{n}} < t\right)$ -t y t

T de Student distrib. simétrica



-t = qt(0.002, df = 24) = -2.17

t = qt(0.962, df = 24) = 2.17

$P\left(-2.17 < \frac{\bar{X} - \mu}{\hat{S}/5} < 2.17\right)$

{

Preguntar!!

$\mu \in (\quad)$

Tema 4: Intervales de Confianza

Intervalos de Confianza Para la Media y Varianza desconocida

Ejercicio:
Nueva York es conocida como "la ciudad que nunca duerme". En cierta encuesta se preguntó a una muestra aleatoria de 25 neoyorquinos cuánto tiempo dormían por la noche. Los datos se encuentra en "new_york.csv". ¿Apoyan los datos la afirmación de que los neoyorquinos duermen menos de 8 horas por noche en promedio? Usa un nivel de confianza del 96 %.

Resuelva con la función t.test

t.test → es para las f(x)s T de Student

1) Discutir asunciones → Distrib Normal (histograma); muestreo con reemplazamiento o población ∞

2) La f(x) t.test ⇒ Mucha info [?t.test]

t.test(new-york$sleep hours) → toda la info

t.test(new-york$sleephours, conf.level = 0.96)

↳ Toda la info. Pero con nivel de confianza 96 %.

TEST DE HIPÓTESIS PARA UNA POBLACIÓN NORMAL

Test de Hipótesis de 2 colas

Ejemplo: Homeopatía y pérdida de peso
Un producto homeopático afirma que **"gracias a su uso, perderás 2 Kg en dos semanas"**.

Escéptico ante esta afirmación, reclutas a 50 personas de tu ciudad para participar en un experimento. Las personas usan el producto homeopático durante dos semanas y reportan su pérdida de peso (por ejemplo, $x_i = 3$ significaría que se han perdido 3 Kg, mientras que $x_i = -3$ significaría que se han ganado 3). Datos en "homeo_weight_loss.csv".

En base a los datos, ¿es creible la afirmación del producto homeopático? Por sencillez, asume que la pérdida de peso tiene desviación típica poblacional $\sigma = 2.5$.

mean(dataframe $ column) = 0.037

▸ Los datos no parecen apoyar al Fabricante

1) Asumimos Inocencia Fabricante: Hipótesis Nula

H0 : mu = 2 (la media de perder kg es 2)

asumiendo Normalidad:

$X_i \sim N(\mu, \sigma^2) \xrightarrow{\text{Pobl } \infty} N\left(2, \sigma^2 = \frac{2.5^2}{50}\right)$

2) Bajo esta hipótesis ⇒ $P(\bar{X} < 0.037) \simeq 0$ (es muy baja respecto a 2)

PROCEDIMIENTO

1. Aceptamos Hipótesis de Partida (Hipótesis nula, Ho). Lo contrario de lo q queremos aceptar (Hipótesis alternativa, Ha)
2. Estadístico de contraste T q nos permita testear la veracidad de Ho.
   ⚠ IMP ⇒ discutir asunciones
3. Calculamos P-valor: Prob. observar estadístico tan extremo ↓P-valor = ↓evidencia Ho
4. Comparamos P-valor con nivel de significancia α
   ▸▸ P-valor < α: Ha ✓✓
   valores más típicos de α: 0.05 y 0.01

```r
library(readr)
homeo_weight_loss <- read_csv("data_2/data_2/homeo_weight_loss.csv")
#View(homeo_weight_loss)
media_muestral = mean(homeo_weight_loss$weight_loss_Kg)

# 1) HIPÓTESIS
  # H0: gracias a su uso, perderás 2kg
  #Ha, H1: el uso del producto no conlleva pérdida de 2kg

  #H0: mu = 2 (mu es la pérdida de peso medio)
  #Ha: mu != 2

#2) Buscamos un estadístico de contraste
  #para adivinar mu, usamos X_gorrito
  #X_gorrito ¿? --> N(media, sd^2/n)
  #Es independiente --> población infinita (muestra pequeña respecto a la población)

  #Observamos histograma --> vemos si tiene distribución normal
  num_bins=nclass.FD(homeo_weight_loss$weight_loss_Kg)
  ggplot(homeo_weight_loss, aes(x= weight_loss_Kg)) + geom_histogram(bins = num_bins)
  # Observo que X es normal y como se cumplen ambas condiciones:
  # X_gorrito ¬ N(media, sd^2/n)

  #Bajo la H0 (asimos cierta): X_gorrito ¬ N(2, (2,5^2)/50)

#3) Calculamos p-valor (confrontar datos vs H0)
  #a) de forma gráfica (dibujar distrib normal N(2, (2,5^2)/50) )
    eje_x = seq(0,5, by = 0.01) #porque se q va a estar centrada en 2
    eje_y = dnorm(eje_x , mean = 2, sd= sqrt(2.5 ^2 / 50))
    plot(eje_x, eje_y, type = "l")
    x_mean = mean(homeo_weight_loss$weight_loss_Kg) #0.037
    #Dibujamos una linea vertical en 0.037
    abline(v = x_mean, col = 2)
  #¿ES PROBABLE OBTENER ROJO SI NEGRO ES CIERTO?

  #b) calculamos p-valor: prob de obtener un evento tan o más extremo que el observado
      #evento observado: 0.037
      #más extremo que 0.037--> más a la izq
       abline(v = 4 - x_mean, col = 2) #tb muy extremo si están a la dcha

      #P(X_gorrito < 0.037 o x > 4-0.037)
      # 2 * P(X_gorrito < 0.037 o x > 4-0.037)
      #multiplicamos por dos porqq la probabilidad de la izq es la misma que la de la dcha
       p_valor = 2 * pnorm(0.037, mean = 2, sd= sqrt(2.5 ^2 / 50) )

#4) comparamos el p-valor con un umbral llamado nicel de significancia: alfa
    #Siginficancia alfa: 0.01
      #Forma antigua de presentarlo:
       # Si p-valor < 0.01 ==> Aceptamos H1, rechazamos H0
       # Si p-valor < 0.01 ==> Aceptamos H0, rechazamos H1

      #Forma moderna de analizarlo:
       # Si p-valor < 0.01 ==> los datos apoyan la H1 mientras que la H0 es poco veraz
       # Si p-valor > 0.01 ==> los datos no apoyan la H1 / no hay suficiente evidencia...

print (
  "En este caso los datos apoyan que mu != 2kg (NO vas a perder 2kg) "
)
```

Tema 4: Intervalos de Confianza

TEST DE HIPÓTESIS PARA UNA POBLACIÓN NORMAL

Cuando no conocemos la Varianza:

```
# ---------SIGMA NO ES CONOCIDA --------
#1)
  #H0: mu = 2 (mu es la pérdida de peso medio)
  #Ha: mu != 2

#2) Distribucion de estadsitico (mirar en la tabla)
  # X_gorrito es una T de Student
  # Asunciones acerca de Normalidad y pobl infinita
  # Podemos calcular T de Students --> t.test

  t.test(homeo_weight_loss$weight_loss_Kg)
    #p_value = 0.9215 (apoya la hipotesis nula)
    # Hipotesis alternativa: verdadera pero mu != 2

  t.test(homeo_weight_loss$weight_loss_Kg, mu = 2)
    # como p-value es muy pequeño, HIpotesis alternativa bien
    # Los datos apyan la hipótesis alternativa
```

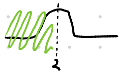
```
#3) Calcular a partir de TContraste p-valor
        ?pt #(calcula la probabilidad a la izq de la línea roja)
        p_valor = 2 * pt(t_contraste, df = n-1)

        #Como p-valor es muy pequeño, Ha es verdadera
```

Test de Hipótesis de una sola cola y tamaño del efecto

Test de Hipótesis de una sola cola

Ejercicio: Test de una sola cola
Un producto homeopático afirma que "gracias a su uso, perderás **al menos** 2 Kg en dos semanas"... ¿Es creíble esta afirmación?

"al menos 2 kg"



$\leq \leftrightarrow <$

1) Hipótesis:

(antes)
Ho: $\mu = 2$
Ha: $\mu \neq 2$

(ahora)
Ho : mu > 2
Ha : mu < 2 (Pierde al menos 2 kg)

2) ¿? un estadístico de Contraste

- VA continua → Distrib. Normal → Pobl ∞ ▷▷ T de Student

```
my_test = t.test(homeo_weight_loss$weight_loss_Kg,
                 alternative = "less",
                 mu = 2)
```

3) Calculamos el p-valor y confrontamos datos con Ho

2 formas:

```
p_valor = my_test$p.value
p_valor2 = pt(t_contraste)
```
→ PREGUNTAR !! no me f(x) na

4) Dibujamos

```
------------DIBUJAMOS-----------------
eje_x = seq(-6,6, by=0.1)
plot(eje_x, dt(eje_x, df = n-1), type = "l")
abline( v = t_contraste, col=2, lwd=2)
```

gráfica :)

Hipótesis y errores

Rechazaremos la Ho (siendo ✓✓) un α% veces

```
set.seed(4)
x <- rnorm(30, mean = 0, sd = 5) #normal cuya media = 0
print(
  t.test(x, mu = 0)$p.value       #H0: mu = 0, Ha != 0

)
#Datos demuestran que la media es != 0 (ERROR)
```

Que Ha sea FALSA $\xrightarrow{\text{NO IMPLICA}}$ Ho VERDADERA

```
set.seed(42)
x <- rnorm(30, mean = 0, sd = 5)
print(c(
  t.test(x, mu = 1)$p.value,     #H0: mu=1, Ha: mu!=1
  t.test(x, mu = -1)$p.value     #H0: mu=-1, Ha: mu!=-1
))
```

```
alpha =0.05 #Falla 5% de las veces

sims <- replicate(5000, {
  # H0: mu = 0
  x <- rnorm(100, mean = 0)            # Genera muestras de H0 (H0 es correcta)
  pvalor <- t.test(x, mu = 0)$p.value     # Testea H0
  pvalor<alpha # "Rechazar H_0 a pesar de ser correcta"
})
print(paste("alpha =", alpha, "| p(incorrectly reject H0) = ", mean(sims)))
```

```
[1] "alpha = 0.05 | p(incorrectly reject H0) =  0.0546"
```
$\underbrace{\text{alpha = 0.05 | p(incorrectly reject H0) = 0.0546}}_{\simeq}$ ✓✓

Errores Tipo I y tipo II

	H_0 es cierta	H_a es cierta
Se escogió H_0	No hay error	Error de tipo II/Falso negativo (β)
Se escogió H_a	Error de tipo I/Falso positivo (α)	No hay error

P(escoger Ha | Ho ✓✓) = α (falso Positivo)
P(escoger Ho | Ha ✓✓) = β (falso Negativo)
P(escoger Ha | Ha ✓✓) = 1 - β "Potencia del Contraste"

Tamaño del efecto

Gracias a tu éxito con el análisis del producto homeopático, una farmacéutica interesada en desarrollar un fármaco para la pérdida de peso te contrata. La empresa quiere comercializar su (carísimo) producto con un eslogan del tipo "Hay evidencia científica de nuestro producto te hará perder peso si lo usas dos meses".

Te facilitan los datos de "pharma_weight_loss.csv". ¿Hay suficiente evidencia de que el fármaco te hace perder peso? (Usa $\alpha = 0.05$.)

1) Hipótesis

Ho: mu < 0 (ganas kg)
Ha: mu > 0

2) ¿? Estadístico de contraste ▶ Distrib Normal + Pobl. ∞ = $\bar{X} \sim$ T de Student

↓ geom_histogram

```
my_test = t.test(pharma_weight_loss$weight_loss_Kg, alternative = "greater", mu = 0)
```

3) Calculamos p-valor

```
pvalor = my_test$p.value
```

4) Comparamos p-valor con α

```
alpha = 0.05
pvalor < alpha #TRUE
```

Los datos apoyan la hipótesis alternativa

Tema 4: Intervalos de Confianza

Para el anterior ejercicio: Calcule Tamaño del Efecto ⚠

```r
#Tamaño del efecto: calcula la relevancia del resultado
  #num pequeños --> poco relevantes
  weight = pharma_weight_loss$weight_loss_Kg
  cohens_d =  mean(weight)/sd(weight)
  #¿0.108571 nº pequeño?
  #no tenemos porq saber programarlo --> utilizamos librería
```

```
> cohens_d
[1] 0.1085716
```

```r
#install.packages("effectsize")
library("effectsize")
effectsize(my_test)
```

```
Cohen's d |       95% CI
------------------------
0.11      | [0.06, Inf]
```

```r
#consultamos tabla --> Cohen's d (t-tests): Small: 0.2 Medium: 0.5 Big: 0.8
print(
  "El tamaño del efecto es pequeño ya que 0.11<0.2,
  el resultado es poco relevante"
)

my_test
# Vemos q el nivel de confianza va de 0.055 a Inf --> De una cola
# Estamos al 95% seguros que la pérdida de peso media es superior a 0.055
```

```
        One Sample t-test

data:  pharma_weight_loss$weight_loss_Kg
t = 3.4333, df = 999, p-value = 0.0003103
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.05542461        Inf
sample estimates:
mean of x
0.1064891
```

↓

```
¿Cómo reportar resultados?

Los datos apoyan que hay una pérdida de peso gracias al uso del producto,
(p-valor = 0.003). Sin embargo, el tamaño del efecto es pequeño, luego los
resultados son poco relevantes (la pérdida de peso es poco relevante).
(Con 95% de confianza, la pérdida > 0.05kg)
```

Potencia de un Test

Ejercicio: Test para la varianza

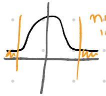
Los test de cociente intelectual (CI) están diseñados para que la desviación típica poblacional sea de 15 puntos. Sin embargo, en los procesos de traducción de un test "oficial" de CI pueden surgir desajustes.

Por ejemplo, "iq_spanish.csv" tiene los resultados de un test de CI traducido del inglés al español. ¿Hay evidencia de que la desviación típica es distinta de 15 y de que, por tanto, debe revisarse la traducción? Usa un nivel de significación del 5%.

1) Discutir Hipótesis

$H_0: \sigma = 15$ → prueba dos colas

$H_a: \sigma \neq 15$



2) Estadístico de contraste:

sd (df$column) → usamos sd ($\hat{S}$) para estimar σ

$\hat{S}$ → ∞ cualquiera:

Varianza muestral $\hat{S}^2$

distrib: $(n-1)\hat{S}^2/\sigma^2 \sim \chi^2_{n-1}$ "Chi-cuadrado"

Reestructurar hipótesis

1) $H_0: \sigma^2 = 15^2$

$H_a: \sigma^2 \neq 15^2$

asumir q H_0 es cierta:

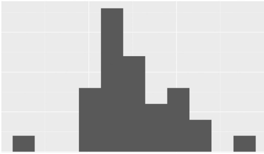
$(n-1) \cdot \hat{S}^2 \sim \chi^2_{n-1}$

$\frac{29 \cdot \hat{S}^2}{15^2} \sim \chi^2_{29}$

2) Estadístico de contraste:

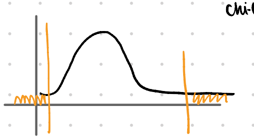
¿Chi-Cuadrado? → Distrib. Normal
→ Población ∞ ✓✓ (todos los españoles)

```r
library(readr)
iq_spanish <- read_csv("data_2/data_2/iq_spanish.csv")

# 1) Hipótesis
  #H0: var = 15^2
  #Ha: var != 15^2

#2) Estadistico de contraste
 sigma =   sd(iq_spanish$iq)
 var_muestral = var(iq_spanish$iq)

 library(ggplot2)
 ggplot(iq_spanish, aes(x=iq)) + geom_histogram(bins = nclass.FD(iq))
 #Distrib Normal + población infinita
```



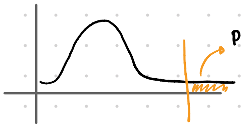
3) Calcular p-valor:

chi-cuadrado: χ^2_{29} (representa H_0)



$\hat{S}$ = var_muestral = 388.7

$\frac{29 \cdot 388.7}{15^2} = 50,1$



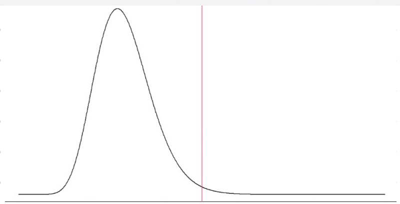
$P(\chi^2_{29} > 50.1) = 1 - P(\chi^2_{29} < 50.1)$

$= \underbrace{1 - \text{pchisq}(50.1,\ df = n-1)}_{\text{Tengo que multiplicarlo por 2}}$ ($n-1$ = 29)

Tema 4: Intervalos de Confianza

Potencia de un Test

```
# 3)
valor_estadistico_datos = 29 * var_muestral/(15^2)

   n = nrow(iq_spanish)
    #valor estadistico en base a datos es 50
   #(irá más o menos en el centro) asiq pondremos como 100 de numero
   eje_x = seq(0,100, by=0.1)
   plot(eje_x, dchisq(eje_x, df = n-1), type = "l")
   abline (v = valor_estadistico_datos , col = 2 )
```



```
  #Tenemos q cuantificar como de raro es:
  pvalor = 2*(1-pchisq(valor_estadistico_datos, df = n-1))

# 4) debo usar un alpha = 0.05%
  alpha = 0.05
  resultado = pvalor<alpha

  ifelse(resultado == TRUE ,
      print("Los datos apoyan la hipotesis alternativa (revisar traducción"),
      print ("Rechazamos la hipotesis alternativa ") )
```

$\alpha = 0.05$	$\alpha = 0.01$
$0.018 < 0.05$	$0.018 > 0.01$
↳ Rechazamos H0	↳ Aceptamos H0
[95% seguros]	[99% seguros]

Para estar totalmente seguros haría un experimento con más muestras

¿Con cuantas? ⇒ Potencia de un test

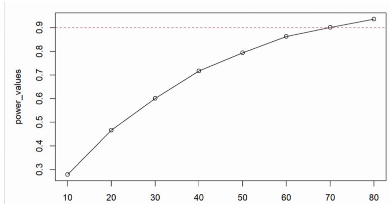
Ejercicio: Potencia de un test

Como $\hat{s}^2 = 388.6902$, escribe una función que calcule la probabilidad de rechazar H_0 si $\sigma^2 = 388.6902$ para un número de muestras n. Completa el siguiente código:

```
power_var_test <- function(n, H0_sigma2 = 15 ^ 2, true_sigma2 = 388.6902,
                           significance = 0.05, N = 5000) {
  sims <- replicate(N, {
    data <- rnorm(n, sd = sqrt(true_sigma2))
    var_stat <- ??
    p_value <- ??
    p_value < significance
  })
}
```

Hemos obtenido: $\sigma = 388.6902$

$H0: \sigma = 15^2$




```
power_var_test <- function(n, H0_sigma2 = 15 ^ 2, true_sigma2 = 388.6902,
                           significance = 0.05, N = 5000) {
  sims = replicate(N, {
    data = rnorm(n, sd = sqrt(true_sigma2)) #[datos donde Ha es cierta]
    var_stat = (n-1) * var(data)/ ( H0_sigma2) # chi-cuadrado
    p_value = 2*(1-pchisq(var_stat, df = n-1))
    p_value < significance #TRUE si rechazo H0 (acepto Ha)
  })
  mean(sims) #P(aceptar Ha | Ha sea cierta (sigma2 = 388.6902))
}

power_var_test(10) #10 muestras --> solo acertaré el 0.27%

#-----------HALLAR NUM DE MUESTRAS--------------------
  #power_var_test(¿?) > 0.9
  power_var_test = Vectorize(power_var_test)
  posibles_num = seq(10,80, by=10)
  power_values = power_var_test(posibles_num)

  plot(posibles_num, power_values, type="o")
  abline(h=0.9, col = 2, lty=2) #línea en y=0.9

  print("Habría que repetir el estudio con unas 70 muestras")
```

Ejercicio: Potencia del T-test: número de muestras

Según la estadísticas oficiales, la media de peso de las mujeres de cierto país es de 63.5 Kg (con desviación típica 4.1). Sin embargo, un equipo de investigadores cree que debido a cambios en la alimentación la media se ha incrementado. ¿Cuántas muestras necesitarán para poder detectar un incremento de medio Kg con un nivel de significación del 1 % y una potencia del 90 %? Usa `power.t.test`

```
     One-sample t test power calculation

              n = 877.9675
          delta = 0.5
             sd = 4.1
      sig.level = 0.01
          power = 0.9
    alternative = one.sided
```

```
power.t.test(delta = 0.5, #diferencia en las medias q queremos detectar
         #(Cuando quier demostrar un cambio más pequeño--> necesito más muestras)
          sd = 4.1,       #desviación típica
          sig.level=0.01, #Nivel de significacion
          power=0.9,      #Detectar esos cambios en el 90% de los casos
          type = "one.sample",  # "one.sample" solo está haciendo referencia a una población
          alternative ="one.sided" ) #Ha de una o dos colas (una cola, Ha = mu > 63.5)
```

```
[1] "nuestro estudio deberia correrse con 878 muestras"
```