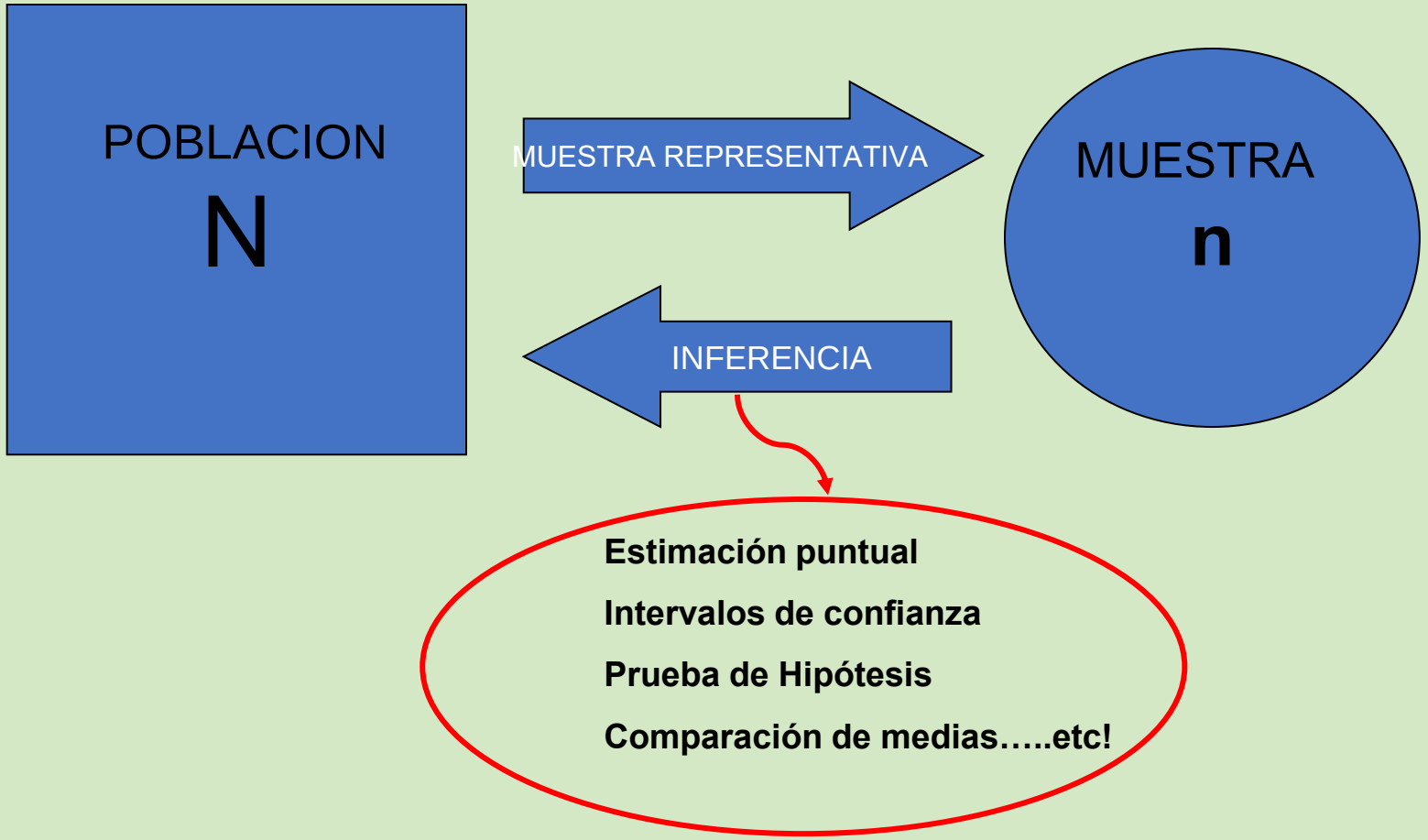
Fundamentos de Estadística

Silvia N. Pérez
Especialización en Ciencia de Datos - UNO

INFERENCIA ESTADISTICA



INFERENCIA ESTADISTICA BASICA

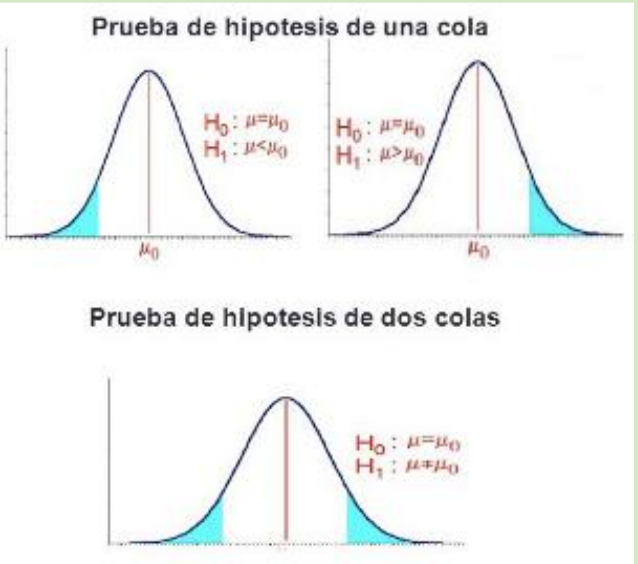
INFERENCIA ESTADISTICA

Estimación de parámetros

Pruebas de hipótesis

Estimación puntual

Intervalos de confianza



Estimación de parámetros

Un parámetro es un valor constante que caracteriza a una población.

Cuando definimos algunas distribuciones de interés, presentamos:

- Binomial (n,p)
- Poisson (μ)
- Normal (μ, σ)
- etc

Por ejemplo, si la distribución es normal, los parámetros de interés podrían ser la esperanza y la varianza, ya que para especificar completamente la distribución es necesario conocer estos dos valores.

Para estimarlos, tomaremos una muestra aleatoria de valores de la variable y los usaremos para construir un valor que consideramos “aproxima” el parámetro de interés.

Estimación de parámetros

Algunos parámetros de interés: alturas medias de habitantes de una región, proporción de votantes a favor de un cierto candidato, etc.

Los parámetros se estiman a partir de estadísticos (variables aleatorias observadas en una muestra).

Definición: Estimador puntual

Dada una variable X cuya distribución depende de un parámetro Θ, y dada una muestra aleatoria X_1, X_2,… X_n, **un estimador de Θ** es una variable aleatoria que sea función de la muestra y que tome valores en el conjunto de valores posibles de Θ.

Por ejemplo:

- si queremos estimar la media μ de una población, utilizamos la media de una muestra: $\bar{X}$

- Si queremos estimar la proporción de voto a un candidato, usamos $\hat{p}$ = #casos de intención de voto / total de muestra

Estimadores más usados

1. Para inferir sobre la media de una población, μ, generalmente se utiliza el estimador $\bar{X}$ por cumplir con las propiedades de un buen estimador.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$\bar{X}$ es una V.A porque depende de las V.A $X_1, \ldots X_n$

- ✓ Si la muestra es pequeña y X puede suponerse normal, $\bar{X}$ sigue distribución normal: $\bar{X} \sim \mathcal{N}(\mu\,; \frac{\sigma}{\sqrt{n}})$
- ✓ Si la muestra es grande, $\bar{X}$ sigue distribución aproximadamente normal aunque las variables de origen no sean normales.

↳ IMPORTANTE P/ INTERVALO de Conf d Tcu 4

- ✓ PROPIEDAD IMPORTANTE: Ley de los grandes números

" $\bar{X} \rightarrow \mu$ cuando n→infinito "…. ¿en qué sentido?

Estimadores más usados

2. Para inferir sobre la varianza de una población, σ^2, se utiliza el estimador S^2.

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

dividiendo por $n-1$ tiene propiedad de ser un "alrededor" de σ^2 verdadero

3. Para inferir sobre una proporción, se utiliza la proporción muestral:

$$\hat{p} = \frac{\#\ éxitos\ en\ la\ muestra}{n}$$

$X \sim N(\mu, \sigma^2) \Rightarrow E(S^2) = E\left(\frac{\Sigma (x_i - \bar{x})^2}{n-1}\right) = \cdots = \sigma^2$

Estimación por intervalo de confianza

Los estimadores puntuales son también variables aleatorias y, por lo tanto, no se puede esperar que en una realización cualesquiera den un valor idéntico al parámetro que estiman.

Se busca entonces que una estimación puntual esté acompañada de alguna medida del posible error de esa estimación. Esto puede hacerse indicando el error estándar del estimador o dando un intervalo que incluya al verdadero valor del parámetro con un cierto nivel de confianza.

Intervalo de confianza para un parámetro

- Es un intervalo [LI, LS] tal que si el parámetro a estimar es θ, entonces:

$$P(LI \leq \theta \leq LS) = 1-\alpha$$

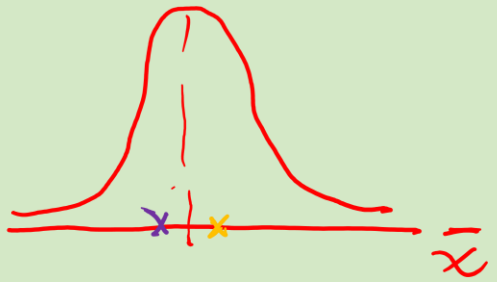
V. A.

- Se lee: “el intervalo de límites aleatorios LI y LS tiene probabilidad (1-α) de contener al parámetro θ”, donde (1-α) denota la confianza de la estimación y se denomina **coeficiente de confianza**.
- Una vez que calculamos los valores observados en la muestra, el intervalo numérico observado se dice que tiene una confianza (1-α) de contener al verdadero valor del parámetro θ.

Intervalos de confianza para la media en X ~ N(μ, σ)

Dada una muestra aleatoria $X_1, X_2, \ldots X_n, \ldots$ se debe encontrar una v. a. que me permita construir LI, LS.

- Si la varianza es conocida, se usa $\bar{X} \sim \mathcal{N}(\mu\ ; \frac{\sigma}{\sqrt{n}})$



→ el intervalo queda : $\left[\bar{X} - z * \frac{\sigma}{\sqrt{n}}\ ; \bar{X} + z * \frac{\sigma}{\sqrt{n}}\right]$

$P(LI < \mu \leq LS) = 1-\alpha$

donde z es el cuantil N(0,1) que deja en el centro una prob $1-\alpha$

- Si la varianza es desconocida, se usa

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

IC para μ cuando X ~ N(μ, σ) y σ es desconocida

Resulta como:

$$I.C.para\ \mu: \left[\bar{X} - t * \frac{S}{\sqrt{n}} ; \bar{X} + t * \frac{S}{\sqrt{n}}\right]$$

desviomuestral

Error estandar

error

Donde:

t = cuantil de la distribución t-Student(n-1)

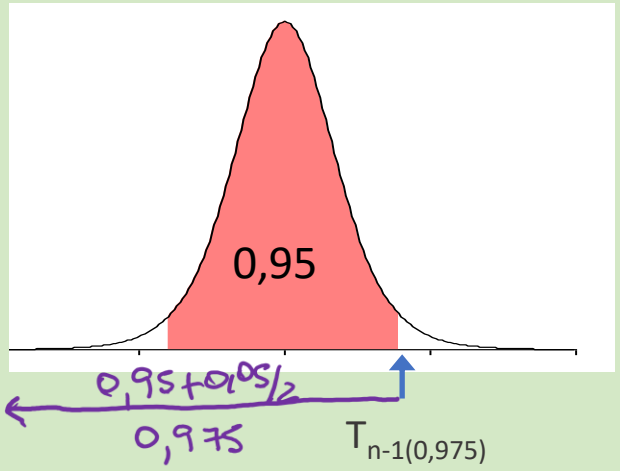
n = tamaño de la muestra.

$\bar{X}$ = promedio muestral.

S = desvío muestral.

Esto supone una distribución Normal de X!!

Hay que verificarlo....



Ejemplo:

Se tomó una muestra de alturas de 500 personas de 30 años de edad. Se supone que la altura tiene distribución normal.

X: altura de persona de 30 años
$X \sim N(\mu, \sigma)$

- A partir de la muestra se calculan los valores:

$\bar{x}$ = 1,7074 mts. ; s = 0,29138

$\bar{x}$: valor obs de $\bar{X}$; s: desvío muestral

y así se determina un intervalo para la verdadera altura media con una confianza del 95%:

I.C. para μ : [1,6859 ; 1,7257]

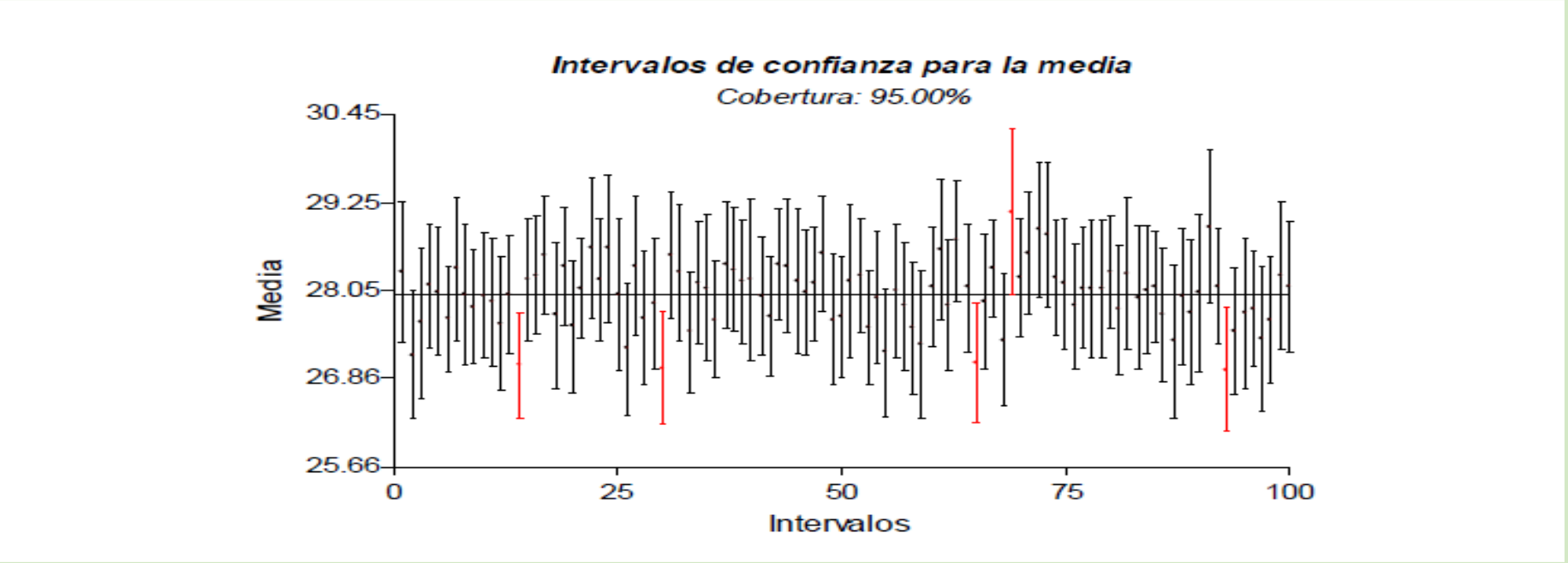
t = cuantil q' deja en cuto 0,95
$t = 1,96\ldots$

- Interpretación:

Se tiene una confianza del 95% de que el intervalo [1,6589 y 1,7257] contiene a la altura media de los habitantes de 30 años de edad de esta población.

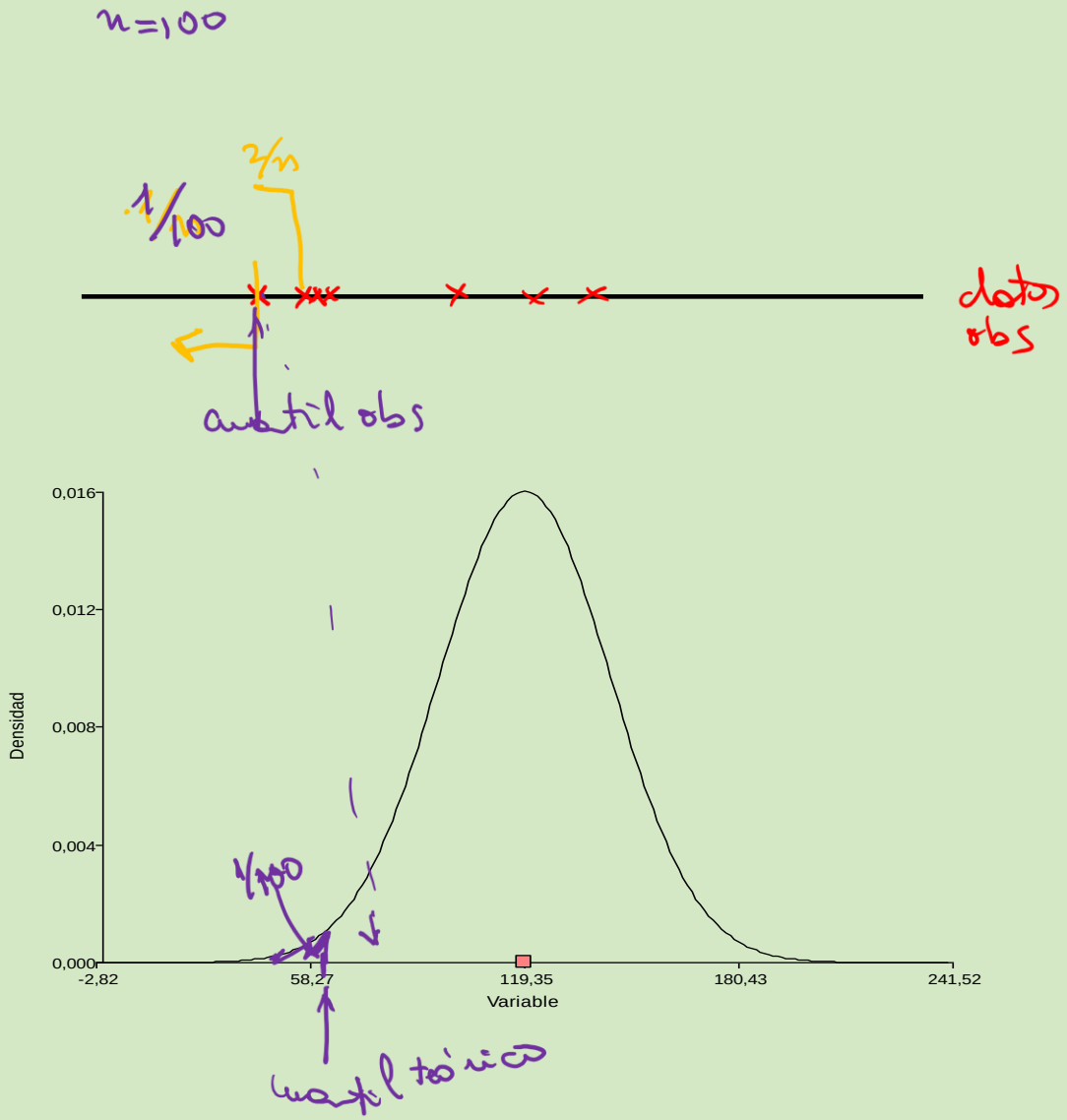
¿Porqué hablamos de “confianza” de un intervalo?

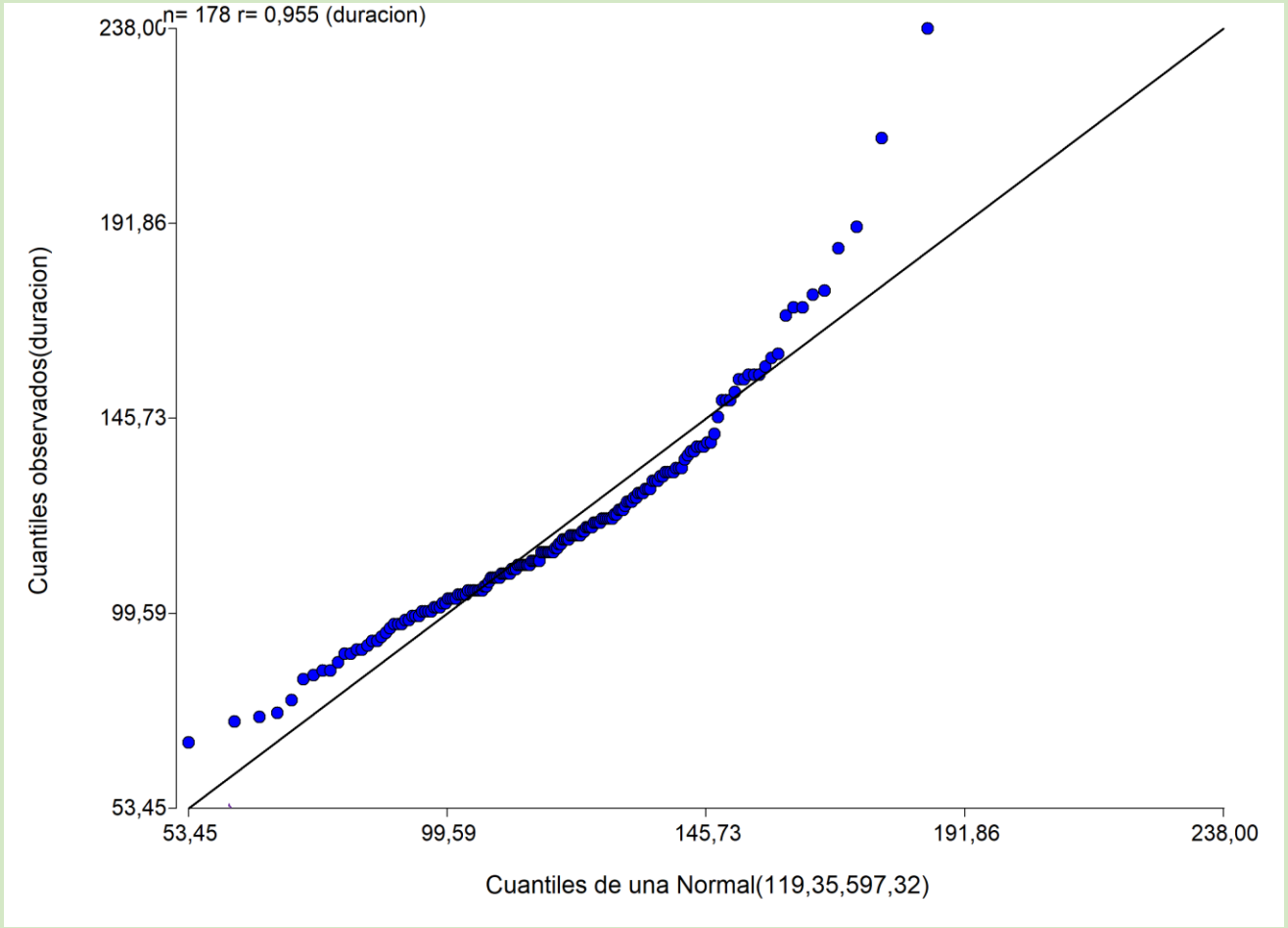
Si de una población con μ=28, se toman 100 muestras de tamaño ***n*** y se construyen para cada una un intervalo de confianza del 95%, entonces ocurrirá que aproximadamente 95 de los 100 intervalos incluyan al valor 28 y que 5 intervalos no lo incluyan, como puede verse en:



Entonces, tendremos CONFIANZA de que UN intervalo en particular que obtuvimos “sea de los buenos”, porque hay muchos que lo son!

Para ver normalidad (1er vistazo): QQplot





Ejemplo: estudiantes

- Hallar estimadores para la media y la varianza de las variables nota y edad.
- Dar intervalos de confianza para la nota media, con niveles 90%, 95% y 99%. ¿En qué se diferencian?
- Hallar intervalos de confianza para la nota según “Colegio”.

Vamos a R!!

IC para σ^2 cuando X ~ N(μ, σ)

Obs: p/armar un IC para un parámetro necesito

m.a. → una v.A q' estime el parám → estadístico q' tenga una distrib bien conocida

$\mu \rightarrow \bar{x} \rightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t$

$\sigma^2 \rightarrow S^2 \rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

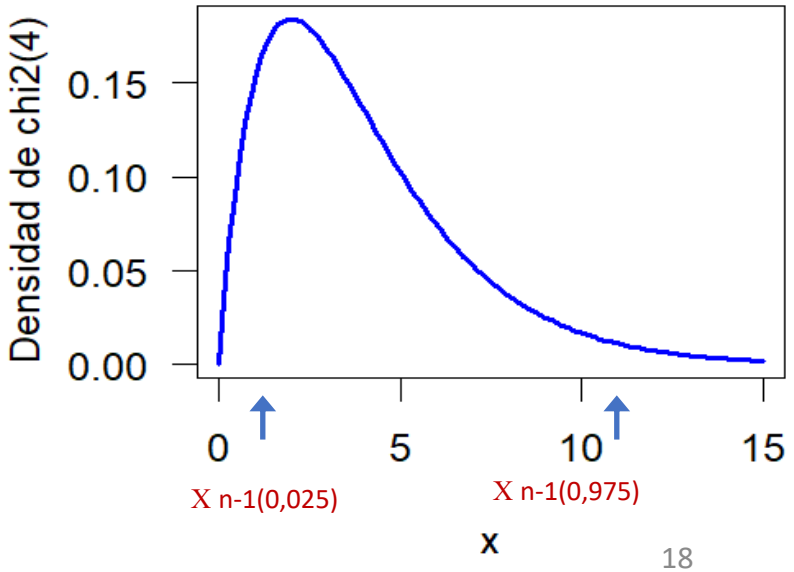
Resulta como:

$$I.C.del\ 95\%\ para\ \boldsymbol{\sigma^2}: \left[\frac{(n-1)\,S^2}{\chi^2_{n-1,\,sup}} ; \frac{(n-1)\,S^2}{\chi^2_{n-1,\,inf}}\right]$$

cuantiles de χ^2_{n-1}

Esto supone una distribución Normal de X!!

Hay que verificarlo….



Intervalos para una proporción

Si queremos estimar la proporción poblacional, un intervalo de confianza asintótico es:

Intervalo de confianza para p:

Límite inferior = $\hat{\boldsymbol{p}} - z. \sqrt{\frac{\hat{\boldsymbol{p}}.(1-\hat{\boldsymbol{p}})}{n}}$

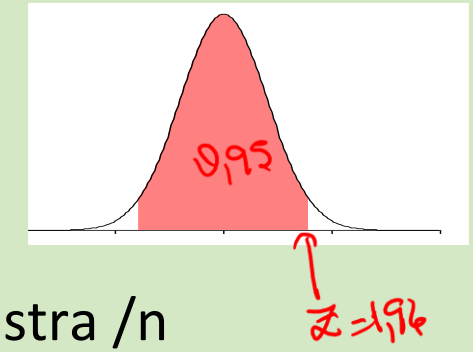
Límite superior = $\hat{\boldsymbol{p}} + z. \sqrt{\frac{\hat{\boldsymbol{p}}.(1-\hat{\boldsymbol{p}})}{n}}$

Donde:

z = cuantil de la distribución Normal(0,1)

n = tamaño de la muestra.

$\hat{p}$= proporción muestral = #casos en la muestra /n



Intervalos de confianza para diferencia de medias de poblaciones independientes y normales

$$[(\overline{X}_1 - \overline{X}_2) \pm t_{df,1-\alpha/2} * \sqrt{\frac{\widehat{S}_1^2}{n_1} + \frac{\widehat{S}_2^2}{n_2}}]$$

df= n1+n2-2 ó min(n1-1, n2-1) ó…

Se debe cumplir la condición de normalidad de ambas poblaciones Y homogeneidad de varianzas.

SI esto último no se cumple, se usa el test de Welch.

Intervalos de confianza para diferencia de medias apareadas

Usamos que

$$D=X-Y \sim N(\mu, \sigma)$$

Con lo que el análisis es igual que para IC para UNA media.

Se debe cumplir la condición de normalidad!

Veamos en R!!