

Temas 7-8-9

Inferencia Estadística

Ingeniería del software

Temario

Distribuciones de muestreo fundamentales

7.1. Modelos bidimensionales. Distribución conjunta

7.2. Distribución Normal multivariante.

7.3. Muestreo aleatorio.

7.4. El Teorema del límite central

7.5. Distribuciones asociadas a poblaciones normales

- Distribución X^2 de Pearson
- Distribución t de Student
- Distribución F de Snedecor

7.6. Distribuciones de estadísticos en el muestreo

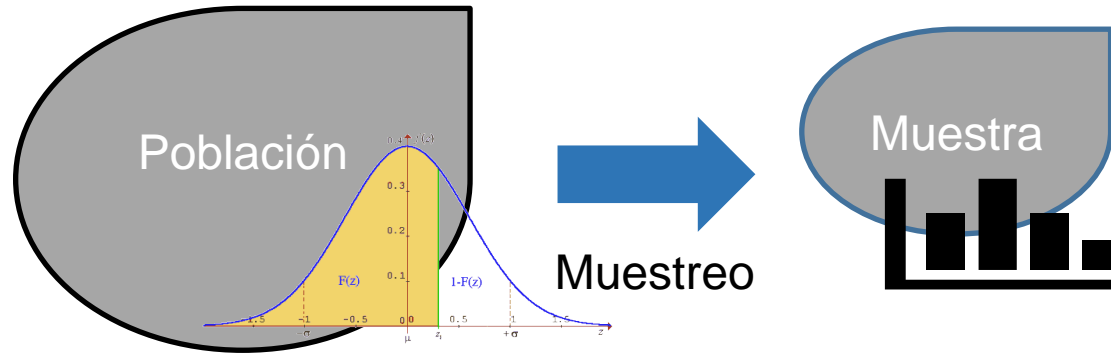
7.5.1. Distribución muestral de medias

7.5.2. Distribución muestral de varianzas

7.5.3. Distribución muestral de proporciones

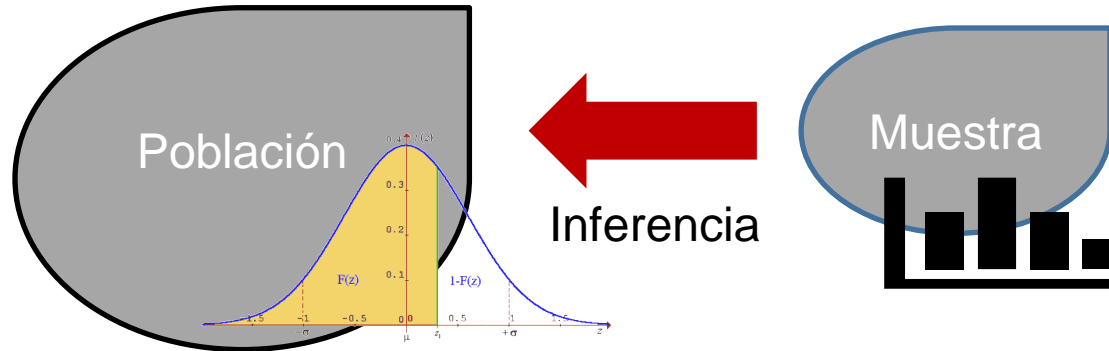
Introducción

En los últimos temas hemos conseguido establecer un marco matemático ajustando el conocimiento sobre fenómenos que tienen distribuciones conocidas. De esta forma, si conocemos el marco probabilístico de una población, seremos pues capaces de calcular probabilidades sobre la muestra aleatoria que podamos tomar.



Introducción

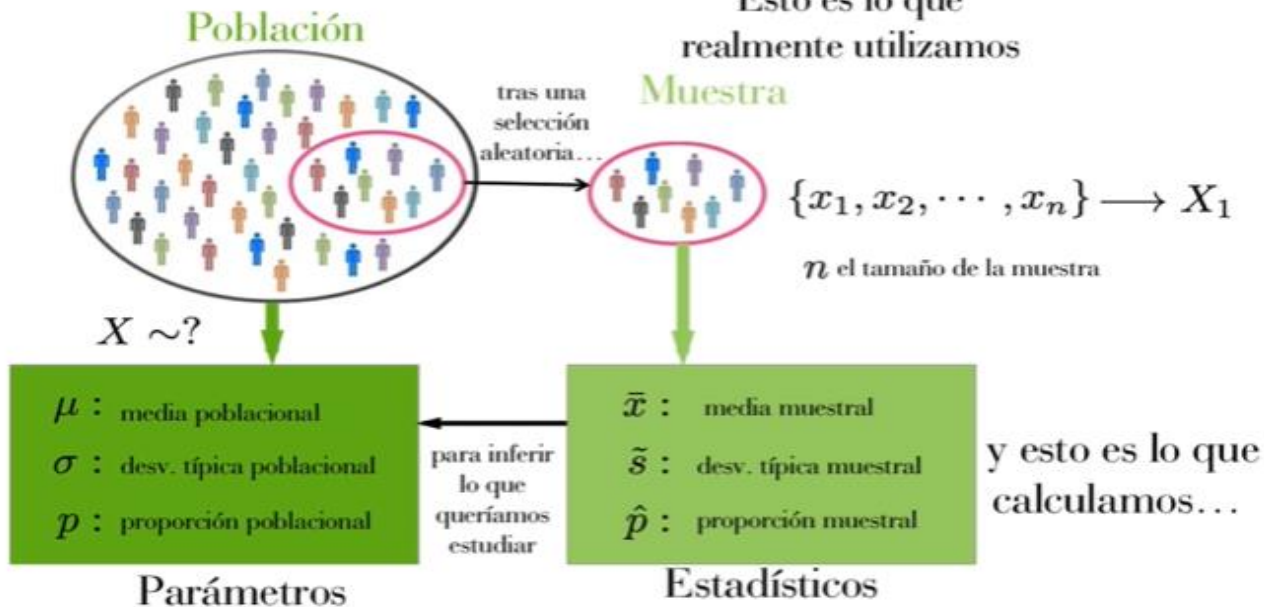
En estos temas que vienen a continuación, el proceso será inverso. Partiremos de información muestral (Temas 1 y 2) y trataremos de **inferir** información relevante de la posible distribución que tiene la población (Temas 3, 4 y 5).



Ejemplos:

- *Información de la muestra:* Una máquina expendedora de bebidas está sirviendo bebidas con un contenido en torno a 240 cc → *Lo que se quiere inferir:* Que la media poblacional es de 240 cc.
- *Información de la muestra:* se toman muestras del contenido de alquitrán de dos marcas de cigarrillos y se obtienen sus respectivas medias muestrales → *Lo que se quiere inferir:* Que el contenido medio en alquitrán de ambas marcas difieren.

¿Qué queremos estudiar?



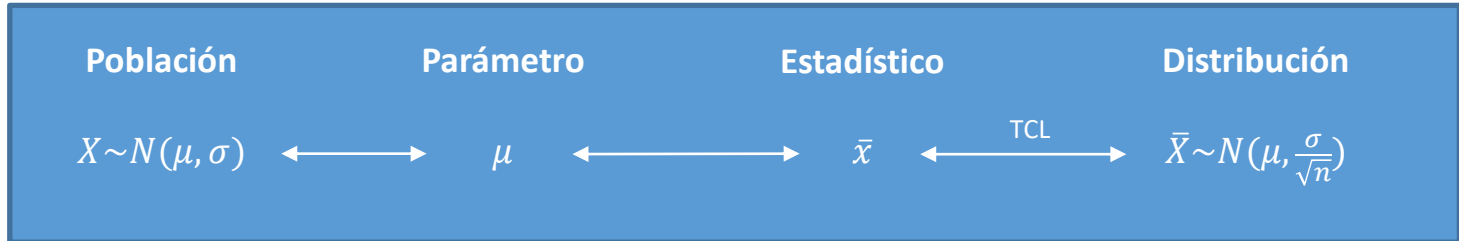
Distribuciones en el muestreo

Mar Angulo Martínez
mar.angulo@u-tad.com

Estadísticos

Parámetro (poblacional) es un número real constante, no aleatorio y único que describe una característica de la población ($\mu, \sigma^2, \lambda, \dots$).

Estadístico (muestral) es una función de la muestra (\bar{x}, s^2, \dots); es por tanto, una variable aleatoria que tiene su propia distribución.



- ☐ **Parámetro** (poblacional) es un número real constante, no aleatorio y único que describe una característica de la población. Problema: generalmente desconocido

- ☐ **Estadístico** (muestral) es una función de la muestra; es por tanto una variable aleatoria
- ☐ El valor de un estadístico cambia de una muestra a otra
- ☐ La distribución muestral es la distribución de probabilidad de todos los posibles valores del estadístico muestral
- ☐ La **distribución de un estadístico** depende de la distribución de la población, del tamaño de la muestra y del método de selección de las muestras

Estadísticos

En distribuciones normales, la media muestral \bar{x} es buen estimador de la media poblacional μ :

- Por una parte es insesgado,

$$E[\bar{x}] = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n} \cdot E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n} \cdot \sum_{i=1}^n E[X_i] = \frac{1}{n} \cdot \sum_{i=1}^n E[X] = \frac{1}{n} \cdot n\mu = \mu$$

- Y por la otra, tiene varianza mínima:

$$Var(\bar{x}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2} \cdot Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \cdot \sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n} \xrightarrow{n \rightarrow +\infty} 0$$

Estadísticos

Sin embargo, la varianza muestral s^2 **no** es el mejor estimador de la varianza poblacional σ^2 :

- Porque es sesgado,

$$\begin{aligned} E[s^2] &= E\left[\left(\frac{\sum_{i=1}^n X_i^2}{n}\right) - \bar{X}^2\right] = \frac{1}{n} \cdot \sum_{i=1}^n E[X_i^2] - E[\bar{X}^2] = \\ &= \frac{1}{n} \cdot \sum_{i=1}^n (Var(X_i) + E[X_i]^2) - (Var(\bar{X}) + E[\bar{X}]^2) = \frac{1}{n} \cdot n(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n} \sigma^2 \end{aligned}$$

Definimos la **cuasivarianza muestral** como:

$$S^2 = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Estadísticos

En este caso, la cuasivarianza muestral S^2 es el mejor estimador de la varianza poblacional σ^2 :

- Porque es insesgado,

$$E[S^2] = E\left[\frac{n}{n-1} s^2\right] = \frac{n}{n-1} E[s^2] = \sigma^2$$

- Y tiene varianza mínima,

$$Var(S^2) = \frac{1}{(n-1)^2} \cdot Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{(n-1)^2} \cdot \sum_{i=1}^n Var(X_i) = \frac{n \cdot \sigma^2}{(n-1)^2} \xrightarrow{n \rightarrow +\infty} 0$$

Para muestras grandes, la cuasivarianza es una buena aproximación de la varianza muestral:

$$S^2 = \frac{n}{n-1} s^2 \xrightarrow{n \rightarrow +\infty} s^2 \text{ (en la práctica, si } n \geq 30 \rightarrow \text{ las podemos considerar iguales } S^2 \approx s^2)$$

Estadísticos

Parámetros

Estadísticos

μ	→	$\bar{x} = \frac{X_1 + X_2 + \dots + X_n}{n}$	(Media muestral)
σ^2	→	$s^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}$	(Cuasivarianza muestral)
π	→	$p = \frac{X_1 + X_2 + \dots + X_n}{n}$	(proporción muestral)



Estadísticos

Un estadístico es una función de la muestra



¿cuál es su distribución de probabilidad?

Distribuciones de estadísticos en el muestreo

- ❑ Las distribuciones muestrales de \bar{X} y S^2 son los mecanismos a partir de los cuales podremos hacer inferencias sobre μ y σ^2 .

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

- ✓ Es la distribución que resulta cuando un experimento se lleva a cabo una y otra vez
- ✓ Describe la variabilidad de los promedios muestrales en torno al valor μ

Si en una población $X \rightarrow N(\mu, \sigma)$ tomamos una muestra aleatoria de n observaciones

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- ✓ Porque la combinación lineal de variables Normales sigue una distribución Normal

Distribuciones de estadísticos en el muestreo

➤ En la práctica...

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$$

❖ Ejemplo 1

El tiempo que requiere un cierto tipo de rata para encontrar la salida de un laberinto es una variable normalmente distribuida de media 1,5m y desviación típica 0,35m. En un grupo de 5 ratas, ¿cuál es la probabilidad de que el tiempo medio que tardan en salir no exceda los 2 m?

- Si $X \equiv$ tiempo que tarda una rata en encontrar la salida $\rightarrow N(1,5; 0,35)$
- $\bar{X} =$ tiempo medio muestral $\rightarrow N(1,5, \frac{0,35}{\sqrt{5}}) = N(1,5; 0,157)$
- $p(\bar{X} \leq 2) = p(Z \leq \frac{2-1,5}{0,157}) = p(Z \leq 3,18) = 0,9993$

El teorema del límite central

- ¿Y si \bar{X} es la media de una muestra aleatoria de tamaño n , tomada de una *población* **no necesariamente normal**?

Teorema central del límite (T.C.L.)

- ❑ Si X_1, \dots, X_n son variables independientes, idénticamente distribuidas (i.i.d.) con media μ y varianza σ^2 . Si n es suficientemente grande, entonces

$$\bar{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \iff \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

- ❑ La aproximación normal es buena si $n \geq 30$ (mejor cuanto mayor sea n) y siempre que la distribución de la población no sea muy asimétrica.

<https://www.youtube.com/watch?v=29z2G6OUa68>

❖ Ejemplo 2

El nº de defectos en un nuevo modelo de automóvil es una variable aleatoria con media 3,2 y desviación estándar de 2,4. Entre 100 vehículos seleccionados al azar, ¿cuál es la probabilidad de encontrar una media muestral de más de 4 defectos?

- $X \equiv$ número de defectos en modelo de automóvil: distribución desconocida con media 3,2 y d.t. 2,4
- $\bar{X} =$ número medio de defectos en una muestra de 100 vehículos
- $\bar{X} \rightarrow N(3,2, \frac{2,4}{\sqrt{100}}) = N(3,2; 0,24)$
- $p(\bar{X} > 4) = p(Z > \frac{4-3,2}{0,24}) = p(Z > 3,33) \approx 0,0004$

❖ Ejemplo 3

Queremos estimar el tiempo medio de viaje en un autobús especial para ir de un campus de una Universidad al campus de otra dentro una ciudad. Sabemos que el tiempo de viaje sigue una distribución normal con desviación típica de 5 minutos.

En cierta semana un autobús hizo el viaje 40 veces y el tiempo medio de viaje fue de 28 minutos.

- a) Obtener un intervalo al 95% de confianza para el tiempo medio de viaje. ¿Y al 99%?
- b) ¿Cuál es el error máximo de estimación en cada uno de los casos?
- c) Calcular el número de viajes que habrá que realizar para estimar el tiempo promedio real de viaje con un error menor a 1 minuto y una confianza del 99%.

X = tiempo (en minutos que el autobús tarda en ir de un campus a otro) $\rightarrow N(\mu, 5)$
¿cuál es la distribución muestral de \bar{X} (tiempo promedio muestral que tarda el autobús)

Recuerda:

Si en una población $X \sim N(\mu, \sigma)$ tomamos una muestra aleatoria de n observaciones,

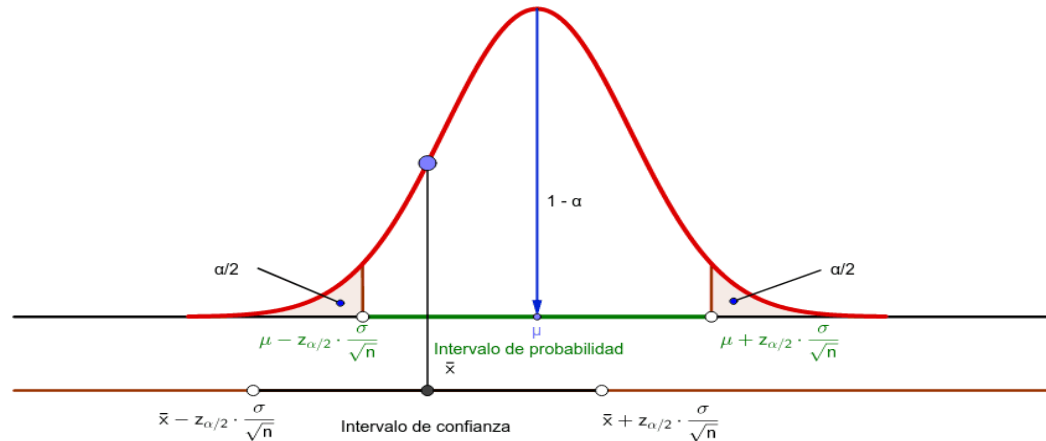
$$\bar{X} \rightarrow N(\mu, \sigma/\sqrt{n}) \longleftrightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

a) IC (95% de confianza): $\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(28 - 1,96 \cdot \frac{5}{\sqrt{40}}, 28 + 1,96 \cdot \frac{5}{\sqrt{40}}\right) = (28 \pm 1,55) = (26,45; 29,55)$

b) IC (99% de confianza): $\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(28 - 2,575 \cdot \frac{5}{\sqrt{40}}, 28 + 2,575 \cdot \frac{5}{\sqrt{40}}\right) = (28 \pm 2,036) = (25,964; 30,036)$

El error máximo de estimación es mayor en el segundo intervalo que en el primero

$$c) Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 1 \implies 2,575 \cdot \frac{5}{1} \leq \sqrt{n} \implies n \geq 165,77 \implies n = 166 \text{ viajes}$$



❑ ¿y si la desviación típica σ es desconocida?

1) Si en una población $X \rightarrow N(\mu, \sigma)$ tomamos una muestra aleatoria de n observaciones, y σ es desconocida, pero n grande ($n \geq 40$)

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

2) Si σ es desconocida, pero n pequeña ($n < 40$)

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

❖ Ejemplo 4

Se obtienen las calificaciones en Matemáticas de una muestra aleatoria de 500 estudiantes. La media y la cuasi desviación típica muestrales han sido respectivamente de 501 y 112 puntos. Las calificaciones siguen una distribución Normal.

- Obtener un intervalo de confianza para la puntuación media poblacional (Nivel de significación 95%). Dar una medida del error máximo de estimación
- Realizar el mismo cálculo en otro centro se toma una muestra de sólo 25 estudiantes que obtienen la misma puntuación media pero con una $s = 12$ puntos. Obtener ahora intervalo al 99% de confianza.
- ¿A cuántos estudiantes habrá que analizar si $s = 12$ y queremos estimar la puntuación media con un nivel de confianza del 99% y un error máximo de estimación de 3 puntos?

- $X \equiv$ calificaciones en Matemáticas: distribución desconocida con media μ y d.t. σ
- \bar{X} = puntuación media en una muestra de 500 estudiantes = 501 puntos
- s = cuasi desviación típica en una muestra de 500 estudiantes = 112 puntos

a) IC (95% de confianza): $\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right) = \left(501 - 1,96 \cdot \frac{112}{\sqrt{500}}, 501 + 1,96 \cdot \frac{112}{\sqrt{500}} \right) = (501 \pm 9,817) = (491,183; 510,817)$

Error máximo de estimación: 9,817 puntos

b) IC (99% de confianza): $\left(\bar{x} - t_{24, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{24, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right) = \left(501 - 2,797 \cdot \frac{12}{\sqrt{25}}, 501 + 2,797 \cdot \frac{12}{\sqrt{25}} \right) = (501 \pm 6,71) = (494,29; 507,71)$

▪ c) $Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq 3 \implies 2,575x \frac{12}{\sqrt{n}} \leq 3 \implies n \geq 106,09 \implies n = 107$

□ Distribución de la varianza muestral: s^2

$$\text{Varianza muestral } S^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^k x_i^2 \cdot n_i}{n} - \bar{x}^2$$

$$\text{Cuasivarianza muestral } s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{n-1}$$

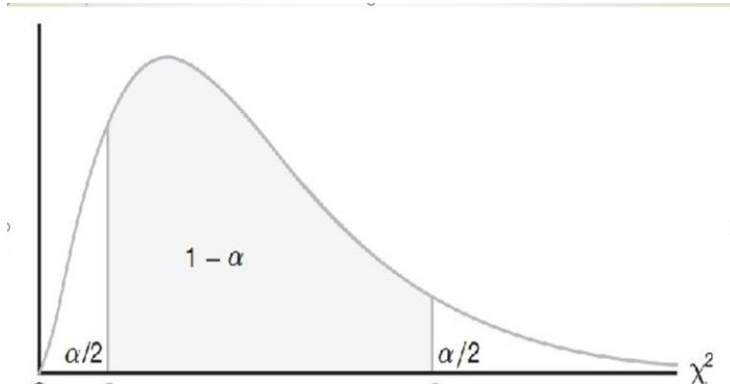
$$\frac{(n-1)s^2}{\sigma^2} \rightarrow \chi^2 \text{ con } n-1 \text{ grados de libertad}$$

❖ Ejemplo 5

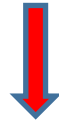
Los datos sobre voltaje de ruptura de circuitos eléctricamente sobrecargados siguen una distribución normal.

- En una muestra de 17 circuitos se ha obtenido una cuasivarianza muestral de 137.324,3. Obtener un intervalo de confianza al 95% para la varianza del voltaje de ruptura en circuitos sobrecargados
- Obtener ahora un intervalo de confianza al 99% para dicha varianza
- Dar una estimación puntual de σ^2

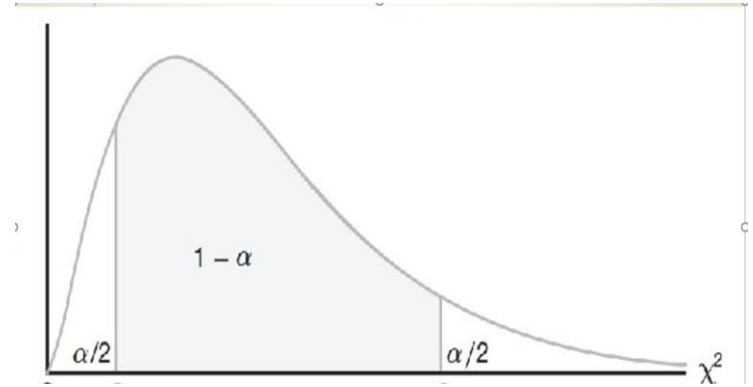
- $X \equiv$ voltaje de ruptura de los circuitos $\rightarrow N(\mu; \sigma)$
- $s^2 =$ cuasivarianza muestral = 137.324,3 $n=17$
- $\frac{(n-1)s^2}{\sigma^2} \rightarrow \chi^2_{16}$
- $p(a < \frac{(n-1)s^2}{\sigma^2} < b) = 0,95$
- Encontramos en la tabla los valores $a = \chi^2_{16;0,025} = 6,91$ y $b = \chi^2_{16;0,975} = 28,8$
- IC: $(\frac{(n-1)s^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1;\frac{\alpha}{2}}}) = (\frac{16 \times 137.324,3}{28,8}, \frac{16 \times 137.324,3}{6,91}) = (76.291,28; 317.972,33)$
- Encontramos en la tabla los valores $a = \chi^2_{16;0,005} = 5,14$ y $b = \chi^2_{16;0,995} = 34,27$
- $p(5,14 < \frac{(n-1)s^2}{\sigma^2} < 34,27) = 0,99 = p(\frac{(n-1)s^2}{34,27} < \sigma^2 < \frac{(n-1)s^2}{5,14}) = p(64.114,06 < \sigma^2 < 427.468,64) = 0,99$
- d) Estimación puntual para σ^2 : $\hat{\sigma}^2 = s^2 = 137.324,3$ voltios²



$$a = \chi^2_{16;0,025} = 6,91$$



$$b = \chi^2_{16;0,975} = 28,8$$



$$a = \chi^2_{16;0,005} = 5,14$$



$$b = \chi^2_{16;0,995} = 34,27$$

□ Distribución de la proporción muestral: p

$X \rightarrow B(n, \pi)$ cuando $n \rightarrow \infty$

la variable

y estimando π por $p = \frac{X}{n}$

$$\frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

$\rightarrow N(0,1)$

$$\frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}$$

$\rightarrow N(0,1)$

- π es la proporción de individuos que cumplen una característica en la población
- p es la proporción de individuos que cumplen una característica en la muestra



❖ Ejemplo 6

Se selecciona una muestra aleatoria de 200 personas en una ciudad y se encuentra que 114 apoyan una decisión del Ayuntamiento.

- a) Calcular entre qué dos valores está comprendida la proporción de “a favor” de toda la población con una probabilidad de 0,95.
- b) ¿Qué tamaño debería tener una muestra si queremos que con un 95% de confianza nuestra estimación difiera de la proporción poblacional verdadera como mucho en un 1%?

- $X \equiv$ número de personas que están a favor de la decisión $\rightarrow B(n; \pi)$
- $\pi =$ proporción poblacional de personas que están a favor de la decisión
- $p \equiv$ proporción muestral de personas que están a favor de la decisión $= \frac{114}{200} = 0,57$
- $p\left(a < \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} < b\right) = 0,95$ en la tabla $N(0,1)$ los valores $a = Z_{0,025} = -1,96$ y $b = Z_{0,975} =$
$$0,95 = p\left(p - 1,96\sqrt{\frac{p(1-p)}{n}} < \pi < 0,57 + 1,96\sqrt{\frac{p(1-p)}{n}}\right)$$

$$= p(0,57 - 1,96 \times 0,035 < \pi < 0,57 + 1,96 \times 0,035) = (0,57 \pm 0,0686) = (0,5014; 0,6386)$$
- Con un 95% de confianza la proporción de personas favorables a la decisión está comprendida entre el 50,14% y el 63,86%. Error máximo de estimación: 6,86%
- b) $1,96\sqrt{\frac{p(1-p)}{n}} \leq 0,01 \rightarrow 1,96\sqrt{\frac{0,57(1-0,57)}{n}} \leq 0,01 \rightarrow 9.415,76 \leq n$
- Hemos de tomar como mínimo 9.416 personas en la muestra

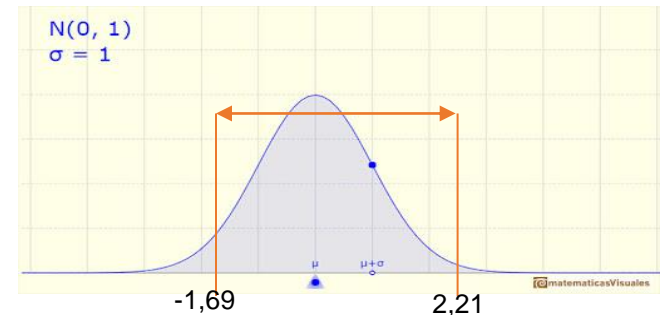
Algunos ejemplos más de distribuciones en el muestreo...

❖ Ejemplo 7

El incremento anual de los altos cargos en España se distribuye normalmente con media 8,3 % y desviación típica 2,3 %. Se toma una muestra aleatoria simple de nueve observaciones.

- a) Calcular la probabilidad de que el incremento medio muestral esté comprendido entre el 7% y el 10%
b) ¿y cuál es la probabilidad de que el incremento medio no supere el 8%? ¿y de que sea exactamente del 8%?

- Si $X \equiv$ incremento de salarios de altos cargos en España $\rightarrow N(8,3; 2,3)$
- $\bar{X} =$ incremento medio muestral $\rightarrow N(8,3, \frac{2,3}{\sqrt{9}}) = N(8,3; 0,77)$
- $p(7 \leq \bar{X} \leq 10) = p(\frac{7-8,3}{0,77} \leq Z \leq \frac{10-8,3}{0,77}) = p(-1,69 \leq Z \leq 2,21) = 0,9864 - 0,0455 = 0,9409$
- $p(\bar{X} \leq 8) = p(Z \leq \frac{8-8,3}{0,77}) = p(Z \leq -0,39) = 0,3483$



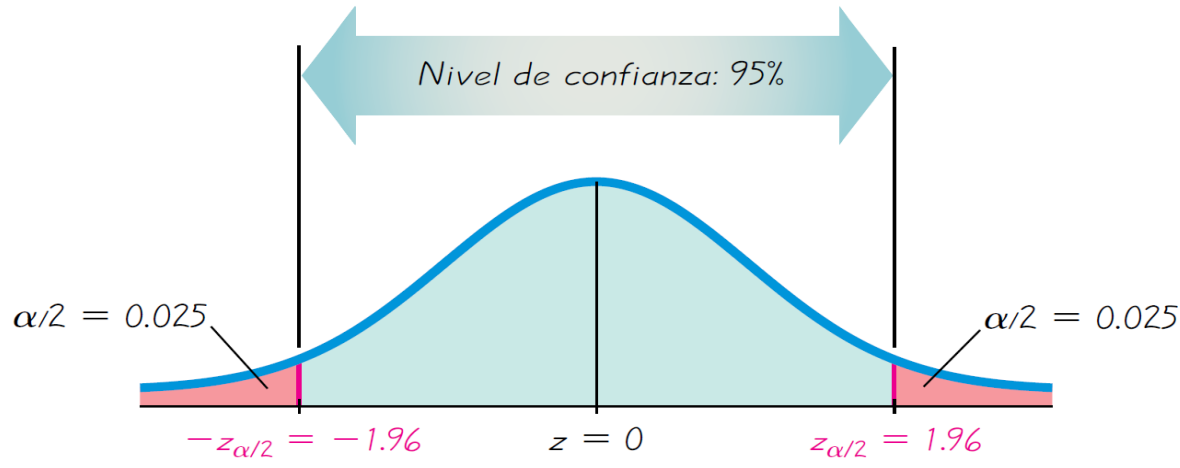
❖ Ejemplo 8

Expertos en TV afirman que un 65% de los hombres con edades comprendidas entre 30 y 70 años ven el partido de liga el sábado por TV. Se toma una muestra aleatoria simple de 200 individuos de esa población. Y se comprueba que han visto el partido 136

- a) Si la cifra de los expertos es cierta, ¿cuál es la probabilidad de que más del 66% de los hombres encuestados vea el partido?
- b) ¿Entre qué dos valores estimarías el porcentaje real de personas que ven el partido del sábado? (Nivel de significación 0,95 ¿cuál es el error máximo de estimación?
- c) ¿y si aumentas el nivel de significación a 0,99? ¿en qué sentido cambian los valores obtenidos en el apartado anterior?
- d) Si se decide reducir ahora el error máximo de estimación a un 3% reduciendo la confianza al 90%, ¿qué tamaño mínimo deberá tener la muestra?

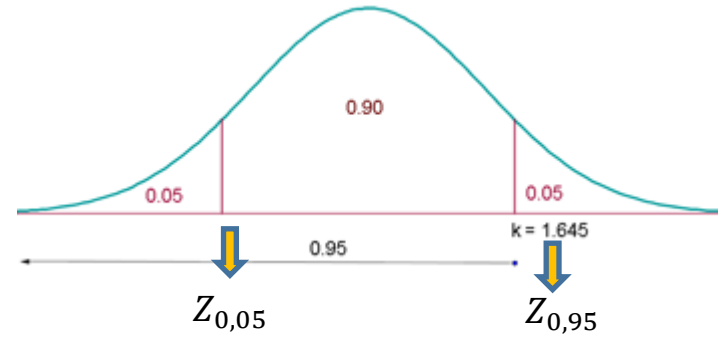
- $X \equiv$ número de hombres que ven en TV el partido de liga del sábado $\rightarrow B(n; \pi)$
- $\pi =$ proporción poblacional de hombres que ven el partido
- $p \equiv$ proporción muestral de hombres que ven el partido $= \frac{136}{200} = 0,68$
- a) $p(p > 0,66) = p\left(\frac{\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}}{\frac{0,66-0,65}{\sqrt{\frac{0,65(1-0,65)}{200}}}} > \frac{0,66-0,65}{\sqrt{\frac{0,65(1-0,65)}{200}}}\right) = p(Z > 0,3) = 1-0,6179 = 0,3821$

- b) 0,95 en la tabla $N(0,1)$ los valores $a = Z_{0,025} = -1,96$ y $b = Z_{0,975} = 1,96$
- $p(-1,96 < \frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}} < 1,96) = 0,95 = p(0,68 - 1,96\sqrt{\frac{0,68(1-0,68)}{200}} < \pi < 0,68 + 1,96\sqrt{\frac{0,68(1-0,68)}{200}}) = (0,68 \pm 0,065) = (0,615; 0,745)$



- 0,99 en la tabla $N(0,1)$ los valores $a = Z_{0,005} = -2,575$ y $b = Z_{0,995} = 2,575$
- $p(-2,575 < \frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}} < 2,575) = 0,99 = p(0,68 - 2,575\sqrt{\frac{0,68(1-0,68)}{200}} < \pi < 0,68 + 2,575\sqrt{\frac{0,68(1-0,68)}{200}}) = (0,68 \pm 0,085) = (0,595; 0,765)$

$$d) 1,645 \sqrt{\frac{0,68(1-0,68)}{n}} \leq 0,03 \longrightarrow n \geq 654,26 \longrightarrow \text{Tomaríamos muestra de 655 personas}$$



Distribuciones

χ^2 t F

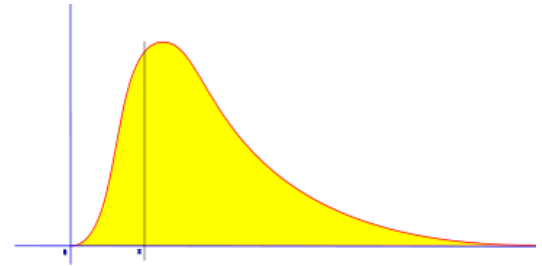
Mar Angulo Martínez
mar.angulo@u-tad.com

Distribución χ^2 de Pearson

La **distribución χ^2 (chi cuadrado) de Pearson** es una distribución continua que se define como la suma de los cuadrados de distribuciones normales estándar de la siguiente forma:

$$\chi_n^2 = X_1^2 + X_2^2 + \dots + X_n^2$$

donde v.a.i.i.d., $X_i \sim N(0,1), i = 1, 2, \dots, n$.



Propiedades:

- Su forma depende de los n grados de libertad (g.d.l.), que se corresponden con el número de distribuciones normales estándar.
- La asimetría va disminuyendo conforme aumentan los grados de libertad.
- Solo toma valores positivos.

Distribución χ^2 de Pearson

Propiedades:

- Cuando n es grande ($n \geq 30$), χ^2 se aproxima a una distribución normal $N(\sqrt{2n-1}, 1)$.
- Dadas dos distribuciones chi, con respectivos g.d.l., $\chi_{n_1}^2$ y $\chi_{n_2}^2$, entonces $\chi_{n_1}^2 + \chi_{n_2}^2 = \chi_{n_1+n_2}^2$.
- En el muestreo, si tomamos muestras de la media, \bar{x} , y la cuasivarianza, S^2 , en una población normal, $N(\mu, \sigma)$, la variable

$$\chi_{n-1}^2 \sim \frac{(n-1) S^2}{\sigma^2}$$

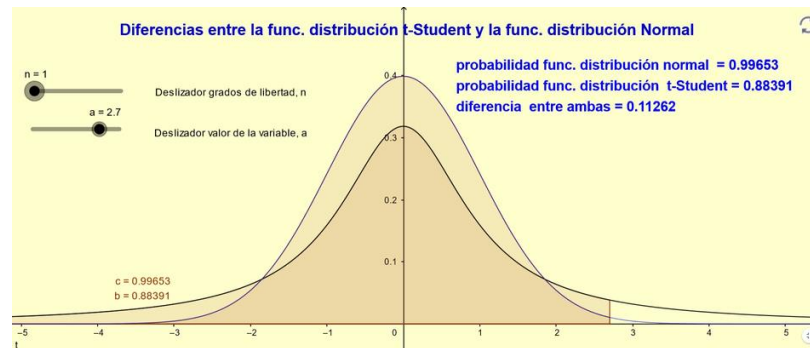
es una chi cuadrado de Pearson con $n - 1$ g.d.l.

Distribución t de Student

La **distribución t de Student** es una distribución continua que se define como el cociente entre una normal estándar y una chi cuadrado:

$$t_n = \frac{X}{\sqrt{\frac{1}{n}(X_1^2 + X_2^2 + \dots + X_n^2)}} = \frac{X}{\sqrt{\frac{1}{n}\chi_n^2}}$$

donde $X \sim N(0,1)$.



Distribución t de Student

Propiedades:

- Es simétrica con respecto al 0 y toma todos los valores de la recta real (es muy similar a la normal en forma).
- Su forma depende de los n g.d.l.
- En el muestreo, si tomamos muestras de la media, \bar{x} , y la varianza, s^2 , en una población normal, $N(\mu, \sigma)$, la variable

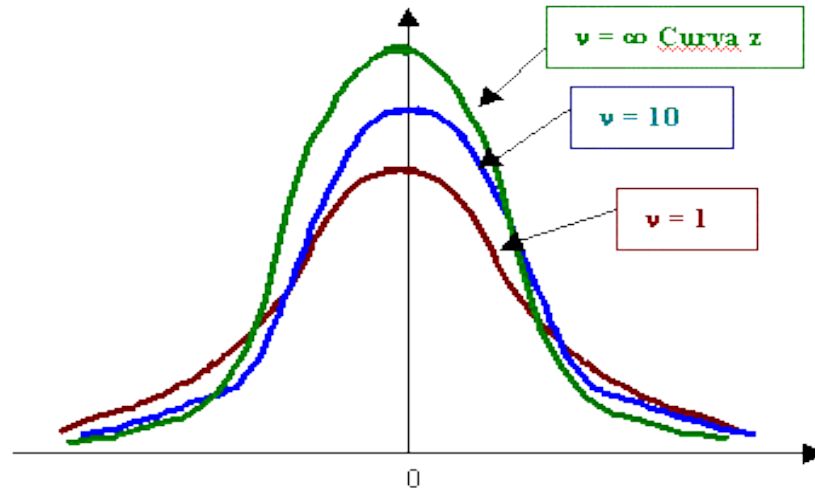
$$t_{n-1} \sim \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

es una t de Student con $n - 1$ g.d.l.

Distribución t de Student

□ Distribución t de Student

- La distribución t de Student se obtiene como cociente entre una $N(0,1)$ y la raíz cuadrada de una Chi-Cuadrado entre el n° de grados de libertad
- Es simétrica respecto al valor 0

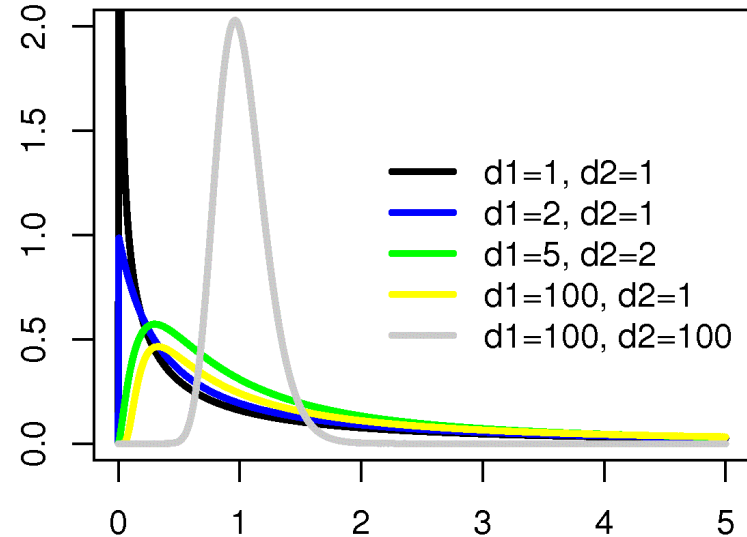


Distribución F de Snedecor

La **distribución F de Snedecor** es una distribución continua que se define como el cociente dos distribuciones chi cuadrado divididas por sendos grados de libertad:

$$F_{n_1; n_2} = \frac{X_1^2/n_1}{X_2^2/n_2}$$

donde $X \sim N(0,1)$.

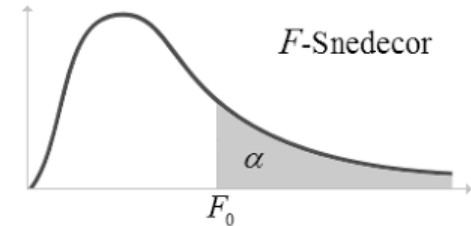


Distribución F de Snedecor

Propiedades:

- Su forma depende de los n_1 y n_2 g.d.l. de las distribuciones chi.
- La asimetría se forma en función de los grados de libertad del numerador y denominador.
- Solo toma valores positivos.
- Cumple la siguiente relación de áreas, útil para el cálculo de probabilidades,

$$F_{n_1;n_2;1-\alpha} = \frac{1}{F_{n_2;n_1;\alpha}}$$



Técnicas de Inferencia

Mar Angulo Martínez
mar.angulo@u-tad.com

Temario

Introducción a la Inferencia. Estimación

- 8.1. Inferencia estadística. Técnicas de inferencia
- 8.2. Estimación puntual y estimación por intervalo
- 8.3. Error máximo de estimación. Determinación del tamaño muestral
- 8.4. Estimación de la media de una población normal
- 8.5. Estimación de la varianza de una población normal
- 8.6. Estimación de una proporción poblacional
- 8.7. Estimación de la diferencia entre dos medias
- 8.8. Estimación de la diferencia entre dos varianzas
- 8.9. Estimación de la diferencia entre dos proporciones

Estimación puntual

- ❑ Una **estimación puntual de un parámetro θ** es el valor que se obtiene seleccionando el estadístico apropiado y calculando su valor con los datos muestrales.
- ❑ El estadístico $\hat{\theta}$ que utilizamos para estimar se llama **estimador puntual de θ**
- ❑ $\hat{\theta}$ es un estimador insesgado de θ si $E(\hat{\theta}) = \theta$.
 - ❑ La diferencia $E(\hat{\theta}) - \theta$ se llama *sesgo del estimador*

- $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ es un estimador insesgado para μ
- $s^2 = \frac{(x_i - \bar{x})^2}{n-1}$ es un estimador insesgado para σ^2
- $P = \frac{\sum_{i=1}^n X_i}{n}$ es un estimador insesgado para π

Intervalos de Confianza

Mar Angulo Martínez
mar.angulo@u-tad.com

Intervalos de confianza

Un **intervalo de confianza** ($1 - \alpha$) es un intervalo en torno a la estimación obtenida donde, con el nivel de significación fijado, tenemos la confianza de encontrar el auténtico valor del parámetro estimado.

Por ejemplo, en el caso de la distribución Normal, si lo que deseamos es estimar el Intervalo de confianza al 95% del valor de la media,

- El parámetro que deseamos estimar es μ
- Y lo que queremos, es hallar un intervalo en la recta real $[a, b]$, tal que

$$P(a < \mu < b) = 1 - \alpha$$

- En otras palabras, el intervalo tal que $\mu \in [a, b]$, con probabilidad 0,95.

❑ Intervalos de Confianza

- ❑ Un intervalo de confianza $(1-\alpha)$ es un intervalo en torno a la estimación obtenida donde, con el nivel de significación fijado, tenemos la confianza de encontrar el auténtico valor del parámetro estimado
- ❑ Un intervalo de confianza del 95% significa que si extraemos un número determinado de muestras del mismo tamaño de una población el 95% de los intervalos de confianza contruidos a partir de esas muestras contendrán el valor del parámetro que buscamos y el 5% restante no lo contendrán.

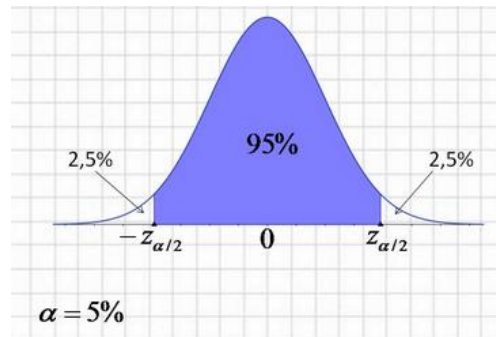
Intervalos de confianza

Si continuamos con el ejemplo, dado que sabemos que $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$,

$$1 - \alpha = P(a < \mu < b) = P\left(\frac{\bar{x} - b}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{\bar{x} - a}{\sigma/\sqrt{n}}\right)$$

Dado que $N(0,1)$ es simétrica respecto a 0, se trata de encontrar

$$P\left(Z \leq \frac{\bar{x} - a}{\sigma/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}$$



Por tanto, trataremos de encontrar, en la tabla de la $N(0,1)$, el valor llamado $z_{\alpha/2}$, tal que $P(Z < z_{\alpha/2}) = 1 - \alpha/2$

Intervalos de confianza


Cálculo de $z_{\alpha/2}$

Ejemplo: $\alpha = 0,05$



$$P(a < \mu < b) = 0,95$$

$$P(Z < z_{\alpha/2}) = 0,975$$

$$z_{\alpha/2} = 1,96$$



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767



Intervalos de confianza

Así,

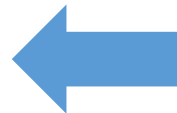
$$\frac{\bar{x} - a}{\sigma/\sqrt{n}} = z_{\alpha/2} \Rightarrow a = \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Y por simetría,

$$b = \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Finalmente, el intervalo de confianza para la media poblacional es

$$I.C. = \left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

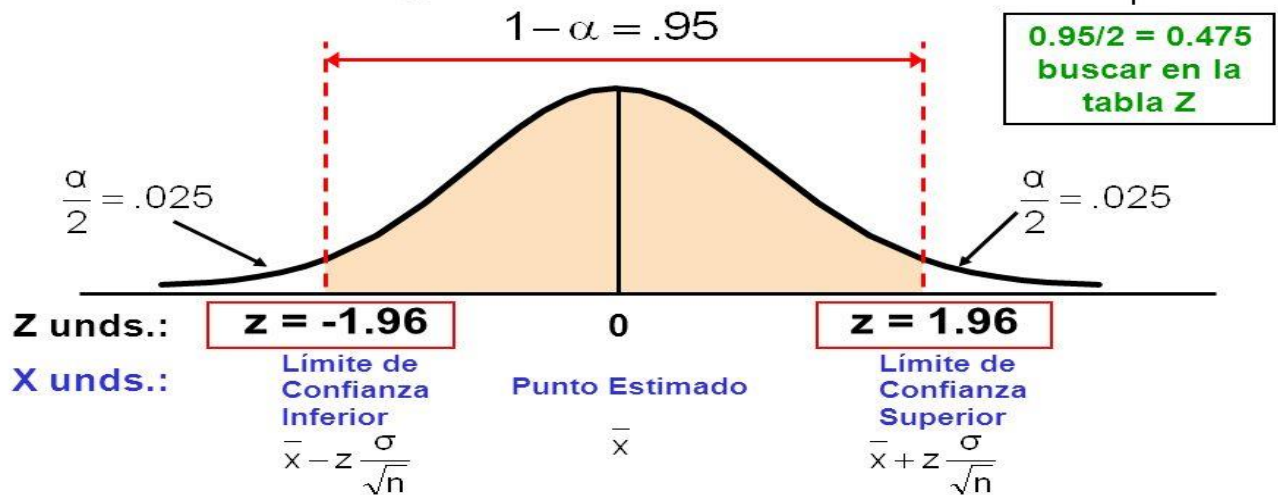


“La media poblacional μ pertenece al intervalo I.C. con probabilidad del $(1 - \alpha)\%$ ”



Hallando el Valor Crítico

- Considerar un intervalo de confianza al 95%:



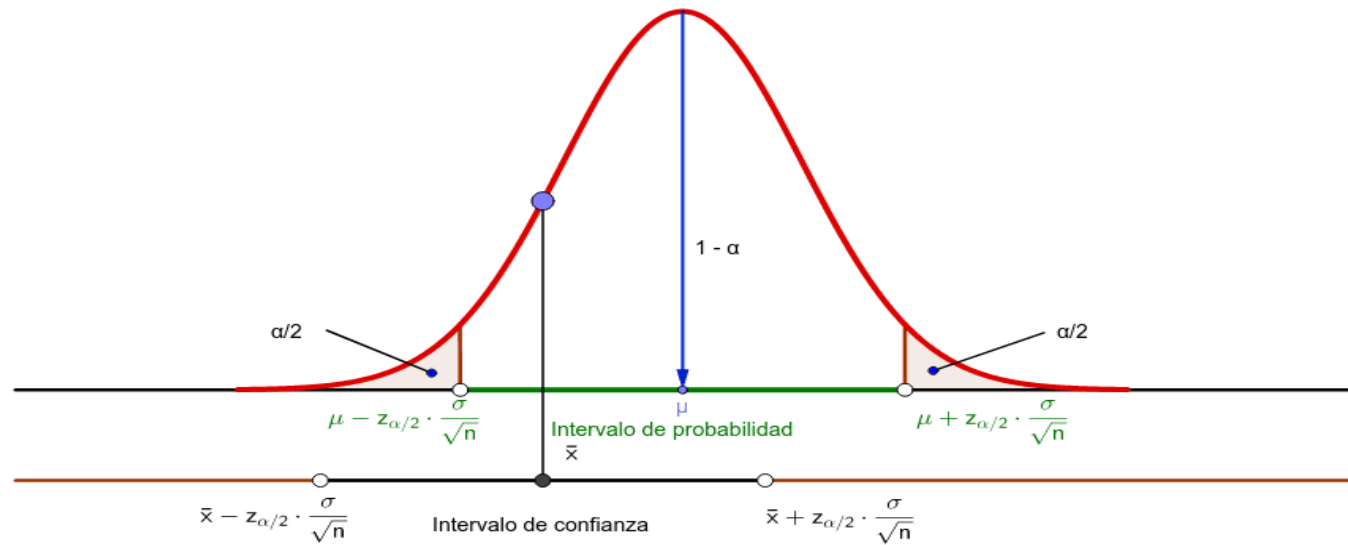
❑ Intervalo de confianza $(1 - \alpha)$ para la media de una distribución Normal (σ conocida)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1) \quad \text{IC: } \left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

- ✓ Nivel de confianza: $(1 - \alpha)$
- ✓ Longitud del intervalo: $2 \cdot Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- ✓ La precisión de la estimación disminuye a medida que aumenta la longitud del intervalo
- ✓ Estrategia habitual: dado unos niveles de confianza y precisión exigidos, calcular el tamaño de muestra mínimo: n

Recuerda

- Error máximo de estimación $Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ nos da una medida de la **precisión** del intervalo (También se llama cota del error de estimación)
- La amplitud del intervalo es el doble del error máximo de estimación: $2Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- El nivel de significación $1 - \alpha$ nos da una medida de la **confianza** del intervalo



- ❑ Intervalo de confianza $(1 - \alpha)$ para la media de una distribución Normal (σ desconocida y muestra grande ($n > 40$))

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow N(0,1) \quad \text{IC: } \left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right)$$

- ❑ Intervalo de confianza $(1 - \alpha)$ para la media de una distribución Normal (σ desconocida y muestra pequeña)

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1} \quad \text{IC: } \left(\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right)$$

- Intervalo de confianza $(1 - \alpha)$ para una varianza σ^2

$$\frac{(n-1)s^2}{\sigma^2} \rightarrow \chi^2_{n-1} \quad \text{IC: } \left(\frac{(n-1)s^2}{\chi^2_{n-1; 1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1; \frac{\alpha}{2}}} \right)$$

- ❑ Intervalo de confianza $(1 - \alpha)$ para una proporción π

$$\frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow N(0,1) \quad \text{IC} \left(p - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, p + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right)$$

❑ **Intervalo de confianza para una diferencia de medias: $\mu_1 - \mu_2$**
(poblaciones Normales)

Consideraciones básicas

- X_1, \dots, X_m es una muestra aleatoria de una distribución Normal con media μ_1 y varianza σ_1^2
- Si Y_1, \dots, Y_n es una muestra aleatoria de una distribución Normal con media μ_2 y varianza σ_2^2
- Las muestras X e Y son independientes entre sí

Estimador puntual de $\mu_1 - \mu_2$: $\bar{X} - \bar{Y}$

❑ Distribución muestral de la diferencia de medias

➤ Permite comparar las medias de dos poblaciones: $\mu_1 - \mu_2$

❑ ¿distribución muestral de $\bar{X} - \bar{Y}$?

✓ Es la distribución que resulta cuando se extraen al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones con medias μ_1 y μ_2 y con desviaciones típicas σ_1 y σ_2 conocidas

$$\bar{X} - \bar{Y} \rightsquigarrow N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \longrightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$$

❑ Caso I (σ_1 y σ_2 conocidas)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1) \quad \text{I.C.} \left(\bar{x} - \bar{y} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x} - \bar{y} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

❑ Caso II (σ_1 y σ_2 desconocidas, $n_1 > 40$ y $n_2 > 40$)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow N(0,1) \quad \text{I.C.} \left(\bar{x} - \bar{y} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x} - \bar{y} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

❑ Caso III (σ_1 y σ_2 desconocidas, n_1 ó $n_2 < 40$)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow t_\varepsilon \quad \text{IC} \left(\bar{x} - \bar{y} - t_{\varepsilon, 1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x} - \bar{y} + t_{\varepsilon, 1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

$$\varepsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}} \quad (\text{redondear siempre } \varepsilon \text{ al entero más cercano hacia abajo})$$

❖ Ejemplo 8

Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura, el A y el B. Con la pintura tipo A se pintan 18 elementos y se registra el tiempo que tarda en secar (en horas). Y lo mismo se hace con la pintura tipo B. Se sabe que la desviación típica de la población es en ambas de 1. Si se supone que el tiempo de secado sigue una distribución normal y que los elementos pintados se secan en el mismo tiempo con los dos tipos de pintura, calcular la probabilidad de que los tiempos medios muestrales de secado difieran en más de una hora.

- $\mu_1 \equiv$ tiempo medio poblacional de secado (en horas) con la pintura A
- $\mu_2 \equiv$ tiempo medio poblacional de secado (en horas) con la pintura B
- $\bar{x}_1 \equiv$ tiempo medio muestral de secado (en horas) con la pintura A
- $\bar{x}_2 \equiv$ tiempo medio muestral de secado (en horas) con la pintura B
- Inferencia para la diferencia de medias con desviaciones típicas poblacionales conocidas

$$\bar{X} - \bar{Y} \rightsquigarrow N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$$

$$\text{▪ } p(\bar{X} - \bar{Y} > 1) = p\left(\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{18} + \frac{1}{18}}} > \frac{1}{\sqrt{\frac{1}{18} + \frac{1}{18}}}\right) = p(Z > 3) = 0,0013$$

❖ Ejemplo 9

Para analizar si la duración de los cinescopios del fabricante A es la misma o no que la duración de los cinescopios que fabrica B se ha tomado una muestra de 36 cinescopios de A y otra muestra de 49 cinescopios de B. En los de A se ha observado una duración media de 6,5 años y una desviación típica de 0,9 años; mientras que en la muestra del fabricante B se ha obtenido una media de 6 años y una desv. Típica de 0,8 años

- Obtener un intervalo de confianza para la diferencia de duraciones medias con un 99% de confianza
- Dar una medida de la precisión de dicho intervalo

- $\mu_1 \equiv$ *duración media (en años)* de los cinescopios del fabricante A
 - $\mu_2 \equiv$ *duración media (en años)* de los cinescopios del fabricante B
 - $\bar{x}_1 \equiv$ *duración media (en años)* de los 36 cinescopios de la muestra del fabricante A = 6,5 años
 - $\bar{x}_2 \equiv$ *duración media (en años)* de la muestra de 49 cinescopios del fabricante B = 6 años
 - $p(a < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < b) = 0,99$ en la tabla $N(0,1)$ los valores $a = Z_{0,005} = -2,575$ y $b = Z_{0,995} = 2,575$
 - $p(-2,575 < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < 2,575) = 0,99 = p((\bar{X} - \bar{Y}) - 2,575 \sqrt{\frac{(0,9)^2}{36} + \frac{(0,8)^2}{49}} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + 2,575 \sqrt{\frac{(0,9)^2}{36} + \frac{(0,8)^2}{49}})$
- = $p(0,5 - 0,486 < \mu_1 - \mu_2 < 0,5 + 0,486) = (0,014 \pm 0,986)$

La diferencia en el número de años de duración está comprendida entre estos dos valores (en todo caso es mayor la duración de los cinescopios del fabricante A) con una probabilidad del 99%. Error máximo de estimación: 0,486 años

❖ Ejemplo 10

Se quiere analizar el impacto del consumo de comida rápida en la dieta de los adolescentes; para ellos se mide el nº de calorías por día de ingesta en dos grupos

Comida rápida	Tamaño de muestra	Media muestral	Desv. Típica muestral
NO	663	2258	1519
SI	413	2637	1138

- ¿Proporcionan estos datos evidencia de que la ingesta de calorías es significativamente mayor en los jóvenes que toman comida rápida? ($\alpha = 0,95$)
- ¿Aceptarías que quienes toman comida rápida ingieren en media hasta 200 calorías más al día? (Realiza el contraste con $\alpha = 0,95$ y con $\alpha = 0,99$)
- Obtener un IC para la diferencia de calorías/día en función del consumo de comida rápida

$$b) H_0: \mu_1 - \mu_2 = 200$$

$$H_A: \mu_1 - \mu_2 > 200$$

$$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{379 - 200}{\sqrt{\frac{(1138)^2}{413} + \frac{(1519)^2}{663}}} = 2,2$$

$$p\text{valor} = p(Z > 2,2) = 0,0139$$

- Como p-valor $< 0,05$: Rechazamos H_0 si $\alpha = 0,05$
 $\alpha = 0,01$

p-valor $> 0,01$: Aceptamos H_0 si

- μ_1 = número medio de calorías por día en la población que consume comida rápida
- $n_1=413$; $\bar{x}=2637$: n° medio calorías muestral en grupo comida rápida $s_1=1138$
- μ_2 = número medio de calorías por día en la población que no consume comida rápida
- $n_2=663$; $\bar{y}=2258$: n° medio calorías muestral en grupo comida rápida $s_2=1519$

a) Planteamos un test para contrastar Hipótesis:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

- Es un contraste sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales grandes (>30)

$$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{379 - 0}{\sqrt{\frac{(1138)^2}{413} + \frac{(1519)^2}{663}}} = 4,66 \quad p\text{valor} = p(Z > 4,66) \approx 0$$

Rechazamos H_0

Concluimos que al 1% el número de calorías es significativamente mayor en el grupo de adolescentes de comida rápida

c) Intervalo de confianza para la diferencia de medias con desviaciones típicas conocidas

$$\text{IC} = \left(\bar{x} - \bar{y} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x} - \bar{y} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = \left(379 - 1,96 \sqrt{\frac{(1138)^2}{413} + \frac{(1519)^2}{663}}, 379 + 1,96 \cdot \sqrt{\frac{(1138)^2}{413} + \frac{(1519)^2}{663}} \right)$$
$$=(379 - 159,43, 379 + 159,43) = (219,57; 538,43)$$

- Con un 95% de confianza la diferencia entre el número medio de calorías diario ingerido por los jóvenes que toman comida rápida y los que no lo hacen está comprendida entre 219,57 y 538,43 calorías.
- Error máximo de estimación: 159,43 calorías

❖ Ejemplo 11

Estudio para analizar el deterioro de las redes de tuberías municipales. Se quiere testar una tecnología que inserta un forro flexible para rehabilitar las tuberías existentes. Los datos de resistencia a la tensión cuando se utilizó dicha técnica de fusión y cuando no se hizo son los que aparecen en la tabla adjunta

	Tamaño de muestra	Media muestral	Desv. típica muestral
SIN fusión	10	2.902,8	277,3
CON fusión	8	3.108,1	205,9

- Obtener un IC para la diferencia de resistencias medias al 99%
- ¿Asegurarías que la técnica de fusión con forro es efectiva al 99%?
- Analizar conjuntamente los resultados

- μ_1 = resistencia media de las tuberías en que se ha aplicado fusión
- $n_1=8$; $\bar{x}=3108,1$: resistencia media en la muestra de tuberías con fusión $s_1=205,9$
- μ_2 = resistencia media de las tuberías en que no se ha aplicado fusión
- $n_2=10$; $\bar{y}=2902,8$ resistencia media en la muestra de tuberías sin fusión $s_2=277,3$

- Inferencia sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales pequeños (<40)

❏ **Caso III** (σ_1 y σ_2 desconocidas, n_1 ó $n_2 < 40$)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow t_{15} \quad \varepsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{205,9^2}{8} + \frac{277,3^2}{10}\right)^2}{\frac{\left(\frac{205,9^2}{8}\right)^2}{7} + \frac{\left(\frac{277,3^2}{10}\right)^2}{9}} = \frac{168.710.977,68}{10.581.747,72} = 15,94$$

$$\begin{aligned} \text{a) } \text{IC}\left(\bar{x} - \bar{y} - t_{\varepsilon, 1 - \frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x} - \bar{y} + t_{\varepsilon, 1 - \frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) = \\ = (205,3 - 2,947 \sqrt{\frac{205,9^2}{8} + \frac{277,3^2}{10}}; 205,3 + 2,947 \sqrt{\frac{205,9^2}{8} + \frac{277,3^2}{10}}) = (205,3 \pm 335,87) = (-130,57; 541,17) \end{aligned}$$

- Con un $\alpha=0,01$ la diferencia entre la resistencia media de tuberías con fusión y sin ella se encuentra entre 130,57 unidades (siendo mayor en el segundo caso) y 541,17 unid. A favor de las tuberías con fusión.
- 0 es un valor dentro del IC: podríamos aceptar la hipótesis de igualdad de medias.

- $\mu_1 =$ resistencia media de las tuberías en que se ha aplicado fusión
- $n_1=8$; $\bar{x} =3108,1$: resistencia media en la muestra de tuberías con fusión $s_1 =205,9$
- $\mu_2=$ resistencia media de las tuberías en que no se ha aplicado fusión
- $n_2=10$; $\bar{y} =2902,8$ resistencia media en la muestra de tuberías sin fusión $s_2= 277,3$

a) Planteamos un test para contrastar Hipótesis:

$H_0: \mu_1 - \mu_2 =0$ la resistencia media de las tuberías con fusión no difiere significativamente de la resistencia media de las tuberías a las que no se ha aplicado

$H_A: \mu_1 - \mu_2 >0$ la resistencia media de las tuberías con fusión es significativamente mayor

- **Es un contraste sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales pequeños (<40)**

$$t_{15 obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{205,3 - 0}{\sqrt{\frac{42394,81}{8} + \frac{76895,29}{10}}} = \frac{205,3 - 0}{113,97} = 1,8$$

- $p\text{valor} = p(t_{15} > 1,8) \in (0,025; 0,05) < 0,05$ Rechazamos H_0 al 5% pero aceptaríamos H_0 al 1% porque nuestro p valor es $> 0,01$
- al 1% concluimos que la resistencia media de las tuberías **no** aumenta significativamente con la técnica de fusión sometida a contraste.

□ ¿y para el cociente de varianzas?

$$\frac{\frac{(n-1)s_1^2}{\sigma_1^2}/(n-1)}{\frac{(m-1)s_2^2}{\sigma_2^2}/(m-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \rightarrow F_{n-1,m-1}$$

- Recuerda

$$F_{m,n,1-\alpha} = \frac{1}{F_{n,m,\alpha}}$$

❑ Intervalo de confianza para el cociente de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \rightarrow F_{n-1, m-1} \quad \text{IC} \left(\frac{s_1^2 / s_2^2}{F_{n-1, m-1, 1-\frac{\alpha}{2}}}, \frac{s_1^2 / s_2^2}{F_{n-1, m-1, \frac{\alpha}{2}}} \right)$$

❑ Intervalo de confianza $(1 - \alpha)$ para una diferencia de proporciones

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \rightarrow N(0, 1) \quad \text{IC} \left(p_1 - p_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}, p_1 - p_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}} \right)$$

❖ Ejemplo 12

Disponemos de los pesos, en kg, de 10 paquetes de semilla para pasto distribuidas por cierta empresa: se ha obtenido una cuasivarianza muestral de $0,43 \text{ kg}^2$. Suponemos una población Normal.

a) Calcular entre qué dos valores estará la varianza de los pesos de todos los paquetes distribuidos por la empresa con probabilidad de 0,95.

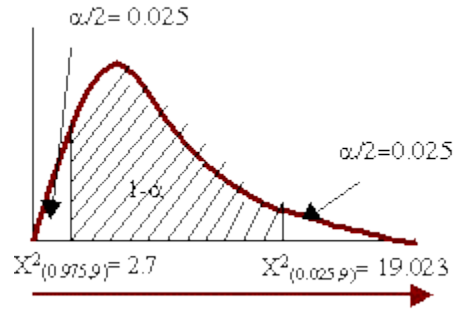
b) Y si tenemos otra muestra de 5 paquetes de semilla de otra empresa cuyos pesos han dado una cuasivarianza muestral de $0,37 \text{ kg}^2$

¿puedes asegurar que hay una diferencia significativa en la dispersión de ambas variables?

- a) Datos: $n=10$
- σ^2 =varianza (dispersión) poblacional de los pesos de los paquetes de semilla (parámetro desconocido a investigar)
- s^2 =varianza muestral de los pesos de los paquetes de semilla = $0,43 \text{ kg}^2$

a) Vamos a calcular un intervalo de confianza para σ^2 (nivel de confianza 0,95):

- $p(a < \frac{(n-1)s^2}{\sigma^2} < b) = 0,95$ Encontramos en la tabla los valores $a = \chi^2_{9;0,025} = 2,7$ y $b = \chi^2_{9;0,975} = 19,02$
- $p(2,7 < \frac{(n-1)s^2}{\sigma^2} < 19,02) = 0,95 = p(\frac{(n-1)s^2}{19,02} < \sigma^2 < \frac{(n-1)s^2}{2,7}) = p(\frac{9 \times 0,43}{19,02} < \sigma^2 < \frac{9 \times 0,43}{2,7}) =$
- $p(0,203 < \sigma^2 < 1,43) = 0,95$
- Con un 95% de confianza la varianza en el peso de los paquetes de semillas está comprendida entre 0,203 y 1,43 kg^2
- Estimación puntual de σ^2 : s^2 = cuasivarianza muestral = $0,43 \text{ kg}^2$



- b) *Datos de la primera empresa:* $n_1=10$
- σ_1^2 =varianza (dispersión) poblacional de los pesos de los paquetes de semilla de la 1ª empresa
- s_1^2 =cuasivarianza muestral de los pesos de los paquetes de semilla de la 1ª empresa = $0,43 \text{ kg}^2$
- *Datos de la segunda empresa:* $n_2=5$
- σ_2^2 =varianza (dispersión) poblacional de los pesos de los paquetes de semilla de la 1ª empresa
- s_2^2 =cuasivarianza muestral de los pesos de los paquetes de semilla de la 2ª empresa = $0,37 \text{ kg}^2$

Planteamos un test para contrastar Hipótesis:

$H_0: \frac{\sigma_1}{\sigma_2} = 1$ ($\equiv \sigma_1 = \sigma_2$) la dispersión en el peso de los paquetes de semillas no presenta diferencias significativas entre la empresa 1 y la empresa 2

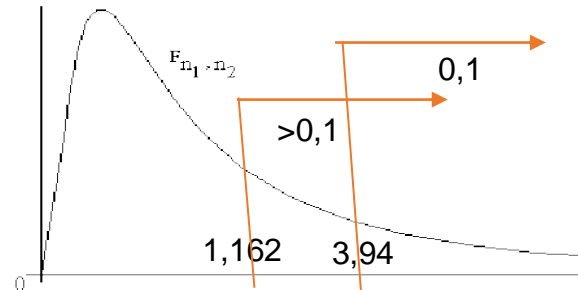
$H_A: \frac{\sigma_1}{\sigma_2} > 1$ ($\equiv \sigma_1 > \sigma_2$) la dispersión en el peso de los paquetes de semilla es significativamente mayor en la empresa 1

- **Es un contraste sobre diferencia de varianzas**

$$F_{obs} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2} = \frac{0,43}{0,37} = 1,162$$

$$p\text{-valor} = p(F_{9,4} > 1,162) > 0,1 (\Rightarrow es > 0,05)$$

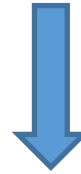
- Aceptamos que la dispersión en el peso de los paquetes de semillas no presenta diferencias significativas



□ y para la diferencia de proporciones...

Estadístico

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \quad N(0,1)$$



□ Intervalo de confianza $(1 - \alpha)$ para una diferencia de proporciones

$$\left(p_1 - p_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 q_2}{n_2}}, p_1 - p_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 q_2}{n_2}} \right)$$

□ o bien...

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad N(0,1)$$



$p = \frac{X_1 + X_2}{n_1 + n_2}$ es la proporción global de individuos que verifican la propiedad en la muestra

□ Intervalo de confianza $(1 - \alpha)$ para una diferencia de proporciones

$$\left(p_1 - p_2 - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, p_1 - p_2 + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right)$$

❖ Ejemplo 13

Una compañía dedicada a la investigación sobre energía solar pretende averiguar el porcentaje de personas dispuestas a adquirir un vehículo con ese tipo de energía. Para ello decide realizar una encuesta en Madrid

a) Si los responsables de la compañía pretenden estimar dicho porcentaje con un error máximo del 3% y una confianza del 99%, ¿a cuántas personas como mínimo deberá presentarse la encuesta para asegurar las condiciones?

b) Se decide finalmente realizar la encuesta a partir de una muestra de 500 personas y resulta que sólo 63 respondieron que estarían dispuestos a comprar un vehículo solar. De manera independiente una encuesta similar fue realizada en Barcelona y de los 700 entrevistados, 79 se mostraron favorables a adquirir un vehículo de ese tipo.

¿Asegurarías al 95% que el vehículo de energía solar tiene una mayor aceptación en una ciudad que en otra? Efectuar un test de hipótesis y explicar la conclusión

c) Construir un intervalo de confianza ($1-\alpha = 0,95$) para la diferencia de proporciones; ¿aceptarías ahora que las proporciones de partidarios del vehículo solar son iguales en Madrid y Barcelona? Razona tus conclusiones.

Partiendo de la idea de **Intervalo de confianza $(1 - \alpha)$ para una proporción**

$\left(p - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}, p + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$: el error máximo de estimación es

$Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$ Como no tenemos aún ninguna estimación de π , utilizamos que $\pi(1-\pi) \leq 0,25$

Entonces exigimos que $Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi(1-\pi)}{n}} \leq 0,03$ Como $Z_{1-\frac{\alpha}{2}} =_{(tablas)} 2,575$

Entonces: $2,575 \sqrt{\frac{0,25}{n}} \leq 0,03 \longrightarrow n \geq 1841,84 \longrightarrow \text{tamaño mínimo de muestra} = 1842$

Ahora ya tenemos las proporciones muestrales:

$$\text{Madrid: } p_1 = \frac{63}{500} = 0,126$$

$$\text{Barcelona: } p_2 = \frac{79}{700} = 0,113$$

b) Planteamos un contraste que nos permita comparar las dos proporciones de personas dispuestas a adquirir vehículo solar (en Madrid y en Barcelona)

Hipótesis: $H_0: \pi_1 - \pi_2 = 0$ No hay diferencias significativas entre Madrid y Barcelona
 $H_A: \pi_1 - \pi_2 > 0$ la proporción de personas dispuestas a comprar un vehículo solar es significativamente mayor en Madrid

$$Z_{obs} = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{0,126 - 0,113}{\sqrt{\frac{0,126 \times 0,874}{500} + \frac{0,113 \times 0,887}{700}}} = 0,68$$

p-valor = $p(Z > 0,68) = 0,2483 > \alpha \longrightarrow$ Aceptamos H_0
Porque tanto si $\alpha = 0,05$ como si $\alpha = 0,01$ el p-valor es un valor mayor que α

C) Intervalo de confianza $(1 - \alpha)$ para una diferencia de proporciones

$$\left(p_1 - p_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}, p_1 - p_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}} \right) = (0,013 - 1,96 \cdot 0,019; 0,013 + 1,96 \cdot 0,019) \\ = (0,013 \pm 0,037) = (-0,024; 0,05)$$

- Con un 99% de confianza, la diferencia de % de personas dispuestas a adquirir un vehículo solar está entre el 2, 4% (siendo mayor en Barcelona) y el 5% siendo mayor en Madrid.
- Como 0 está en el intervalo, en un test bilateral, aceptaríamos la hipótesis de que esas proporciones son iguales (no son significativamente distintas) si trabajamos con $\alpha=0,05$

Contrastes de Hipótesis

Mar Angulo Martínez
mar.angulo@u-tad.com

Temario

Pruebas de hipótesis

- 9.1. Prueba de hipótesis: conceptos generales.
- 9.2. El p-valor. Aplicación en la toma de decisiones
- 9.3. Errores de tipo I y II en una prueba de hipótesis. Potencia del test
- 9.4. Prueba de hipótesis sobre una media poblacional
- 9.5. Prueba de hipótesis sobre una varianza poblacional
- 9.6. Prueba de hipótesis sobre una proporción poblacional
- 9.7. Prueba de hipótesis sobre una diferencia entre dos medias
- 9.8. Prueba de hipótesis sobre la diferencia entre dos varianzas
- 9.9. Prueba de hipótesis para la diferencia entre dos proporciones

❑ **Hipótesis estadística** es una afirmación o aseveración sobre el valor de un parámetro, sobre los valores de varios parámetros o sobre la forma de una distribución completa

❑ **Hipótesis nula:** H_0 es la aseveración que queremos contrastar; la afirmación que inicialmente se cree cierta “creencia previa”

❑ **Hipótesis alternativa:** H_A es una aseveración que contradice la hipótesis nula

❑ La hipótesis nula será rechazada en favor de la hipótesis alternativa sólo si los datos muestran una evidencia de que H_0 es falsa

❑ **Contraste de hipótesis** (prueba de hipótesis) es una técnica que consiste en utilizar los datos muestrales para decidir si la hipótesis nula debe ser aceptada o rechazada

Elementos de un contraste de hipótesis

- ✓ **Estadístico de contraste:** es una función de los datos muestrales en los que deberemos basar la decisión de rechazar o aceptar H_0
- ✓ **Región crítica** (región de rechazo): conjunto de todos los valores para los que H_0 será rechazada

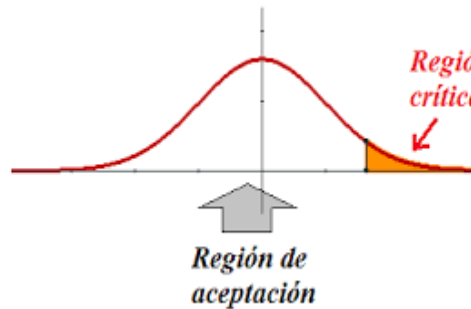
Recuerda:

- Rechazar H_0 no equivale a decir que es falsa, sino que los datos evidencian una diferencia significativa respecto a ese valor contrastado, que resulta muy improbable que H_0 sea cierta a la vista de los resultados que ofrece la muestra
- No rechazar H_0 no significa decir que H_0 es cierta sino que no hay evidencia suficiente en los datos muestrales para rechazarla.

❑ Contrastes de un lado y de dos lados

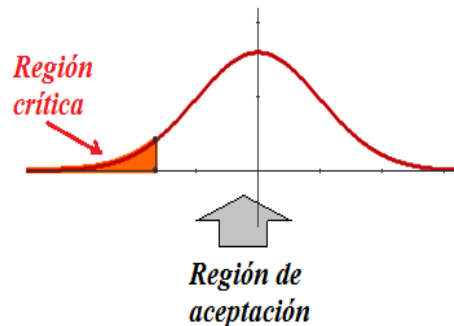
$$H_0: \theta = \theta_0$$

$$H_A: \theta > \theta_0$$



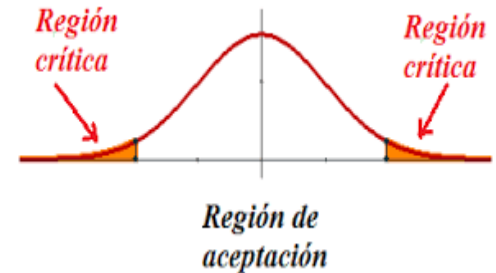
$$H_0: \theta = \theta_0$$

$$H_A: \theta < \theta_0$$



$$H_0: \theta = \theta_0$$

$$H_A: \theta \neq \theta_0$$



❑ Errores en un contraste de hipótesis

	H_0 cierta	H_0 falsa
Rechazamos H_0	Error tipo I α	Decisión correcta
Aceptamos H_0	Decisión correcta	Error tipo II β

Error tipo I: α es la probabilidad de rechazar H_0 cuando es cierta: $p_{H_0}(C)$

Error tipo II: β es la probabilidad de aceptar H_0 cuando es falsa: $p_{H_A}(\bar{C})$

Potencia del contraste: $1 - \beta$ es la probabilidad de rechazar H_0 cuando es falsa

□ p-valor (p value) es la probabilidad de que, siendo cierta la hipótesis nula, obtengamos unos valores más extremos (en el sentido de la región crítica) que los que obtenidos en nuestra muestra.

- Es por tanto la probabilidad de obtener más discrepancia con H_0 que la obtenida con la muestra
- Cuanto menor el p-valor, más extremo es el resultado muestral y por tanto más evidencia muestran los datos para rechazar H_0
- Es por tanto el mínimo nivel de significación al que rechazaríamos H_0 con la muestra obtenida.

❑ Método de NEYMAN-PEARSON


- ✓ La única posibilidad de reducir simultáneamente los errores de tipo I y II es aumentar el tamaño muestral
- ✓ Se considera un valor pequeño de α (0.01 , 0.05, 0,1) que se llama Nivel de significación. **Este es el valor máximo permitido de P(error tipo I)**
- ✓ Entre todas las regiones críticas de nivel α se elige aquella que minimiza la probabilidad de Error de Tipo II (β), es decir, se incrementa al máximo la potencia del test para ese valor de α

❑ Criterio de decisión

❑ Si $p\text{-valor} < \alpha$ *Se rechaza H_0*

❑ Si $p\text{-valor} > \alpha$ *Se acepta H_0*

□ Test de hipótesis: pasos a seguir

- 1) Describir el modelo, identificar el parámetro de interés y describirlo en el contexto del problema
 - 2) Formular la hipótesis nula y la hipótesis alternativa apropiada
 - 3) Seleccionar el estadístico de contraste apropiado que cuantifique la discrepancia entre los datos y la hipótesis nula, y calcular el valor observado (suponiendo H_0 cierta)
 - 4) Definir la región crítica (de rechazo) para el nivel de significación α .
 - 5) Determinar si el valor observado está en la región crítica o en la región de aceptación
 - 6) Calcular el p-valor y comparar con el nivel de significación
 - 7) Decidir si H_0 debe ser rechazada o aceptada y expresar la conclusión en el contexto del problema
- 

¿Y existe alguna relación
entre los contrastes de
hipótesis y los intervalos de
confianza?



Si el valor θ_0 cae dentro de un
intervalo de confianza
(con nivel de confianza $1 - \alpha$)

En el contraste de hipótesis

$$H_0: \theta = \theta_0$$

$$H_A: \theta \neq \theta_0$$

Aceptaríamos H_0

Con error tipo I: α



❖ Ejemplo 14

Análisis de la vacuna de Salk contra la polio en 1954

Se comenzó a administrar una nueva vacuna cuya efectividad se consideraría significativa si lograba reducir a la mitad la incidencia de la polio que se estimaba inicialmente en 30 por cada 100.000.

Se tomó una muestra de 200.745 niños a los que se administró la vacuna y entre ellos se observaron 33 casos de polio

A otro grupo de 201.229 niños se les administró placebo y en este grupo hubo 110 casos de polio.

- a) Construir un intervalo de confianza del 99% para la diferencia de proporciones
- b) ¿Qué puedes concluir sobre la efectividad de la vacuna?

Disponemos de datos de las correspondientes proporciones muestrales:

$$p_2 = \frac{33}{200.745} = 0,000164 \text{ proporción muestral de niños vacunados que enfermaron de polio}$$

$$p_1 = \frac{110}{200.745} = 0,00055 \text{ proporción muestral de niños sin vacunar que enfermaron de polio}$$

- a) Queremos hacer inferencias sobre el parámetro $\pi_1 - \pi_2$ la diferencia entre las respectivas proporciones poblacionales de niños que enfermaron, que nos da una medida de la efectividad de la vacuna

- Queremos construir un Intervalo de Confianza para la diferencia de proporciones
Vamos a trabajar con $\alpha = 0,99$

- $p(a < \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < b) = 0,99$ en la tabla $N(0,1)$ los valores $a = Z_{0,005} = -2,575$ y $b = Z_{0,995} = 2,575$

- $p(-2,575 < \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < 2,575) = 0,99 =$

$$p((p_1 - p_2) - 2,575 \sqrt{\frac{0,00055 \times 0,99945}{201229} + \frac{0,000164 \times 0,999836}{200745}} < \pi_1 - \pi_2 < (p_1 - p_2) + 2,575 \sqrt{\frac{0,00055 \times 0,99945}{201229} + \frac{0,000164 \times 0,999836}{200745}})$$

$$p((0,000386) - 2,575 \times 0,0000286 < \mu_1 - \mu_2 < (0,000386) + 2,575 \times 0,0000286)$$

$$IC: (0,000386 \pm 0,0000736) = (0,00031; 0,00046)$$

- Con un 99% de confianza, la proporción de niños no vacunados que contrajeron la polio estuvo entre 31 y 46 por cada 100.000 por encima de la proporción de niños vacunados que enfermaron; es decir la vacuna consiguió reducir entre 31 y 46 casos de cada 100.000

- b) Vamos a plantear un contraste de hipótesis para comprobar si la vacuna logra que la proporción niños vacunados que se contagian esté en el 15%
- Se trata de plantear un contraste de hipótesis para π_2

Hipótesis: $H_0: \pi_2 = 0,00015$ la vacuna logra reducir la incidencia a la mitad
 $H_A: \pi_2 > 0,00015$ la vacuna no logra el objetivo

- $$Z_{obs} = \frac{p_2 - \pi_2}{\sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}} = \frac{0,000164 - 0,00015}{\sqrt{\frac{0,00015 \times 0,99985}{200745}}} = 0,0000273$$
- P-valor = $p(Z > 0,0000273) = 0,5 > \alpha \longrightarrow$ Aceptamos H_0
- Es decir, con un 1% de margen de error podemos admitir que la incidencia de la polio en niños a los que se administró la vacuna fue de 15 por cada 100.000 (y por tanto se redujo el valor inicial a la mitad, luego se considera efectiva)

❖ Ejemplo 15

En un estudio sobre los préstamos concedidos por dos entidades financieras se toma una muestra aleatoria de 41 préstamos de la primera entidad observando un importe medio de 15.000 euros y una desviación típica de 9.800 euros. Al obtener los datos de una muestra aleatoria de 49 préstamos de la segunda entidad se comprobó un importe medio de 13.000 euros y una desviación típica de 9.300 euros

- a) ¿Proporcionan los datos evidencia ($\alpha = 0,01$) de que el importe medio de los préstamos concedidos es significativamente mayor en la primera entidad?
- b) Construir un intervalo con un 99% de confianza para el importe medio de los préstamos concedidos por la segunda entidad. Interpretar el resultado.
- c) ¿Cuántos créditos como mínimo habrá que analizar para estimar el importe medio de los créditos de la primera entidad con un error máximo de 300 euros? (nivel de confianza: 0,99)

- $\mu_1 =$ importe medio de todos los préstamos de la entidad A
- $n_1=41$; $\bar{x} =15$: importe medio muestral en A (m euros) $s=9,8$ (m euros)
- $\mu_2 =$ importe medio de todos los préstamos de la entidad B
- $n_2=49$; $\bar{x} =13$: importe medio muestral en A (m euros) $s= 9,3$ (m euros)

a) Planteamos un test para contrastar Hipótesis:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

Es un contraste sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales grandes (>40)

$$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(15 - 13) - 0}{\sqrt{\frac{96,04}{41} + \frac{86,49}{49}}} = 0,987 \quad \text{pvalor} = p(Z > 0,987) = 0,1611 > 0,01$$

Aceptamos H_0

Concluimos que al 1% no existen diferencias significativas entre los importes medios de los créditos concedidos por las dos entidades.

b) Intervalo de confianza $(1 - \alpha)$ para el importe medio de los préstamos de la entidad B (σ desconocida y muestra grande)

$$\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right) = \left(13 \mp 2,575 \cdot \frac{9,3}{\sqrt{49}} \right) \\ = (13 \mp 3,42) = (9,58; 16,42)$$

- Con una confianza del 99% el importe medio de los créditos de la entidad B está comprendido entre 9.580 euros y 16.420 euros
- Mejor interpretación: si tomamos muestras de 49 créditos de la entidad B, el 99% de los intervalos de confianza contruidos a partir de esas muestras contendrán el valor del parámetro que buscamos y el 5% restante no lo contendrán.

c) Exigimos que el error máximo de estimación $Z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq 0,3$ (m euros)

$$\text{Entonces } 2,575 \cdot \frac{9,8}{\sqrt{n}} \leq 0,3 \quad n \geq 7075,61$$

Hemos de tomar una muestra de 7076 créditos de la primera entidad para lograr esas condiciones de estimación

❖ Ejemplo 16

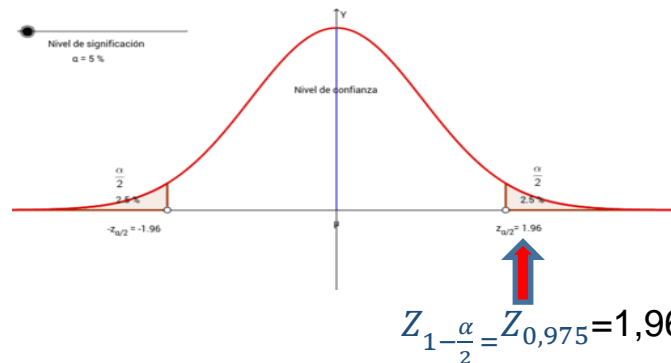
La porosidad de muestras de carbón en una costura particular está normalmente distribuida con desviación típica verdadera de 0,75.

- a) ¿entre qué dos valores estará la porosidad media poblacional si la porosidad media en 20 elementos ha sido de 4,85? Confianza:0,95
- b) Y si aumentamos la confianza al 98%?
- c)Cuál deberá ser el tamaño de la muestra para que la amplitud del intervalo sea de 0,40 con un 95% de confianza?
- d) ¿y cuál deberá ser el tamaño mínimo de muestra para estimar la media verdadera con error $<0,2$ y confianza del 99%?
- e) ¿y cuál deberá ser el tamaño mínimo de muestra para estimar la media verdadera manteniendo la confianza del 99% pero reduciendo la cota de error a la mitad?

- $\mu = \text{porosidad media de las costuras de carbón}$ $\sigma=0,75$
- $n=20$; $\bar{x}=4,85$: porosidad media muestral

a) **Intervalo de confianza ($1 - \alpha = 0,95$) para la porosidad media de costuras de carbón (σ conocida)**

$$\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(4,85 - 1,96 \cdot \frac{0,75}{\sqrt{20}}, 4,85 + 1,96 \cdot \frac{0,75}{\sqrt{20}} \right) = (4,85 - 0,33, 4,85 + 0,33) = (4,52; 5,18)$$



a) **b) Si aumentamos la confianza $1 - \alpha = 0,98$**

$$Z_{1-\frac{\alpha}{2}} = Z_{0,99} = 2,33$$

$$\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(4,85 - 2,33 \cdot \frac{0,75}{\sqrt{20}}, 4,85 + 2,33 \cdot \frac{0,75}{\sqrt{20}} \right) = (4,85 - 0,39; 4,85 + 0,39) = (4,46; 5,24)$$

c) **¿Cuál deberá ser el tamaño de la muestra para que la amplitud del intervalo sea de 0,40 con un 95% de confianza?**

- Amplitud del intervalo: $2Z_{0,975} \cdot \frac{\sigma}{\sqrt{n}} = 0,40 \implies 2 \times 1,96 \times \frac{0,75}{\sqrt{n}} = 0,40 \implies n = 54,02 \approx 54$
- Se deben analizar 54 elementos para conseguir un 95% de confianza y una amplitud del intervalo de 0,4 (equivalente a decir que la cota del error de estimación sea de 0,2 unidades)

d) **¿y cuál deberá ser el tamaño mínimo de muestra para estimar la media verdadera con error <0,2 y confianza del 99%?**

- Error máximo de estimación $Z_{0,995} \cdot \frac{\sigma}{\sqrt{n}} < 0,2 \implies 2,575 \times \frac{0,75}{\sqrt{n}} < 0,2 \implies n = 93,24 \approx 94$
- Se deben analizar 94 elementos para conseguir un 99% de confianza y una amplitud del intervalo de 0,4 (equivalente a decir que la cota del error de estimación sea de 0,2 unidades)

e) **¿y cuál deberá ser el tamaño mínimo de muestra para estimar la media verdadera manteniendo la confianza del 99% pero reduciendo la cota de error a la mitad?**

- Error máximo de estimación $Z_{0,995} \cdot \frac{\sigma}{\sqrt{n}} < 0,1 \implies 2,575 \times \frac{0,75}{\sqrt{n}} < 0,1 \implies n = 324,9 \approx 325$
- Se deben analizar 325 elementos para conseguir un 99% de confianza y una cota del error de estimación sea de 0,1 unidades.

❖ Ejemplo 17

Estudio sobre 215 médicos; 125 en servicio de tiempo completo vivieron un promedio de 48,9 años después de su graduación, en tanto que 90 médicos con trabajo académico vivieron un promedio de 43,2 años.

- a) Obtener un IC al 95% para $\mu_1 - \mu_2$ suponiendo que son conocidas $\sigma_1 = 14,6$ y $\sigma_2 = 14,4$
- b) Analizar si el tiempo de supervivencia tras la graduación es significativamente mayor en alguno de los colectivos de médicos (Contraste de un lado)
- c) Obtener la estimación puntual para la diferencia de años de supervivencia tras la graduación

- $\mu_1 =$ tiempo medio de vida (en años) tras la graduación de los médicos en servicio completo;
- $n_1=125$; $\bar{x}=48,9$: tiempo medio de vida de los médicos en servicio completo de la muestra
 $\sigma_1 =14,6$ (años)
- $\mu_2 =$ tiempo medio de vida (en años) tras la graduación de los médicos con trabajo académico;
- $n_2=9$; $\bar{y}=43,2$: tiempo medio de vida muestral para los médicos con trabajo académico
 $\sigma_2 = 14,4$ años

a) Intervalo de confianza ($1 - \alpha = 0,95$) para la diferencia entre el tiempo medio de vida tras la graduación entre ambos tipos de médicos

❑ Caso I (σ_1 y σ_2 conocidas)

$$\begin{aligned} \text{I.C. } \left(\bar{x} - \bar{y} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x} - \bar{y} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) &= (5,7 - 1,96 \sqrt{\frac{(14,6)^2}{125} + \frac{(14,4)^2}{90}}; 5,7 + 1,96 \sqrt{\frac{(14,6)^2}{125} + \frac{(14,4)^2}{90}}) \\ &= (5,7 - 3,924; 5,7 + 3,924) = (1,776; 9,624) \end{aligned}$$

$$Z_{1-\frac{\alpha}{2}} = Z_{0,975} = 1,96$$

- Con un 95% de confianza los médicos en servicio completo viven de media entre 1,776 y 9,624 años más tras la graduación. Error máximo de estimación: 3,924 años
- d) La estimación puntual para la diferencia de tiempos medios de vida tras graduación es de 5,7 años.

b) Analizar si el tiempo de supervivencia tras la graduación es significativamente mayor en alguno de los colectivos de médicos (Contraste de un lado)

- **Es un contraste sobre diferencia de medias con varianzas poblaciones conocidas**

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

- $$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{5,7 - 0}{\sqrt{\frac{(14,6)^2}{125} + \frac{(14,4)^2}{90}}} = 2,85$$

- $p\text{valor} = p(Z > 2,85) = 1 - 0,9978 = 0,0022 < 0,01 \longrightarrow \text{Rechazamos } H_0$

- Concluimos que al 1% el tiempo medio de supervivencia tras la graduación es significativamente mayor en el grupo de médicos que están en servicio completo.

❖ Ejemplo 18

Una muestra aleatoria de 10 relámpagos en cierta región dieron un eco de radar promedio de 0,81 sg y una desviación típica de 0,34 sg. El eco de radar sigue una distribución normal.

Calcular entre qué dos valores se encuentra la desviación típica del eco e interpretar el resultado

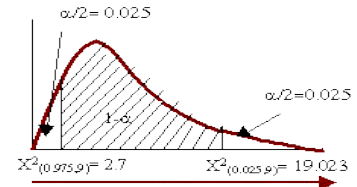
- $X \equiv$ eco de radar de los relámpagos $N(\mu; \sigma)$
- $\sigma \equiv$ desviación típica poblacional del eco de radar de los relámpagos
- $s =$ desviación típica de los 110 relámpagos de la muestra = 0,34 sgs

- **Intervalo de confianza $(1 - \alpha)$ para una varianza σ^2**

$$\text{IC: } \left(\frac{(n-1)s^2}{\chi^2_{n-1; 1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1; \frac{\alpha}{2}}} \right) = \left(\frac{9 \times (0,34)^2}{19,02}, \frac{9 \times (0,34)^2}{2,71} \right) = (0,055; 0,38)$$

Con un 95% de probabilidad la varianza de los relámpagos en esa región está comprendida entre 0,055 y 0,38, la desviación típica estará entre 0,2345 sg y 0,616 sg.

$$\chi^2_{n-1; \frac{\alpha}{2}} = \chi^2_{9; 0,025} = 19,02 \quad \chi^2_{n-1; 1-\frac{\alpha}{2}} = \chi^2_{9; 0,975} = 2,71$$



❖ Ejemplo 19

En una muestra de 1000 consumidores seleccionados al azar 250 aseguraron haber solicitado reembolso después de haber comprado un producto. Estimar al 95% la proporción de consumidores que solicitaron reembolso en la población

¿Dirías que hay evidencia para aceptar que esa proporción es menor que 1/3?

- $\pi \equiv$ proporción poblacional de consumidores que solicitaron reembolso
- $p \equiv$ proporción muestral de consumidores que solicitaron reembolso $= \frac{250}{1000} = 0,25$

□ Intervalo de confianza $(1 - \alpha)$ para una proporción π

$$\left(p - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, p + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) = \left(0,25 - 1,96 \cdot \sqrt{\frac{0,25 \times 0,75}{1000}}, 0,25 + 1,96 \sqrt{\frac{0,25 \times 0,75}{1000}} \right) = (0,25 \pm 0,027) = (0,2225; 0,277)$$

- Con un 95% de confianza la proporción de consumidores que solicitó reembolso está entre el 22,25% y el 27,7%)



Error máx. de estimación: 2,7%

$$Z_{1-\frac{\alpha}{2}} = 1,96$$

a) ¿Dirías que hay evidencia para aceptar que esa proporción es menor que 1/3?

- **Es un contraste sobre una proporción**

$$H_0: \pi = 1/3$$

$$H_A: \pi < 1/3$$

- $$Z_{obs} = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0,25 - 0,33}{\sqrt{\frac{0,33 \times 0,67}{1000}}} = \frac{0,25 - 0,33}{\sqrt{\frac{0,33 \times 0,67}{1000}}} = \frac{-0,08}{0,015} = -5,38$$
- $p\text{valor} = p(Z < -5,38) \approx 0 < 0,01 \longrightarrow \text{Rechazamos } H_0$
- Con un 1% de margen de error, podemos asegurar que el porcentaje de consumidores que solicitaron reembolso no es de 1/3; es significativamente menor

❖ Ejemplo 20

❖ Una consultora encuesta a los residentes de un distrito sobre la proporción de electorado satisfecha con los equipamientos públicos del mismo.

a) ¿Qué tamaño de muestra es necesario si el intervalo de confianza del 95% debe tener amplitud como mucho de 0,1?

b) la consultora está firmemente convencida de que por lo menos 2/3 del electorado está conforme con los servicios Toma una muestra de 400 personas y obtiene respuestas positivas de 265 ¿qué opinión te merece la afirmación de la consultora?

$\pi \equiv$ proporción poblacional de personas satisfechas con el equipamiento del distrito

- Buscamos el tamaño muestral mínimo para estimar la proporción con una confianza de 0,95 y una amplitud del intervalo de 0,1 \longrightarrow error máximo de estimación =0,05

- **Recuerda:** Cuando no tenemos ningún valor de p (porque no hemos tomado aún ninguna muestra) podemos utilizar que $\pi(1 - \pi) \leq 0,25$

- Amplitud del intervalo: $2 Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \leq 0,1 \longrightarrow 2 \times 1,96 \times \sqrt{\frac{0,25}{n}} \leq 0,1 \longrightarrow n \geq 384,16 \approx 385$
- **Tendremos que consultar a 385 personas de ese distrito** para conseguir un intervalo con un **95% de confianza** y una amplitud del intervalo de 0,1 (equivalente a decir que la **cota del error de estimación sea de 0,05, es decir del 5%**)

b) Vamos a contrastar si el criterio de la consultora es o no acertado

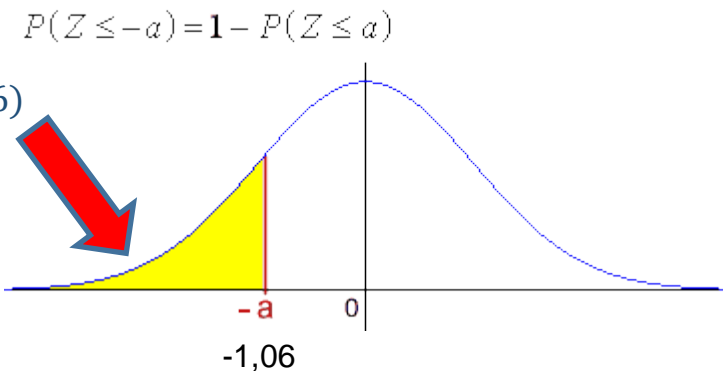
El dato obtenido es $p = \frac{258}{400} = 0,645$

- **Es un contraste sobre una proporción**

$$H_0: \pi = 2/3$$

$$H_A: \pi < 2/3$$

$$p\text{-valor} = p(Z < -1,06)$$



- $$Z_{obs} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0,645 - 0,67}{\sqrt{\frac{0,67 \times 0,33}{400}}} = \frac{-0,025}{0,0235} = -1,06$$

- $p\text{valor} = p(Z < -1,06) = 0,1446 > 0,05 \longrightarrow \text{Aceptamos } H_0$

- Con un 1% de margen de error, podemos asegurar que el porcentaje de personas del distrito que están conformes con los equipamientos públicos del mismo no difiere significativamente de 2/3; admitimos por tanto el criterio de la consultora

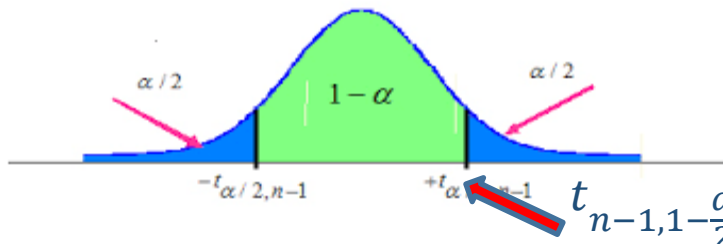
❖ Ejemplo 21

Considera una muestra de 10 perritos calientes para tratar de evaluar la cantidad de grasa que contienen (que sigue una distribución normal). Se ha obtenido una cantidad media de 21,9 grs con desviación típica de 4,134 grs. Estimar el nivel medio de grasa en los perritos calientes. Dar una medida de la precisión de tu estimación.

- μ = cantidad media de grasa de los perritos calientes
- $n=10$ $\bar{x}=21,9$ grs; es la cantidad media de grasa en los 10 perritos calientes de la muestra
- $S=4,134$ es la desviación típica muestral

❑ Queremos obtener un Intervalo de confianza $(1 - \alpha = 0,95)$ para la media de una distribución Normal (σ desconocida y muestra pequeña)

$$\begin{aligned} \text{IC: } \left(\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right) &= \left(21,9 - t_{9, 0,975} \cdot \frac{4,134}{\sqrt{10}}, 21,9 + t_{9, 0,975} \cdot \frac{4,134}{\sqrt{10}} \right) = \\ &= \left(21,9 \pm 2,262 \cdot \frac{4,134}{\sqrt{10}} \right) = (21,9 \pm 2,96) = (18,94; 24,86) \end{aligned}$$



- Con un 95 % de confianza, la cantidad media de grasa de los perritos calientes está comprendida entre 18,94 y 24,86 grs. Error máx. de estimación: 2,96 grs

❖ Ejemplo 22

El tiempo medio de secado de pintura está normalmente distribuido con media 75 minutos y se cree que la desviación estándar es de 9 m. Algunos químicos proponen un aditivo para reducir el tiempo de secado. Se toma una muestra de 25 elementos y se pintan con el nuevo compuesto, resultando un tiempo medio de 70,8 minutos. ¿Existe razón para adquirir el nuevo producto?

- a) Valorar la veracidad de la afirmación al 5% y al 1% de significación.
- b) Obtener un Intervalo de Confianza al 99% y valorar la hipótesis desde dicho intervalo.

- $\mu =$ tiempo medio de secado con la pintura con aditivo
- $n=25$ $\bar{x}=70,8$ grs; es el tiempo medio de secado de los elementos de la muestra
- $\sigma=9$ es la desviación típica (conocida)

a) Contraste de hipótesis para una media con σ conocida

$H_0: \mu=75$ El aditivo no reduce significativamente el tiempo medio de secado

$H_A: \mu<75$ El aditivo reduce significativamente el tiempo de secado

$$Z_{obs} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{70,8-75}{9/\sqrt{25}} = -2,33 \quad p\text{-valor} = p(Z < -2,33) = 0,0099 < 0,01 \longrightarrow \text{Rechazamos } H_0$$

- Con un margen de error del 1% podemos afirmar que el tiempo medio de secado se reduce significativamente con el aditivo.

- ❑ **b) Queremos obtener un Intervalo de confianza ($1 - \alpha = 0,99$) para la media de una distribución Normal (σ conocida)**
- ❑ $IC: \left(\bar{x} - Z_{0,995} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{0,995} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(70,8 - 2,575 \cdot \frac{9}{\sqrt{25}}, 70,8 + 2,575 \cdot \frac{9}{\sqrt{25}} \right) = (70,8 \pm 4,635) = (66,165; 75,435)$

$$Z_{0,995} = 2,575$$

- *Con un 99 % de confianza, el tiempo medio de secado con la pintura que incorpora nuevo aditivo está comprendido entre 66,165 y 75,435 minutos*
- ***¿Aceptaríamos que el tiempo medio se mantiene en 75 minutos o que se ha reducido de forma significativa?***

En un test bilateral con $\alpha = 0,01$ aceptaríamos que no hay una diferencia significativa, aunque comprobamos de hecho que el valor 75 está justo en el límite superior del intervalo, lo que indica que hay una reducción que ha de tenerse en cuenta.

❖ Ejemplo 23

- ❖ Una consultora encuesta a los residentes de un distrito sobre la proporción de electorado satisfecha con los equipamientos públicos del mismo.
- a) ¿Qué tamaño de muestra es necesario si el intervalo de confianza del 95% debe tener amplitud como mucho de 0,1?
- b) la consultora está firmemente convencida de que por lo menos 2/3 del electorado está conforme con los servicios Toma una muestra de 400 personas y obtiene respuestas positivas de 265 ¿qué opinión te merece la afirmación de la consultora?

- $\pi \equiv$ *proporción poblacional de personas satisfechas con el equipamiento del distrito*
- Buscamos el tamaño muestral mínimo para estimar la proporción con una confianza de 0,95 y una amplitud del intervalo de 0,1 \longleftrightarrow error máximo de estimación =0,05
- **Recuerda:** Cuando no tenemos ningún valor de p (porque no hemos tomado aún ninguna muestra) podemos utilizar que $\pi(1 - \pi) \leq 0,25$

- Amplitud del intervalo: $2 Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}} \leq 0,1 \longrightarrow 2 \times 1,96 \times \sqrt{\frac{0,25}{n}} \leq 0,1 \quad n \geq 384,16 \approx 385$
- **Tendremos que consultar a 385 personas de ese distrito** para conseguir un intervalo con un **95% de confianza** y una amplitud del intervalo de 0,1 (equivalente a decir que la **cota del error de estimación sea de 0,05, es decir del 5%**

b) Vamos a contrastar si el criterio de la consultora es o no acertado

El dato obtenido es $p = \frac{258}{400} = 0,645$

Es un contraste sobre una proporción

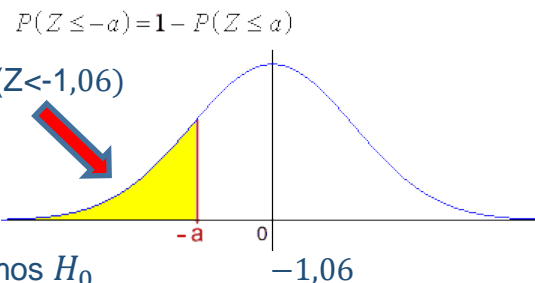
$$H_0: \pi = 2/3$$

$$H_A: \pi < 2/3$$

$$Z_{obs} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0,645 - 0,67}{\sqrt{\frac{0,67 \times 0,33}{400}}} = \frac{-0,025}{0,0235} = -1,06$$

$$p\text{-valor} = p(Z < -1,06) = 0,1446 > 0,05 \Rightarrow$$

Aceptamos H_0



- Con un 1% de margen de error, podemos asegurar que el porcentaje de personas del distrito que están conformes con los equipamientos públicos del mismo no difiere significativamente de 2/3; admitimos por tanto el criterio de la consultora

❖ Ejemplo 25

Los autores de un estudio concluyeron que la distribución de ferritina en adultos tenía una varianza menor que en los jóvenes. Para una muestra de 31 adultos mayores, la varianza fue de 52,6 mg/l y en una muestra de 26 jóvenes fue de 84,2

- a) ¿Puedes asegurar que la afirmación es cierta al 1%?
- b) Obtener un Intervalo de confianza con el 95%

- b) *Datos de la muestra de jóvenes: $n_1=26$*
- σ_1^2 =varianza (dispersión) poblacional de la ferritina en los jóvenes
- s_1^2 =varianza muestral de la ferritina en los jóvenes de la muestra = 84,2 mg/l
- *Datos de la muestra de adultos: $n_2=31$*
- σ_2^2 =varianza (dispersión) poblacional de la ferritina en adultos
- s_2^2 =varianza muestral de la ferritina en los adultos de la muestra = 52,6 mg/l

a) Planteamos un test para contrastar Hipótesis:

$H_0: \frac{\sigma_1}{\sigma_2} = 1$ ($\equiv \sigma_1 = \sigma_2$) la dispersión en la ferritina no presenta diferencias significativas entre jóvenes y adultos

$H_A: \frac{\sigma_1}{\sigma_2} > 1$ ($\equiv \sigma_1 > \sigma_2$) la dispersión en la ferritina es significativamente mayor en jóvenes que en adultos

▪ **Es un contraste sobre diferencia de varianzas**

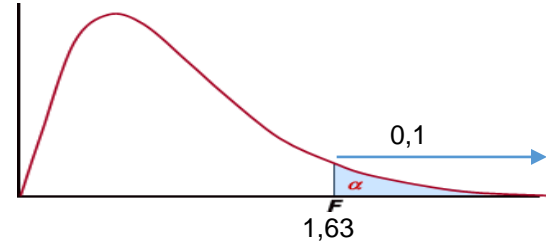
$$F_{obs} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2} = \frac{84,2}{52,6} = 1,6$$

p-valor = $p(F_{25,30} > 1,6) > 0,1 \longrightarrow$ Aceptamos H_0

Al 1% podemos asegurar que no hay diferencia significativa en la dispersión en los niveles de ferritina de jóvenes y adultos

$$b) IC: \left(\frac{s_1^2 / s_2^2}{F_{n-1, m-1, 1-\frac{\alpha}{2}}}, \frac{s_1^2 / s_2^2}{F_{n-1, m-1, \frac{\alpha}{2}}} \right) = \left(\frac{84,2 / 52,6}{F_{25,30,0,025}}, \frac{84,2 / 52,6}{F_{25,30,0,975}} \right) = \left(\frac{1,6}{F_{25,30,0,025}}, \frac{1,6}{F_{25,30,0,975}} \right) = (0,46; 2,12)$$

$$utilizando que F_{25,30,0,025} = \frac{1}{F_{30,25,0,975}} = \frac{1}{2,18} = 0,46$$



❖ Ejemplo 24

Un servicio de mensajería anuncia que al menos el 90% de todos los paquetes recibidos en su oficina en torno a las 9.00 de la mañana son entregados antes del mediodía (siempre dentro de la misma ciudad).

- Si en una muestra de 225 paquetes sólo el 80% han sido entregados en esas horas ¿aceptarías la afirmación de la empresa?
- Calcular n para estimar la verdadera proporción con un error máximo del 3%. ($1-\alpha = 0,95$)

- $\pi = \text{proporción de paquetes que son entregados antes del mediodía}$
- $p \equiv \text{proporción muestral de paquetes entregados antes del mediodía} \quad p=0,8$

$$H_0: \pi=0,9 \quad Z_{obs} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} > \frac{0,8-0,9}{\sqrt{\frac{0,9(1-0,9)}{225}}} = \frac{-0,01}{0,02} = -0,5$$
$$H_A: \pi < 0,9$$

$$p\text{-valor} = p(Z < -0,5) = 0,3085 > 0,01 \longrightarrow \text{Aceptamos } H_0$$

Admitimos por tanto al 1% el anuncio del servicio de que consiguen entregar en ese tiempo al menos un 90% de los paquetes recibidos

$$b) \text{ Exigimos que } 1,96 \sqrt{\frac{\pi(1-\pi)}{n}} \leq 0,03 \quad 1,96 \sqrt{\frac{0,8(1-0,8)}{n}} \leq 0,03 \longrightarrow n \geq 682,95 \quad \text{Tomaríamos } 683$$

❖ Ejemplo 25

Se pretende evaluar un nuevo proceso de fabricación de componentes de un producto.

Para determinar si el nuevo proceso supone una mejora en la calidad que ofrece el producto se toman muestras de partes fabricadas con el proceso antiguo y con el actual.

Se encuentra que 75 de 1500 artículos elaborados con el proceso actual están defectuosos y 80 de los 2000 manufacturados con el nuevo también.

- a) ¿Cómo informarías el cambio de proceso de fabricación?
- b) Estimar mediante un intervalo la diferencia de proporciones de partes defectuosas e interpretar el resultado.
- c) A la vista del intervalo ¿qué conclusión extraerías respecto al apartado a?
Utilizar en todo caso nivel de significación 0,05

$\pi_1 \equiv$ proporción de artículos defectuosos con el proceso actual

$\pi_2 \equiv$ proporción de artículos defectuosos con el proceso nuevo

- Las proporciones de defectuosos en las respectivas muestras han sido de.

$$p_1 = \frac{75}{1500} = 0,05 \text{ con el proceso actual}$$

$$p_2 = \frac{80}{2000} = 0,04 \text{ con el proceso nuevo}$$

- Planteamos un contraste que nos permita comparar las dos proporciones de elementos defectuosos con uno y otro procesos de fabricación:

Hipótesis: $H_0: \pi_1 - \pi_2 = 0$ No hay diferencias significativas entre ambos
 $H_A: \pi_1 - \pi_2 > 0$ La proporción de piezas defectuosas es menor con el proceso nuevo

$$Z_{obs} = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{0,05 - 0,04}{\sqrt{\frac{0,05 \times 0,95}{1500} + \frac{0,04 \times 0,96}{2000}}} = \frac{0,01}{0,07} = 0,14$$

p-valor = $p(Z > 0,14) = 0,4443 > \alpha \longrightarrow$ Aceptamos H_0

Porque tanto si $\alpha = 0,05$ como si $\alpha = 0,01$ el p-valor es un valor mayor que α

Concluimos que no hay diferencia significativa: los datos no proporcionan evidencia de que la proporción de piezas defectuosas sea significativamente menor con el proceso nuevo: no invertiríamos en él, por tanto.

b) Intervalo de confianza $(1 - \alpha)$ para una diferencia de proporciones

$$\left(p_1 - p_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}, p_1 - p_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}} \right) =$$
$$= (0,01 - 1,96 \cdot 0,07; 0,013 + 1,96 \cdot 0,07) = (0,01 \pm 0,137) = (-0,127; 0,147)$$

c) Con un 95% de confianza, la diferencia de % de elementos defectuosos con el actual y con el nuevo proceso de fabricación está comprendido entre un 12,7% (siendo mayor con el nuevo) y un 14,7% (siendo mayor con el actual).

Como 0 está en el intervalo, en un test bilateral y con $\alpha=0,05$, aceptaríamos la hipótesis de que esas proporciones no son significativamente distintas

❖ Ejemplo 26

Para probar la diferencia en el desgaste abrasivo de dos materiales se prueban 12 piezas de material 1 para medir el desgaste; las muestras revelaron un desgaste promedio de 85 unidades con una desviación típica de 4; las 10 piezas de una segunda muestra de material 2 revelaron un promedio de 81 y una desviación de 5 unidades;

a) Podríamos concluir, a un nivel de significación de 0,05, que el desgaste abrasivo del material 2 excede en más de 2 unidades al del material 1?

- $\mu_1 = \text{desgaste medio de las piezas del primer material}$ $n_1=12$;
- $\bar{x}=85$: desgaste medio en la muestra del primer material $s_1=4$
- $\mu_2 = \text{desgaste medio de las piezas del segundo material}$ $n_2=10$;
- $\bar{y}=81$ desgaste medio en la muestra del segundo material $s_2=5$

a) Planteamos un test para contrastar Hipótesis:

$H_0: \mu_1 - \mu_2 = 2$ el desgaste medio del primer material excede en 2 unidades al desgaste medio del segundo

$H_A: \mu_1 - \mu_2 > 2$ el desgaste medio del primer material excede en más de 2 unidades al desgaste medio del segundo

- Inferencia sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales pequeños (<40)

□ **Caso III** (σ_1 y σ_2 desconocidas, n_1 ó $n_2 < 40$)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow t_{17}$$

$$\varepsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{16}{12} + \frac{25}{10}\right)^2}{\frac{\left(\frac{16}{12}\right)^2}{11} + \frac{\left(\frac{25}{10}\right)^2}{9}} = \frac{14,694}{0,854} = 17,19$$

- $t_{17 \text{ obs}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{85 - 81 - 2}{\sqrt{\frac{16}{12} + \frac{25}{10}}} = \frac{2}{1,958} = 1,02$
- $p\text{valor} = p(t_{17} > 1,02) \in (0,15; 0,20) > 0,05 \rightarrow$ aceptaríamos H_0 al 5% y también al 1% porque nuestro p valor es $> 0,01$ y también $> 0,05$
- al 1% concluimos que el desgaste medio del primer material excede en dos unidades (y no más) al desgaste medio del segundo material

❖ Ejemplo 27

En una muestra aleatoria de 500 familias con televisor en una ciudad de Canadá se encuentra que 340 familias están suscritas a HBO.

- a) Calcular un Intervalo de Confianza del 95% para la proporción real de familias suscritas a HBO en toda la ciudad
- b) Queremos reducir la cota del error de estimación a la mitad del que resulta en el apartado a); ¿qué tamaño mínimo de muestra se requiere para conseguirlo?
- c) Se asegura en la publicidad de la cadena que el % de habitantes de la ciudad que tienen suscripción supera el 70%. Contrastar esa hipótesis

- Se mide la proporción de familias en toda la ciudad con televisor que cuentan con suscripción a HBO: π

$$p = \text{proporción muestral de familias con suscripción} = \frac{340}{500} = 0,68$$

a) Calculamos un Intervalo de confianza (0,95) para π

$$\left(p - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}, \quad p + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \right) = \left(0,68 - 1,96 \cdot \sqrt{\frac{0,68 \times 0,32}{500}}, 0,68 + 1,96 \cdot \sqrt{\frac{0,68 \times 0,32}{500}} \right) = \\ = (0,68 \pm 0,04) = (0,64, 0,72)$$

El % de familias con suscripción en dicha ciudad está entre el 64% y el 72% (confianza de 0,95)

Error máximo de estimación: 0,04 (4%)

b) Si queremos reducir a la mitad el error máximo de estimación (cota del error de estimación), debe pasar a ser del 2% (de 0,02)

$$\sqrt{\frac{p(1-p)}{n}} \leq 0,02 \quad 1,96 \sqrt{\frac{0,68 \times 0,32}{n}} \leq 0,02 \quad \longrightarrow \quad n \geq 2.089,83$$

El tamaño muestral mínimo es de 2.090 personas

c) Planteamos un test para contrastar $H_0: \pi = 0,7$ $H_A: \pi < 0,7$

- Calculamos el valor observado sustituyendo en el estadístico de contraste (suponiendo cierta H_0)

$$\frac{(p-\pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} \rightarrow N(0,1)$$

$$Z_{obs} = \frac{(p-\pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0,68-0,7}{\sqrt{\frac{0,7 \times 0,3}{500}}} = -1 \quad \text{p-valor} = p(Z < -1) = 0,1587 > 0,05 \quad \longrightarrow \text{Acept. } H_0$$

Podemos admitir con un error del 5% que la proporción de suscritos a HBO en la ciudad es del 70% (los datos no proporcionan evidencia de que esa proporción sea significativamente menor del 70%)

Examen final MAIS 2 febrero 2021

Parcial 2

Problema 1 [6 ptos]

Para poder estimar la calidad de los materiales de dos tipos de disipadores X e Y se realiza un experimento de estrés al procesador, tratando de sobrecalentarlo de forma continuada con 5 y 6 ordenadores, respectivamente, con disipadores X y disipadores Y . Al finalizar el experimento, se ha medido el tiempo transcurrido hasta el primer fallo. Los resultados que se han obtenido son:

Disipador X : $\bar{x} = 15$ días, $S_x^2 = 16$

Disipador Y : $\bar{y} = 12$ días, $S_y^2 = 16$

- Para poder estimar estadísticamente cuál de los dos dura más, se pide realizar un contraste de hipótesis a un nivel de significación de $1 - \alpha = 0,95$. Razonar si tiene sentido o no el enunciado.
- En los disipadores de tipo X , ¿con cuántos ordenadores tendríamos que realizar el experimento para estimar el tiempo medio poblacional con un error máximo de 1,25 días alrededor de su media muestral? (nivel de significación de $1 - \alpha = 0,95$)
- Suponiendo que se decide realizar otro experimento con 12 ordenadores usando el disipador Y , ¿cuál es la probabilidad de que la cuasivarianza del experimento sea inferior a 4 asumiendo que la cuasivarianza poblacional es 2,25?

▪ Datos:

- $\mu_1 =$ número medio de días antes del 1º fallo en los disipadores de tipo X
- $n_1=5$
- $\bar{x}=15$ días : nº medio de días antes del primer fallo en la muestra de disipadores X $s_x^2=16$
- $\mu_2 =$ número medio de días antes del 1º fallo en los disipadores de tipo Y
- $n_2=6$
- $\bar{y}=12$ días : nº medio de días antes del primer fallo en la muestra de disipadores Y $s_y^2=16$

- Inferencia sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales pequeños (<40)

□ Caso III (σ_1 y σ_2 desconocidas, n_1 ó $n_2 < 40$). Obtenemos el estadístico

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow t_8$$

$$\varepsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{16}{5} + \frac{16}{6}\right)^2}{\frac{\left(\frac{16}{5}\right)^2}{4} + \frac{\left(\frac{16}{6}\right)^2}{5}} = \frac{34,42}{2,56 + 1,42} = 8,65$$

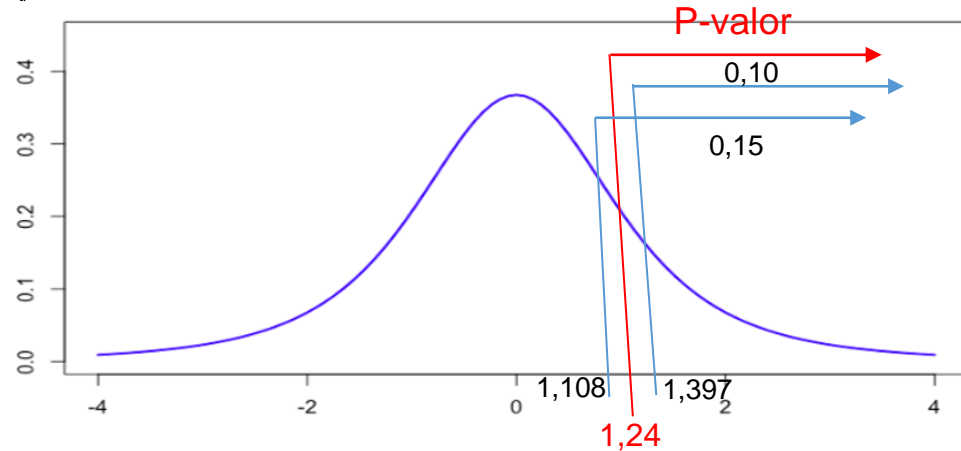
a) Planteamos un test para contrastar Hipótesis:

$H_0: \mu_1 - \mu_2 = 0$ el tiempo medio antes del primer fallo no difiere significativamente en los dos tipos de disipadores
 $H_A: \mu_1 - \mu_2 > 0$ el tiempo medio antes del primer fallo es significativamente mayor en los disipadores de tipo A

- $t_{\varepsilon, 1 - \frac{\alpha}{2}} = t_{8; 0,95} = 2,947$

- Es un contraste sobre diferencia de medias con varianzas poblaciones desconocidas y tamaños muestrales pequeños (<40)

$$t_{8\text{ obs}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3-0}{\sqrt{\frac{16}{5} + \frac{16}{6}}} = 1,24$$



- $p\text{valor} = p(t_8 > 1,24) \in (0,1; 0,15) > 0,05$ Aceptamos H_0 al 5% y también aceptaríamos H_0 al 1% porque nuestro p valor es $> 0,01$
- concluimos que la duración media antes del primer fallo **se puede considerar igual para los dos tipos de disipadores.**

Inferencia sobre la diferencia de medias en poblaciones normales

b) En los disipadores de tipo X, ¿con cuántos ordenadores tendremos que realizar el experimento para estimar el tiempo medio poblacional con un error máximo de 1,25 días? $1 - \alpha = 0,95$

- $IC\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{n_1}, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{n_1}\right)$
- Error máximo de estimación: $Z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{\sqrt{n}} = 1,96x \frac{4}{\sqrt{n}} \leq 1,25 \quad n \geq 20,34$
- Habría que tomar por tanto un mínimo de 40 ordenadores para lograr esos niveles de confianza y precisión.

c) Suponiendo que se realiza el experimento con 12 ordenadores usando el disipador Y ¿cuál es la probabilidad de que la cuasivarianza sea inferior a 4 asumiendo que la varianza poblacional es 2,25?

- $Y \equiv n^\circ$ de días antes del primer fallo de los disipadores tipo Y
- s_y^2 : *cuasi varianza muestral* σ_y^2 es conocida=2,25
- $\frac{(n-1)s^2}{\sigma^2} \rightarrow \chi_{11}^2$
- $p(s_y^2 < 4) = p\left(\frac{(n-1)s_y^2}{\sigma^2} < \frac{11 \times 4}{2,25}\right) = p(\chi_{11}^2 < 19,56) \in (0,9; 0,95)$

