eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea
The University of the Basque Country

E.U.I.T.I. Bilbao

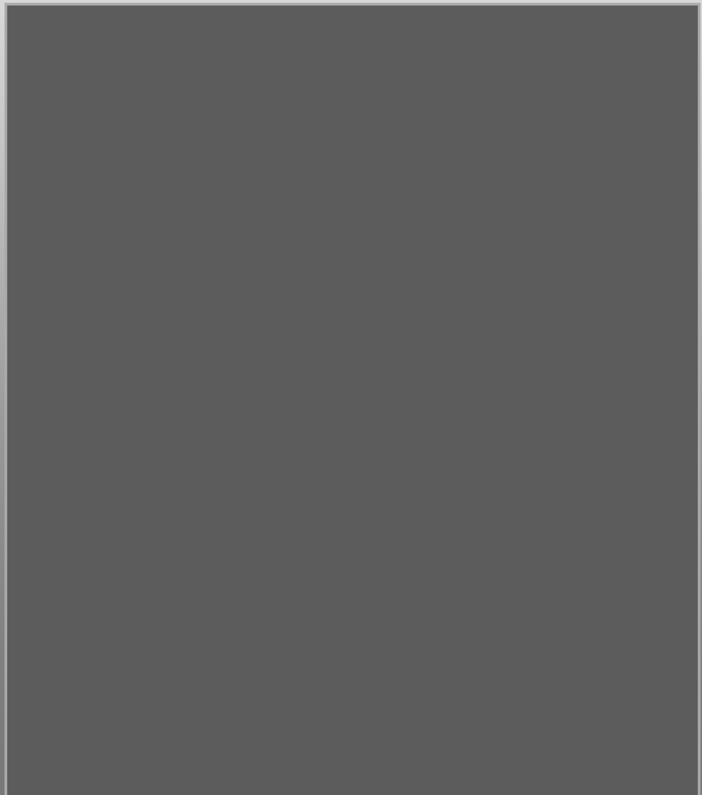
Asignatura: MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

E.U.I.T.I. Bilbao

Asignatura:
MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

TEMA 6: ESTIMACIÓN DE PARÁMETROS DE UNA DISTRIBUCION

¡Datos, datos, datos! – gritó impacientemente. No puedo hacer ladrillos sin arcilla.



Sherlock Holmes

Las aventuras de los bombachos de cobre

Arthur Conan Doyle

Actor: Basil Rathbone

1. RESUMEN

Se describen las técnicas más usuales para estimar la media, la varianza y otros parámetros poblacionales con valores aislados (estimación puntual) o mediante intervalos de confianza.

Palabras clave:

- estimador puntual
- método de los momentos
- método de máxima verosimilitud
- intervalo de confianza
- nivel de confianza

2. ÍNDICE DEL TEMA

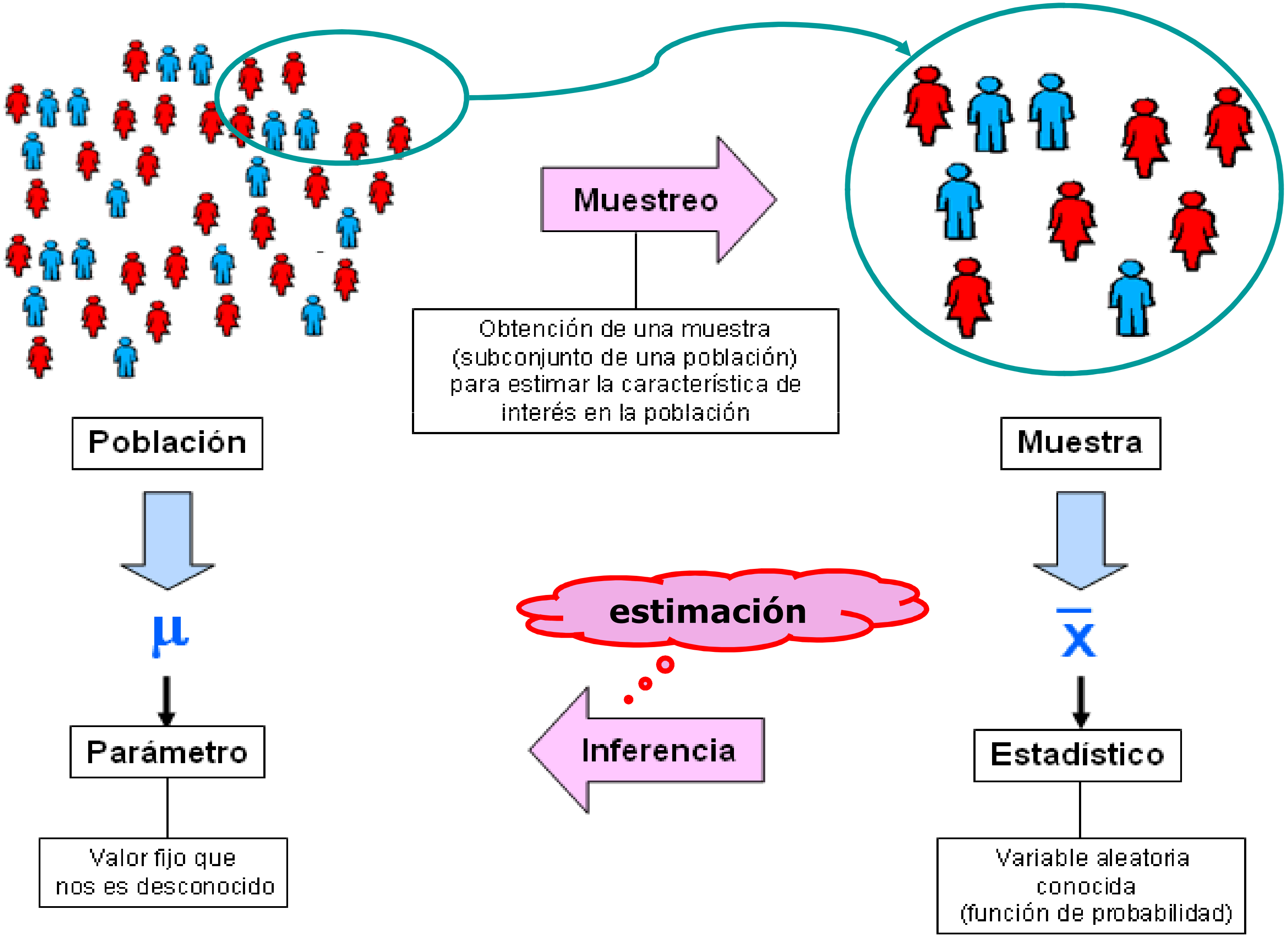
2. ÍNDICE DEL TEMA

3. INTRODUCCIÓN

- en Estadística, una **estimación** consiste en la realización de una inferencia sobre el valor del parámetro de una población usando la información obtenida a través de una muestra

- en Estadística hay tres formas de inferir un valor de un parámetro de una población:
  - **estimación puntual**: regla ó fórmula que permite calcular un valor concreto de ese parámetro basándose en la información contenida en una muestra
  - **estimación por intervalos de confianza**: regla ó fórmula que usa la información muestral para hallar un intervalo ó región de confianza para el valor de dicho parámetro
  - **contraste de hipótesis**: tomando una decisión sobre un valor hipotético del parámetro

3. INTRODUCCIÓN



3. INTRODUCCIÓN

- **ejemplo**: el rendimiento de un equipo de trabajo en una cadena de producción puede estar representado por el número medio de componentes producidas.
  Supóngase que un ingeniero quiere proporcionar información acerca de este promedio en su equipo. Posibilidades:
  - tratar de estimar el promedio de componentes producidas a través de un único valor estimado
  - proporcionar un intervalo de valores en el que tenga mucha confianza de que se encuentre el valor promedio
  - comparar el valor promedio de su equipo con un valor hipotético para, por ejemplo, demostrar que tiene un mejor rendimiento que el promedio general de la empresa
- **notación**:
  - parámetro de la población que se quiere estimar: θ
  - estimador puntual de dicho parámetro: $\widehat{\theta}$

ESTIMACIÓN PUNTUAL

4. ESTIMACIÓN PUNTUAL

<u>Estimador puntual</u>: regla que indica cómo calcular una estimación numérica de un parámetro poblacional desconocido, θ, a partir de los datos de una muestra

- notación: $\widehat{\theta}$

<u>Estimación puntual</u>: número concreto que resulta del cálculo de la regla para una muestra dada

- **ejemplo**: si se desea obtener estimaciones de la media de una v.a. parece lo más lógico utilizar como estimador la media muestral
  - cada media muestral (una por muestra) es una estimación puntual de la media poblacional

4. ESTIMACIÓN PUNTUAL

- entonces, un estimador puntual
  - usa la información aportada por una muestra para calcular un **único valor** que se utiliza para estimar el parámetro de una población
  - como se calcula a partir de una muestra tiene una distribución muestral que describe por completo sus propiedades

- **Ejemplo**. Según el teorema central del límite, bajo ciertas hipótesis:

$$\boxed{\bar{X} \approx N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right)}$$

  - μ: media de la población
  - σ: desviación típica de la población
  - n: tamaño de la muestra

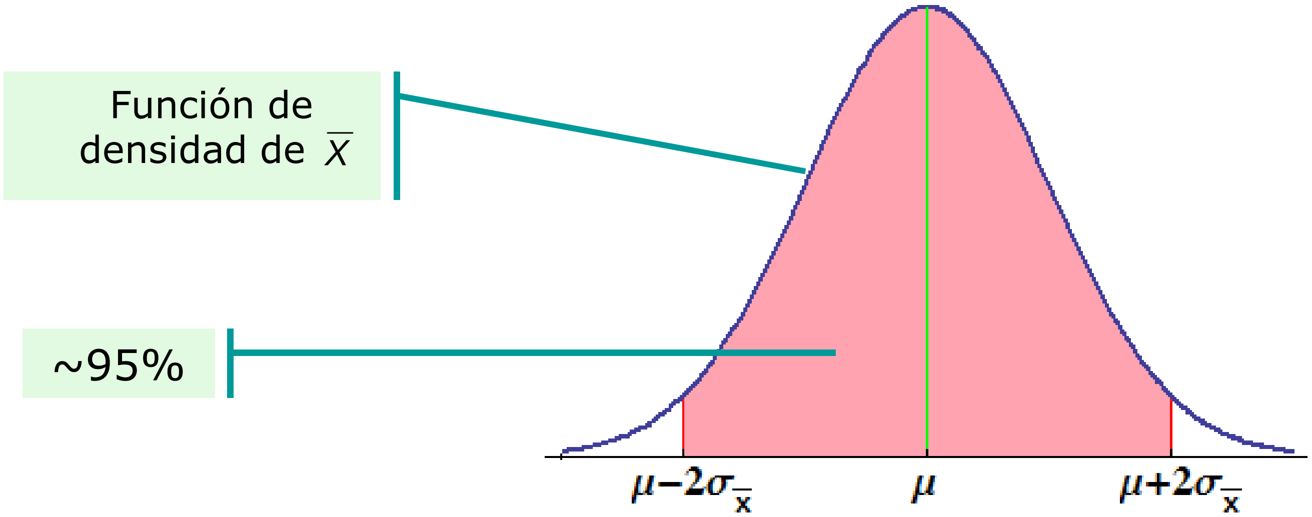
4. ESTIMACIÓN PUNTUAL

► **Ejemplo**.

$$\overline{X} \approx N\left(\mu;\ \frac{\sigma}{\sqrt{n}}\right)$$

- por la simetría de la gráfica de la distribución normal, la media muestral tiene la misma probabilidad de quedar "por encima" ó "por debajo" de la media poblacional

- además: $$P\left(\left|\overline{X}-\mu\right|<\frac{2\sigma}{\sqrt{n}}\right)\approx 0.95$$



4. ESTIMACIÓN PUNTUAL

Propiedades

- Las características mostradas en la figura anterior identifican las dos propiedades más deseables de los estimadores puntuales

- **Propiedad 1**

  - aunque el estimador no proporcione siempre el valor exacto del parámetro, al menos debe *equivocarse* en igual medida por exceso que por defecto
  - lo ideal es que la distribución muestral de un estimador esté centrada en el parámetro que se desea estimar
  - este tipo de estimadores se denominan insesgados

4. ESTIMACIÓN PUNTUAL

Propiedades

▶ Propiedad 1

- un **estimador**, $\widehat{\theta}$, de un parámetro θ se dice **insesgado** si la media de la distribución muestral de dicho estimador puntual es igual que el parámetro de la población que estima, θ

$$\boxed{E\left[\widehat{\theta}\right] = \theta}$$

- en caso contrario, el **estimador** se dice **sesgado**: $\boxed{E\left[\widehat{\theta}\right] \neq \theta}$

- el **sesgo** B de un estimador puntual, $\widehat{\theta}$, se define como la diferencia entre la media de la distribución muestral del estimador puntual y el parámetro estimado, θ

en valor absoluto

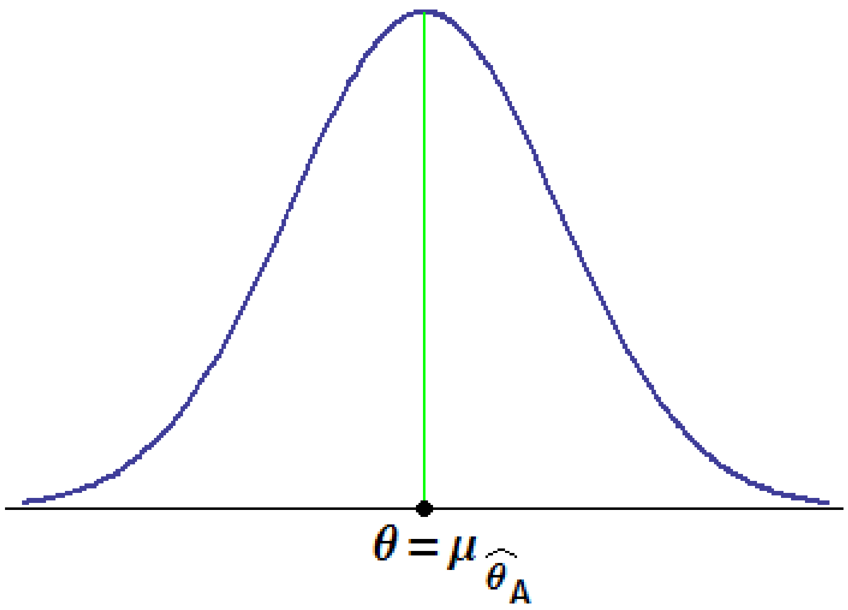
$$\boxed{B\left(\widehat{\theta}\right) = E\left[\widehat{\theta}\right] - \theta} \longrightarrow \boxed{B\left(\widehat{\theta}\right) = \left| E\left[\widehat{\theta}\right] - \theta \right|}$$

4. ESTIMACIÓN PUNTUAL

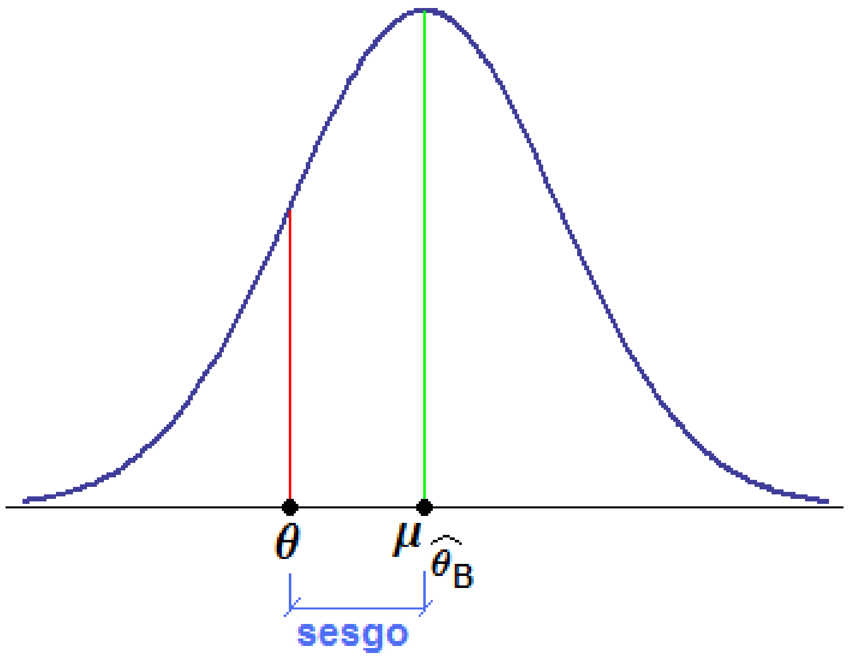
Propiedades

► **Propiedad 1**

distribución muestral de un estimador insesgado $\hat{\theta}_A$ de un parámetro θ



distribución muestral de un estimador sesgado $\hat{\theta}_B$ de un parámetro θ



4. ESTIMACIÓN PUNTUAL

Propiedades

Propiedad 2

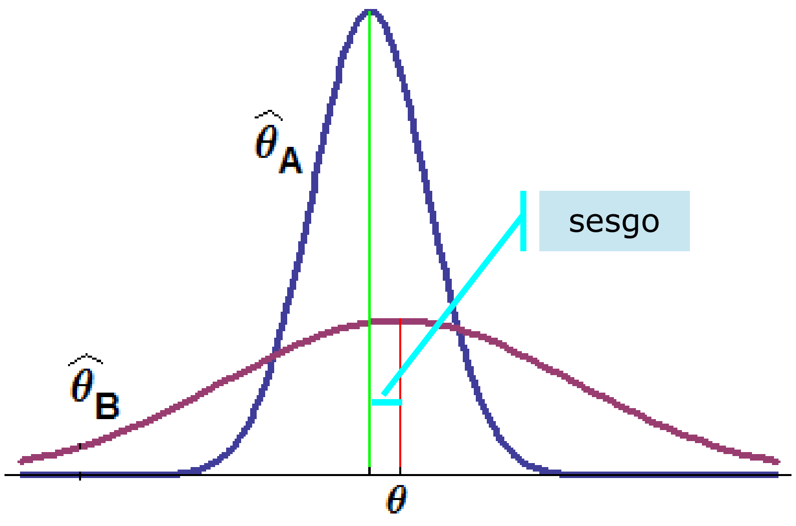
- además de la falta de sesgo, es deseable que la distribución muestral de un estimador tenga varianza mínima
- es decir, que la dispersión de la distribución muestral sea lo más pequeña posible con lo que las estimaciones tiendan a ser cercanas al valor del parámetro θ que se estima
- **estimador insesgado de mínima varianza** de un parámetro θ es aquél estimador, $\widehat{\theta}$, que tiene la varianza más pequeña de entre todos los estimadores insesgados
- **error estándar de un estimador:** desviación típica de dicho estimador

4. ESTIMACIÓN PUNTUAL

Propiedades

► Propiedad 2

- no siempre es fácil encontrar este estimador y, en ocasiones, se admite un ligero sesgo con tal de que la varianza del estimador sea mínima
- en casos así, se elige el estimador que minimiza el error cuadrático medio



4. ESTIMACIÓN PUNTUAL

Propiedades

Propiedad 2

- **error cuadrático medio de un estimador** ($\widehat{\theta}$): media del cuadrado de la desviaciones entre el estimador puntual y el parámetro estimado, θ

$$\boxed{MSE\left(\widehat{\theta}\right)=E\left[\left(\widehat{\theta}-\theta\right)^2\right]}$$

- se puede demostrar que: $\boxed{MSE\left(\widehat{\theta}\right)=Var\left[\widehat{\theta}\right]+\left(B\left(\widehat{\theta}\right)\right)^2}$

- entonces, si $\widehat{\theta}$ es un estimador puntual insesgado se tiene:

$$\boxed{MSE\left(\widehat{\theta}\right)=Var\left[\widehat{\theta}\right]}$$

4. ESTIMACIÓN PUNTUAL

Propiedades

► **Ejemplo.** Para estimar la cantidad media reclamada por incendio en pisos de tamaño medio se ha muestreado los ficheros de una compañía de seguros y se han obtenido las siguientes cantidades reclamadas en 10 incendios medidas en miles de euros:

12.1, 5.5, 6.3, 1.2, 8.0, 14.1, 4.2, 5.1, 6.6, 10.3

Solución

- estimador (insesgado) de la media de cantidades reclamadas:

$$\overline{X} = \frac{12.1+5.5+6.3+1.2+8.0+14.1+4.2+5.1+6.6+10.3}{10} = \frac{73.4}{10} = 7.34$$

- se estima que la cantidad media reclamada por incendio es de 7340€ ya que, como se ha visto: $\boxed{E\left[\overline{X}\right] = \mu}$

4. ESTIMACIÓN PUNTUAL

Propiedades

▶ **Ejemplo.** Para estimar la cantidad media reclamada por incendio en pisos de tamaño medio se ha muestreado los ficheros de una compañía de seguros y se han obtenido las siguientes cantidades reclamadas en 10 incendios medidas en miles de euros:

12.1, 5.5, 6.3, 1.2, 8.0, 14.1, 4.2, 5.1, 6.6, 10.3

Solución

- error estándar: $$SD\left[\overline{X}\right] = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- ¿cómo puede reducirse el error estándar a la mitad?
  - ◇ multiplicando por 4 el tamaño de la muestra

4. ESTIMACIÓN PUNTUAL

Estimación de la media de una v.a.

- sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de un v.a. X normal con media μ y desviación típica σ
- como ya se ha visto:
  - la media muestral es un estimador insesgado de la media de X

$$\boxed{\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \Rightarrow E\left[\overline{X}\right] = \mu}$$

  - el error estándar es:

$$\boxed{SD\left[\overline{X}\right] = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

4. ESTIMACIÓN PUNTUAL

Estimación de la varianza de una v.a.

- sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una v.a. X normal con media μ y desviación típica σ
- se demuestra que:
  - la cuasivarianza muestral es un estimador insesgado de la varianza de la v.a. X

$$\boxed{\hat{s}^2 = s_{n-1}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i - \overline{X}\right)^2 \Rightarrow E\left[S^2\right] = \sigma^2}$$

4. ESTIMACIÓN PUNTUAL

Estimación de la varianza de una v.a.

▶ **Ejemplo (ejercicio 1).** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una población normal con media μ y varianza σ^2

1. demostrar que la cuasivarianza muestral es un estimador insesgado de la varianza de la población
2. ¿es la varianza muestral un estimador insesgado de la varianza de la población?

Solución

- en el tema *Distribuciones muestrales* se vio:

$$\boxed{\widehat{s}^{\,2} = \frac{n}{n-1} \cdot s^2}$$

$$\boxed{\frac{(n-1)\cdot \widehat{s}^{\,2}}{\sigma^2} \approx \chi^2_{n-1}}$$

4. ESTIMACIÓN PUNTUAL

Estimación de la varianza de una v.a.

▶ **Ejemplo (ejercicio 1).**

Solución

1.

$$\boxed{\frac{(n-1)\cdot \widehat{s}^{\,2}}{\sigma^2} \approx \chi^2_{n-1}}$$

$$\left.\begin{aligned} E\left[\frac{(n-1)\cdot \widehat{s}^{\,2}}{\sigma^2}\right] &= E\left[\chi^2_{n-1}\right] = n-1 \\ E\left[\frac{(n-1)\cdot \widehat{s}^{\,2}}{\sigma^2}\right] &= \frac{(n-1)}{\sigma^2}\cdot E\left[\widehat{s}^{\,2}\right] \end{aligned}\right\} \Rightarrow \frac{(n-1)}{\sigma^2}\cdot E\left[\widehat{s}^{\,2}\right] = n-1$$

$$\Downarrow$$

$$\boxed{E\left[\widehat{s}^{\,2}\right] = \sigma^2}$$

4. ESTIMACIÓN PUNTUAL

Estimación de la varianza de una v.a.

▶ **Ejemplo (ejercicio 1).**

Solución

2.

$$\left.\begin{aligned} &\frac{(n-1)\cdot \hat{s}^2}{\sigma^2} \approx \chi^2_{n-1} \\ &\hat{s}^2 = \frac{n}{n-1}\cdot s^2 \end{aligned}\right\} \Rightarrow \boxed{\frac{n\cdot s^2}{\sigma^2} \approx \chi^2_{n-1}}$$

$$\left.\begin{aligned} &E\left[\frac{n\cdot s^2}{\sigma^2}\right] = E\left[\chi^2_{n-1}\right] = n-1 \\ &E\left[\frac{n\cdot s^2}{\sigma^2}\right] = \frac{n}{\sigma^2}\cdot E\left[s^2\right] \end{aligned}\right\} \Rightarrow \frac{n}{\sigma^2}\cdot E\left[s^2\right] = n-1$$

$$\Downarrow$$

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

La varianza muestral NO es un estimador insesgado de σ^2

4. ESTIMACIÓN PUNTUAL

Estimación de una proporción poblacional

- se desea estimar una proporción desconocida, *p*, que representa la probabilidad de un suceso dentro de un espacio muestral
- sea $X_1, X_2, \ldots, X_n$ una muestra de resultados del experimento asociado al espacio muestral denotando $X_i=1$ el éxito del suceso en el *i*-ésimo experimento y $X_i=0$ el fracaso
- en este caso:
  - estimador insesgado de *p*: $$\boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^{i=n} X_i}$$
  - error estándar: $$\boxed{s.e.(\hat{p}) = SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}}$$

4. ESTIMACIÓN PUNTUAL

Estimación de una proporción poblacional

- error estándar:

$$\boxed{s.e.\left(\widehat{p}\right)=SD\left[\widehat{p}\right]=\sqrt{\frac{p\left(1-p\right)}{n}}}$$

- no puede evaluarse ya que *p* es desconocido
- si el tamaño de la muestra, *n*, es grande se utiliza el valor del estimador en lugar de *p* en la expresión que da el error estándar
- cota superior:

$$si\ \ 0\leq p\leq 1: \quad p\left(1-p\right)\leq \tfrac{1}{4} \Rightarrow \boxed{s.e.\left(\widehat{p}\right)\leq\sqrt{\frac{1}{4n}}=\frac{1}{2\sqrt{n}}}$$

4. ESTIMACIÓN PUNTUAL

Estimación de una proporción poblacional

- **Ejemplo.** En una muestra de 1000 individuos de una población el número de varones es de 510.
  - estimación de la proporción de varones en la población total:

$$\widehat{p} = \frac{510}{1000} = 0.51$$

  - cota superior del error estándar: $s.e.\left(\widehat{p}\right) \leq \dfrac{1}{2\sqrt{1000}} = 0.01581139$
  - estimación del error estándar:

$$s.e.\left(\widehat{p}\right) \approx \sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}} = 0.0158082$$

5. OBTENCIÓN DE ESTIMADORES PUNTUALES: MÉTODOS

- se ha estimado la media y la varianza de una v.a. (población) con una distribución normal mediante la media muestral y la cuasivarianza muestral, respectivamente
- sin embargo, hay distribuciones teóricas que no dependen directamente de la media o la varianza:
  - la binomial depende de p
  - la gamma de k y λ
- existen varias maneras de estimar estos parámetros, se van a ver dos de los más sencillos:
  - método de los momentos
  - método de máxima verosimilitud

6. MÉTODO DE LOS MOMENTOS

► se va a explicar el método sólo para distribuciones de uno o dos parámetros poblacionales, que son del tipo de las que se han visto

► sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una v.a. X

1. si la distribución de X depende de un parámetro, θ

- la media poblacional de X es función θ: $\boxed{E[X] = \mu = f(\theta)}$
- el estimador del parámetro θ, $\widehat{\theta}$, propuesto por el método de los momentos se obtiene despejándolo (si es posible) de la ecuación:

$$\boxed{\overline{x} = f\left(\widehat{\theta}\right)}$$

6. MÉTODO DE LOS MOMENTOS

- sea X_1, X_2, … , X_n una muestra aleatoria simple de una v.a. X

1. si la distribución de X depende de un solo parámetro θ

- **ejemplo**: sea $X \sim B(n;p)$
  - media de la población binomial: $\boxed{E[X] = \mu = n \cdot p = f(p)}$
  - estimador de p propuesto por el método:

$$\boxed{\bar{x} = n \cdot \hat{p} = f(\hat{p}) \Rightarrow \hat{p} = \frac{\bar{x}}{n}}$$

- **ejemplo**: sea $X \sim \mathcal{E}(\lambda)$
  - media de la población exponencial: $\boxed{E[X] = \mu = \frac{1}{\lambda} = f(\lambda)}$
  - estimador de λ propuesto por el método:

$$\boxed{\bar{x} = \frac{1}{\hat{\lambda}} = f(\hat{\lambda}) \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}}$$

6. MÉTODO DE LOS MOMENTOS

► sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una v.a. X

2. si la distribución de X depende de dos parámetros, θ_1 y θ_2

- la media poblacional de X es función de ambos:

$$\boxed{E[X] = \mu = f(\theta_1, \theta_2)}$$

- igualmente, la varianza poblacional de X está expresada como función de esos parámetros:

$$\boxed{Var[X] = \sigma^2 = g(\theta_1, \theta_2)}$$

- según este método, los estimadores de θ_1 y θ_2 , $\widehat{\theta}_1$ y $\widehat{\theta}_2$, se obtienen despejándolos (si es posible) del sistema de ecuaciones:

$$\boxed{\begin{cases} \overline{x} = f(\widehat{\theta}_1, \widehat{\theta}_2) \\ s_{n-1}^2 = g(\widehat{\theta}_1, \widehat{\theta}_2) \end{cases}}$$

6. MÉTODO DE LOS MOMENTOS

▶ sea X_1, X_2, … , X_n una muestra aleatoria simple de una v.a. X

2. si la distribución de X depende de dos parámetros, θ_1 y θ_2

- **ejemplo**: sea $X \sim Gamma(k;\lambda)$
  - ◇ media de la población: $\boxed{E[X] = \mu = \frac{k}{\lambda} = f(k,\lambda)}$
  - ◇ varianza de la población: $\boxed{Var[X] = \sigma^2 = \frac{k}{\lambda^2} = g(k,\lambda)}$
  - ◇ estimadores de k y λ propuestos por el método:

$$\left.\begin{array}{l} (1):\ \bar{x} = f\left(\hat{k},\hat{\lambda}\right) = \dfrac{\hat{k}}{\hat{\lambda}} \\[2ex] (2):\ s_{n-1}^2 = g\left(\hat{k},\hat{\lambda}\right) = \dfrac{\hat{k}}{\hat{\lambda}^2} \end{array}\right\} \Rightarrow \frac{(1)}{(2)}:\ \boxed{\hat{\lambda} = \frac{\bar{x}}{s_{n-1}^2} \xrightarrow{a\,(1)} \hat{k} = \frac{\bar{x}^2}{s_{n-1}^2}}$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

► el método obedece a un principio muy lógico: dada una muestra se escogen como estimadores aquellos valores de los parámetros que hagan *más creíbles* (o *más verosímiles*) los datos de la muestra

► en el siguiente ejemplo se ven las ideas básicas del método de máxima verosimilitud que nos llevarán a la definición general

► **Ejemplo.** Se sabe que en una urna hay 4 bolas (que pueden ser blancas o negras). La proporción de bolas blancas, θ, es desconocida y puede tomar los valores:

$$\theta = \left\{0, \tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}, 1\right\}$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

► **Ejemplo.** Para obtener más información, se extraen de la urna 2 bolas con reemplazamiento (con lo que se garantiza la independencia de las observaciones). Supóngase que la primera bola es blanca y la segunda negra, denotándolo como *BN*.

$$\boxed{\theta = \left\{0, \tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}, 1\right\}}$$

- la probabilidad de obtener, precisamente, esa muestra depende de la proporción de bolas blancas en la urna, θ
- la idea del método de máxima verosimilitud es muy sencilla y razonable; tomar como estimación de θ aquel valor que da la mayor probabilidad a la muestra obtenida

7. MÉTODO DE MÁXIMA VEROSIMILITUD

▶ **Ejemplo.** $\theta = \left\{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\right\}$

- se calcula la probabilidad de obtener la muestra *BN*

$$P\left(BN \mid \theta\right) = \begin{cases} 0 & si \quad \theta = 0 \\ \dfrac{1}{4}\cdot\dfrac{3}{4} = \dfrac{3}{16} & si \quad \theta = \dfrac{1}{4} \\ \dfrac{2}{4}\cdot\dfrac{2}{4} = \dfrac{4}{16} & si \quad \boxed{\theta = \dfrac{2}{4}} \\ \dfrac{3}{4}\cdot\dfrac{1}{4} = \dfrac{3}{16} & si \quad \theta = \dfrac{3}{4} \\ 0 & si \quad \theta = 1 \end{cases}$$

Mayor probabilidad para la muestra obtenida

Si la muestra obtenida es *BN* el estimador de máxima verosimilitud es

$$\widehat{\theta} = \frac{1}{2}$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

- sea x_1, x_2, … , x_n una muestra aleatoria simple con n observaciones de una v.a. X
  - si X es **discreta**:
    - función de masa de probabilidad: ***p(x)***
    - función de probabilidad de la muestra:

$$\boxed{p(x_1, x_2, \cdots, x_n) = p(x_1) \cdot p(x_2) \cdot \cdots \cdot p(x_n)}$$

  - si X es **continua**:
    - función de densidad de probabilidad: ***f(x)***
    - función de densidad de la muestra:

$$\boxed{f(x_1, x_2, \cdots, x_n) = f(x_1) \cdot f(x_2) \cdot \cdots \cdot f(x_n)}$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

► sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria simple con n observaciones de una v.a. X

► la función de probabilidad (masa ó densidad) depende de k parámetros denotados:

$$\boxed{\theta = \left(\theta_1, \theta_2, \cdots, \theta_k\right)}$$

► **estimador de máxima verosimilitud de $\boldsymbol{\theta}$** es el formado por los valores:

$$\boxed{\widehat{\theta} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k\right)}$$

► ese estimador maximiza la función de verosimilitud

7. MÉTODO DE MÁXIMA VEROSIMILITUD

► **función de verosimilitud**, ***L*(*θ*):** función de probabilidad de la muestra (masa ó densidad)

- si *X* es **discreta**:

$$\boxed{L(\theta) = p(x_1)\cdot p(x_2)\cdot \;\cdots\; p(x_n)}$$

- si *X* es **continua**:

$$\boxed{L(\theta) = f(x_1)\cdot f(x_2)\cdot \;\cdots\; f(x_n)}$$

- **Nota.** Como la estimación de θ es el valor que da la mayor probabilidad a la muestra, será aquel valor que maximice la función de verosimilitud.

7. MÉTODO DE MÁXIMA VEROSIMILITUD

Procedimiento

1. escribir la función de verosimilitud, $L(\theta)$

- si X es **discreta**: $$L(\theta)=\prod_{i=1}^{i=n} p(x_i)$$
- si X es **continua**: $$L(\theta)=\prod_{i=1}^{i=n} f(x_i)$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

Procedimiento

2. maximizar la función de verosimilitud, $L(\theta)$

- **Nota.** Dado que el máximo de una función coincide con el de su logaritmo para el mismo valor de la variable resulta muy útil, en general, maximizar el logaritmo de la función $L(\theta)$ en lugar de la propia función de verosimilitud

- si X es **discreta**: $$ln\left(L(\theta)\right) = \sum_{i=1}^{i=n} ln\left(p\left(x_i\right)\right)$$

- si X es **continua**: $$ln\left(L(\theta)\right) = \sum_{i=1}^{i=n} ln\left(f\left(x_i\right)\right)$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

Ejemplo

Obtener el estimador de máxima verosimilitud del parámetro λ de una distribución de Poisson

Solución

- X es discreta: $X \sim P(\lambda)$
- función de masa de probabilidad:

$$P\left(X = i\right) = \frac{e^{-\lambda} \cdot \lambda^{i}}{i\,!} \quad \left(i = 0, 1, 2, \ \ldots\right)$$

- sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria simple de la v.a. X con n observaciones

7. MÉTODO DE MÁXIMA VEROSIMILITUD

Ejemplo

Obtener el estimador de máxima verosimilitud del parámetro λ de una distribución de Poisson

Solución

- función de verosimilitud, $L(\lambda)$:

$$P\left(X = x_i\right) = p\left(x_i\right) = \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i\,!} \quad \left(i = 1, 2, \ldots, n\right)$$

$$\boxed{L\left(\lambda\right) = \prod_{i=1}^{i=n} p\left(x_i\right) = e^{-n\lambda} \frac{\cdot\, \lambda^{\sum_{i=1}^{i=n} x_i}}{\prod_{i=1}^{i=n} \left(x_i\,!\right)}}$$

7. MÉTODO DE MÁXIMA VEROSIMILITUD

Solución

- logaritmo neperiano de la función de verosimilitud, ln[$L(\lambda)$]:

$$ln\left[L\left(\lambda\right)\right] = -n\lambda + \left(\sum_{i=!}^{i=n} x_i\right) \cdot ln\,\lambda - ln\left[\prod_{i=1}^{i=n}\left(x_i\,!\right)\right]$$

- maximización de la función, ln[$L(\lambda)$]:

$$\frac{d}{d\lambda}\left\{ln\left[L\left(\lambda\right)\right]\right\} = -n + \frac{1}{\lambda} \cdot \left(\sum_{i=!}^{i=n} x_i\right) = 0$$

- estimador de máxima verosimilitud:

$$\boxed{\lambda = \frac{1}{n} \sum_{i=!}^{i=n} x_i = \overline{x}}$$

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

8. ESTIMACIÓN POR INTERVALOS

- cuando se estima un parámetro mediante un estimador puntual no se puede esperar que el estimador sea exactamente igual al parámetro sino que esté *próximo*
- en la práctica, además de dar una estimación puntual de un parámetro, interesa facilitar un intervalo que permita precisar la incertidumbre existente en la estimación
- un **estimador por intervalo** es una regla ó fórmula que, usando la información muestral, permite hallar dos valores que definen un intervalo utilizado como estimación del parámetro poblacional (se predice que el parámetro está contenido en él) con cierto grado de confianza

9. INTERVALOS DE CONFIANZA

- sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria simple con n observaciones de una v.a. X cuya distribución depende de un parámetro desconocido θ

Un **intervalo de confianza para $\boldsymbol{\theta}$** con un **nivel de significación $\boldsymbol{\alpha}$**, $\boldsymbol{I}(x_1, x_2, \ldots, x_n)$, es un intervalo real que depende de la muestra, pero que no depende de θ, tal que:

$$\boxed{P\left[\theta \in I\left(x_1, x_2, \cdots, x_n\right)\right] = 1-\alpha}$$

- **nivel** (o coeficiente) **de confianza**: 1-α
- para cada muestra particular se tiene un intervalo de confianza:

$$\boxed{I\left(x_1, x_2, \cdots, x_n\right) = \left(T_1\left(x_1, x_2, \cdots, x_n\right), T_2\left(x_1, x_2, \cdots, x_n\right)\right)}$$

9. INTERVALOS DE CONFIANZA

► **Observación:**

- el objetivo de un intervalo de confianza es proporcionar, en base a los datos de la muestra, una región en la que se tenga un determinado nivel de confianza de que se encuentre el parámetro
- como en el caso de los estimadores puntuales, el intervalo de confianza es aleatorio ya que depende de los valores de la muestra
- además, se da por hecho que existe la posibilidad de que el verdadero parámetro θ no quede situado dentro del intervalo de confianza, algo que ocurre con probabilidad α

9. INTERVALOS DE CONFIANZA

▶ **<u>Procedimiento</u>:**

- identificar un **estadístico pivotal**, es decir, un estadístico que depende de los valores muestrales y del parámetro que se está estimando
- seleccionar un **coeficiente de confianza** (denotado $1-\alpha$)
  - ◇ normalmente: $\alpha=0.05$ ó $\alpha=0.01$
  - ◇ suele considerarse en tanto por ciento: $(1-\alpha)\times 100\%$
- construir el intervalo de confianza

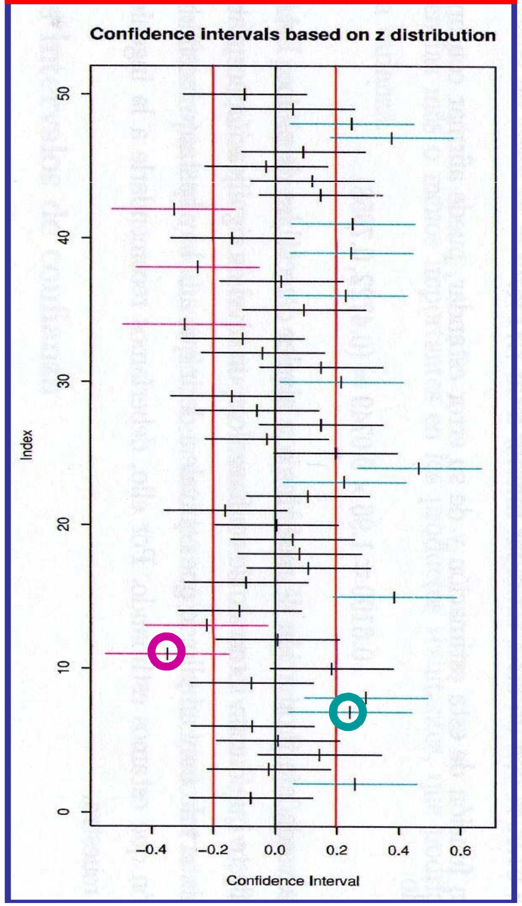
10. NIVEL DE CONFIANZA

- **<u>Nota</u>:**
  - adoptando una interpretación frecuentista de la probabilidad, un intervalo de confianza al 95%, por ejemplo, garantiza que si se toman 100 muestras el parámetro poblacional estará en el interior del intervalo en, aproximadamente, 95 de intervalos construidos
  - en la práctica, esto resulta absurdo porque no se tienen 100 muestras sino solamente una
  - entonces, a partir de los datos de una muestra se construye un intervalo de confianza; caben dos posibilidades:
    - el parámetro está dentro del intervalo o no lo está
  - el parámetro es constante y el intervalo también (no se repite el experimento); por ello, se habla de intervalos de confianza interpretando que se tiene una *confianza* del 95% en que el parámetro estará dentro del intervalo

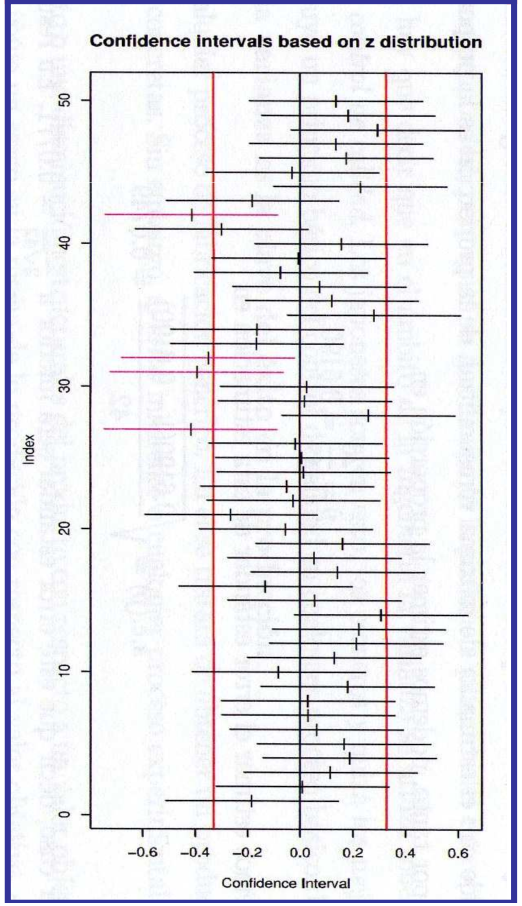
10. NIVEL DE CONFIANZA

EJEMPLO: intervalo de confianza para una media

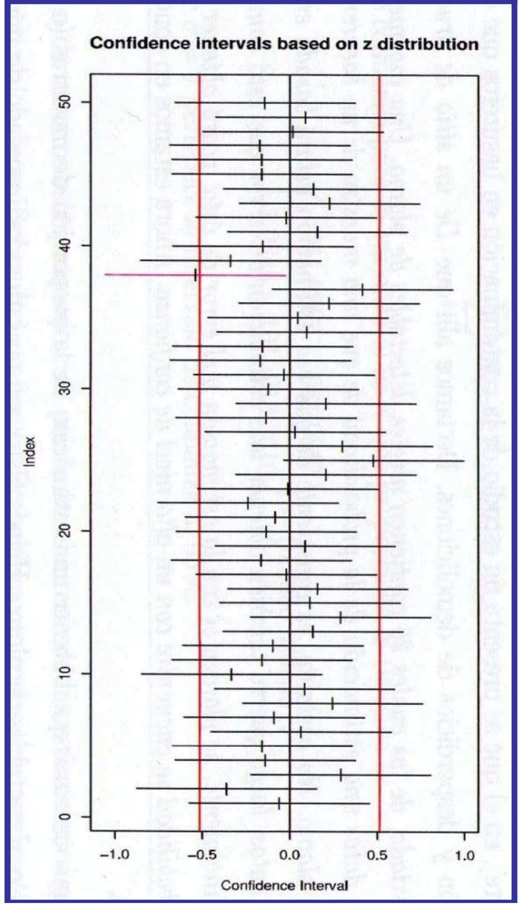
nivel de confianza $(1-\alpha)\cdot 100\% = 68\%$



nivel de confianza $(1-\alpha)\cdot 100\% = 90\%$



nivel de confianza $(1-\alpha)\cdot 100\% = 99\%$



http://www4.ujaen.es/~svilchez/metodos/Apuntes.pdf

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- **Hipótesis adicionales**
  - población normal
  - media, μ, desconocida
  - desviación típica, σ, conocida
  - nivel de confianza: $1-\alpha$

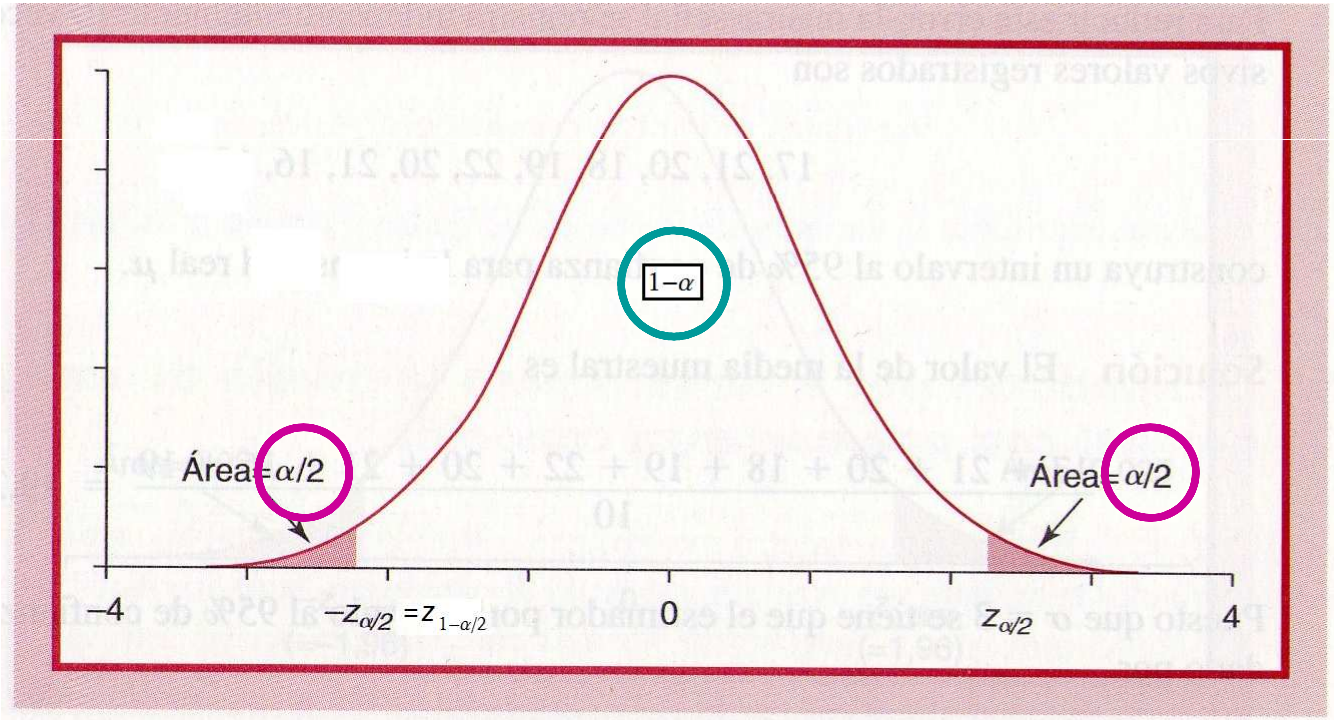
$$\overline{X} \approx \left(\mu ; \frac{\sigma}{\sqrt{n}} \right)$$

- Sea $X_1, X_2, \ldots, X_n$ una muestra de tamaño n de la población
- Tipificando: $\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} = Z \approx N\left(0;1\right)$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- se verifica $\forall\, \alpha \in [0,1]$ (ver figura): $P\left(|Z| \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$



$P\{|Z| \le z_{\alpha/2}\} = P\{-z_{\alpha/2} \le Z \le z_{\alpha/2}\} = 1 - \alpha.$

Ross, M.S.; Introducción a la Estadística; Ed. Reverté S.A. (2005)

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- multiplicando ambos miembros de la desigualdad por

$$\frac{\sigma}{\sqrt{n}}$$

$$P\left(|Z| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha \longrightarrow P\left(\left|\overline{X} - \mu\right| \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- por tanto, con probabilidad 1-α la media poblacional μ está en el intervalo:

$$\boxed{\overline{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- se denomina estimador por intervalo al 100(1-α)% de confianza al intervalo:

$$\boxed{\left(\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)}$$

- si el valor observado de la media muestral es $\overline{x}$ el intervalo de confianza para la media poblacional, μ, es, con una confianza del 100(1-α)%:

$$\boxed{\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)}$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

► **Ejemplo.** Si se quiere obtener un estimador por intervalo con 95% de confianza

$$P\left(|Z| \le z_{\frac{\alpha}{2}}\right) = 0.95$$

?

Excel: =DISTR.NORM.ESTAND.INV(0,975)=1.95996398~1.96

$$P\left(|Z| \le z_{\frac{\alpha}{2}}\right) = 0.95 \quad \Rightarrow P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right) = 0.95$$

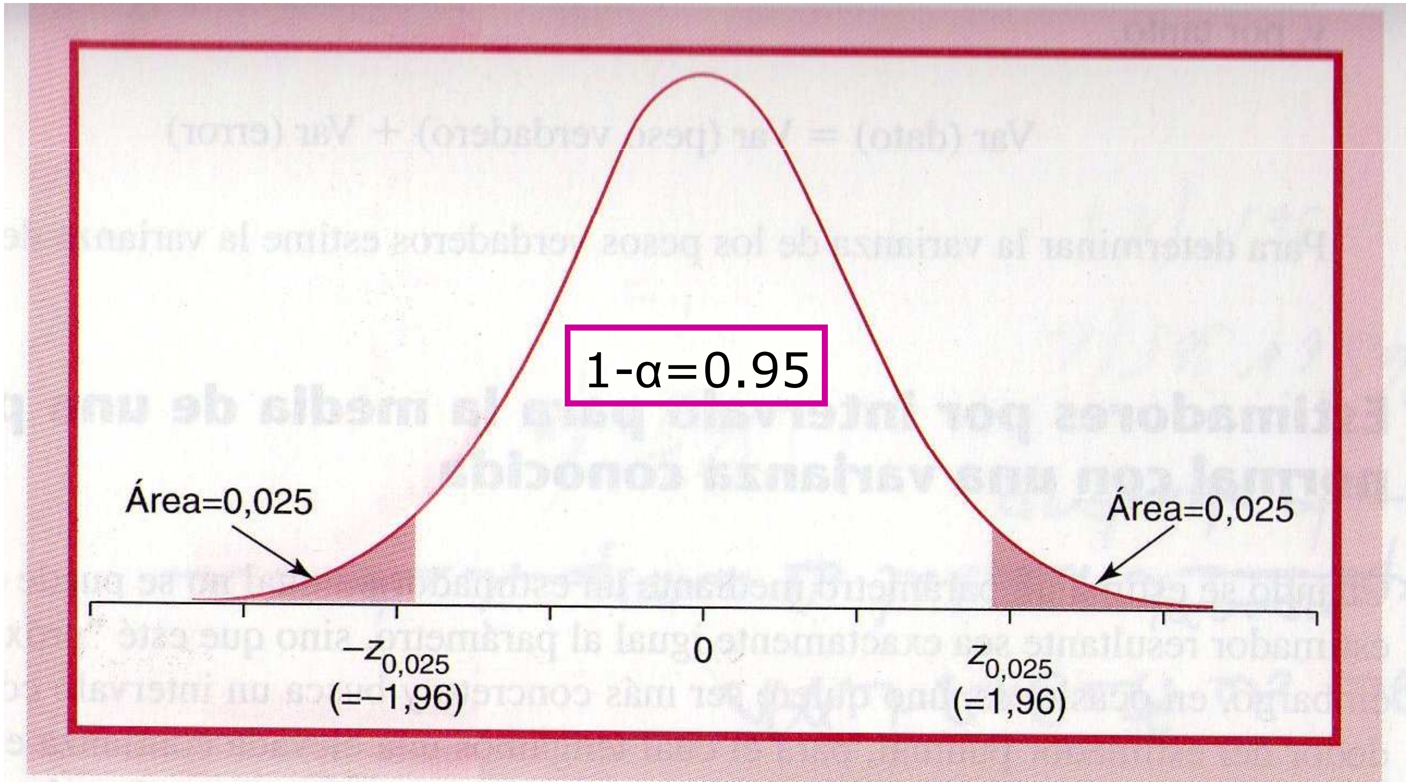
$$\boxed{\times \frac{\sigma}{\sqrt{n}}}$$

$$P\left(|\overline{X} - \mu| \le 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- **Ejemplo.** Si se quiere obtener un estimador por intervalo con 95% de confianza



$$P\{|Z| \leq 1{,}96\} = P\{\,-1{,}96 \leq Z \leq 1{,}96\} = 0{,}95.$$

Ross, M.S.; Introducción a la Estadística; Ed. Reverté S.A. (2005)

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

► **Ejemplo.** Si se quiere obtener un estimador por intervalo con 95% de confianza

$$P\left(\left| \overline{X} - \mu \right| \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

$$P\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

- entonces, con un 95% de confianza, la media poblacional, μ, se encuentra en el intervalo:

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \, , \; \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right)$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ **Ejemplo.** Una señal de intensidad μ emitida desde un punto A se registra en un punto B con una intensidad que se distribuye según una normal de media μ y σ=3 (esto es, debido al "ruido" la intensidad registrada difiere de la real en una cantidad que sigue la distribución normal indicada). Para reducir el error, la misma señal se registra, independientemente, 10 veces tal que los valores registrados son:

17, 21, 20, 18, 19, 22, 20, 21, 16, 19

Construir intervalos al 90%, 95% y 99% de confianza para la intensidad real μ

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ **Ejemplo.**

Solución

- media muestral : $\bar{x} = \dfrac{17+21+20+18+19+22+20+21+16+19}{10} = 19.3$

- confianza del **95%**:

$$\left(19.3 - 1.96 \cdot \frac{3}{\sqrt{10}}\ ,\ 19.3 + 1.96 \cdot \frac{3}{\sqrt{10}}\right) = (19.3 - 1.86\ ,\ 19.3 + 1.86) = (17.44\ ,\ 21.16)$$

- confianza del **90%**:

Excel: =DISTR.NORM.ESTAND.INV(0,95)=1.64485363~1.645

$$\left(19.3 - 1.645 \cdot \frac{3}{\sqrt{10}}\ ,\ 19.3 + 1.645 \cdot \frac{3}{\sqrt{10}}\right) = (17.74\ ,\ 20.86)$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ **Ejemplo.**

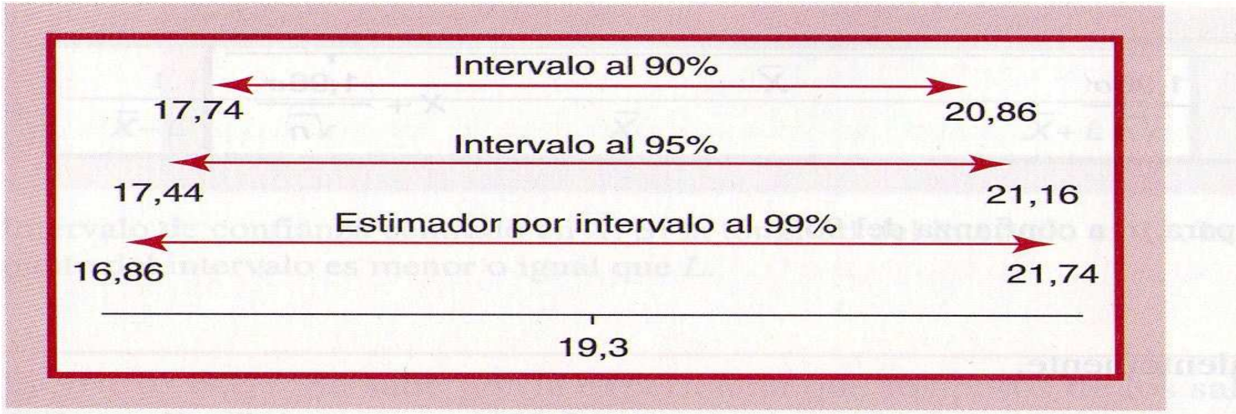
Solución

- confianza del **99%**:

Excel: =DISTR.NORM.ESTAND.INV(0,995)=2.5758293~2.576

$$\left(19.3 - 2.576 \cdot \frac{3}{\sqrt{10}}\ ,\ 19.3 + 2.576 \cdot \frac{3}{\sqrt{10}}\right) = (16.86\ ,\ 21.74)$$

Ross, M.S.; Introducción a la Estadística; Ed. Reverté S.A. (2005)



Estimadores por intervalo, con confianzas del 90, el 95 y el 99 por ciento.

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- **Hipótesis adicionales**
  - población normal
  - media, μ, desconocida
  - desviación típica, σ, desconocida
  - tamaño muestral pequeño: $n \leq 30$
  - nivel de confianza: 1-α

$$\Rightarrow \boxed{\overline{X} \approx \left(\mu; \frac{\sigma}{\sqrt{n}}\right)}$$

- en la práctica es poco probable que se desconozca μ y se conozca σ con lo que la aplicación del teorema es limitada
- el siguiente resultado responde a la necesidad de extender el resultado anterior cuando no se conoce σ

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- cuando el valor de σ es desconocido, en la anterior fórmula de un intervalo de confianza para la media, puede utilizarse para aproximarlo la cuasidesviación típica muestral, $\hat{s}$
- esa aproximación es bastante buena siempre que n sea suficientemente grande; si el tamaño muestral es pequeño la aproximación de la desviación típica poblacional por la cuasidesviación típica muestral no es buena
- para compensar esa mala aproximación se tipifica para utilizar la distribución t de Student:

$$\frac{\overline{X}-\mu}{\hat{s}/\sqrt{n}} = t_{n-1}$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- sea $X_1, X_2, \ldots, X_n$ una muestra de tamaño n de la población indicada
- entonces: $$\boxed{P\left(\mu \in \left(\bar{x} - t_{n-1;\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}},\ \bar{x} + t_{n-1;\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}}\right)\right) = 1 - \alpha}$$
  - $\bar{x}$: media muestral
  - n : tamaño muestral
  - $\hat{s}$: cuasidesviación típica muestral
  - $t_{n-1;\frac{\alpha}{2}}$: es el valor de una variable aleatoria t, con distribución de Student con n-1 grados de libertad, que deja a su derecha un área igual a $\alpha/2$; es decir: $$\boxed{P\left(t_{n-1} > t_{n-1;\frac{\alpha}{2}}\right) = \frac{\alpha}{2}}$$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- los dos resultados que se han enunciado se basan en que es conocida la distribución de la muestra (normal, lo que permite deducir que la media muestral sigue una distribución normal de media μ y varianza σ^2/n)
- teorema central del límite: sea cual sea la distribución de las variables de la muestra aleatoria simple, la media muestral sigue, aproximadamente, una distribución normal de media μ y varianza σ^2/n
- por tanto, se puede obtener un intervalo de confianza aproximado para cualquier media de cualquier distribución

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

- sea $X_1, X_2, \dots, X_n$ una muestra de tamaño n de una v.a. X que sigue una distribución cualquiera con una media, μ, desconocida y con una desviación típica σ
- siendo $\bar{x}$ la media muestral, si n es suficientemente grande (n>30) entonces:

$$\boxed{P\left(\mu \in \left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,,\ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha}$$

- si σ es desconocida puede sustituirse por la cuasidesviación típica muestral, $\hat{s}$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ tabla resumen

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\overline{X}$	$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$	◇ población normal ◇ σ conocida
$\overline{X}$	$\left(\overline{x} - t_{n-1;\,\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t_{n-1;\,\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$	◇ población normal ◇ σ desconocida
$\overline{X}$	$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$	◇ n>30 ◇ σ conocida
$\overline{X}$	$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$	◇ n>100 ◇ σ desconocida

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ **Ejemplo.** El tiempo de fallo de una componente electrónica sigue una distribución exponencial de parámetro λ desconocido. Se toma una muestra de 50 tiempos de fallo y su media muestral es de 17,5 siendo la cuasidesviación típica muestral de 19,2. Calcular un intervalo de confianza para λ con un nivel de significancia α=0,1

Solución

- nivel de confianza del **90%**: $1-\alpha=1-0.1=0.9$
- $\alpha/2=0.05$
- media muestral: $\bar{x}=17.5$
- cuasidesviación típica muestral: $\hat{s}=19.2$
- tamaño muestral: $n=50$

11. ESTIMACIÓN DE LA MEDIA

Intervalo de confianza para la media poblacional

▶ **Ejemplo.**

Solución

$$z_{\frac{\alpha}{2}} = z_{0.05} \rightarrow z_{0.95} = 1.645$$

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}} \right) \Rightarrow \left(17.5 - 1.645 \cdot \frac{19.2}{\sqrt{50}}, 17.5 + 1.645 \cdot \frac{19.2}{\sqrt{50}} \right)$$

- entonces: $\mu = E[X] \in (13.033, 21.967)$
- por otra parte: $\mu = E[X] = \dfrac{1}{\lambda}$
- intervalo de confianza al 90% de λ:

$$\boxed{\lambda \in \left(\tfrac{1}{21.967}, \tfrac{1}{13.033} \right) = (0.0455, 0.0767)}$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: muestras independientes

- se consideran dos poblaciones:
  - población 1: media μ_1 y desviación típica σ_1
  - población 2: media μ_2 y desviación típica σ_2
- se seleccionan dos muestras aleatorias simples de forma independiente:
  - población 1: tamaño muestral n_1 → $\boxed{\bar{X}_1 \approx N\left(\mu_1; \frac{\sigma_1}{\sqrt{n_1}}\right)}$
  - población 2: tamaño muestral n_2 → $\boxed{\bar{X}_2 \approx N\left(\mu_2; \frac{\sigma_2}{\sqrt{n_2}}\right)}$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: muestras independientes

- se va a utilizar la información suministrada por las dos muestras (seleccionadas independientemente) para estimar la diferencia de las medias poblacionales

$$\boxed{\mu_1 - \mu_2}$$

- se sabe que:
  - $E\left[\overline{X}_1 - \overline{X}_2\right] = \mu_1 - \mu_2$
  - $Var\left[\overline{X}_1 - \overline{X}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: muestras independientes

► **tabla resumen.** Intervalo de confianza para $\boxed{\mu_1 - \mu_2}$

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\bar{x}_1 - \bar{x}_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	◇ poblaciones normales ó n_1, n_2 >15 ◇ σ_1, σ_2 conocidas
$\bar{x}_1 - \bar{x}_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$	◇ n_1, n_2 >30 ◇ σ_1, σ_2 desconocidas

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: muestras independientes

► **tabla resumen.** Intervalo de confianza para $\boxed{\mu_1 - \mu_2}$

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\bar{x}_1 - \bar{x}_2$	$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{n_1+n_2-2\ ;\ \frac{\alpha}{2}} \cdot S_P \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	◇ poblaciones normales ◇ σ_1, σ_2 desconocidas ◇ $\sigma_1 = \sigma_2$ ◇ n_1, n_2 pequeños

$$\boxed{S_P^2 = \frac{\left(n_1 - 1\right) \cdot \hat{s}_1^2 + \left(n_2 - 1\right) \cdot \hat{s}_2^2}{n_1 + n_2 - 2}}$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: muestras independientes

► **tabla resumen.** Intervalo de confianza para $\boxed{\mu_1 - \mu_2}$

- en el último caso, como se supone que las varianzas de ambas poblaciones son iguales aunque desconocidas, se utiliza la información proporcionada por las dos muestras para construir un estimador de σ^2
- ese estimador se denota como: S_P^2
- S_P^2 : media ponderada de las dos cuasivarianzas muestrales con pesos proporcionales a sus respectivos tamaños muestrales (n_1 y n_2)

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

- **muestras dependientes e independientes**
  - si la selección de los datos de una población no está relacionada con la de los datos de la otra se trata de muestras independientes
  - si las muestras se seleccionan de manera que cada medida en una de ellas pueda asociarse naturalmente con una medida en la otra muestra se llaman muestras dependientes
- diferencia de medias en **muestras independientes**
  - se estudia la diferencia de medias de dos muestras indepentes extraídas aleatoriamente de poblaciones normales

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

► diferencia de medias con **datos pareados**

- se estudia la diferencia de medias antes y después de cierto tratamiento al que son sometidos los mismos sujetos que constituyen una muestra extraída aleatoriamente de una población normal
- si dos medidas se obtienen de la misma fuente se puede pensar que las medidas están pareadas; por tanto, dos medidas que se obtienen de la misma fuente son dependientes
- si dos muestras son dependientes entonces, necesariamente, tienen el mismo tamaño

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

► **ejemplo 1**

- se quiere estudiar el efecto de una dieta al cabo de 6 meses; para ello, se mide el peso de los sujetos antes de iniciar la dieta y después de realizarla durante los 6 meses

- en este caso ambos pesos no son variables independientes ya que es claro que el peso al cabo de los 6 meses de dieta dependerá del peso que se tenía anteriormente

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pa-reados (ó pares coincidentes)

► ejemplo 2

- se quiere comparar la velocidad de ejecución computacional de dos algoritmos
- se podría hacer ejecutando los algoritmos sobre dos conjuntos aleatorios de problemas independientes pero, en este caso, el experimento podría verse afectado por la naturaleza de los problemas escogidos para cada algoritmo
- este problema puede evitarse si a cada tipo de problema se le aplica un algoritmo y después el otro midiendo los tiempos de ejecución correspondientes

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

▶ procedimiento

- se estudia el caso en el que *X* e *Y* son dos variables aleatorias que representan características diferentes de la misma población y tales que:

$$X \approx N\left(\mu_X ; \sigma_X\right) \quad Y \approx N\left(\mu_Y ; \sigma_Y\right)$$

- para analizar sus diferencias se toman muestras pareadas:
  - ◇ se obtienen los valores de *X* e *Y* sobre los **mismos** sujetos de la población
  - ◇ el tamaño de la muestra, ***n***, es igual para *X* e *Y*

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

► procedimiento

- se define la variable aleatoria **diferencia**, ***D = X−Y***, que sigue un modelo de distribución normal:

$$\boxed{D \approx N\left(\mu_D ; \sigma_D\right) \quad donde \quad \mu_D = \mu_X - \mu_Y}$$

- el problema se convierte en una estimación por intervalo para la variable ***D***:
  - ◇ $D_i = X_i - Y_i \quad \left(i = 1, 2, \cdots, n\right)$: es una muestra aleatoria simple de la variable diferencia *D*
  - ◇ $\overline{D} = \overline{X} - \overline{Y}$: es la media muestral de la variable diferencia *D*
  - ◇ $\hat{s}_D^2 = S_D^2$: es la cuasivarianza muestral de la variable *D*

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

▶ procedimiento

- estimación por intervalo para la variable ***D***:

  ◇ estimador: $\overline{D} = \overline{X} - \overline{Y}$

  ◇ tamaño muestral grande (n≥30), σ conocida

$$\boxed{\frac{\overline{D} - \mu_D}{\sigma_D / \sqrt{n}} = Z \approx N(0;1)}$$

  ◇ tamaño muestral pequeño (n<30), σ desconocida

$$\boxed{\frac{\overline{D} - \mu_D}{\hat{s}_D / \sqrt{n}} = T \approx t_{n-1}}$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

- **Ejercicio (Junio 2012).**
  Para contrastar la efectividad de una herramienta nueva en la realización de una tarea se ha pedido a 9 operarios que realicen la tarea con la herramienta nueva y la antigua. Se han medido los tiempos de realización en minutos y el resultado es le siguiente

	Operario								
Herramienta	1	2	3	4	5	6	7	8	9
Antigua	39.5	35.4	35	41.3	44	32.9	37.7	41	38
Nueva	43.2	30.5	31.5	38	44.5	33	36.2	38.4	35

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

- **Ejercicio (Junio 2012).**

  (1º) Encuentre un intervalo de confianza de 95 % para la diferencia media de tiempos de realización de la tarea con ambas herramientas,suponiendo que las diferencias de tiempos se distribuyen aproximadamente de forma normal.
  (2º) ¿Cuál tendría que ser el número de operarios si se desea que la desviación típica del estimador sea como máximo de 0.5 minutos?

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pa-reados (ó pares coincidentes)

- **Ejercicio (Junio 2012).**

  (3º) ¿Es el resultado de la prueba evidencia suficien-te de que la herramienta nueva reduce al menos en un minuto el tiempo medio de realización de la tarea con un nivel de significación α=5%?

  NOTA. Este apartado se podrá resolver cuando se estudie el tema relativo a los *Contrastes de hipótesis*

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

▶ **Ejercicio (Junio 2012).**

- estadísticos muestrales:

Herramienta	Operario									Estadísticos	
	1	2	3	4	5	6	7	8	9	MEDIA	ŝ
Antigua: A	39,5	35,4	35	41,3	44	32,9	37,7	41	38	38,3111	3,5201
Nueva: N	43,2	30,5	31,5	38	44,5	33	36,2	38,4	35	36,7000	4,8731
D=A–N	-3,7	4,9	3,5	3,3	-0,5	-0,1	1,5	2,6	3	1,6111	2,6398

- ***Excel***:
  - ◇ $\bar{x}_D = \bar{D}$ =PROMEDIO(B5:J5)=1.6111
  - ◇ $\hat{s}_D$ =DESVEST(B5:J5)=2.6398

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

► **Ejercicio (Junio 2012). Apartado 1º.**

- se trata de un problema de muestras pequeñas (n<30) y pares coincidentes ya que la herramienta se mide para cada operario
- estimación por intervalo para la variable ***D***:
  - ◇ estimador insesgado de varianza mínima:

$$\overline{D} = \overline{x}_A - \overline{x}_B = 1.6111$$

  - ◇ error probable:

$$\sigma_D = \frac{\hat{s}_D}{\sqrt{n}} = \frac{2.6398}{\sqrt{9}} = 0.8799$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

▶ **Ejercicio (Junio 2012). Apartado 1º.**

- como se ha de estimar una media poblacional se utiliza la distribución de probabilidad ***t*** de Student
  - ◇ nivel de confianza: α = 0.95 ≈ 95%
  - ◇ grados de libertad: $gl = \nu = n - 1 = 8$
  - ◇ valor de la variable (dos colas): $t_{\nu;\frac{\alpha}{2}} = t_{8;0,975} = 2.3060$
  - ◇ ***Excel***: =DISTR.T.INV(0,05;8)=2.3060
- Intervalo de confianza: $IC = \overline{D} \pm t_{8;0,975} \cdot \sigma_D = 1.611 \pm 2.3060 \cdot 0.8799$

$$\boxed{IC = \left(-0.4180\,,\,3.6403\right)}$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pareados (ó pares coincidentes)

- **Ejercicio (Junio 2012). Apartado 2º.**
  - la desviación típica del estimador viene dada por:

$$\boxed{t_{\nu;\frac{\alpha}{2}} \cdot \sigma_D = t_{\nu;\frac{\alpha}{2}} \cdot \frac{\hat{s}_D}{\sqrt{n}} \leq e_{\max} = 0.5}$$

  - por lo tanto:

$$n \geq \left(t_{\nu;\frac{\alpha}{2}} \cdot \frac{\hat{s}_D}{e_{\max}} \right)^2 = 148.2267 \quad \Rightarrow \quad \boxed{n_{\min} = 149 \;\; operarios}$$

12. DIFERENCIA DE MEDIAS

Estimación de la diferencia de medias: datos pa-reados (ó pares coincidentes)

- **Ejercicio (Junio 2012). Apartado 3º.**
  - se propone como ejercicio para el alumno cuando se estudie el tema relativo a *Contrates de hipótesis*

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para la varianza poblacional

- análogamente a lo ya estudiado, pueden obtenerse intervalos de confianza para la varianza, con la media conocida ó desconocida, pero sólo cuando la variable aleatoria observada sigue una distribución gaussiana

- en el capítulo anterior se vio que:

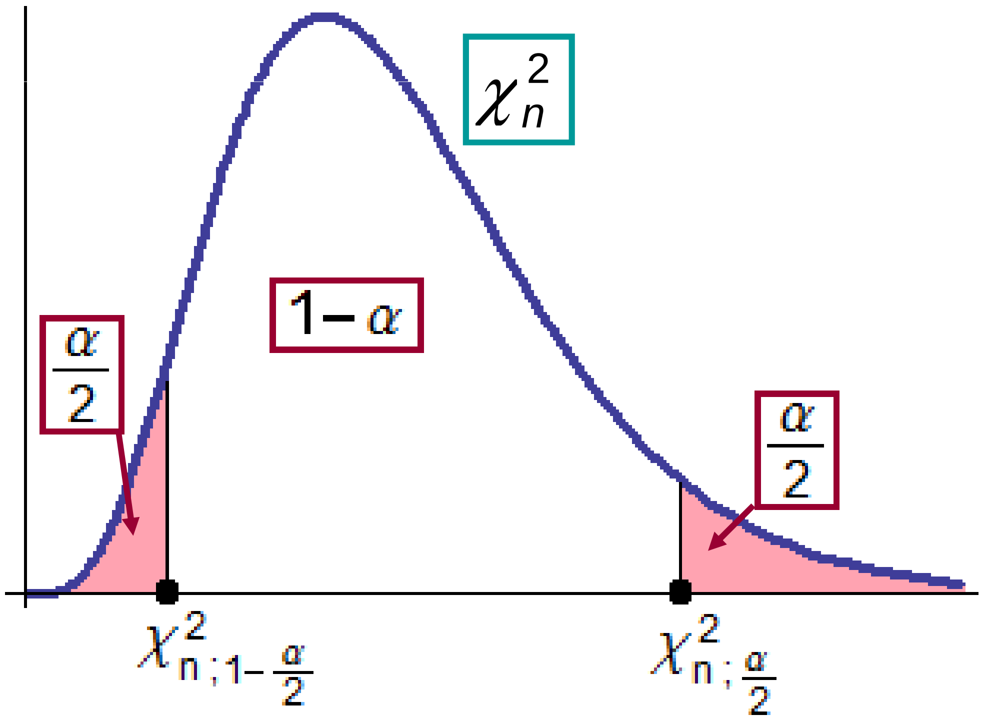
$$\boxed{\frac{(n-1)\cdot\hat{s}^2}{\sigma^2} \approx \chi^2_{n-1}}$$

  - n: tamaño muestral
  - σ^2: varianza de la población normal que se desea estimar
  - $\hat{s}$: cuasivarianza muestral

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para la varianza poblacional

► la distribución χ^2 no es simétrica respecto al 0



$$P\left(\chi^2_n > \chi^2_{n\,;\,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

$$P\left(\chi^2_n > \chi^2_{n\,;\,\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para la varianza poblacional

- se tiene que: $P\left(\chi^2_{n-1\,;\,1-\frac{\alpha}{2}} \leq \chi^2_{n-1} \leq \chi^2_{n-1\,;\,\frac{\alpha}{2}}\right) = 1-\alpha$

- sustitución: $\dfrac{(n-1)\cdot \hat{s}^2}{\sigma^2} \approx \chi^2_{n-1}$

- operando:

$$\boxed{P\left(\frac{(n-1)\cdot \hat{s}^2}{\chi^2_{n-1\,;\,\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\cdot \hat{s}^2}{\chi^2_{n-1\,;\,1-\frac{\alpha}{2}}}\right) = 1-\alpha}$$

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para la varianza poblacional

► **tabla resumen.** Intervalo de confianza para σ^2

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\hat{s}^2$	$\left(\dfrac{(n-1)\cdot\hat{s}^2}{\chi^2_{n-1\,;\,\frac{\alpha}{2}}} , \dfrac{(n-1)\cdot\hat{s}^2}{\chi^2_{n-1\,;\,1-\frac{\alpha}{2}}} \right)$	◇ población normal ◇ μ desconocida
$\dfrac{\sum_{i=1}^{i=n}(x_i-\mu)^2}{n}$	$\left(\dfrac{\sum_{i=1}^{i=n}(x_i-\mu)^2}{\chi^2_{n\,;\,\frac{\alpha}{2}}} , \dfrac{\sum_{i=1}^{i=n}(x_i-\mu)^2}{\chi^2_{n\,;\,1-\frac{\alpha}{2}}} \right)$	◇ población normal ◇ μ conocida

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para el cociente de dos varianzas poblacionales

► se consideran dos poblaciones:

- población 1: varianza σ_1^2
- población 2: varianza σ_2^2

► para comparar las varianzas de las dos poblaciones realizan inferencias sobre su cociente: $\boxed{\dfrac{\sigma_1^2}{\sigma_2^2}}$

► la distribución muestral del anterior estimador es muy conocida cuando las muestras son aleatorias simples y extraídas de dos poblaciones normales independientes

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para el cociente de dos varianzas poblacionales

- bajo las hipótesis indicadas, se utiliza la distribución *F* de Fisher-Snedecor para obtener un intervalo de confianza para el estimador: $\boxed{\dfrac{\sigma_1^2}{\sigma_2^2}}$

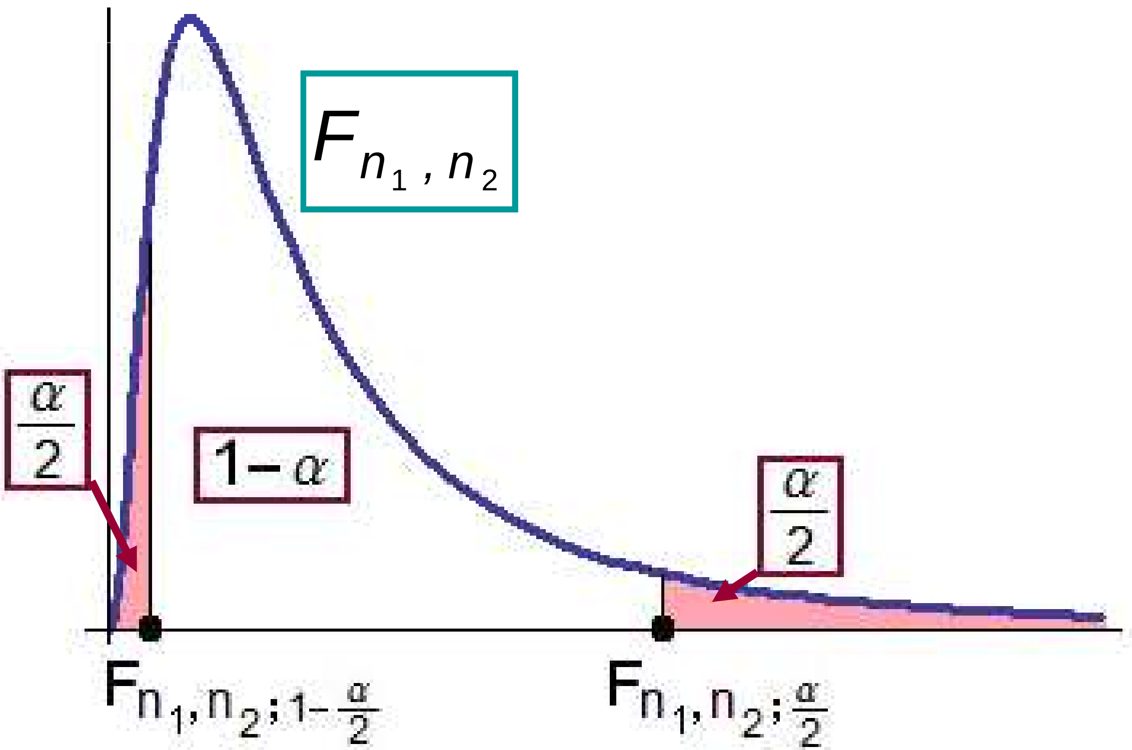
$$\boxed{\frac{\dfrac{\hat{s}_1^2}{\sigma_1^2}}{\dfrac{\hat{s}_2^2}{\sigma_2^2}} \approx F_{n_1-1,\, n_2-1;\, 1-\alpha}}$$

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para el cociente de dos varianzas poblacionales

► para establecer los límites del intervalo de confianza del estimador debe recordarse que:

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



13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para el cociente de dos varianzas poblacionales

- se tiene que: $P\left(F_{n_1-1,\, n_2-1\,;\, 1-\frac{\alpha}{2}} \le F_{n_1-1,\, n_2-1} \le F_{n_1-1,\, n_2-1\,;\, \frac{\alpha}{2}}\right) = 1-\alpha$

- sustitución: $\dfrac{\dfrac{\hat{s}_1^2}{\sigma_1^2}}{\dfrac{\hat{s}_2^2}{\sigma_2^2}} \approx F_{n_1-1,\, n_2-1;\, 1-\alpha}$

- operando:

$$\boxed{P\left(\frac{\hat{s}_1^2 / \hat{s}_2^2}{F_{n_1-1,\, n_2-1\,;\, \frac{\alpha}{2}}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\hat{s}_1^2 / \hat{s}_2^2}{F_{n_1-1,\, n_2-1\,;\, 1-\frac{\alpha}{2}}}\right) = 1-\alpha}$$

13. ESTIMACIÓN DE LA VARIANZA

Intervalo de confianza para el cociente de dos varianzas poblacionales

► **tabla resumen.** Intervalo de confianza para $\boxed{\dfrac{\sigma_1^2}{\sigma_2^2}}$

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\dfrac{\hat{s}_1^2}{\hat{s}_2^2}$	$\left(\dfrac{\hat{s}_1^2 / \hat{s}_2^2}{F_{n_1-1,\, n_2-1;\, \frac{\alpha}{2}}} , \dfrac{\hat{s}_1^2 / \hat{s}_2^2}{F_{n_1-1,\, n_2-1;\, 1-\frac{\alpha}{2}}} \right)$	◇ poblaciones normales ◇ μ_1, μ_2 desconocidas

14. ESTIMACIÓN DE UNA PROPORCIÓN

Intervalo de confianza para una proporción

- estimación de la proporción p de éxitos, es decir, la proporción p de individuos de una población que presentan una determinada característica
- sea p la probabilidad desconocida de un determinado suceso, denominado *éxito*, que puede ocurrir en un experimento binomial
- sea cierta muestra $\{x_1, x_2, \ldots, x_n\}$ de realizaciones independientes del experimento, donde
  - si se da el éxito: $x_i = 1$
  - proporción de éxitos de la muestra: $\hat{p} = \frac{1}{n} \sum_{i=1}^{i=n} x_i$

14. ESTIMACIÓN DE UNA PROPORCIÓN

Intervalo de confianza para una proporción

► **tabla resumen.** Intervalo de confianza para *p*

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\hat{p} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$	$\hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$	◇ $n \cdot \hat{p} \geq 4$ ◇ $n \cdot \left(1-\hat{p}\right) \geq 4$

14. ESTIMACIÓN DE UNA PROPORCIÓN

Intervalo de confianza para la diferencia de proporciones de dos poblaciones

- sean dos experimentos binomiales independientes
- sea considera una muestra de cada uno :
  - muestra 1: tamaño n_1
  - muestra 2: tamaño n_2
  - proporción de éxitos de la muestra 1: $\hat{p}_1 = \dfrac{1}{n_1} \sum_{i=1}^{i=n_1} x_i$
  - proporción de éxitos de la muestra 2: $\hat{p}_2 = \dfrac{1}{n_2} \sum_{i=1}^{i=n_2} x'_i$

14. ESTIMACIÓN DE UNA PROPORCIÓN

Intervalo de confianza para la diferencia de proporciones de dos poblaciones

► **tabla resumen.** Intervalo de confianza para p

Estimador	Intervalo de confianza: (1-α)100%	Hipótesis adicionales
$\hat{p}_1 - \hat{p}_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	◇ $n_1 \cdot \hat{p}_1 \geq 4$, $n_1 \cdot (1-\hat{p}_1) \geq 4$ ◇ $n_2 \cdot \hat{p}_2 \geq 4$, $n_2 \cdot (1-\hat{p}_2) \geq 4$