# Tema 4

# Muestreo y Estimación

## 4.1. Conceptos generales

Abordaremos en este tema y en los siguientes un problema que surge, de manera natural, cuando se trata de estudiar una característica en un conjunto de individuos muy numeroso.

**Definición 4.1.1.–** *Se llama* **población** *o* **universo** *al conjunto de todos los individuos o elementos objeto de estudio.*

Cuando la población es muy grande generalmente no es posible estudiar a todos los individuos, ya sea por cuestiones económicas (podría ser muy costoso o necesitar demasiado tiempo), porque el estudio de cada individuo conlleva su destrucción (pruebas de resistencia de un material) o, simplemente, porque no tendría sentido (en el caso de los sondeos electorales, el sondeo no puede implicar a toda la población porque, en ese caso, serían unas elecciones en sí mismos). En estos casos el objetivo es obtener conclusiones, sobre todos los individuos que forman la población, pero estudiando únicamente una parte de la misma.

**Definición 4.1.2.–** *Una* **muestra** *es un subconjunto, más o menos representativo, de la población. Se llama* **tamaño** *de la muestra al número de elementos que la forman.*

**Definición 4.1.3.–** *Se denomina* **muestreo** *al proceso por el cual extraemos una muestra de una población.*

Hay distintos tipos de muestreo que se adaptan a distintas situaciones experimentales. En éste y en los sucesivos temas trabajaremos bajo la hipótesis de que las muestras se recogen mediante un procedimiento de 'muestreo aleatorio simple'.

**Definición 4.1.4.–** *Una* **muestra aleatoria simple de tamaño** $n$ *es una muestra que se consigue eligiendo* $n$ *elementos de la población al azar y sin reposición, esto es, se elige al azar uno de los elementos de la población, de los que quedan se elige otro, y así hasta obtener los* $n$ *elementos que forman la muestra.*

De este modo, si $X$ es una variable aleatoria que representa la medida de la característica bajo estudio, al repetir el experimento $n$ veces en las mismas condiciones, las $X_1, X_2, \ldots, X_n$ medidas de $X$ serán variables aleatorias **independientes**, que siguen la misma distribución que $X$. Constituyen, de este modo, una muestra aleatoria simple de tamaño $n$. Notesé que, si $X_1, X_2, \ldots, X_n$ forman una muestra aleatoria simple de una variable $X$, puesto que cada $X_i$ sigue la misma distribución que $X$, entonces $E[X_i] = E[X]$, $E[X_i^2] = E[X^2], Var[X_i] = Var[X]$ y, en general, cualquier característica de una variable aleatoria $X_i$ coincidirá con la misma característica medida sobre $X$.

De manera general, la característica que se desea estudiar en la población se modela mediante una variable aleatoria que viene caracterizada por uno o más parámetros. Por ejemplo, la longitud de una pieza manufacturada se podría modelar mediante una variable $X \sim N(\mu, \sigma^2)$. De este modo, al ser $E[X] = \mu$, estudiar la longitud media de las piezas nos lleva necesariamente a tratar de determinar el valor del parámetro $\mu$.

Dicha estimación se puede realizar de dos formas: puntualmente o mediante intervalos. En el primer caso se trata de encontrar un valor aproximado del valor del parámetro. En el segundo se trata de encontrar un intervalo que, con una cierta confianza, contenga el valor real del parámetro.

## 4.2. Estimación puntual

**Definición 4.2.1.–** *Dada una variable aleatoria $X$, que mide una característica en una población, un parámetro es una caracterización numérica de la distribución de la población que determina total o parcialmente la función de densidad (o de probabilidad) de la variable aleatoria $X$.*

**Ejemplo 4.2.2.–**

- Si $X \sim \mathcal{P}(\lambda)$, la distribución queda determinada cuando se conoce el valor de $\lambda$.
- Si $X \sim N(\mu, \sigma^2)$, la distribución queda determinada cuando se conocen los valores de $\mu$ y de $\sigma^2$.

**Definición 4.2.3.–** *Al conjunto de valores que puede tomar un parámetro $\theta$ se le llama* **espacio paramétrico** *y lo denotaremos por $\Theta$.*

**Ejemplo 4.2.4.–**

- Si $X \sim \mathcal{P}(\lambda)$ entonces $\Theta = \mathbb{R}^+$.
- Si $X \sim N(\mu, \sigma^2)$ entonces $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$.

### 4.2.1. Estadísticos y estimadores.

**Definición 4.2.5.–** *Sea $X$ una variable aleatoria cuya función de densidad o de probabilidad es $f(x, \theta)$ y $X_1, \ldots, X_n$ una m.a.s. Un estadístico es una función $T(X_1, \ldots, X_n)$ que no depende del parámetro $\theta$.*

**Ejemplo 4.2.6.–** Si $X \sim N(\mu, \sigma^2)$ con $\mu$ y $\sigma$ desconocidas, las siguientes funciones son estadísticos:

$$T_1(X_1, \ldots, X_n) = \tfrac{X_1+\cdots+X_n}{n}; \qquad T_2(X_1, \ldots, X_n) = \tfrac{X_1^2+\cdots+X_n^2}{n}$$

$$T_3(X_1, \ldots, X_n) = max\{X_1, \ldots, X_n\}$$

Sin embargo, $T_4(X_1, \ldots, X_n) = \dfrac{X_1 + \cdots + X_n}{n\sigma}$ no es un estadístico, por ser $\sigma$ desconocido.

Nótese que un estadístico es función de la muestra, y por tanto es una variable aleatoria.

**Definición 4.2.7.–** *Un estadístico $T(X_1, \ldots, X_n)$ se dice que es un estimador de $\theta$ si toma valores en el espacio paramétrico de $\theta$. Para una realización $x_1, \ldots, x_n$ de una muestra, el valor de $T(x_1, \ldots, x_n)$ se llama estimación de $\theta$.*

Por lo tanto, un estimador de un parámetro no es más que una función de la muestra que no depende de parámetros desconocidos y toma valores admisibles para el parámetro.

**Ejemplo 4.2.8.–** Sea $X$ una variable aleatoria con $E[X] = \mu$ y $Var(X) = \sigma^2$ y sea $X_1, \ldots, X_n$ es una m.a.s. de $X$. Entonces:

- $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ es un estimador de $\mu$ y se denomina **media muestral**.
- $S^2 = \dfrac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ es un estimador de $\sigma^2$ y se denomina **varianza muestral**.
- $S_c^2 = \dfrac{n}{n-1}S^2 = \displaystyle\sum_{i=1}^{n} \dfrac{(X_i - \overline{X})^2}{n-1}$ también es un estimador de $\sigma^2$ y se denomina **cuasivarianza muestral**.
- La raíz cuadrada positiva de la varianza muestral, $\sqrt{S^2}$, se denomina **desviación típica muestral** y se denota por $S$. La raíz cuadrada positiva de la cuasivarianza muestral, $\sqrt{S_c^2}$, se denomina **cuasidesviación típica muestral** y se denota por $S_c$.

Se verifica que $(n-1)S_c^2 = nS^2$ y $S^2 = \dfrac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2$. A la hora de calcular la varianza y la cuasivarianza de un conjunto de datos, resulta más *cómodo* calcularlo a partir de estas expresiones que utilizando la definición.

### 4.2.2. Propiedades deseables en los estimadores

Puesto que un estimador es una variable aleatoria, no siempre nos devolverá el valor exacto del parámetro que deseamos estimar. De hecho, cuando el estimador sea una variable aleatoria continua, la probabilidad de que el estimador proporcione el valor real del parámetro será 0. Más aún, aunque en algún caso el estimador nos proporcione el valor exacto del parámetro no podríamos conocer esta circunstancia dado que el parámetro es desconocido. Por este motivo, existen una serie de propiedades que debe cumplir cualquier estimador para garantizar que la estimación que proporciona es buena *en algún sentido*. Entre las propiedades deseables de un buen estimador se encuentran las siguientes: **carencia de sesgo, consistencia y eficiencia**.

• **Carencia de Sesgo**

Un buen estimador debería, en promedio, determinar el verdadero valor del parámetro. Esto nos lleva a la siguiente definición.

**Definición 4.2.9.–** *Se dice que $T$ es un* **estimador insesgado** *de $\theta$ si $E[T] = \theta$. En otro caso se dice que $T$ es un estimador sesgado y se llama sesgo de $T$ a $|E[T] - \theta|$.*

**Ejemplo 4.2.10.–** Sea $X$ una variable aleatoria de la que sabemos que $E[X] = \mu$ y $Var(X) = \sigma^2$, y $X_1, \ldots, X_n$ una muestra aleatoria simple de $X$. Entonces:

a) $\overline{X}$ es un estimador insesgado de $\mu$.

b) $S^2$ es un estimador sesgado de $\sigma^2$.

c) $S_c^2$ es un estimador insesgado de $\sigma^2$.

Comprobaremos a continuación que las afirmaciones realizadas en el ejemplo anterior son correctas. Para ello vamos a calcular la esperanza de cada uno de ellos, utilizando las propiedades de la esperanza de una variable aleatoria. Recuérdese también que, al ser $X_1, \ldots, X_n$ una muestra aleatoria de $X$, cada $X_i$ sigue la misma distribución que $X$ y, por lo tanto, $E[X_i] = E[X]$ y $E[X_i^2] = E[X^2]$, para cualquier $1 = 1, \ldots, n$.

a) $$E[\,\overline{X}\,] = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n}E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n}n\mu = \mu.$$

b) $$E[S^2] = E\left[\frac{\sum_{i=1}^n X_i^2}{n} - \overline{X}^2\right] = E\left[\frac{\sum_{i=1}^n X_i^2}{n}\right] - E\left[\overline{X}^2\right] = \frac{1}{n}\sum_{i=1}^n \left(E[X_i^2]\right) - E\left[\overline{X}^2\right] =$$
$$= \frac{1}{n}\sum_{i=1}^n \left(E[X^2]\right) - E\left[\overline{X}^2\right] = \frac{1}{n}\cdot n \cdot E[X^2] - E\left[\overline{X}^2\right] = E[X^2] - E\left[\overline{X}^2\right]$$

Calcularemos ahora $E[X^2]$ y $E[\overline{X}^2]$:

- Al ser $Var[X] = E[X^2] - E^2[X]$, despejando $E[X^2] = Var[X] + E^2[X] = \sigma^2 + \mu^2$.

- Del mismo modo, $E[\overline{X}^2] = Var[\overline{X}] + E^2[\overline{X}]$.

  Entonces, usando las propiedades de la varianza,

  $$Var[\overline{X}] = Var\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}Var\left[\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\sum_{i=1}^n Var\left[X_i\right] =$$
  $$\frac{1}{n^2}\sum_{i=1}^n Var\left[X\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

  Además, por el apartado a), $E[\overline{X}] = \mu$ y, por tanto, $E^2[\overline{X}] = \mu^2$.

  De ambas cosas, $E[\overline{X}^2] = \dfrac{\sigma^2}{n} + \mu^2$.

  Finalmente, obtenemos que $E[S^2] = \sigma^2 + \mu^2 - \left(\dfrac{\sigma^2}{n} + \mu^2\right) = \dfrac{n-1}{n}\sigma^2$.

c) $E[S_c^2] = E\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}E[S^2] = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2.$

• **Consistencia**

Puesto que la muestra contiene información sobre el parámetro sería deseable que a medida que aumente el tamaño de la muestra mejoren las estimaciones obtenidas. Esto se traduce en la siguiente definición.

**Definición 4.2.11.–** *Diremos que $T$ es un* **estimador consistente** *de cara a estimar el parámetro $\theta$ si*

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} P\left(|T-\theta| < \varepsilon\right) = 1.$$

El estudio de la consistencia mediante la definición puede ser complicado. El siguiente teorema nos proporciona una condición suficiente para el estudio de la consistencia. Por lo tanto, si se cumplen las condiciones teorema podremos afirmar la consistencia del estimador en cuestión. Si no se cumplen, no podremos afirmar nada.

**Teorema 4.2.12.–** *Sea $X_1, \ldots, X_n$ una muestra aleatoria simple y $T(X_1, \ldots, X_n)$ un estimador del parámetro $\theta$. Si se verifica que*

$$\lim_{n\to\infty} E[T] = \theta \quad y \qquad \lim_{n\to\infty} Var(T) = 0$$

*entonces $T$ es un estimador consistente para $\theta$.*

**Ejemplo 4.2.13.–**

- Si $X$ es una variable aleatoria tal que $E(X) = \mu$, entonces $\overline{X}$ es un estimador consistente de $\mu$.
- Si $X \sim N(\mu, \sigma^2)$ entonces $S^2$ y $S_c^2$ son estimadores consistentes de $\sigma^2$.

• **Eficiencia**

Puesto que un estimador es una variable aleatoria sería deseable que los valores que proporciona presenten la menor dispersión posible. De este modo, dados dos estimadores insesgados del parámetro a estimar, será preferible utilizar el que presente una menor dispersión de sus valores, esto es, el de menor varianza. De este modo se garantiza que, aunque generalmente el estimador no 'acierta' de manera exacta con el valor del parámetro que se desea estimar, al menos se *queda cerca*.

**Definición 4.2.14.–** *Sean $T_1$ y $T_2$ dos estimadores insesgados de $\theta$. Diremos que $T_1$ es más eficiente que $T_2$ si $Var(T_1) < Var(T_2)$.*

Dados dos estimadores insesgados de $\theta$, será preferible utilizar el de menor varianza.

### 4.2.3. Algunos estimadores usuales

La siguiente tabla recoge estimadores para los parámetros de algunas de las distribuciones más usuales.

| Distribución | Estimador |
|---|---|
| $Be(p)$ | $\hat{p} = \overline{X}$ |
| $Ge(p)$ | $\hat{p} = 1/\overline{X}$ |
| $\mathcal{P}(\lambda)$ | $\hat{\lambda} = \overline{X}$ |
| $N(\mu, \sigma^2)$ | $\hat{\mu} = \overline{X}$<br>$\hat{\sigma^2} = S^2$ (No insesgado)<br>$\hat{\sigma^2} = S_c^2$ (Insesgado) |
| $Exp(\lambda)$ | $\hat{\lambda} = 1/\overline{X}$ |

## 4.3. Distribuciones asociadas al muestreo en poblaciones Normales

**Definición 4.3.1.–** *Una variable aleatoria $X$ se dice que sigue una distribución normal de parámetros $\mu$ y $\sigma^2$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, y se denota por $X \sim N(\mu, \sigma^2)$, si su f.d.d. es:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad\qquad \forall\ x \in \mathbb{R}$$

Se denota por $z_{1-\alpha}$ al cuantil $1-\alpha$ de la distribución, esto es, a aquel valor tal que $P[Z \leq z_{1-\alpha}] = 1-\alpha$.

**Definición 4.3.2.–** *Si $X_1, \ldots, X_n$ son variables aleatorias independientes idénticamente distribuidas según $N(0,1)$, entonces $X_1^2 + \ldots + X_n^2$ sigue una* **distribución Chi-cuadrado** *con $n$ grados de libertad. Lo denotaremos $X \sim \chi_n^2$.*

**•Propiedades:**

1. Si $X \sim \chi_n^2$ e $Y \sim \chi_m^2$, $X$ e $Y$ independientes, entonces $X + Y \sim \chi_{n+m}^2$
2. $E[\chi_n^2] = n$ y $Var[\chi_n^2] = 2n$

Se denota $\chi_{n,1-\alpha}^2$ al cuantil $1-\alpha$ de la distribución, esto es, a aquel valor tal que $P[X \leq \chi_{1-\alpha,n}^2] = 1-\alpha$.

**Definición 4.3.3.–** *Sean* $X \sim N(0,1)$ *e* $Y \sim \chi^2_n$, $X$ *e* $Y$ *independientes, entonces* $T = \frac{X}{\sqrt{Y/n}}$ *sigue una* **distribución T-Student** *con* $n$ *grados de libertad. Lo denotaremos* $T \sim t_n$

**•Propiedades:**

1. $E[T] = 0$ y $Var[T] = \frac{n}{n-2}$
2. Si $T \sim t_n$, entonces $\lim\limits_{n\to+\infty} T \sim N(0,1)$. La aproximación es buena para $n \geq 30$.
3. Además, $t_{n,1-\alpha} = -t_{n,\alpha}$.

Se denota por $t_{n,1-\alpha}$ al cuantil $1-\alpha$ de la distribución, esto es, a aquel valor tal que $P[T \leq t_{n,1-\alpha}] = 1-\alpha$.

**Definición 4.3.4.–** *Sean* $X \sim \chi^2_n$ *e* $Y \sim \chi^2_m$, $X$ *e* $Y$ *independientes, entonces* $F = \frac{X/n}{Y/m}$ *sigue una* **distribución F-Snedecor** *con* $n$ *y* $m$ *grados de libertad. Lo denotaremos como* $F \sim \mathcal{F}_{n,m}$

**•Propiedades**

1. $E[F] = \frac{m}{m-2}$ y $Var[F] = \frac{m^2(2n+2m-4)}{n(m-2)^2(m-4)}$
2. Si $F \sim \mathcal{F}_{n,m}$, entonces $\frac{1}{F} \sim \mathcal{F}_{m,n}$

Se denota por $f_{n,m,1-\alpha}$ al valor tal que $P[F \leq f_{n,m,1-\alpha}] = 1-\alpha$.

Se verifica que $f_{n,m,\alpha} = 1/f_{m,n,1-\alpha}$.

## 4.4. Estimación por regiones de confianza

Aunque bajo ciertas condiciones los estimadores puntuales gozan de buenas propiedades y proporcionan buenas estimaciones no nos permitirán obtener, en general, el valor exacto del parámetro a estimar.

Nuestro objetivo ahora es encontrar un intervalo aleatorio del que podamos afirmar, con una probabilidad prefijada, que contiene el verdadero valor del parámetro.

**Definición 4.4.1.–** *Sea* $X$ *una variable aleatoria cuya función de distribución depende de un parámetro* $\theta$, *y* $l_1(X_1,\ldots,X_n)$ *y* $l_2(X_1,\ldots,X_n)$ *dos estadísticos tales que* $l_1 \leq l_2$. *Un intervalo de la forma* $(l_1(X_1,\ldots,X_n), l_2(X_1,\ldots,X_n))$ *se llama intervalo aleatorio de límite inferior* $l_1$ *y límite superior* $l_2$.

**Definición 4.4.2.–** *Un intervalo aleatorio* $(l_1(X_1,\ldots,X_n), l_2(X_1,\ldots,X_n))$ *se llama intervalo de confianza para* $\theta$ *a nivel de confianza* $(1-\alpha)\cdot 100\,\%$ *con* $0 < \alpha < 1$ *si*

$$P\left[l_1(X_1,\ldots,X_n) < \theta < l_2(X_1,\ldots,X_n)\right] = 1-\alpha.$$

La siguiente tabla recoge la expresión de intervalos a nivel de confianza $(1-\alpha)\cdot 100\,\%$ para la media y la varianza de poblaciones normales y para la diferencia de medias y el cociente de varianzas de dos poblaciones normales independientes. En las expresiones de los intervalos de confianza para los parámetros de una población normal, el tamaño muestral se representa por $n$. En el caso de los intervalos asociados a los parámetros de dos poblaciones normales, los tamaños muestrales de ambas poblaciones se representan por $n_x$ y $n_y$.

# Intervalos de confianza, en poblaciones normales, al $(1-\alpha)\cdot 100\,\%$

| Parámetro | Casos | Intervalo |
|---|---|---|
| $\mu$ | $\sigma$ conocida | $\overline{X} \pm \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$ |
| | $\sigma$ desconocida | $\overline{X} \pm \frac{S_c}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}$ |
| $\sigma^2$ | | $\left(\frac{(n-1)S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$ ó $\left(\frac{nS^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{nS^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$ |
| $\mu_x - \mu_y$ | $\sigma_x$, $\sigma_y$ conocidas | $(\overline{X} - \overline{Y}) \pm \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \cdot z_{1-\frac{\alpha}{2}}$ |
| $\mu_x - \mu_y$ | $\sigma_x = \sigma_y$ desconocidas * | $(\overline{X} - \overline{Y}) \pm \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \cdot S_p \cdot t_{n_x+n_y-2,1-\frac{\alpha}{2}}$ |
| $\mu_x - \mu_y$ | $\sigma_x \neq \sigma_y$ desconocidas ** | $(\overline{X} - \overline{Y}) \pm \sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}} \cdot t_{g,1-\alpha/2}$ |
| $\sigma_y^2/\sigma_x^2$ | | $\left(\frac{S_{cy}^2}{S_{cx}^2}\frac{1}{f_{n_y-1,n_x-1,1-\frac{\alpha}{2}}}, \frac{S_{cy}^2}{S_{cx}^2} f_{n_x-1,n_y-1,1-\frac{\alpha}{2}}\right)$ |

$$^{*}S_p^2 = \frac{(n_x-1)S_{cx}^2+(n_y-1)S_{cy}^2}{n_x+n_y-2} \qquad {}^{**}g = \frac{\left(\frac{S_{cx}^2}{n_x}+\frac{S_{cy}^2}{n_y}\right)^2}{\frac{(S_{cx}^2/n_x)^2}{n_x+1}+\frac{(S_{cy}^2/n_y)^2}{n_y+1}} - 2$$

## 4.5. ANEXO. Distribuciones de funciones muestrales

### 4.5.1. Distribución de la media muestral

**Definición 4.5.1.–** *Sea $X_1, \ldots, X_n$ una muestra aleatoria simple procedente de una variable aleatoria $X$. Se define la media muestral como:*

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

*Nota: $\overline{X}$ es una variable aleatoria, no una constante.*

**Teorema 4.5.2.–** *Si $X_1, \ldots, X_n$ es una muestra aleatoria simple procedente de una variable aleatoria $X$ con media $\mu$ y varianza $\sigma^2$, se tiene que: $E[\overline{X}] = \mu$ y $Var[\overline{X}] = \frac{\sigma^2}{n}$.*

*Se define la desviación típica de la muestra (error estándar de la media) como* $S_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

**Teorema 4.5.3.–** *Sea $X_1, \ldots, X_n$ una muestra aleatoria simple procedente de una variable aleatoria $X \sim N(\mu, \sigma^2)$. Entonces*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

*En consecuencia:*

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

**Teorema 4.5.4.–** *Si $X_1, \ldots, X_n$ es una muestra aleatoria simple de una variable aleatoria $X$, de distribución no especificada, con media $\mu$ y varianza $\sigma^2$ finita, entonces:*

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

En general, la aproximación es buena para $n \geq 30$

### 4.5.2. Distribución de la varianza muestral

**Definición 4.5.5.–** *Si $X_1 \ldots X_n$ es una muestra aleatoria simple procedente de una variable aleatoria $X$, se define la* **varianza muestral** *como:*

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

*y la* **cuasivarianza muestral** *como:*

$$S_c^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

Se verifica que $(n-1)S_c^2 = nS^2$ y $S^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \overline{X}^2$

**Teorema 4.5.6.–** *Sea $X_1, \ldots, X_n$ una muestra aleatoria simple procedente de una población $N(\mu, \sigma^2)$ de media y varianzas desconocidas. Entonces:*

1. *$\overline{X}$ y $S_c^2$ son independientes.*
2. *Si $Y = \frac{(n-1)S_c^2}{\sigma^2}$, entonces $Y \sim \chi^2_{n-1}$*
3. *Si $Y = \frac{nS^2}{\sigma^2}$, entonces $Y \sim \chi^2_{n-1}$*

**Proposición 4.5.7.–** *Sea $X_1, \ldots, X_n$ una muestra aleatoria simple procedente de una población $N(\mu, \sigma^2)$ de media y varianzas desconocidas. Entonces:*

1. *$\frac{\overline{X}-\mu}{S_c}\sqrt{n} \sim t_{n-1}$*
2. *$\frac{\overline{X}-\mu}{S}\sqrt{n-1} \sim t_{n-1}$*

### 4.5.3. Distribución de la diferencia de medias muestrales

Sea $X_1, \ldots, X_{n_x}$ una muestra aleatoria simple procedente de una variable aleatoria $X$ e $Y_1, \ldots, Y_{n_y}$ una muestra aleatoria simple procedente de una variable aleatoria $Y$, tales que $X$ e $Y$ son independientes y además $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$. Entonces,

si $\sigma_x^2 = \sigma_y^2 = \sigma^2$ conocida,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim N(0,1)$$

si $\sigma_x^2 \neq \sigma_y^2$, siendo ambas conocidas,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$$

si $\sigma_x^2 = \sigma_y^2$ pero ambas son desconocidas,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x-1)S_{cx}^2 + (n_y-1)S_{cy}^2}{n_x+n_y-2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim t_{n_x+n_y-2}$$

si $\sigma_x^2 \neq \sigma_y^2$, ambas desconocidas,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}}} \sim t_g$$

siendo

$$g = \frac{\left(\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}\right)^2}{\frac{(S_{cx}^2/n_x)^2}{n_x+1} + \frac{(S_{cy}^2/n_y)^2}{n_y+1}} - 2$$

### 4.5.4. Distribución del cociente de varianzas muestrales

Sea $X_1, \ldots, X_{n_x}$ una muestra aleatoria simple procedente de una variable aleatoria $X$ e $Y_1, \ldots, Y_{n_y}$ una muestra aleatoria simple procedente de una variable aleatoria $Y$, tales que $X$ e $Y$ son independientes y además $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$. Entonces:

$$\frac{S_{c_x}^2/\sigma_x^2}{S_{c_y}^2/\sigma_y^2} \sim \mathcal{F}_{n_x-1, n_y-1}$$

En el caso en que $\sigma_x = \sigma_y$ entonces:

$$\frac{S_{c_x}^2}{S_{c_y}^2} \sim \mathcal{F}_{n_x-1, n_y-1}$$