# 24. Estimación por intervalos

Al comienzo del capítulo del teorema central del límite (TCL) nos preguntamos: ¿Cuál es el peso medio de los recién nacidos? y ¿Cuál es la proporción de alumnos que no entienden estadística? Son muchas más las preguntas que podríamos responder si conociéramos los parámetros de diferentes poblaciones. En la práctica, lo único que podemos hacer es estimarlos; pero las estimaciones tienen error.

¿Entonces?

En vez de **estimar** un parámetro poblacional **mediante** un único número, podemos construir **un intervalo con** una **"garantía"** de cubrir al valor a estimar.

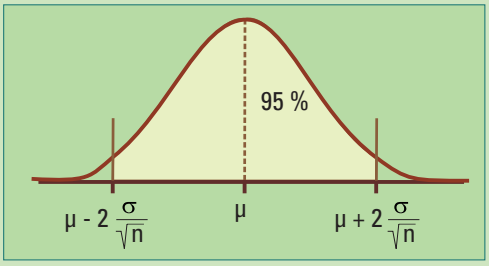
En este capítulo daremos **estimaciones por intervalos** de:

- la media de una población,
- la diferencia de medias de dos poblaciones,
- una proporción poblacional,
- la diferencia de dos proporciones poblacionales.

En todos los casos utilizaremos el TCL para construirlos.

## □ 24.1 Intervalos de confianza para la media

Si una variable se distribuye en la población con media **μ** y desvío estándar σ, se realizan muestreos aleatorios simples y se registran los valores ($\mathbf{x_1}$, **...** $\mathbf{x_n}$,) de dicha variable, entonces por el TCL la distribución de muestreo de su **media muestral** ($\bar{\mathbf{x}}$) es aproximādamente Normal, con media **μ** y desvío estándar $\frac{\sigma}{\sqrt{n}}$ .




***24.1.*** *En el eje horizontal se representan los valores de $\bar{x}$. La distribución aproximada es Normal. Cerca del 95% de los valores de muestreo de $\bar{x}$ se encuentran en el intervalo* $\left[\mu - 2\frac{\sigma}{\sqrt{n}};\ \mu + 2\frac{\sigma}{\sqrt{n}}\right]$

Cerca del 95% de las veces que calculemos una media muestral $\bar{\mathbf{x}}$ ésta se encontrará dentro del intervalo

$$\left[\mu - 2\frac{\sigma}{\sqrt{n}};\ \mu + 2\frac{\sigma}{\sqrt{n}}\right]$$

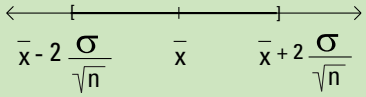
(fig24.1). O sea $\bar{\mathbf{x}}$ cumplirá las siguientes desigualdades:

$$\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 2\frac{\sigma}{\sqrt{n}}$$

Restando miembro a miembro $\bar{x}+\mu$, y multiplicando miembro a miembro por -1 en las desigualdades anteriores, obtenemos un intervalo para **μ** cuyos límites dependen de $\bar{x}$, **n** y σ:

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2\frac{\sigma}{\sqrt{n}}$$

Este intervalo está **centrado en $\bar{x}$** y tiene una semiamplitud de **2 desvíos estándar de la media muestral** ($2\frac{\sigma}{\sqrt{n}}$):




Cerca del 95% de las veces el intervalo $\left[\ \bar{x} - 2\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + 2\frac{\sigma}{\sqrt{n}}\ \right]$ **contiene a la verdadera media poblacional μ.**

¿A qué "veces" se refiere la frase anterior? A todas las veces que se realice un muestreo aleatorio simple y se calcule el intervalo.

El 95% de las veces el intervalo contiene a la verdadera media poblacional y 5% no. **Confiamos** en que el intervalo que construimos a partir de una única muestra efectivamente **contenga a μ,** pero no lo podemos saber con certeza. Decimos que:

$IC\ (\mu) = \left[\ \bar{x} - 2\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + 2\frac{\sigma}{\sqrt{n}}\ \right]$ es un intervalo de confianza para **μ** del 95% aproximadamente.

El intervalo puede escribirse en forma compacta observando que el mismo término, $2\frac{\sigma}{\sqrt{n}}$, suma o resta $\bar{x}$.

La **media muestral** es un **estimador puntual** de la media poblacional **μ,** es decir que obtenemos un número como estimación y no podemos decir nada respecto a si la media se parece o no se parece a **μ.** Un **intervalo de confianza** también es un **estimador de μ**, pero esta vez tenemos un rango de valores posibles y un grado de confianza (el % de veces que al obtener el intervalo este contendrá a **μ**). La figura 24.2 representa la construcción de intervalos de confianza de la forma $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$

para la media de una población muchas veces. Conocemos **μ** en este ejemplo y podemos saber cuáles son los intervalos de confianza que contienen a **μ** y cuáles no.

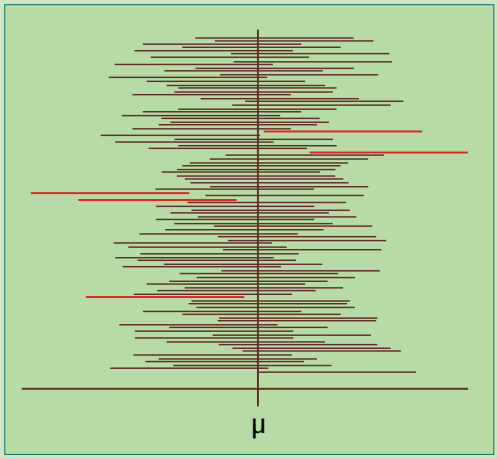



***Figura 24.2.*** *Cien intervalos de confianza del 95% para μ, obtenidos de la misma población. Con rojo aparecen los intervalos que no contienen a μ.*

Cinco de los 100 intervalos no contienen el valor verdadero del parámetro poblacional que estamos estimando mediante el intervalo.

Un intervalo de confianza, calculado a partir de los datos de una muestra, es uno de los tantos que podríamos obtener a partir de diferentes muestras. Podemos imaginar la construcción de un intervalo de confianza del 95% para μ como una selección al azar de un intervalo, entre todos los intervalos posibles (la figura 24.2 muestra 100 ellos). Algunos intervalos contienen a **μ** son los "buenos".

> ¡ Los intervalos de confianza no están centrados en **μ**; están centrados en $\bar{x}$.

> **Confiamos que el intervalo elegido sea uno de los "buenos"** y no uno "malo" porque estos son sólo el 5%. Pero, **no podemos saber si μ** pertenece al intervalo particular que construimos.

Si queremos aumentar el **nivel de confianza** a 99,7%, aumentamos la longitud del intervalo de confianza tomando 3 desvíos estándar:

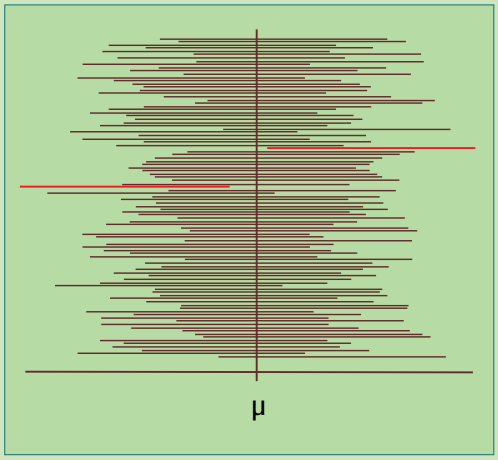



***Figura 24.3.*** *Cien intervalos de confianza del 99,7 %. Con rojo aparecen los intervalos que no contienen a μ. Esperamos que a la larga el 99,7% de los intervalos contengan a μ.*

$$\left[ \bar{x} - 3\frac{\sigma}{\sqrt{n}} ; \ \bar{x} + 3\frac{\sigma}{\sqrt{n}} \right]$$

es un **IC (μ) del 99,7%** aproximadamente.

Los intervalos de la figura 24.3 fueron construidos con los mismos datos que los de la figura 24.2 pero tomando $\bar{x} \pm 3\frac{\sigma}{\sqrt{n}}$

de manera que la longitud se multiplicó por un factor 3/2. **Perdemos en precisión** pero **ganamos en confianza.** Dos de los cien intervalos no contienen a **μ.** Si tomáramos más y más intervalos, en el largo plazo, tendríamos un 99,7% de intervalos "buenos".

**Ejemplo 3:** Supongamos que la fábrica que produce garrafas de gas comprimido de uso doméstico con 20 kg de capacidad nominal está realizando un nuevo proceso de producción, el anterior tenía una media de 47 $dm^3$. Interesa saber si la media del nuevo proceso ha cambiado, para ello obtiene las capacidades en $dm^3$ de 50 garrafas:

45,93 47,08 46,32 45,87 46,74 46,44 46,84 45,57 47,09 45,34 46,17 46,45 45,96 47,69 47,51 46,92 46,72 46,39 45,96 46,05 45,94 46,97 46,14 46,79 47,08 47,46 46,45 46,39 47,37 45,98 45,19 45,50 46,14 47,73 46,48 46,71 46,03 45,74 45,99 47,24 45,40 46,56 45,54 46,46 45,44 46,71 46,44 46,67 46,91 45,27

Por las características del proceso se sabe que $\sigma$**=0,75** no ha cambiado. Interesa estimar la capacidad media de las garrafas mediante un intervalo de confianza del 95%.

La media muestral de los 50 datos es $\bar{\mathbf{x}}$ = 46,3952 $dm^3$. Con una confianza del el 95% aproximadamente, podemos decir que la media de las capacidades de todas las garrafas producidas con el mismo proceso se encuentra dentro del intervalo

$$\left[ \bar{x} - 2\frac{\sigma}{\sqrt{n}};\ \bar{x} + 2\frac{\sigma}{\sqrt{n}} \right] = [46{,}3952 - 2 \times 0{,}75 / \sqrt{50}\ ;\ 46{,}3952 + 2 \times 0{,}75 / \sqrt{50}]$$

$$= [46{,}3952 - 2 \times 0{,}1061;\ 46{,}3952 + 2 \times 0{,}1061]$$
$$= [46{,}3952 - 0{,}2122;\ 46{,}3952 + 0{,}2122]$$
$$= [46{,}183;\ 46{,}607]$$

El valor 47 $dm^3$ no pertenece al intervalo. Si la media de la capacidad en $dm^3$ no cambió, el 95% de las veces que obtuviéramos un intervalo con otras muestras, pero con el mismo cálculo, el intervalo contendría el valor 47. Sospechamos que el nuevo proceso produce garrafas con menor capacidad.

¿Y si $\sigma$ es desconocida?

Rara vez el desvío estándar de la distribución de una variable en la población, $\sigma$, es conocido. Se utiliza **s** para estimar $\sigma$ y el **error estándar** ( $\frac{s}{\sqrt{n}}$, sección 23.2) para estimar $\frac{\sigma}{\sqrt{n}}$ obteniendo:

$$\left[ \bar{x} - 2\frac{s}{\sqrt{n}};\ \bar{x} + 2\frac{s}{\sqrt{n}} \right]$$ un IC (**μ**) **del 95%** aproximadamente.

y

$$\left[ \bar{x} - 3\frac{s}{\sqrt{n}};\ \bar{x} + 3\frac{s}{\sqrt{n}} \right]$$ un IC (**μ**) **del 99,7%** aproximadamente.

Podemos escribir los intervalos para **μ** de aproximadamente del **95% y 99,7%** nivel de confianza, respectivamente, en forma más compacta:

$\bar{\mathbf{x}}$**±2**.(error estándar) y $\bar{\mathbf{x}}$**±3**.(error estándar).

**Nos interesa que la longitud de los intervalos de confianza sea pequeña.** Si sumamos y restamos a la media muestral un término muy grande, significa que $\bar{x}$ es una estimación poco precisa de la media poblacional **μ.**

Tres factores afectan la longitud del intervalo de confianza:

- **El nivel de confianza.** Para 95% el desvío estándar de la media se multiplica por 2, y para 99% por 3.
- **El tamaño de la muestra n.** Aparece como $\frac{1}{\sqrt{n}}$ . Si n =9, $\frac{1}{\sqrt{n}}$ =1/3=0,333.
- Si n=36, $\frac{1}{\sqrt{n}}$ =1/6=0,166. Para reducir la longitud a la mitad, el tamaño de la muestra debe multiplicarse por 4.
- **La variabilidad de la variable en la población (σ).** A mayor variabilidad de los datos individuales en la población tendremos una longitud mayor del intervalo de confianza.

**Ejemplo 3:** Continuación.

La encargada del control de procesos en la fábrica de garrafas necesita estimar el volumen de las garrafas con mayor precisión. Desea saber qué tamaño de muestra debe elegir para obtener un intervalo de longitud 0,3 $dm^3$ con una confianza del 95%.

Sabemos que el intervalo es $\mathbf{x \pm 2 \frac{\sigma}{\sqrt{n}}}$ y **σ=0,75.** Por lo tanto, la longitud del intervalo es:

$$L = 4 \frac{\sigma}{\sqrt{n}}$$

Pero interesa que L=0,3 $dm^3$, o sea:

$$L = 4 \frac{\sigma}{\sqrt{n}}$$
$$L = 0{,}3$$

Por lo tanto:

$$n = \left(4 \frac{0{,}75}{0{,}3}\right)^2$$
$$n = 100$$

Originalmente se había elegido una muestra de tamaño **n=50,** para obtener la precisión deseada se **debe duplicar ese tamaño de la muestra.**

Podemos preguntarnos si cuánto más pequeña sea la longitud del intervalo, siempre es mejor. Para obtener intervalos de confianza cada vez más estrechos es necesario aumentar más y más el tamaño de la muestra. Esto encarece y dificulta el estudio. En cada situación, el investigador deberá decidir el tamaño de la muestra equilibrando la longitud tolerable para sus intervalos de confianza y sus posibilidades. Estas consideraciones son válidas para todos los intervalos de confianza que presentaremos en el capítulo.

## □ 24.2. Intervalos de confianza para la diferencia de medias

El objetivo de muchas encuestas y estudios es comparar dos poblaciones, como los hombres frente a mujeres, los pacientes tratados con la droga A con los tratados con la droga B, las familias de bajos ingresos con las familias de ingresos altos, los individuos con estudios secundarios completos y con estudios secundarios incompletos, respecto de algunas variables. Cuando estas son numéricas (por ejemplo, altura, peso, nivel de colesterol en sangre o ingresos) los parámetros de interés suelen ser sus medias en cada población. Sus estimadores son medias muestrales.

Nuevamente, podemos utilizar el teorema central del límite para hallar la distribución de muestreo de la diferencia de medias muestrales.

Supongamos que una variable numérica se distribuye en una población con media $\mu_X$ y desvío estándar $\sigma_X$, y en otra población con media $\mu_Y$ y desvío estándar $\sigma_Y$. Si se realizan muestreos aleatorios simples de tamaño $n_X$ y $n_Y$ de cada una de las poblaciones respectivamente, se registran los valores (**x**) e (**y**) de dicha variable en cada una de las poblaciones y se calcula la diferencia ($\bar{x}$-□) de las correspondientes medias muestrales entonces:

- La distribución de muestreo de $\bar{x}-\bar{y}$, es **aproximadamente Normal** con tal de tomar tamaños de muestra $n_X$ y $n_Y$ suficientemente grandes.
- Cuanto mayor sean los tamaños de las muestras ($n_X$ y $n_Y$) tanto mejor será la aproximación. Si $n_X$ y $n_Y$ son por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de muestreo de $\bar{x}-\bar{y}$, es $\mu_X-\mu_Y$.
- El desvío estándar de la distribución de muestreo de $\bar{x}-\bar{y}$, es $\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$. Decrece cuando aumentan los tamaños de las muestras ($n_X$ y $n_Y$).
- La distribución de muestreo de $\frac{\bar{x}-\bar{y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}$ es **aproximadamente Normal** Estándar con tal de tomar tamaños de muestra $n_X$ y $n_Y$ suficientemente grandes.

**Observación:** Nuevamente, **la distribución de la variable no es necesariamente Normal para ninguna** de las dos poblaciones estudiadas. La **distribución** de muestreo de la diferencia de las medias muestrales **sí lo es** aproximadamente, cuando los tamaños de las muestras son suficientemente grandes.

Frecuentemente $\sigma_X$ y $\sigma_Y$ no se conocen, entonces se los estima respectivamente, por los desvíos estándar de cada una de las muestras (sección 18.2.3):

$$s_X = \sqrt{\frac{1}{n_X - 1} \sum_{i=1}^{n_X} (x_i - \bar{x})^2} \quad \text{y} \quad s_y = \sqrt{\frac{1}{n_Y - 1} \sum_{i=1}^{n_y} (y_i - \bar{y})^2}$$

El desvío estándar estimado de la distribución de $\bar{x}$-□ es:

**El error estándar de** $\bar{x} - \bar{y} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$

Una estimación para la diferencia de dos medias poblacionales se obtiene tomando la diferencia de medias muestrales (una de cada una de las poblaciones) y sumándole y restándole un margen de error en forma similar a lo que vimos en la sección 24.1 para una única media.

El **intervalo** de nivel **de confianza** aproximado de 95 % **para la diferencia de medias** ($\mu_x$-$\mu_y$) tiene la misma estructura que para una media:

$$IC\ (\mu_x - \mu_y):\ \bar{x} - \bar{y} \pm 2 \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

**Ejemplo 4:** Llamamos X=peso de un varón de 16 años, Y=peso de una mujer de 16 años. Interesa hallar un intervalo de confianza del 95% para la diferencia de medias de los pesos de varones y mujeres de esa edad.

A partir de los datos del ejemplo 16.4. (supondremos que se trata de pesos provenientes de **muestras representativas** de las poblaciones de varones y mujeres de 16 años de una gran ciudad) obtenemos los siguientes resultados:

| Variable | n | Media muestral | Desvío estándar muestral (s) | Error estándar $s/\sqrt{n}$ |
|---|---|---|---|---|
| X | 52 | 66,096 | 6,6488 | 0,9220 |
| Y | 49 | 51,265 | 6,2141 | 0,8877 |

Calculemos primero el error estándar de la diferencia de medias

$$\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} = \sqrt{(0,9220)^2 + (0,8877)^2}$$

$$= \sqrt{1,6381}$$

$$= 1,28$$

Por lo tanto, un intervalo de confianza del 95% para la diferencia de medias de los pesos de varones y mujeres de 16 años es:

$$\bar{x} - \bar{y} \pm 2\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

$$66{,}1 - 51{,}3 \pm 2 \cdot 1{,}28$$

$$14{,}8 \pm 2{,}56$$

Estimamos con una confianza aproximada del 95% que la diferencia de pesos de varones y mujeres de 16 años se encuentra en el intervalo (12,24; 17,36 ).

## □ 24.3. Intervalos de confianza para una proporción

Para cualquier población con una proporción **p** de éxitos y una muestra de tamaño **n** siempre que $np>10$ y $n(1-p)>10$, si $\hat{p} = \frac{\text{cantidad de éxitos en la muestra}}{\text{tamaño de la muestra}}$ , el TCL asegura que:

- La distribución de $\hat{\mathbf{p}}$ es **aproximadamente Normal.**
- La **media** de la distribución de $\hat{\mathbf{p}}$ es **p.**
- El **desvío estándar** de la distribución de **p** es $\sqrt{\frac{p(1-p)}{n}}$ **.**

Por lo tanto, por las propiedades de la curva Normal (sección 20.2.3),

- Cerca del 95% de las proporciones muestrales están entre **p** menos 2 veces el desvío estándar y **p** más 2 veces el desvío estándar, o sea dentro de $p \pm 2\sqrt{\frac{p(1-p)}{n}}$ .

- El 99,7% (casi todas) de las proporciones muestrales se encuentran en el intervalo $\left( p - 3\sqrt{\frac{p(1-p)}{n}} \; , \; p + 3\sqrt{\frac{p(1-p)}{n}} \right)$ , o sea dentro de $p \pm 3\sqrt{\frac{p(1-p)}{n}}$ .

Podemos realizar cálculos similares a los realizados para la media muestral, pero esta vez para obtener un **intervalo con nivel de confianza de aproximadamente 95% para p:**

Sabemos que cerca del 95% de las veces que calculemos $\hat{\mathbf{p}}$

éste se encontrará dentro del intervalo $\left[ p - 2\sqrt{\frac{p(1-p)}{n}} \; ; \; p + 2\sqrt{\frac{p(1-p)}{n}} \right]$

O sea, $\hat{\mathbf{p}}$ cumplirá las siguientes desigualdades:

$$p - 2\sqrt{\frac{p(1-p)}{n}} \leq \hat{p} \leq p + 2\sqrt{\frac{p(1-p)}{n}}$$

Restando miembro a miembro $\hat{\mathbf{p}}$-**p** y multiplicando miembro a miembro por -1 en las desigualdades anteriores, obtenemos un intervalo para **p** cuyos límites dependen de $\hat{\mathbf{p}}$ y **n**:

$$\hat{p} - 2\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}$$

Este intervalo está centrado en $\hat{p}$ y tiene una semiamplitud de **2 desvíos estándar** $2\sqrt{\frac{p\,(1-p)}{n}}$

$$\hat{p} - 2\sqrt{\frac{p\,(1-p)}{n}} \qquad \hat{p} \qquad \hat{p} + 2\sqrt{\frac{p\,(1-p)}{n}}$$

Cerca del 95% de las veces el intervalo $\left[\ \hat{p} - 2\sqrt{\frac{p\,(1-p)}{n}}\ ;\ \hat{p} + 2\sqrt{\frac{p\,(1-p)}{n}}\ \right]$ **contiene a la verdadera proporción poblacional p.**

El intervalo también puede escribirse como $\hat{p} \pm 2\sqrt{\frac{p\,(1-p)}{n}}$. Pero este intervalo no sirve en la práctica porque no conocemos **p.**

¿Qué se puede hacer? Veamos 2 opciones:

**Procedimiento 1.** Estimar el desvío estándar $\left(\sqrt{\frac{p\,(1-p)}{n}}\right)$ por el error estándar $\left(\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right)$ obteniendo el intervalo:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

**Procedimiento 2.** Como $\sqrt{p(1-p)} \leq 0{,}5$ (figura 24.4) podemos acotar el desvío estándar eligiendo el valor más grande de $\sqrt{p(1-p)}$ $(0{,}5)$. De esta manera podemos decir: el desvío estándar $\leq 0{,}5\,\frac{1}{\sqrt{n}}$. Si tomamos error estándar $= 0{,}5\,\frac{1}{\sqrt{n}}$, resulta el intervalo: $\hat{p} \pm \frac{1}{\sqrt{n}}$

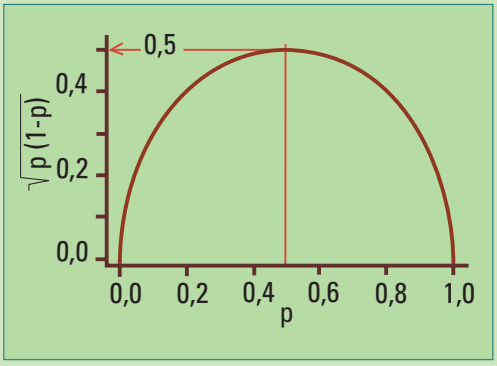



***Figura 24.4.*** *Gráfico de la función* $f(p) = \sqrt{p(1-p)}$ *para* $0 \leq p \leq 1$.

Se denomina **margen de error** de $\hat{p}$, a la semiamplitud del intervalo de confianza del 95%.

Del procedimiento 1 resulta que el margen de error (sobreentendiendo que se trata del 95%) de $\hat{p}$ es:

$$2 \times 0{,}5\,\sqrt{\frac{1}{n}} = \sqrt{\frac{1}{n}}$$

cualquiera sea **p.**

Obtenemos así el

**método rápido** para el cálculo del **margen de error** de $\hat{p}$ $\qquad \sqrt{\frac{1}{n}}$

**Ejemplo 5.** Retomemos nuevamente el ejemplo del Club Grande de Fútbol (capítulo 9) la investigadora tomó una muestra de 538 socios y obtuvo $\hat{p}$**=0,51.** Utilicemos ambos procedimientos para calcular el margen de error:

1. Reemplazar **p** por $\hat{p}$: $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2\sqrt{\frac{0,51\ (1-0,51)}{538}}$

$$= 2\sqrt{\frac{0,2499}{538}}$$

$$= 0,04310$$

2. Método rápido. El margen de error $= \sqrt{\frac{1}{538}}$

$$= 0,04311$$

Obtuvimos prácticamente el mismo resultado con los dos procedimientos, esto se debe a que $\hat{p}$**=0,51** está muy cerca de 0,5 donde se alcanza el máximo de $\sqrt{p(1-p)}$ para valores de **p** entre 0 y 1 (figura 24.2). Cuando **p** es pequeño $\frac{1}{\sqrt{n}}$ sobreestima el desvío estándar.

El intervalo de confianza del 95% para la proporción de socios a favor de Rolando Forzudo es: 0,51±0,04, cualquiera sea el procedimiento empleado. El intervalo puede expresarse en porcentajes: 51 %±4 %. Decimos que el resultado de la encuesta es 51% a favor de Rolando Forzudo con un margen de error del 4%. **Confiamos** que el verdadero porcentaje se encuentre dentro del intervalo [50,6; 51,4]. **¿Por qué confiamos?** Porque el 95% de las veces que utilicemos este procedimiento para obtener un intervalo de confianza para una proporción, el mismo contendrá al valor verdadero, pero **no podemos estar seguros** que eso ocurrió **esta vez.**

Cualquier intervalo de confianza para una proporción puede expresarse en porcentajes, simplemente hay que multiplicar los extremos del intervalo por 100.

El margen de error del 95% obtenido con el método rápido, $\sqrt{\frac{1}{n}}$ es siempre **mayor o igual** al margen de error calculado con el valor de $\hat{p}$: $2\sqrt{\frac{\hat{p}(1-p)}{n}}$

## □ 24.4. Intervalos de confianza para la diferencia de proporciones

Muchos estudios requieren comparar proporciones entre dos poblaciones. Por ejemplo, para comparar la proporción de mujeres y hombres a favor de establecer un salario de desempleo, o la proporción de argentinos que prefiere los automóviles chicos en comparación

con los españoles, o la proporción de pacientes tratados con la droga A que reduce su dolor de cabeza en comparación con los tratados con la droga B, etc.

En todos los ejemplos anteriores tenemos **dos poblaciones** (hombres-mujeres; argentinos-españoles; pacientes tratados con la droga A-pacientes tratados con la droga B) y una característica de interés que llamamos **"éxito".** Éxito puede representar estar a favor un salario de desempleo, preferir automóviles chicos, reducir su dolor de cabeza.

Sean:
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1
$\mathbf{n_2}$= tamaño de la muestra de la población 2
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2

También se puede aplicar el TCL para la distribución de muestreo de la diferencia de proporciones:

- Si $\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ son todos mayores a 10, entonces la distribución de muestreo de $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ es **aproximadamente Normal.**
- La **media** de la distribución de $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ es $\mathbf{p_1 - p_2}$.
- El desvío estándar de la distribución de $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ es $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ .
- La distribución de muestreo de $\frac{\hat{p}_1 - \hat{p}_1 - (p_1 - p_1)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ es **aproximadamente Normal Estándar** si $\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ son todos mayores a 10.

En la práctica no se conocen $\mathbf{p_1}$ ni $\mathbf{p_2}$, por lo cual la condición de los tamaños muestrales ($\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ todos mayores a 10) se verifica reemplazando $\hat{\mathbf{p}}_1$ y $\hat{\mathbf{p}}_2$ por **y.**

**Las muestras deben:**
- ser independientes entre sí,
- ser a lo sumo un 10% de la población,
- provenir de un muestreo aleatorio simple.

Luego el intervalo de aproximadamente el 95% de confianza, para la diferencia de proporciones ($\mathbf{p_1\text{-}p_2}$) basado en las proporciones muestrales, es:

$$IC\,(p_1\text{-}p_2): \hat{p}_1\text{-}\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

y el intervalo es aproximadamente el 99,7%, para la diferencia de proporciones ($p_1$ -$p_2$) basado en las proporciones muestrales es:

$$IC\,(p_1\text{-}p_2):\hat{p}_1\text{-}\hat{p}_2 \pm 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

La diferencia de proporciones muestrales ($\hat{\mathbf{p}}_1$- $\hat{\mathbf{p}}_2$) es un **estimador** de la diferencia de las proporciones poblacionales $\mathbf{p_1}$-$\mathbf{p_2}$ cuyo **desvío estándar** es:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

**Ejemplo 6:** Supongamos que en un muestreo aleatorio simple de 125 choferes de taxis de la Ciudad de Buenos Aires, el 64% opina que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año. Mientras que de una encuesta similar realizada entre 150 choferes de taxis del Conurbano Bonaerense, el 76% dio esa misma respuesta. Construya un intervalo de confianza del 95% para la verdadera diferencia de proporciones.

Sean:
**Población 1:** todos los choferes de taxis de la Ciudad de Buenos Aires.
**Población 2:** todos los choferes de taxis del Conurbano Bonaerense.
**"Éxito"** = opinar que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1 =125
$\mathbf{n_2}$= tamaño de la muestra de la población 2 =150
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1 = 0,64
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2 = 0,76

Primero:

$n_1\,\hat{p}_1 = (125)(0,64)$ $\qquad$ $n_2\,\hat{p}_2 = (150)(0,76)$

$n_1\,\hat{p}_1 = 80$ $\qquad$ $n_2\,\hat{p}_2 = 114$

$n_1\,(1\text{-}\hat{p}) = (125)(1\text{- }0,64)$ $\qquad$ $n_2\,(1\text{-}\hat{p}) = (150)(0,76)$

$n_1\,(1\text{-}\hat{p}) = 45$ $\qquad$ $n_2\,(1\text{-}\hat{p}) = 36$

Son todos mayores a 10.

Segundo: Las muestras son menos del 10% de la población de choferes de taxis, tanto para la Ciudad de Buenos Aires como para el Conurbano Bonaerense.

Tercero:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}=\sqrt{\frac{0,64\,(1-0,64)}{125}+\frac{0,76\,(1-0,76)}{150}}$$

$$= 0,0553$$

$\hat{p}_1 - \hat{p}_2 = 0{,}64 - 0{,}76$
$= -0{,}12$

¡La diferencia de proporciones puede ser negativa!

Finalmente, el intervalo resulta $\hat{p}_1 - \hat{p}_2 \pm 2 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}}{n_2}}$

$-0{,}12 \pm 2 \times 0{,}0553$

$-0{,}12 \pm 0{,}1106$

$(-0{,}2306 ; -0{,}0094)$

Podemos asegurar con un 95% de confianza que la diferencia de proporciones de los choferes de taxis que opinan que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año, se encuentra entre -0,2306 y -0,0094. Con un 95% de confianza podemos decir que la proporción es menor en la Ciudad de Buenos Aires.

**Ejemplo 7:** Retomemos el ejemplo 10 de la sección 22.5. (Dos variables Categóricas) en el que interesa estudiar si existe asociación entre dos variables categóricas: "come rápido" y "sobrepeso" ambas con categorías "sí, "no". Dentro del grupo de individuos que come rápido el 30 % tiene sobrepeso, mientras que entre los que no comen rápido ese porcentaje se reduce al 10 %. Los porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido son bastante diferentes. Necesitamos saber si esa diferencia puede atribuirse a la variabilidad que surge del muestreo para decidir que las variables están asociadas.

Sean:
**Población 1:** todos los individuos que comen rápido.
**Población 2:** todos los individuos que no comen rápido.
**"Éxito" =** tener sobrepeso.
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1 = 250
$\mathbf{n_2}$= tamaño de la muestra de la población 2 = 220
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1 = 0,3
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2 = 0,1

Primero:

$n_1 \hat{p}_1 = (250)(0{,}30)$ $\quad$ $n_2 \hat{p}_2 = (220)(0{,}10)$
$n_1 \hat{p}_1 = 75$ $\quad$ $n_2 \hat{p}_2 = 22$

$n_1 (1-\hat{p}_1) = (250)(0{,}70)$ $\quad$ $n_2 (1-\hat{p}_2) = (220)(0{,}90)$
$n_1 (1-\hat{p}_1) = 175$ $\quad$ $n_2 (1-\hat{p}_2) = 198$

son todos mayores a 10.

Segundo: Las muestras son menos del 10% de las poblaciones en consideración

Tercero:
$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-p)}{n_2}} = \sqrt{\frac{0{,}30(1-0{,}30)}{250}+\frac{0{,}10(1-0{,}10)}{220}}$$
$$= 0{,}0939$$

$\hat{p}_1 - \hat{p}_2 = 0{,}30 - 0{,}10$

$\hat{p}_1 - \hat{p}_2 = 0{,}20$

Finalmente, el intervalo resulta: $\hat{p}_1 - \hat{p}_2 \pm 2 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$0{,}20 \pm 2 \times 0{,}0939$

$0{,}20 \pm 0{,}1878$

$(0{,}0122 ; 0{,}3878)$

Como el cero no es un valor del intervalo, podemos asegurar con un 95% de confianza que las proporciones poblacionales $\mathbf{p_1}$ **y** $\mathbf{p_2}$ son distintas. Esto es que las variables "sobrepeso" y "come rápido" están asociadas.

## □ 24.5. Consideraciones generales sobre intervalos de confianza

**Todos los intervalos de confianza** presentados tienen **la misma forma:**

**Estimador ± K x desvío estándar** del estimador.

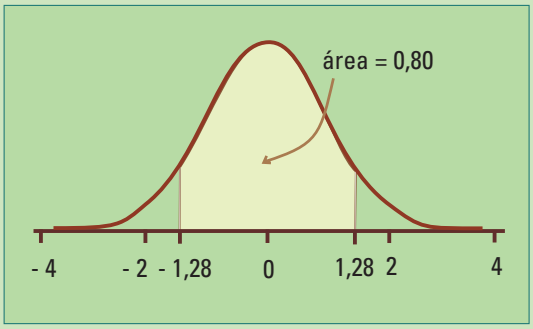
Hemos utilizado la regla 68-95-99,7 para obtener intervalos con niveles aproximados de confianza del 95% y 99,7% tomando **K=2** y **K=3** respectivamente. Pero, puede ocurrir que consideremos que un 95% de confianza es un criterio demasiado exigente para nuestro problema en particular, o que necesitamos un nivel de confianza mayor. Es posible obtener intervalos con **cualquier nivel de confianza** entre 0 y 100% utilizando diferentes valores de **K.** La tabla 24.1 presenta diferentes valores de **K,** y sus correspondientes niveles de confianza.

En particular, la figura 24.3 muestra el área de 80% bajo la curva Normal entre -1,29 y 1,29; si se elige K=1,29 para calcular el intervalo de confianza para un parámetro utilizando la forma:

**Estimador ± K x desvío estándar** del estimador.

## Niveles de confianza y valores de K obtenidos de la distribución N(0,1). TABLA 24.1

| Niveles de confianza | K | Niveles de confianza | K |
|---|---|---|---|
| **50 %** | 0,67 | 95,000% | 1,96 |
| **60 %** | 0,84 | 99,000% | 2,58 |
| **70 %** | 1,04 | 99,000% | 3,29 |
| **80 %** | 1,28 | 99,990% | 3,89 |
| **90 %** | 1,64 | 99,999% | 4,42 |




***24.5.*** *Área bajo la curva Normal entre -1,28 y 1,28.*

La regla 68-95-99,7 utiliza el 2 como aproximación al valor **K=1,96.** En la mayoría de las aplicaciones esta aproximación es adecuada, pero la tabla 23.2 muestra el valor más exacto 1,96.

**Ejemplo 7.** Continuación.
Construyamos el intervalo del 95% de confianza para la diferencia de porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido, pero esta vez utilizando 1,96 en vez de 2:

$$0{,}20 \pm 1{,}96 \times 0{,}0939$$
$$0{,}20 \pm 0{,}1840$$
$$(0{,}016;\ 0{,}384)$$

El resultado es esencialmente el mismo que habíamos obtenido antes.

## 24.6. Actividades y ejercicios

1. Una fábrica produce dardos con diámetros que tienen un desvío estándar $\sigma$= 0,25mm. De un lote grande, se seleccionó una muestra aleatoria simple de 40 dardos, y se obtuvo un promedio de 3,09 mm en sus diámetros.

a. Obtenga un intervalo de confianza de aproximadamente el 95% para la media de los diámetros de todos los dardos producidos de la misma forma.
b. Idem b pero con una confianza aproximada del 99.7%
c. De qué tamaño debe ser la muestra para que la longitud del intervalo de confianza del 95% sea 0,1 mm.

En los ejercicios 2 a 7 se presentan varias respuestas, elija la correcta y explique brevemente.

2. Cambiar de 95% a 99,7% el nivel de confianza de un intervalo para una proporción, dejando el resto igual.

a. Aumenta la longitud del intervalo en 4,7%
b. Reduce la longitud del intervalo en 4,7%
c. Aumenta la longitud del intervalo en 50%
d. Reduce la longitud del intervalo en 50%
e. No puede saberse sin conocer el tamaño de la muestra.

3. Se prueban 49 autos de un nuevo modelo y se registran los litros de nafta consumidos en un recorrido de 100 km, obteniéndose una media muestral, **x=6,8** litros y un desvío estándar muestral, **s=1,4** litros. Obtenga un intervalo de aproximadamente 95% de confianza para la cantidad media de litros de nafta consumida por ese tipo de vehículo en 100 km.

a. [5,4; 8,2]
b. [6,6; 7,0]
c. [6,4; 7,2]
d. [6,2; 7,4]

4. Se sabe que el 82 % de los alumnos del último año de las escuelas secundarias dependientes de alguna universidad planean seguir estudios superiores. Supongamos que se selecciona una muestra aleatoria simple de alumnos del último año de dichas escuelas y se obtiene un intervalo de confianza en base a la proporción que manifiesta tener interés en continuar sus estudios. Entonces:

a. El centro del intervalo de confianza es 0,82.
b. El intervalo de confianza contiene el valor 0,82.
c. Un intervalo de confianza del 99,7% contiene el valor 0,82.

5. En general, ¿cómo cambia la longitud del intervalo de confianza si se duplica el tamaño de la muestra y todo lo demás queda igual?

a. Se duplica la longitud.
b. La longitud se reduce a la mitad
c. La longitud se multiplica por 1,414
d. La longitud se divide por 1,414
e. No se puede saber.

6. Una encuesta reveló que el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes de anticipación es 68% con un margen de error del ± 5%. ¿Qué significa el ± 5%?

a. Se encuestó al 5 % de la población.
b. En la muestra, el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes anticipación se encontró entre 63% y 73%.
c. En la población, el porcentaje de mujeres que no le gusta planificar sus vacaciones más de un mes con anticipación está entre 63% y 73%.
d. Se encuestó entre 63% y 73% de la población.
e. Sería raro que en la población el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes anticipación esté fuera del intervalo de 63% a 73%.

7. En un muestreo aleatorio simple de 300 hombres mayores de 70 años, el 48 % eran viudos y en un muestreo aleatorio simple de 400 mujeres en el mismo rango de edades, 65% eran viudas. Halle un intervalo de confianza para la diferencia entre el porcentaje de viudas y viudos.

a) 17% ± 0,38 %
b) 17% ± 7,48 %
c) 55,6 % ± 7,48 %
d) 56,5 % ± 0,74 %
e) 56,5 % ± 0,38 %

8. Se realizó un muestreo aleatorio simple, de un embarque de 50.000 piezas delicadas, registrándose 16 piezas dañadas de un total de 220 observadas. Obtenga un intervalo del 95% de confianza para estimar la verdadera proporción y a partir de él la cantidad de piezas dañadas.

9. El desvío estándar de la distribución de muestreo de $\hat{\mathbf{p}}$ es $\sqrt{\frac{p(1-p)}{n}}$ , depende de **p**. Para entender cómo se comporta para distintos valores de **p**, grafique en el eje vertical $\sqrt{p(1-p)}$ y en el eje horizontal **p** para los siguientes valores de **p**: 0 0,1 0,2 0,3 ... 0,9 1. Trace una curva que una los puntos. Observe que el gráfico alcanza su máximo para **p=0,5.** El margen de error calculado con $\hat{\mathbf{p}}$ ¿será menor o igual que el obtenido por el procedimiento rápido?