23. Teorema central del límite (TCL)

Uno de los resultados más importantes de la teoría estadística.

En algunas ocasiones nos interesa saber, por ejemplo, cuál es el peso medio de los recién nacidos, o cuál es la proporción de alumnos que no entienden estadística, sin embargo, en realidad no lo podemos saber exactamente. Aún cuando se recolecte con cuidado una muestra tan grande y representativa como sea posible, el error debido al muestreo es inevitable.

Nos interesa conocer los **parámetros poblacionales,** la media μ_X y el desvío σ_X estándar de una variable X en la población (es decir de todos los valores que toma la variable en la población); pero sólo podemos hallar la **media muestral** ($\bar{x}$) y el desvío estándar muestral (s). Se trata de **estadísticos** cuyos valores cambian entre una muestra y otra.

El resultado descripto en este capítulo permite cuantificar el error debido al muestreo. Comenzamos con preguntar: ¿qué pasaría si tomásemos muchas muestras de la misma población?

Ya planteamos esa pregunta en el capítulo 10 para proporciones, ahora consideraremos primero la media muestral y luego volveremos con las proporciones.

□ 23.1. Distribución de muestreo de la media muestral

Describiremos la **distribución de muestreo** de la **media muestral.** ¿Qué significa esto? ¿Cuál es la población? La **población** de muestreo de la media muestral está compuesta por **todas las medias muestrales** que se podrían obtener de una población utilizando **muestreos aleatorios simples,** observando los valores de alguna variable (**X**) y calculando la media muestral para cada una de los muestreos realizados. La distribución de muestreo de la media muestral tiene, como cualquier otra distribución, forma, centro y una medida de variabilidad. Conocer la distribución de muestreo de la media muestral permite estimar su error.

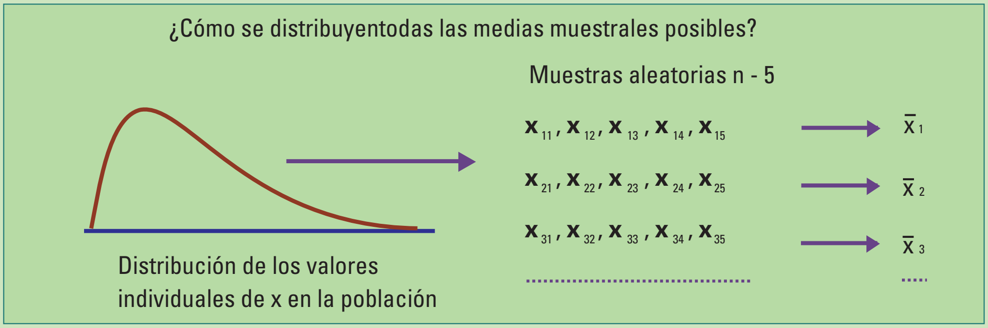


Figura 23.1. *Repetición del muestreo aleatorio simple de tamaño n=5.*

Muchas veces diremos: **"distribución de $\bar{x}$"**, sobreentendiendo que se trata de la **distribución de las medias muestrales en un muestreo aleatorio simple** de muestras de tamaño **n.**

Para tener una idea más concreta acerca de cómo se compara la distribución de una variable (**X**) en la población y la distribución de muestreo de $\bar{\mathbf{x}}$, veamos un caso particular de cómo se distribuyen 20 medias muestrales, calculadas a partir de muestras de tamaño n=4 y todos los valores individuales.

Ejemplo 1: volvamos a la fábrica que produce **garrafas de gas comprimido** de uso doméstico con 20 kg de capacidad nominal (sección 21.1.1.1), pero esta vez, para estudiar como se comporta la **media muestral.** Seleccionamos una muestra de 4 garrafas de un lote grande y calculamos la media de los volúmenes obtenidos. Repetimos este procedimiento 20 veces:

¿Cómo se distribuyen los volúmenes de las garrafas seleccionadas? ¿Cómo se distribuyen las medias muestrales?

Los valores individuales se encuentran entre 45,37 y 48,48 (dentro de un intervalo de longitud 3,19), las medias están entre 46,59 y 47,42 (ahora, el intervalo es de longitud menor a 1).

VOLÚMENES (dm^3) DE 80 GARRAFAS DE 20 kg.

TABLA 23.1

					Media muestral x
Muestra 1	46.40	47.62	48.39	47.75	47.54
Muestra 2	47.08	46.39	45.37	48.08	46.73
Muestra 3	47.05	47.54	47.34	47.79	47.43
Muestra 4	46.68	47.51	46.89	47.39	47.12
Muestra 5	46.91	48.00	47.97	46.82	47.42
Muestra 6	47.29	46.73	46.47	47.03	46.88
Muestra 7	46.92	47.07	47.23	47.36	47.14
Muestra 8	47.76	46.88	48.16	47.26	47.51
Muestra 9	46.30	47.76	47.18	46.65	46.97
Muestra 10	46.79	45.83	45.70	46.75	46.27
Muestra 11	47.08	45.57	46.45	48.48	46.90
Muestra 12	46.57	46.54	46.98	46.35	46.61
Muestra 13	48.37	48.04	46.65	46.01	47.27
Muestra 14	47.61	47.02	47.44	46.87	47.24
Muestra 15	46.92	47.52	46.80	46.50	46.94
Muestra 16	47.07	47.90	46.33	46.46	46.94
Muestra 17	46.76	47.17	47.33	46.34	46.90
Muestra 18	47.75	47.39	47.14	46.18	47.12
Muestra 19	46.67	47.04	48.15	47.50	47.34
Muestra 20	46.28	47.12	46.23	47.52	46.79

Para tener una imagen visual construimos dos histogramas en escala densidad (Recordemos que en escala densidad el área de cada rectángulo coincide con la proporción de datos que pertenecen al correspondiente intervalo de clase, sección 17.2). El primero con todos los datos individuales (columnas 2, 3, 4 y 5 de la tabla 23.1) y el segundo con las medias muestrales (columna 6). Utilizamos las mismas escalas en los ejes de ambos histogramas para poder compararlos.

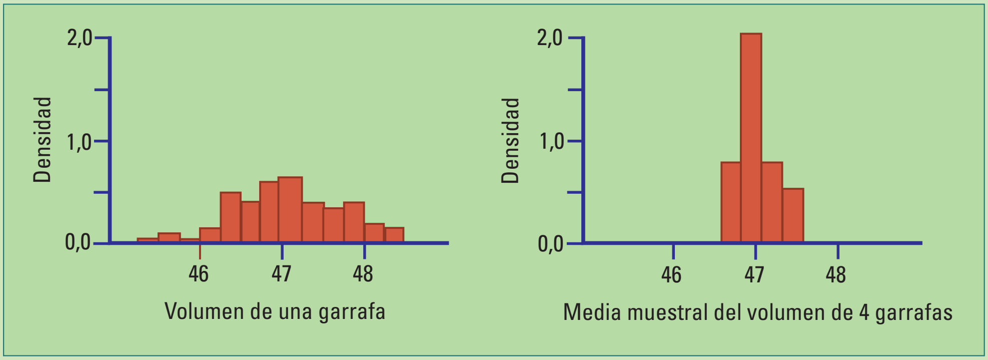
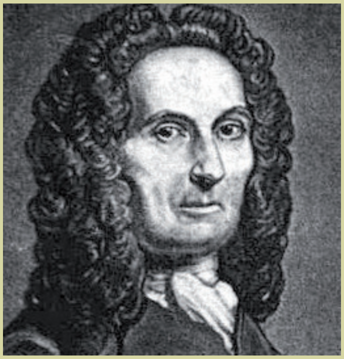


Figura 23.2. *Histogramas de datos de volumen de garrafas de 20 kg. Las medias muestrales (histograma lado derecho) están más concentradas alrededor del valor central 47 que los valores individuales (histograma lado izquierdo).*



Abraham de Moivre 1667 - 1754. Matemático francés.

En 1733, postuló la primera versión del Teorema Central del Límite en el marco de los juegos de azar. Específicamente, lo hizo como descripción de la distribución de "la cantidad de caras" que resultan al arrojar "una moneda equilibrada" muchas veces.

Muchos años después (1812), el tema fue retomado, generalizado a la "cantidad de caras de cualquier moneda" (aunque sea una moneda cargada) y demostrado por otro matemático francés: Pierre-Simon Laplace.

Las medias muestrales tienen menor variabilidad que los valores individuales.

¡Por eso promediamos!

Le creemos más al resultado de un promedio que a un resultado individual.

Además de tener menor variabilidad, la distribución de las medias muestrales tiene otra propiedad muy importante. Esta propiedad la establece el Teorema Central del Límite.

□ 23.2. Enunciado del TCL

Supongamos que una variable se distribuye en la población con media μ y desvío estándar σ. Si se realizan muestreos aleatorios simples y se registran los valores ($x_1, ..., x_n$) de dicha variable entonces:

- La distribución de muestreo de $\bar{x}$, es **aproximadamente Normal** con tal de tomar tamaños de muestra suficientemente grandes.
- Cuanto mayor sea el tamaño de la muestra (**n**) tanto mejor será la aproximación. Si **n** es por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de $\bar{x}$ también es μ.
- El desvío estándar de la distribución de $\bar{x}$, es $\frac{\sigma}{\sqrt{n}}$. Decrece cuando aumenta el tamaño de la muestra (**n**).
- La distribución de muestreo de $\sqrt{n}\,\frac{\bar{x}-\mu}{\sigma}$, es **aproximadamente Normal Estándar** siempre que se utilicen tamaños de muestra suficientemente grandes

El TCL puede enunciarse, bajo los mismos supuestos, para la suma:

- La distribución de muestreo de $\sum_{i=1}^{n} x_i$, es **aproximadamente Normal** siempre que se utilicen tamaños de muestra suficientemente grandes.
- Cuanto mayor sea el tamaño de la muestra (**n**) tanto mejor será la aproximación. Si **n** es por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de $\sum_{i=1}^{n} x_i$ también es n μ.
- El desvío estándar de la distribución de $\sum_{i=1}^{n} x_i$, es $\sigma\sqrt{n}$.

Observación. Como $\sum_{i=1}^{n} x_i = n\bar{x}$, los valores de muestreo de $\sum_{i=1}^{n} x_i$ y $\bar{x}$ difieren únicamente en un factor multiplicativo (**n**). Por lo tanto, la forma de sus distribuciones es la misma; aproximadamente Normal aunque con diferente media y diferente desvío estándar.

Frecuentemente no se conoce σ, entonces se lo estima por **s** el desvío estándar de la muestra (sección 18.2.3):

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

El desvío estándar estimado de la distribución de $\bar{x}$ utilizando **s** se denomina error estándar.

- Se denomina **error estándar** a $\frac{s}{\sqrt{n}}$ y se lo utiliza para estimar $\frac{\sigma}{\sqrt{n}}$.

El TCL, además de mostrar cómo la media muestral da una estimación más precisa cuanto más grande sea la muestra de la cual proviene, con un desvío estándar de $\frac{\sigma}{\sqrt{n}}$ describe que esas medias muestrales se distribuyen en forma **aproximadamente Normal,** cualquiera sea la forma de la distribución de los datos individuales.

Generalmente, se considera que **n=30** es un tamaño de muestra adecuado para que la distribución de $\bar{x}$ pueda aproximarse por la Normal. En realidad, ese tamaño puede ser bastante menor si la distribución de los datos individuales es simétrica, y debe ser mayor si esa la distribución es muy diferente de la Normal.

Si los datos provienen de una variable que tiene distribución Normal en la población, $\bar{x}$ tendrá una **distribución Normal, cualquiera sea n.**

Si los datos provienen de una variable que **no tiene** distribución Normal en la población, $\bar{x}$ tendrá distribución a**proximadamente Normal** siempre que **n sea grande.**

Siguiendo con el ejemplo de las garrafas, supongamos que los volúmenes, de **todas las posibles garrafas producidas de la misma forma,** se distribuyen de forma Normal con media **μ=47** dm^3 y desvío **σ=0,75** dm^3. La figura 23.3 muestra la distribución de $\bar{x}$ para una muestra de tamaño 4 junto con la distribución de las observaciones individuales. El desvío estándar de la distribución de $\bar{x}$ con **n=4** es la mitad que el desvío estándar de distribución de las observaciones individuales.

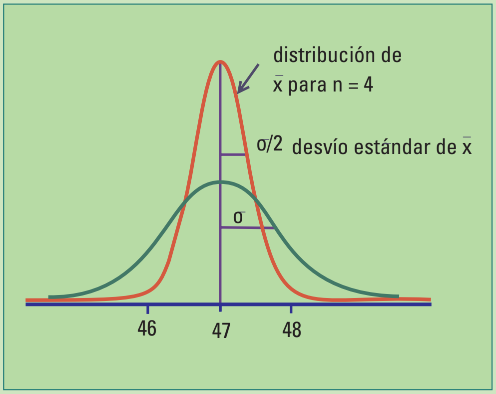


Figura 23.3. *Distribución de $\bar{x}$ para un muestreo aleatorio simple con n=4, junto con la distribución de las observaciones individuales.*

Ejemplo 2: La distribución de la longitud de las palabras puede distinguir idiomas, estilos de escritura e incluso hasta autores. La figura 23.4 muestra la distribución presentada en la resolución del ejercicio 4 de la sección 17.3. Supondremos que esta es efectivamente la distribución de la longitud de las palabras de todo un libro. Esas longitudes tienen media **μ=5,51** letras por palabra y desvío estándar **σ=3,56**, valores que en lo que sigue, se consideran parámetros poblacionales. A partir de ahora, pensamos en una "población de palabras" con media **μ=5,51** letras por palabra y desvío estándar σ=**3,56.**

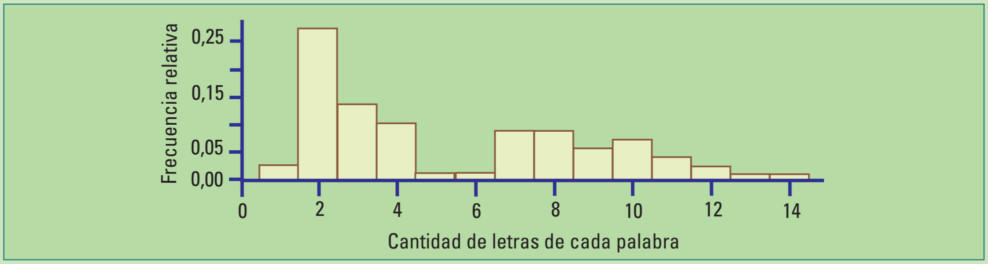


Figura 23.4. *Distribución de la cantidad de letras por palabra de un libro en castellano.*

Supongamos que elegimos al azar 1.000 muestras de 3 palabras de ese libro, registramos la cantidad de letras que tiene cada una de las palabras elegidas y calculamos la media de cada una de las 1.000 muestras. Tendremos así 1.000 números cuyo histograma mostrará su distribución. Repetimos el mismo procedimiento anterior pero con muestras de tamaño 30. Las figuras siguientes muestran los resultados:

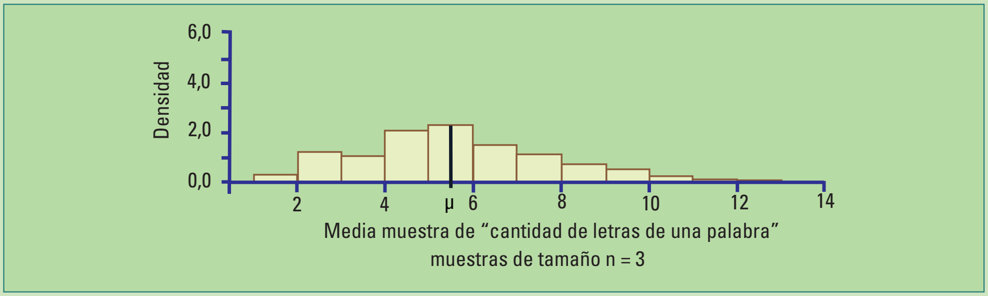


Figura 23.5. *Distribución de la media muestral para tamaño de muestra 3.*

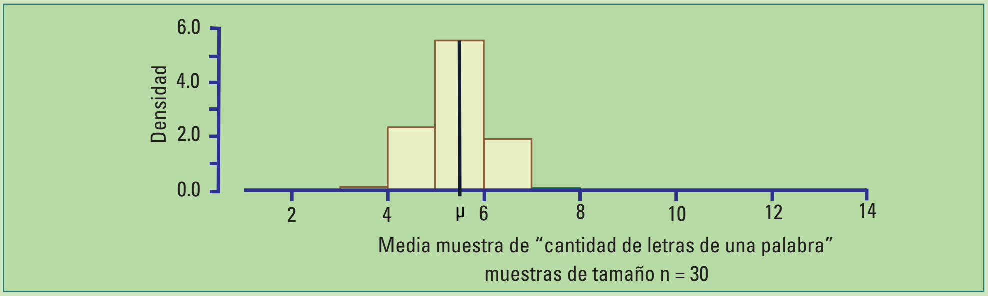
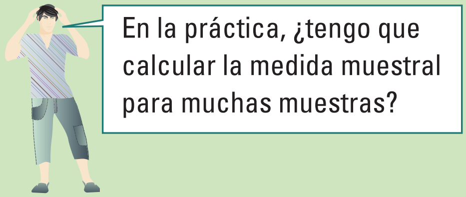
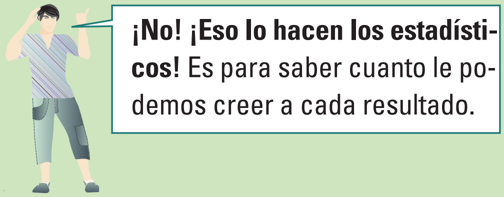


Figura 23.6. *Distribución de la media muestral para tamaño de muestra 30.*

Los valores de las medias muestrales son más simétricos y están más concentrados alrededor de la media poblacional (**μ=5,51**) cuando se toman muestras de tamaño 30, que cuando solamente las muestras tienen 3 palabras cada una (figuras 23.5 y 23.6).





En los ejemplos 1 y 2 se generaron muchas medias muestrales para conocer cómo se distribuyen.

En la práctica, la media muestral se calcula una única vez y no conocemos la distribución poblacional.

Entonces, ¿para qué sirve lo anterior?

Sirve para:

- Ver que las medias muestrales obtenidas con más datos, en general estarán más cerca de la media poblacional aunque no la conozcamos.
- Conocer la distribución y determinar un rango de valores donde se pueden encontrar la mayoría de las medias muestrales, **aunque no conozcamos la distribución poblacional** de la que provienen los datos.

□ 23.3. Distribución de muestreo de la proporción muestral

El TCL no se aplica sólo a la media muestral de una variable numérica (ejemplos 1 y 2), también se utiliza para la proporción muestral $\hat{\mathbf{p}}$ (se lee p-sombrero). Se trata, en este caso, de un análisis cualitativo observando la presencia o ausencia de una característica para estimar su proporción p en la población. La población se divide en dos partes, los que tienen y los que no tienen la característica en cuestión.

Vimos esta situación con el ejemplo del "Club Grande de Fútbol":

- En el capítulo 9 la investigadora tomó una **única muestra** de 538 socios, para cada uno se observó si estaba a favor (presencia) o no estaba a favor (ausencia) del candidato 1; obteniendo $\hat{\mathbf{p}}$ **=0,51.**
- En el capítulo 10 se presentaron los resultados de seleccionar **1.000 muestras** aleatorias simples, cada una de **tamaño 538,** de una población para la cual la verdadera proporción es **p=0,5.**

La **distribución de muestreo** de $\hat{\mathbf{p}}$ es la distribución de **todas** las **proporciones muestrales** posibles, calculadas a partir de muestras aleatorias simples del mismo tamaño **n.**

Diremos: **"distribución de $\hat{p}$"**, sobreentendiendo que se trata de la **distribución de la proporción muestral en un muestreo aleatorio simple.**

Para cualquier población que tiene una proporción **p** de individuos con una característica (éxito) y

$$\hat{p} = \frac{\text{cantidad de éxitos en la muestra}}{\text{tamaño de la muestra}}$$

es la proporción de éxitos en una muestra aleatoria simple de **n** individuos, entonces el TCL asegura que:

- La **media** de la distribución de $\hat{p}$ es **p**, ($\mu_{\hat{p}} = p$).
- El **desvío estándar** de la distribución de $\hat{p}$ es $\sqrt{\frac{p(1-p)}{n}}$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- La distribución de $\hat{p}$ es **aproximadamente Normal,** siempre que tanto n p y n (1 - p) sean grandes (mayores a 10) y la muestra sea una muestra aleatoria simple.
- La distribución de muestreo de $\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}}$, es **aproximadamente Normal Estándar** siempre
- que, **n p** y **n (1-p)** sean grandes (mayores a 10) y la muestra sea una muestra aleatoria simple.

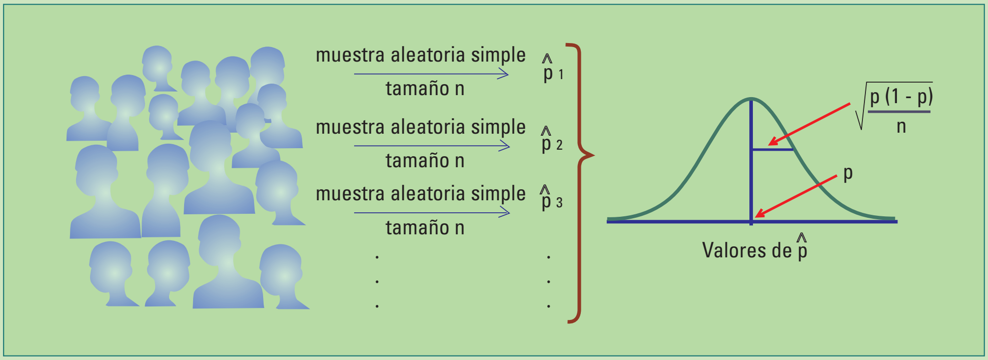


Figura 23.7. *Repetición del muestreo aleatorio simple de tamaño n. Los valores de $\hat{p}$ tienen distribución Normal con media p y desvío estándar* $= \sqrt{\frac{p(1-p)}{n}}$

Como **p** es desconocida, una estimación razonable es utilizar $\hat{\mathbf{p}}$ para estimar el desvío estándar, esta estimación se denomina error estándar:

$$\textbf{Error estándar} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

☐ 23.4. Corrección por tamaño de población

Hemos realizado muestreos aleatorios simples sin tener en cuenta el tamaño de la población (**N**), esto es válido siempre que el **tamaño de la muestra** (**n**) sea sólo **una pequeña parte de la población.** Ya utilizamos esta propiedad al introducir el concepto de margen de error en la sección 10.2.

Cuando **la muestra no es una pequeña parte** de la **población,**

- la media de la distribución de muestreo de $\hat{p}$ ($\mu_{\hat{p}}$) sigue siendo p, pero
-
- el desvío estándar de la distribución de muestreo ($\sigma_{\hat{p}}$) es $\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$

El **factor de corrección** $\sqrt{\frac{N-n}{N-1}}$ que multiplica a $\sqrt{\frac{p(1-p)}{n}}$ (el desvío estándar de $\hat{p}$ dado en la sección 23.3) para corregir el desvío estándar **es el mismo** que debe utilizarse para el **error estándar** y el **margen de error.**

En el ejemplo del Club Grande de Fútbol, **N=58.210** y **n=538**, **no es necesario utilizar el factor de corrección** ya que es muy cercano a 1:

$$\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{58.210-538}{58.210-2}} = 0{,}996$$

Distribución de muestreo de $\hat{P}$ para n=3 y N=5. TABLA 23.2

		Proporción de alumnos que no estudió
Muestra 1	ABC	$\hat{p}_1 = 0$
Muestra 2	ABD	$\hat{p}_2 = 1/3$
Muestra 3	ABE	$\hat{p}_3 = 1/3$
Muestra 4	ACD	$\hat{p}_4 = 1/3$
Muestra 5	ACE	$\hat{p}_5 = 1/3$
Muestra 6	ADE	$\hat{p}_6 = 2/3$
Muestra 7	BCD	$\hat{p}_7 = 1/3$
Muestra 8	BCE	$\hat{p}_8 = 1/3$
Muestra 9	BDE	$\hat{p}_9 = 2/3$
Muestra 10	CDE	$\hat{p}_{10} = 2/3$

El siguiente es un ejemplo hipotético. El tamaño de la muestra es más de la mitad de la población. La población es conocida y pequeña para que los cálculos sean razonablemente cortos. **No vale la aproximación de la distribución de muestreo** de $\hat{p}$ por la Normal, pero sí las conclusiones respecto de su media y su desvío estándar.

Ejemplo 3: Una población está formada por cinco alumnos (**N=5**) A, B, C, D, E, que van a rendir examen de estadística, en la que 2 de ellos (**D** y **E**) no estudiaron. La proporción de alumnos que no estudió es **p=2/5=0,4.** Se eligen 3 alumnos al azar para saber qué proporción de alumnos estudió.

La tabla 23.2 muestra la lista de las **10 muestras posibles** de tamaño **n=3** de la población y la proporción muestral de alumnos que no estudió.

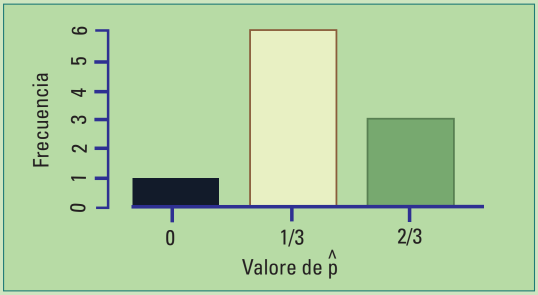


Figura 23.8. Distribución de muestreo de $\hat{p}$, datos de la tabla 23.2.

Utilizamos los valores de la tabla 23.2 para calcular la **media** de la distribución de muestreo de $\hat{\mathbf{p}}$ (promedio de **todos** los **valores posibles** de $\hat{\mathbf{p}}$):

$$\mu_{\hat{p}} = \frac{\sum_{i=1}^{10} \hat{p}_i}{10}$$

$$\mu_{\hat{p}} = \frac{0+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+\frac{2}{3}}{10}$$

$$\mu_{\hat{p}} = \frac{6\times\frac{1}{3}+3\times\frac{2}{3}}{10}$$

$$= \frac{4}{10}$$

$$= 0{,}4$$

$\mu_{\hat{p}}$ es un parámetro poblacional (de la distribución de muestreo de $\hat{\mathbf{p}}$) y coincide con la proporción poblacional p=0,4.

Ahora, calculamos el desvío estándar de los valores de $\hat{\mathbf{p}}$, $\sigma_{\hat{p}}$.

Calculemos el desvío estándar poblacional (es un parámetro) para los valores de $\hat{\mathbf{p}}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i-\mu_{\hat{p}})^2}$$

$$= \sqrt{\frac{1}{10}\sum_{i=1}^{10}(\hat{p}_i-0{,}4)^2}$$

$$= 0{,}2$$

No debe preocupar el cálculo anterior, las calculadoras lo hacen bien.

Llama la atención que $\sigma_{\hat{p}}$ **no es** $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}4\ (1-0{,}4)}{3}} = 0{,}2828$ (el desvío es tándar de $\hat{\mathbf{p}}$ dado en la sección 23.3), sino un número más pequeño (0,2) que puede obtenerse mediante el siguiente cálculo:

$$\sigma_{\hat{p}} = \sqrt{\frac{N-n}{N-1}}\ \sqrt{\frac{p\,(1-p)}{n}}$$

$$= \sqrt{\frac{5-3}{5-1}}\ \sqrt{\frac{0{,}4\,(1-0{,}4)}{3}}$$

$$= 0{,}2$$

$\sqrt{\frac{5-3}{5-1}} = 0{,}7071$ es el factor de corrección en este ejemplo.

En general:

El factor de corrección es $\sqrt{\frac{N-n}{N-1}}$

Lo anterior es para proporciones, pero ¿qué ocurre con la media muestral cuando la muestra no es una pequeña parte de la población?

- La media de la distribución de $\bar{x}$ ($\mu_{\bar{X}}$) sigue siendo $\boldsymbol{\mu}$, pero
-
- El desvío estándar de la distribución de $\bar{x}$ ($\sigma_{\bar{X}}$) es $\sqrt{\frac{N-n}{N-1}} \; \frac{\sigma}{\sqrt{n}}$

> **El mismo factor de corrección** se utiliza corregir el desvío estándar de la distribución de $\hat{p}$ y de $\bar{x}$.

El factor de corrección es:

- Menor o igual a 1, siempre.
- Cercano a 1, si el tamaño de la población (**N**) es muy grande comparado con el tamaño de la muestra (**n**) y no es necesario utilizarlo.
- Cercano a 0, si el tamaño de la muestra se acerca al tamaño de la población, (**n≈ N**).
- Cero, si **n = N.**

Por lo tanto, como era de esperar:

> **Si la muestra es toda la población,** el **desvío estándar** de la distribución de $\bar{p}$ y de $\bar{x}$, y **el margen de error valen** 0.

□ 23.5. El TCL y el mundo real

El Teorema Central del Límite sirve para interpretar los resultados de una encuesta o un estudio.

Permite aplicar, tanto para la **media muestral** como para la **proporción muestral,** las propiedades que vimos (sección 20.2.3) sobre datos que se distribuyen de acuerdo con la curva Normal; siempre que el tamaño de la muestra sea suficientemente grande. Por supuesto, las observaciones deben ser mediciones válidas y no tener sesgo.

Por ejemplo, cuando se trata de medias muestrales cuya distribución tienen media $\boldsymbol{\mu}$ y desvío estándar $\frac{\sigma}{\sqrt{n}}$, permite afirmar que:

- Cerca del 95% de las medias muestrales están entre $\boldsymbol{\mu}$ menos 2 veces su desvío estándar, y μ más 2 veces su desvío estándar, o sea dentro de $\boldsymbol{\mu \pm 2} \frac{\sigma}{\sqrt{n}}$.

- El 99,7% (casi todas) de las medias muestrales se encuentran en el intervalo $(\boldsymbol{\mu-3}\frac{\sigma}{\sqrt{n}} \, , \boldsymbol{\mu+3}\frac{\sigma}{\sqrt{n}})$, o sea dentro de $\boldsymbol{\mu\pm3}\frac{\sigma}{\sqrt{n}}$.

Ya hemos utilizado este último resultado al establecer los límites de control en un gráfico equis barra (sección 21.2.2):

- Límite inferior de control (LIC) = $\mu - 3\frac{\sigma}{\sqrt{n}}$
- Límite superior de control (LSC) = $\mu + 3\frac{\sigma}{\sqrt{n}}$

De esta manera si el proceso está en control, solamente 0,3 % de las veces una media muestral excederá los límites dando una alarma falsa.

El Teorema Central del Límite se utiliza también como argumento para tratar de explicar porqué frecuentemente los errores de medición, y muchas variables relacionadas con fenómenos naturales, tienen distribuciones que pueden aproximarse por curvas de Gauss.

Por ejemplo, podemos pensar el error de medición como una suma de pequeños errores, luego por el TCL, es razonable que estos errores tengan distribución Normal. La idea puede aplicarse también a cualquier variable que se pueda considerar como una suma de pequeñas contribuciones independientes (no sirve sumar siempre el mismo número). Generalizaciones del teorema central del límite a promedios pesados y a sumandos parcialmente independientes muestran cuándo se obtiene una buena aproximación a la Normal y cuando no.

Desde que Abraham de Moivre postuló la primera versión del TCL en 1733, muchos matemáticos célebres -Pierre Simon, marqués de Laplace (1812), Siméon Denis Poisson (1824), Pafnuty Tchebyshev (1887), Aleksandr Lyapunov (1901), Jarl Waldemar Lindeberg (1922), para mencionar algunos de los más famosos- obtuvieron resultados respecto a la aproximación por la Normal a la distribución de medias o sumas. Más que un único teorema, se trata de muchos dependiendo de las condiciones que se requieren para su validez. Estas generalizaciones siguen siendo aún temas de investigación en estadística. Las demostraciones utilizan herramientas matemáticas avanzadas y resultados de la teoría de probabilidad.

Utilizaremos los resultados del TCL en los capítulos 24 y 25 en la construcción de dos nuevas herramientas estadísticas.

23.6. Actividades y ejercicios

1. Indique cuáles de las siguientes afirmaciones son verdaderas y justifique brevemente.

   a) Cuanto mayor sea el tamaño de la muestra mayor será el desvío estándar de la distribución de muestreo de $\bar{\mathbf{x}}$.
   b) El desvío estándar de la distribución de muestreo de $\bar{\mathbf{x}}$ sólo depende del tamaño de la muestra.
   c) La distribución de muestreo de $\bar{\mathbf{x}}$ es Normal si la población tiene una distribución Normal.
   d) Cuando n es grande la distribución de muestreo de $\bar{\mathbf{x}}$ es aproximadamente Normal, aún cuando la población no tenga una distribución Normal.

2. Indique cuales de las siguientes afirmaciones son verdaderas y justifique brevemente.

   a) La distribución de muestreo de $\hat{\mathbf{p}}$ tiene media igual a la proporción poblacional **p**.
   b) La distribución de muestreo de $\hat{\mathbf{p}}$ tiene un desvío estándar igual a $\sqrt{np(1-p)}$.
   c) Se considera que la distribución de muestreo de $\hat{\mathbf{p}}$ es aproximadamente Normal cuando $\mathbf{n \geq 30}$.

3. Realice los siguientes experimentos, puede pedir ayuda a sus compañeros y realizar una repetición cada uno:

   a) Arroje un dado 3 veces, si sale 1 ó 2 registre éxito, en caso contrario registre fracaso. Cuente la cantidad de éxitos. Repita 60 veces.
   b) Arroje un dado 30 veces, si sale 1 ó 2 registre éxito, en caso contrario registre fracaso.

   Para cada uno de los experimentos anteriores construya la tabla siguiente:

Número de repetición	Cantidad de éxitos	Proporción de éxitos
1		
2		
3		
.		
.		
60		

   Halle el histograma de las proporciones de éxitos obtenidas. Compare.

   En la práctica, la distribución de los valores de una variable en la poblción no cambia si se quita de ella una proporción muy pequeña. Los experimentos anteriores representan los resultados de muestreos aleatorios simples con reposición (o sea, la población no cambia con cada extracción) para estimar la proporción de éxitos (**p**).

En el ejemplo del dado, se registra éxito cuando al arrojarlo sale 1 ó 2. Si el dado está equilibrado, 2 resultados entre 6 posibles dan éxito en una proporción **p=1/3**. Los resultados de arrojar el dado representan muestreos aleatorios simples con reposición de una población cuya proporción de "éxitos" es 1/3. En general, no conocemos la verdadera proporción de "éxitos" en una población real.

4. Interesa saber cómo se comporta la media y el desvío de la distribución de muestreo de $\bar{\mathbf{x}}$ cuando el tamaño de la muestra es una parte importante del tamaño de la población.

   Considere una población de tamaño **N=6** y una variable en la población toma los valores 2 4 6 9 12 18.

   a) Calcule la media y el desvío poblacional de la variable ($\boldsymbol{\mu}$ y σ).
   b) Enumere los valores de la variable de todas las posibles muestras de tamaño **n=2** y calcule $\bar{\mathbf{x}}$ para cada una de ellas.
   c) Muestre que la media de las 15 medias muestrales ($\bar{\mathbf{x}}$) es $\boldsymbol{\mu}$**.**
   d) Muestre que el desvío estándar de las 15 medias muestrales es:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

   e) ¿Qué puede decirse, en general, sobre el valor anterior cuando el tamaño de la población **N** es muy grande en comparación con el tamaño muestral **n**?