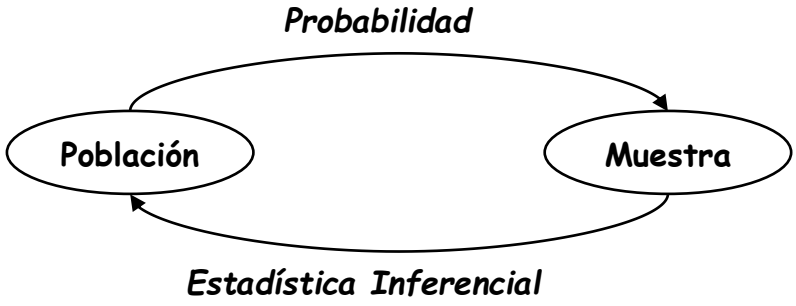
4: DISTRIBUCIONES FUNDAMENTALES DEL MUESTREO

Este es un capítulo de mucha importancia para entender las aplicaciones en los capítulos siguientes. ¡Léalo detenidamente!

Introducción

Hasta ahora se trabajado con variables aleatorias teniendo por conocidos sus parámetros. Así, en Ingeniería, eso es posible cuando se trabaja con procesos productivos controlados. Por ejemplo, el error en el diámetro de engranajes producidos bajo especificaciones se modela adecuadamente con una distribución normal con parámetros μ y σ^2 conocidos, lo cual equivale a identificar a la variable aleatoria con la población. El procedimiento seguido para analizar la probabilidad de ocurrencia de las características poblacionales en la muestra es un camino deductivo (se va de lo general, la población, a lo particular, la muestra). Esto es exactamente lo que se realiza en un control de calidad, por ejemplo.
Pero ahora se invertirá el camino, es decir, a partir de una muestra se tratará de conocer las características de la población (los parámetros), justamente cuando no es posible, por motivos de tiempo o económicos, relevar a todos los individuos de ésta. Esto es, en esencia, un camino inductivo o **inferencial** (se va de lo particular, la muestra, a lo general, la población) y es este razonamiento el que distinguirá al resto de los conceptos por venir.

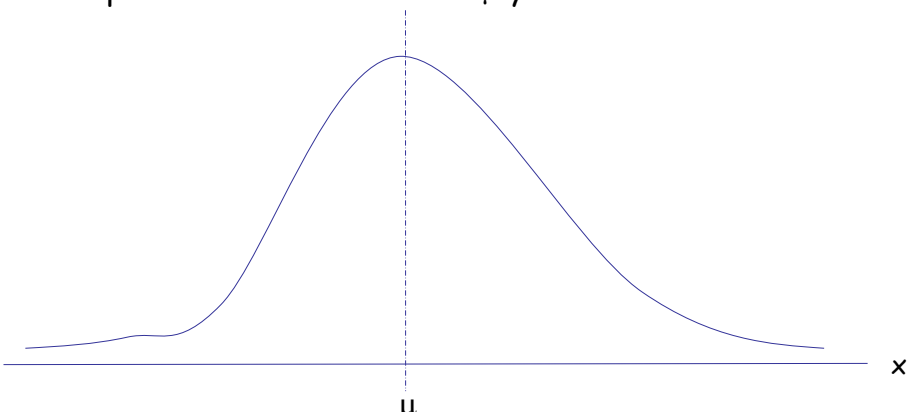


Como ejemplo, si se desea conocer la resistencia media a la tracción del hormigón producido por una empresa proveedora, se analizará la resistencia promedio en una muestra con la esperanza de poder inferir alguna conclusión con respecto al

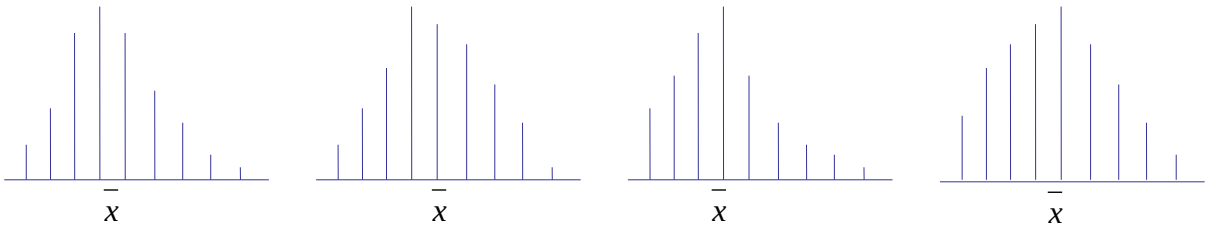
parámetro poblacional, es decir, la verdadera resistencia promedio a la tracción del hormigón producido por la empresa.
En todo este proceso será fundamental definir precisamente los conceptos de población, muestra y distribución de muestreo. Primero se abordará el tema de manera intuitiva y gráfica, pues es necesario "ver" las distribuciones muestrales (o de muestreo) antes de definirlas y desarrollarlas formalmente.

Base conceptual para muestrear poblaciones

Suponga que una población está constituida por todos los filtros de un gran sistema industrial de control de contaminación y que la variable en estudio es la cantidad de horas de operación antes de que un filtro quede obstruido. La distribución de las horas de operación tiene una media μ y una desviación estándar σ.



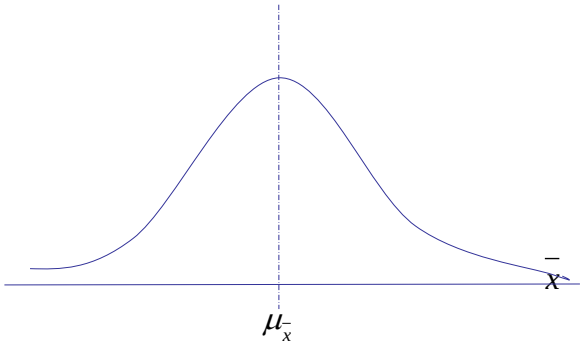
Supóngase que pueden tomarse todas las muestras posibles de tamaño 9 de filtros de la población. A continuación se calcularía la media y la desviación estándar en cada una de las muestras[1].
Como resultado, cada muestra tendría su propia media $\bar{x}$ y su propia desviación estándar s tal como puede verse a continuación en la gráfica para algunas de esas muestras.



Ninguna de las medias individuales sería la misma que la poblacional (o, más concretamente, la probabilidad de serlo se considera prácticamente nula). Éstas tenderían a estar cerca de la media poblacional, pero rara vez coincidirían exactamente con este valor.

[1] Note que a pesar de tratarse de una distribución continua para la población, los datos muestrales se comportan como discretos para un tamaño de muestra n.

Si la media en la muestra es una función de la misma muestra y su resultado puntual depende enteramente del azar, es acertado pensar que puede tratarse como una variable aleatoria. Y si tal es el caso, entonces es coherente pensar que, como tal, tendrá una distribución de probabilidad para todos sus posibles valores. Así se podría elaborar una distribución de todas las medias de cada muestra que se puedan tomar.



Esta distribución de las medias de la muestra es conocida como distribución teórica de muestreo o distribución muestral de la media y, como es lógico pensar, tiene su propia media $\mu_{\bar{x}}$ y su propia desviación estándar $\sigma_{\bar{x}}$ (conocida como error estándar de la media).[2]

En el caso puntual descripto, tendríamos la distribución para las horas promedio de uso de los filtros antes de una obstrucción.

Ya se tiene, entonces, una idea bastante aproximada de lo que es una distribución muestral. En este caso se ha hecho una ejemplificación para la media, pero el razonamiento es extensivo para la varianza, para proporciones, diferencias de medias, etcétera.

Por último, no se ha hecho hincapié, aún, en la forma específica que tiene la distribución de la población, ni mucho menos en la muestral. Éste será un problema a tratar más formalmente.

Se procederá, ahora, a ampliar estos conceptos de una manera formal.

Población y muestra

Definición 4.1: Se llama **población** al conjunto total de elementos en discusión y sobre los cuales se quiere tener alguna información.

El problema de la inferencia inductiva en Estadística se podría plantear de la siguiente manera:

Se tiene una población de la cual se quiere tener alguna información. Como se dijo antes, a veces es imposible o poco práctico, observar toda la población, entonces

[2] Observe los símbolos empleados y entienda la lógica empleada en su construcción, como forma de recordar su significado.

se toma parte de ella (**muestra**) y después de analizar esta parte se infieren los resultados a la población total.[3]

Como la inferencia estadística se formula con base en una muestra de objetos de la población de interés, el proceso por medio del cual se obtiene será aquél que asegure la selección de una "buena" muestra.[4] Una manera de obtenerla es cuando el proceso de muestreo proporciona, a cada objeto en la población, una oportunidad igual e independiente de ser incluido en la muestra. Este concepto conduce a lo que se conoce como **muestra aleatoria**.

Si la población consiste en *N* objetos y de éstos se seleccionan *n*, el proceso de muestreo debe asegurar que cada muestra de tamaño *n* tenga la misma probabilidad de ser seleccionada. Para esto se deben elegir los *n* elementos con reposición o bien, considerar a la población infinita (es decir *N* suficientemente grande) y utilizar una selección sin reemplazo. Si cada uno de los *n* valores proviene del resultado de un experimento, como por ejemplo arrojar una moneda y observar la cara que muestra, este experimento debe repetirse *n* veces bajo las mismas condiciones para asegurar la independencia en los valores obtenidos. En esta unidad se considerará que la población bajo estudio es infinita.

En la práctica se obtienen valores numéricos $x_1, x_2, \ldots, x_n$, que en el concepto de muestra aleatoria deben ser considerados como los valores observados, o realizaciones, de *n* variables aleatorias $X_1, X_2, \ldots X_n$ con la misma distribución de la población. Para que la muestra sea aleatoria el valor observado de cada X_i debe ser independiente de los valores observados de las otras variables aleatorias. Es decir $X_1, X_2, \ldots X_n$ deben ser estocásticamente independientes. El conjunto de los valores observados de dichas variables constituye la muestra, se indica $x_1, x_2, \ldots, x_n$.

Definamos formalmente el concepto de muestra aleatoria.

Definición 4.2: Una **muestra aleatoria** de tamaño "n" de una población con función (densidad) de distribución de probabilidad *f* es un conjunto de "n" variables aleatorias independientes y cada una con idéntica distribución de la población (IID).

Simbólicamente se indica:

$$X_1, X_2, \ldots X_n \overset{IID.}{\sim} f_X(\underline{x}; \theta)$$

donde el símbolo θ indica el o los parámetros poblacionales y $\underline{x} = x_1, x_2, \ldots, x_n$ son las n observaciones de las X_i variables aleatorias. De acuerdo con las propiedades

[3] ¡Imagínese probando una población de 1. 000.000 de circuitos hasta que fallen antes de comercializarlos! Mejor es tomar algunos de ellos, observar la proporción que falla y luego inferir este resultado al total circuitos. Evidentemente este resultado no será nunca "exacto" pero puede resultar interesante si se lo relaciona con el concepto de probabilidad. Es decir, si se puede establecer una cierta confianza en nuestra inferencia.

[4] Los métodos de muestreo se describen al final de este capítulo.

de las distribuciones de probabilidad conjunta, la función (densidad) de probabilidad conjunta para n variables aleatorias independientes viene dada por:

$$f(x_1, x_2, ..., x_n; \theta) = g(X_1) \cdot h(X_2) \cdot ... \cdot k(X_n)$$

Estadísticos y parámetros

En los comentarios introductorios se mencionó de manera breve que las características muestrales se emplean para realizar inferencias con respecto a las características de la población. A las primeras se las denomina "estadísticas" o "estadísticos", mientras que las segundas reciben el nombre de "parámetros". Así, para estudiar las características de una población cuyo parámetro θ es desconocido, se evaluará una realización particular $\hat{\theta}$ del estadístico muestral $\hat{\Theta}$.
El objetivo de esta sección será el de examinar con detalle el papel que desempeñan las estadísticas en relación con la inferencia. En particular, se desarrollará la noción de una distribución de muestreo de una estadística, que es uno de los conceptos más importantes en inferencia estadística.
Para colocar a las estadísticas en una mejor perspectiva se debe definir y analizar, de manera formal, un parámetro de población.

Definición 4.3: Un **parámetro** es una caracterización numérica de la distribución de la población de manera que describe, parcial o completamente, la función de densidad de probabilidad de la característica de interés. La oración "describe de manera completa" sugiere que una vez que se conoce el valor de θ entonces puede formularse cualquier proposición probabilística de interés.

Definición 4.4: Un **estadístico** o **estadística** es cualquier función de las variables aleatorias que se observaron en la muestra de manera que esta función no contiene cantidades desconocidas.

Considérese la muestra $X_1, X_2, ...X_n \overset{IID.}{\sim} f_X(\underline{x}; \theta)$ (que como se vio consiste de *n* variables aleatorias (IID) con una función de densidad de probabilidad *f(x;* θ) que depende de un parámetro desconocido θ).
Supóngase que se definen funciones como:

$$\hat{\Theta}_1 = X_1 + X_2 + n$$
$$\hat{\Theta}_2 = X_1^2 + \ln X_2$$
$$\hat{\Theta}_3 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

y como tantas otras que se pueden definir (¡Preste especial atención al último estadístico dado en la lista!).
Todos ellos son estadísticos porque se determinan de manera completa por las variables aleatorias que contiene la muestra. De manera general, denótese una

estadística por $\hat{\Theta} = u(X)$. Dado que $\hat{\Theta}$ es una función de variables aleatorias, es en sí misma una variable aleatoria y su valor específico $\hat{\theta} = u(X)$ puede determinarse cuando se conozcan las realizaciones x de X.

Si se emplea una estadística $\hat{\Theta}$ para estimar un parámetro desconocido θ, $\hat{\Theta}$ recibe el nombre de *estimador* de Θ y el valor específico de $\hat{\theta}$, como un resultado de los datos muestrales, recibe el nombre de *estimación puntual* de Θ.

Esto es, un *estimador* es una *estadística* que identifica al mecanismo funcional por medio del cual, una vez que las observaciones en la muestra se realizan, se obtiene una *estimación*.

Algunos ejemplos de parámetros, estadísticos y estimaciones puntuales son:

θ	$\hat{\Theta}$	$\hat{\theta}$
μ	$\overline{X}$	$\overline{x}$
σ^2	S^2	s^2
p	$\hat{P}$	$\hat{p}$

Una estadística es sustancialmente diferente de un parámetro. Un parámetro es una constante y una estadística es una variable aleatoria. Además, dado un valor del parámetro se describe de manera completa un modelo de probabilidad (suponiendo una distribución uniparamétrica); ningún valor de una estadística puede desempeñar tal papel si cada uno de éstos depende del valor de las observaciones de las muestras. Y dado que las muestras se toman en forma aleatoria, ninguna muestra es más válida que cualquier otra que se haya tomado con el mismo fin.

Definición 4.5: La distribución de muestreo de una estadística $\hat{\Theta}$ es la distribución de probabilidad de $\hat{\Theta}$ que puede obtenerse como resultado de un número infinito de muestras aleatorias independientes, cada una de tamaño n, provenientes de la población de interés.

Dado que se supone que las muestras son aleatorias, la distribución de una estadística es un tipo de modelo de probabilidad conjunta para variables aleatorias independientes, en donde cada variable posee una función de densidad de probabilidad igual a la de las demás. De manera general, la distribución de muestreo de una estadística no tiene la misma forma que la función de densidad de probabilidad en la distribución de la población.

Estadísticos particulares y sus distribuciones muestrales

La primera distribución muestral importante a considerar es la de $\overline{X}$

1. Media muestral

Si $X_1, X_2, \ldots X_n$ representan una muestra aleatoria de tamaño n, entonces la media de la muestra se define mediante la estadística:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ahora suponga que se toma una muestra aleatoria de n observaciones de una población normal con media μ y varianza σ^2. Cada observación $X_i\,(i = 1,2,\ldots,n)$ tendrá entonces la misma distribución normal que la población que se muestrea. Entonces, por la propiedad reproductiva de la distribución normal [5] se concluye que:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

tiene **distribución normal** con media:

$$\mu_{\overline{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \frac{n\mu}{n} = \mu$$

y varianza:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

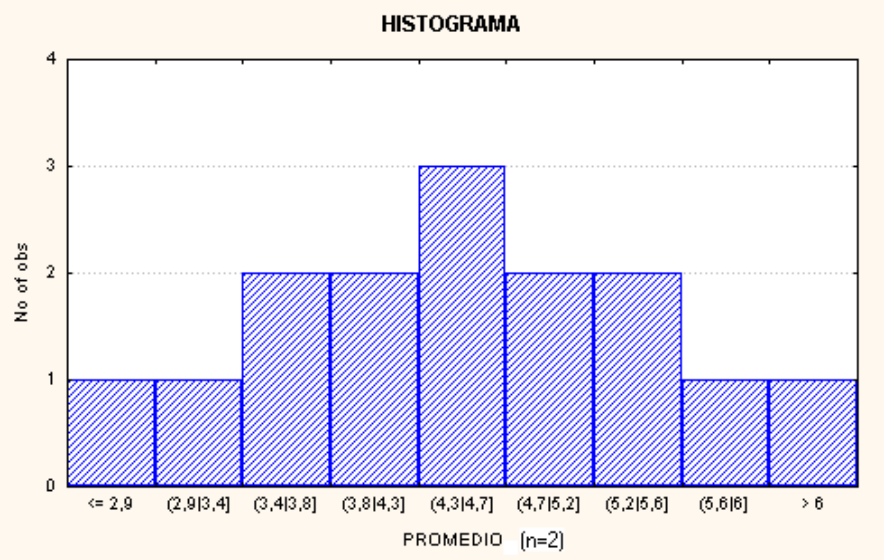
Ejemplo 1:

Se determinará la distribución de $\overline{X}$ a partir de muestras de tamaño 2 tomadas de los precios de un cierto producto. Para determinar $f_{\overline{X}}(\overline{x})$ se tiene en cuenta las muestras obtenidas y en cada una de ellas determinamos el valor observado de $\overline{X}$.

Muestra	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
x_i	2	2	2	2	2	3	3	3	3	4	4	4	5	5	6
x_i	3	4	5	6	7	4	5	6	7	5	6	7	6	7	7
$\overline{x}$	2,5	3	3,5	4	4,5	3,5	4	4,5	5	4,5	5	5,5	5,5	6	6,5

El histograma correspondiente es:

[5] Recordar que si $X_1, X_2, \ldots X_n$ son variables aleatorias, cada una con distribución normal con media y varianza μ_i y σ_i^2 con (i=1,2,…,n), respectivamente, entonces la variable aleatoria $Y = a_1X_1 + a_2X_2 + \ldots a_nX_n$ tiene distribución normal con media $\mu_Y = a_1\mu_1 + a_2\mu_2 + \ldots a_n\mu_n$ y varianza $\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots a_n^2\sigma_n^2$.



Puede verse que los valores de $\bar{x}$ se concentran alrededor de $\mu = 4.5$; el problema es que se observa una gran variabilidad, Esto hace que los valores observados de $\bar{X}$ en muestras de tamaño 2 no presenten un buen comportamiento para darnos información respecto del parámetro desconocido μ de la población bajo estudio, Si tomáramos muestras más grandes, la distribución $\bar{X}_n$ tendría mejores características, Esto puede verse en el hecho de que la varianza $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ y, por lo tanto, su raíz cuadrada, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, llamada error estándar, disminuyen a medida que aumenta el tamaño de muestra, Sin embargo, la desviación estándar de la distribución muestral de $\bar{X}$ siempre es menor que la de X debido a que está multiplicada por el factor $1/\sqrt{n}$, Esto es así debido a que se promedian valores de $\bar{x}$, los cuales tienden a estar más cerca entre sí.
Este comportamiento se describe formalmente en el siguiente teorema:

TEOREMA DEL LÍMITE CENTRAL

Sea X una variable aleatoria con función densidad con media μ y varianza σ^2 finitas, si se toma una muestra aleatoria de tamaño n y se obtiene $\bar{X}$, se puede definir una nueva variable aleatoria Z como sigue

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

La distribución de Z tiende a una distribución normal estándar cuando $n \to \infty$, Es decir, $\bar{X}$ es asintóticamente normal con media μ y varianza $\frac{\sigma^2}{n}$.

El teorema del límite central se puede aplicar para una muestra aleatoria de cualquier distribución siempre que μ y σ^2 sean finitos y el tamaño de la muestra sea grande.
En general, la aproximación será buena si $n \geq 30$.
Si $n < 30$, la distribución muestral de $\overline{X}$ será normal **sólo si** la distribución de X es normal.

Ejemplo 2:
Se ha medido las alturas de cuatro personas, en centímetros, que serán nuestra "población"[6], encontrándose una altura promedio $\mu = 186{,}5$ con un desvío estándar $\sigma = 2{,}6926$, Siendo esta población de tamaño $N = 4$, podemos seleccionar 16 muestras aleatorias[7] de tamaño $n = 2$,

1	*2*	*3*	*4*	**μ**	**σ**
183	185	188	190	**186,5**	**2,6926**

Entonces:

Muestra	Observación 1	Observación 2	$\overline{x}_i$
n_1	183	183	183,0
n_2	183	185	184,0
n_3	183	188	185,5
n_4	183	190	186,5
n_5	185	183	184,0
n_6	185	185	185,0
n_7	185	188	186,5
n_8	185	190	187,5
n_9	188	183	185,5
n_{10}	188	185	186,5
n_{11}	188	188	188,0
n_{12}	188	190	189,0
n_{13}	190	183	186,5
n_{14}	190	185	187,5
n_{15}	190	188	189,0
n_{16}	190	190	190,0

así, si calculamos la media y la desviación estándar de los valores obtenidos para las medias muestrales, obtenemos $\mu_{\overline{X}} = 186{,}5$ y $\sigma_{\overline{X}} = 1{,}9039$.

Observamos que $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2{,}6926}{\sqrt{2}} = 1{,}9039$.

[6] Para una población de tamaño cuatro no sería necesario trabajar con muestras, por lo que se usará el ejemplo únicamente con propósitos didácticos.

[7] Recuerde que todas las posibles muestras de tamaño 2 serían, $C'^4_2 = 2^4 = 16$ mientras que las posibles muestras sin reemplazo serían, $C^4_2 = \dfrac{4!}{2!(4-2)!} = 6$

Vemos que habiendo partido de una distribución uniforme para X con f(x) = 1/4, obtenemos para la $\overline{X}$ una distribución simétrica con media $\mu_{\overline{X}} = \mu$ y $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, como detallamos a continuación:

$\overline{x}_i$	$f(\overline{x}_i)$
183,0	0,063
184,0	0,125
185,0	0,063
185,5	0,125
186,5	0,250
187,5	0,125
188,0	0,063
189,0	0,125
190,0	0,063

Los resultados anteriores se han obtenido suponiendo un muestreo con reemplazo o que las muestras se han extraído de una población infinita.
Muchas veces no se muestrea con reemplazo y, en muchas ocasiones se muestrea a partir de poblaciones finitas.
Así, en nuestro ejemplo, bajo un muestreo sin reemplazo, el número de muestras posibles es 6:

Muestra	Observación 1	Observación 2	$\overline{x}_i$
n_1	183	185	184,0
n_2	183	188	185,5
n_3	183	190	186,5
n_4	185	188	186,5
n_5	185	190	187,5
n_6	188	190	189,0

así, si calculamos la media y la desviación estándar de los valores obtenidos para las medias muestrales, tenemos $\mu_{\overline{X}} = 186{,}5$ y $\sigma_{\overline{X}} = 1{,}5546$. Advertimos que el valor calculado para la desviación estándar no coincide con $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2{,}6926}{\sqrt{2}} = 1{,}9039$.

En este caso, la varianza de la media muestral no es igual a la varianza poblacional dividido el tamaño de la muestra. Sin embargo, existe una relación entre éstas y está dada por $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$.

En nuestro ejemplo vemos que $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{2{,}9626}{\sqrt{2}} \cdot \sqrt{\frac{4-2}{4-1}} = 1{,}5546$.

Ejemplo 3:
Retomamos el ejemplo 1, Se determinará la distribución de $\overline{X}$ a partir de muestras de tamaño 2 tomadas de los precios de cierto producto: $2, $3, $4, $5, $6,

\$7, Para determinar $f_{\bar{X}}(\bar{x})$ se tendrá en cuenta las $C_2^6 = 15$ muestras (sin reemplazo) obtenidas y en cada una de ellas el valor observado de $\bar{X}$.

Muestras	2 - 3	2 - 4	2 - 5	2 - 6	2 - 7	3 - 4	3 - 5	3 - 6	3 - 7	4 - 5	4 - 6	4 - 7	5 - 6	5 - 7	6 - 7
$\bar{x}$	2,5	3	3,5	4	4,5	3,5	4	4,5	5	4,5	5	5,5	5,5	6	6,5

Se deduce que la distribución muestral observada del estadístico $\bar{X}$ es:

$\bar{x}$	2,5	3	3,5	4	4,5	5	5,5	6	6,5
$f_{\bar{X}}(\bar{x})$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

Recordando que $\mu_{\bar{X}} = E(\bar{X}) = \sum_{i=1}^{n} \bar{x}_i \cdot f_{\bar{x}}$, entonces $\mu_{\bar{X}} = 4{,}5$ (que coincide con la media poblacional $\mu = 4{,}5$).

Vemos que de una distribución uniforme para X: "Precio de cierto producto" con $\mu = 4{,}5$ y $\sigma = 1{,}7078$, se llega a una distribución simétrica para $\bar{X}$ con media $\mu_{\bar{X}} = 4{,}5$ y $\sigma_{\bar{X}} = 1{,}0801$.

La diferencia entre $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 1{,}2076$ y $\sigma_{\bar{X}} = 1{,}0801$ se debe al muestreo sin reemplazo.

Para acercarnos a los valores poblacionales sería necesario aplicar un factor de corrección por finitud, debido a que nuestra población es finita. Este factor está dado por $\sqrt{\frac{N-n}{N-1}}$. De esta manera $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{1{,}7078}{\sqrt{2}} \cdot \sqrt{\frac{6-2}{6-1}} = 1{,}0801$.

Ejemplo 4:

De acuerdo con la información que suministra la compañía telefónica, el pago mensual promedio de todos los abonados de la Ciudad de Mendoza es de \$153 con una desviación estándar de \$41, Se toma una muestra de tamaño 36 de esa población ¿cuál es la probabilidad de que el pago promedio sea inferior a \$140?

Como el tamaño de muestra es $n = 36$ se puede considerar que la distribución de $\bar{X}$ es aproximadamente normal, entonces

$$P(\bar{X} < 140) = P\left(Z < \frac{140 - 153}{6{,}83}\right) \cong 0{,}0287$$

ya que

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{41}{\sqrt{36}} = 6{,}83$$

2. Diferencia de medias muestrales

Sea X_1 y X_2 variables aleatorias con función de densidad con medias μ_1 y μ_2, y varianzas finitas σ_1^2 y σ_2^2, respectivamente, Si se toman muestras aleatorias independientes de tamaño n_1 y n_2 y se obtienen $\overline{X}_1$ y $\overline{X}_2$, entonces, para la variable aleatoria $\overline{X}_1 - \overline{X}_2$ se puede definir una nueva variable aleatoria Z como sigue:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

La distribución de Z tiende a una distribución normal estándar cuando $n \to \infty$. Es decir $\overline{X}_1 - \overline{X}_2$ es asintóticamente normal con media $\mu_1 - \mu_2$ y desviación estándar $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Las consideraciones para el tamaño de muestra son las mismas que para el caso de la media muestral $\overline{X}$.

3. Varianza muestral

Para empezar, será de utilidad recordar los siguientes teoremas y su corolario:

Teorema 1: Sea X una variable aleatoria con media μ y desvío estándar σ, la variable $Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$ tiene distribución ji-cuadrado con un grado de libertad

Teorema 2: Si $X_1, X_2, \ldots, X_n$ son *n* variables aleatorias independientes con distribución ji-cuadrado con $v_1, v_2, \ldots, v_n$ grados de libertad, respectivamente, entonces la variable aleatoria $Y = X_1 + X_2 + \ldots X_n$ tiene distribución ji-cuadrado con $\nu = \nu_1 + \nu_2 + \ldots \nu_n$ grados de libertad.

Corolario: sean $X_1, X_2, \ldots, X_n$ n variables aleatorias independientes con distribución normal, entonces, la variable aleatoria $Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ tiene distribución ji-cuadrado con $v = n$ grados de libertad.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con función densidad de probabiliad f, la varianza muestral S^2 se define como:

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

Se puede probar[8] que $\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2$ (1)

Como $S^2=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$, podemos decir que $\sum_{i=1}^{n}(X_i-\bar{X})^2=S^2.(n-1)$

de aquí, podemos expresar a (1) como $S^2(n-1)+n(\bar{X}-\mu)^2=\sum_{i=1}^{n}(X_i-\mu)^2$

dividiendo todos los términos por σ^2, queda $\dfrac{S^2(n-1)}{\sigma^2}+\dfrac{n(\bar{X}-\mu)^2}{\sigma^2}=\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$

y reordenando, $\dfrac{S^2(n-1)}{\sigma^2}+\left(\dfrac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}}\right)^2=\sum_{i=1}^{n}\left(\dfrac{X_i-\mu}{\sigma}\right)^2$ (2)

Por el corolario antes mencionado, el término del miembro derecho de la igualdad (2) tiene una distribución ji-cuadrado con $\nu=n$ grados de libertad y el segundo término del miembro izquierdo de la igualdad es una variable aleatoria ji-cuadrado con $\nu=1$ grados de libertad, Entonces, por el teorema 2, el primer término del miembro izquierdo es una variable aleatoria ji-cuadrado con $v=n-1$ grados de libertad, Así, podemos decir que:

> Si S^2 es la varianza muestral en una muestra aleatoria de tamaño n tomada de una población normal con varianza σ^2, entonces la estadística
>
> $$X^2=\frac{S^2(n-1)}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$$
>
> tiene distribución ji-cuadrado con $\nu=n-1$ grados de libertad.

Distribuciones del muestreo

Significado de los grados de libertad

La variable aleatoria $\sum_{i=1}^{n}\left(\dfrac{X_i-\mu}{\sigma}\right)^2$ calculada de una muestra aleatoria tomada de una población normal tiene una distribución ji-cuadrado con *n* grados de libertad, Bajo las mismas condiciones, la variable aleatoria $\sum_{i=1}^{n}\left(\dfrac{X_i-\bar{X}}{\sigma}\right)^2$ tiene una distribución ji-

[8] $\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i-\mu-\bar{X}+\mu)^2=\sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2=$

$=\sum_{i=1}^{n}\left[(X_i-\mu)^2-2(\bar{X}-\mu)(X_i-\mu)+(\bar{X}-\mu)^2\right]=\sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\mu)}_{=n(\bar{X}-\mu)}+\underbrace{\sum_{i=1}^{n}(\bar{X}-\mu)^2}_{=n(\bar{X}-\mu)^2}=\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2$

cuadrado con $v = n-1$ grados de libertad, lo cual resulta de reemplazar μ por $\overline{X}$ (debido a que se desconoce μ), por lo que se pierde un grado de libertad al estimar μ usando información muestral (es decir, $n-1$ piezas independientes de información)

Características de la distribución muestral de S^2

1. Por ser ji-cuadrado, es sesgada a derecha.
2. La probabilidad de que una muestra aleatoria produzca un valor χ^2 mayor que algún específico es igual a α.
3. Exactamente el 95% de una distribución ji-cuadrado yace entre $\chi^2_{0,975}$ y $\chi^2_{0,025}$, Mirando al estadístico χ^2 puede verse que un valor que cae a la derecha de $\chi^2_{0,025}$ es poco probable a menos que σ^2 sea demasiado pequeña; igualmente, un valor que cae a la izquierda de $\chi^2_{0,975}$ es poco probable a menos que σ^2 sea demasiado grande, Este razonamiento será de gran utilidad a la hora de realizar inferencias con respecto a la varianza poblacional.

Ejemplo 5:

Un fabricante de baterías afirma que la duración promedio de sus baterías es de tres años con una desviación estándar de uno, Si se toma una muestra aleatoria de cinco de estas baterías y se encuentra los siguientes valores: 1,9; 2,4; 3,0; 3,5 y 4,2, ¿qué puede decirse de la afirmación del fabricante con respecto a la desviación estándar?

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \equiv \frac{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}{n(n-1)} = \frac{5(48{,}26) - 15^2}{5(4)} = 0{,}815$$

$$\chi^2 = \frac{S^2(n-1)}{\sigma^2} = \frac{0{,}815 \cdot 4}{1^2} = 3{,}26$$

Como $\chi^2_{4;0,975} = 0{,}484$ y $\chi^2_{4;0,025} = 11{,}143$, entonces $\chi^2_{4;0,975} < \chi^2 < \chi^2_{4;0,025}$ y $\sigma^2 = 1$ (y por lo tanto $\sigma = 1$) es razonable.

4. Distribución t

Muchas veces, la misma información muestral que produce $\overline{X}$ debe usarse para estimar σ debido al desconocimiento de la población o del proceso, En este caso, la estadística apropiada para realizar inferencias sobre μ es:

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Si la muestra proviene de una población normal, puede escribirse:

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \cdot \frac{\sigma}{\sigma} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{1}{\frac{s}{\sigma}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{1}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}},$$

donde $V = \frac{S^2(n-1)^2}{\sigma^2} \sim \chi^2_{v=n-1}$ y $Z \sim N(z;0,1)$

Entonces:

Teorema: Si Z es una variable aleatoria con distribución normal estándar y V una variable aleatoria con distribución ji-cuadrado con $v = n$ grados de libertad, y además son independientes entre sí, entonces la distribución de la variable aleatoria T, donde:

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}}$$

está dada por

$$f(x;v) = \frac{1}{v\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \cdot \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{n}}, \quad -\infty < x < \infty, \; v > 0$$

Esta distribución se conoce como t de student con $v = n$ grados de libertad.

Corolario: Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes con distribución normal con media μ y varianza σ^2 y, además.

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

entonces la variable aleatoria $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$

tiene distribución t con $v = n - 1$ grados de libertad.

Para muestras grandes, los valores de S^2 no difieren significativamente de los de σ^2, por lo que la distribución de probabilidad de la estadística T tiende a la normalidad conforme $n \to \infty$.

Para muestras pequeñas, los valores de S^2 fluctúan demasiado de una muestra a otra y la distribución de T se aparta considerablemente de la normalidad. En estos casos, debe tratarse con la distribución exacta.

En términos concretos, si $n \geq 30$, la distribución de T se aproxima suficientemente bien mediante la distribución normal. Si $n < 30$, entonces debe tratarse con la distribución apropiada, que es la t de student. El valor límite $n = 30$ **no está relacionado** con el teorema del límite central, sino más bien con la propiedad de consistencia[9] de S^2 como estimador de σ^2.
Es importante destacar que la distribución t de student es simétrica con forma de campana, aunque, comparada con la normal estándar, aquélla es más dispersa (es decir, más "achatada"), Como puede verse para la estadística T, a medida que aumenta el tamaño de muestra, y por ende, el número de grados de libertad, la distribución t tiende a parecerse cada vez más a la normal estándar, de tal forma que cuando $\nu \to \infty$, $t \to Z$.

Ejemplo 6:
Cierto producto concentrado se encuentra dentro de especificaciones si su rendimiento **promedio** está dentro 500 ± 0,4 (g/ml). El rendimiento sigue una distribución normal. Para controlar que así sea se toma una muestra de 25 paquetes encontrándose un rendimiento promedio de 518 (g/ml) y una desviación estándar de 40 (g/ml). ¿Qué puede concluirse?

Datos: $\mu = 500;\ n = 25;\ \bar{x} = 518;\ s = 40$

Como $X \sim N(x;\mu,\sigma)$, σ es desconocida (o no se tienen datos) y $n < 30$, entonces

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2{,}25$$

El criterio es $-t_{0,05;\nu=n-1} < t < t_{0,05;\nu=n-1}$ y por la tabla sabemos que $t_{0,05;24} = \pm 1{,}711$.

Por lo tanto, como $2{,}25 > 1{,}711$, el producto está fuera de especificaciones (aunque debe notar que el rendimiento es mayor de lo esperado). La gerencia deberá decidir si *demasiada* calidad no tiene un costo no recuperable.

5. Distribución *F*

Así como T tiene utilidad en problemas relacionados con inferencias acerca de la media poblacional y χ^2 en inferencias acerca de la varianza poblacional, la distribución *F* es útil en la comparación de varianzas muestrales para realizar inferencias sobre las varianzas de dos poblaciones distintas.

[9] Un estadístico es un ***estimador consistente*** de un parámetro si al aumentar el tamaño de muestra su valor se aproxima al del parámetro. Simbólicamente, $\hat{\theta} \to \theta$ cuando $n \to \infty$.

La estadística *F* se define como $F = \frac{U/v_1}{V/v_2}$

donde *U* y *V* son variables aleatorias independientes con distribución ji-cuadrado, divididas por sus respectivos grados de libertad.

Teorema: Sean *U* y *V* son variables aleatorias con distribución ji-cuadrado independientes con ν_1 y ν_2 grados de libertad, respectivamente. Entonces, la distribución de la variable aleatoria

$$F = \frac{U/v_1}{V/v_2}$$

está dada por:

$$f(x;v_1,v_2) = \begin{cases} \dfrac{\Gamma\left(\dfrac{v_1+v_2}{2}\right).v_1^{\frac{v_1}{2}}.v_2^{\frac{v_2}{2}}}{\Gamma\left(\dfrac{v_1}{2}\right)\Gamma\left(\dfrac{v_2}{2}\right)} \cdot \dfrac{x^{\frac{v_1-2}{2}}}{(v_2+v_1.x)^{\frac{v_1+v_2}{2}}} & \text{para } x>0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Esta distribución se conoce como F de Fischer-Snedecor, con ν_1 y ν_2 grados de libertad.

La curva de *F* no sólo depende de sus grados de libertad ν_1 y ν_2, sino del orden en que se establecen. La distribución f de Fischer es asimétrica, con sesgo positivo (al igual que la ji-cuadrado) y por lo tanto un valor de f que deje un área $\alpha/2$ a su derecha será distinto de aquél otro que deje un área $\alpha/2$ a su izquierda. Para encontrar dichos valores, se hace uso del siguiente teorema:

Teorema: Si *F* tiene una distribución f con ν_1 y ν_2 grados de libertad, entonces *F'* = 1/*F* tiene una distribución f pero con ν_2 y ν_1 grados de libertad (en ese orden), de tal forma que

$$f_{1-a};v_1,v_2 = \frac{1}{f_a;v_2,v_1}$$

(¡Preste atención al orden de los grados de libertad!)

Por último, suponga que las muestras de tamaño n_1 y n_2 se seleccionan de poblaciones normales independientes con varianzas σ_1^2 y σ_2^2 y además S_1^2 y S_2^2 son las respectivas varianzas muestrales, entonces la estadística

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene distribución f con $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$ grados de libertad.

Ejemplo 6:

Halle entre qué valores yace el 90% de la distribución f cuando se toman dos muestras, de tamaños 7 y 11 respectivamente, de sendas poblaciones normales.

Buscando en la tabla de valores para $\alpha = 0{,}05$

$f_{0,05;6,10} = 3{,}22$

Entonces:

$$f_{0,95;6,10} = \frac{1}{f_{0,05;10,6}} = \frac{1}{4{,}06} = 0{,}246$$

Técnicas de muestreo

Hemos hablado de muestras desde el comienzo de nuestros estudios y hemos definido características que debe tener una muestra. Para determinar la manera en que serán seleccionados de la población los elementos de la muestra es necesario recurrir a las llamadas *técnicas de muestreo*.

Se denomina *muestreo* al procedimiento mediante el cual se obtiene una muestra de la población.

Existen dos tipos de muestreo: el ***probabilístico*** y el ***no probabilístico***.

Con el muestreo probabilístico, todos los sujetos tienen la misma probabilidad de formar parte del estudio. El no probabilístico es aquel en el que no todos los sujetos tienen la misma probabilidad de formar parte de la muestra de estudio.

Muestreo probabilístico

Muestreo aleatorio simple

Para poder realizar este tipo de muestreo, todos los individuos de la población deben estar numerados en un listado. Normalmente, se hace a partir de un listado de números aleatorios, disponible en casi todos los libros de estadística, con un programa estadístico, o con alguno de los programas para calcular el tamaño de la muestra que tenga la opción de generar listados de números aleatorios.

Si no se dispone del listado de individuos, no se podrá utilizar esta técnica de muestreo, por lo que se debe recurrir a otro tipo de muestreo que no precise tener a los individuos identificados.

Muestreo aleatorio sistemático

Los individuos deben estar identificados, pero no es necesario disponer de un listado. Éstos no se eligen a partir de un listado de números aleatorios, sino que se hace sistemáticamente eligiendo a uno de cada cierto número de sujetos.

Este número se denomina *razón de muestreo* (k) y se calcula dividiendo el total de elementos de la población por el tamaño de la muestra:

Por ejemplo, si se tiene una población de 8000 individuos y el tamaño de la muestra necesario es de 400, se seleccionará uno de cada 20, que será la razón de

muestreo (8000/400). Para decidir por cuál se ha de comenzar, se selecciona aleatoriamente, un número del 1 al 20, y a partir de dicho número se va seleccionando a un sujeto de cada 20.
En este caso, si por azar se elige el 7º elemento para comenzar, el segundo será el 27º, el tercero será el 47º y así, el último será el valor que ocupe el 7987º lugar.

Muestreo aleatorio estratificado
En este tipo de muestreo se divide a la población en subgrupos o estratos que tienen alguna característica común y teniendo en cuenta que, además, interesa mantener estos estratos en la muestra, para que se mantenga la composición de la población.
La selección de sujetos dentro de cada estrato se realizará aleatoriamente.
La estratificación se suele hacer en función de diferentes variables o características de interés: género, edad, situación laboral, etcétera.
Si se desea efectuar una estratificación por género y se sabe que en la población la distribución es del 55% de mujeres y 45% de hombres, la muestra seleccionará de cada estrato esta misma proporción. Por tanto, si el tamaño de la muestra es de 400, se elegirán aleatoriamente 220 mujeres y 180 hombres.
Si bien no es obligatorio mantener la proporción de los estratos en la muestra, el muestreo estratificado proporcional es el que menor error de muestreo produce.

Muestreo por conglomerados
Los conglomerados son lo contrario de los estratos. Mientras los estratos son homogéneos internamente y heterogéneos entre ellos, los conglomerados son heterogéneos en su interior y bastante homogéneos entre ellos.
Este tipo de muestreo también se denomina en "etapas múltiples o multietápico". Se emplea cuando se desea estudiar una población grande y dispersa, y no se dispone de ningún listado para poder aplicar las técnicas anteriores.
La diferencia con los estratos del tipo de muestreo anterior es que los conglomerados ya están agrupados así de forma natural (escuelas, barrios, etcétera).
Algunos autores proponen que, por ejemplo, entre todos los barrios de cierto nivel socioeconómico (que serían los conglomerados) se elige uno al azar y se estudia a todos los individuos del mismo.
Otros autores consideran que en lugar de seleccionar sujetos, se empieza por seleccionar subgrupos o conglomerados a los que se da el nombre de "unidades de primera etapa" o "unidades primarias". En una segunda etapa, se seleccionan, de manera aleatoria, las "unidades de segunda etapa" o "unidades secundarias", a partir de las unidades primarias. Así, sucesivamente, se van eligiendo hasta llegar a las unidades de análisis, que serán los individuos que compongan la muestra de estudio.

Muestreo no probabilístico

Los tipos de muestreo no probabilístico más utilizados son: accidental, de conveniencia, por cuotas y por bola de nieve.

Muestreo accidental

Este tipo de muestreo se denomina también "consecutivo", ya que la selección de los sujetos de estudio se hace en función de su presencia o no en un lugar y momento determinados. Es el caso, por ejemplo, de la inclusión de las mujeres a medida que van acudiendo al hospital, o el de un encuestador que, en la calle, entrevista a las personas que pasan en ese momento por allí.

Aunque puede parecer similar al muestreo probabilístico, es evidente que no todas las personas tienen la misma probabilidad de estar en el momento y el lugar donde se selecciona a los sujetos.

Muestreo de conveniencia

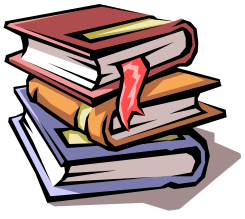
Los investigadores deciden, según sus criterios de interés y basándose en los conocimientos que tienen sobre la población, qué elementos entrarán a formar parte de la muestra de estudio. En este muestreo no probabilístico es muy importante definir con claridad los criterios de inclusión y exclusión, y cumplirlos rigurosamente.

Muestreo por cuotas

Consiste en seleccionar la muestra considerando una serie de características específicas presentes en la población, por lo que la muestra habrá de tenerlas en la misma proporción. Las cuotas se establecen a partir de variables consideradas relevantes: grupos de edad, género, categoría laboral, etcétera.

Muestreo por bola de nieve

Se utiliza cuando la población es difícil de identificar o cuando es complicado acceder a ella porque tiene ciertas características que no son muy aceptadas socialmente. Consiste en ir seleccionando los individuos a partir de un solo elemento o de un grupo reducido, que va conduciendo a otros individuos que reúnen las características de estudio; éstos, a su vez, conducen a otros y así se va obteniendo el número de individuos necesario.



Actividad bibliográfica

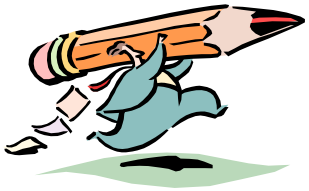
1. Lea en las páginas 215 a 228, los apartados 8.4, 8.5 y 8.6 del libro *Probabilidad y Estadística para Ingenieros* de Walpole, Myers y Myers.
   Tenga en cuenta las siguientes recomendaciones al estudiar este material:

- Página 217: El teorema 8.2 es de suma importancia, debe recordarlo (¡y muy bien!) para poder aplicarlo adecuadamente.
- Página 217: En el párrafo anterior al ejemplo 8.13 dice m < 30 y debe decir n < 30.
- Página 219: En el párrafo que dice: "En otras palabras, si la media μ es 5, ¿cuál es la posibilidad de que $\overline{X}$ se desvíe a lo más en 0,027 milímetros?" convendría, para que se entienda mejor, reemplazar "a lo más" por "al menos", Quedando, entonces: "En otras palabras, si la media μ es 5, ¿cuál es la posibilidad de que $\overline{X}$ se desvíe al menos en 0,027 milímetros?"
- Página 220: En las fórmulas anteriores al Teorema 8.3.
  - Corrija el subíndice de la segunda media muestral en la expresión de la media poblacional, debe ser 2 en lugar de 1.
  - Corrija el exponente del cociente entre la varianza de la primera población y el tamaño de muestra n_1, debe ser 2 en lugar de 1.
- Página 220: Debe recordar el teorema 8.3.
- Página 221: Reemplace Pr por P en el segundo párrafo de la página y en los otros lugares que aparezca.
- Página 221: Reemplace *sabemos que* por *suponemos que* en la primera oración de la solución del Ejemplo 8.15.
- Página 225: En el segundo párrafo de la página dice: *…y se calcula la varianza muestral σ^2 obtenemos…*, pero allí hay un error porque el símbolo de la varianza muestral debe ser S^2.
- Página 226: Debe recordar el teorema 8.4.
- Página 226: Como en capítulos anteriores, el libro hace referencia a sus propias tablas, pero nosotros realizaremos todos los cálculos con las tablas de la cátedra.

¡A repasar,,,!

Sabemos que ha encarado solo este tema y que puede tener algunas dudas.
Para autoevaluarse, responda las preguntas que están a continuación, Puede hacerlo con el material de estudio, pero asegurándose que "entiende" cada palabra, a tal punto que usted podría explicarle a un amigo, que no conoce el tema, de manera simple, los conceptos estudiados:

- ☑ **¿Recuerda la diferencia entre parámetro y estadístico o estadística?**
- ☑ **¿Qué es una distribución muestral?**
- ☑ **¿Qué dice el Teorema del límite central?**
- ☑ **¿Cuándo la aproximación normal para la media muestral es buena y cuándo no lo es, para distintos valores del tamaño de muestra?**
- ☑ **¿Cómo se distribuye la estadística media muestral?**
- ☑ **¿Cómo se distribuye la estadística diferencia entre medias muestrales?**
- ☑ **¿Cuándo la aproximación normal para la diferencia entre medias muestrales es buena y cuándo no lo es, para distintos valores del tamaños de muestra?**
- ☑ **¿Cómo se distribuye la estadística varianza muestral?**
- ☑ **¿Cuáles son las técnicas de muestreo?**



Por favor, no avance al siguiente tema si tiene dudas o no recuerda las nociones aquí volcadas, Pero si se siente listo para continuar, es hora de empezar a trabajar con las **autoevaluaciones**,,,

Aclaración: En esta unidad no hay **aplicaciones prácticas**.