Libro 4: Convergencia y distribución muestral

Sitio: Agencia de Habilidades para el Futuro
Curso: Estadística y probabilidades para el desarrollo del soft 1° D
Libro: Libro 4: Convergencia y distribución muestral

Imprimido por: Eduardo Moreno
Día: lunes, 19 de mayo de 2025, 01:20

Tabla de contenidos

Población y muestra

Ideas iniciales

Supongamos que una empresa está analizando lanzar una ayuda económica para sus empleados que tengan hijos. Para lo cual, deben primero hacer un análisis en este aspecto.

Si la empresa tiene dos mil empleados, podríamos armar una tabla como la siguiente:

Empleados	1	2	3	4	5	6	7	8	9	10	$\ldots$	2000
Cantidad de hijos	3	2	1	1	0	3	4	2	1	3	$\ldots$	2

Y así hasta el empleado número 2000. Con esta información podríamos realizar diferentes tareas, como por ejemplo:

- Armar la función de distribución de probabilidad puntal.
- Calcular la probabilidad de que si seleccionamos un empleado al azar, tenga menos de tres hijos
- Construir un histograma.
- Calcular su media y varianza. **¡Vamos a focalizar en este punto!**

Este conjunto de elementos, llamado población ya que estamos considerando a todos los empleados, tiene una media μ y una varianza σ^2.

Ahora bien, si la empresa quiere hacer entrevistas personales, sería muy engorroso hacerlo para los 2000 empleados. Sin embargo, puede tomar una muestra representativa y trabajar con este subgrupo de empleados. Acá es importante entender que la muestra tomada también va a tener su propia media y varianza, las cuales estarán relacionadas de alguna u otra manera con las de la población, como explicaremos en las próximas secciones.

Antes de seguir con otros ejemplos, definamos estos dos conceptos centrales para este libro: población y muestra.

Definición de población

Una población consta de la totalidad de las observaciones en las que estamos interesados.

El número de observaciones en la población se define como el **tamaño de la población** y se lo suele simbolizar con la letra N. En nuestro ejemplo, el tamaño de la población es de $N = 2000$.

El tamaño de la población puede ser finito (como el de nuestro ejemplo) o infinito, si pensamos en las observaciones que se obtienen al medir diariamente la presión atmosférica desde el pasado hasta el futuro o todas las mediciones de la profundidad de un lago desde cualquier posición concebible.

Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$.

Notación: la media poblacional se denota con μ y la varianza con σ^2.

Definición de muestra

Una muestra es un subconjunto de una población. El tamaño muestral se lo suele indicar con la letra n.

Por ejemplo, si de los 2000 elegimos de manera aleatoria un grupo representativo de 100 empleados, entonces decimos que hemos tomado una muestra cuyo tamaño es de $n = 100$.

Notación: la media muestral se denota con $\bar{x}$ y la varianza con s^2.

¿Qué es una población binomial? ¿O una población normal?

Ejemplo 1:

Si se inspeccionan artículos que salen de una línea de ensamble para buscar defectos, entonces cada observación en la población podría ser un valor 0 o 1 de la variable aleatoria X de Bernoulli, con una distribución de probabilidad

$$b(x, 1, p) = p^x q^{1-x}, \text{ con } x = 0, 1$$

donde 0 indica un artículo sin defecto y 1 indica un artículo defectuoso. De hecho, se supone que p, la probabilidad de que cualquier artículo esté defectuoso, permanece constante de una prueba a otra.

Ejemplo 2:

El tiempo de duración de las baterías de almacenamiento son valores que toma una variable aleatoria continua que podría llegar a tener una distribución normal.

De ahora en adelante, cuando nos refiramos a una "**población binomial**", a una "**población normal**" o, en general, a la "**población** $f(x)$", aludiremos a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o la distribución de probabilidad $f(x)$, respectivamente.

Muestreo aleatorio

Retomemos el ejemplo sobre la cantidad de hijos de los empleados de una empresa. ¿Observaron que acá tenemos el total de la población y tomamos una muestra representativa de ella para hacer más fácil el análisis?

Ahora pensemos lo siguiente. Supongamos que tenemos una fábrica de focos y queremos indicar cuál es su tiempo promedio de funcionamiento.

¿Sería posible encender todos los focos fabricados y calcular el promedio de duración? La respuesta es bastante obvia, ya que si hacemos esto no tendríamos focos para vender. La pregunta recae entonces en lo siguiente: ¿cómo hacemos para estimar el promedio de duración de todos los focos sin tener que analizarlos a todos?

Para responder este interrogante surge la **inferencia estadística**, donde el estadístico se interesa, a partir de una muestra, en llegar a conclusiones respecto a una población, cuando es imposible o poco práctico conocer todo el conjunto de observaciones que la constituyen.

Muestreo aleatorio

Para que las inferencias que hacemos sobre la población a partir de la muestra sean válidas, debemos obtener muestras que sean representativas de ella. Se dice que cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población, está **sesgado**. Para eliminar cualquier posibilidad de sesgo en el procedimiento de muestreo es deseable elegir una muestra aleatoria, lo cual significa que las observaciones se realicen de forma independiente y al azar.

Para seleccionar una muestra aleatoria de tamaño n de una población $f(x)$, definimos la variable aleatoria X_i, con $i = 1, 2, \ldots, n$, que representa la i-ésima medición o valor de la muestra que observamos. Si las mediciones se obtienen repitiendo el experimento n veces independientes en, esencialmente, las mismas condiciones, las variables aleatorias $X_1, X_2, \ldots, X_n$ constituirán entonces una muestra aleatoria de la población $f(x)$ con valores numéricos $x_1, x_2, \ldots, x_n$.

Debido a las condiciones idénticas en las que se seleccionan los elementos de la muestra, es razonable suponer que las n variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes y que cada una tiene la misma distribución de probabilidad $f(x)$.

Definición de muestra aleatoria

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes n, cada una con la misma distribución de probabilidad $f(x)$. Definimos $X_1, X_2, \ldots, X_n$ como una muestra aleatoria de tamaño n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

Algunos estadísticos importantes

Nuestro principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros desconocidos de la población. Suponga, por ejemplo, que deseamos concluir algo respecto a la proporción de consumidores de mate en Argentina que prefieren cierta marca de yerba. Sería imposible interrogar a cada consumidor argentino de mate para calcular el valor del parámetro p que representa la proporción de la población. En vez de esto se selecciona una muestra aleatoria grande y se calcula la proporción $\hat{p}$ de personas en esta muestra que prefieren la marca de yerba en cuestión.

El valor $\hat{p}$ se utiliza para hacer una inferencia respecto a la proporción p verdadera. Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria, ya que es posible tomar muchas muestras aleatorias de la misma población y esperaríamos que $\hat{p}$ variara un poco de una a otra muestra. Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P. Tal variable aleatoria se llama **estadístico**.

Definición de estadístico

Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama estadístico.

Algunos estadísticos importantes

En los libros anteriores, analizamos los parámetros μ y σ^2, que miden el centro y la variabilidad de una distribución de probabilidad. Éstos son parámetros de población constantes y de ninguna manera se ven afectados o influidos por las observaciones de una muestra aleatoria.

Definiremos algunos estadísticos importantes que describen las medidas correspondientes de una muestra aleatoria, los cuales se pueden clasificar en dos grandes grupos:

- Aquellos que se utilizan para medir el centro de un conjunto de datos: media, mediana y moda.
- Aquellos que se utilizan para medir la variabilidad: varianza, desvío estándar y rango.

Medidas de tendencia central

Como su nombre lo indica, estos parámetros sirven para analizar cuál es el centro en que se distribuyen los datos muestrales. Aquí los más utilizados son la media, la mediana y la moda. Vamos a definirlos considerando que $X_1, X_2, \ldots, X_n$ representan n variables aleatorias.

a) Definición de media muestral

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

Observar que el estadístico $\bar{X}$ toma el valor $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ cuando X_1 toma el valor x_1, X_2 toma el valor x_2 y así sucesivamente. El término media muestral se aplica tanto al estadístico $\bar{X}$ como a su valor calculado $\bar{x}$.

b) Definición de mediana muestral

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & si\ n\ es\ impar \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & si\ n\ es\ par \end{cases}$$

La mediana muestral es el número que se encuentra en el medio de todas las observaciones, ordenadas de menor a mayor o viceversa, es decir, es el valor que deja el 50% de los datos a su izquierda y a su derecha.

c) Definición de moda o modo

Es el valor que ocurre con mayor frecuencia en la muestra, es decir, el que tiene mayor frecuencia absoluta.

Medidas de variabilidad

La variabilidad en la muestra refleja cómo se dispersan las observaciones respecto de su centro.

Es posible tener dos conjuntos de observaciones con las mismas media o mediana que difieran de manera considerable en la variabilidad de sus mediciones sobre el promedio.

Considere las siguientes mediciones, en litros, para dos muestras de jugo envasado por las empresas A y B:

Muestra A	$0,97$	1	$0,94$	$1,03$	$1,06$
Muestra B	$1,06$	$1,01$	$0,88$	$0,91$	$1,14$

Pueden hacer el cálculo de la media y verán que ambas tienen la misma que es de 1 litro. Sin embargo, es evidente que la empresa A envasa el jugo con un contenido más uniforme y cercano a la media que la B. Decimos entonces que la variabilidad o la dispersión de las observaciones a partir del promedio es menor para la muestra A que para la B. Por lo tanto, al comprar jugo tendríamos más confianza en que el envase que seleccionemos se acerque al promedio anunciado si se lo compramos a la empresa A.

Vamos a definir las medidas de variabilidad teniendo en cuenta que $X_1, X_2,...,X_n$ representan n variables aleatorias.

a) Definición de la varianza muestral:

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

El valor calculado de S^2 para una muestra dada se denota con s^2.

b) Definición de la desviación estándar muestral:

$$S= \sqrt[]{S^2}$$

donde S^2 es la varianza muestral.

En nuestro ejemplo, la desviación estándar de la empresa A es menor que la de la empresa B.

c) Definición de rango muestral:

Si $X_{máx}$ denota el valor más grande de X_i y $X_{mín}$ el más pequeño, entonces el rango muestral se define como la diferencia entre el valor más grande y el más pequeño de toda la muestra. En símbolos nos queda:

$$R=X_{máx}-X_{mín}$$

Teorema sobre la varianza muestral

Si S^2 es la varianza de una muestra aleatoria de tamaño n, podemos escribir:

$$ S^2= \frac{1}{n(n-1)} \left[n \displaystyle\sum_{i=1}^{n} X_i^2- \left(\displaystyle\sum_{i=1}^{n} X_i \right)^2 \right] $$

Si bien acá hemos definido brevemente los seis estadísticos, nos seguiremos enfocando en la media y la varianza. Los demás serán retomados en el Bloque 2 sobre Estadística.

Ejemplo

Ejemplo:

Una consultora local hizo un relevamiento del precio del kilogramo de pan en diez panaderías de dos barrios cercanos. Los datos recolectados fueron los siguientes:

Barrio 1	\($710 \)	\($790 \)	\($720 \)	\($780 \)	\($750 \)	\($740 \)	\($760 \)	\($750 \)	\($735 \)	\($765 \)
Barrio 2	\($745 \)	\($755 \)	\($750 \)	\($740 \)	\($760 \)	\($750 \)	\($735 \)	\($765 \)	\($748 \)	\($752 \)

a) ¿En qué barrio el precio del kilogramo de pan, en promedio, es más barato?
b) ¿En qué barrio sería más esperable que el precio que nos cobren por el kilo de pan esté cercano a la media?

Solución:
a) Para responder a esta pregunta, debemos calcular la media muestral en cada caso:

- \(\bar{x_1}= \frac{1}{10} \cdot ($710+$790+$720+$780+$750+$740+$760+$750+$735+$765)=$750 \)

- \(\bar{x_2}= \frac{1}{10} \cdot ($745+$755+$750+$740+$760+$750+$735+$765+$748+$752)=$750 \)

Podemos ver que la media en ambas muestras coincide. Por lo tanto, no hay un barrio que tenga el precio más barato en promedio.

b) Para responder a la segunda pregunta, vamos a calcular la varianza muestral en cada caso:

- \(s^2_1= \frac{1}{9} \displaystyle\sum_{i=1}^{10} (x_i-750)^2= \frac{(710-750)^2+(790-750)^2+(720-750)^2+(780-750)^2+(750-750)^2+(740-750)^2+(760-750)^2+(750-750)^2+(735-750)^2+(765-750)^2}{9} \approx 627,78 \)

Por lo tanto, \(s_1 \approx 25,06 \).

- \(s^2_2= \frac{1}{9} \displaystyle\sum_{i=1}^{10} (x_i-750)^2= \frac{(745-750)^2+(755-750)^2+(750-750)^2+(740-750)^2+(760-750)^2+(750-750)^2+(735-750)^2+(765-750)^2+(748-750)^2+(752-750)^2}{9} \approx 78,67\)

Por lo tanto, \(s_2 \approx 8,87 \).

Algo que se podía intuir por los valores de las tablas, lo hemos corroborado con los cálculos. Los valores del barrio \(2 \) están más centrados al rededor de la media que los del barrio \(1 \), ya que su desviación estándar es menor. Por lo tanto, es más esperable que en el barrio \(2 \) consigamos el kilogramo de pan a un precio más cercano a la media muestral, es decir, a los \($750 \).

Ley de los Grandes Números

Lanzamiento de cuatro monedas

Supongamos que tenemos cuatro monedas equilibradas, las lanzamos simultáneamente y anotamos la cantidad de caras que salen en cada lanzamiento.

caras: 2

Por lo tanto, según lo que vimos antes, podemos llamar a la variable aleatoria discreta X como la cantidad de caras que salen en cada lanzamiento. Además, es posible armar su función de distribución de probabilidad y calcular su valor esperado, de la siguiente manera:

x	0	1	2	3	4
$f(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

$\mu = E(X)=0 \cdot \frac{1}{16} + 1 \cdot \frac{1}{4}+2 \cdot \frac{3}{8}+ 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{16} =2$

Esto significa que se espera obtener dos caras en cada lanzamiento que se realice, ¿lógico no?

Poner a prueba el experimento

Lo que hicimos anteriormente fue resolver el ejercicio de manera teórica (por decirlo de alguna manera). Pero ahora les proponemos simular este experimento, para lo cual deberán ingresar al siguiente enlace:

https://www.geogebra.org/m/pmxXRa55

En él podrán simular el experimento de lanzar las cuatro monedas e ir observando la cantidad de caras que salen en cada uno. Verán que el software va registrando toda la información en una tabla, donde aparece la frecuencia absoluta (cantidad de veces que salió ese valor) y la frecuencia relativa (cociente entre la frecuencia absoluta y la cantidad de lanzamientos). Además, en la parte derecha aparece un histograma que se va actualizando con cada lanzamiento, en el cual pueden tildar la opción "ver probabilidades teóricas", que son los cálculos que hicimos en la función de distribución de probabilidad.

Por lo tanto, los invitamos a realizar lo siguiente:

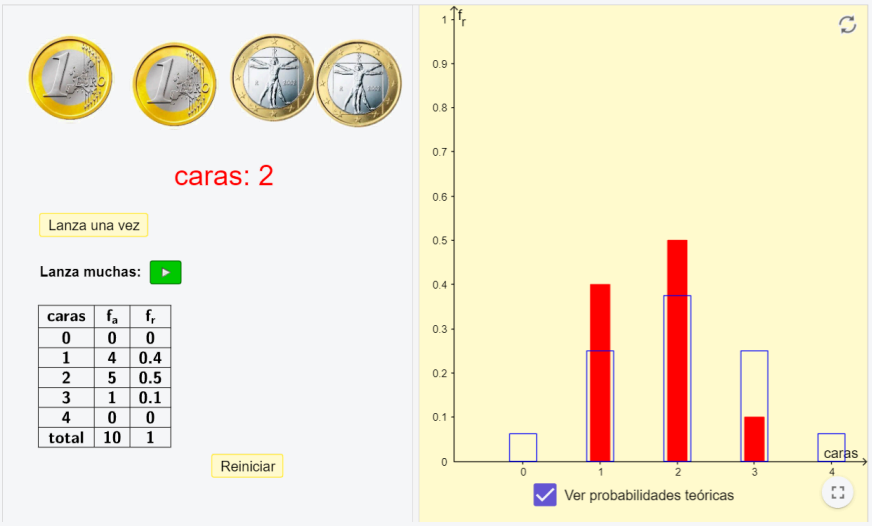
1. Poner en funcionamiento la simulación del lanzamiento de las cuatro monedas.
2. Calcular la media para los 10, 50, 500 y 1000 lanzamientos.
3. Responder la siguiente pregunta: ¿observan alguna relación entre la media calculada para cada cantidad de lanzamiento del punto 2 con el valor esperado calculado de "manera teórica" ($\mu = 2$)?

Una vez que tengan anotado todo esto, los invitamos a que continuen con la lectura del libro.

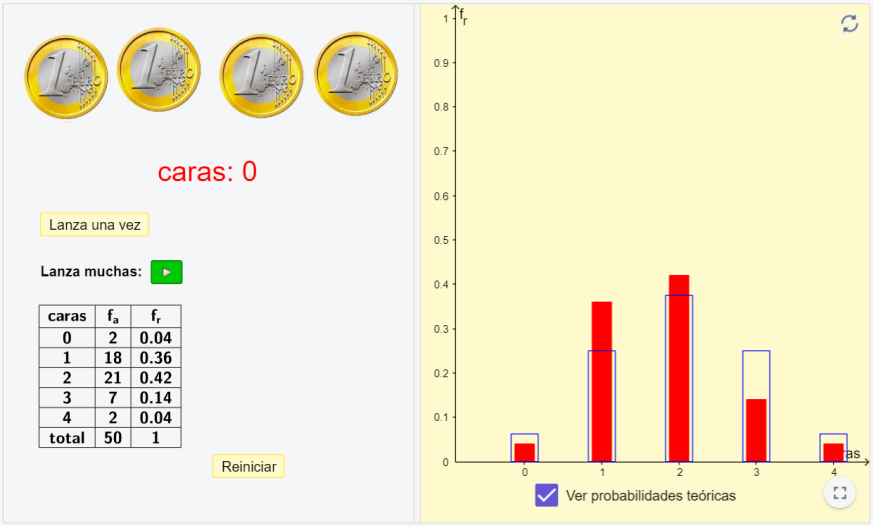
Simulación del experimento

Aquí les dejemos los resultados que nos arrojó a nosotros el software cuando lo pusimos en funcionamiento, siendo \(n \) la cantidad de observaciones o las veces en que se repitió el experimento en las mismas condiciones, correspondiente con el tamaño muestral.

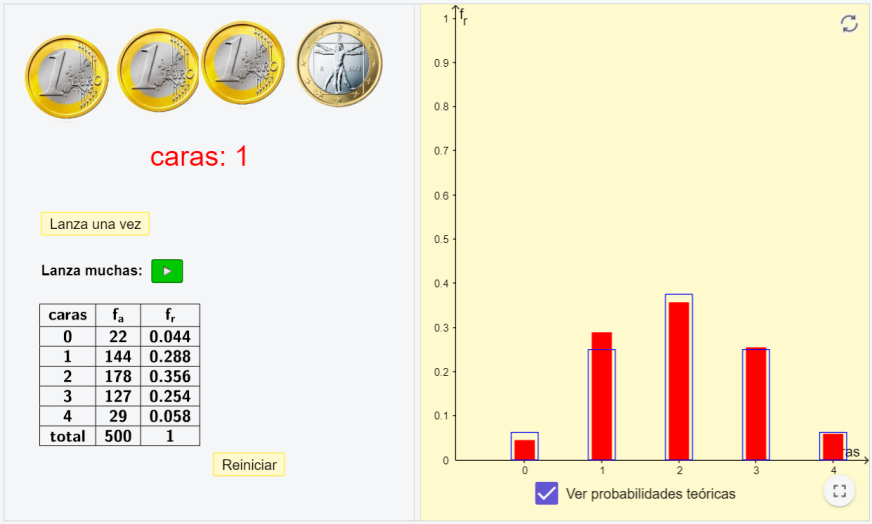
- **Para \(n=10 \), la media es de \(1,7 \):**



- **Para \(n=50 \), la media es de \(1,78 \):**

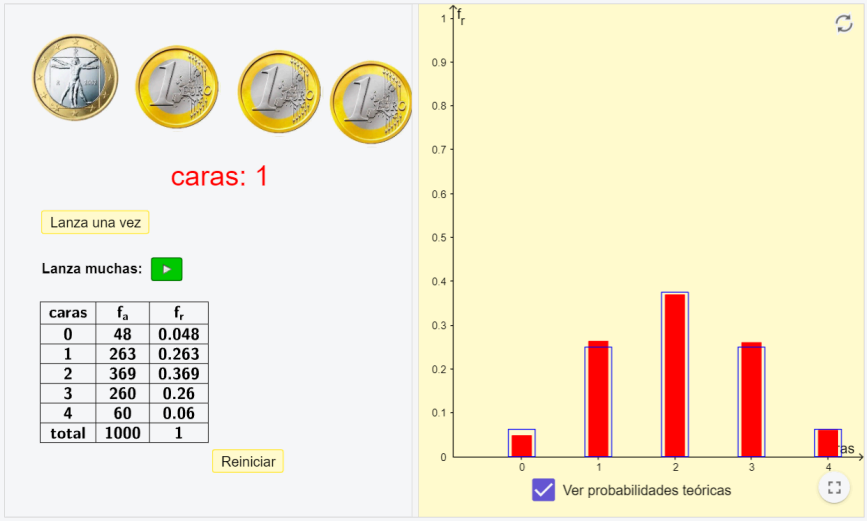


- **Para \(n=500 \) , la media es de \(1,994 \):**



caras	f_a	f_r
0	22	0.044
1	144	0.288
2	178	0.356
3	127	0.254
4	29	0.058
total	500	1

- **Para $n=1000$, la media es de $2,021$:**



caras	f_a	f_r
0	48	0.048
1	263	0.263
2	369	0.369
3	260	0.26
4	60	0.06
total	1000	1

Algunas conclusiones:

- Podemos observar que a medida que repetimos el experimento una mayor cantidad de veces, la frecuencia relativa f_r de cada valor de la variable aleatoria (barras de color rojo en el histograma), tiende a la probabilidad teórica p (rectángulos azules en el histograma). Esto es similar a decir que la diferencia entre estas dos cantidades tiende a cero cuando n tiende a infinito.
- El promedio del número de caras observado cuando cuatro monedas equilibradas son arrojadas, se aproxima al valor medio $\mu = 2$ de la distribución, cuando el número de repeticiones del experimento se hace cada vez más grande. Observen que para $n=1000$, la media muestral nos dio por encima del valor esperado $\mu = 2$, lo cual es totalmente posible.

Por lo tanto, de acuerdo a nuestro experimento, podemos decir que cuando $n \to \infty$,

$$f_r \to p \quad \text{y} \quad \bar{X_n} \to \mu$$

Todo esto lo establece la **Ley de los Grandes Números** que definiremos a continuación.

Definición

Ley de los Grandes Números

Sea X una variable aleatoria con función de densidad o de distribución de probabilidad $f(x)$ y con $E(X) = \mu$. Supongamos que se desea "estimar" μ. Como hemos visto que la esperanza de una variable aleatoria se puede pensar como un promedio de sus valores, parece razonable estimarla mediante el promedio de valores observados de X. Por supuesto que en una situación real solo tendremos un número finito de observaciones y nos preguntamos: usando solo un número finito de valores de X, ¿puede hacerse inferencia confiable respecto de $E(X)$?

Según lo que vimos en nuestro ejemplo anterior, para un número considerable de observaciones del experimento, el promedio muestral converge a μ. Esto es lo que afirme la **Ley de los Grandes Números**, la cual nos dice (en forma coloquial) que el promedio $\bar{X}$ converge a μ cuando el número de observaciones (o tamaño de la muestra) tiende a infinito.

¿En qué sentido converge $\bar{X}$ a μ?

Sea X_n, con $n \geq 1$, una sucesión de variables aleatorias, diremos que X_n converge en probabilidad a la variable aleatoria X si

$$\displaystyle\lim_{n \to \infty }P(\left| \begin{matrix} X_n - X \end{matrix} \right|> \epsilon)=0 \), \(\forall \epsilon >0$$

Definición

Sean $X_1, X_2, \ldots$ variables aleatorias independientes e idénticamente distribuidas (muestra aleatoria) con $E(X)=\mu$ y $V(X)= \sigma ^2 < \infty$, entonces

$$\bar{X}_n \rightarrow \mu,$$

siendo $\bar{X}_n = \frac{\displaystyle\sum_{i=1}^{n} X_i}{n}$ el denominado promedio muestral.

El caso Bernoulli

Versión Bernoulli de la Ley de los Grandes Números

Consideremos n repeticiones independientes de un experimento aleatorio y sea A un suceso con probabilidad $P(A) = p$, constante en las n repeticiones. Si llamamos f_a a la frecuencia absoluta de A (número de veces que ocurre A en las n repeticiones) y $f_r = \frac{f_a}{n}$ a la frecuencia relativa, entonces

$$f_r \rightarrow p$$

es decir,

$$P(\left| \begin{matrix} f_r - p \end{matrix} \right| > \epsilon) \rightarrow 0), \forall \epsilon > 0,$$

cuando $n \rightarrow \infty$.

Observen que esto es lo que dijimos al comienzo en nuestro ejemplo del lanzamiento de las cuatro monedas. Al efectuar el experimento, tanto en la tabla como en el gráfico, se pudo observar cómo las frecuencias relativas tendían a la probabilidad teórica.

El ejemplo de los dos dados

En libros anteriores hemos explicado que si lanzamos un dado equilibrado, la probabilidad de que salga un número es $\frac{1}{6} \approx 0,167$.

Ahora supongamos que en las mismas condiciones repetimos el experimento una cantidad de veces y calculamos la frecuencia relativa en que sale cada número. **¿Qué sucede con esa frecuencia relativa cuando la cantidad de repeticiones es cada vez más grande?**

Para responder a esta pregunta, los invitamos a ingresar al siguiente simulador:

https://www.geogebra.org/m/cSfwCRkM

En él, deberán apretar primero el botón de "reiniciar" y luego, varias veces la opción de "lanzar un dado". Verán que aparecerá el número obtenido y se irá armando un histograma con la frecuencia relativa de cada número.

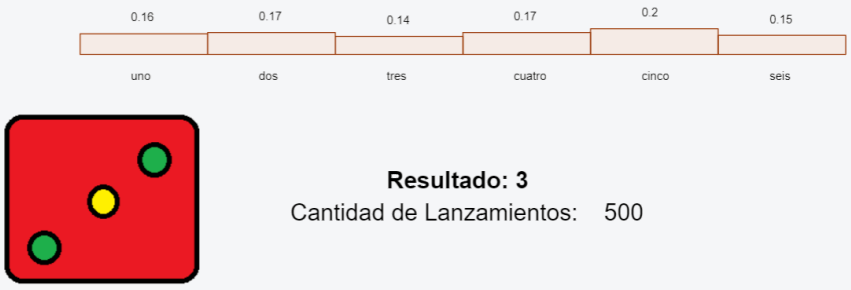
Los invitamos a que interactúen con el simulador e intenten responder la pregunta que formulamos arriba. Una vez lo hayan hecho, podremos continuar con el desarrollo del libro.

Realización del experimento

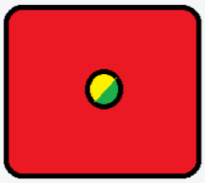
La realización de nuestro experimento arrojó estos resultados para los diferentes valores de n:

- **$n=50$:**

- **$n=500$:**



- **$n=1000$:**



Resultado: 1
Cantidad de Lanzamientos: 1000

Análisis de lo obtenido:

Analicemos qué sucedió por ejemplo con el número 6. Cuando se repitió el experimento 50 veces, fue el número con la frecuencia relativa más baja, esto significa que de los 50 números que salieron, fue el que menos veces salió. Sin embargo, a medida que el experimento se siguió repitiendo, su frecuencia relativa logró equipararse hasta llegar a $0,16$ con 1000 lanzamientos. ¿No les resulta familiar este número? ¡Exacto! Es aproximadamente la probabilidad teórica que tiene el número 6 de salir en el dado. Y esto ocurre con todos los demás elementos del espacio muestral equiprobable.

Claramente podrán advertir que si seguimos repitiendo una mayor cantidad de veces el experimento, las frecuencias relativas van a tender a la probabilidad de que salga ese número.

Distribuciones muestrales

Volvamos al ejemplo del embotellamiento de jugo. Supongamos que ahora un ejecutivo le pide al encargado del sector que calcule la media de 50 botellas servidas, el cual obtiene que $\bar{x} = 0,98$. Con base a este valor, el ejecutivo decide que la máquina está sirviendo bebidas con un contenido promedio de $\mu = 1$ litro. Las 50 botellas servidas representan una muestra de la población infinita de posibles bebidas que despachará esta máquina.

El ejecutivo de la empresa decide que la máquina despachadora está sirviendo bebidas con un contenido promedio de 1 litro, aunque la media de la muestra fue de $0,98$ litros, porque conoce la teoría del muestreo según la cual, si $\mu = 1$ litro, tal valor de la muestra podría ocurrir fácilmente. De hecho, si realiza pruebas similares, cada hora por ejemplo, esperaría que los valores del estadístico $\bar{x}$ fluctuaran por arriba y por abajo de $\mu = 1$ litro. Solo cuando el valor de $\bar{x}$ difiera considerablemente de 1 litro, el ejecutivo de la empresa tomará medidas para ajustar la máquina.

Definición de distribución muestral

Como un estadístico es una variable aleatoria que depende solo de la muestra observada, debe tener una distribución de probabilidad. La distribución de probabilidad de un estadístico se denomina distribución muestral.

¿De qué depende la distribución muestral de un estadístico?

La distribución muestral de un estadístico depende:

- de la distribución de la población,
- del tamaño de las muestras y
- del método de selección de las muestras.

En las próximas secciones estudiaremos la distribución muestral de uno de los estadísticos más importantes: la media. La varianza también tiene su propia distribución muestral pero no la trabajaremos en este curso.

Distribución muestral de la media

Ideas iniciales

Se deberían considerar las distribuciones muestrales de $\bar{X}$ y S^2 como los mecanismos a partir de los cuales se puede hacer inferencias acerca de los parámetros μ y $\sigma ^2$.

La distribución muestral de $\bar{X}$ con tamaño muestral n es la distribución que resulta cuando un experimento se lleva a cabo una y otra vez (siempre con una muestra de tamaño n)) y resultan los diversos valores de $\bar{X}$. Por lo tanto, esta distribución muestral describe la variabilidad de los promedios muestrales alrededor de la media de la población μ. En el caso de la máquina despachadora de bebidas, el conocer la distribución muestral de $\bar{X}$, le permite al analista encontrar una discrepancia "típica" entre un valor $\bar{x}$ observado y el verdadero valor de μ. Se aplica el mismo principio en el caso de la distribución de S^2. La distribución muestral produce información acerca de la variabilidad de los valores de s^2 alrededor de $\sigma ^2$ en experimentos que se repiten.

Distribución muestral de la media $\bar{X}$

Suponga que de una población normal con media μ y varianza $\sigma ^2$ se toma una muestra aleatoria de n observaciones. Cada observación X_i, $i = 1, 2,..., n$, de la muestra aleatoria tendrá entonces la misma distribución normal que la población de donde se tomó. Así, por la propiedad reproductiva de la distribución normal concluimos que:

$$\bar{X}= \frac{1}{n} \left(X_1+X_2+...+X_n\right)$$

tiene una distribución normal con media

$$\mu_{\bar{X}}= \frac{1}{n} (\mu + \mu + ... + \mu) = \mu$$

y varianza

$$\sigma^2_{\bar{X}}= \frac{1}{n^2} (\sigma^2 + \sigma^2 + ... + \sigma^2) = \frac{\sigma^2}{n}$$

donde $\mu + \mu + ... + \mu$ y $\sigma^2 + \sigma^2 + ... + \sigma^2$ son sumas de n términos.

Podemos agregar también que la desviación estándar es:

$$\sigma_{\bar{X}} = \frac{\sigma}{ \sqrt[]{n} }$$

Ejemplo

Una empresa de material eléctrico fabrica focos que tienen una duración que se distribuye aproximadamente de forma normal, con media de 800 horas y desviación estándar de 40 horas.

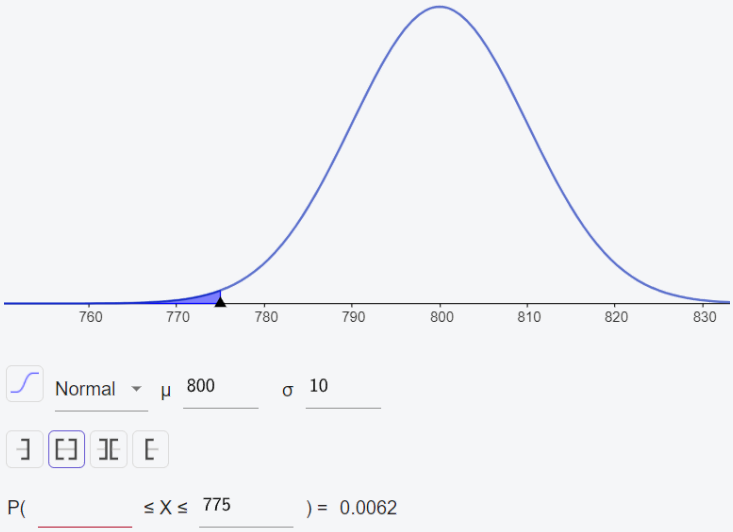
Calcular la probabilidad de que una muestra aleatoria de 16 focos tenga una vida promedio de menos de 775 horas.

Solución:

Observen que ahora nos están pidiendo calcular la probabilidad en relación a un valor promedio, entonces debemos pensar en la distribución de la media muestral. Como la muestra se toma de una población normal, cada observación de la muestra aleatoria tendrá entonces la misma distribución normal de la población de donde se tomó.

También dijimos que $\mu_{\bar{X}}= \mu = 800$ y que $\sigma_ {\bar{X}}= \frac{\sigma}{ \sqrt[]{n} } = \frac{40}{ \sqrt[]{16} }=10$.

Entonces podemos indicar que $\bar{X} \sim N(800,10)$ y debemos calcular $P(\bar{X} < 775)$. La probabilidad que se desea es determinada por el área de la región sombreada de la siguiente imagen:



En conclusión, $P(\bar{X} < 775) \approx 0,006$.

Teorema del límite central

Ideas iniciales

Si tomamos muestras de una población con distribución desconocida, ya sea finita o infinita, la distribución muestral de $\bar{X}$ aún será aproximadamente normal con media μ y varianza $\frac{\sigma^2}{n}$, siempre que el tamaño de la muestra sea grande. Este asombroso resultado es una consecuencia inmediata del siguiente teorema, que se conoce como **teorema del límite central**.

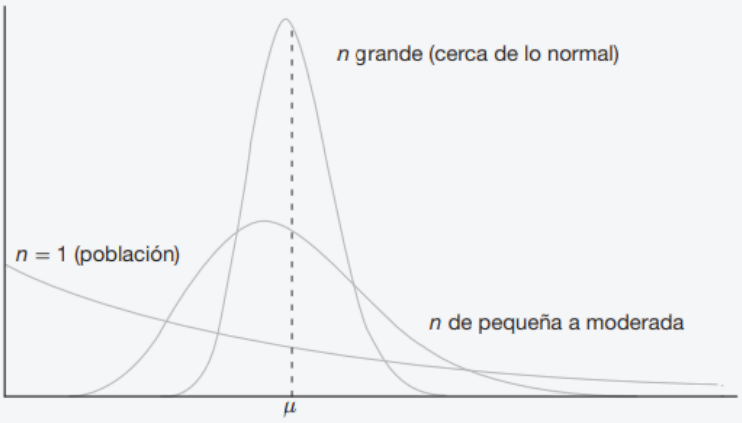
Enunciación del Teorema del límite central

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n, tomada de una población con media μ y varianza finita $\sigma ^2$, entonces la forma límite de su distribución a medida que $n\rightarrow \infty$ se aproxima a la distribución normal.

La aproximación normal para $\bar{X}$ por lo general será buena si $n \geq 30$, siempre y cuando la distribución de la población no sea muy asimétrica. Si $n < 30$, la aproximación será buena solo si la población no es muy diferente de una distribución normal y, como antes se estableció, si se sabe que la población es normal, la distribución muestral de $\bar{X}$ seguirá siendo una distribución normal, sin importar qué tan pequeño sea el tamaño de las muestras.

El tamaño de la muestra $n = 30$ es un lineamiento para el teorema del límite central. Sin embargo, como indica el planteamiento del teorema, la suposición de normalidad en la distribución de $\bar{X}$ se vuelve más precisa a medida que n se hace más grande.

Para entender cómo funciona este teorema veamos la siguiente imagen:



La figura indica cómo la distribución de $\bar{X}$ se acerca más a la normalidad a medida que aumenta n, empezando con la distribución claramente asimétrica de una observación individual $(n = 1)$. También ilustra que la media de $\bar{X}$ sigue siendo μ para cualquier tamaño de la muestra y que la varianza de $\bar{X}$ se vuelve más pequeña a medida que aumenta n.

Inferencias sobre la media de la población

Una aplicación muy importante del teorema del límite central consiste en determinar valores razonables de la media de la población μ. Temas como prueba de hipótesis, estimación, control de calidad y muchos otros utilizan el teorema del límite central. Sin embargo, estos conceptos no serán trabajados en este curso, por lo que los invitamos a seguir investigando en este sentido.

Ejemplo

Retomemos nuestro ejemplo de la cantidad de hijos por empleado de la empresa.

Si llamamos \(X \) a la variable aleatoria discreta que representa la cantidad de hijos por empleado, su distribución de probabilidad es:

\(x \)	\(0 \)	\(1 \)	\(2 \)	\(3 \)	\(4 \)
\(P(X=x) \)	\(0,17 \)	\(0,42 \)	\(0,24 \)	\(0,1 \)	\(0,07 \)

Con esta información, se nos pide:
a) Calcular la media \(\mu \) y la varianza \(\sigma ^2 \) de \(X \).
b) Calcular la media \(\mu_{\bar{X}} \) y la varianza \(\sigma^2_{\bar{X}} \) de la media \(\bar{X} \) para muestras aleatorias de 49 empleados.
c) Calcular la probabilidad de que el número promedio de hijos en 49 empleados sea menor que \(1 \).

Solución:

a) Para el primer apartado, calculamos tal como se hizo en los libros anteriores:

$$\mu = E(X)= 0 \cdot 0,17 + 1 \cdot 0,42 + 2 \cdot 0,24 + 3 \cdot 0,1 + 4 \cdot 0,07 = 1,48$$

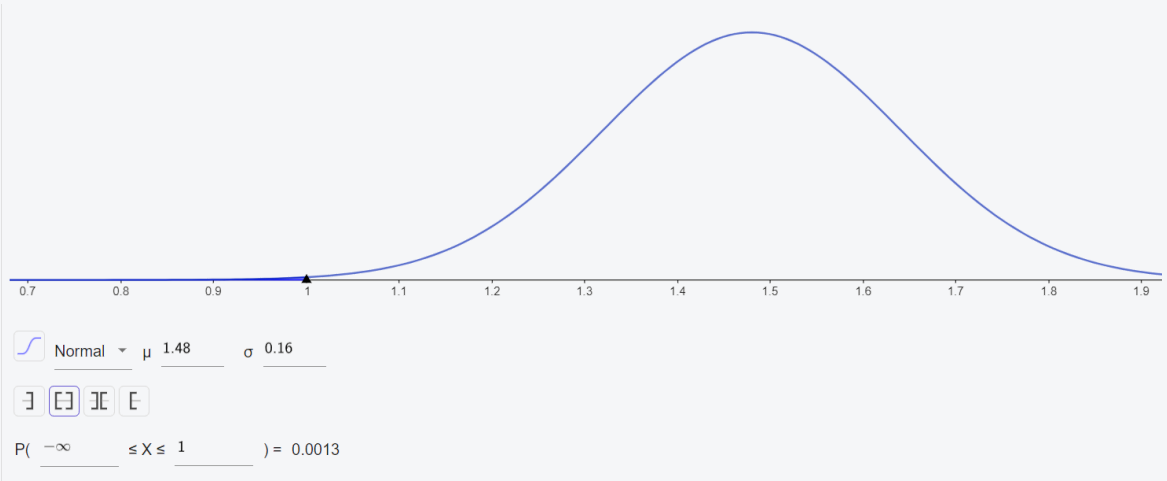
$$\sigma^2 = (0 - 1,48)^2 \cdot 0,17 + (1 - 1,48)^2 \cdot 0,42 + (2 - 1,48)^2 \cdot 0,24 + (3 - 1,48)^2 \cdot 0,1+ (4 - 1,48)^2 \cdot 0,07 = 1,2096$$

b) Como $n = 49 >30$, según el teorema del límite central, $\bar{X} \sim N$, por lo que:

$$\mu_{\bar{X}}= \mu = 1,48 \quad \text{y} \quad \sigma^2_{\bar{X}} = \frac{1,2096}{49} \approx 0,025$$

De lo anterior se desprende que la desviación estándar es de aproximadamente $0,16$.

c) Para responder este ítem, debemos calcular $P(\bar{X} < 1)$. Utilizando GeoGebra vemos que $P(\bar{X} <1) \approx 0,001$, una probabilidad muy baja, casi nula.



Distribución muestral de la diferencia entre dos medias

Suponga que tenemos dos poblaciones, la primera con media μ_1 y varianza σ^2_1, y la segunda con media μ_2 y varianza σ^2_2. Representemos con el estadístico $\bar{X_1}$ la media de una muestra aleatoria de tamaño n_1, seleccionada de la primera población, y con el estadístico $\bar{X_2}$ la media de una muestra aleatoria de tamaño n_2 seleccionada de la segunda población, independiente de la muestra de la primera población.

¿Qué podríamos decir acerca de la distribución muestral de la diferencia $\bar{X_1} - \bar{X_2}$ para muestras repetidas de tamaños n_1 y n_2?

De acuerdo con el teorema del límite central, tanto la variable $\bar{X_1}$ como la variable $\bar{X_2}$ están distribuidas más o menos de forma normal con medias μ_1 y μ_2 y varianzas $\frac{\sigma^2_1}{n_1}$ y $\frac{\sigma^2_2}{n_2}$, respectivamente. Esta aproximación mejora a medida que aumentan n_1 y n_2.

Al elegir muestras independientes de las dos poblaciones nos aseguramos de que las variables $\bar{X_1}$ y $\bar{X_2}$ sean independientes. Además, concluimos que $\bar{X_1} - \bar{X_2}$ se distribuye aproximadamente de forma normal con media

$$\mu_{\bar{X_1}-\bar{X_2}}= \mu_{\bar{X_1}}-\mu_{\bar{X_2}}= \mu_1 - \mu_2$$

y varianza

$$\sigma^2_{\bar{X_1}-\bar{X_2}}= \sigma^2_{\bar{X_1}} + \sigma^2_{\bar{X_2}}= \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}$$

El teorema del límite central se puede ampliar fácilmente al caso de dos muestras y dos poblaciones.

Enunciación del Teorema del límite central para dos muestras

Si se extraen al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_1 y μ_2 y varianzas σ^2_1 y σ^2_2, respectivamente, entonces la distribución muestral de las diferencias de las medias, $\bar{X_1} - \bar{X_2}$, tiene una distribución aproximadamente normal, con media y varianza dadas por

$$\mu_{\bar{X_1}-\bar{X_2}}= \mu_1 - \mu_2 \quad y \quad \sigma^2_{\bar{X_1}-\bar{X_2}}= \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}$$

Algunas apreciaciones sobre el tamaño de n

Si tanto n_1 como n_2 son mayores o iguales que 30, la aproximación normal para la distribución de $\bar{X_1} - \bar{X_2}$ es muy buena cuando las distribuciones subyacentes no están tan alejadas de la normal. Sin embargo, aun cuando n_1 y n_2 sean menores que 30, la aproximación normal es hasta cierto punto buena, excepto cuando las poblaciones no son definitivamente normales. Por supuesto, si ambas poblaciones son normales, entonces $\bar{X_1} - \bar{X_2}$ tiene una distribución normal sin importar de qué tamaño sean n_1 y n_2.

Ejemplo

Los cinescopios para televisor del fabricante A tienen una duración media de $6,5$ años y una desviación estándar de $0,9$ años. Mientras que los del fabricante B tienen una duración media de 6 años y una desviación estándar de $0,8$ años.

¿Cuál es la probabilidad de que una muestra aleatoria de 36 cinescopios del fabricante A tenga por lo menos 1 año más de vida media que una muestra de 49 cinescopios del fabricante B?

Solución:

Primero vamos a organizar la información que nos brinda el enunciado del problema en una tabla como la siguiente:

Población 1	Población 2
$\mu_1=6,5$	$\mu_2=6$
$\sigma_1=0,9$	$\sigma_1=0,8$
$n_1=36$	$n_2=49$

Como los tamaños muestrales superan los 30, podemos utilizar el teorema anterior y la distribución muestral de $\bar{X_1} - \bar{X_2}$ será aproximadamente normal con una media

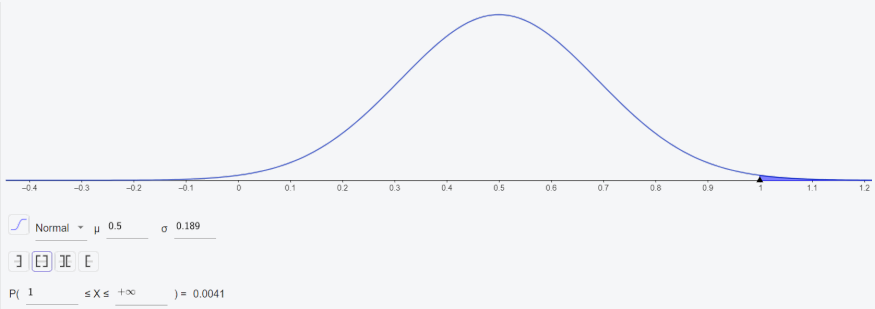
$$\mu_{\bar{X_1}-\bar{X_2}}=6,5-6=0,5$$

y una desviación estándar de

$$\sigma_{\bar{X_1}-\bar{X_2}}= \sqrt[]{ \frac{0,81}{36}+ \frac{0,64}{49} } \approx 0,189$$

Por lo tanto, podemos indicar que $\bar{X_1} - \bar{X_2} \sim N(0,5;0,189)$ y la probabilidad que nos piden calcular es $P(\bar{X_1} - \bar{X_2} \geq 1)$.

Empleando el GeoGebra la calculamos y nos queda:



Por lo tanto, $P(\bar{X_1} - \bar{X_2} \geq 1) \approx 0,004$.