Distribución de muestreo

Aquí se expondrá el fundamento de la estimación, por qué pueden hacerse inferencias desde estadísticos de la muestra hasta los parámetros poblacionales correspondientes. Existen diferencias entre el estadístico calculado en la muestra y el correspondiente parámetro poblacional inferido. Por lo tanto, será necesario tener en cuenta la forma en que varían los estadísticos muestrales. La variabilidad producida en los datos estadísticos de las distintas muestras está relacionada con las características de la población, el tamaño de la muestra y el tipo de parámetro en estudio. Para analizar esta mutabilidad se hace necesario estudiar las distribuciones de muestreo y el teorema del límite central, herramientas imprescindibles para observar la variabilidad de los estadísticos y hacer inferencias estadísticas justificadas.

Distribución de muestreo para la media

Caso Productora de Té

La productora de té Ackerman e Hijos es oriunda del litoral y hace envíos a todo el país. Necesita saber cuál es el porcentaje del consumo de té por hogar en la República Argentina. Considera, de acuerdo a los estudios realizados con anterioridad, que anualmente, por hogar, el consumo está normalmente distribuido, con una desviación estándar de 1,25 Kg, pero desconoce la media μ.

La productora le encarga que realice un estudio estadístico para determinar:

1. Si se toma una muestra de 36 hogares seleccionados azarosamente y se calcula el promedio del consumo de té durante un año ¿cuál es la probabilidad de que la media de esa muestra no difiera respecto del consumo promedio de toda la población en más de 0,5 Kg?
2. ¿Cuál será la probabilidad de que el consumo medio de la muestra supere al consumo promedio poblacional en más de 0,75 kg?
3. Si se conoce que en la población el 60 % de las familias consumen la marca de té de la productora Ackerman, ¿qué probabilidades hay de que, en la muestra seleccionada, la media de la proporción de las familias que consumen su marca de té no difiera de la media poblacional más allá de ± 5 %?

Antes de comenzar, vale recordar:

- Al inicio de este módulo, se estudió la manera de determinar las muestras de una población con el objeto de definir los datos estadísticos para inferir sobre los parámetros de la población. Es lógico pensar que los estadísticos calculados en cada muestra no serán iguales entre sí y, por lo tanto, tampoco se debe esperar que los estadísticos de una muestra sean iguales a los parámetros poblacionales.
- Los parámetros poblacionales más usuales en la estimación son: a) la media, b) el desvío estándar y c) la proporción de los elementos de la población que cumplen con determinada característica.

Es decir, la media de la muestra permite inferir sobre la media poblacional, como también, el desvío estándar o la proporción de los elementos con la característica en estudio de la muestra permite inferir sobre el comportamiento de los parámetros poblacionales.

El objetivo es hacer buenas estimaciones. Para ello se define qué es una **distribución de muestreo de la media**, ya que es la base de la estimación estadística y, a la vez, uno de los temas centrales de esta lectura.

Definición de distribución de muestreo de la media

Ejemplo 1

En el caso hipotético de producción de autopartes, se toman muestras de un tipo de arandelas. Las muestras constan de 10 arandelas cada una. La población se considera infinita por ser un proceso continuo. Se desea medir el diámetro interior de la arandela. Al calcular la media y la desviación estándar correspondiente a cada una de estas muestras, pronto se ve que la media y la desviación estándar de cada una serían diferentes.

Definición

"Una distribución de probabilidad de todas las medias posibles de las muestras es una distribución de las medias de las muestras. Los especialistas en estadística la conocen como distribución de muestreo de la media" (Levan-Rubín, 2004, p. 247).

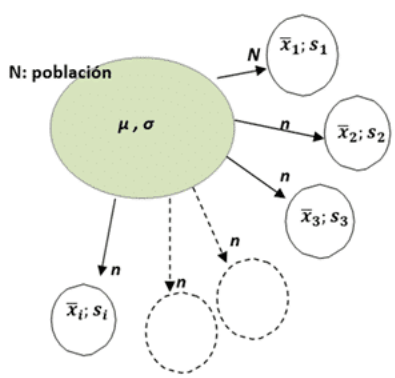
Generalización

Ahora se plantea un ejemplo teórico, no porque no exista un caso práctico para aplicar, sino porque, habitualmente, suele ser tedioso y costoso comprobarlo.

- Se considera una población infinita o finita, pero con un N lo suficientemente grande de forma tal que una muestra extraída sin reposición no modifique las probabilidades que tienen otras muestras de ser elegidas.
- La población en estudio tiene una media μ y un desvío estándar σ.
- De esa población se extraen todas las muestras posibles de un mismo tamaño n, cada una.
- De cada muestra se calcula:

la media: $\bar{x}$ y la desviación estándar s.

Figura 1: Esquema del proceso de obtención de una distribución de muestreo para la media



Fuente: elaboración propia.

Descripción de la figura. Esquema del proceso de obtención de una distribución de muestreo para la media. Es una esquematización teórica sobre la extracción de todas las posibles muestras del mismo tamaño en una población y cómo llegar a la distribución de muestreo de la media a partir del cálculo de la media y la desviación estándar de cada una de las muestras.

- [ ] Si se tienen en cuenta las medias de todas esas muestras y se genera con ellas una distribución, se obtendrá una **distribución de medias** o una **distribución de muestreo de las medias** o también, una **distribución de probabilidad de las medias muestrales**. Las tres formas de nombrarla son correctas.

- [ ] A cada media muestral, se la considera como una **variable aleatoria**, pues es el resultado numérico de un experimento aleatorio que consiste en tomar muestras azarosamente de una población y calcular la media de cada muestra. Esto permite confeccionar una tabla de distribución de frecuencias e, incluso, un gráfico con todas las medias de todas las muestras posibles tomadas de la población. Este procedimiento se denomina **distribución de muestreo de la media**.

- [ ] Podrá observarse que la distribución de medias muestrales tendrá una forma aproximadamente normal, independiente de la forma original de la población (haya sido normal o no).

- [ ] Como toda distribución de variables aleatorias, tendrá una media que, en este caso, se calcula como la **medida de todas las medias** y es un valor esperado.

- [ ] Al desvío estándar de esta distribución de medias muestrales se lo denomina **error muestral estándar**. Cuanto más pequeño sea este valor, menor será el error que se cometa al momento de inferir sobre la media poblacional a través de la media de una muestra.

De la misma manera, se puede obtener una **distribución de muestreo para la proporción**, con base en la proporción en que una característica se encuentra en todas las muestras posibles tomadas de la población.

El concepto de error muestral estándar

Estos conceptos expresan lo mismo: –

La desviación estándar de la distribución de las medias muestrales **es lo mismo que** el error estándar de la media.

Estos conceptos son distintos: –

El error estándar de la muestra **es distinto a** la desviación estándar de la población. Aunque existe una relación entre ellos.

Características del error muestral estándar –

- El error estándar indica qué tan dispersas (separadas) están las medias de las muestras entre sí.
- EL término **error estándar** se utiliza para explicar que la variabilidad de los datos estadísticos calculados en las muestras provienen de un error de muestreo. Este ocurre por la aleatoriedad de las muestras. Como ya se dijo, hay diferencias entre cada muestra y la población, y entre las diversas muestras, únicamente debido a los elementos que decidimos escoger.

Características de las distribuciones de muestreo para la media

En el tema anterior se dieron algunas características de las distribuciones de muestreo de la media, en esta sección se afianzarán esos conceptos y se profundizarán a través de un ejemplo concreto.

En el **ejemplo 1**, se supone una población infinita, para facilitar los cálculos y poder explicar matemáticamente a qué conclusiones se llega. Luego de visualizar y definir a esas características, se generalizarán a todo tipo de poblaciones con algunas salvedades.

Ejemplo 2. Análisis a partir de una población, para extraer conclusiones sobre las distribuciones de muestreo de la media.

Aquí se explicará, paso a paso, la variabilidad de los datos estadísticos a medida que aumenta la muestra seleccionada. Se analizará por etapas para extraer conclusiones en cada una. El objetivo no es memorizar estos pasos, sino que se ponga atención en los análisis que se realizan y a las conclusiones a las que se llega.

Una empresa de ventas cuenta con cinco vendedores a los que designaremos como A, B, C, D y E, cuyos salarios mensuales son de $40.000, $40.000, $50.000, $60.000 y $60.000, respectivamente. Las remuneraciones constituirán la población de estudio, en este caso, dicha población es finita y consta de solo 5 elementos.

Etapa 1: calcular los parámetros de la población

Se diseña una tabla con los datos del ejemplo 2 y se calculan los siguientes parámetros: la media poblacional μ, la varianza poblacional Var(x) o (σ^2) y la desviación poblacional σ.

$$\text{El salario promedio será: } \mu = \frac{\sum x_i}{N} = \frac{250000}{5} = \$\ 50000$$

Tabla 1: Salarios de los vendedores del ejemplo 2 con los cálculos para la varianza y la desviación estándar

Descripción de la figura: la tabla resume los datos de los cinco vendedores de la empresa del ejemplo 2 y cuenta con las columnas necesarias para calcular la varianza y la desviación estándar.

Trabajadores	Salarios en $ x_i	Media μ	$(x_i - \mu)$	$(x_i - \mu)^2$
A	40000	50000	-10000	100.000.000
B	40000	50000	-10000	100.000.000
C	50000	50000	0	0
D	60000	50000	10000	100.000.000
E	60000	50000	10000	100.000.000
Suma	**250000**			**400.000.000**

Fuente: elaboración propia.

Varianza: $Var(x) = \sigma^2 = \frac{\sum(x_i-\mu)^2}{N} = \frac{400.000.000}{5} = 80.000.000 = 8.10^7$

Desviación estándar: $\sigma = \sqrt{80.000.000} \cong 8944{,}27$

Nota: esta es una situación ideal de una población muy pequeña, por eso no se considera N-1 en el denominador para la varianza y la desviación estándar.

Resumen: $\begin{cases} \mu = \$\ 50000 \\ \sigma = \$\ 8944{,}27 \end{cases}$

Se pueden resumir los datos en una tabla de distribución de frecuencias, para luego ver la forma de la distribución de la población.

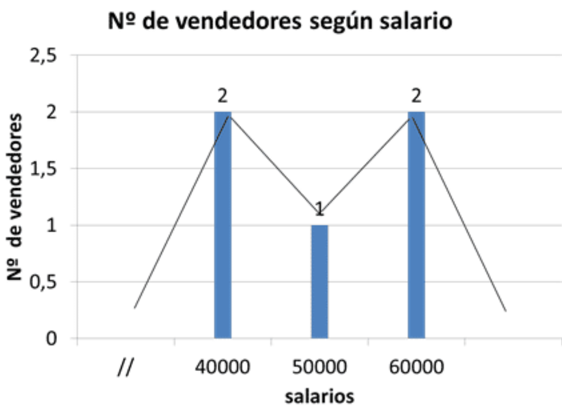
Tabla 2: Distribución de frecuencias para los vendedores del ejemplo 2

Descripción de la figura: se muestra la distribución de las frecuencias absolutas de la variable en estudio que son los sueldos de los vendedores del ejemplo 2.

Salarios en $ x_i	nº vendedores f_i
40000	2
50000	1
60000	2
suma	**5**

Fuente: elaboración propia.

Figura 2: Diagrama de bastones correspondiente a la tabla 2 del ejemplo 2



Fuente: elaboración propia.

Descripción de la figura. Diagrama de bastones correspondiente a la tabla 2 del ejemplo 2. Se graficó el diagrama de bastones correspondiente al ejemplo 2. Se elige este tipo de gráfico con bastones, porque las presentadas en el caso son variables discretas (aunque el sueldo generalmente se considera variable continua, pero aquí se simplificó). También se le asoció un polígono para ver la forma que adoptaría la distribución si se tratara de variables continuas.

Etapa 2: seleccionar todas las muestras posibles de tamaño 2 de la población

Se consideran todas las muestras posibles de dos elementos cada una de ellas. Hay que formar grupos de dos personas dentro de un grupo de cinco. No importa si una persona está en más de una pareja. Lo que diferencia un grupo de otro es que, por lo menos, haya un integrante distinto. Tampoco tiene relevancia en qué orden se nombran los grupos. Se trata de combinaciones de cinco elementos tomados de dos en dos, tal como se usó en la distribución de probabilidad binomial en el módulo 3. Por lo tanto, es necesario calcular , que, como ya se vio, es igual a 10 grupos de personas.

Tabla 3: Muestras de dos personas tomadas de la población de cinco con sus promedios de sueldo

Descripción de la figura: se escriben todos los posibles grupos de vendedores, combinándolos de a 2. Se calculan los promedios de los sueldos de cada muestra.

	Muestras	Salarios en miles de $	Media muestral $\bar{x}$
1	A - B	40; 40	40.000
2	A - C	40; 50	45.000
3	A - D	40; 60	50.000
4	A - E	40; 60	50.000
5	B - C	40; 50	45.000
6	B - D	40; 60	50.000
7	B - E	40; 60	50.000
8	C - D	50; 60	55.000
9	C - E	50; 60	55.000
10	D - E	60; 60	60.000

Fuente: elaboración propia.

Etapa 3: distribución de medias muestrales de tamaño 2

Con las medias de todas las muestras, se puede generar una nueva distribución, a la que se denominará **distribución de medias muestrales de tamaño 2** (o distribución de muestreo de la media). Primero se ordenan los valores de la variable de menor a mayor y luego se cuenta las veces que se repite (fi).

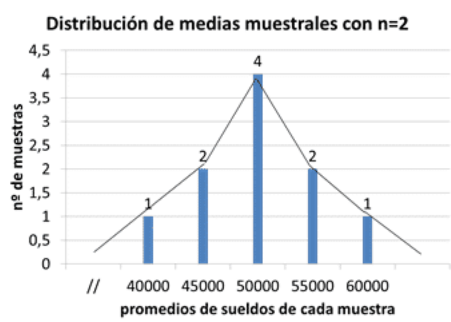
Tabla 4: Distribución de frecuencias para las muestras de vendedores del ejemplo 2, tomados de a 2

Descripción de la figura: se muestra la distribución de las frecuencias absolutas de la variable aleatoria. En este caso, la variable aleatoria es el promedio de los dos sueldos correspondientes a los dos vendedores que forman cada muestra.

Medias muestrales (promedio de salarios en $) $\bar{x}_i$	Cantidad de muestras de vendedores f_i
40000	1
45000	2
50000	4
55000	2
60000	1
total	**10**

Fuente: elaboración propia

Figura 3: Diagrama de bastones correspondiente a la tabla 4 del ejemplo 2



Fuente: elaboración propia

Descripción de la figura. Diagrama de bastones correspondiente a la tabla 4 del ejemplo 2. Se graficó el diagrama de bastones correspondiente a la tabla 4 con las frecuencias respectivas. También se le asoció un polígono para ver la forma que adopta la distribución de medias.

Etapa 4: calcular los estadísticos de la distribución de medias muestrales de la etapa 3

Ya que se cuenta con una distribución de una variable aleatoria representada por la media de cada una de las muestras, se calculan algunos de sus estadísticos, los que convienen para el ejemplo que se analiza son:

1 La media de la distribución muestral:

$$\bar{\bar{x}} = \mu_{\bar{x}} = \frac{\sum x_i . f_i}{\sum f_i} = \frac{500000}{10} = \$50000.$$

2 La desviación estándar de las medias muestrales o error estándar (o error muestral estándar, como vimos anteriormente).

$$\textbf{Varianza}: Var(x) = \sigma_{\bar{x}}^2 = \frac{\sum (x_i - \mu)^2 . f_i}{\sum f_i} = \frac{300.000.000}{10} = 30.000.000$$

$$\textbf{Desviación estándar}: \sigma_{\bar{x}} = \sqrt{30.000.000} \cong 5477,23$$

Además, interesa calcular:

3 la probabilidad de ocurrencia de las medias muéstrales y

4 la probabilidad de ocurrencia de las medias muestrales viene dada por las frecuencias relativas

$$fr_i = \frac{f_i}{\sum f_i}.$$

Así, la probabilidad de que una muestra de tamaño 2, tomada de esa población, proporcione un promedio salarial de $50000 es de: P(x=50000) = 0,4

$$\textbf{Resumen}: \begin{cases} \mu_{\bar{x}} = \$\ 50000 \\ \sigma_{\bar{x}} = \$\ 5477,23 \end{cases}$$

Conclusiones parciales 1

Si de la población se extraen todas las muestras posibles de tamaño 2 y de cada una de ellas se determina su media, la distribución generada con las medias de todas las muestras presenta las siguientes características:

1 La media de la distribución de las medias de todas las muestras posibles de la población de tamaño (n = 2) es igual a la media de la población (ver etapa 4).

2 El desvío estándar de dicha distribución, también denominado error muestral estándar, es menor que el desvío estándar poblacional (ver etapa 4).

3 Mientras que la población presenta un diagrama de frecuencias bimodal, en este caso, el correspondiente a la distribución de las medias muestrales tiende a ser normal (ver figura 3).

Etapa 5: seleccionar de la población todas las muestras posibles de tamaño n = 3

Con la atención puesta en la población, se consideran todas las muestras posibles de tres elementos cada una (n = 3). El número de muestras posibles estará dado por las combinaciones de 5 tomadas de 3 en 3.

Por lo tanto, se necesita calcular

$$C_{5;3}.$$

$$C_{5;3} = 10$$

Tabla 5: Muestras de 3 personas tomadas de la población de 5 y sus promedios de sueldo (ejemplo 2)

Descripción de la figura: se escriben todos los posibles grupos de vendedores, combinándolos de a 3. Se calculan los promedios de sueldos de cada muestra.

Muestras		Salarios en miles de $	Media muestral $\bar{x}$
1	A – B – C	40; 40; 50	433333,33
2	A – B – D	40; 40; 60	46666,67
3	A – B – E	40; 40; 60	46666,67
4	A – C – D	40; 50; 60	50000,00
5	A – C – E	40; 50; 60	50000,00
6	A – D – E	40; 60; 60	53333,33
7	B – C – D	40; 50; 60	50000,00
8	B – C – E	40; 50; 60	50000,00
9	B – D – E	40; 60; 60	53333,33
10	C – D – E	50: 60; 60	56666,67

Fuente: elaboración propia.

Etapa 6: Distribución de medias muestrales de tamaño n = 3

Con las medias de todas las muestras se genera una nueva distribución: **distribución de medias muestrales de tamaño 3** (o distribución de muestreo de la media). Hay que recordar que primero se ordenan los valores de la variable de menor a mayor y luego se cuentan las veces que se repite (fi).

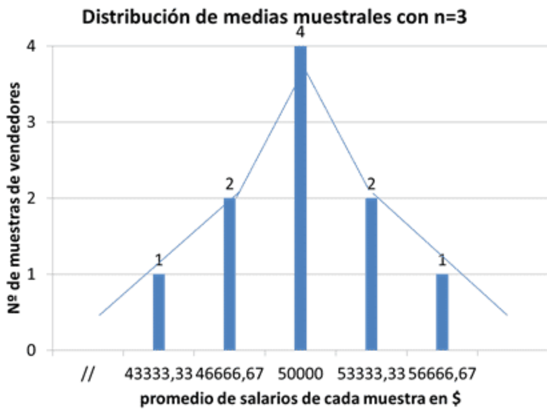
Tabla 6: Distribución de frecuencias para las muestras de vendedores del ejemplo 2, tomados de a 3

Descripción de la figura: muestra la distribución de las frecuencias absolutas de la variable aleatoria. En este caso, la variable aleatoria es el promedio de los 3 sueldos correspondientes a los tres vendedores que forman cada muestra.

Medias muestrales (promedio de salarios en $) $\overline{x}_i$	Cantidad de muestras de vendedores f_i
43333,33	1
46666,67	2
50000,00	4
53333,33	2
56666,67	1
total	**10**

Fuente: elaboración propia.

Figura 4: Diagrama de bastones correspondiente a la tabla 6



Fuente: elaboración propia.

Descripción de la figura. Diagrama de bastones correspondiente a la tabla 6. Se graficó el diagrama de bastones con las frecuencias respectivas correspondiente a la tabla 6. También se le asoció un polígono para ver la forma que adopta la distribución de medias.

Si en lugar de escribir en el eje vertical las frecuencias absolutas, se pusieran las frecuencias relativas, se estaría hablando de una distribución de probabilidades, tal como se estudió en el módulo 3. Si se tratara de una población más grande que la tomada en el ejemplo 2 y la variable aleatoria fuera continua, el polígono se suaviza. La curva en la figura 4 va tomando forma acampanada. También se cambió la escala del eje vertical por las frecuencias relativas para observar mejor la distribución de probabilidades.

Etapa 7: Calcular los estadísticos de la distribución de medias muestrales de la etapa 6

Ya que existe una distribución de una variable aleatoria representada por la media de cada una de las muestras, se calculan algunos de sus estadísticos. Lo conveniente para el ejemplo 2 son (ver tabla 6 para realizar los cálculos):

1 La media de la distribución muestral

$$\bar{\bar{x}} = \mu_{\bar{x}} = \frac{\sum x_i.f_i}{\sum f_i} = \frac{500000}{10} = \$50000$$

2 La desviación estándar de las medias muestrales o error estándar (o error muestral estándar).

Varianza: $Var(x) = \sigma_{\bar{x}}{}^2 = \frac{\sum (x_i - \mu)^2.f_i}{\sum f_i} = \frac{133.333.333,33}{10} = 13.333.333,33$

Desviación estándar: $\sigma_{\bar{x}} = \sqrt{13.333.333,33} \cong 3651,48$

Resumen: $\begin{cases} \mu_{\bar{x}} = \$\ 50000 \\ \sigma_{\bar{x}} = \$\ 3651,48 \end{cases}$

Conclusiones parciales 2

1 La media de la distribución de las medias de todas las muestras posibles de n = 3 elementos cada una es igual a la media de la población, lo que también ocurría en el caso de n = 2.

2 La distribución tiende a adquirir las características propias de una distribución normal.

3 El desvío estándar de la distribución es menor que el desvío estándar poblacional.

4 El desvío estándar de la distribución muestral de n = 3 es menor que el desvío estándar de la distribución muestral de n = 2. Es decir, a medida que el número de elementos que componen la muestra aumenta, disminuye la dispersión de la distribución muestral.

Conclusiones generales

Tabla 7: Resumen de los estadísticos calculados

Descripción de la figura: se sintetizan los estadísticos: media, error estándar y rango de las muestras; tomadas según tamaño n = 2 y n = 3.

Tamaño de la muestra: n	Media $\mu_{\bar{x}}$	Error estándar $\sigma_{\bar{x}}$	Rango
2	50000	5477,23	60000 – 40000 =20000
3	50000	3651,48	56666,67 – 43333,33 = 13333,34

Fuente: elaboración propia.

Del análisis del ejemplo 2 y del cuadro 7 se sacan las siguientes conclusiones:

1

La distribución de las medias muestrales para n grande, tiene forma normal:

1. Cuando la población es grande y está normalmente distribuida, esa distribución de las medias muestrales será normal.
2. Cuando la población no está distribuida normalmente, la distribución de las medias muestrales se aproximará a una distribución normal, si el tamaño de la muestra es suficientemente grande (treinta o más elementos).

2

La distribución de las medias muestrales tiene una media igual a la media poblacional:

$$\bar{\bar{x}} = \mu.$$

3

La desviación estándar de las distribuciones de las medias muestrales, denominada error estándar, está dada por la expresión:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Las conclusiones extraídas del análisis realizado con las distribuciones de las medias muestrales, a medida que se aumenta el tamaño de la muestra, lleva al enunciado del teorema fundamental de la estadística inferencial: **el teorema del límite central.**

Teorema del límite central

Si de una población de media μ y desviación estándar σ, se extraen todas las muestras posibles del mismo número de elementos n cada una de ellas, de cada muestra obtenemos su media

$$\bar{x},$$

la distribución de todas esas medias será del tipo normal, independientemente del tipo de distribución que sea la población, con una media

$$\bar{\bar{x}},$$

igual a la media poblacional μ y un desvío estándar

$$\sigma_{\bar{x}}$$

menor al desvío estándar poblacional, denominado error muestral estándar, cuyo valor es igual a:

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

(cociente entre el desvío estándar poblacional sobre la raíz cuadrada de n, tamaño de muestra).

Factor de corrección para poblaciones finitas

Poblaciones infinitas y finitas

Ya se habló de poblaciones infinitas, pero en términos generales, se trata de comparar a la población con la muestra.

1. La población se considera infinita si es mayor al 5 % de la muestra: $N > 0{,}05\ n$
2. Análogamente, la población es finita, si es menor o igual al 5 % de la muestra: $N \leq 0{,}05\ n$
   O lo que es lo mismo:

$$\frac{n}{N} \geq 0{,}05$$

Factor de corrección

Si estamos frente a una población finita, es decir, menor o igual al 5 % de la muestra, es necesario introducir un factor de corrección para poblaciones finitas para el cálculo del error estándar:

$$fc = \sqrt{\frac{N-n}{N-1}}$$

Por lo tanto, la fórmula de cálculo para el error estándar, en caso que la población sea finita, es:

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

Aplicación del teorema del límite central en el caso del ejemplo 2

Para n = 2:

1. El error estándar a partir de la distribución de medias muestrales, cuando n = 2, dio (ver tabla 7): $ **5477,23.**
2. El error estándar por el teorema del límite central. Se aplica, en este caso, el factor de corrección para poblaciones finitas (ver etapa 1 del ejemplo 2):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{8944{,}27}{\sqrt{2}} \times \sqrt{\frac{5-2}{5-1}} = 6324{,}55 \times \sqrt{\frac{3}{4}} = \mathbf{5477,22}$$

1

Para n = 3:

1. El error estándar a partir de la distribución de medias muestrales, cuando n = 3, dio (ver tabla 7): **$ 3651,48**
2. El error estándar por el teorema del límite central. Se aplica, en este caso, el factor de corrección:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{8944{,}27}{\sqrt{3}} \times \sqrt{\frac{5-3}{5-1}} = 5163{,}98 \times \sqrt{\frac{2}{4}} = \mathbf{3651,48}$$

Como puede comprobarse, los errores estándares dan el mismo resultado en los dos casos, por lo que se deduce que el teorema del límite central permite extraer una muestra representativa de una población e inferir, a partir de los estadísticos (la media y la desviación estándar, en este caso), los parámetros poblacionales correspondientes con un cierto grado de error manejable.

Resolución del caso de la productora de té, preguntas a) y b)

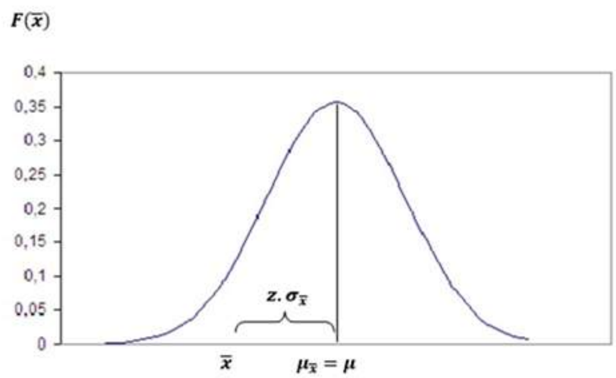
Por el teorema del límite central, se sabe que si se pudieran extraer de la población todas las muestras posibles de 36 hogares cada una de ellas y de cada una obtuviéramos el consumo medio anual de té con las medias de todas las muestras, se generaría una distribución muestral de consumos medios con forma aproximadamente normal, pues n > 30.

Por otra parte, la media de dicha distribución de medias coincide con la media poblacional, la cual es desconocida. Además, la media de la muestra diferirá respecto a la media de la distribución muestral en

$z.\sigma_{\bar{x}}$.

- Para una distribución normal: $z = \frac{x-\mu}{\sigma}$.
- En el caso de una distribución de medias muestrales: $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$.
- Entonces: $z.\sigma_{\bar{x}} = \bar{x} - \mu$ cómo se observa en la figura 5.

Figura 5: Esquema genérico de la distribución muestral de medias para el caso de la productora de té



Fuente: elaboración propia.

Descripción de la figura. Esquema genérico de la distribución muestral de medias para el caso de la productora de té. Representación de los datos del caso de la productora de té, en la que se visualiza cómo quedaría la distribución de medias muestrales con forma acampanada, con la media poblacional y la diferencia entre cualquier media muestral y la media poblacional.

Como en la primera pregunta del caso planteado, se pide cuál es la probabilidad de que la diferencia entre la media de la muestra tomada de 36 familias y la media poblacional no sea mayor a 0,5 kg, se plantea:

- $|\bar{x} - \mu| \leq 0{,}5$; en valor absoluto, pues lo que se pretende es que la media muestral no esté más allá de ± 0,5.

También se lo puede interpretar de la siguiente manera:

- Si la media poblacional es desconocida, podemos plantear que el intervalo a considerar es el que está dentro $[\mu - 0{,}5\ ;\ \mu + 0{,}5]$.

En resumen, se estandariza y se trabaja con la tabla de la distribución normal (tal como se expuso en el módulo 3), pero teniendo en cuenta que, aquí, la variable aleatoria es una media muestral, que la media poblacional es la media de todas las medias muestrales (desconocida en este caso) y que la desviación estándar se llama, en este supuesto, error estándar y tiene su propia fórmula en función inversa al tamaño de la muestra.

Se resuelve, seguidamente, la pregunta a). Aquí existe un problema: no se tiene ni la media de la muestra, ni la media poblacional. Pero sí se cuenta con la diferencia entre ambas, entonces se dilucidan dos casos:

- si $\bar{x} < \mu$: $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{-0{,}5}{\frac{\sigma}{\sqrt{n}}} = \frac{-0{,}5}{\frac{1{,}25}{\sqrt{36}}} = \frac{-0{,}5}{0{,}21} = -2{,}38$

Y, por otra parte,

- si $\bar{x} > \mu$: $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{0{,}5}{\frac{\sigma}{\sqrt{n}}} = \frac{0{,}5}{\frac{1{,}25}{\sqrt{36}}} = \frac{0{,}5}{0{,}21} = 2{,}38$

Por último, hay que buscar en la tabla de la distribución normal acumulada:

$$P(-2{,}38 \leq z \leq 2{,}38) = P(\mu - 0{,}5 < \bar{x} < \mu + 0{,}5) = 0{,}9913 - 0{,}0087 = 0{,}9826$$

Respuesta: la probabilidad de que la media muestral esté a 0,5 kg (o menos) de la media poblacional es 0,9826. El porcentaje es del 98,26 %.

Ahora es el turno de resolver la pregunta b). Acá se quiere saber cuál será la probabilidad de que el consumo medio de la muestra supere al consumo promedio poblacional en más de 0,75 kg. Si la media poblacional es desconocida, podemos plantear lo siguiente:

- Se quiere averiguar $P[(\bar{x} - \mu) > 0{,}75]$

Entonces, se debe estandarizar. Como la media muestral debe ser mayor a la media poblacional en 0,75, (por lo menos), se puede considerar lo siguiente:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{0{,}75}{\frac{\sigma}{\sqrt{n}}} = \frac{0{,}75}{\frac{1{,}25}{\sqrt{36}}} = \frac{0{,}75}{0{,}21} = 3{,}57$$

La P(z>3,57)=0, pues se consideran las desviaciones estándar hasta 3,5 aproximadamente, más allá las probabilidades son nulas. Hay que recordar que la tabla da P(z<3,57)=1. Entonces, se tiene que restar a 1, para obtener la probabilidad buscada: 1 - 1 = 0.

Respuesta: la probabilidad de que la muestra supere a la media poblacional en 0,75 kg es 0. El porcentaje es 0 %.

En el caso de la productora de té, si la media poblacional se conoce y es de 300 kg por familia por año, con una desviación estándar poblacional de 82,2 kg, ¿cuál es la probabilidad de que, al elegir una muestra aleatoria de 64 familias, la media muestral obtenida sea menor a 270 kg?

- La probabilidad de que el consumo medio de las 64 familias sea menor a 270 kg es de 0,0018.
- La probabilidad de que el consumo medio de una de las familias de la ciudad sea menor a 270 kg es de 0,0018.
- La probabilidad de que el consumo medio de las 64 familias sea menor a 270 kg es 0, pues el error estándar es menor a la desviación poblacional.
- La probabilidad de que el consumo medio de las 64 familias sea menor a 82,2 kg es de 0,0018.
- La probabilidad de que el total de consumos de las 64 familias sea menor a 270 kg es de 0,018.

SUBMIT

Devolución a la actividad

Justificación: datos: $\mu = 300$; $\sigma = 82{,}2$; n=64 ; Calcular $P(\bar{x} < 270)$

Estandarización: $\bar{x} = 270$. Entonces $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{270-300}{\frac{82{,}2}{\sqrt{64}}} = \frac{-30}{10{,}275} = -2{,}92$

Según la tabla de **la normal**: $P(z < -2{,}92) = P(\bar{x} < 270) = 0{,}0018$

Distribución de muestreo para la proporción

¿Qué es una distribución muestral para la proporción?

En muchas oportunidades es necesario determinar una proporción en una población. Por ejemplo:

- porcentaje o proporción de votantes que se estima votarán en la próxima elección por un cierto candidato,
- porcentajes de alumnos del nivel primario que no terminan el ciclo,
- porcentaje de enfermos con HIV,
- porcentaje de niños de una determinada edad que estén vacunados contra el sarampión.

Se trata ahora de extraer una muestra, calcular la proporción de la característica de interés en dicha muestra y analizar qué sucede en la población con esa característica, con base en los datos muestrales.

La fórmula para calcular la proporción muestral es:

$$\bar{p} = \frac{x}{n}$$

x = número de elementos de la muestra que poseen la característica de interés.

n = tamaño de la muestra.

La proporción muestral $\bar{p}$ es una variable aleatoria y su distribución de probabilidad se conoce como **distribución muestral de $\bar{p}$**

Propiedades de la distribución muestral de la proporción. Aplicación al caso de la productora de té

Al igual que en la distribución de medias, para determinar cuán cerca está la proporción de una muestra

$\bar{p}$.

de la proporción poblacional p, se necesita saber cuáles son las propiedades de la distribución muestral de

$\bar{p}$:

- el valor esperado de $\bar{p}$,
- la desviación estándar de $\bar{p}$ y
- la forma de la distribución muestral de $\bar{p}$.

Valor esperado o media de la distribución muestral de la proporción P

El valor esperado de $\bar{p}$ es la media de todos los posibles valores de $\bar{p}$. Que, a su vez, es igual a la proporción poblacional p.

$$E(\bar{p}) = p$$

En el caso de la productora de té el resultado es 0,60.

Desviación estándar de la proporción muestral

También llamado error estándar de la proporción.

Al igual que en la distribución de medias muestrales, se distingue entre poblaciones infinitas y finitas.

- Error estándar de la proporción para poblaciones infinitas:

$$\sigma_{\bar{p}} = \sqrt{\frac{p.q}{n}}.$$

p: proporción en que se encuentra la característica en estudio en la población.
Q = 1-p.
n= tamaño de la muestra.

- Error estándar de la proporción para poblaciones finitas:

$$\sigma_{\bar{p}} = \sqrt{\frac{p.q}{n}}.\sqrt{\frac{N-n}{N-1}},$$

se le agrega el factor de corrección.

Para el caso de la productora de té, se calcula el error estándar de la distribución de la proporción de todas las familias que consumen la marca de té Ackerman. Como se trata de una población infinita frente a la muestra (es una muestra de 36 familias, representativa de todas las familias argentinas), no se agregará el factor de corrección:

$$\sigma_{\bar{p}} = \sqrt{\frac{p.q}{n}} = \sqrt{\frac{0{,}60 \times 0{,}40}{36}} = 0{,}0816$$

El error estándar de la distribución de muestras de la proporción de clientes que consumen la marca de té Ackerman es de 0,0816.

Forma de la distribución muestral de la proporción

Se conocen la media y la desviación estándar de la distribución muestral de $\bar{p}$. Ahora queda por analizar la forma de esa distribución muestral.

La fórmula para calcular la proporción muestral es $\bar{p} = \frac{x}{n}$.

Se sabe que x es una variable aleatoria binomial que indica el número de los elementos de la muestra que tienen la característica de interés. La muestra fue tomada de forma aleatoria de una población grande (ver *Características de la distribución binomial* en la lectura 3). Los eventos se consideran independientes y, como n es una constante, la probabilidad de x/n es la misma que la probabilidad de x. Esto significa que la distribución muestral de $\bar{p}$ también es una distribución de probabilidad discreta y que la probabilidad de cada x/n es la misma que la probabilidad de x.

En el módulo 3, se estudia que una distribución binomial se aproxima mediante una distribución normal, siempre que el tamaño de la muestra sea lo suficientemente grande para satisfacer las siguientes condiciones:

$$n.p \geq 5 \; y \; n.q \geq 5$$

Se afirma, entonces, que la distribución muestral de $\bar{p}$ se aproxima mediante una distribución normal siempre:

$$n.p \geq 5 \; y \; n.q \geq 5$$

Estos conceptos se aplican, ahora, al caso de la productora de té:

- La proporción poblacional de las familias que consumen el té de la productora Ackerman es de 0,60.
- Con una muestra aleatoria de $n = 36$:

$$n.p = 36 \times 0{,}60 = 21{,}6 \; que \; es > 5$$

- Además, debe cumplirse:

$$n.q = 36 \times 0{,}40 = 14{,}4 \; que \; es > 5$$

Por tanto, la distribución muestral se calcula mediante la distribución normal.

La proporción y el teorema del límite central

A continuación, se recapitula lo estudiado sobre distribución muestral de la proporción en los siguientes conceptos:

- La proporción se encuentra en una población como un nuevo parámetro, el que se determinará en la mayoría de los casos, con una inferencia a través del estadístico correspondiente de una muestra.

- El valor práctico de la distribución muestral de

$$\bar{p}$$

es que permite obtener información probabilística acerca de la diferencia entre la proporción muestral y la proporción poblacional.

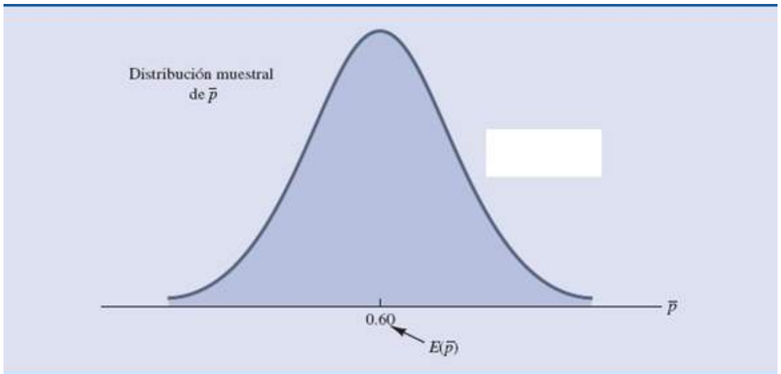
- Para la obtención y el análisis del estadístico, se tendrá presente que el teorema del límite central puede ser aplicado para las proporciones y, por lo tanto, si de una población con una proporción p (de los elementos con la característica en estudio), se extraen todas las muestras posibles del mismo tamaño y de cada muestra se determina la proporción; con las proporciones de todas las muestras posibles se genera una distribución de proporciones muestrales, la que será normal, ajustándose a las pautas ya mencionadas de las medias muestrales. Además, la media de esa distribución de proporciones será igual a la proporción de la población y el desvío estándar de la distribución de proporciones muestrales será igual a

$$\sigma_{\bar{p}} = \sqrt{\frac{p.q}{n}}$$

Resolución del caso de la productora de té, pregunta c)

De acuerdo con las propiedades descriptas anteriormente, se visualiza el caso de la productora de té en la figura 5.

Figura 5: Distribución muestral de p para el caso de la productora de té



Fuente: elaboración propia, con base en Anderson, D. R., Sweeney, D. J. y Williams, T. A., 2008, p. 282.

Descripción de la figura. Distribución muestral de para el caso de la productora de té. Se muestran las propiedades de una distribución muestral de la proporción: la forma normal, la media y el error estándar.

Conociendo la media y el error estándar de la muestra, hay que averiguar la proporción de las familias que consumen la marca de té y que no difiera de la media poblacional en más de ± 5 %.

Primero se toma la muestra, se calculan la proporción de interés y el error estándar, para comenzar a sacar conclusiones sobre la proporción de la población.

Si se desea que el estadístico proporción muestral no difiera de la media poblacional más allá del 5 %, es decir de 0,05, entonces se debe calcular la probabilidad de que la proporción poblacional esté dentro del intervalo 0,60 - 0,05; 0,60 + 0,05 o, lo que es lo mismo, 0,55; 0,65.

Dicho de otra forma, ¿cuál es la probabilidad de obtener una muestra en la que el valor de

$$\bar{p}$$

esté entre 0,55 y 0,65?

Como la forma de la distribución de la muestra de la proporción se puede aproximar a la normal, se aplica la estandarización a la proporción muestral:

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}}$$

$\bar{p}$: es cualquier proporción de la distribución de proporciones de la muestra. En el caso de estudio, son los límites del intervalo $p \pm 0{,}05$.

p: es la proporción de la población.

$\sigma_{\bar{p}}$: es el error muestral de la distribución de proporciones de la muestra. En este caso 0,0816.

Se comienza estandarizando $\bar{p} = 0,65$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0,65 - 0,60}{0,0816} = 0,61$$

Luego se estandariza $\bar{p} = 0,55$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0,55 - 0,60}{0,0816} = -0,61$$

En la tabla de probabilidad normal estándar aparece que la probabilidad acumulada que corresponde a z = 0,61 es 0,7291. En la misma tabla, la probabilidad correspondiente a z = - 0,61 es 0,2709. De esta manera, la probabilidad de seleccionar una muestra en la cual el valor de

$\bar{p}$

no difiera más de 0.05 de la proporción poblacional p está dada por 0,7291 - 0,2709 = 0,4582. O como se dice cotidianamente es del 45,82 %.

Importante: se puede verificar, analizando las fórmulas respectivas, que si se aumenta el tamaño de la muestra, el error estándar de la proporción disminuye y la probabilidad de que la proporción muestral no esté más allá del 5 % de la proporción de la población aumenta.

Referencias

Anderson, D. R., Sweeney, D. J. y Williams, T. A. (2008). Muestreo y distribuciones muestrales en Anderson, D. R., Sweeney, D. J. y Williams, T. A., Estadística para Administración y Economía (pp. 267-285). México: Ed. Cengage Learning Editores, S.A.

Levin, R. y Rubin, D. (2004). Estadística para Administración y Economía (7ª edición). México: Pearson.