

3. CONTAR. ESTIMAR INTERVALOS DE CONFIANZA

3.1 Contar y estimar intervalos de confianza

La toma de decisiones es algo que, inevitablemente, debemos enfrentarnos día a día. Es clave contar con herramientas que nos ayuden a realizar una acción por sobre otra. El objetivo de esta parte, frente a la información estadística, es poder tomar, con fundamentos, un camino u otro.

“La única prueba relevante de la validez de una hipótesis es la comparación de sus predicciones con la experiencia” (Milton Friedman Friedman)

Introducción

En esta instancia, se profundizará sobre la estadística inferencial basada en el modelo de probabilidades, lo que significa que los datos sobre la muestra estudiada han sido generados mediante algún mecanismo aleatorio. En ese sentido, se repasarán los conceptos de qué es contar y se introducirá, en esta instancia, cuánto contar. Esto nos servirá como materia prima para trabajar sobre los grados de confianza a la hora de hacer inferencias sobre poblaciones, basadas en datos de una muestra y establecer diferencias entre muestras.

Contar: ¿cómo y cuánto?

Contar es algo que parece común, intuitivo y fácil. ¿Es esto siempre así? Hay casos que contar no es solamente decir cuántas personas realizaron una actividad determinada. Por ejemplo, si hay que elegir para armar un equipo con 4 jugadores que están alineados... ¿cuántas maneras distintas existen

para hacer dicha selección desde el primero al último? Esta respuesta no es trivial y requiere cierto grado de análisis. En este caso, son 4 opciones para la primera posición, 4 – 1 opciones para la segunda, 4 – 2 opciones para la tercera, 4 – 3 opciones para la cuarta. De esta manera, el número de posibilidades será $\#posiciones = 4 \times (4 - 1) \times (4 - 2) \times (4 - 3)$. Este tipo de expresión se conoce como factorial del número; en este caso, 4 (se representa como 4!), que se entiende como el número de permutaciones de n elementos, esto es: $4! = 1 \times 2 \times 3 \times 4 = 24$.

En este sentido, la siguiente tabla representa y resume diferentes formas de contar, dependiendo la situación.

Tabla 1: Resumen de diferentes formas de contar

Nombre	Ecuación	Importa el orden	Se repiten elementos	Denominación
Permutaciones ordinarias	$n!$	Sí	No	Permutaciones de n elementos.
Variaciones	$\frac{n!}{(n - k)!}$	Sí	No	Variaciones ordinarias de n elementos tomados de k en k .
Variaciones con repetición	n^k	Sí	Sí	Variaciones con repetición de n elementos tomados de k en k .
Combinaciones	$\frac{n!}{k!(n - k)!}$	No	No	Combinaciones de n elementos tomados de k en k .
Combinaciones con repetición	$\frac{(n + k - 1)!}{k!(n - 1)!}$	No	Sí	Combinaciones con repetición de n elementos tomados de k en k .

Fuente: Elaboración propia.

Tamaño de muestra para estimar la media de una población

Uno de los principales temas, y no tan evidente, es el hecho de cuántos datos son necesarios recabar en estadística inferencial para que el parámetro estadístico que es preciso determinar sea óptimo para poder realizar una conclusión o comparación exacta y confiable (error mínimo). A esto se lo denomina tamaño de la muestra.

Si, por ejemplo, se desea calcular el valor medio de una muestra que presenta una distribución normal y con una desviación estándar determinada, ¿cuál será el número mínimo de datos que se deben registrar? La respuesta a esta pregunta dependerá de *fiabilidad* y *exactitud* que se deseen para la medida. En la estimación estadística, la exactitud se mide por los márgenes

de error o el intervalo de confianza definido para una determinada fiabilidad. Como ambas medidas dependen del tamaño de la muestra, especificarlas permite determinar el tamaño de esta.

Por ejemplo, se desea estimar el rendimiento promedio diario (μ) para una empresa productora de papel donde se necesita que el error (E) por determinar sea menor a 5 toneladas y con una fiabilidad del 95 % (probabilidad $z = 1,96$). Esto, suponiendo una distribución normal para el proceso de producción, significa que, en el 95 % del tiempo en la toma de datos (fiabilidad), la distancia entre la media de la muestra \bar{x} y de la población (μ) será menor a $1,96 \sigma_{\bar{x}}$, donde $\sigma_{\bar{x}}$ es el error estándar de la media y es igual a $\frac{\sigma}{\sqrt{n}}$ (con σ la desviación estándar poblacional). Entonces, lo que se busca es el número tal de muestras donde se cumpla que $1,96 \sigma_{\bar{x}} < 5$, o bien reemplazando $1,96 \sigma_{\bar{x}} = 1,96 \frac{\sigma}{\sqrt{n}} < 5$. De esta manera, se puede deducir n como $\left(\frac{z\sigma}{E}\right)^2 = \left(\frac{1,96}{5}\right)^2 \sigma^2 < n$ ó $0,1537 \sigma^2 < n$. Suponiendo que la compañía ha determinado, midiendo el proceso, una desviación estándar de 20 toneladas, entonces n será igual a $0,1537 \sigma^2 = 0,1537 20^2 = 62$. ¿Qué significa este valor? Que, usando un tamaño de muestra de 62 datos o mayor, la empresa se asegura, con una certeza razonable (probabilidad del 95 %), que la estimación del rendimiento promedio del proceso estará dentro de las 5 toneladas respecto al promedio actual.

Es común que, en realidad, σ no sea conocida y, por ende, hay que, de alguna manera, determinarlo o estimarlo. Muchas veces lo que generalmente se hace es realizar un análisis sobre una muestra reducida y determinar la desviación estándar para la muestra y usarla en lugar de σ (o sea, se considera la desviación estándar de la muestra igual a la de la población).

Tamaño de muestra para proporciones

La forma general para estimar un intervalo de la proporción de una población p es $\bar{p} \pm \text{margen de error}$; para ello, la distribución de la muestra es fundamental para el cálculo del margen de error. Considerando una distribución normal la determinación del tamaño de la muestra y similar a lo visto para estimar la media poblacional, se puede plantear que el margen de error (E) asociado a un intervalo en la proporción de la población

viene dado por $E = z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$. El margen de error dependerá del valor de $z_{\alpha/2}$ por considerar (confiabilidad), de \bar{p} y del tamaño de la muestra, lo que será, evidentemente, menor a mayor número de datos.

Entonces, se puede decir que el número mínimo, n , de datos por recabar en una muestra será $n = \frac{z_{\alpha/2}^2 \bar{p}(1-\bar{p})}{E^2}$, donde se tendrá un nivel de confianza de

$z_{\alpha/2}$ y un error E . En la práctica, el valor de \bar{p} puede ser adoptado siguiendo alguno de los siguientes métodos:

- usar la proporción de una muestra anterior con unidades similares;
- al igual que para la población, se puede realizar un estudio de prueba seleccionando una muestra reducida;
- frente a la experiencia del analista, tomar un valor predictivo de \bar{p} ;
- adoptar un valor arbitrario de \bar{p} (generalmente, un buen comienzo es el valor 0,50).

Un ejemplo práctico puede ser una encuesta que desea realizar una determinada universidad sobre sus estudiantes y la satisfacción que muestran a la hora del almuerzo. ¿Cuál debe ser el tamaño de la muestra si el decano de dicha institución desea estimar la proporción de la población con un margen de error del 0,035 y una fiabilidad del 95 % si -sobre 872 encuestados- 248 mostraron un alto nivel de satisfacción? En este caso, no se tiene toda la población evaluada, sino una muestra. En este caso, entonces, se puede calcular $\bar{p} = 248/872 = 0,2844$. Ahora, con este valor, se puede calcular $n = (z_{\alpha/2})^2 \bar{p}(1-\bar{p}) / E^2 = ([1,96])^2 0,2844 (1-0,2844) / [0,035]^2 = 638$. En este sentido, la muestra debería contar con, al menos, 638 encuestados. Si se utiliza la opción de considerar una proporción de 0,50, el número de encuestados debería ser, al menos, de 784.

Intervalos de confianza

La necesidad de la estadística se hace evidente cuando, basado en un conjunto de datos, se trata de hacer algún pronóstico o inferencia sobre una población mayor. En esta parte se discutirá cómo basar estas inferencias considerando el modelo probabilístico que implica que los datos son generados o adquiridos de una forma aleatoria. Generalmente, se pueden utilizar dos mecanismos, a saber:

1. **Muestreo al azar reducido de una población grande:** a este caso se lo puede correlacionar con una empresa de ventas por Internet que tiene miles de clientes y desea saber el tiempo que tardan en elegir un producto. Entonces, la empresa selecciona y evalúa, aleatoriamente, el comportamiento de alguno de sus clientes para inferir sobre el comportamiento global de la población.
2. **Experimento aleatorio:** un ejemplo de esta situación es cuando una compañía está considerando dos maneras diferentes de premiar a sus trabajadores para obtener como resultado una mayor productividad

por empleado. Para esto, la empresa realizó un “experimento” donde les aplicó *aleatoriamente* a distintos trabajadores uno u otro mecanismo de premio durante un período determinado y midió su productividad. Luego, evidentemente, pretenderá inferir cuál de los mecanismos utilizados es el que mejor efecto surtió para, así, implementarlo al resto del personal. En este caso, se debería determinar la diferencia entre las productividades medias de *todos* trabajadores para cada mecanismo de premio. Este procedimiento es extremadamente difícil de llevar a cabo y, por este motivo, se han tomado dos sub-grupos de trabajadores para cada situación. El hecho de que esto se haya realizado de manera azarosa permite descartar varios factores que pueden surgir de seleccionar grupos en una población (por ejemplo, al ser al azar, se asegura que la proporción de géneros se mantiene para cada grupo) y, también, es posible aplicar las reglas de las probabilidades para inferir si las diferencias que surjan de cada situación se pueden trasladar al total del personal.

En estadística se utilizan, principalmente, los siguientes análisis para inferir sobre poblaciones:

- la determinación del intervalo de confianza;
- las pruebas de hipótesis.

Distribuciones de muestreo

Valor z

Nuevamente, comenzaremos con la media muestral, \bar{x} . El teorema del límite central (lectura 2 de este curso) dice que, si el tamaño de una muestra es lo suficientemente grande, para cualquier distribución de la población, la distribución del muestreo de \bar{x} será aproximada a la normal con una media poblacional, μ , y una desviación estándar de la población igual a σ/\sqrt{n} . Uno puede definir un parámetro de estandarización, z , también conocido como z -score igual a $z=(\bar{x}-\mu)/(\sigma/\sqrt{n})$, el cual es aproximadamente normal con una media igual a 0 y una desviación estándar igual a 1. Como hemos discutido anteriormente, esto es utilizado para hacer inferencias acerca de la media poblacional, pero el problema es que, generalmente, la desviación estándar de la población nunca se conoce. Y, nuevamente, mencionamos que, en general, se puede reemplazar σ por s , que es la desviación estándar de la muestra. El inconveniente de esta consideración es que se introduce una nueva fuente de variabilidad y puede ocasionar que ahora el muestreo no presente una distribución normal. Esto se denomina distribución t , la cual es una pariente cercana de la distribución normal, que es ampliamente utilizada en diversas aplicaciones estadísticas.

Distribución t , valor t o prueba de student

Este análisis se basa en la estimación de la media poblacional, μ , con un tamaño de muestra, n . Se asume que la población tiene una distribución normal con una desviación estándar poblacional, σ , desconocida. Ahora se plantea basar las inferencias sobre el valor estandarizado de \bar{x} , donde σ ahora es reemplazado por s . Esto define el valor de distribución estándar t como $t = (\bar{x} - \mu) / (s / \sqrt{n})$.

En este caso, el valor t o parámetro t de student tiene una distribución del tipo t con un grado de libertad $n-1$ el cual define la forma de la curva de distribución muestral o poblacional. En cada caso que se trabaje con una distribución t se deben especificar los grados de libertad. En este contexto de muestreo, como estamos basando inferencias sobre μ desde la distribución de una muestra \bar{x} , se pierde un grado de libertad para un tamaño de muestra n .

El valor t representa el número de errores estándar por los cuales la media muestral difiere de la media poblacional. Por ejemplo, si $t=2$, la media de la muestra es 2 errores estándar por encima de la media poblacional. Si $t=-2$, esto dice que la media muestral está 2 errores estándar por debajo de la media poblacional.

Intervalos de fiabilidad o confianza

Para la media, con σ conocido

El intervalo de confianza (IC) para un parámetro de la población es un intervalo de valores construido de una manera que, con un grado de fiabilidad dado, el parámetro de la población está comprendido dentro él. En este sentido, surge un coeficiente de confianza que denota la probabilidad que IC contenga al parámetro de la población en muestreos repetidos. El valor porcentual de este coeficiente es el nivel de fiabilidad deseado. El objetivo es poder inferir los valores de uno o varios parámetros de una población como la media, la desviación estándar y/o la proporción muestral.

Para obtener un IC para μ , se debe especificar el nivel de fiabilidad deseado. En general, los niveles utilizados, dependiendo del análisis por realizar, son del 90 %, 95 % y 99 %. Esto permite determinar el múltiplo del error estándar, lo cual define el intervalo de confianza como $IC = \text{parámetro estándar} * \text{error estándar} (\bar{x})$. Luego, la expresión para denotar este intervalo será $\mu \pm IC$. El valor de μ , generalmente, no se conoce, ya que, en estadística, se utiliza una muestra aleatoria, la cual es parte de la población. Por este motivo, es razonable, cuando el tamaño de la muestra es adecuado, utilizar la media muestral, \bar{x} . Cuando se conoce el valor de la desviación estándar de la población, se lo define como el intervalo de confianza para la media como $\bar{x} \pm z_{(\alpha/2)} \sigma / \sqrt{n}$.

Para la media, con σ desconocido

Pero, como se ha descrito arriba, generalmente el valor de la desviación estándar poblacional no se conoce. Por este motivo, es necesario considerar que el muestreo presente una distribución del tipo t. Para obtener el valor indicado de t, consideremos a α igual a 1 menos el nivel de fiabilidad (expresado como decimales). Por ejemplo, si se plantea un nivel del 90 % (en decimales 0,90) de fiabilidad, α será igual a 0,10. Entonces, el valor de t es aquel que excluye a cada cola de la distribución t para la probabilidad con un valor igual a $\alpha/2$ (dos colas de la campana) con n-1 grados de libertad. Esto se esquematiza en la figura 1.

Figura 1: Distribución t con una probabilidad o área igual a $\alpha/2$ en la cola superior



Fuente: Elaboración propia.

Similar a lo considerado para el IC de media, con desviación estándar conocida, cuando σ no se conoce, se puede utilizar, en su lugar, la desviación estándar de la muestra, s ; de esta manera, la expresión para el IC será $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$.

Cabe recalcar que, a mayor nivel de fiabilidad, el ancho del intervalo de confianza también aumenta, donde lo deseable es que sea lo más angosto posible. En este sentido, hay un compromiso entre el nivel de fiabilidad y el IC. Cuando se trabajan con niveles de confianza altos, lo ideal es aumentar el tamaño de la muestra, ya que el valor t tiende a disminuir.

Para el total

Existen situaciones donde la media poblacional no es el parámetro de mayor interés sobre el cual uno desea basar alguna decisión o estrategia. A veces, es necesario determinar el IC para el *total de la población*. En este sentido,

definiendo a T como el total poblacional y a \hat{T} como el punto estimativo de T basado en una muestra de tamaño n y tomada al azar, es razonable sumar todos los valores en la muestra, T_s , y luego *proyectar* este total a la población entera de la manera $\hat{T} = \frac{N}{n} T_s = N\bar{x}$. La segunda igualdad ($N\bar{x}$) se debe a que la sumatoria total, T_s , dividida por el tamaño de la muestra, n , es igual a la media de la muestra, \bar{x} .

Al igual que \bar{x} , \hat{T} también presentará una distribución muestral con una media, $E_{(\hat{T})}$, igual a $E_{(\hat{T})} = T$, y un error estándar, ES , igual a $ES_{(\hat{T})} = N\sigma/\sqrt{n}$. Como, en general, se desconoce el valor de σ , se utiliza, en su lugar, s (desviación estándar de la muestra) para el cálculo del error estándar del total, lo cual queda como $ES_{(\hat{T})} = \frac{N\sigma}{\sqrt{n}} = N ES_{(\bar{x})}$.

El valor \hat{T} es un estimador *imparcial* de T , lo que implica que no habrá una tendencia de sub o sobreestimar el total de la población.

De estas definiciones, el intervalo de confianza para el total poblacional se puede determinar mediante el cálculo del IC para la muestra, tal como se describió anteriormente, y multiplicar cada extremo del intervalo por el tamaño de la población, N .

Para la proporción

Normalmente, las encuestas cotidianas que vemos se basan en estimar proporciones, como, por ejemplo, la proporción de personas socias de un club deportivo que está de acuerdo con cómo se manejan las políticas económicas del club. Por este motivo, es necesario saber cómo calcular el IC para cualquier población con una proporción p . Nuevamente, el procedimiento es similar al realizado para la media poblacional. Requiere de un parámetro estimativo conocer el error estándar de dicho parámetro y un múltiplo, el cual dependerá del nivel de fiabilidad deseado. Esto deja una expresión para el nivel de fiabilidad del tipo *parámetro estimativo \pm multiplicador \times error estándar*.

Si definimos a A como cualquier propiedad de los miembros de una población y p como la proporción de la población con la propiedad A , $p = A/N$. Ahora, si se obtienen datos de una muestra de la población con un tamaño n , la proporción de la muestra será $\hat{p} = A/n$, la cual puede ser utilizada como parámetro estimador de p . Por ejemplo, si 15 personas de una muestra de 60 miembros del club deportivo están de acuerdo con sus políticas económicas, la proporción de personas con la característica A (de acuerdo) será $\hat{p} = \frac{15}{60} = 0,25$.

Para una muestra con un tamaño suficiente (n alto), la distribución de \hat{p} es aproximadamente normal con una media p y un error estándar igual a $\sqrt{p(1-p)/n}$. Y como p es un parámetro desconocido, se reemplaza por \hat{p} , por lo que queda como $ES_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$.

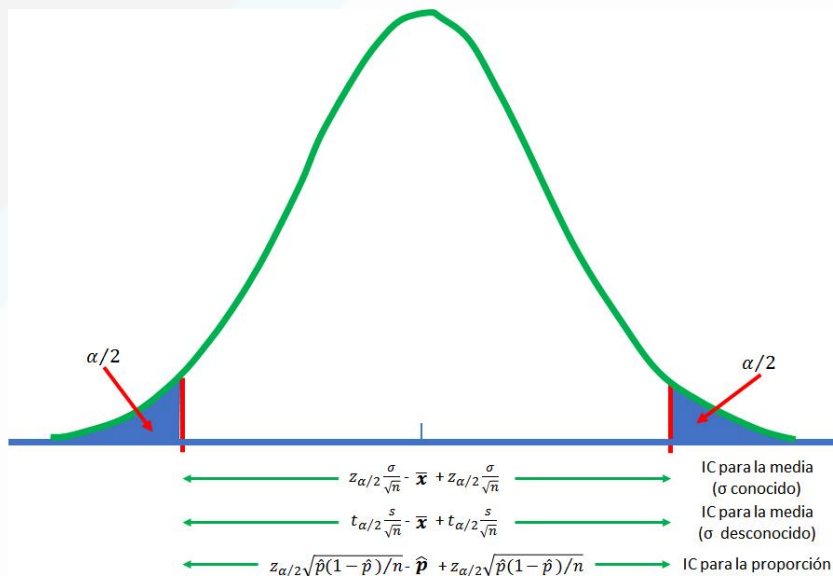
Finalmente, el multiplicador usado para el cálculo del IC para p es el valor z (no el valor t). Como ya hemos visto, el valor de z para un nivel de confianza del 95 % es igual a 1,96, ya que este es el valor donde la probabilidad es del 0,025 para cada cola de la distribución normal (figura 1).

En resumen, el intervalo de confianza tiene una expresión:

$$\hat{p} \pm \text{valor } z \times \sqrt{\hat{p}(1-\hat{p})/n}.$$

En la figura 2, se resumen los distintos intervalos de confianza con sus respectivas expresiones matemáticas.

Figura 2: Representación de los intervalos de confianza descriptos



Fuente: Elaboración propia.

Ejemplos

Para hallar los valores de $z_{\alpha/2}$, se utilizará la función de Excel “=INV.NORM.ESTAND (probabilidad)” (en la versión inglesa, “NORM.S.INV (probability)”). En probabilidad, se debe asignar el nivel de confianza deseado. Si, por ejemplo, en una distribución normal con dos colas se plantea un 95 % de fiabilidad, el límite superior acumulativo será $0,95 + \alpha/2$ (donde $\alpha = 1 - IC$). En este caso, será $0,95 + \frac{1-0,95}{2} = 0,975$ y la función en Excel se completa como =INV.NORM.ESTAND (0,975), de lo cual se obtiene el valor de $z_{\alpha/2}$ igual a 1,96. El límite inferior (igual a $\alpha/2$) es -1,96 (que se obtiene con la función =INV.NORM.ESTAND (0,025)). De una manera similar, para obtener el valor de $t_{\alpha/2}$, se utilizará la función de Excel “=DISTR.T.INV(probabilidad, grados de libertad)” o, en inglés, “=T.INV(probability, freedom degree)”, donde, para el límite superior, probabilidad será el valor del nivel de confianza deseado más $\alpha/2$ (donde $\alpha = 1 - IC$), en tanto que para el límite inferior será $\alpha/2$. Los grados de libertad, como hemos visto, son iguales al tamaño de la muestra menos 1 ($n - 1$). Así, el valor de superior e inferior $t_{\alpha/2}$, para el nivel de confianza del 95 % de una muestra de 28 mediciones, se obtiene en Excel aplicando la fórmula =DISTR.T.INV(0,975, 27) y =DISTR.T.INV(0,025, 27), respectivamente. En este caso, los valores serán $t_{\alpha/2; sup} = 2,05$ y $t_{\alpha/2; inf} = -2,05$.

1. **IC para la media:** un estudio con 130 observaciones determinó que el tiempo medio para finalizar un examen en una universidad fue de

98,2 minutos. La desviación estándar calculada de estas observaciones es igual a 0,73. Para encontrar el intervalo de confianza para un nivel de fiabilidad del 95 % para la media muestral de 98,2 minutos, primero debemos encontrar el valor $t_{\alpha/2}$ para 129 grados de libertad. Los valores son de 1,98 y -1,98 para los límites superior e inferior, respectivamente. El error estándar de la media muestral es igual a $\frac{s}{\sqrt{n}} = \frac{0,73}{\sqrt{130}} = 0,064$; de esta manera, el intervalo de confianza para el estudio realizado será $98,2 \pm 1,98 \times 0,064 = 98,2 \pm 0,1$.

2. **IC para proporciones:** un relevamiento al azar mostró que, de 59 jugadores de fútbol, 15 patean con la pierna izquierda. Esto nos dice que la proporción, \hat{p} , de jugadores izquierdos es $\hat{p} = \frac{15}{59} = 0,2542$ (25 %). Si deseamos estimar, para una fiabilidad del 95 %, en qué porcentaje estarán comprendidos los jugadores izquierdos en toda la liga de fútbol, debemos calcular el error estándar para la población ($ES_{\hat{p}}$) multiplicado por el valor $z_{\alpha/2}$ para el nivel de confianza del 95 %. Esto es igual a $z_{\alpha/2} \times ES_{\hat{p}} = \pm 1,96 \sqrt{\hat{p}(1 - \hat{p})/n} = \pm 1,96 \sqrt{0,2542 \times \frac{1-0,2542}{59}} = 0,111$. De esta manera, el intervalo de confianza será de $0,2542 \pm 0,1111$, lo cual indica que, con una fiabilidad del 95 %, el porcentaje de jugadores izquierdos está comprendido entre un 14 % y un 37 % (límite inferior: $0,2542 - 0,1111 = 0,1431$; límite superior: $0,2542 + 0,1111 = 0,3653$).

3.2 Intervalos de confianza para diferencias

Intervalos de confianza para la diferencia entre medias

Una de las aplicaciones más importantes de la estadística inferencial es la comparación de dos medias poblacionales. Los siguientes ejemplos ilustran esta parte de la estadística:

- El dueño de un local de vestimentas desea conocer cuánto más o menos gasta en promedio una mujer en comparación con un varón.
- Una organización de clientes desea conocer cuál es el promedio de atraso en los arribos para dos empresas aéreas que poseen rutas de vuelo similares.
- Una compañía de telefonía tiene su servicio al cliente para responder a preguntas y quejas de los clientes. Su personal está compuesto por un primer grupo, el cual tiene poca experiencia en atención al consumidor, pero que ha realizado un curso de cómo tratar con clientes, mientras que el segundo grupo se trata de empleados con vasta experiencia, pero que no ha realizado ningún tipo de especialización. La empresa desea evaluar diferencias entre ambos grupos midiendo el promedio de quejas de los clientes debido al desempeño del servicio de atención durante el último trimestre.
- Una empresa de venta de automóviles desea ver cómo es la reacción de esposas y maridos frente a la presentación de los productos. Ambos, marido y mujer, son encuestados, por separado, acerca de la calidad de la presentación de los productos.

Estos ejemplos se basan en la comparación de la diferencia entre promedios (medias) de dos poblaciones. Pero existe una diferencia entre los tres primeros casos y el último. En el caso de la reacción del marido y la mujer, existe una dependencia entre muestras, ya que no se eligen “maridos” o “esposas” al azar, sino que cada uno está ligado al otro. A este tipo de muestras se las denomina dependientes, en tanto que en los otros casos se trata de muestras independientes.

Muestras independientes

Cuando se desea estimar la diferencia entre dos medias poblacionales μ_1 y μ_2 , que presentarán sus respectivas desviaciones estándar σ_1 y σ_2 , se pueden tomar aleatoriamente muestras de tamaño n_1 y n_2 , no necesariamente iguales, e inferir sobre la diferencia poblacional basada en las diferencias de las medias de las muestras, \bar{x}_1 y \bar{x}_2 . Sobre la base de esta estimación, se

desea dar un intervalo de confianza para la diferencia de las medias poblacionales, la cual es incógnita. En este caso, la distribución de muestra de la diferencia de las medias muestrales presenta una distribución del tipo t , donde los grados de libertad vienen dados por $n_1 + n_2 - 2$. En este caso, el intervalo de confianza se representa como:

$$\bar{x}_1 - \bar{x}_2 \pm \text{valor } t \times \text{error estándar} (\bar{x}_1 - \bar{x}_2).$$

Como se ha mencionado, el parámetro t de student es el valor que corta la probabilidad, en donde cada cola de la distribución t (dependiendo del nivel de fiabilidad deseado) presentará $n_1 + n_2 - 2$ grados de libertad (esto asume que las distribuciones de las poblaciones son normales y los tamaños de las muestras son razonablemente grandes). Por ejemplo, para un nivel de confianza del 95 % y $n_1 = n_2 = 40$, el parámetro t de student será 1,99. Este valor se puede obtener utilizando la función de Excel “=DISTR.T.INV(0,975, 78)” (o, en inglés, “=T.INV(0,975, 78)”), donde 0,975 será la probabilidad acumulativa hasta $\alpha/2$ y 78 los grados de libertad.

Para el cálculo del *error estándar* ($\bar{x}_1 - \bar{x}_2$), se puede asumir que las desviaciones estándar poblacionales son iguales, $\sigma_1 = \sigma_2$, lo cual permite calcular la desviación estándar agrupada de la muestra, S_p , como

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}, \text{ esto es, con } S_1 \text{ y } S_2 \text{ las desviaciones estándar de las dos muestras. De esta función se deduce que el error estándar de } \bar{x}_1 - \bar{x}_2 \text{ es igual: } \text{error estándar} (\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Ejemplo de aplicación

Se plantea si hay diferencia entre los patrones de viaje a la universidad entre estudiantes y profesores. Para ello, un estudio midió el promedio de kilómetros recorridos por 11 estudiantes, el cual fue de 5,6 km, con una desviación estándar de 2,8 km. Por otro lado, 14 profesores mostraron un promedio de 14,3 km recorridos, con una desviación estándar de 9,1.

¿Cuál es el intervalo de confianza, con un nivel de fidelidad del 95 %, para la diferencia entre los promedios de kilómetros recorridos por alumnos y profesores?

La tabla 2 muestra los datos con el respectivo análisis para el problema planteado.

Tabla 2: Problema de aplicación para la determinación de intervalo de confianza para la diferencia de medias muestrales

	Alumnos	Profesores	Determinación de intervalos de confianza	
Tamaño muestra	12	15	Nivel de confianza	95 %
Media muestra	6,1	15,2	Diferencia medias	9.1
Desviación estándar Muestra	2,7	8,8	Grados de libertad	25
$(n_i - 1)s_i^2$	80,19	1084,16	<i>t</i> student	2.06
s_p	6,82		Límite inferior	-5,44
<i>Error estándar</i> $(\bar{x}_1 - \bar{x}_2)$	2,64		Límite superior	5,44

Fuente: Elaboración propia.

De la tabla 2, se considera que, para los límites inferior y superior, la diferencia de distancias recorridas por estudiantes y profesores es entre 3,7 km y 14,5 km. En este caso, se asume que las desviaciones estándar son aproximadamente iguales y que las dos distribuciones son similares a una distribución normal.

Muestras dependientes

Cuando las muestras por ser comparadas son, de alguna manera, dependientes, existe un análisis específico para realizar, en lugar del análisis de dos muestras que se mostró arriba. Un ejemplo es cuando comparamos el desempeño de empleados de una fábrica, los cuales, al ingresar, realizaron un primer curso de capacitación y a los 6 meses un segundo curso. Se puede suponer que aquellas personas que tuvieron una calificación baja en el primer curso son propensas a tener, también, un bajo desempeño en el segundo y una situación inversa puede plantearse para aquellos que tuvieron alta calificación. El tratamiento de dos muestras no considera esta *historia previa* y, por lo mismo, no considera la dependencia entre las muestras. Por ello, para analizar este tipo de datos se verá cómo utilizar esta información de una forma ventajosa.

El análisis no estudia dos variables por separado (digamos, de nuestro ejemplo, el primer y segundo curso), sino que analiza sus diferencias. Para cada par en la muestra, se calcula la diferencia entre las dos calificaciones. Luego, esto genera una sola muestra sobre la cual se aplica un análisis simple como el que se mostró para la determinación del intervalo de confianza para la media.

Ejemplo de aplicación

Se plantea buscar un indicador fisiológico para una patología mental. Por dicho motivo, se realizó un estudio con 15 pares de gemelos, donde un miembro del par padece la patología y el otro no. En el estudio se registraron los volúmenes de distintas regiones y subregiones cerebrales de las personas. Los datos obtenidos para las dimensiones del hipocampo se muestran en la tabla 3.

Tabla 3: Ejemplo para la determinación de intervalo de muestras dependientes

Par	Sano (cm ³)	Enfermo (cm ³)	$d_i = x_i - y_i$		
1	1,98	1,33	0,66	\bar{d}	0,17
2	1,47	1,70	-0,23	s_d	0,25
3	1,60	1,53	0,06	<i>valor t</i>	2,1448
4	1,62	1,45	0,17		
5	2,11	2,01	0,09		
6	1,70	1,31	0,38		
7	1,79	1,78	0,01		
8	1,81	1,74	0,07		
9	1,82	1,34	0,49		
10	1,96	1,93	0,03		
11	1,28	1,06	0,21		
12	1,97	1,40	0,58		
13	2,09	2,11	-0,02		
14	1,66	1,66	0,00		
15	2,13	2,06	0,07		

Fuente: Elaboración propia.

Las medidas no son independientes, ya que han sido medidas sobre el mismo par (par i) de gemelos; por tanto, para considerar esta dependencia y poder calcular el intervalo de confianza, se debe calcular la diferencia entre los pares de medidas definiendo, ahora, $d_i = x_i - y_i$ (tabla 3). El hecho de generar esta diferencia es para tener, si lo hay, un índice directo de la patología para cada par de gemelos. De esta manera, se reduce a una muestra (diferencia) independiente y se trata como se vio anteriormente, donde, en este caso, el intervalo de confianza será dado por $\bar{d} \pm \text{valor } t \times \frac{s_d}{\sqrt{n}}$, con \bar{d} igual a la diferencia promedio y s_d la desviación estándar muestral de la diferencia (ver tabla 3). De este análisis, podemos ver que el intervalo de confianza para \bar{d} será:

$$\begin{aligned} \bar{d} \pm \text{valor } t \times \frac{s_d}{\sqrt{n}} &= 0,17 \text{ cm}^3 \pm 2,1448 \times \frac{0,25 \text{ cm}^3}{\sqrt{15}} \\ &= (0,17 \pm 0,14) \text{ cm}^3 \end{aligned}$$

De esto se interpreta que se puede tener un 95 % de confianza en que el tamaño medio para individuos sin la patología será entre 0,03 y 0,31 cm^3 mayor en relación con el de los que presentan la enfermedad.

Intervalos de confianza para la diferencia entre proporciones

El procedimiento para estimar intervalos de confianza para una diferencia de proporciones es el mismo que el representado para la diferencia de medias muestrales. Sin embargo, en vez de comparar dos medias, ahora se comparan dos proporciones. Formalmente, si p_1 y p_2 representan dos proporciones desconocidas de la población y definimos a \hat{p}_1 y a \hat{p}_2 como las dos proporciones de la muestra, basados en los tamaños de muestra n_1 y n_2 , se puede considerar que la estimación puntual $p_1 - p_2$ es comparable a la diferencia de las proporciones de la muestra $\hat{p}_1 - \hat{p}_2$. También, si el tamaño de la muestra es lo suficientemente grande, la distribución muestral será normal. Con estas consideraciones, se puede definir como intervalo de confianza para la diferencia de proporciones poblacionales $p_1 - p_2$ como:

$$\hat{p}_1 - \hat{p}_2 \pm \text{valor } z \times \text{error estándar } (\hat{p}_1 - \hat{p}_2).$$

El error estándar $(\hat{p}_1 - \hat{p}_2)$ se define como:

$$\text{error estándar } (\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

Este tipo de análisis se aplica a diversos casos de la vida cotidiana y de los negocios, como, por ejemplo:

- Una empresa tiene dos plantas que fabrican el mismo producto. La gerencia desea conocer en cuánto difiere, entre plantas, la proporción del producto fuera de la calidad esperada.
- Un laboratorio desea conocer la efectividad de una nueva droga somnifera; para ello, mide la diferencia de proporciones en una muestra donde personas tomaron la nueva droga y a otras se les administró un placebo.

Ejemplo de aplicación

Una fábrica de electrodomésticos necesita contratar a una empresa de reparto de este tipo de productos; por ello, evalúa el desempeño de dos compañías de envíos, A y B. El desempeño se midió determinando el número de envíos, los cuales incurrieron en un reclamo por un monto igual o mayor a 120 USD por pérdida o rotura del artículo. Para hacer la comparación, se tomaron dos muestras al azar y se observó para A que, de 391 envíos, hubo un reclamo en 32 casos, mientras que para la empresa B, de 442 envíos, 20 fueron reclamados. La fábrica desea estimar la diferencia entre las tasas de reclamos con un nivel de confianza del 90 %.

En este caso, lo que puede interesarle a la fábrica es el parámetro $p_1 - p_2$, la diferencia en las proporciones de reclamos de las empresas A versus B.

Por este motivo, se debe, primeramente, calcular $\hat{p}_1 = \frac{32}{391} = 0,0818$ y $\hat{p}_2 = \frac{20}{442} = 0,0452$. El parámetro, para el nivel de confianza del 90 %, es $z_{\alpha/2} = \pm 1,645$ (en Excel se utiliza “=INV.NORM.ESTAND (0,95)”, donde 0,95 es el nivel de confianza + $\alpha/2$ para el límite superior, en tanto que para el límite inferior será “=INV.NORM.ESTAND (0,05)”, donde 0,05 es $\alpha/2$). Con estos valores, ahora, se puede obtener el *error estándar* ($\hat{p}_1 - \hat{p}_2$):

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

$$= \sqrt{\frac{0,0818(1 - 0,0818)}{391} + \frac{0,0452(1 - 0,0452)}{442}} = 0,017.$$

Como se definió arriba, el intervalo va a estar dado por lo siguiente:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \times \text{error estándar} (\hat{p}_1 - \hat{p}_2) = 0,037 \pm 0,028.$$

Esto da un rango de 0,009 y 0,065; en otras palabras, con un nivel de confianza del 90 %, la empresa A tiene un nivel de reclamo mayor que el de la empresa B, de entre el 0,9 % y 6,5 %.

Referencias

Agresti, A., y Franklin, C. (2013). *Statistics: The Art and Science of Learning from Data* (traducción propia). Boston, USA: Pearson Education, Inc.
Albright, S. C., y Winston, W. L. (2013). *Business Analytics: Data Analysis and Decision Making* (A. Apellido del traductor o Traducción propia). Stamford, USA: Cengage Learning

Anderson, D. R., Sweeney, D. J., y Williams, T. A. (2011). *Statistics for Business and Economics* (Traducción propia). Mason, OH: Cengage Learning.

Levine, D. M., Krehbiel, T. C., y Berenson, M. L. (2014). *Estadística para administración* (Traducción propia). México: Pearson Educación.

Mendenhall, W., Beaver, R. J., y Beaver, B. M. (2013). *Introduction to Probability and Statistics* (Traducción propia). Boston, USA: Brooks/Cole, Cengage Learning.