AWAKELAB

BASECAMP

Ciencia de Datos

Inferencia Estadística

Objetivo de la jornada

- Realizar cálculos de probabilidad utilizando la distribución muestral para resolver un problema.

Muestras y muestreo

Los investigadores utilizan ampliamente diferentes métodos de muestreo en los estudios de mercado, de modo que no necesitan investigar a toda la población para recoger percepciones procesables.

Hoy conoceremos las características de cada uno de estos métodos para que decidas cuál es el que necesitas llevar a cabo para que tu proyecto de investigación sea todo un éxito.

Definición de muestreo

El muestreo es una técnica de selección de miembros individuales o de un subconjunto de la población para hacer inferencias estadísticas a partir de ellos y estimar las características de toda la población.

También es un método conveniente en cuanto al tiempo y eficaz en función de los costos y, por lo tanto, constituye la base de cualquier diseño de investigación. Las técnicas de muestreo pueden utilizarse en un programa informático de encuestas de investigación para una derivación óptima.

Por ejemplo, si un fabricante de medicamentos desea investigar los efectos secundarios adversos de un medicamento en la población del país, es casi imposible llevar a cabo un estudio de investigación en el que participen todos. Para ello, el investigador elige tener una muestra de

personas de cada grupo demográfico para luego investigar, dándole una retroalimentación indicativa sobre el comportamiento del medicamento.

Métodos de muestreo

Existen dos métodos de muestreo: El muestreo probabilístico y el no probabilístico:

Muestreo probabilístico: El muestreo probabilístico es una técnica de muestreo en la que un investigador establece una selección de unos pocos criterios y elige al azar a los miembros de una población. Todos los miembros tienen la misma oportunidad de formar parte de la muestra con este parámetro de selección.

Muestreo no probabilístico: En el muestreo no probabilístico, el investigador elige al azar a los miembros de la investigación. Este método de muestreo no es un proceso de selección fijo o predefinido. Esto dificulta que todos los elementos de una población tengan las mismas posibilidades de ser incluidos en una muestra.

Ejemplos de métodos de muestreo

Conozcamos diversos tipos de muestreo probabilístico y no probabilístico que pueden aplicarse en cualquier estudio de investigación de mercado.

El muestreo de probabilidad es una técnica de muestreo en la que los investigadores eligen muestras de una población más grande utilizando un método basado en la teoría de la probabilidad. Este es uno de los métodos de muestreo que considera a todos los miembros de la población y forma muestras basadas en un proceso fijo.

Por ejemplo, en una población de 1000 miembros, cada miembro tendrá una probabilidad de 1/1000 de ser seleccionado para formar parte de una muestra. El muestreo probabilístico elimina el sesgo en la población y da

a todos los miembros una oportunidad justa de ser incluidos en la muestra.

Muestreo probabilístico

Muestreo Aleatorio Simple

El muestreo aleatorio simple es un procedimiento de muestreo probabilístico que da a cada elemento de la población objetivo y a cada posible muestra de un tamaño determinado, la misma probabilidad de ser seleccionado.

El muestreo aleatorio simple no es tan utilizado en investigaciones del consumidor, sobre todo porque es complicado obtener un marco de muestreo donde extraer al azar y no querrás darle a todas las unidades de la muestra una probabilidad igual de ser elegidas, ya que usualmente para hacer una investigación de este tipo se requiere a usuarios de tiendas o consumidores de ciertos productos o ciertas áreas específicas para ser las unidades de muestreo.

No olvidemos que una parte muy importante del muestreo consiste en tener el tamaño de la muestra correcta, para no tener un error de muestreo, el cual debe ser el mínimo posible.

Pasos para seleccionar una muestra aleatoria simple

1. Define la población objetivo. Quizá quieras leer: ¿Cómo encontrar a tu mercado objetivo?

2. Identifica un marco de muestreo actual de la población objetivo o desarrollar uno nuevo.

3. Evalúa el marco de muestreo para la falta de cobertura, cobertura excesiva, cobertura múltiple y la agrupación, y haz los ajustes que consideres necesarios.

4. Asigna un número único a cada elemento de la trama.

5. Determina el tamaño de la muestra.

6. Selecciona al azar el número específico de elementos de la población.

Para seleccionar el número de elementos de la población puedes recurrir al método de lotería, una tabla de números aleatorios y los números generados de forma aleatoria mediante un programa de computadora, es decir, al azar.

El método de lotería sólo funciona bien con pequeñas poblaciones de la muestra, es poco práctico para su uso con poblaciones más grandes.

Un ejemplo del uso del método de lotería sería la selección de una muestra aleatoria de entre un grupo de 100 miembros. Se ponen todos los nombres en un recipiente y se van sacando uno por uno hasta tener el tamaño suficiente de nuestra muestra.

Al utilizar el sistema de sorteo, los números que representan cada elemento de la población objetivo son colocados en tarjetas, papel u otros objetos.

Los chips se colocan entonces en un recipiente y se mezclan. A continuación, a ciegas se seleccionan las fichas desde el recipiente hasta que se haya obtenido el tamaño de muestra deseado. Las desventajas de este método de selección es que consume mucho tiempo, y se limita a poblaciones pequeñas.

El uso de números aleatorios, un método alternativo implica también la numeración de miembros de la población de 1 a N. Luego, el tamaño de

muestra de n tiene que ser determinada por selección de los números al azar.

El uso de la tabla de números aleatorios similar a la que aparece a continuación puede ayudar en gran medida con la aplicación de esta técnica de muestreo.

TABLA DE NÚMEROS ALEATORIOS											
20	17	42	01	72	33	94	55	89	65	58	60
72	49	04	27	56	49	11	63	77	79	23	00
94	70	49	05	74	64	00	26	07	23	60	31
22	15	78	49	74	37	50	94	13	90	08	14
93	29	12	20	26	22	66	98	37	53	82	62
45	04	77	48	87	77	33	58	12	08	91	12
16	23	91	95	97	98	52	49	99	78	30	37
04	50	65	37	99	57	74	98	93	99	78	30
03	64	59	55	85	63	49	46	61	89	33	79
62	49	00	67	28	96	19	65	13	44	78	39
89	03	90	40	10	18	43	37	68	97	28	19

En una tabla de números aleatorios no se sigue un patrón particular. Pueden ser leídos de cualquier manera, es decir, horizontal, vertical, diagonal, hacia delante o hacia atrás. El número de dígitos que se utiliza debe corresponder al tamaño total de la población objetivo.

Los números que el investigador encuentra que no concuerdan con los números asignados a elementos de la población objetivo son ignorados.

Este proceso de la tabla de números aleatorios es un proceso tedioso, consume tiempo, y no se recomienda para grandes poblaciones.

En su lugar, se pueden utilizar softwares estadísticos u hojas de cálculo para generar números aleatorios. Los elementos de las poblaciones cuyos números asignados coinciden con los números generados por el software son incluidos en la muestra. Se puede seleccionar un número de una tabla de números aleatorios para usarlo como el número de partida para el procedimiento.

Ejemplo de muestreo aleatorio simple

Una empresa tiene 120 empleados. Se quiere extraer una muestra de 30 de ellos.

- Enumera a los empleados del 1 al 120
- Sortea 30 números entre los 120 trabajadores
- La muestra estará formada por los 30 empleados que salieron seleccionados de los números obtenidos.

En resumen el muestreo aleatorio simple es como hacer un sorteo, afortunadamente existen herramientas que facilitan la automatización y le da seriedad a este proceso.

Muestreo aleatorio clusters multi-etapas

El muestreo por conglomerados nos ayuda cuando es imposible o poco práctico crear un marco de muestreo de una población objetivo debido a que está muy dispersa geográficamente y el costo de la recopilación de datos es relativamente alto.

El muestreo por conglomerados, también conocido como muestreo por racimos, es un procedimiento de muestreo probabilístico en que los elementos de la población son seleccionados al azar en forma natural por agrupaciones (clusters). Los elementos del muestreo se seleccionan de la población de manera individual, uno a la vez.

Las unidades de muestreo o grupos pueden ser espaciados, tal como ocurre naturalmente en las unidades geográficas o físicas (por ejemplo: estados, delegaciones o distritos); en base a una organización como escuelas, grado escolar; o servicio telefónico tales como códigos de área o el cambio de las claves lada de los números de teléfono.

La heterogeneidad del grupo es fundamental para un buen diseño del muestreo por conglomerados. Por otra parte, los elementos dentro de cada grupo debe ser tan heterogéneos como la población objetivo

Pasos para seleccionar un muestreo por conglomerados

1. Definir la población objetivo.

2. Determinar el tamaño de la muestra.

3. Identificar un marco de muestreo existente o desarrollar un nuevo marco de muestreo de grupos de la población objetivo.

4. Evaluar el marco de muestreo para la falta de cobertura, cobertura excesiva, múltiple cobertura, y la agrupación, y hacer los ajustes cuando sea necesario.Idealmente, los grupos serían tan heterogéneos como la población, mutuamente excluyentes, y colectivamente exhaustivos. La duplicación de elementos de la muestra puede aparecer si elementos de la población pertenecen a más de un grupo. La omisión dará lugar a un sesgo de cobertura.

5. Determinar el número de grupos que se seleccione. Esto se puede hacer dividiendo el tamaño de la muestra por el número promedio estimado

de elementos de la población en cada grupo. En la medida en que la homogeneidad y la heterogeneidad de los grupos sean diferentes a la de la población, el número del grupo aumenta e incrementa la precisión. Por otra parte, si las diferencias aumentan, la precisión disminuye.

6. Seleccionar al azar el número previsto de las agrupaciones.

Muestreo aleatorio estratificado

El muestreo estratificado es uno de los tipos de muestreo probabilístico del que podemos hacer uso.

El muestreo estratificado es un procedimiento de muestreo en el que el objetivo de la población se separa en segmentos exclusivos, homogéneos (estratos), y luego una muestra aleatoria simple se selecciona de cada segmento (estrato).

Las muestras seleccionadas de los diversos estratos se combinan en una sola muestra. Este procedimiento de muestreo se refiere a veces como ”muestreo de cuota aleatorio”.

Pasos de selección para un muestreo estratificado

Hay ocho pasos principales en la selección de una muestra aleatoria estratificada:

1. Define la población objetivo.

2. Identifica variable o variables de estratificación y determinar el número de estratos a usarse. Las variables de estratificación deben estar relacionadas con el propósito de estudio. Si el propósito del estudio es hacer estimaciones de los subgrupos, las variables de estratificación deben estar relacionadas con esos subgrupos.

   La disponibilidad de información auxiliar a menudo determina las variables de estratificación que se utilizan. Puede ser utilizada más de una variable de estratificación. Considera que a medida que el número de variables de estratificación aumenta, incrementa la probabilidad de que algunas de las variables cancelen los efectos de otras variables, no más de cuatro a seis variables de estratificación y no se deben utilizar más de seis estratos de una variable en particular.

3. Identifica un marco de muestreo existente o desarrollar uno que incluya información sobre la o las variables de estratificación para cada elemento de la población objetivo. Si el marco de la muestra no incluye la información en las variables de estratificación, la estratificación no sería posible.

4. Evalúa el marco de muestreo para la falta de cobertura, cobertura excesiva, múltiple, y la agrupación, y haz los ajustes cuando sea necesario.

5. Divide el marco de muestreo en estratos, categorías de la estratificación de la o las variables, creando un marco de muestreo para cada estrato. Dentro del estrato las diferencias deben reducirse al mínimo, y las diferencias entre los estratos deben maximizarse. Los estratos no

deben estar superpuestos, en conjunto, debe constituir toda la población. Los estratos deben ser independientes y mutuamente exclusivos del subconjunto de la población. Cada elemento de la población debe estar en un sólo estrato.

6. Asigna un número único a cada elemento.

7. Determina el tamaño de la muestra para cada estrato. La distribución numérica de los elementos incluidos en la muestra a través de los diversos estratos determina el tipo de muestreo a implementar. Puede ser un muestreo proporcional estratificado o uno de los diversos tipos de muestreo estratificado desproporcionado.

8. Selecciona al azar el número específico de elementos de cada estrato. Al menos un elemento se debe seleccionar de cada estrato para la representación de la muestra; y por lo menos dos elementos deben ser elegidos de cada estrato para el cálculo del margen de error de las estimaciones calculadas a partir de los datos recogidos.

Desvíos en las muestras

El muestreo sistemático es un tipo de muestreo probabilístico donde se hace una selección aleatoria del primer elemento para la muestra, y luego se seleccionan los elementos posteriores utilizando intervalos fijos o sistemáticos hasta alcanzar el tamaño de la muestra deseado.

1. Pasos para la selección de un muestreo sistemático

2. Definir la población objetivo.

3. Determinar el tamaño deseado de la muestra (n).

4. Identificar el marco muestreo existente o desarrollar un marco de muestreo de la población objetivo.

5. Evaluar el marco muestral por falta de cobertura, cobertura excesiva, múltiple cobertura, agrupación, periodicidad, y hacer los ajustes cuando sea necesario. Idealmente, la lista estará en un orden aleatorio con respecto al estudio variable o, mejor aún, ordenados en función de la variable de interés o su correlación, creando así estratificación implícita.

6. Determinar el número de elementos en el marco de la muestra (N).

7. Calcular el intervalo de muestreo (i) dividiendo el número de elementos en el marco de muestreo (N) por el tamaño de la muestra específica (n). Uno debería ignorar el resto y redondear o terminar en el número entero más próximo. El redondeo hacia abajo y truncando puede hacer que el tamaño de la muestra sea más grande de lo deseado. Si es así, se puede eliminar de forma aleatoria las selecciones adicionales. Si no se conoce el tamaño exacto, o es poco práctico determinar, se puede fijar una fracción de muestreo.

8. Seleccionar al azar un número, r, de “1” mediante i.

9. Selecciona para la muestra, r, r + i, r + 2i, r ,+3i, y así sucesivamente, hasta agotar el marco.

A nivel técnico, el muestreo sistemático no crea una muestra verdaderamente aleatoria. Sólo la selección del primer elemento de muestreo sistemático es una selección de probabilidad. Una vez que el primer elemento es seleccionado, algunos de los elementos tendrán una probabilidad cero de selección.

Además, cierta combinación de elementos, como los elementos que son adyacentes entre sí en el marco de muestreo, pueden no ser seleccionados. Muestreos sistemáticos repetidos pueden utilizarse para abordar este problema.

Distribución muestral y teorema del límite central

La distribución de muestreo

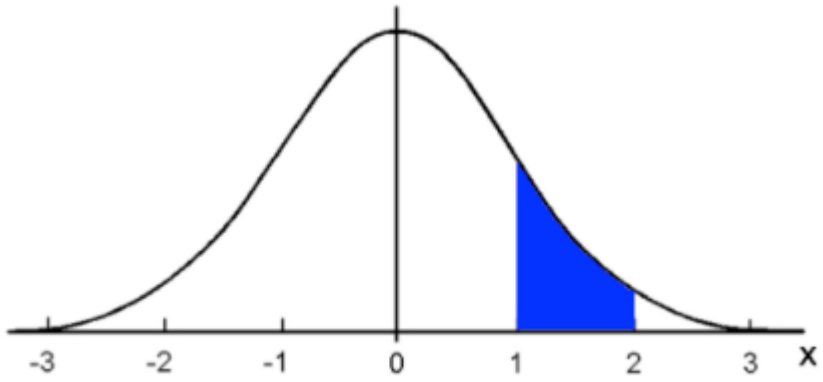
En cada una de las distintas muestras que pueden ser extraídas de una población se pueden calcular la media aritmética o la proporción de elementos que presentan cierta característica; por ejemplo, la media de estaturas o la proporción de licenciados universitarios. Cuando los elementos son escogidos de manera aleatoria, los estadísticos pueden tomar distintos valores en cada una de las muestras, cada uno de ellos con distinta probabilidad. En los ejemplos del inicio de esta sección ya vimos que los valores de la media en diferentes muestras aleatorias se encontraban con mayor probabilidad cerca del valor de la media poblacional, y que era menos probable que se encontrasen muy alejados de ella.

La probabilidad de cada uno de los posibles valores que puede tomar un estadístico en muestras extraídas al azar viene dada por una función matemática denominada **distribución muestral**, que depende del estadístico en cuestión. Se habla así, por ejemplo, de la distribución muestral de la media aritmética o de la distribución muestral de la proporción.

Una distribución muestral es una función de probabilidad, ya que asigna a cada posible valor de un estadístico su probabilidad de aparecer en una muestra extraída al azar. En realidad, esta definición es estrictamente cierta sólo cuando la variable toma valores discretos; por ejemplo, cuando procede de un contaje y sus posibles valores son 0, 1, 2, 3, etc. Cuando el valor del estadístico muestral es una variable continua, la distribución muestral correspondiente se denomina función de densidad de probabilidad. La probabilidad en este caso corresponde gráficamente a un área bajo la curva de esa función, delimitada por un cierto intervalo de la variable. Analíticamente, esa área se calcula como la integral de la función entre los límites del intervalo de la variable, que en la práctica se obtiene con un ordenador o se consulta en una tabla. El área total bajo la

curva, que se extiende a todos los posibles valores de la variable, es siempre uno, que corresponde a la probabilidad de un suceso seguro.

La siguiente gráfica muestra la curva de una función de densidad de probabilidad para una variable x, y en ella se señala la probabilidad de que esa variable se encuentre entre los valores 1 y 2, que corresponde al área bajo la curva marcada en azul:



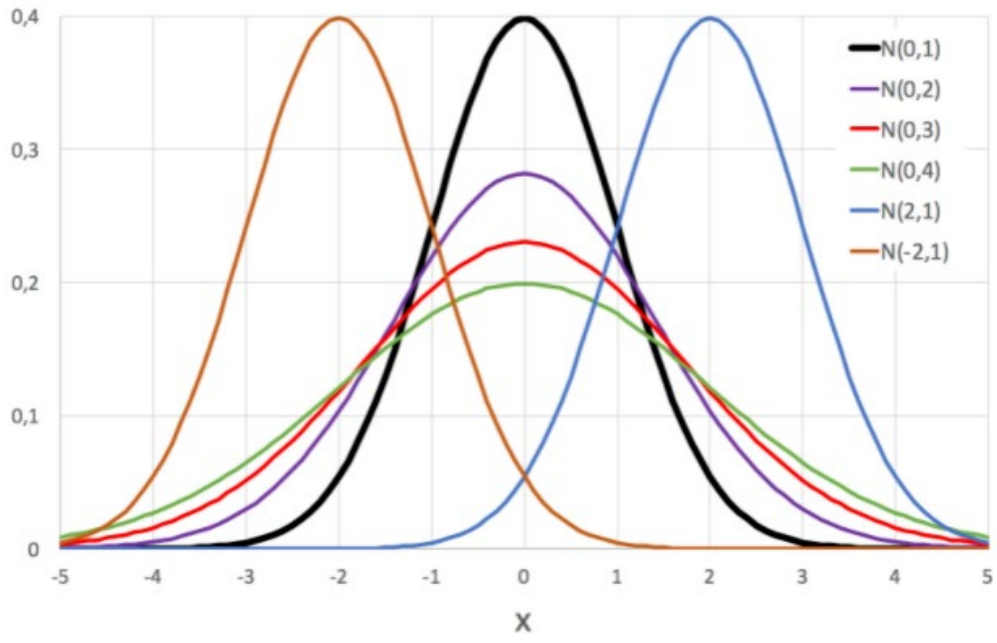
La función de densidad de probabilidad más importante en estadística se llama distribución normal o distribución gaussiana, o también campana de Gauss, por la forma que toma cuando es representada gráficamente (como aparece en la figura anterior). Su forma concreta depende de dos parámetros, la media y la varianza. La curva alcanza un máximo cuando la variable toma el valor de la media, y es simétrica respecto a ese valor, aproximándose a cero indefinidamente conforme la variable se aleja de la media por ambos lados. La desviación típica, que es la raíz cuadrada de la varianza, está relacionada con la anchura de la campana: a mitad de altura del máximo, la anchura de la campana es aproximadamente 2,36 σ. Una distribución normal de media μ y varianza σ2 se puede simbolizar como N(μ,σ2), y así lo usaremos aquí.

Para una variable aleatoria x se define su variable tipificada como:

$$z = \frac{x - \mu}{\sigma}$$

Si la variable x sigue una distribución N(μ,σ2), su variable tipificada correspondiente sigue una distribución N(0,1), que se denomina distribución normal estándar, con media 0 y varianza 1 (es la representada en la figura de arriba). Existen infinitas distribuciones normales distintas, tantos como posibles valores de la media y la varianza, pero las áreas bajo la curva normal estándar son las únicas que se pueden encontrar en tablas para su consulta.

La siguiente figura recoge la representación gráfica de distribuciones normales con distintos valores de media y de varianza, según la notación N(μ,σ2); la curva negra más gruesa es la distribución normal estándar. Se puede observar cómo la posición del máximo coincide con la media y cómo la anchura de las campanas es proporcional a la desviación; el área total bajo cualquiera de las curvas es siempre uno.



Ejemplo:

Se tiene una variable aleatoria x con distribución normal de media 2 y varianza 9, N(2,9). Se quiere calcular la probabilidad de que la variable tome valores entre 1 y 2.

La probabilidad pedida corresponde al área bajo la curva normal especificada y limitada por los valores dados de la variable. Para poder consultar una tabla de áreas bajo la curva normal estándar, se tipifican los valores límite de la variable:

$$z_A = (x_A - \mu)/\sigma = (1-2)/\sqrt{9} = -0{,}333$$

$$z_B = (x_B - \mu)/\sigma = (2-2)/\sqrt{9} = 0$$

El área bajo la curva normal N(2,9) delimitada por los valores de la variable 1 y 2 es la misma que el área bajo la curva normal N(0,1) delimitada por los valores de la variable −0,333 y 0. Si se busca en una tabla de áreas bajo la curva normal (como se describe en los ejemplos posteriores), se obtiene 0,13. Así, la probabilidad de que una variable aleatoria distribuida según N(2,9) tome valores entre 1 y 2 es de 0,13, es decir, un 13%.

Teorema del límite central

El teorema central del límite (TCL), es una teoría estadística que establece que, dada una muestra aleatoria suficientemente grande de la población, la distribución de las medias muestrales seguirá una distribución normal.

Además, el TCL afirma que a medida que el tamaño de la muestra se incrementa, la media muestral se acercará a la media de la población. Por tanto, mediante el TCL podemos definir la distribución de la media muestral de una determinada población con una varianza conocida. De manera que la distribución seguirá una distribución normal si el tamaño de la muestra es lo suficientemente grande.

El teorema central del límite tiene una serie de propiedades de gran utilidad en el ámbito estadístico y probabilístico. Las principales son:

- Si el tamaño de la muestra es suficientemente grande, la distribución de las medias muestrales seguirá aproximadamente una distribución normal. El TCL considera una muestra como grande cuando el tamaño de la misma es superior a 30. Por tanto, si la muestra es superior a 30, la media muestral tendrá una función de distribución próxima a una normal. Y esto se cumple independientemente de la forma de la distribución con la que estamos trabajando.

- La media poblacional y la media muestral serán iguales. Es decir, la media de la distribución de todas las medias muestrales será igual a la media del total de la población.

- La varianza de la distribución de las medias muestrales será σ^2/n. Que es la varianza de la población dividido entre el tamaño de la muestra.

Que la distribución de las medias muestrales se parezca a una normal es tremendamente útil. Porque la distribución normal es muy fácil de aplicar para realizar contrastes de hipótesis y construcción de intervalos de confianza. En estadística que una distribución sea normal es bastante importante, dado que muchos estadísticos requieren este tipo de distribución. Además, el TCL nos permitirá hacer inferencia sobre la media poblacional a través de la media muestral. Y esto es de gran utilidad cuando por falta de medios no podemos recolectar datos de toda una población.

Ejemplo:

Imaginemos que queremos analizar las rentabilidades medias históricas del índice S&P 500, que como sabemos, tiene unas 500 compañías dentro del mismo. Pero no tenemos suficiente información como para analizar la totalidad de las 500 compañías del índice. En este caso la rentabilidad media del S&P 500 sería la media poblacional.

Ahora bien, siguiendo al TCL podemos coger una muestra de estas 500 empresas para realizar el análisis. La única limitación que tenemos es que en la muestra tiene que haber más de 30 compañías para que se cumpla el teorema. Entonces imaginemos que cogemos 50 compañías del índice de manera aleatoria y repetimos el proceso varias veces. Los pasos a seguir el ejemplo serían los siguientes:

- Elegimos la muestra de unas 50 compañías y obtenemos la rentabilidad media de la totalidad de la muestra.
- De manera continuada seguimos escogiendo 50 compañías y obtenemos la rentabilidad media.
- La distribución de todas las rentabilidades medias de todas las muestras escogidas se aproximará a una distribución normal.
- Las rentabilidades medias de todas las muestras seleccionadas se aproximará a la rentabilidad media del total del índice. Tal y como demuestra el teorema Central del Límite.

Por tanto mediante inferencia de la rentabilidad media de la muestra podemos acercarnos a la rentabilidad media del índice.

Distribución muestral de la media

La media m de las muestras extraídas al azar de una población con media μ y varianza σ2 es una variable aleatoria que sigue una distribución de probabilidad normal caracterizada por:

- Su media, que coincide con la media poblacional μ;
- Su varianza, que viene dada por la varianza de la población σ2 dividida entre el tamaño de la muestra n, es decir, σ2/n.

Así pues, la distribución muestral de la media m es N(μ, σ2/n). La media muestral tipificada sigue una distribución normal estándar, N(0,1), y se calcula como:

$$m^* = \frac{m - \mu}{\sigma / \sqrt{n}}$$

Con estos resultados se puede calcular fácilmente la probabilidad de que la media de una muestra extraída al azar se encuentre en un cierto intervalo, conociendo la media y la varianza de la población de la que se ha extraído la muestra. Para ello, se calcula el área bajo la curva de la distribución normal $N(\mu, \sigma 2/n)$ entre los límites del intervalo, o bien se consulta en una tabla el área bajo la curva de la distribución normal estándar $N(0,1)$ entre los límites tipificados del intervalo.

Ejemplo

Si los individuos de una población tienen un peso medio de 70 kg, con una desviación típica de 10 kg, ¿cuál es la probabilidad de que la media de peso de los 121 pasajeros de un avión, que se supone que representan una muestra al azar, esté entre 72 y 73 kg?

Se puede suponer que la media muestral sigue una distribución normal de media 70 kg y varianza 102/121 = 0,826 kg2, es decir, $N(70, 0{,}826)$. Los valores tipificados de los límites del intervalo de la variable son:

$$m^*A = (72-70)/\sqrt{0{,}826} = 2{,}20$$

$$m^*B = (73-70)/\sqrt{0{,}826} = 3{,}30$$

La variable tipificada sigue una distribución $N(0,1)$ cuyas áreas bajo la curva se encuentran tabuladas. Por ejemplo, si se mira en esta tabla, se pueden encontrar las áreas bajo la curva que quedan a la izquierda de un cierto valor de la variable. El área bajo la curva comprendida entre m^*A y m^*B es entonces el área a la izquierda de m^*B menos el área a la izquierda de m^*A.

En esa tabla la cifra de las unidades de la variable se encuentra en la primera fila y las cifras del primer y segundo decimal se encuentran en la primera columna. Una vez localizada la fila y la columna que corresponde al valor buscado, la casilla en la que se cruzan contiene el área bajo la curva normal tipificada que queda a la izquierda de ese valor. Se procede entonces de la siguiente manera:

- Para m^*A = 2,20 se localiza la columna 2,00 y la fila 0,20, en cuya intersección aparece el área 0,986096.

- Para m^*B = 3,30 se localiza la columna 3,00 y la fila 0,30, en cuya intersección aparece el área 0,999516.

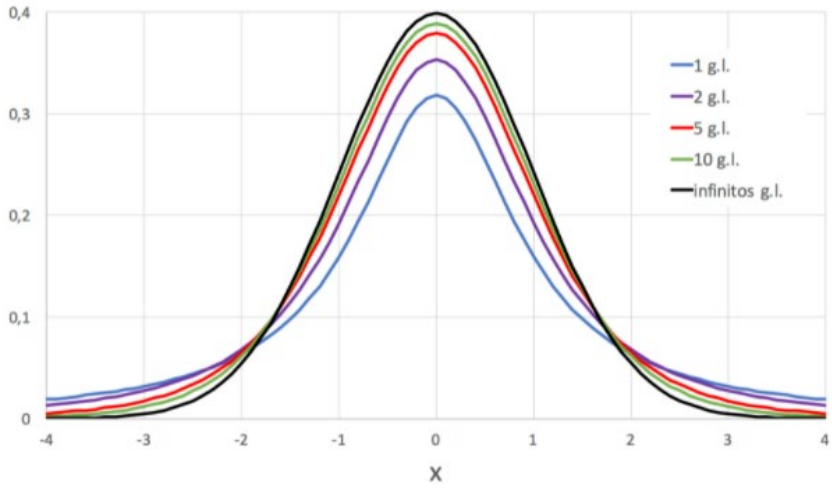
La diferencia entre ambas áreas es 0,01342, que corresponde al área bajo la curva delimitada entre m^*A y m^*B. Por tanto, la probabilidad de que la media de peso de los pasajeros del avión se encuentre entre 72 y 73 kg es de 0,013 aproximadamente, o de 1,3 %.

Cuando la varianza de la población $\sigma 2$ no se conoce, que es lo más habitual, ha de estimarse a partir de la varianza de la propia muestra que se ha extraído, $s2$. La media muestral tipificada se obtiene entonces sustituyendo la varianza poblacional por la muestral:

$$m_t^* = \frac{m - \mu}{s/\sqrt{n}}$$

Este nuevo estadístico t contiene la media muestral m y la varianza muestral s2, esta última con su propia distribución muestral. El cociente entre ambas ya no sigue la distribución normal estándar, sino otra distribución denominada t de Student, que depende de los grados de libertad de la muestra (número de elementos que contiene menos uno, n−1). Cuanto mayor es el tamaño de la muestra, más se parece esta distribución a la normal estándar, por lo que a menudo se emplea esta última en muestras grandes incluso si la varianza poblacional es desconocida.

La siguiente gráfica recoge la representación gráfica de distribuciones t de Student con distintos grados de libertad (g.l.); la curva negra, que corresponde a infinitos grados de libertad, coincide con la distribución normal estándar N(0,1):



Distribución muestral de la diferencia de medias

También se puede analizar la distribución muestral de la diferencia entre dos medias, m1−m2, obtenidas de muestras extraídas al azar de dos poblaciones distintas, una de ellas con media μ1 y varianza σ12 y otra con media μ2 y varianza σ22. La distribución muestral de la diferencia de medias sigue una distribución normal caracterizada por:

- Su media, que coincide con la diferencia de las medias poblacionales μ1−μ2;
- Su varianza, que depende de las varianzas de ambas poblaciones y del tamaño de ambas muestras, n1 y n2, y que viene dada por σ12/n1 + σ22/n2.

Así pues, la distribución muestral de la diferencia de medias m1−m2 es N(μ1−μ2, σ12/n1+σ22/n2). La diferencia de medias muestrales tipificada sigue una distribución normal estándar, N(0,1), y se calcula como:

$$(m_1 - m_2)^* = \frac{(m_1 - m_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Ejemplo:

Se tienen dos poblaciones distintas, una con media de edad de 20 años y desviación típica 1,5 años y la otra con media 25 años y desviación típica 1,7 años. Se reúnen en un aula 50 individuos escogidos al azar de la primera población y en otra aula, 60 individuos de la segunda población. ¿Cuál es la probabilidad de que la diferencia entre las medias de edad de ambas aulas esté entre 4,5 y 5,5 años?

Se puede suponer que la diferencia de medias muestrales sigue una distribución normal de media 25−20 = 5 años y varianza 1,52/50+1,72/60 = 0,0932 años2, es decir, $N(5, 0{,}0932)$. Los valores tipificados de los límites del intervalo de la variable son:

$$(m1-m2)^*A = (4{,}5-5)/\sqrt{0{,}0932} = -1{,}64 \text{ años}$$

$$(m1-m2)^*B = (5{,}5-5)/\sqrt{0{,}0932} = 1{,}64 \text{ años}$$

Procediendo de manera análoga al ejemplo anterior, para $(m1-m2)^*B = 1{,}64$ se localiza en la tabla la columna 1,00 y la fila 0,64, en cuya intersección aparece el área 0,949497. En la tabla no aparecen valores negativos de la variable, pero como la curva normal estándar es simétrica respecto al valor 0, se deduce que el área a la izquierda de $(m1-m2)^*A = -1{,}64$ es igual al área a la derecha del valor +1,64; esta última se puede obtener como el área total bajo la curva, que es 1, menos el área a la izquierda de +1,64, es decir, 1−0,949497 = 0,050503.

El área bajo la curva delimitada entre $(m1-m2)^*A$ y $(m1-m2)^*B$ corresponde entonces a la diferencia 0,949497−0,050503 = 0,898994. Por tanto, la probabilidad de que la diferencia entre las medias de edad de las personas de ambas aulas se encuentre entre 4,5 y 5,5 años es de casi 0,9, o del 90%.

Distribución muestral de la proporción

La proporción es el número de elementos de un cierto tipo presentes en un conjunto dividido entre el número total de elementos. En una muestra se suele denotar por p, y la proporción de elementos que no presentan la característica en cuestión es entonces 1−p. En una población, se suelen denotar π y 1−π, respectivamente.

La proporción p en muestras grandes (n>30) extraídas al azar de una población con proporción π tiende a seguir una distribución de probabilidad normal, caracterizada por:

- Su media, que coincide con la proporción poblacional π;
- Su varianza, que viene dada por el producto de las proporciones poblacionales de los elementos que presentan y que no presentan la característica en cuestión, dividido entre el tamaño de la muestra n, es decir, π(1−π)/n.

Así pues, la distribución muestral de la proporción p es N(π, π(1−π)/n). La proporción tipificada sigue una distribución normal estándar, N(0,1), y se calcula como:

$$p^* = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$$

Ejemplo:

El porcentaje de población española que fuma es del 22%. En un grupo de 100 personas escogido al azar, ¿cuál es la probabilidad de que fumen entre 25 y 30 personas?

La proporción de fumadores en la población es 0,22. Se puede suponer entonces que la proporción muestral sigue una distribución normal de media 0,22 y varianza 0,22 (1−0,22) / 100 = 0,001716, es decir, *N*(0,22, 0,001716). Los límites del intervalo de la variable, expresados en proporciones, son 0,25 y 0,30, cuyos valores tipificados son:

$$p^*1 = (0{,}25 - 0{,}22)/\sqrt{0{,}001716} = 0{,}72.$$

$$p^*2 = (0{,}30 - 0{,}22)/\sqrt{0{,}001716} = 1{,}93.$$

Procediendo de manera análoga a los ejemplos anteriores, para $p^*1 = 0{,}72$ se localiza en la tabla la columna 0,00 y la fila 0,72, en cuya intersección aparece el área 0,764237, y para $p^*2 = 1{,}93$ se localiza en la tabla la columna 1,00 y la fila 0,93, en cuya intersección aparece el área

0,973196. El área bajo la curva delimitada entre *p**1 y p*2 corresponde entonces a la diferencia entre ambas, 0,208959. Por tanto, la probabilidad de que en esa muestra fume entre un 25% y un 30% de las personas es de casi 0,21, o del 21%.

Referencias

[1] Técnicas de muestreo

https://scielo.conicyt.cl/scielo.php?pid=S0717-95022017000100037&script=sci_arttext

[2] Distribución muestral

https://www.ecotec.edu.ec/material/material_2021A1_MAT165_01_147222.pdf

https://bookdown.org/dietrichson/metodos-cuantitativos/distribucion-muestral.html

https://www.ugr.es/~mvargas/tema6sd.pdf

[3] muestreo Probabilístico.

https://www.questionpro.com/blog/es/como-realizar-un-muestreo-probabilistico/

[4] muestreo NO Probabilístico.

https://www.questionpro.com/blog/es/muestreo-no-probabilistico/

[5] Teorema del Limite central.

https://bookdown.org/aquintela/EBE/el-teorema-central-del-limite.html