AWAKELAB

BASECAMP

Ciencia de Datos

Inferencia Estadística

Objetivo de la jornada

- Realizar estimaciones de la media de una población utilizando intervalos de confianza a partir de una muestra aleatoria.

Inferencia e Intervalos de Confianza para la media

Inferencia Estadística

La inferencia estadística es el conjunto de métodos que permiten inducir, a través de una muestra estadística, el comportamiento de una determinada población. La inferencia estadística, estudia entonces cómo, a través de la aplicación de dichos métodos sobre los datos de una muestra, se pueden extraer conclusiones sobre los parámetros de la población de datos. De la misma manera estudia también el grado de fiabilidad de los resultados extraídos del estudio.

Para entender el concepto es importante entender tres conceptos:

1. **Inferencia**: Inferir significa, literalmente, extraer juicios o conclusiones a partir de ciertos supuestos, sean estos generales o particulares.
2. **Población**: Una población de datos, es el conjunto total de datos que existen sobre un variable.
3. **Muestra estadística**: Una muestra es una parte de la población de datos.

Teniendo claro a lo que nos referimos con el concepto de inferir, una de las dudas fundamentales recae en el hecho de elegir una muestra en lugar de una población.

Normalmente, en estadística, se trabaja con muestras debido a la gran cantidad de datos que tiene una población. Por ejemplo, si queremos sacar conclusiones, esto es, inferir, los resultados de las elecciones generales, es imposible preguntar a toda la población del país. Para solventar ese problema se escoge una muestra variada y representativa. Gracias a la cual se puedan extraer una estimación del resultado final. Escoger una muestra adecuada corre a cargo de las distintas técnicas de muestreo.

Métodos de inferencia estadística.

Los métodos y técnicas de la inferencia estadística se pueden dividir en dos: métodos de estimación de parámetros y métodos de contraste de hipótesis.

- **Métodos de estimación de parámetros:** Se encarga de asignar un valor al parámetro o al conjunto de parámetros que caracterizan el campo sujeto a estudio. Claro que al ser una estimación existe cierto error. Para obtener estimaciones adaptadas a esa realidad, se crean **intervalos de confianza**.
- **Métodos de contraste de hipótesis**: Su objetivo es comprobar si una estimación corresponde con los valores poblacionales. En todo contraste de hipótesis existen dos supuestos. La hipótesis nula (H0) que recoge la idea de que un valor tiene un valor predeterminado. Si se rechaza la hipótesis nula (H0), entonces se acepta la hipótesis alternativa (H1).

Intervalo de confianza

Un intervalo de confianza, es una técnica de estimación utilizada en inferencia estadística que permite acotar un par o varios pares de valores, dentro de los cuales se encontrará la estimación puntual buscada (con una determinada probabilidad)

Intervalo de confianza = media +- margen de error

Conocer el verdadero parámetro poblacional, por lo general, suele ser algo muy complicado. Pensemos en una población de 4 millones de personas. ¿Podríamos saber el gasto medio en consumo por hogar de esa población? En principio sí. Simplemente tendríamos que hacer una encuesta entre todos los hogares y calcular la media. Sin embargo, seguir ese proceso sería tremendamente laborioso y complicaría bastante el estudio.

Ante situaciones así, se hace más factible seleccionar una muestra estadística. Por ejemplo, 500 personas. Y sobre dicha muestra, calcular la media. Aunque seguiríamos sin saber el verdadero valor poblacional, podríamos suponer que este se va a situar cerca del valor muestral. A esa media le sumamos el margen de error y tenemos un valor del intervalo de confianza. Por otro lado, le restamos a la media ese margen de error y tendremos otro valor. Entre esos dos valores estará la media poblacional.

En conclusión, el intervalo de confianza no sirve para dar una estimación puntual del parámetro poblacional, si nos va a servir para hacernos una idea aproximada de cuál podría ser el verdadero de este. Nos permite acotar entre dos valores en dónde se encontrará la media de la población.

Factores de los que depende un intervalo de confianza

El cálculo de un intervalo de confianza depende principalmente de los siguientes factores:

- **Tamaño de la muestra seleccionada**: Dependiendo de la cantidad de datos que se hayan utilizado para calcular el valor muestral, este se acercará más o menos al verdadero parámetro poblacional.
- **Nivel de confianza**: Nos va a informar en qué porcentaje de casos nuestra estimación acierta. Los niveles habituales son el 95% y el 99%.
- **Margen de error de nuestra estimación**: Este se denomina como alfa y nos informa de la probabilidad que existe de que el valor poblacional esté fuera de nuestro intervalo.
- **Lo estimado en la muestra (media, varianza, diferencia de medias…):** De esto va a depender el estadístico pivote para el cálculo del intervalo.

Intervalo de confianza para la media con desviación Estándar de la población conocida

Dada una muestra, x1, ..., xn de una población con media (desconocida) μ y varianza conocida σ 2 , un intervalo de 95 % de confianza para la media poblacional μ es:

$$\bar{x} \pm 1{,}96\sigma/\sqrt{n}.$$

El cálculo del intervalo se puede hacer con calculadora o a través del Excel.

Ejemplo:

Se quería estimar la velocidad media en una calle con un límite teórico de 50km por hora. Con un radar oculto, se observó que la velocidad media de una muestra de **25** coches fue de **58km/hora**. Si la desviación típica de

la velocidad en esta calle es de 6km/hora, calcular un intervalo de **95 %** de confianza para la verdadera velocidad media.

Un intervalo de confianza es:

$$58 \pm 1{,}96 * 6/\sqrt{25} = 58 \pm 2{,}35 = (55{,}65,\ 60{,}35).$$

Se estima que la verdadera velocidad media en esta calle es entre 55,65km/hora y 60,35km/hora

Intervalo de confianza para la media con desviación Estándar de la población desconocida

El supuesto que se conoce la desviación típica de velocidades en la calle cuando no se conoce la media es poco realista en la práctica. Una alternativa en esta situación es usar la (cuasi) desviación típica muestral, para estimar la desviación típica de la población. Ahora si la muestra es de tamaño grande, el intervalo es:

$$\bar{x} \pm 1{,}96 s/\sqrt{n}.$$

Ejemplo:

En 100 pruebas de alcoholemia de conductores que han saltado un semáforo en Aranjuez el nivel medio de alcohol en aire era de 0,65 mg/litro con una cuasi desviación típica de 0,1mg/litro. Hallar un intervalo de 95 % de confianza para el verdadero nivel medio de alcohol en el aire para conductores que saltan el semáforo.

El intervalo es 0,65 ± 0, 02 = (0,63, 0,67).

Intervalo de confianza para una proporción

También, se puede hacer un intervalo típico para una proporción poblacional p dada una proporción muestral pˆ con la siguiente fórmula:

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Sería bastante complicado llevar esto a mano, pero se puede hacer con Excel con un pequeño truco.

Recordando el intervalo para una media.

$$\bar{x} \pm 1{,}96\sigma/\sqrt{n}.$$

Sustituimos σ por pˆ(1 − pˆ) es la parte marcada en rojo.

Ejemplo:

De una muestra de 100 pruebas aleatorias de alcoholemia, 10 conductores dieron positivo. Hallar un intervalo de confianza para la verdadera proporción de conductores en España que van borrachos.

$n = 100$ y $\hat{p} = 10/100 = 0{,}1$. Luego el intervalo es

$$0{,}1 \pm 1{,}96\sqrt{0{,}1 * 0{,}9/100} = 0{,}1 \pm 0{,}059 = (0{,}041,\ 0{,}159).$$

Se estima que entre 4 % y 16 % de los conductores van bebidos.

Niveles de confianza

El nivel de confianza, en estadística, es la probabilidad máxima con la que podríamos asegurar que el parámetro a estimar se encuentra dentro de nuestro intervalo estimado.

El nivel de confianza se define como 1-alfa y sus valores más comunes son 90%, 95% y 99%.

En estadística es común tener que estimar parámetros, los cuales, nunca vamos a poder afirmar al 100% que son el valor real que buscamos. Por ejemplo, observando a simple vista la altura de 10 alumnos en una clase podríamos estimar que la altura está entre 1,70 y 1,75.

Sería difícil saber con un 100% de certeza la altura media si no medimos a cada alumno y hacemos los cálculos. Por el contrario, sí podríamos acotar un intervalo y situar el valor dentro de este.

El nivel de confianza, sería el porcentaje máximo con el que podríamos asegurar que el parámetro real se encuentra dentro del intervalo acotado.

El nivel de confianza está directamente relacionado con el nivel de significación. En la expresión 1-alfa, el valor de alfa no es más que el nivel de significación.

El nivel de significación (o alfa) es la probabilidad máxima que asumimos de forma voluntaria de equivocarnos al rechazar la hipótesis nula cuando en realidad es cierta. En otras palabras, el nivel de significación es el máximo error que queremos cometer en nuestra estimación o contraste.

Dicho así quizá pueda sonar algo confuso pero pensemos que sí podemos afirmar con un 95% de probabilidad que nuestro valor estimado estará

dentro de nuestro intervalo, el restante 5% será la probabilidad de que el valor estimado, no esté en ese intervalo.

Ejemplo:

El estadístico pivote utilizado para el cálculo sería el siguiente:

$$\frac{\bar{X}-\mu}{6/\sqrt{n}} \sim N(0,1)$$

$\bar{X}$ = Media muestral
μ = Media poblacional
6 = Desviación típica (conocida)
$\sqrt{n}$ = Raíz cuadrada del tamaño muestral

El intervalo resultante está expresado a continuación:

$$P\left(\bar{X}-6/\sqrt{n} * Z\,\alpha/2 < \mu < \bar{X}+6/\sqrt{n} * Z\,\alpha/2\right) = 1-\alpha$$

Si nos fijamos en el intervalo de la izquierda y de la derecha de la desigualdad, podemos notar que tenemos la cota inferior y superior respectivamente. Por tanto, la expresión nos dice que la probabilidad de que la media de la población se encuentre entre esos dos valores es de 1-alfa (nivel de confianza).

Para entender mejor, vamos a resolver un ejemplo relacionado con el concepto de nivel de confianza.

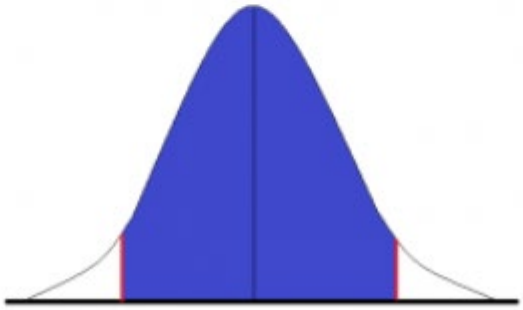
Imaginemos que queremos realizar una estimación del tiempo medio en el que un corredor recorre una maratón. Para tal fin, hemos cronometrado 10 maratones y hemos obtenido una media de 4 horas con una desviación estándar de 33 minutos (lo que en realidad serían 0,55 horas). Nos proponen obtener un intervalo con un nivel de confianza del 95%.

$\bar{X}$ = 4
μ = ¿?
6 = 0,55 (ya que está medida en minutos)
n = Tamaño muestral
$Z\,\alpha/2$ = Valor que deja a la derecha un valor de $\alpha/2$

Dado que lo que queremos es obtener el intervalo, lo que debemos hacer es sustituir las variables en la fórmula que hemos presentado anteriormente:

$$\left(4 - {}^{0,55}\Big/_{\sqrt{10}} * 1,96 < \mu < \bar{4} + {}^{0,55}\Big/_{\sqrt{10}} * 1,96\right)$$

$$(3,7 < \mu < 4,3)$$



El intervalo de confianza está representado en la imagen anterior con el color azul. Los dos valores acotados por este son los correspondientes a las dos líneas de color rojo. La línea central es la media de la población.

En conclusión, con un nivel de confianza del 95% podemos afirmar que el tiempo medio en el que recorrerá la maratón se encontrará entre 3,7 horas y 4,3 horas.

Elección del tamaño muestral

Una encuesta es realmente valiosa cuando es confiable y representativa. Uno de los factores para lograr esto es el tamaño de muestra, encontrar a nuestra población ideal puede resultar verdaderamente difícil.

Existen diferentes aspectos que tienes que considerar para obtener la muestra correcta para tu investigación.

Una muestra es una selección de los encuestados elegidos y que representan a la población total. El tamaño de la muestra es una porción significativa de la población que cumple con las características de la investigación reduciendo los costos y el tiempo.

¿Qué es el tamaño de la muestra?

Saber cómo determinar el tamaño de la muestra antes de comenzar una investigación es un principio estadístico que nos ayuda a evitar el sesgo en la interpretación de los resultados obtenidos.

¿Cómo calcular el tamaño de la muestra?

El tamaño de la muestra de una encuesta es muy importante para poder realizar una investigación de manera correcta, por lo que hay que tener en cuenta los objetivos y las circunstancias en que se desarrolle la investigación.

Recuerda que la finalidad es que las personas completen la encuesta y te otorguen los datos que estás buscando.

Una muestra demasiado grande dará lugar a la pérdida de valiosos recursos como tiempo y dinero, mientras que una muestra pequeña puede no proporcionar información confiable.

¿Entonces de qué tamaño debe ser una muestra? Esto sin duda depende de qué tan exactos necesites que sean los datos obtenidos en tu encuesta, que tan cercanos quieres que sean a los de la población total.

El tamaño de la muestra puede ser:

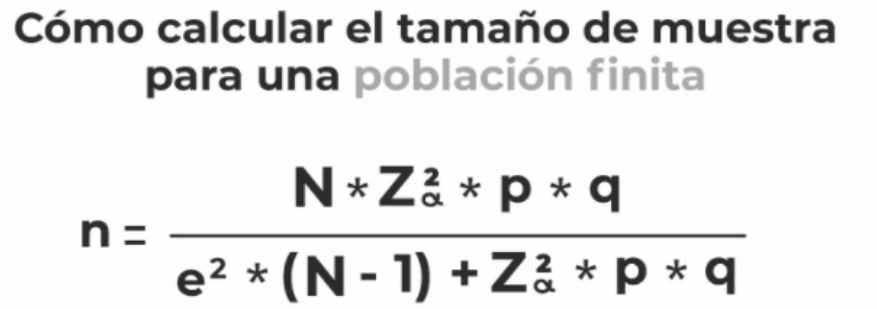
Representativa: Hace referencia a que todos los miembros de un grupo de personas tengan las mismas oportunidades de participar en la investigación.

Adecuada: Se refiere a que el tamaño de la muestra debe de ser obtenido mediante un análisis que permite resultados como disminuir el margen de error.

Ejemplo:
Si quieres realizar una investigación dentro de una universidad que ofrece 10 carreras diferentes y cada una tiene 700 alumnos, no querrás hacer 7000 mil encuestas, bastará con determinar el tamaño de la muestra. Sin embargo, debemos considerar el margen de error.

Ejemplo de cómo calcular el tamaño de muestra finita

Es hora de aprender a determinar el tamaño de la muestra mediante un cálculo de la misma.



n = Tamaño de muestra buscado

N = Tamaño de la Población o Universo

z = Parámetro estadístico que depende el Nivel de Confianza (NC)

e = Error de estimación máximo aceptado

p = Probabilidad de que ocurra el evento estudiado (éxito)

q = (1 - p) = Probabilidad de que no ocurra el evento estudiado

Tu nivel de confianza corresponde a una puntuación Z. Este es un valor constante necesario para esta ecuación. Aquí están las puntuaciones Z para los niveles de confianza más comunes:

90% - Puntuación Z = 1,645

95% - Puntuación Z = 1.96

99% - Puntuación Z = 2.576

Supongamos que nos piden calcular el tamaño para una población de 543.098 consumidores de una marca de bebidas energéticas, donde el investigador asigna un nivel de confianza de 95% y un margen de error de 3%. Donde se desconoce la probabilidad “p” del evento.

Basándonos en este ejemplo, y en nuestra fórmula, el "N" será 543.098, nuestro Z será 1.96 (recuerda que el investigador asignó un nivel de confianza de 95%) y “e” será de 3%. Y como nuestro ejemplo dice que se desconoce la probabilidad de que ocurra el evento, se asigna un 50% a "p" y un 50% a "q".

El resultado de nuestro tamaño de muestra sería: 1065.2, y tendría que ser redondeado pues estamos hablando de personas.

Ejemplo de cómo calcular el tamaño de muestra infinita

Si necesitas calcular el tamaño de muestra de una población desconocida, donde el investigador necesite un nivel de confianza del 95%, un margen de error del 3% y se desconoce la probabilidad “p” del evento que se está estudiando, sigue la siguiente fórmula:

Formula para calcular el tamaño de muestra infinita

$$n = \frac{Z_{\alpha}^{2} * p * q}{e^{2}}$$

n = Tamaño de muestra buscado

N = Tamaño de la Población o Universo

z = Parámetro estadístico que depende el Nivel de Confianza (NC)

e = Error de estimación máximo aceptado

p = Probabilidad de que ocurra el evento estudiado (éxito)

q = (1 - p) = Probabilidad de que no ocurra el evento estudiado

Donde "Z" es el intervalo de confianza al cuadrado, en este caso se pide que sea del 95%, lo que indica que sería 1,96 al cuadrado, y cómo no sabemos la probabilidad de que ocurra el evento, "p" y "q" sería 50%.

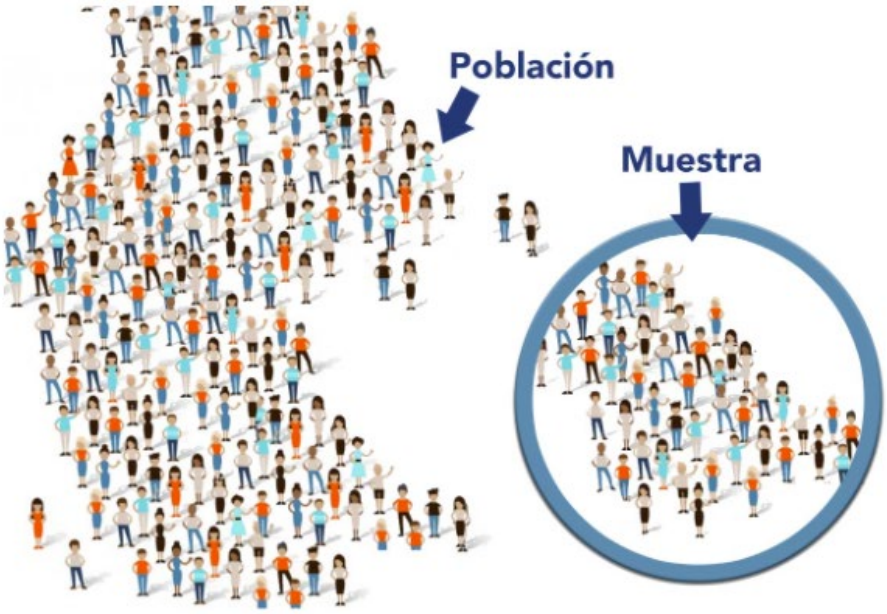
Entre el margen de error al solicitado al cuadrado. El resultado sería 1067,11, que también lo debemos redondear.

Para calcular el número de participantes que necesitas en tu próxima investigación de manera más rápida y fácil te invitamos a usar nuestra calculadora de muestra para que agilices este proceso.

¿Cómo podemos tener una muestra representativa y adecuada?

Es mucho mejor tener a las personas adecuadas para contestar nuestra encuesta, que tener una gran cantidad de personas equivocadas que no nos van a aportar la información que necesitamos.

Una muestra representativa está integrada por personas con intereses similares a nuestro objeto de estudio, no tiene que ver, en este caso, con el tamaño.



Lo ideal es poder seleccionar a los encuestados de una población representativa de una manera aleatoria, por ejemplo, seleccionar a cada 5 miembros de una lista de correos de usuarios que sean realmente representativos para nuestra investigación.

Una vez que tengas la muestra adecuada, hay que decidir el tamaño de la muestra que desees estudiar. Cuanto más precisa quieres que sea, más grande debe ser el tamaño.

Referencias

[1] Inferencia estadística.

https://www.sdelsol.com/glosario/interferencia-estadistica/

https://www.superprof.es/apuntes/escolar/matematicas/estadistica/inferencia/ejercicios-y-problemas-de-inferencia-estadistica.html

[2] Intervalo de confianza.

https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/IntervalosConfianza/3IntervalodeConfianza.html

[3] Intervalo de confianza para la media con desviación típica poblacional conocida

https://www.youtube.com/embed/QC79MnHLYLU

[4] Intervalo de confianza para la media con desviación típica poblacional desconocida

https://www.youtube.com/embed/aVNXCgR9eYA

[5] Determinar tamaño de una muestra.

https://www.questionpro.com/blog/es/como-determinar-el-tamano-de-una-muestra/

[6] Calculadora de Muestra

https://www.questionpro.com/es/calculadora-de-muestra.html