INSTITUTO de ENSEÑANZAS a DISTANCIA de ANDALUCÍA

2º de Bachillerato

Matemáticas Aplicadas a las Ciencias Sociales II

Contenidos

Estadística inferencial: Intervalos de confianza II

1. Introducción

El objetivo básico de la Inferencia Estadística es hacer inferencias o sacar conclusiones sobre la población a partir de la información contenida en una muestra aleatoria de la población.

Cualquier inferencia o conclusión obtenida de la población, necesariamente, estará basada en un estadístico muestral, es decir, en la información proporcionada por la muestra

Las inferencias sobre el valor de un parámetro poblacional, como la media de una población μ, se pueden obtener básicamente de dos maneras: a partir de la estimación o a partir del contraste de hipótesis.

En este tema, como en el anterior, nosotros vamos a estudiar la estimación estadística, que se divide en dos grandes grupos: la estimación puntual y la estimación por intervalos. La estimación puntual consiste en obtener un único número calculado a partir de las observaciones muestrales, y que es utilizado como estimación del valor del parámetro μ. Se le llama estimación puntual porque a ese número, que se utiliza como estimación del parámetro μ, se le puede asignar un punto sobre la recta real. En la estimación por intervalos se obtienen dos puntos (un extremo inferior y un extremo superior) que definen un intervalo sobre la recta real, el cual contendrá con cierta seguridad el valor del parámetro μ.

Curiosidad

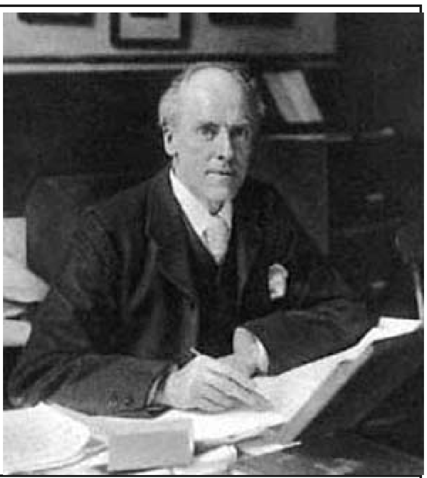


Imagen de Karl Pearson en Wikimedia Commons

Karl Pearson (Londres 27 de marzo de 1857- Londres, 27 de abril de 1936) fue un prominente científico, matemático y pensador británico, que estableció la disciplina de la *estadística matemática*. Desarrolló una intensa investigación sobre la aplicación de los métodos estadísticos en la biología y fue el fundador de la bioestadística.

En 1901, junto con Galton y Walter Frank Raphael Weldon, fundó la revista *Biometrika*, desde entonces una de las más importantes en el campo de la estadística.

En 1911 creó el primer departamento de estadística en la Universidad de Londres, en el que Pearson incorporó el laboratorio biométrico y el laboratorio de Francis Galton.

Cuando Einstein -de 23 años- comenzó un grupo de estudio sugirió que el primer libro a leer debía ser *La Gramática de la Ciencia*, de Pearson. Este libro trataba sobre varios temas que más tarde se convirtieron en parte de las teorías de Einstein y otros científicos.

Pearson aseveró que las leyes de la naturaleza son relativas a la habilidad perceptiva del observador. La irreversibilidad de los procesos naturales, decía Pearson, es puramente una concepción relativa. Un observador que viaja a exactamente la velocidad de la luz vería un eterno momento, o una ausencia de movimiento.

2. Estimación puntual

Imagen de YamilGonzales con licencia Creative Commons

A nuestros amigos de la empresa TisBet Survey les ha encargado el Ayuntamiento de su ciudad, un estudio sobre el Cociente Intelectual (C.I.) medio de los alumnos y alumnas del IES "Benito V.", ya que están elaborando un proyecto de atención a alumnado con altas capacidades.

Como no conocen la media del C.I. del alumnado, toman una muestra aleatoria con alumnos y alumnas del instituto, a través de la cual calcularán una aproximación de la media.

Esta aproximación es lo que llamamos **estimación**.

Cuando tenemos que hacer el estudio de una determinada característica de una población y no se puede realizar el estudio a todos los individuos que la componen, bien porque son muchos y es imposible realizar el estudio para todos o porque económicamente supone un gasto excesivo. En este caso, como vimos en la unidad anterior, seleccionamos la muestra y realizamos el estudio para los individuos escogidos.

¿Qué ocurre una vez calculado el C.I. medio de la muestra?

Supongamos que los trabajadores de TisBet Survey han seleccionado una muestra de 100 alumnos y alumnas y han obtenido que la media muestral es de 106 ($\overline{x} = 106$).

Parece lógico estimar que la media μ de todo el alumnado del instituto será aproximadamente igual que la media de la muestra, 106. El hecho de decir que el valor de μ es aproximadamente $\overline{x}$, significa que estamos haciendo una **estimación puntual**.

El resultado que hemos obtenido lo hemos generalizado a toda la población, esto es lo que llamamos inferir los resultados de la muestra a la población.

Importante

Una **estimación puntual** del valor de un parámetro poblacional desconocido, como puede ser la media μ o la desviación típica σ, es un número que se utiliza para aproximar el verdadero valor de dicho parámetro poblacional.

Para calcular ese número, tomaremos una muestra de la población y calcularemos el estimador muestral asociado al parámetro poblacional. Por ejemplo el estimador de la media poblacional μ, es la media muestral $\overline{X}$.

El valor de este estimador muestral para una muestra concreta será la

estimación puntual del parámetro poblacional.

Por lo tanto para estimar un parámetro tenemos que saber cuál es el mejor estimador muestral de ese parámetro.

En el tema anterior hemos visto que el estimador del parámetro p de una distribución binomial es $\hat{p}$, la proporción muestral.

En este tema vamos a usar como estimadores muestrales de la media μ y la diferencia de las medias μ_1-μ_2 de una población a los estadísticos $\overline{X}$ y $\overline{X}_1-\overline{X}_2$, respectivamente.

¿Cómo sabemos que son los mejores? Porque tienen unas propiedades determinadas que tiene que tener todo estadístico para ser el que mejor aproxima al verdadero valor de la variable.

Como vimos en la unidad anterior los estadísticos son variables aleatorias (cuando realizo el estudio del C.I. medio de una muestra, no sé que valor va a salir) y, como tales, tienen sus propios parámetros. Por lo tanto, podemos hablar de la media del estadístico $\mu_{\overline{X}}$.

Pues bien, cuando la media del estadístico coincide con el verdadero valor del parámetro diremos que el **estadístico es centrado o insesgado**.

Eso le ocurre a los estadísticos anteriores. Pero tenemos un problema con la varianza y es que la varianza muestral no coincide con el valor de σ. Por ello el estadístico insesgado de σ^2 es:

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

llamado **cuasivarianza muestral**. Observa que es muy parecido a la varianza muestral, la diferencia es que está dividido por n-1 en vez de n.

Como consecuencia de este resultado la estimación de la desviación típica es la raíz cuadrada de ese valor, es decir, $\hat{S}$

Importante

La diferencia entre el verdadero valor del parámetro que se estima y la media del estimador, mide el error cometido al utilizar el estimador y se denomina **sesgo**.

Un estadístico es insesgado cuando la media de su distribución muestral asociada coincide con la media de la población.

Esto ocurre, por ejemplo, con el estimador $\overline{X}$ ya que su media muestral coincide con μ. En el caso de la varianza σ^2, el estimador insesgado es la cuasivarianza muestral $\hat{S}^2$.

Veamos una tabla que nos aclare un poco todo este lío:

Parámetro poblacional	Estimador muestral o Estadístico	Estimación puntual (Valor concreto)
Media μ	$\hat{\mu}=\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	$\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}$
Varianza σ^2	$\hat{\sigma}^2=\hat{S}^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	$\hat{s}^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$

Ejercicio resuelto

La empresa TisBet Survey quiere estimar la media de estatura de los alumnos y alumnas de 1º de ESO en el IES "Benito V."

Para ello, escoge una muestra al azar de 10 alumnos y los mide. Los resultados en cm fueron:

160, 170, 170, 150, 160, 180, 160, 170, 130, 150.

Haz una estimación puntual de la media y la varianza.

Imagen de tiffa 130 con licencia Creative Commons

Mostrar retroalimentación

Para estimar la media, utilizamos el parámetro de la media muestral:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{160+170+170+150+160+180+160+170+130+150}{10} = 160\ cm$$

.

Luego la media poblacional es μ=160 cm.

Para estimar la varianza, utilizamos el parámetro de la cuasivarianza muestral:

$$\hat{s}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\hat{s}^2 = \frac{0^2+10^2+10^2+10^2+0^2+20^2+0^2+10^2+30^2+10^2}{9} = 200$$

Comprueba lo aprendido

Blanco

Rellena los espacios en blanco con las siguientes palabras:

Parámetro, Estadístico, Estimador y Estimación.

________ puntual : Es el estadístico que se usa para estimar un parámetro poblacional. Por lo tanto, es una variable aleatoria en el muestreo que tiene su correspondiente distribución muestral.

_______ : Es un valor numérico que describe una característica de la población.

_______ puntual : Es el valor numérico que toma el estimador puntual para una muestra determinada.

_______ : Es un valor numérico que describe una característica de la muestra.

Enviar

Para saber más

Cuando dos estimadores de un parámetro tengan la misma media, el que tenga menor varianza será el estimador que escojamos. La varianza de un estimador muestral nos va a medir la eficiencia de ese estadístico. Cuando en un estimador la varianza es mínima, decimos que el estimador es **eficiente**.

La desviación típica del estimador se va a llamar **error estándar.** En el caso del estimador $\overline{X}$, su error estándar es $\frac{\sigma}{\sqrt{n}}$.

Cuanto mayor sea el tamaño de la muestra **n**, menor será la varianza del estimador $\overline{X}$ y , por tanto, mejor serán nuestras estimaciones.

3. Intervalo de confianza para la media de una población

La obsolescencia programada nos indica que se fabrican productos tecnológicos con fecha de caducidad interna para poder mantener el sistema de producción. El vídeo que te mostramos corresponde al documental "Comprar, tirar, comprar" emitido en la noche temática de La 2 (aquí puedes verlo completo)

Para comprobar la veracidad de este problema una asociación de consumidores ha contratado a nuestra empresa TisBet Survey para que haga un estudio de la vida media útil de un pequeño electrodoméstico.

Nuestra empresa realiza un estudio con 100 personas que han comprado dicho aparato. En este caso, se obtiene una media de vida útil de 880 días y una desviación típica σ de 180 días.

Los técnicos de TisBet Survey han decidido dar como respuesta a la asociación un intervalo de vida media útil. Es decir, algo parecido a "la vida media útil de este electrodoméstico está entre 2,5 años y 2,6 años con una probabilidad del 95%".

¿Por qué lo hacen así? La estimación puntual es poco útil, pues solo obtenemos un valor como aproximación al que tratamos de estimar. Es mucho más interesante obtener un intervalo dentro del cual se tiene una cierta confianza de que se encuentre el parámetro que tratamos de estimar. Estos intervalos ya los habíamos estudiado en el tema anterior cuando calculábamos la estimación de la proporción de un determinado aspecto de una población.

El objetivo que se pretende con los intervalos de confianza es obtener un intervalo de poca amplitud y con una alta probabilidad de que el parámetro que buscamos, en este caso μ, se encuentre en su interior. Así pues, elegiremos probabilidades cercanas a la unidad, que se representan por 1-α y cuyos valores más frecuentes suelen ser 0,90; 0,95 y 0,99.

Veamos como sería en el caso del estudio de la media, es decir, vamos a estimar mediante un intervalo el parámetro media poblacional **μ**, suponiendo que la población de partida es N(μ, σ). Recuerda que el Teorema central del límite nos permite suponer que esto es cierto para valores del tamaño muestral superiores a 30 individuos.

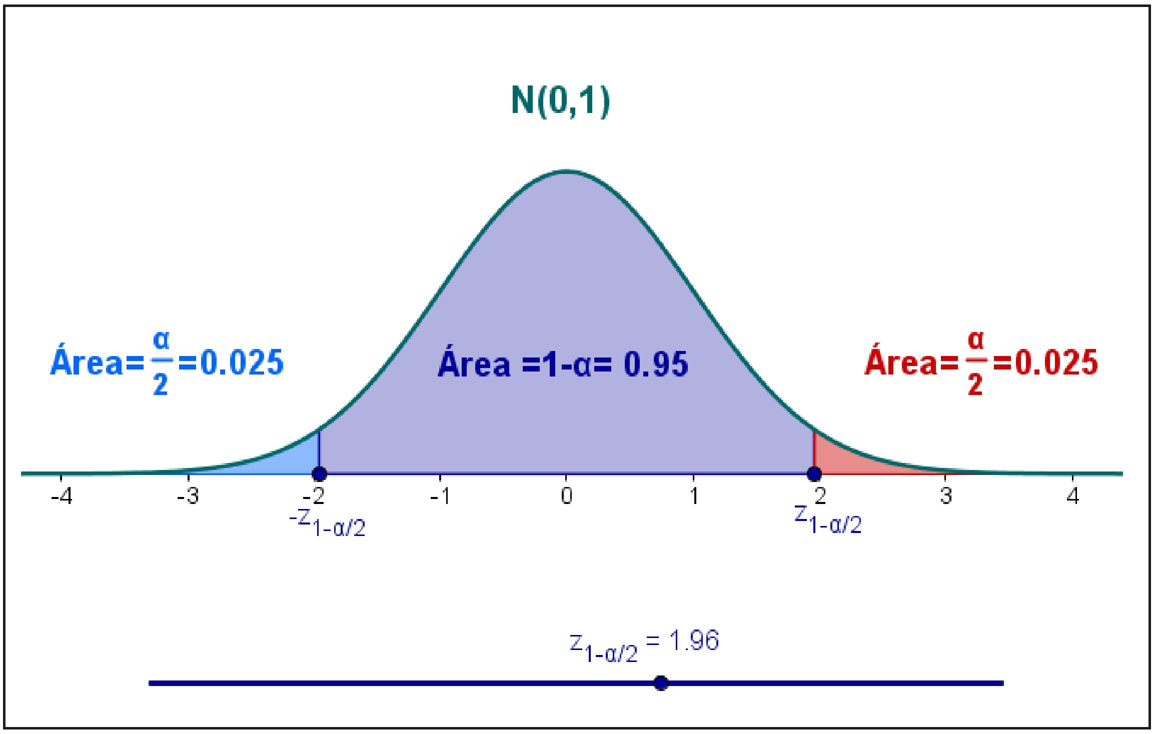
Recordemos que para una muestra de tamaño n suficientemente grande la distribución del estadístico $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Veamos como conseguir nuestro intervalo. Tipificando la variable aleatoria $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$, que se aproxima a una N(0,1).

Queremos que la vida media útil de nuestro electrodoméstico esté en el intervalo que vamos a construir con una probabilidad del 95%. A este valor, en tanto por uno, lo vamos a llamar 1-α, es decir, **α=0,05**.

Como queremos que el intervalo esté centrado en la media (que como está tipificada la variable es 0) tendremos que el intervalo tiene los extremos iguales en valor absoluto:

$$P(-a<Z<a)=1-\alpha$$



Como la distribución N(0,1) siempre toma los mismos valores, el valor que buscamos tiene que dejar a la izquierda de z=a exactamente un área de α/2+(1-α)=0,025+0,95=0,975. Vamos a llamar a ese número valor crítico y lo vamos a representar como $\mathbf{a=z_{1-\alpha/2}}$. Date cuenta que el subíndice de z coincide con la probabilidad que buscaremos en la tabla de la N(0,1). Con ayuda de esta tabla podemos ver que $\mathbf{z_{0,975}=1,96}$. El intervalo de confianza tipificado es $\mathbf{(-z_{0,975},z_{0,975})=(-1.96,1.96)}$ **.** Veamos como se calcula el valor crítico en la siguiente película:

Vamos a sustituir el valor de Z en la fórmula original, para calcular el intervalo centrado en la media estimada, por lo que tendremos:

$$P(-a<Z<a)=P\left(-z_{1-\frac{\alpha}{2}}<\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<z_{1-\frac{\alpha}{2}}\right)\Rightarrow$$

$$\Rightarrow P\left(-z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}<\overline{X}-\mu<z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)\Rightarrow P\left(\overline{x}-z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}<\mu<\overline{x}+z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

Importante

El **intervalo de confianza para el parámetro media μ de una población N(μ, σ)** al nivel de significación α viene dado por:

$$IC=\left(\overline{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

Si el nivel de significación o de riesgo es α, el nivel de confianza será de 1-α.

Al valor de la abscisa que deja a su izquierda un área igual a $1-\frac{\alpha}{2}$ siendo 1 - α el nivel de confianza se le llama valor crítico.

Si σ es desconocida y n es grande (n ≥30), el intervalo de confianza viene dado por:

$$IC=\left(\overline{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\hat{s}}{\sqrt{n}}\right)$$

donde $\hat{s}^2$ es la cuasivarianza muestral.

Para saber más

El siguiente vídeo dura 38:34 min, pero te aconsejamos que no dejes de verlo. En él puedes ver un resumen de los que llevamos visto hasta ahora y una explicación práctica sobre intervalos de confianza que aclaran muy mucho en que consisten.

El vídeo está realizado por el profesor D. Francisco Javier Barón López del Departamento de Bioestadística de la Universidad de Málaga.

Puedes ver el vídeo pinchando en el siguiente enlace.

Reflexiona

¿Recuerdas el problema con el que habíamos iniciado este punto sobre la obsolescencia programada? Habíamos dejado a nuestros amigos de TisBet Survey con un trabajo a medio hacer.

Nuestra empresa realizó un estudio con 100 personas que habían comprado un determinado electrodoméstico.

Obteniendo una duración media en la vida útil del aparato de 880 días. La varianza era de σ=180 días. Realizaremos el estudio para un nivel de confianza del 95%.

Tenemos que calcular el intervalo de confianza para los datos obtenidos.

Imagen de Lee J Haywood con licencia Creative Commons

Mostrar retroalimentación

Como nos piden un 95% de nivel de confianza, esto quiere decir que $\alpha=0{,}05$ y por lo tanto $z_{1-\alpha/2}=z_{0{,}975}=1{,}96$, por lo que sustituyendo el intervalo quedará como sigue:

$$IC=\left(880\pm 1{,}96\cdot\frac{180}{\sqrt{100}}\right)=(880\pm 1{,}96\cdot 18)=(880\pm 35{,}28)=(844{,}72\,;\,915{,}28)$$

Podemos afirmar que la vida útil media de un aparato estará comprendida entre 845 y 915 días con una probabilidad del 95%.

Comprueba lo aprendido

Blanco

Tienes que realizar una tabla con los valores críticos $z_{1-\alpha/2}$ más usuales. Es decir. Para un nivel de confianza del 80%, 90%, 95% y 99%, en cada caso tendrás que calcular α, $\alpha/2$ y $z_{1-\alpha/2}$.

Puedes ayudarte de la tabla de la normal N(0,1)

Nivel de Confianza	80%	90%	95%	99%
$1-\alpha$			0,95	
α			0,05	
$1-\alpha/2$			0,975	
$z_{1-\alpha/2}$			1,960	

Enviar

3.1. Error y tamaño de la muestra

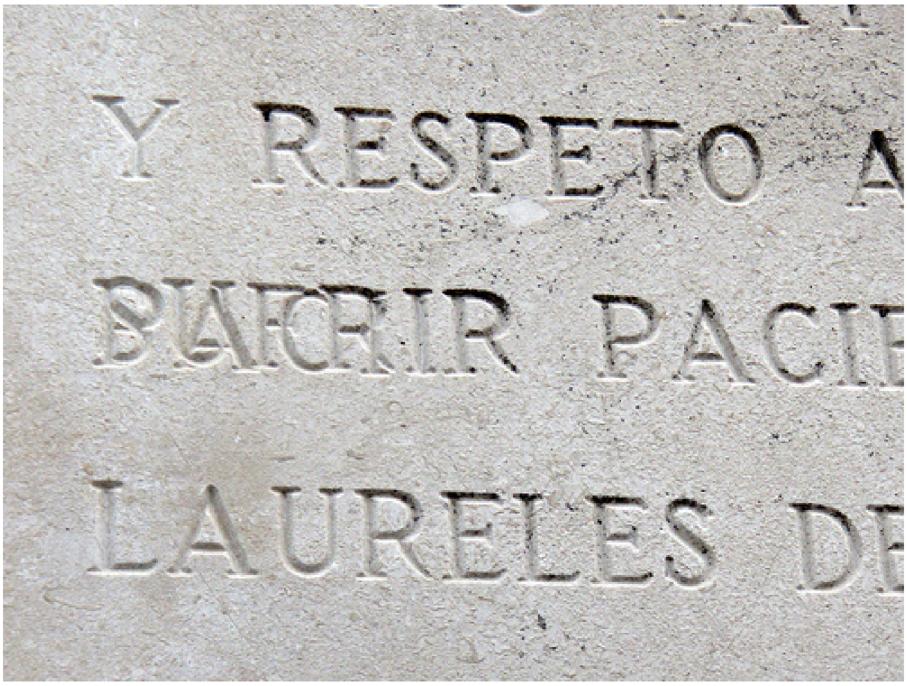


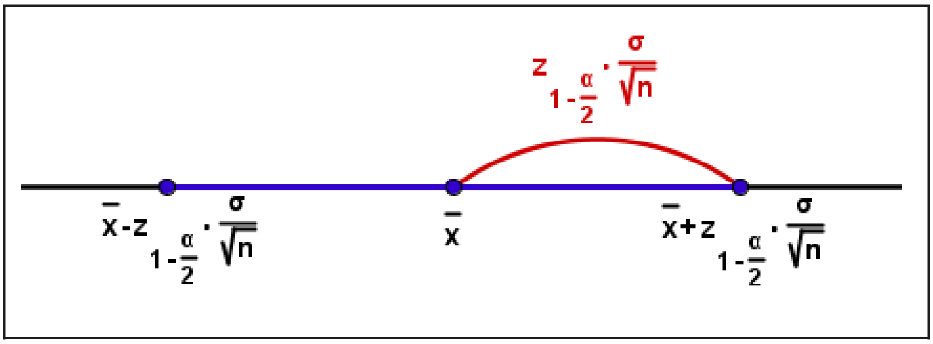
Imagen de Daquella manera con licencia Creative Commons

Los errores aparecen y más cuando estamos realizando cálculos. En este caso, saber cuál es el error que queremos cometer de antemano, nos va a ayudar a decidir cual es el tamaño de la muestra que necesitamos para equivocarnos menos de ese error.

Cuando calculamos un intervalo de confianza para la media estamos indicando un intervalo donde puede que se encuentre la media de la población con una cierta probabilidad, que hemos llamado nivel de confianza.

Es evidente que se comete un error. Puede que la media real de la población esté fuera de ese intervalo. ¿Cuál es el error máximo admisible?

La forma de nuestro intervalo cuando estamos infiriendo la media está centrado en $\bar{x}$ y le sumamos y restamos una misma cantidad:



El error máximo que se comete, E, es el radio del intervalo centrado en la media y el margen de error es la amplitud del intervalo, 2E.

¡Lo mismo que en el tema anterior, vaya!

Importante

El **Error máximo admisible** en el cálculo de un intervalo de confianza para la media es:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

El margen de error admisible es la amplitud del intervalo, 2E.

Si queremos disminuir el error una vez escogido el nivel de significación la única variable que nos queda, en la fórmula anterior del Error máximo admisible, es el tamaño de la muestra. A mayor tamaño muestral el Error disminuirá.

Por lo tanto, si sé cuál es el error máximo que quiero cometer, puedo saber cuál es el tamaño de la muestra que tengo que seleccionar para conseguirlo. Basta con despejar n en la fórmula anterior, con lo que nos queda:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow E^2 = z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{n} \Rightarrow n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2}{E^2} \Rightarrow n = \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{E}\right]^2$$

¡Muy parecida a la formula del error máximo para un intervalo de confianza con la proporción!

Importante

El **tamaño** de la muestra para que en el intervalo de confianza se cometa un **error prefijado** al nivel de significación α es:

$$n = \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{E}\right]^2$$

Ejercicio resuelto

Imagen de Ignacio Sanz con licencia Creative Commons

En una comunidad autónoma se está haciendo un estudio sobre el número de días que dura un contrato temporal.

Se sabe que la desviación típica de los contratos es igual a 57 días.

El estudio se quiere hacer con un nivel de confianza del 95%.

Indica el número mínimo de contratos en los que se ha de mirar su duración para que el intervalo que da la duración media de un contrato de ese tipo tenga una amplitud no mayor de 10 días.

Mostrar retroalimentación

Del enunciado deducimos que:

σ =57 días.

α=0,05 por lo que $z_{1-\alpha/2}$=1,96

Como la amplitud del intervalo ≤ 10 días ⇒ Error máximo admisible: E ≤ 5.

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 5 \;\Rightarrow\; 1{,}96 \cdot \frac{57}{\sqrt{n}} \leq 5$$

Despejando se obtiene que:

$$n \geq \left[\frac{1{,}96 \cdot 57}{5}\right]^2 \;\Rightarrow\; n \geq 499{,}254$$

Por lo tanto, se ha de mirar la duración media de, al menos, 500 contratos temporales.

Reflexiona

La longitud de los cables de los auriculares que fabrica una empresa es una variable aleatoria que sigue una ley Normal con desviación típica 4,5 cm. Para estimar la longitud media se han medido los cables de una muestra aleatoria de 9 auriculares y se han obtenido las siguientes longitudes, en cm:

205, 198, 202, 204, 197, 195, 196, 201, 202.

a) Halla un intervalo de confianza, al 97%, para la longitud media de los cables.

b) Determina el tamaño mínimo que debe tener una muestra de estos auriculares para que el error de estimación de la longitud media sea inferior a 1 cm, con el mismo nivel de confianza del apartado anterior.

Imagen de Aditza 121 con licencia Creative Commons

Mostrar retroalimentación

a) Recordemos que el intervalo de confianza para la media con desviación típica conocida tiene la siguiente fórmula:

$$IC = \left(\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Empecemos calculando la estimación de la media:

$$\bar{x} = \frac{205+198+202+204+197+195+196+201+202}{9} = \frac{1800}{9} = 200$$

Calculemos el valor crítico. Como el nivel de confianza es del 97% ⇒ 1-α es 0,97 ⇒ α es 0,03 ⇒ α/2 es 0,015 ⇒ 1-α/2 es 0,985 ⇒ $z_{1-\alpha/2} = z_{0,985} = 2{,}17$

Por lo que:

$$IC = \left(\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(200 \pm 2{,}17 \cdot \frac{4{,}5}{3}\right) = \left(200 \pm 2{,}17 \cdot 1{,}5\right) =$$
$$= \left(200 \pm 3{,}255\right) = \left(196{,}745 \;;\; 203{,}255\right)$$

La longitud media de los cables se encontrará en el intervalo (196,745 ; 203,255) con un nivel de confianza del 97%

b) Nos piden que el error debe ser menor que 1 cm. Por lo que:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 1 \;\Rightarrow\; 2{,}17 \cdot \frac{4{,}5}{\sqrt{n}} \leq 1$$

Despejando se obtiene que:

$$n \geq \left[\frac{2{,}17 \cdot 4{,}5}{1}\right]^2 \Rightarrow n \geq 95{,}355$$

Por lo tanto, se ha de tomar una muestra de, al menos, 96 auriculares para que el error de estimación sea inferior a 1 cm.

Curiosidad

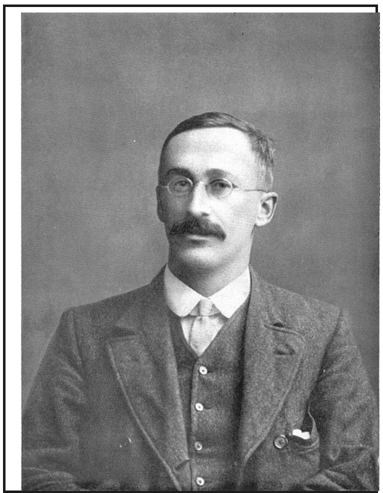


Imagen de William Sealy Gosset de Wikimedia Commons

William Sealy Gosset (11 de junio de 1876 – 16 de octubre de 1937) fue un estadístico, mejor conocido por su sobrenombre literario ***Student***. Asistió a la famosa escuela privada Winchester College, antes de estudiar química y matemáticas en el New College de Oxford. Tras graduarse en 1899, se incorporó a las destilerías Guinness en Dublín.

Guinness era un negocio agroquímico progresista y Gosset podría aplicar sus conocimientos estadísticos tanto a la destilería como a la granja (para seleccionar las mejores variedades de cebada).

Para evitar el espionaje industrial, Guinness había prohibido a sus empleados la publicación de artículos independientemente de la información que contuviesen.

Esto significaba que Gosset no podía publicar su trabajo usando su propio nombre. De ahí el uso de su seudónimo *Student* en sus publicaciones, para evitar que su empleador lo detectara. Por tanto, su logro más famoso se conoce ahora como la distribución t de Student, que de otra manera hubiera sido la distribución t de Gosset.

La distribución t de Student se usa como distribución del estimador muestral de la media cuando la muestra tiene muy pocos individuos, es decir, n es muy pequeño.

4. Intervalo de confianza para la diferencia de medias

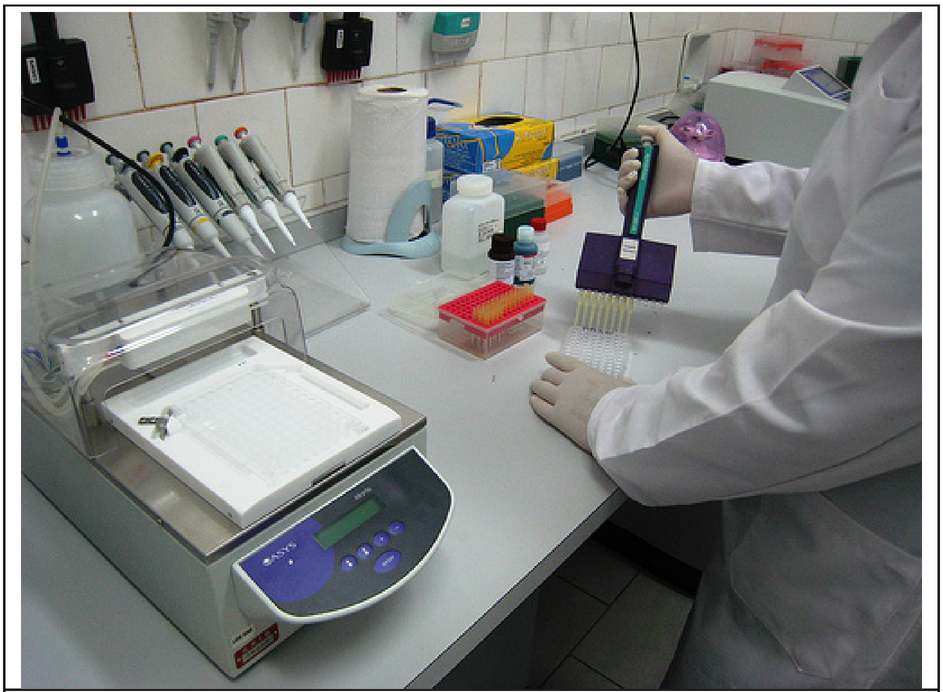


Imagen de Agriculturasp con licencia Creative Commons

Comparar es algo fundamental en algunos casos. Cuando hacemos investigaciones científicas comparar nos va a servir para ver cuál de las posibilidades de que dispongo es la mejor o la más eficiente.

Podemos ver un ejemplo de esto en los estudios de laboratorio de un determinado medicamento. El mismo medicamento puede ser inoculado a dos series de ratones en distintas condiciones para ver en cual de ellas se obtienen mejores resultados.

Otro ejemplo lo podemos ver en la medida de resultados académicos de dos centros de estudio distintos. Se eligen dos muestras al azar de cada uno de ellos y se les pasa el mismo test para estudiar en cual de los dos los resultados académicos son mejores.

Los políticos, por interés propio, junto con los medios de comunicación encargados de cubrir estas noticias son dados a encargar encuestas a empresas como nuestra TisBet Survey para comparar, frente a unas próximas elecciones, la diferencia de intención de voto entre los dos partidos políticos principales de una ciudad, comunidad autónoma o de todo el país.

En general los estudios irán encaminados a comparar las medias de dos poblaciones referidas a un mismo aspecto. Comparar las medias equivale a hacer un estudio de la diferencia de las mismas, para ver cual de los resultados es mejor. En cualquier caso los estudios se deben hacer escogiendo una muestra de un determinado tamaño de cada población y aplicar las matemáticas para inferir esos resultados a toda la población de la forma más fiable posible.

Obtendremos un intervalo de confianza en el cual se encontrará la diferencia de las medias con una cierta probabilidad que llamaremos nivel de confianza. Generalmente, en los estudios, este nivel de confianza será del 90%, 95% o 99%. Evidentemente, a mayor nivel de confianza mayor será el intervalo obtenido.

En este caso tenemos que comparar las medias de dos poblaciones. Para ello tendremos que usar el estadístico que estudiamos en el tema del Teorema Central del Límite para la diferencia de las medias.

Importante

Sean dos poblaciones $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$; de cada una de ellas se elige una muestra de tamaños n_1 y n_2, respectivamente de tamaños n_1 y n_2, respectivamente.

Sean $\overline{x}_1$ y $\overline{x}_2$ las medias de la primera y de la segunda muestra respectivamente. Para un nivel de significación α, se tiene que

- Si σ_1 y σ_2 son conocida, el intervalo de confianza viene dado por:

$$IC=\left[\overline{x}_1-\overline{x}_2 \pm z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

- Si σ_1 y σ_2 son desconocidas y n_1 y n_2 son grandes ($n_1 \geq 30$ y $n_2 \geq 30$), el intervalo de confianza viene dado por:

$$IC=\left[\overline{x}_1-\overline{x}_2 \pm z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\tilde{s}_1^2}{n_1}+\frac{\tilde{s}_2^2}{n_2}}\right]$$

donde $\tilde{s}_1^2$ y $\tilde{s}_2^2$ son las cuasivarianzas de cada muestra.

Ejercicio resuelto

A nuestras oficinas de TisBet Surbey llegó el siguiente encargo de los laboratorios MIPASTILLA S.A.

Imagen de juanpol con licencia Creative Commons

En sus laboratorios habían realizado un estudio consistente en la medición, en dos grupos de pacientes, de la cantidad de cierta proteína en sangre en pacientes tras dos tratamientos con distintos medicamentos.

Nos solicitaban que realizásemos una comparativa de los resultados clínicos y para ello nos enviaban los datos del experimento.

El estudio consistía en dos muestras de 50 individuos con los siguientes resultados.

	Media($\overline{x}$)	Cuasivarianza ($\tilde{s}^2$)
Medicamento 1	3,2	1,5
Medicamento 2	4	2,3

El encargado de realizar el estudio decide calcular el intervalo de confianza para la diferencia de las medias con un nivel de significación del 95%.

Ayuda a nuestros amigos de TisBet Survey a realizar el estudio ayudándote del intervalo de confianza para la media que acabas de estudiar.

Mostrar retroalimentación

En este caso los datos que tenemos son:

$\overline{x}_1$ =3,2 $\overline{x}_2$ =4 $\tilde{s}_1^2$ =1,5 $\tilde{s}_2^2$ =2,3

Nivel de confianza: 1-α=95% ⇒ Nivel de significación: α=0,05

$z_{1-\alpha/2}=z_{0,975}=1,96$.

$n_1=n_2=50$

Aplicamos la fórmula del intervalo de confianza:

$$IC=\left[\bar{x}_1-\bar{x}_2 \pm z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

y sustituyendo:

$$IC=\left[3{,}2-4 \pm 1{,}96\cdot\sqrt{\frac{1{,}5}{50}+\frac{2{,}3}{50}}\right]=\left[-0{,}8 \pm 1{,}96\cdot 0{,}2757\right]=\left[-0{,}8 \pm 0{,}5404\right]$$

por lo que, el intervalo de confianza quedará **IC=(-1,3404, -0,2596)**

Reflexiona

Imagen de xornalcerto con licencia Creative Commons

Una planta de tratamiento de aguas residuales, recibe diariamente el agua de dos pequeñas poblaciones próximas, Villalta y Villabajo.

Desde la empresa de tratamiento deciden estudiar, para comparar la calidad de las mismas, la concentración de nitratos disueltos. Para ello se extraen dos muestras al azar y se analiza el contenido de nitratos, obteniéndose los siguientes resultados expresados en mg/l.

Población Villalta:

$\bar{x}_1$ =87 mg/l, σ_1=3,19mg/l, n_1=33 l

Población Villabajo:

$\bar{x}_2$ =109 mg/l, σ_2=4,15mg/l, n_2=27 l

Calcula un intervalo de confianza para la diferencia de las medias con un nivel de significación del 10%.

Mostrar retroalimentación

En este caso los datos que tenemos son:

$\bar{x}_1$ =87 $\bar{x}_2$ =109 σ_1=3,19 σ_2=4,15

Nivel de significación del 10% ⇒ α=0,1 ⇒ $\alpha/2$=0,05 ⇒ $1-\alpha/2$=0,95 ⇒ $z_{1-\alpha/2}=z_{0,95}$=1,645

n_1=33 y n_2=27

Aplicamos la fórmula del intervalo de confianza:

$$IC=\left[\bar{x}_1-\bar{x}_2 \pm z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

y sustituyendo:

$$IC=\left[87-109 \pm 1{,}645\cdot\sqrt{\frac{3{,}19}{33}+\frac{4{,}15}{27}}\right]=\left[-22 \pm 1{,}645\cdot 0{,}5004\right]=\left[-22 \pm 0{,}8231\right]$$

por lo que el intervalo de confianza quedará **IC=(-22,8231, -21,1769)**

4.1. Error y tamaño de la muestra

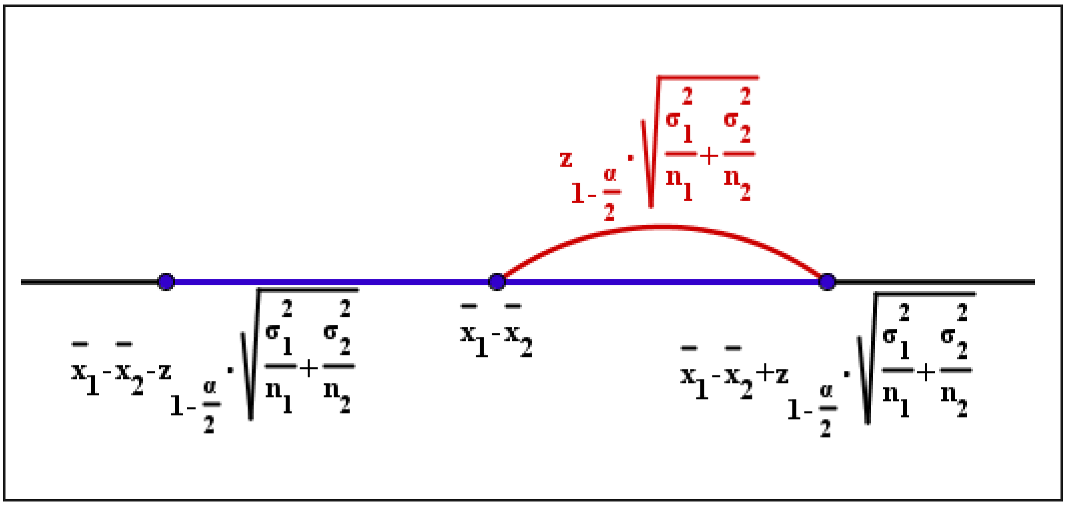
En el siguiente vídeo puedes comprobar que hasta los cineastas más grandes cometen errores

Igual que los cineastas nosotros al calcular el intervalo de confianza para la diferencia de las medias también cometemos errores.

Ya nos pasaba con el intervalo de confianza para la proporción, para el de la media, y cuando calculamos un intervalo de confianza para la diferencia de las medias, también cometemos un error.

El intervalo de confianza nos indica que la diferencia de las medias estará en ese intervalo con una cierta probabilidad, concretamente la que nos marca el nivel de confianza, (1- α). ¿Pero cual es el error máximo admisible?

El intervalo de confianza para la diferencia de las medias podemos representarlo en la recta real de la siguiente forma:



El error máximo que se comete, E, es el radio del intervalo centrado en la media y el margen de error es la amplitud del intervalo, 2E.

Importante

El **Error máximo admisible** en el cálculo de un intervalo de confianza para la diferencia de las medias es:

$$E = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

El margen de error admisible es la amplitud del intervalo, 2E.

Si queremos disminuir el error una vez escogido el nivel de significación la única variable que nos queda, en la fórmula anterior del Error máximo admisible, es el tamaño de la muestra. A mayor tamaño muestral el Error disminuirá.

Además, en este caso vamos a suponer que el tamaño muestral es el mismo en las dos muestras, es decir, $n_1 = n_2 = n$. Por lo que la fórmula nos quedaría:

$$E = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}$$

Por lo tanto, si sé cuál es el error máximo que quiero cometer, puedo saber cuál es el tamaño de la muestra que tengo que seleccionar para conseguirlo. Basta con despejar n en la fórmula anterior, con lo que nos queda:

$$E = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \Rightarrow E^2 = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\sigma_1^2 + \sigma_2^2}{n} \Rightarrow n = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\sigma_1^2 + \sigma_2^2}{E^2} \Rightarrow n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_1^2}{E^2} + \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_2^2}{E^2} \Rightarrow$$

$$\Rightarrow n = \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma_1}{E}\right]^2 + \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma_2}{E}\right]^2$$

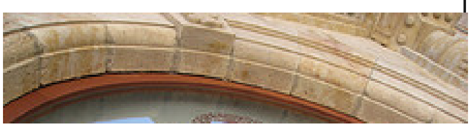
Importante

El tamaño mínimo de las muestras para que en el intervalo de confianza se cometa un error menor que una cantidad prefijada al nivel de significación α es:

$$n \geq \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma_1}{E}\right]^2 + \left[\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma_2}{E}\right]^2$$

Ejercicio resuelto

En la Universidad de Nagora están preocupados con las notas de selectividad del último año, por lo que encargan un estudio al Departamento de Estadística para que calcule la diferencia de la



media de las notas con la prestigiosa universidad de Agalam.

El investigador necesita saber de que tamaño escoger la muestra, aunque descubre que las notas de las dos universidades se distribuyen según una normal. La de Nagora tiene una desviación típica de 2,25, mientras que en la de Agalam es del 1,69.

Si quiere cometer un error menor de 0,25 puntos ayúdale a calcular el tamaño de las muestras para los niveles de confianza:

a) 90%

b) 95%

c) 99%

Imagen de Gabi Rondón con licencia Creative Commons

Mostrar retroalimentación

Con los datos que nos dan tenemos que aplicar la fórmula anterior:

$$n=\left[\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma_1}{E}\right]^2+\left[\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma_2}{E}\right]^2$$

Sólo nos falta calcular los valores críticos para los niveles de confianza pedidos.

Recordemos la tabla que construimos en el apartado 2, concretamente en el de la obsolescencia programada y tendremos:

a) Para el nivel de confianza del 90%, α=0,10; por lo que $z_{1-\alpha/2}=z_{0,95}=1,645$.

Por lo que sustituyendo los datos del problema en la fórmula tenemos que:

$$n=\left[\frac{1,645\cdot 2,25}{0,25}\right]^2+\left[\frac{1,645\cdot 1,69}{0,25}\right]^2=14,805^2+11,1202^2=342,846$$

Por lo que el tamaño muestral n de cada una de las muestras debe ser superior a 343 individuos.

b) Para el nivel de confianza del 95%, α=0,05; por lo que $z_{1-\alpha/2}=z_{0,975}=1,960$.

Por lo que sustituyendo los datos del problema en la fórmula tenemos que:

$$n=\left[\frac{1,960\cdot 2,25}{0,25}\right]^2+\left[\frac{1,960\cdot 1,69}{0,25}\right]^2=17,64^2+13,2496^2=486,72$$

En este caso, el tamaño muestral n de cada una de las muestras debe ser superior a 487 individuos.

c) Para el nivel de confianza del 99%, α=0,01; por lo que $z_{1-\alpha/2}=z_{0,995}=2,575$.

Por lo que sustituyendo los datos del problema en la fórmula tenemos que:

$$n=\left[\frac{2,575\cdot 2,25}{0,25}\right]^2+\left[\frac{2,575\cdot 1,69}{0,25}\right]^2=23,175^2+17,407^2=840,084$$

En este último caso al tener un nivel de confianza del 99%, el tamaño muestral n de cada una de las muestras aumenta considerablemente y debe ser superior a 841 individuos.

Reflexiona

Durante la realización de un estudio encargado a la empresa Tisbet Survey han tenido un problema con la tinta de la impresora y en el informe sólo han aparecido los siguientes datos:

μ_1 =47

σ_1 =3.5

σ_2 =4.1

n_1=35

n_2=40

El intervalo de confianza para la diferencia de las medias es **(3,1 ; 6,9)**

El resto se ha borrado, y tienen que presentar el informe en un plazo muy corto, por lo que no les da tiempo a realizar de nuevo el estudio.

Ayúdales a averiguar los datos que le faltan.

a) Calcula la media muestral de la segunda muestra.

b) Halla el nivel de confianza para el que se realizó el estudio.

Mostrar retroalimentación

a) Para el cálculo de la media muestral que nos falta nos vamos a apoyar en dos cosas:

- La diferencia de las medias muestrales estimada es el centro del intervalo de confianza. Por lo tanto:

$$\mu_1-\mu_2=\frac{6{,}9+3{,}1}{2}=\frac{10}{2}=5$$

- Como conocemos μ_1

$$47-\mu_2=5\Rightarrow\mu_2=42$$

Por lo tanto, la media muestral de la segunda muestra es igual a 42

b) Ahora nos vamos a apoyar en el error máximo que hemos cometido para averiguar el valor crítico y a partir de ahí cual es el nivel de confianza.

El Error E es igual al radio del intervalo, por lo que:

$$E=\frac{6{,}9-3{,}1}{2}=\frac{3{,}8}{2}=1{,}9$$

Por otro lado tenemos que:

$$E=z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Sustituimos todos los datos y despejamos $z_{1-\alpha/2}$

$$1{,}9=z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{3{,}5^2}{35}+\frac{4{,}1^2}{40}}\;\Rightarrow\;1{,}9=z_{1-\frac{\alpha}{2}}\cdot 0{,}88\;\Rightarrow\;z_{1-\frac{\alpha}{2}}=\frac{1{,}9}{0{,}88}=2{,}17$$

Buscamos la probabilidad de este valor en la tabla de la distribución normal y ese, será el valor de $1-\alpha/2$

Para $z=2{,}17$ ⇒ $1-\alpha/2=0{,}985$ y despejando tenemos que $\alpha=0{,}03$, por lo que el valor del nivel de significación es del 3%.

El dato pedido, nivel de confianza, es del 97%.

5. Ejemplo Selectividad

Imagen de ph-stop con licencia Creative Commons

Como en el tema anterior, terminamos con un problema de selectividad relativo a los contenidos tratados en el tema. En concreto, este problema es del examen de Selectividad de junio de 2010 en Andalucía.

En él vamos a calcular un intervalo de confianza para la media y vamos a determinar el tamaño muestral mínimo para que el error sea menor que una cierta cantidad.

Utilizaremos dos niveles de confianza distintos: el 94% para el intervalo y el 90% para el error.

El problema nos plantea la estimación de la media en el tiempo de reacción a un determinado estímulo medido en segundos.

Ejercicio resuelto

JUNIO 2010. ANDALUCÍA

Se sabe que el tiempo de reacción a un determinado estímulo se distribuye según una ley Normal de media desconocida y desviación típica 0,2 segundos.

a) Observada una muestra aleatoria de tamaño 25 se ha obtenido una media muestral de 0,3 segundos. Obtenga un intervalo de confianza para la media de la población con un nivel de confianza del 94%.

b) A un nivel de confianza del 90%, ¿cuál será el tamaño muestral mínimo si el error cometido es inferior a 0,05?

Mostrar retroalimentación

Los datos del problema son que el tiempo de reacción sigue una N(μ ; 0,2)

a) Recordemos que el intervalo de confianza para la media con desviación típica conocida tiene la siguiente fórmula:

$$IC=\left(\bar{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

Como conocemos todos los datos, menos el valor crítico, comencemos calculando este error:

Nos piden un nivel de confianza del 94% ⇒ 1-α es 0,94 ⇒ α es 0,06 ⇒ α/2 es 0,03 ⇒ 1-α/2 es 0,97 ⇒ $z_{1-\alpha/2}=z_{0,97}=1,88$

Por lo que:

$$IC=\left(\overline{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right) = \left(0,3\pm 1,88\cdot\frac{0,2}{5}\right) = \left(0,3\pm 1,88\cdot 0,04\right) = \left(0,3\pm 0,075\right) = \left(0,225,\ 0,375\right)$$

El tiempo de reacción medio se encuentra en el intervalo (0,225 ; 0,375) con un nivel de confianza del 94%

b) Nos piden el tamaño de la muestra para que el error sea menor que 0,05 segundos, pero ahora para un nivel de confianza del 90%.

Como el error es igual a:

$$E=z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\le 0,05$$

Lo primero que tenemos que calcular es el valor crítico:

El nivel de confianza deseado es del 90% ⇒ 1-α es 0,90 ⇒ α es 0,1 ⇒ α/2 es 0,05 ⇒ 1-α/2 es 0,95 ⇒ $z_{1-\alpha/2}=z_{0,95}=1,645$. Sustituyendo tenemos que:

$$1,645\cdot\frac{0,2}{\sqrt{n}}\le 0,05$$

Despejando se obtiene que:

$$n\ge\left[\frac{1,645\cdot 0,2}{0,05}\right]^2 \Rightarrow\ n\ge 43,2964$$

El tamaño muestral solicitado tiene que ser igual o superior a 44 individuos.

Resumen

Importante

Una **estimación puntual** del valor de un parámetro poblacional desconocido, como puede ser la media μ o la desviación típica σ, es un número que se utiliza para aproximar el verdadero valor de dicho parámetro poblacional.

Para calcular ese número, tomaremos una muestra de la población y calcularemos el estimador muestral asociado al parámetro poblacional. Por ejemplo el estimador de la media poblacional μ, es la media muestral $\overline{X}$.

El valor de este estimador muestral para una muestra concreta será la estimación puntual del parámetro poblacional.

Importante

El **intervalo de confianza para el parámetro media μ de una población N(μ, σ)** al nivel de significación α viene dado por:

$$IC=\left(\overline{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

Si el nivel de significación o de riesgo es α, el nivel de confianza será de 1-α.

Al valor de la abscisa que deja a su izquierda un área igual a $1-\frac{\alpha}{2}$ siendo 1 - α el nivel de confianza se le llama valor crítico.

Si σ es desconocida y n es grande (n ≥30), el intervalo de confianza viene dado por:

$$IC=\left(\overline{x}\pm z_{1-\frac{\alpha}{2}}\cdot\frac{\widehat{s}}{\sqrt{n}}\right)$$

donde $\widehat{s}^2$ es la cuasivarianza muestral.

Importante

El **Error máximo admisible** en el cálculo de un intervalo de confianza para la media es:

$$E=z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$

El margen de error admisible es la amplitud del intervalo, 2E.

Importante

El **tamaño** de la muestra para que en el intervalo de confianza se cometa un **error prefijado** al nivel de significación α es:

$$n=\left[\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma}{E}\right]^2$$

Importante

Sean dos poblaciones $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$; de cada una de ellas se elige una muestra de tamaños n_1 y n_2, respectivamente.

Sean $\overline{x}_1$ y $\overline{x}_2$ las medias de la primera y de la segunda muestra respectivamente. Para un nivel de significación α, se tiene que

- Si σ_1 y σ_2 son conocida, el intervalo de confianza viene dado por:

$$IC=\left[\overline{x}_1-\overline{x}_2\ \pm\ z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

- Si σ_1 y σ_2 son desconocidas y n_1 y n_2 son grandes ($n_1 \geq 30$ y $n_2 \geq 30$), el intervalo de confianza viene dado por:

$$IC=\left[\overline{x}_1-\overline{x}_2\ \pm\ z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\widehat{s}_1^2}{n_1}+\frac{\widehat{s}_2^2}{n_2}}\right]$$

donde $\widehat{s}_1^2$ y $\widehat{s}_2^2$ son las cuasivarianzas de cada muestra.

Importante

El **Error máximo admisible** en el cálculo de un intervalo de confianza para la diferencia de las medias es:

$$E=z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

El margen de error admisible es la amplitud del intervalo, 2E.