10 Inferencia estadística. Intervalo de confianza y contraste de hipótesis

Se proyecta crear un centro comercial en una ciudad, como el de la foto, y se quiere saber el poder adquisitivo de los habitantes de la ciudad, dato que influirá en el tamaño del centro comercial. Se establece una variable aleatoria X que indica el salario mensual de las personas mayores de edad en la ciudad. Se supone que X sigue una distribución normal, pero se ignora cuál es su media, μ, y su desviación típica, σ, para disponer de un modelo completamente determinado de la distribución de X con el que poder hacer cálculos de probabilidad del poder adquisitivo, predicciones de consumo, etc.

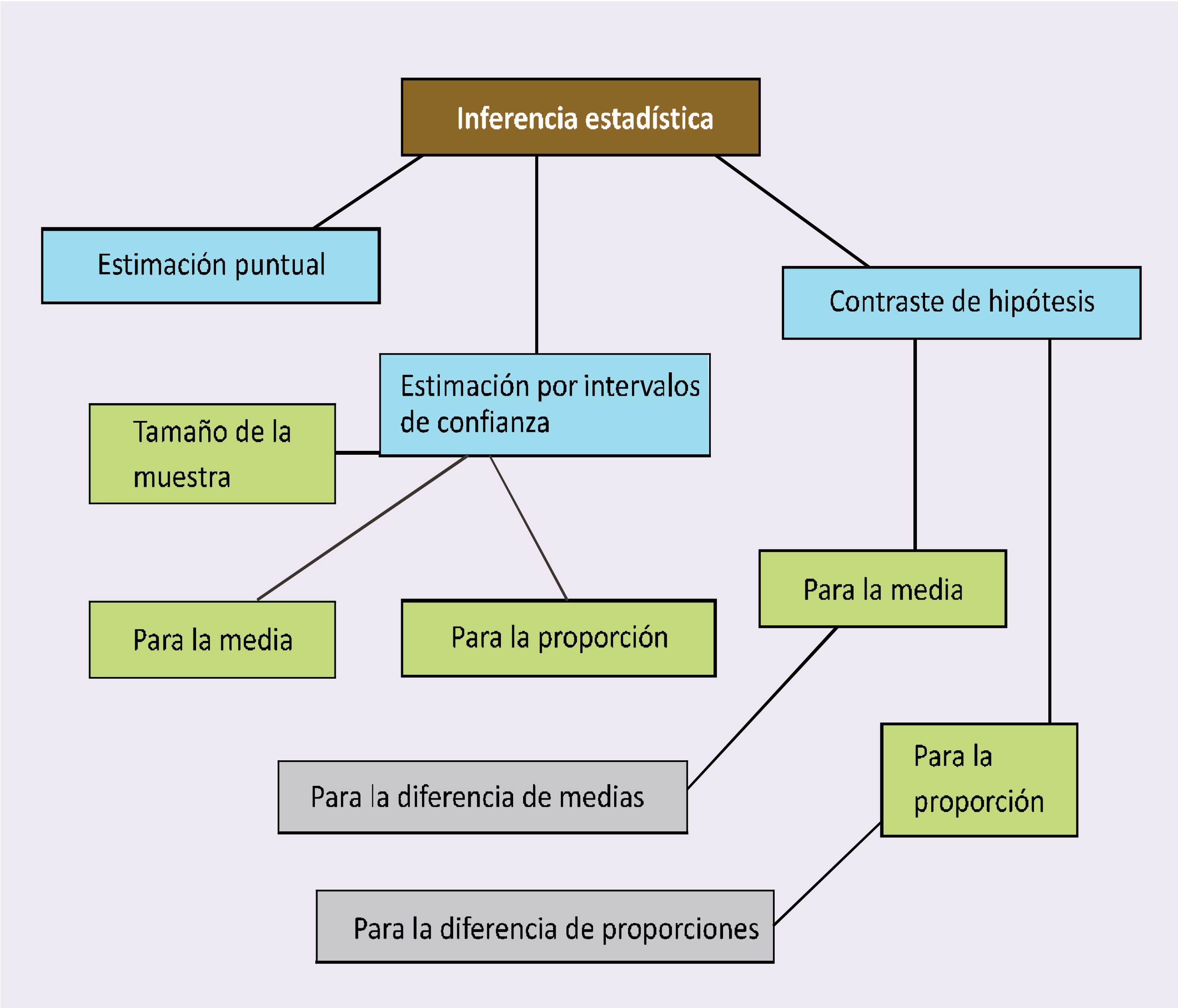
El modo de conocer, o mejor dicho, de ajustar los parámetros, μ y σ, de la distribución de la variable aleatoria X es extrayendo una muestra. Estimar un parámetro consiste en obtener una aproximación de su valor basándose en los datos de la muestra.

• *Vista general de la zona cubierta del barrio Chinatown de Singapur (ISFTIC. Banco de imágenes)*

Hay tres cosas que se pueden hacer para ajustar el valor del parámetro de una población: a) estimar el parámetro desconocido por un número que se obtiene en función de los datos de la muestra (esto se llama estimación por punto o puntual); b) determinar un intervalo al cual es muy probable que pertenezca el parámetro (esto se llama estimación por intervalo de confianza); c) aventurar, a modo de hipótesis, un número como valor del parámetro buscado y contrastar con datos de la muestra si éstos apoyan o rechazan la hipótesis formulada.

Esta Unidad didáctica tiene como **objetivos** los siguientes:

1. Construir intervalos de confianza para la media y la proporción.
2. Calcular el tamaño de la muestra para que el margen de error sea menor que una cantidad prefijada.
3. Contrastar las hipótesis sobre los valores de la media.
4. Contrastar hipótesis sobre los valores de la proporción.



ÍNDICE DE CONTENIDOS

1. Intervalo de confianza para la media

Es muy improbable que la media poblacional, μ, coincida con el valor medio de una muestra $\overline{X}$ (estimación puntual); sin embargo, es más interesante determinar un intervalo centrado en la media muestral, de modo que el valor desconocido de μ se encuentre en dicho intervalo con una probabilidad tan grande como deseemos. A esta probabilidad la llamamos **nivel de confianza**, y la simbolizamos por $1-\alpha$, mientras que a α la denominamos **nivel de significación** o nivel de riesgo.

Concretando, tenemos que buscar un número c, de modo que μ pertenezca al intervalo $(\overline{X}-c, \overline{X}+c)$ con probabilidad $1-\alpha$; es decir, que cumpla que:

$$P\left[\overline{X}-c<\mu<\overline{X}+c\right]=1-\alpha.$$

Despejando c en las desigualdades del corchete anterior,

$$\overline{X}-c<\mu \Rightarrow \overline{X}-\mu<c,$$
$$\mu<\overline{X}+c \Rightarrow -c<\overline{X}-\mu,$$

podemos escribir la probabilidad así:

$$P\left[-c<\overline{X}-\mu<c\right]=1-\alpha\ ,$$

y empleando el valor absoluto queda:

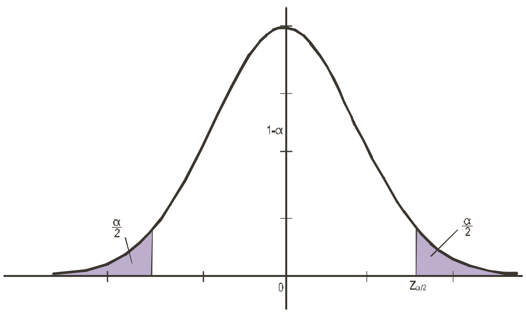
$$P\left[\left|\overline{X}-\mu\right|<c\right]=1-\alpha.$$

Suponiendo que conocemos σ, dado $\overline{X}$ que se distribuye según una $N\left(\mu,\dfrac{\sigma}{\sqrt{n}}\right)$, al dividir la desigualdad por $\sigma/\sqrt{n}$ tipificamos la variable $\overline{X}$,

$$P\left[\frac{\left|\overline{X}-\mu\right|}{\sigma/\sqrt{n}}<\frac{c}{\sigma/\sqrt{n}}\right]=P\left[|Z|<\frac{c}{\sigma/\sqrt{n}}\right]=1-\alpha$$

donde Z es la $N(0,1)$. Buscamos en las tablas el valor de $\dfrac{c}{\sigma/\sqrt{n}}$ correspondiente a la probabilidad $1-\alpha$ y llamamos a ese valor $z_{\alpha/2}=\dfrac{c}{\sigma/\sqrt{n}}$, despejando c obtenemos

$$c=\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}$$



¿Qué es $z_{\alpha/2}$?

Es el valor de una abscisa de la $N(0, 1)$ que deja a su derecha un área de probabilidad $\alpha/2$.

En consecuencia, el intervalo buscado es

$$\left(\overline{X}-\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \overline{X}+\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)$$

que obviamente depende de la probabilidad elegida. Se trata, por tanto, de un intervalo centrado en la media de la muestra tomada, $\overline{X}$, y de radio $\frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$. Este radio del intervalo de confianza se denomina **error máximo**, y es la máxima diferencia que puede existir entre μ y la media de la muestra elegida, $\overline{X}$, para un nivel de confianza $1 - \alpha$. Se simboliza por

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

y de la simple observación de esta fórmula vemos que al aumentar el tamaño de la muestra, n, disminuye el error E; mientras que al aumentar el nivel de confianza, aumenta $z_{\alpha/2}$ y por tanto el error cometido.

Ejemplo

1. En un laboratorio se obtuvieron 6 estimaciones de PH de una solución con los resultados siguientes:

7,91 7,94 7,90 7,93 7,89 7,91

Se supone que la población de todas las determinaciones del PH de la solución tiene una distribución normal de media desconocida y desviación típica 0,02.

a) Determínese un intervalo de confianza al 98% para la media de todas las determinaciones del PH de la misma solución obtenidas con el mismo método.

b) Con el mismo nivel de confianza anterior, ¿cuál debe ser el tamaño mínimo de la muestra para que la amplitud del intervalo de confianza sea a lo sumo 0,01?

Solución:

La variable aleatoria X es el PH de la solución que se distribuye normalmente según la $N(\mu;\ 0{,}02)$.

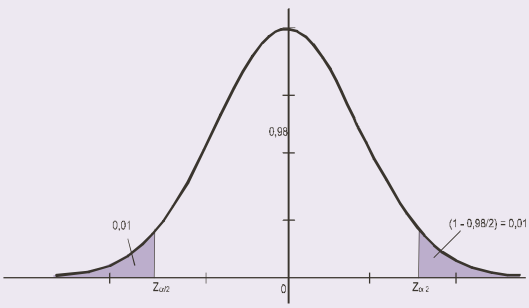
a) Determinamos el intervalo de confianza al 98% para la media. Sabemos que

nivel de confianza = $1 - \alpha$ = 98% = 0,98.

Tenemos que hallar c tal que $P[|\overline{X} - \mu| < c] = 0{,}98$. Además conocemos que se distribuye según una normal $N(\mu;\ 0{,}02/\sqrt{6})$. Tipificamos $\overline{X}$

$$P\left[\frac{|\overline{X} - \mu|}{0{,}02/\sqrt{6}} < \frac{c}{0{,}02/\sqrt{6}}\right] = P\left[|Z| < \frac{c}{0{,}02/\sqrt{6}}\right] = 0{,}98$$

Llamamos $z_{\alpha/2} = \frac{c}{\sigma/\sqrt{n}}$, y buscamos en las tablas de la $N(0,1)$ el valor de $z_{\alpha/2}$, que corresponde a la abscisa de la $N(0,1)$ que vemos en la figura



y que calculamos de la misma figura sabiendo que

$$P\left[Z < z_{\alpha/2}\right] = 0{,}01 + 0{,}98 = 0{,}99.$$

En las tablas de la $N(0, 1)$ leemos $z_{\alpha/2} = 2{,}33$, entonces $2{,}33 = \dfrac{c}{0{,}02/\sqrt{6}}$,

despejando c obtenemos $c = \dfrac{2{,}33 \cdot 0{,}02}{\sqrt{6}} = 0{,}019$

Como la media de la muestra es

$$\overline{X} = \frac{7{,}91+7{,}94+7{,}90+7{,}93+7{,}89+7{,}91}{6} = 7{,}91,$$

podemos escribir el intervalo buscado

$$\left(\overline{X} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \overline{X} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right) = (7{,}91-0{,}019;\ 7{,}91+0{,}019) = (7{,}891;\ 7{,}929).$$

Esto significa que el 98% de las muestras de tamaño 6 tienen una media muestral que difiere o dista de la media poblacional menos que 0,019.

b) Este apartado lo resolveremos en el primer ejemplo del epígrafe siguiente.

Actividades

1. La desviación típica de una variable aleatoria es $\sigma = 6$. Para estimar la media de dicha variable se extrae una muestra de tamaño 100 y se obtiene una media muestral $\overline{X} = 100$. Construir un intervalo de confianza del 95% para estimar la media μ de la población.

2. La duración de las llamadas de teléfono, en una oficina comercial, sigue una distribución normal con desviación típica 10 segundos. Se toma una muestra de 50 llamadas y la duración media de las llamadas de esa muestra es 35 segundos. Calcular el intervalo de confianza al 99% para la duración media de las llamadas.

3. Se supone que los gastos corrientes por empleado de los distintos departamentos de una empresa siguen una distribución normal con desviación típica 450 euros. De los datos disponibles para 16 departamentos se ha obtenido un gasto medio de 1650 euros. Determínese un intervalo de confianza al 99% para el gasto corriente medio por empleado en la empresa.

4. Se probaron 10 automóviles, escogidos aleatoriamente de una misma marca y modelo, por conductores con la misma forma de conducir y en carreteras similares. Se obtuvo que el consumo medio de gasolina, en litros, por cada 100 kilómetros fue de 6,5. Estudios previos indican que el consumo de gasolina tiene una distribución normal de desviación típica 2 litros. Determinar un intervalo de confianza al 95% para la media del consumo de gasolina de estos automóviles.

1.1. Tamaño de la muestra

Otro problema relacionado con el intervalo de confianza es: ¿cuál debe ser el tamaño de la muestra para que con un nivel de confianza determinado 90%, 95%, 99% u otro, la diferencia entre la media muestral y la media poblacional sea menor que un valor fijo? O, lo que es lo mismo, ¿cuál debe ser el tamaño de la muestra para que con un nivel de confianza establecido el error máximo cometido sea menor que una cantidad dada?

En los problemas del intervalo de confianza nos dan el nivel de confianza, el tamaño de la muestra y nos piden hallar la diferencia (el error) entre la media poblacional y la media muestral. En los problemas de tamaño de la muestra nos dan el nivel de confianza, la diferencia máxima entre $\overline{X}$ y μ, es decir, el error máximo admitido, y tenemos que hallar el tamaño de la muestra. En la práctica, únicamente tenemos que despejar n en la fórmula del error máximo

$$E = \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{z_{\alpha/2}\cdot\sigma}{E} \Rightarrow n = \left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2$$

Ejemplos

2. Ahora resolvemos el apartado b) del problema del ejemplo 1, que había quedado pendiente.

b) Con el mismo nivel de confianza, ¿cuál debe ser el tamaño mínimo de la muestra para que la amplitud del intervalo de confianza o radio o error máximo sea a lo sumo 0,01?

Solución:

En la fórmula del error máximo despejamos n:

$$E = \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{z_{\alpha/2}\cdot\sigma}{E} \Rightarrow n = \left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2.$$

Recordamos que σ = 0,02 y que 1 – α = 98% = 0,98 por lo que $z_{\alpha/2}$ = 2,33, como hemos visto en el apartado a), y además queremos hallar el tamaño de la muestra, n, para que a lo sumo el error sea E = 0,01, es decir,

$$\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}} < 0{,}01 \Rightarrow \sqrt{n} > \frac{z_{\alpha/2}\cdot\sigma}{0{,}01}.$$

Sustituyendo en la fórmula, $n = \left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2$, resulta $n = \left(\frac{2{,}33\cdot 0{,}02}{0{,}01}\right)^2 = 21{,}715$.

Como ha de ser n > 21, entonces el tamaño mínimo de la muestra es n = 22.

3. El tiempo de vida de una clase de depuradoras de agua utilizadas en una planta industrial se distribuye normalmente, con una desviación típica de 2000 horas. En un ensayo realizado con una muestra aleatoria de 9 depuradoras, se obtuvieron los siguientes tiempos de vida en miles de horas

9,5 10 7,5 10,5 16,5 10 12 32 18

a) Hállese un intervalo de confianza al 99% para la vida media de las depuradoras.

b) Calcúlese el tamaño mínimo que debería tener la muestra, en el caso de admitir un error máximo de 500 horas, con un grado de confianza del 95%.

Solución:

La variable X es el tiempo de vida de las depuradoras que se distribuye según una $N(\mu, 2000)$. Vamos a resolver este problema de una manera más directa que en el ejemplo 1.

a) El **intervalo de confianza para la media** viene dado por la expresión:

$$\left(\overline{X} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \overline{X} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)$$

Vamos a calcular cada uno de ingredientes de la expresión anterior. Como $1-\alpha$ = 99% = 0,99, entonces de $P[Z < z_{\alpha/2}]$ = 0,005 + 0,99 = 0,995 y obtenemos en las tablas que $z_{\alpha/2}$ = 2,575; además, σ = 2000 horas, el tamaño de la muestra n = 9 y

$$\overline{X} = \frac{9{,}5+10+7{,}5+10{,}5+16{,}5+10+12+32+18}{9} = 14, \text{ en miles de horas, } \overline{X} = 14000 \text{ horas.}$$

Sustituyendo en la expresión del intervalo de confianza queda

$$\left(\overline{X}-\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \overline{X}+\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)=\left(14000-\frac{2{,}575\cdot 2000}{\sqrt{9}}, 14000+\frac{2{,}575\cdot 2000}{\sqrt{9}}\right)= (14000-1716{,}67,\ 14000+1716{,}67)=$$
$$=(12283{,}33,\quad 15716{,}67).$$

Esto significa que el 99% de las muestras de tamaño 9 tienen una media muestral que difiere o dista de la media poblacional menos que 1716,66 horas.

b) En la fórmula del error máximo despejamos n:

$$E=\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}} \Rightarrow \sqrt{n}=\frac{z_{\alpha/2}\cdot\sigma}{E} \Rightarrow n=\left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2.$$

Recordamos que σ = 2000 horas, ahora el grado de confianza o nivel de confianza es $1-\alpha$ = 95% = 0,95 por lo que de $P[Z < z_{\alpha/2}]$ = 0,025 + 0,95 = 0,975, obtenemos que $z_{\alpha/2}$ = 1,96. Queremos hallar el tamaño de la muestra, n, para que el error máximo admitido sea E = 500 horas. Sustituyendo en la última fórmula,

$$n=\left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2, \text{ resulta } n=\left(\frac{1{,}96\cdot 2000}{500}\right)^2=61{,}46.$$

El tamaño mínimo de la muestra es n = 62.

Actividades

5. Se estima que el tiempo de reacción de un conductor ante un obstáculo imprevisto tiene una distribución normal con desviación típica 0,05 segundos. Si se quiere conseguir que el error de estimación de la media no supere los 0,01 segundos con un nivel de confianza del 99%, ¿qué tamaño mínimo ha de tener la muestra de tiempos de reacción?

6. Se sabe que la estatura de los individuos de una población es una variable aleatoria que sigue una distribución normal con desviación típica 6 cm. Se toma una muestra de 225 individuos que dan una media de 176 cm.

a) Obténgase un intervalo con 99% de confianza para la estatura media de la población.

b) Calcule el tamaño mínimo de la muestra para estimar la estatura media de la población con un error inferior a 1 cm y un nivel de confianza del 95%.

7. El precio de ciertos electrodomésticos puede considerarse una variable aleatoria con distribución normal de desviación típica 100 euros. Los precios en euros correspondientes a una muestra de 9 de estos electrodomésticos son:

255 85 120 290 80 80 275 290 135

a) Construir un intervalo de confianza al 98% para la media poblacional.

b) Hallar el tamaño mínimo que debe tener la muestra, para que con un nivel de confianza del 99%, el error de estimación del precio medio no supere los 50 euros.

8. El peso de los perros adultos de una cierta raza es una variable aleatoria que se distribuye normalmente con desviación típica 0,6 kg. Una muestra aleatoria de 30 animales ha dado un peso medio de 7,4 kg.

a) Hállese un intervalo de confianza al 99% para el peso medio de los perros adultos de esta raza.

b) ¿Qué tamaño mínimo debe tener la muestra para tener una confianza del 95% de que la media muestral no se diferencie en más de 0,3 kg de la media de la población?

9. Se sabe que los estudiantes de cierta universidad duermen un número de horas diarias que se distribuye según una ley Normal de media μ y desviación típica $\sigma = 2$ horas.

a) A partir de una muestra de 64 alumnos se ha obtenido el intervalo de confianza (7,26; 8,14) para la media de horas de sueño. Determinar el nivel de confianza con que se ha construido dicho intervalo.

b) Determinar el tamaño muestral mínimo necesario para que el error que se cometa al estimar la media de la población por un intervalo de confianza sea, como máximo, de 0,75 horas con un nivel de confianza del 98%.

1.2. Intervalo de confianza para la media de una distribución normal con σ desconocida

En el desarrollo anterior hemos supuesto que la desviación típica σ es conocida, claro que alguien puede pensar, ¿cómo es posible conocer σ sin conocer μ?

Cuando se desconoce el valor de la desviación típica poblacional, σ, ésta se puede estimar por la **cuasi desviación típica muestral** cuya fórmula es $S = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$ y que a veces se llama simplemente desviación típica muestral, y que en algunas calculadoras se halla con la tecla $\boxed{\sigma_{n-1}}$. Cuando hacemos una estimación puntual de la σ desconocida por la cuasi desviación típica muestral, S, el tamaño de la muestra ha de ser grande ($n \geq 30$) porque si n es menor que 30 al tipificar la variable $\overline{X}$ obtenemos

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Esta variable tipificada no sigue una distribución normal $N(0,1)$ sino una distribución t de Student que no es objetivo de este curso.

Resumiendo: cuando n es grande y desconocemos σ, podemos estimar la desviación típica poblacional por la cuasi desviación típica muestral (o simplemente desviación típica muestral), S, como se muestra en el siguiente ejemplo.

Ejemplos

4. Se ha registrado la edad de una muestra de 200 personas que hacían la compra en un supermercado, y se ha encontrado que la media de edad es 45,3 años, con una desviación típica muestral de 16,8 años. Determinar un intervalo de confianza, con un nivel de significación del 5%, para la edad media de las personas que hacen sus compras en ese supermercado.

Solución:

Como σ es desconocida, y n = 200, la estimamos por la desviación típica muestral S = 16,8 años.

El intervalo de confianza para la media viene dado por la expresión:

$$\left(\overline{X} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \ \overline{X} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)$$

Donde α = 0,05 y 1 – α = 95%, de $P\left[Z < z_{\alpha/2}\right] = 0{,}025 + 0{,}95 = 0{,}975$ obtenemos en las tablas que $z_{\alpha/2}$ = 1,96.

Además, como $\overline{X}$ = 45,3, sustituyendo en la expresión del intervalo de confianza resulta:

$$\left(\overline{X} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \ \overline{X} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right) = \left(45{,}3 - \frac{1{,}96\cdot 16{,}8}{\sqrt{200}},\ \ 45{,}3 + \frac{1{,}96\cdot 16{,}8}{\sqrt{200}}\right) = (42{,}971;\ \ 47{,}628).$$

Actividades

10. El salario medio mensual de una muestra de 100 personas de un determinado barrio ha resultado ser 1080 euros, con una desviación típica muestral de 160 euros. Determinar un intervalo de confianza para el salario de los habitantes del barrio con un nivel de confianza del 99%.

11. Se ha registrado el peso de 200 personas mayores de edad, residentes en una determinada villa, y el peso medio ha resultado ser 85 kg, con una desviación típica muestral de 9 kg. Determinar un intervalo de confianza para el peso de los habitantes de esa villa con un nivel de confianza del 95%.

2. Intervalo de confianza para la proporción

Sabemos que si una población tiene una proporción poblacional p de una determinada característica, entonces la variable aleatoria $\hat{p}$, de las proporciones muestrales, cuando n es grande, tiende a una distribución normal de media $\mu_{\hat{p}} = p$ y desviación típica $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$; es decir, la distribución de $\hat{p}$ se aproxima mucho a una $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Se considera n grande cuando $n \geq 30$.

Conocido el valor de la proporción $\hat{p}$ de una muestra, podemos determinar un intervalo para p, la proporción poblacional de la característica, del mismo modo que hicimos en el caso del intervalo de confianza para la media. Buscamos un intervalo en el que el valor desconocido de p se encuentre con una probabilidad o nivel de confianza $1-\alpha$, donde α indica el nivel de significación.

Sin embargo, en vez de seguir el mismo camino que en la deducción del intervalo de confianza para la media vamos a hacer otra cosa. Se puede demostrar que si n, tamaño de la muestra, es grande, el cociente

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

sigue aproximadamente una distribución normal $N(0,1)$; es decir, que

$$P\left[|Z| < z_{\alpha/2}\right] = P\left[\frac{|\hat{p}-p|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right] = 1-\alpha$$

De la desigualdad del segundo corchete

$$\frac{|\hat{p}-p|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}$$

obtenemos $|\hat{p}-p| < z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ o $-z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \hat{p}-p < z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Multiplicando por –1 las desigualdades anteriores cambian de sentido

$$z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > -\hat{p}+p > -z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

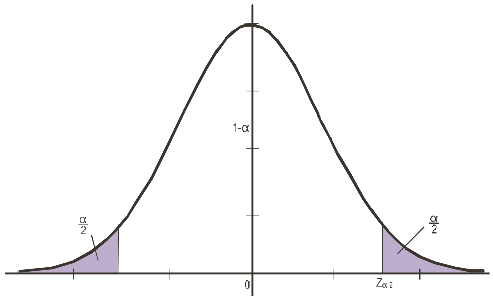
y sumando $\hat{p}$ a todas las desigualdades se llega a $\hat{p}+z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p > \hat{p}-z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$,

que, cambiando el orden, también se puede escribir $\hat{p}-z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p}+z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Luego el **intervalo de confianza para la proporción** cuando n es grande es aproximadamente

$$\left(\hat{p}-z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Donde $z_{\alpha/2}$, como en el caso del intervalo de confianza para la media, es el valor de una abscisa de la $N(0,1)$ que deja a su derecha un área de probabilidad $\alpha/2$.



Se trata de un intervalo centrado en la proporción de la muestra elegida, $\hat{p}$, y de radio $z_{\alpha/2}\cdot\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$. A este radio, como en el caso del intervalo para la media, se denomina error máximo; y es el error que se comete al estimar p mediante $\hat{p}$, y también se simboliza por

$$E = z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Como en el apartado anterior, ahora estamos en condiciones de hallar el tamaño de la muestra para la proporción. Conocemos el nivel de confianza, la diferencia máxima entre $\hat{p}$ y p, es decir, el error máximo admitido, y queremos determinar el tamaño de la muestra. Esto se consigue despejando n en la fórmula:

$$E = z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow E^2 = (z_{\alpha/2})^2\,\frac{\hat{p}(1-\hat{p})}{n} \Rightarrow n = (z_{\alpha/2})^2\,\frac{\hat{p}(1-\hat{p})}{E^2}$$

Ejemplo

5. En una comunidad autónoma se hace una encuesta para conocer si los contribuyentes están a favor de la implantación de un nuevo impuesto. Se interroga a una muestra de 600 contribuyentes y el resultado es favorable sólo en 225 casos.

a) Con un nivel de significación del 5% establecer un intervalo de confianza para la proporción de contribuyentes favorables al impuesto.

b) ¿Cuál ha de ser el tamaño de la muestra para que la diferencia entre la proporción poblacional p y la proporción muestral $\hat{p}$ sea menor que 0,02, con el mismo nivel de confianza?

Solución:

a) intervalo de confianza para la proporción poblacional p es

$$\left(\hat{p} - z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \ \hat{p} + z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Calculamos los ingredientes del intervalo de confianza: α = 5% = 0,05, 1 – α = 0,95; además n = 600 ≥ 30, $\hat{p}$ = 225/600 = 0,375 y $z_{\alpha/2}$ lo calculamos de las tablas, sabiendo que $P[Z < z_{\alpha/2}] = 0{,}95 + 0{,}025 = 0{,}975$, y resulta $z_{\alpha/2}$ = 1,96. El intervalo buscado será:

$$\left(0{,}375 - 1{,}96\cdot\sqrt{\frac{0{,}375\cdot 0{,}625}{600}};\ \ 0{,}375 + 1{,}96\cdot\sqrt{\frac{0{,}375\cdot 0{,}625}{600}}\right) = (0{,}375 - 0{,}039;\ 0{,}375 + 0{,}039) = (0{,}336;\ 0{,}414).$$

Esto significa que el intervalo de confianza en el 95% de las muestras de tamaño 600 contiene a la proporción poblacional buscada; tan solo el 5% restante no contiene a dicha proporción. A la vista del intervalo obtenido no parece que los ciudadanos estén muy entusiasmados con el impuesto.

b) Debemos calcular n en la fórmula $E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$,

donde E = 0,02, $z_{\alpha/2}$ = 1,96, lo hemos calculado antes, porque el nivel de confianza 1 – α = 0,95, es el mismo y $\hat{p}$ = 0,375. Sustituyendo y despejando n, obtenemos

$$0{,}02 = 1{,}96 \cdot \sqrt{\frac{0{,}375 \cdot 0{,}625}{n}} \Rightarrow 0{,}02^2 = 1{,}96^2 \cdot \frac{0{,}375 \cdot 0{,}625}{n} \Rightarrow n = 1{,}96^2 \cdot \frac{0{,}375 \cdot 0{,}625}{0{,}02^2} = 2250{,}94.$$

Como ha de ser n > 2250,94 el tamaño mínimo de la muestra es n = 2251.

Actividades

12. En una comarca ganadera se quiere estimar la proporción de ovejas que sufren una enfermedad endémica. Calcular el tamaño de la muestra necesario para determinar esta proporción con error menor que 0,04 para un nivel de confianza del 95%, sabiendo que en una muestra de 30 ovejas de la comarca resultaron dos enfermas.

13. Tomada al azar una muestra de 60 estudiantes de una Universidad se encontró que un tercio hablaba inglés.

a) Hallar, con un nivel de confianza del 90%, un intervalo para estimar la proporción de estudiantes que hablan el idioma inglés entre los estudiantes de esa Universidad.

b) A la vista del resultado anterior se pretende repetir la experiencia para conseguir una cota de error del 0,01 con el mismo nivel de confianza del 90%. ¿Cuántos individuos ha de tener la muestra?

14. En un Instituto se toma al azar una muestra de 100 estudiantes y se encuentra que sólo 22 han aprobado todas las asignaturas en la última evaluación. Se pide hallar:

a) Con un nivel de confianza del 95%, un intervalo para estimar el porcentaje de estudiantes que aprueban todas las asignaturas.

b) A la vista del resultado anterior se decide repetir la experiencia para conseguir una cota de error del 0,03, con el mismo nivel de confianza del 95%. ¿Cuántos individuos ha de tener la muestra?

15. A partir de la información proporcionada por una muestra aleatoria de 500 familias de cierta ciudad se ha determinado el intervalo de confianza (0,58; 0,64) con un nivel de confianza del 95% para la proporción de familias de la ciudad que disponen de ordenador en casa. Se pide determinar:

a) La estimación puntual que daríamos, a partir de la información recogida, para la proporción de familias en la ciudad que disponen de ordenador en casa.

b) El número mínimo de familias que tendríamos que seleccionar para conseguir, con un nivel de confianza del 95%, que el error máximo en la estimación de dicha proporción sea inferior a 0,01.

3. Contraste de hipótesis para la media

Otra forma de abordar el problema de estimar la media de una población consiste en formular una hipótesis sobre la media y después usar la información que proporciona una muestra para confirmar o rechazar dicha hipótesis. Esta técnica de inferencia estadística se conoce con el nombre de contraste de hipótesis o test de hipótesis. Y, por supuesto, supondremos que estamos ante una variable aleatoria X que sigue una distribución normal de desviación típica, σ, conocida.

El proceso del contraste de hipótesis se realiza en los 4 pasos que indicamos a continuación.

1. Establecer la **hipótesis** que provisionalmente se considera verdadera. Simbolizamos esta hipótesis por H_0, y consiste en que μ, media poblacional, tenga un valor μ_0; es decir,

$$H_0 : \mu = \mu_0$$

Esta H_0 se denomina **hipótesis nula** porque se parte del supuesto de que la diferencia entre el valor verdadero de la media y su valor hipotético es cero.

La **hipótesis alternativa**, simbolizada por H_1, podrá ser de tres tipos diferentes en función de la naturaleza del problema:

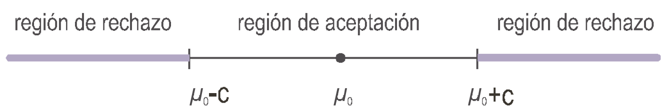
$H_1 : \mu \neq \mu_0$ (contraste bilateral)
$H_1 : \mu > \mu_0$ (contraste unilateral por la derecha)
$H_1 : \mu < \mu_0$ (contraste unilateral por la izquierda)

2. Fijar el **nivel de significación** α, que indica la probabilidad de rechazar H_0 aun siendo verdadera, o establecer el nivel de confianza $1-\alpha$, que indica la probabilidad de aceptar H_0 cuando es cierta. Los valores de α más comunes son 5% , 1% y 10% (o 0,05, 0,01 y 0,1) y los de $1-\alpha$, 95% , 99% y 90% (o 0,95, 0,99 y 0,90).

3. Determinar la **región de aceptación** para el nivel de significación α o nivel de confianza $1-\alpha$, de tal manera que la probabilidad de aceptar H_0 cuando sea cierta coincida con $1-\alpha$. Luego la región de aceptación será el intervalo $(\mu_0 - c,\ \mu_0 + c)$ en el que el número c verifica que:

$$P\left[\left|\bar{X}-\mu_0\right|<c\right]=1-\alpha$$

Gráficamente la región de aceptación y la región de rechazo serían las de la figura.



4. Se extrae una muestra y se calcula la media muestral, $\bar{X}$; a continuación se comprueba si cae dentro o fuera de la región de aceptación. Si cae dentro se acepta H_0 y si no, se rechaza.

3.1. Contraste bilateral

Aparece cuando la hipótesis alternativa es simplemente distinta de la hipótesis nula.

Paso 1. En este caso la hipótesis nula y alternativa son:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Paso 2. Fijamos un nivel de significación α o un nivel de confianza $1-\alpha$.

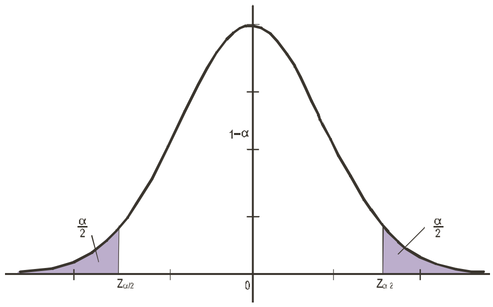
Paso 3. Determinación de la región de aceptación $(\mu_0 - c, \mu_0 + c)$. El valor de c en la región de aceptación, para un nivel de significación α o un nivel de confianza $1-\alpha$, se calcula de la probabilidad:

$$P\left[\left|\overline{X}-\mu_0\right|<c\right]=1-\alpha$$

Dividiendo la desigualdad del corchete por $\sigma/\sqrt{n}$ tipificamos $\overline{X}$,

$$P\left[\frac{\left|\overline{X}-\mu_0\right|}{\sigma/\sqrt{n}}<\frac{c}{\sigma/\sqrt{n}}\right]=1-\alpha,$$

$$P\left[|Z|<\frac{c}{\sigma/\sqrt{n}}\right]=1-\alpha.$$



Buscamos en las tablas un valor, $z_{\alpha/2}$, de la $N(0,1)$ tal que

$$P\left[|Z|<z_{\alpha/2}\right]=1-\alpha$$

Como $z_{\alpha/2}=\dfrac{c}{\sigma/\sqrt{n}}$, entonces $c=\dfrac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}$ y la **región de aceptación** será:

$$\left(\mu_0-\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \mu_0+\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)$$

Paso 4. Se extrae una muestra, se calcula su media $\overline{X}$ y se comprueba si cae dentro o fuera de la región anterior, si cae dentro aceptamos H_0 y si no, se rechaza.

Ejemplos

6. Los diámetros de unas piezas cilíndricas producidas por una máquina siguen una distribución normal de media desconocida y desviación típica σ = 2 mm. Se toma una muestra de 25 piezas y se obtiene un diámetro medio de 36 mm. ¿Se puede afirmar con un nivel de significación de 0,01 que la media del diámetro de estas piezas es 40 mm?

Solución:

Paso 1. La hipótesis nula es $H_0 : \mu = 40$ y la hipótesis alternativa, $H_1 : \mu \neq 40$. Estamos ante un contraste bilateral.

Paso 2. Fijamos un nivel de significación $\alpha = 0{,}01$ o un nivel de confianza $1-\alpha = 0{,}99$.

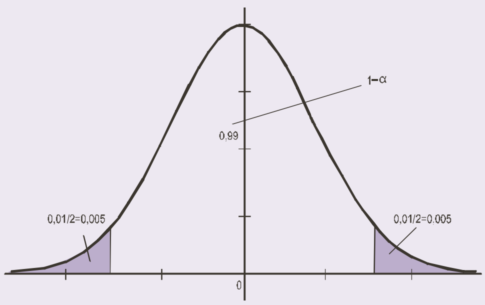
Paso 3. Determinamos la región de aceptación:

$$\left(\mu_0-\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \mu_0+\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\right)$$

Como $\alpha = 0{,}01$, $\alpha/2 = 0{,}005$ y el tamaño de la muestra $n = 25$, según el gráfico y las fórmulas

$$P[|Z|<z_{\alpha/2}] = 1-\alpha,$$

$$P[Z<z_{\alpha/2}] = 0{,}99 + 0{,}005 = 0{,}995;$$



obtenemos el valor de $z_{\alpha/2}$ en las tablas; resultando $z_{\alpha/2} = 2{,}575$; en consecuencia, la región de aceptación será:

$$\left(40 - \frac{2{,}575 \cdot 2}{\sqrt{25}};\ \ 40 + \frac{2{,}575 \cdot 2}{\sqrt{25}}\right) = (38{,}97;\ \ 41{,}03).$$

Paso 4. Para una muestra de tamaño $n = 25$ ha resultado que $\overline{X} = 36$ mm. Es evidente que 36 no pertenece al intervalo (38,97; 41,03), luego rechazamos H_0, la media de la población no es 40 mm.

7. Se quiere comprobar si una máquina destinada al llenado de botellas de agua mineral ha sufrido un desajuste. Una muestra aleatoria de 10 botellas salidas de esa máquina proporciona los siguientes datos:

0,49 0,52 0,51 0,48 0,53 0,55 0,49 0,50 0,52 0,49

Suponiendo que la cantidad de líquido que la máquina deposita en cada envase sigue una distribución normal de media 0,5 l y una desviación típica de 0,02 l, se desea contrastar si el contenido medio de las botellas salidas de esa máquina es 0,5 l con un nivel de significación de 5%. Se pide:

a) Plantear la hipótesis nula y alternativa.

b) Determinar la región crítica del contraste.

c) Realizar el contraste.

Solución:

a) Paso 1. La hipótesis nula y la alternativa son:

$$H_0 : \mu = 0{,}5 \text{ l}$$
$$H_1 : \mu \neq 0{,}5 \text{ l}$$

b) Paso 2. Fijamos el nivel de significación $\alpha = 5\% = 0{,}05$ y, por consiguiente, el nivel de confianza $1 - \alpha = 1 - 0{,}05 = 0{,}95$.

Paso 3. Determinación de la región de aceptación cuando $\sigma = 0{,}02$ l y el tamaño de la muestra $n = 10$,

$$\left(0{,}5 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}};\ 0{,}5 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right) = \left(0{,}5 - \frac{z_{\alpha/2} \cdot 0{,}02}{\sqrt{10}};\ 0{,}5 + \frac{z_{\alpha/2} \cdot 0{,}02}{\sqrt{10}}\right).$$

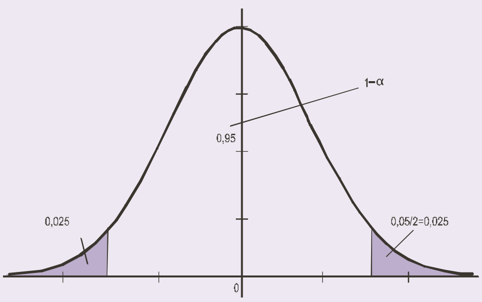
Como $\alpha = 5\% = 0{,}05$, $\alpha/2 = 0{,}025$, según el gráfico

vemos que $P[|Z| < z_{\alpha/2}] = 1 - 0{,}05 = 0{,}95$.

$P[Z < z_{\alpha/2}] = 0{,}95 + 0{,}025 = 0{,}975$

Buscando en las tablas el valor de $z_{\alpha/2}$ obtenemos que

$z_{\alpha/2} = 1{,}96$; en consecuencia, la región de aceptación será:



$$\left(0{,}5 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}};\ 0{,}5 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right) = \left(0{,}5 - \frac{1{,}96 \cdot 0{,}02}{\sqrt{10}};\ 0{,}5 + \frac{1{,}96 \cdot 0{,}02}{\sqrt{10}}\right) = (0{,}5 - 0{,}012;\ \ 0{,}5 + 0{,}012) = (0{,}488;\ 0{,}512).$$

La región crítica o región de rechazo está formada por el complementario de la región de aceptación, es decir, los intervalos $(-\infty,\ 0{,}488)$ y $(0{,}512,\ \infty)$.

d) Paso 4. Calculamos $\overline{X}$, $\overline{X} = \dfrac{0{,}49 + 0{,}52 + \ldots + 0{,}49}{10} = 0{,}508$. Como 0,508 está dentro del intervalo (0,488; 0,512), aceptamos la hipótesis nula.

Actividades

16. En una comunidad autónoma se estudia el número medio de hijos por mujer a partir de los datos disponibles en cada municipio. Se supone que este número sigue una distribución normal con desviación típica igual a 0,08. El valor medio de estos datos para 36 municipios resulta ser igual a 1,17 hijos por mujer. Se desea contrastar, con un nivel de significación de 0,01, si el número medio de hijos por mujer en la comunidad es de 1,25.

17. Un fabricante garantiza a un laboratorio farmacéutico que sus máquinas producen comprimidos con un diámetro medio de 25 mm. Una muestra de 100 comprimidos dio como media de los diámetros 25,18 mm. Suponiendo que el diámetro de los comprimidos es una variable aleatoria con distribución normal de desviación típica 0,89 mm, se desea contrastar con un nivel de significación del 5%, si el diámetro medio que afirma el fabricante es correcto. Para ello:

a) Plantéese la hipótesis nula y la hipótesis alternativa del contraste.

b) Realícese el contraste al nivel de significación indicado.

18. Un establecimiento vende paquetes de carbón para barbacoa de peso teórico 10 Kg. Se supone que el peso de los paquetes sigue una distribución normal con desviación típica 1 Kg. Para contrastar la citada hipótesis, frente a que el peso teórico sea distinto de 10 Kg, se escogen al azar 4 paquetes que pesan en Kg, respectivamente: 8, 10, 9, 8.

Se desea que la probabilidad de aceptar la hipótesis nula, si ésta es cierta, sea 0,95. Se pide:

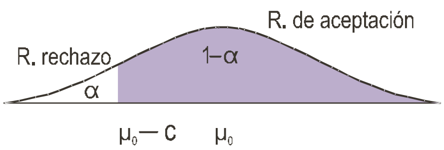
a) La región crítica del contraste.

b) ¿Se debe rechazar la hipótesis nula?

19. El peso medio de una muestra aleatoria de 81 personas de una determinada población es 63,5. Se sabe que la desviación típica poblacional es de 6 Kg. Con un nivel de significación del 0,05, ¿hay suficientes evidencias para rechazar la afirmación de que el peso medio poblacional es de 65 Kg.?

3.2. Contraste unilateral por la izquierda

Hay problemas en que nos interesa decidir si la hipótesis alternativa es menor que el valor de la hipótesis nula. Estamos entonces ante un contraste unilateral por la izquierda cuyas hipótesis son:



$$H_0 : \mu \geq \mu_0$$

$$H_1 : \mu < \mu_0$$

En este caso la región de aceptación es el intervalo $(\mu_0 - c, \infty)$ y la región crítica o región de rechazo $(-\infty, \mu_0 - c)$ está situada a la izquierda de la distribución de las medias muestrales

Para calcular c partimos de $P[\overline{X} > \mu_0 - c] = 1 - \alpha$. Restando μ_0 y dividiendo por $\sigma/\sqrt{n}$ tipificamos $\overline{X}$:

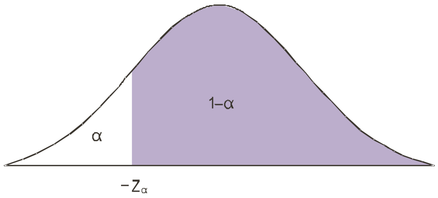
$$P\left[\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\mu_0 - c - \mu_0}{\sigma/\sqrt{n}}\right] = 1-\alpha,$$

$$P\left[Z > \frac{-c}{\sigma/\sqrt{n}}\right] = 1-\alpha \quad \text{ó} \quad P[Z > -z_\alpha] = 1-\alpha.$$

Buscamos en las tablas un valor, $-z_\alpha$, de la $N(0,1)$ tal que $P[Z > -z_\alpha] = P[Z < z_\alpha] = 1 - \alpha$.

Entonces, como $z_\alpha = \dfrac{c}{\sigma/\sqrt{n}} \Rightarrow c = \dfrac{z_\alpha \cdot \sigma}{\sqrt{n}}$,

la región de aceptación es: $\left(\mu_0 - \dfrac{z_\alpha \cdot \sigma}{\sqrt{n}},\ \infty\right)$.



Ejemplo

8. La vida media de las bombillas de 60 W de una determinada marca está garantizada por el fabricante en 800 horas, con una desviación típica de 120 horas. Se escoge una muestra de 50 bombillas después de dejarlas encendidas ininterrumpidamente se encuentra que tienen una vida media de 750 horas. ¿Habrá que rechazar la afirmación del fabricante con un nivel de confianza del 95%?

Solución: Seguiremos el mismo procedimiento de resolución que en el contraste bilateral.

Paso 1. Las hipótesis son:

$$H_0 : \mu \geq 800$$
$$H_1 : \mu < 800$$

Se trata de un contraste unilateral por la izquierda.

Paso 2. Fijamos el nivel de significación $1 - \alpha = 95\% = 0{,}95$, $\alpha = 0{,}05$.

Paso 3. Determinamos la región de aceptación. De la probabilidad $P[Z > -z_\alpha] = P[Z < z_\alpha] = 0{,}95$, obtenemos en las tablas $z_\alpha = 1{,}645$. Entonces la región de aceptación es:

$$\left(\mu_0 - \frac{z_\alpha \cdot \sigma}{\sqrt{n}};\ \infty\right) = \left(800 - 1{,}645 \cdot \frac{120}{\sqrt{50}};\ \infty\right) = (800 - 27{,}9;\ \infty) = (772{,}08;\ \infty).$$

Paso 4. Para aceptar H_0 tiene que ocurrir que $\overline{X}$ caiga en el intervalo $(772{,}08;\ \infty)$.

Evidentemente $\overline{X} = 750$ no pertenece al intervalo $(772{,}08;\ \infty)$. Luego se rechaza la hipótesis nula. No es cierto que las bombillas tengan una vida media de 800 horas.

Actividades

20. En una ciudad se sabe que la edad en que los hijos se independizan de sus padres es una variable de distribución normal con media 29 años y desviación típica 3 años. Se sospecha que la media anterior ha descendido debido a medidas liberalizadoras del mercado de trabajo, mientras que la desviación típica se mantiene. Un encuesta reciente a 100 jóvenes que se acaban de independizar revela una media de edad de 28,1 años.

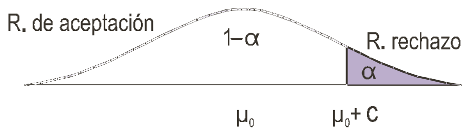
Con un nivel de confianza del 99% ¿puede mantenerse que la edad media de independencia de los jóvenes sigue siendo de 29 años? Plantear el contraste de hipótesis y resolverlo.

21. El coeficiente intelectual de una universidad sigue una distribución normal de media desconocida y desviación típica 14. Si en una muestra de 64 alumnos se observó un coeficiente intelectual medio de 106 puntos, ¿se puede aceptar, con un nivel de significación de 0,01, que el coeficiente intelectual medio de la población de esa universidad es $\mu = 110$ puntos?

3.3. Contraste unilateral por la derecha

En otros casos debemos contrastar si la hipótesis alternativa es mayor que la hipótesis nula. Estamos entonces ante un contraste unilateral por la derecha cuyas hipótesis son: $\begin{cases} H_0 : \mu \le \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$

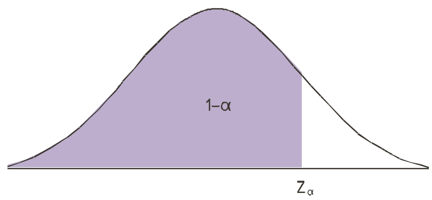
En este caso la región aceptación es el intervalo $(-\infty,\ \mu_0 + c)$ y la región crítica o región de rechazo $(\mu_0 + c,\ \infty)$ está situada a la derecha de la distribución de las medias muestrales.



Para calcular c partimos de $P[\overline{X} < \mu_0 + c] = 1 - \alpha$ Restando μ_0 y dividiendo por $\sigma/\sqrt{n}$ tipificamos $\overline{X}$:

$$P\left[\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{\mu_0 + c - \mu_0}{\sigma/\sqrt{n}}\right] = 1-\alpha,$$

$$P\left[Z < \frac{c}{\sigma/\sqrt{n}}\right] = 1-\alpha \quad \text{o} \quad P[Z < z_\alpha] = 1-\alpha.$$



Buscamos en las tablas un valor, z_α, de la $N(0, 1)$ tal que $P[Z < z_\alpha] = 1 - \alpha$.

Entonces, como $z_\alpha = \dfrac{c}{\sigma/\sqrt{n}} \Rightarrow c = \dfrac{z_\alpha \cdot \sigma}{\sqrt{n}}$, la región de aceptación es: $\left(-\infty,\ \ \mu_0 + \dfrac{z_\alpha \cdot \sigma}{\sqrt{n}}\right)$.

Ejemplos

9. Un servicio telefónico de atención al cliente asegura que el tiempo medio de espera antes de ser atendidos es de 5 minutos y desviación típica 0,6 minutos. Se toma una muestra de 36 llamadas y arrojan una media de espera de 5,2 minutos. ¿Existen razones para creer, con un nivel de significación de 0,05, que el tiempo de espera es mayor que 5 minutos?

Solución:

Paso 1. Las hipótesis son:

$H_0 : \mu \le 5$
$H_1 : \mu > 5$

Se trata de un contraste lateral por la derecha.

Paso 2. Fijamos el nivel de significación $\alpha = 0{,}05$.

Paso 3. Determinamos de $P[Z \le z_\alpha] = 0{,}95$, la región de aceptación y en las tablas obtenemos $z_\alpha = 1{,}645$.

Entonces la región de aceptación es: $\left(-\infty;\ \ \mu_0 + \dfrac{z_\alpha \cdot \sigma}{\sqrt{n}}\right) = \left(-\infty;\ \ 5 + 1{,}645 \cdot \dfrac{0{,}6}{\sqrt{36}}\right) = (-\infty;\ 5{,}1645)$.

Paso 4. Para aceptar H_0 tiene que ocurrir que la media de la muestra, $\overline{X}$, caiga en el intervalo $(-\infty;\ 5{,}1645)$. Como $\overline{X} = 5{,}2$ no pertenece al intervalo $(-\infty;\ 5{,}1645)$, entonces se rechaza la hipótesis nula. No es cierto que el tiempo medio de espera sea 5 minutos.

Actividades

22. Una muestra de 64 soldados de un regimiento ha dado una altura media de 174 cm ¿Se puede aceptar con un nivel de significación de 0,01 que la talla media de los soldados del regimiento sigue siendo los 172 cm y no ha aumentado con los últimos alistamientos? Se supone que la desviación típica poblacional sigue siendo de 8 cm.

23. Un estudio entre la población de deportistas de élite indica que su peso se distribuye normalmente con media 70 Kg y desviación típica 10 kg. Se elige al azar una muestra de 64 deportistas de una determinada disciplina y ha dado un peso medio de 75 kg. Con un nivel de significación de 0,05, ¿puede decirse que los deportistas que practican este disciplina deportiva pesan más que el resto de deportistas de élite?

3.4. Contraste de hipótesis para la media con σ desconocida

Si de la variable X desconocemos σ, desviación típica poblacional, pero el tamaño de la muestra es suficientemente grande ($n \geq 30$), la cuasi desviación típica $S = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$ que en muchas ocasiones se llama simplemente **desviación típica muestral** es un buen estimador de la desviación típica poblacional.

Ejemplos

10. En un servicio telefónico de atención al cliente se asegura que el tiempo medio de espera antes de ser atendidos es igual o inferior a 12 minutos. Se toma una muestra de 300 llamadas y arrojan una media de espera de 14 minutos y una desviación típica muestral de 5 minutos. ¿Existen razones para creer, con un nivel de significación de 0,05, que el tiempo de espera es mayor que 12 minutos?

Solución:

Paso 1. Las hipótesis son: $H_0 : \mu \leq 12$ y $H_1 : \mu > 12$.

Se trata de un contraste lateral por la derecha.

Paso 2. Fijamos el nivel de significación $\alpha = 0,05$.

Paso 3. De $P[Z < z_\alpha]$, determinamos $z_\alpha = 1,645$. Si tomamos $\sigma = S = 5$, entonces la región de aceptación es:

$$\left(-\infty;\ \mu_0 + \frac{z_\alpha \cdot \sigma}{\sqrt{n}}\right) = \left(-\infty;\ 12 + 1,645 \cdot \frac{5}{\sqrt{300}}\right) = (-\infty;\ 12,474).$$

Paso 4. Como $\overline{X} = 14$ no pertenece al intervalo $(-\infty;\ 12,474)$, entonces se rechaza la hipótesis nula. No es cierto que el tiempo medio de espera sea como mucho 12 minutos.

4. Contraste de hipótesis para la proporción

El contraste de hipótesis para una proporción o porcentaje se basa en los mismos principios del contraste de hipótesis para la media; tan solo tenemos que recordar que si una población tiene una proporción poblacional p de una determinada característica, entonces la variable aleatoria $\hat{p}$, de las proporciones muestrales, cuando $n \geq 30$, se aproxima a una distribución normal $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Si ponemos en vez de p, proporción poblacional, el valor de p_0, hipótesis que queremos contrastar, entonces debe aproximarse a una normal $N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

El proceso se realiza en los 4 pasos que formulamos en el contraste de la media.

1. Establecer la hipótesis que provisionalmente se considera verdadera, H_0, que p, media poblacional, tenga un valor p_0; es decir,

$$H_0 : p = p_0$$

La hipótesis complementaria de la hipótesis nula es la hipótesis alternativa, H_1, y que puede ser de tres tipos diferentes:

$H_1 : p \neq p_0$ (contraste bilateral)

$H_1 : p < p_0$ (contraste unilateral por la izquierda)

$H_1 : p > p_0$ (contraste unilateral por la derecha)

2. Fijar el nivel de significación α, que indica la probabilidad de rechazar H_0 aun siendo verdadera, o establecer el nivel de confianza $1 - \alpha$, que indica la probabilidad de aceptar H_0 cuando es cierta.

3. Determinar la región de aceptación para el nivel de significación α supone determinar un intervalo ($p_0 - c$, $p_0 + c$) al que pertenezca la proporción muestral con probabilidad $1 - \alpha$. Luego c es un número que cumple que: $P[|\hat{p} - p_0| < c] = 1 - \alpha$.

Dividiendo la desigualdad por $\sqrt{\frac{p_0(1-p_0)}{n}}$, tipificamos $\hat{p}$; y llamando $z_{\alpha/2} = \frac{c}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, nos queda

$P\left[\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{\alpha/2}\right] = 1 - \alpha$. De la desigualdad del corchete obtenemos $|\hat{p} - p_0| < z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$

que conduce, cuando el contraste es bilateral, a la región de aceptación buscada:

$$\left(p_0 - z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}},\quad p_0 + z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right).$$

En los contrastes unilaterales las regiones de aceptación son:

- por la izquierda $\left(p_0 - z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}},\quad \infty\right)$;
- por la derecha $\left(-\infty,\quad p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

4. Se extrae una muestra y se calcula la proporción muestral, $\hat{p}$; a continuación se comprueba si cae dentro o fuera de la región de aceptación. Si cae dentro se acepta H_0 y si no, como es muy improbable que la proporción muestral obtenida siga la distribución normal de las proporciones muestrales, se rechaza.

Ejemplos

11. Un comerciante de productos informáticos asegura que el 45% de los hogares de cierta ciudad poseen ordenador. Se extrae una muestra de 300 hogares y resulta que 120 poseen ordenador. ¿Se puede rechazar la afirmación del comerciante con un nivel de significación del 5%?

Solución:

Paso 1. Las hipótesis del contraste son:

$H_0 : p = 0{,}45$ y $H_1 : p \neq 0{,}45$ (contraste bilateral).

Paso 2. El nivel de significación es $\alpha = 0{,}05$ y $1 - \alpha = 0{,}95$.

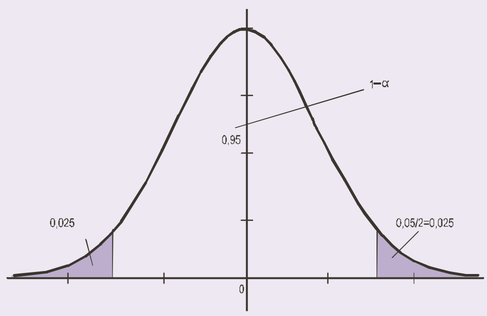
Paso 3. Determinamos la región de aceptación:

$$\left(p_0 - z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}, \quad p_0 + z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}} \right)$$

Como α = 5% = 0,05, $\alpha/2$ = 0,025, según el gráfico

vemos que $P\left[|Z| < z_{\alpha/2}\right] = 1 - 0{,}05 = 0{,}95.$

$P\left[Z < z_{\alpha/2}\right] = 0{,}95 + 0{,}025 = 0{,}975.$



Buscando en las tablas el valor de $z_{\alpha/2}$ obtenemos que $z_{\alpha/2}$ = 1,96; en consecuencia, como n = 300 y p_0 = 0,45, la región de aceptación será:

$$\left(0{,}45 - 1{,}96 \cdot \sqrt{\frac{0{,}45 \cdot 0{,}55}{300}}; \quad 0{,}45 + 1{,}96 \cdot \sqrt{\frac{0{,}45 \cdot 0{,}55}{300}} \right) = (0{,}394; \quad 0{,}506).$$

Paso 4. De la muestra $\hat{p}$ = 120/300 = 0,4, y además cae dentro del intervalo (0,393; 0,506); por tanto, aceptamos la afirmación del comerciante.

12. La experiencia de anteriores elecciones muestran que cierto partido obtiene el 15% de los votos en una ciudad. ¿Se puede aceptar esta afirmación, para un nivel de significación de 0,1, si en la última encuesta sólo 205 personas, entre 1500, se mostraron favorables a dicho partido?

Solución:

Paso 1. Las hipótesis del contraste son:

$H_0 : p \geq 0{,}15$ y $H_1 : p < 0{,}15$ (contraste unilateral por la izquierda).

Paso 2. El nivel de significación es $\alpha = 0{,}1$ y $1 - \alpha = 0{,}9$.

Paso 3. Determinamos la región de aceptación: $\left(p_0 - z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}, \quad \infty \right)$ de $P\left[Z > -z_\alpha\right] = P\left[Z < z_\alpha\right] = 0{,}9.$

obtenemos en las tablas z_α = 1,28 y como p_0 = 0,15 y n = 1500, la región de aceptación es

$$(0{,}15 - 1{,}28 \cdot \sqrt{\frac{0{,}15 \cdot 0{,}85}{1500}};\ \infty) = (0{,}138;\ \infty).$$

Paso 4. De la muestra $\hat{p}$ = 205/1500 = 0,136, y como vemos no cae dentro del intervalo (0,138; ∞). Por tanto, rechazamos la hipótesis nula.

13. Un fabricante de calentadores de gas afirma que como máximo el 2% de los calentadores de gas que comercializa tienen una avería durante el periodo de garantía. Sin embargo, en una muestra de 400 calentadores se encontró que 12 de ellos han tenido una avería durante el primer año de garantía, ¿se puede aceptar la afirmación del fabricante con un nivel de significación del 0,05?

Solución:

Paso 1. Las hipótesis del contraste son:

$H_0 : p \leq 0{,}02$ y $H_1 : p > 0{,}02$ (contraste unilateral por la derecha).

Paso 2. El nivel de significación es α = 0,05 y 1 – α = 0,95.

Paso 3. Determinamos la región de aceptación:

De $P[Z < z_\alpha]$ = 0,95, obtenemos en las tablas z_α = 1,645 y como p_0 = 0,02 y n = 400, la región de aceptación es

$$(-\infty;\ 0{,}02 + 1{,}645 \cdot \sqrt{\frac{0{,}02 \cdot 0{,}98}{400}}) = (-\infty;\ 0{,}031).$$

Paso 4. En la muestra, $\hat{p}$ = 12/400 = 0,03, y como vemos cae dentro del intervalo ($-\infty$; 0,031). Por tanto, aceptamos la hipótesis nula.

Actividades

24. Un dentista asegura que el 35% de los niños de 10 años presenta algún tipo de caries dental. Una muestra de 100 niños reveló que 32 presentaban algún tipo de caries. Comprueba con un nivel de significación de 0,05 si el resultado de la muestra confirma o no la afirmación del dentista.

25. Un ayuntamiento asegura que el 40% de los hogares de la ciudad tiene calefacción de gas. La compañía suministradora de gas sospecha que no son tantos. Se toma una muestra de 200 hogares y resulta que 76 tienen instalación de calefacción a gas. Con un nivel de significación del 0,05, ¿confirma el resultado de la muestra la afirmación del ayuntamiento?

26. Un vendedor de periódicos afirma que 3 de cada 10 habitantes de una determinada ciudad lee el diario LA NACIONCITA. Se elige una muestra de 144 habitantes de la citada ciudad y resulta que 32 admiten leerlo. Con un nivel de significación del 5% contrasta si la afirmación del vendedor de periódicos es exagerada.

5. Comparación de dos medias

La comparación de dos medias es un caso particular del contraste de hipótesis. Supongamos que estamos interesados en conocer si el salario medio de los dependientes de comercio en una ciudad A es igual, o diferente, al salario de los dependientes de comercio de otra ciudad B. Para resolver esta cuestión extraemos dos muestras aleatorias, de tamaños n_1 y n_2, de los salarios de los dependientes, una en cada ciudad, y comparamos las medias muestrales. Esta comparación se hace sometiendo a un contraste la hipótesis de la igualdad de los dos salarios medios.

Sabemos que los salarios de los dependientes de la ciudad A se distribuyen según una $N(\mu_1, \sigma_1)$ y que los de la ciudad B lo hacen según una $N(\mu_2, \sigma_2)$. Vimos en la Unidad didáctica 9, que la variable $\overline{X}_1 - \overline{X}_2$, que es la diferencia de las medias muestrales de dos distribuciones, se distribuye como una $N\left(\mu_1-\mu_2, \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$; entonces los pasos del contraste son los siguientes:

Paso 1. Como $\overline{X}_1 - \overline{X}_2$ tiene media $\mu_1-\mu_2$, entonces las hipótesis nula y alternativa son:

$$H_0: \mu_1-\mu_2=0 \quad \text{o} \quad \mu_1=\mu_2$$
$$H_1: \mu_1-\mu_2\neq 0 \quad \text{o} \quad \mu_1\neq\mu_2$$

Paso 2. Fijamos un nivel de significación α o un nivel de confianza $1-\alpha$.

Paso 3. La región de aceptación, cuando la desviación típica de $\overline{X}_1 - \overline{X}_2$ es $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$, viene dada por:

$$\left(\mu_1-\mu_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\quad \mu_1-\mu_2+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=\left(-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\quad z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

en donde $z_{\alpha/2}$ es una abscisa de la $N(0,1)$ que deja a su derecha un área de probabilidad $\alpha/2$.

Paso 4. Se extraen las muestras y calculamos la diferencia muestral $\overline{X}_1 - \overline{X}_2$ y se comprueba si cae dentro o fuera de la región de aceptación. Si cae dentro, aceptamos que $\mu_1=\mu_2$ y si no, se rechaza.

Todo este proceso se puede abreviar considerablemente. Si $\overline{X}_1 - \overline{X}_2$ cae dentro del intervalo $\left(-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\quad z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$, podemos escribir:

$$-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\overline{X}_1-\overline{X}_2<z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

Dividiendo las desigualdades anteriores por: $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$. Se obtiene entonces $-z_{\alpha/2}<\dfrac{\overline{X}_1-\overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}<z_{\alpha/2}$.

Luego, usando el valor absoluto, resulta: $\left|\dfrac{\overline{X}_1-\overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right|<z_{\alpha/2}$ o $\dfrac{\left|\overline{X}_1-\overline{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}<z_{\alpha/2}$.

Esta última desigualdad la emplearemos como un criterio rápido para decidir si debemos aceptar o rechazar la hipótesis nula. Y el **criterio para la comparación de dos medias** es el siguiente:

- Si $\dfrac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} < z_{\alpha/2}$, se acepta la hipótesis nula
- Si $\dfrac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$, se rechaza la hipótesis nula.

Ejemplo

14. De un test de memoria que se aplica a estudiantes de bachillerato se sabe, porque se ha aplicado muchas veces, que la desviación típica de los alumnos es 32, y 36 para las alumnas. Se escoge una muestra de 40 alumnos y 45 alumnas y se obtienen unas puntuaciones medias muestrales de 172 y 178 respectivamente. Con un nivel de significación del 5%, ¿se puede asegurar que las puntuaciones medias poblacionales de chicos y chicas son iguales?

Solución: Se trata de un contraste de diferencia de medias con desviaciones típicas conocidas. Llamamos μ_1 = *puntuación media de la población de alumnos* y μ_2 = *puntuación media de la población de alumnas*. Las hipótesis nula y alternativa son:

$$H_0: \mu_1 = \mu_2 \text{ o } \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 \neq \mu_2 \text{ o } \mu_1 - \mu_2 \neq 0$$

Si $\sigma_1 = 32$, $\sigma_2 = 36$, $n_1 = 40$, $n_2 = 45$, $\overline{X}_1 = 172$, $\overline{X}_2 = 178$, $\alpha = 0{,}05$ y $z_{\alpha/2} = 1{,}96$, calculamos:

$$\frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} = \frac{|172-178|}{\sqrt{\dfrac{32^2}{40}+\dfrac{36^2}{45}}} = 0{,}81 < 1{,}96.$$

Al ser el valor de la fracción menor que 1,96, se acepta la hipótesis nula. Las puntuaciones medias de los alumnos y de las alumnas son iguales.

Actividades

27. Se han registrado los resultados de las pruebas de selectividad de 100 estudiantes que han asistido a clases particulares y han obtenido una puntuación media de 4,96 con una desviación típica muestral de 0,98. Otro grupo de 100 estudiantes que no fueron a clases particulares obtuvieron una media de 4,77 con una desviación típica muestral de 1,02. Con un nivel de significación del 5%, ¿se puede afirmar que no hay diferencia significativa entre las calificaciones de los estudiantes que van a clases particulares y los que no van? (Nota: se desconocen σ_1 y σ_2, pero se pueden estimar por las desviaciones típicas muestrales).

28. Una empresa multinacional quiere comparar el nivel educativo de sus empleados en dos países diferentes, A y B. Se ha realizado un test idéntico a 140 empleados en cada país y las puntuaciones medias obtenidas figuran en la tabla siguiente:

A	$\overline{X}_1 = 78$	$S_1 = 11{,}5$
B	$\overline{X}_2 = 86$	$S_2 = 13$

Sabiendo que S_1 y S_2 son las desviaciones típicas muestrales, y son un buen estimador de las desviaciones típica poblacionales, con nivel de significación del 5%, averiguar si son iguales o no las puntuaciones medias registradas en los dos países.

6. Comparaciones de dos proporciones

Estamos interesados en saber si la proporción de individuos que poseen una cierta característica en una población es igual a la proporción de individuos con esa característica en otra población.

Se trata, por tanto, de comparar dos porcentajes y para efectuar la comparación hacemos un contraste de hipótesis para determinar si la diferencia de porcentajes es significativa o se debe a aleatoriedad en la extracción de las muestras.

Supongamos una característica que tiene un porcentaje p_1, en la primera población y p_2 en la segunda población. Si tomamos dos muestras, n_1 y n_2, respectivamente en cada población y con tamaño suficientemente grande, entonces la distribución de las diferencias de las proporciones muestrales $\hat{p}_1 - \hat{p}_2$ se distribuye normalmente, según vimos en la Unidad didáctica 9, como una:

$$N\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Para comparar las dos proporciones en poblaciones diferentes contrastamos la hipótesis de la igualdad de las dos proporciones $p_1 = p_2$. Es decir, efectuamos un contraste de hipótesis

Paso 1. Las hipótesis nula y alternativa son:

$$H_0: p_1 - p_2 = 0 \text{ o } p_1 = p_2$$
$$H_1: p_1 - p_2 \neq 0 \text{ o } p_1 \neq p_2$$

Paso 2. Fijamos un nivel de significación α o un nivel de confianza $1 - \alpha$.

Paso 3. La región de aceptación, cuando la desviación típica de $\hat{p}_1 - \hat{p}_2$ es $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$, viene dada por:

$$\left((p_1 - p_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},\ \ p_1 - p_2 + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) =$$
$$= \left(-z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},\ \ z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right);$$

donde $z_{\alpha/2}$ es una abscisa de la $N(0, 1)$ que deja a su derecha un área de probabilidad $\alpha/2$.

Paso 4. Se extraen las muestras, n_1 y n_2, y se calcula la diferencia muestral $\hat{p}_1 - \hat{p}_2$ y se comprueba si cae dentro o fuera de la región de aceptación, en cuyo caso se acepta o se rechaza la hipótesis nula.

Todo este proceso se puede abreviar considerablemente. Si $\hat{p}_1 - \hat{p}_2$ cae en la región de aceptación se cumple que:

$$-z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, < \hat{p}_1 - \hat{p}_2 < z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Dividiendo por $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ resulta: $-z_{\alpha/2} < \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} < z_{\alpha/2}$.

O dicho de otro modo, que

$$\left|\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right| < z_{\alpha/2} \quad \text{ó} \quad \frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} < z_{\alpha/2}$$

dado que en la fórmula de la desviación típica el radicando siempre es positivo.

Esta última desigualdad la emplearemos como criterio cómodo y rápido para decidir si debemos aceptar o rechazar la hipótesis nula. Y el **criterio para la comparación de dos proporciones** es el siguiente:

- Si $\dfrac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} < z_{\alpha/2}$, se acepta la hipótesis nula.
- Si $\dfrac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} > z_{\alpha/2}$, se rechaza la hipótesis nula.

Claro que para emplear este criterio y utilizar la fórmula que en él aparece necesitamos conocer p_1 y p_2. Y si conocemos p_1 y p_2, podemos compararlos sin contraste ni otros requisitos. Por lo tanto, estos valores serán siempre desconocidos y debemos estimar la desviación típica de la $N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1}+\sqrt{\frac{p_2(1-p_2)}{n_2}}}\right.$ por $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, que cuando las muestras son grandes no difieren mucho. Con lo cual los criterios anteriores debemos escribirlos así:

- Si $\dfrac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < z_{\alpha/2}$, se acepta la hipótesis nula.
- Si $\dfrac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_{\alpha/2}$, se rechaza la hipótesis nula.

Ejemplos

15. Se quiere saber si el porcentaje de estudiantes de una universidad A que practican habitualmente deporte es el mismo que el porcentaje de los que hacen deporte en otra universidad B. Se toman muestras de 150 alumnos en cada universidad y se encuentra que en A, 75 practican deporte, mientras que en B sólo lo hacen 65. ¿Se puede afirmar, con un nivel de significación del 5%, que el porcentaje de alumnos que practican deporte es el mismo en ambas universidades?

Solución:

Se trata de contrastar si la diferencia de porcentajes es cero o no. Las hipótesis nula y alternativa son:

$$H_0: p_1 = p_2 \text{ o } p_1 - p_2 = 0$$
$$H_1: p_1 \neq p_2 \text{ o } p_1 - p_2 \neq 0$$

Si $\hat{p}_1 = 75/150 = 0{,}5$, $\hat{p}_2 = 65/150 = 0{,}43$, $n_1 = 150$, $n_2 = 150$, $\alpha = 0{,}05$ y $z_{\alpha/2} = 1{,}96$, calculamos:

$$\frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{\left|0{,}5 - 0{,}43\right|}{\sqrt{\dfrac{0{,}5(1-0{,}5)}{150} + \dfrac{0{,}43(1-0{,}43)}{150}}} = 1{,}23 < 1{,}96$$

Se acepta la hipótesis nula, el porcentaje de alumnos que practican deporte es el mismo en ambas universidades.

Actividades

29. Una tienda de teléfonos móviles ha vendido 160 aparatos de la marca A de los que 14 de ellos han presentados averías en el periodo de garantía. Simultáneamente ha vendido 135 aparatos de la marca B de los que 15 han presentado fallos durante el periodo de garantía. ¿Es cierta la afirmación del vendedor de que el porcentaje de averías en ambas marcas es el mismo?

30. Una fábrica de botellas de vidrio tiene dos sistemas distintos para la fabricación de las botellas. Se tomó una muestra de 300 botellas, de un sistema de producción, y aparecieron 25 defectuosas, mientras que en una muestra de 200 botellas del otro sistema aparecieron 13 defectuosas. ¿Se puede afirmar que los porcentajes de botellas defectuosas por ambos sistemas de producción son iguales?

RECUERDA

- Intervalo de confianza para la media: $\left(\bar{X} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \bar{X} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right)$

- Tamaño de la muestra para la media: $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$

- Intervalo de confianza para la proporción: $\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

- Tamaño de la muestra para la proporción: $n = (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{E^2}$

- Región de aceptación para el contraste bilateral de la media: $\left(\mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right)$

- Región de aceptación para el contraste bilateral de la proporción: $\left(p_0 - z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}},\ p_0 + z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right)$

- Criterio para la comparación de dos medias:
  - Si $\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}$, se acepta la hipótesis nula.
  - Si $\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$, se rechaza la hipótesis nula.

- Criterio para la comparación de dos proporciones:
  - Si $\frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < z_{\alpha/2}$, se acepta la hipótesis nula.
  - Si $\frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_{\alpha/2}$, se rechaza la hipótesis nula.