Intervalos de Confianza

NATALIA ACEVEDO PRINS

natalia.acevedop@udea.edu.co

INTERVALOS DE CONFIANZA

Conceptos básicos:

ESTIMADOR PUNTUAL.

Estadístico calculado a partir de información de la muestra para estimar el parámetro poblacional.

INTERVALO DE CONFIANZA

Conjunto de valores formado a partir de una muestra de datos, tal que, exista la posibilidad de que el parámetro poblacional ocurra dentro del conjunto con una probabilidad específica*.

*La probabilidad específica recibe el nombre de **nivel de confianza.**

ESTIMACIÓN DE LOS INTERVALOS DE CONFIANZA

Para calcular el intervalo de confianza consideraremos dos situaciones:

σ conocida

- Usamos los datos de la muestra para calcular **μ** con $\overline{X}$, y desviación estándar de la población **(σ) es conocida**.

σ desconocida

- Usamos los datos de la muestra para calcular **μ** con $\overline{X}$, y la desviación estándar de la población es desconocida.
- Sustituimos la desviación estándar de la población **(σ)** por la desviación estándar de la muestra.

Estimación con σ conocida

Conocer la desviación estándar de la población, σ permite simplificar el calculo del intervalo de confianza, porque podemos utilizar la distribución normal.

Recuerde que:

La distribución muestral de la media es la distribución de todas las medias muestrales, $\bar{X}$, con tamaño de la muestra, n, de una población.

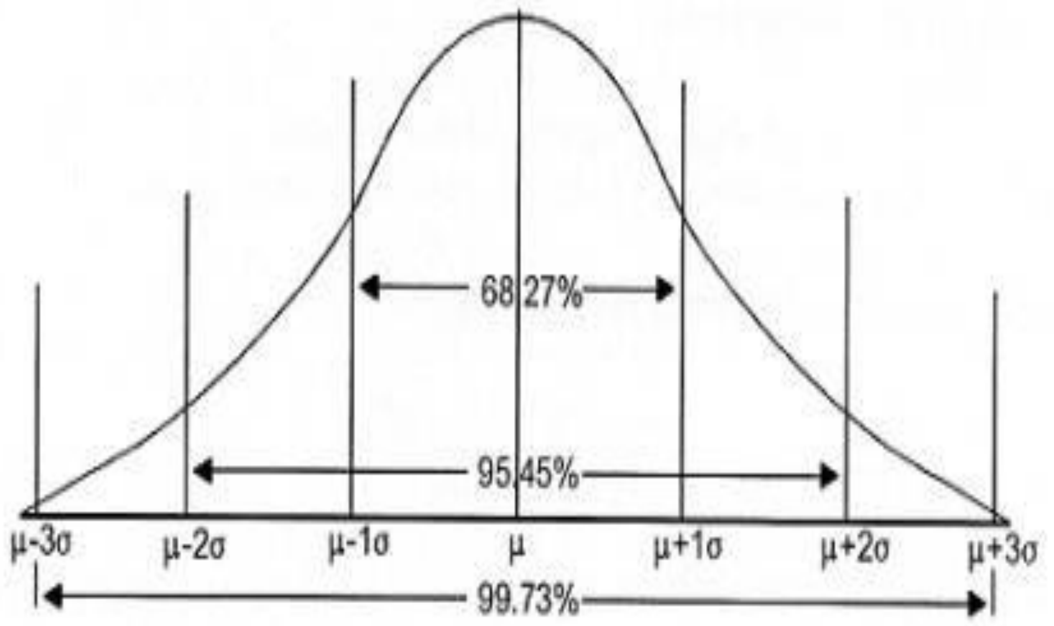
Se conoce la desviación estándar de la población, σ. Por teorema central del límite, la distribución muestral sigue una distribución de probabilidad normal con una media μ y una desviación estándar $\sigma / \sqrt{n}$.

Estimación con σ conocida

Por teorema central del límite afirmamos que:

95% de las medias muestrales de una población se encontrará dentro de 1.96 errores estándares (μ).

99% de las medias muestrales se encontrará a 2.58 errores estándares de la media poblacional.



La amplitud del intervalo depende del nivel de confianza y de la magnitud del error estándar de la media.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Ejemplo Cálculo del intervalo

Una empresa que distribuye duraznos en latas de 4 onzas. Para asegurarse de que cada lata contenga por lo menos la cantidad que se requiere, la empresa establece que el proceso de llenado debe verter 4.01 onzas de duraznos y almíbar en cada lata. 4.01 es la media poblacional. Por supuesto, no toda lata contendrá exactamente 4.01 onzas de duraznos y almíbar. Algunas latas contendrán más y otras menos.

- Si se selecciona una muestra aleatoria de 64 latas y se determina la media de la muestra. Ésta es de 4.015 onzas de duraznos y almíbar. La desviación estándar del proceso es de 0.04 onzas.
- El proceso se rige por la distribución de probabilidad normal.

Solución Cálculo del intervalo

El intervalo de confianza de 95% de la media poblacional de esta muestra particular es:

$$\bar{x} = 4.015$$

$$\sigma = 0.04$$

$$n = 64$$

$$Z_{\alpha/2} = 1.96$$

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$4.015 \pm 1.96 \frac{0.04}{\sqrt{64}} =$$

$$4.015 \pm 0.0098$$

Desviación estándar poblacional σ desconocida

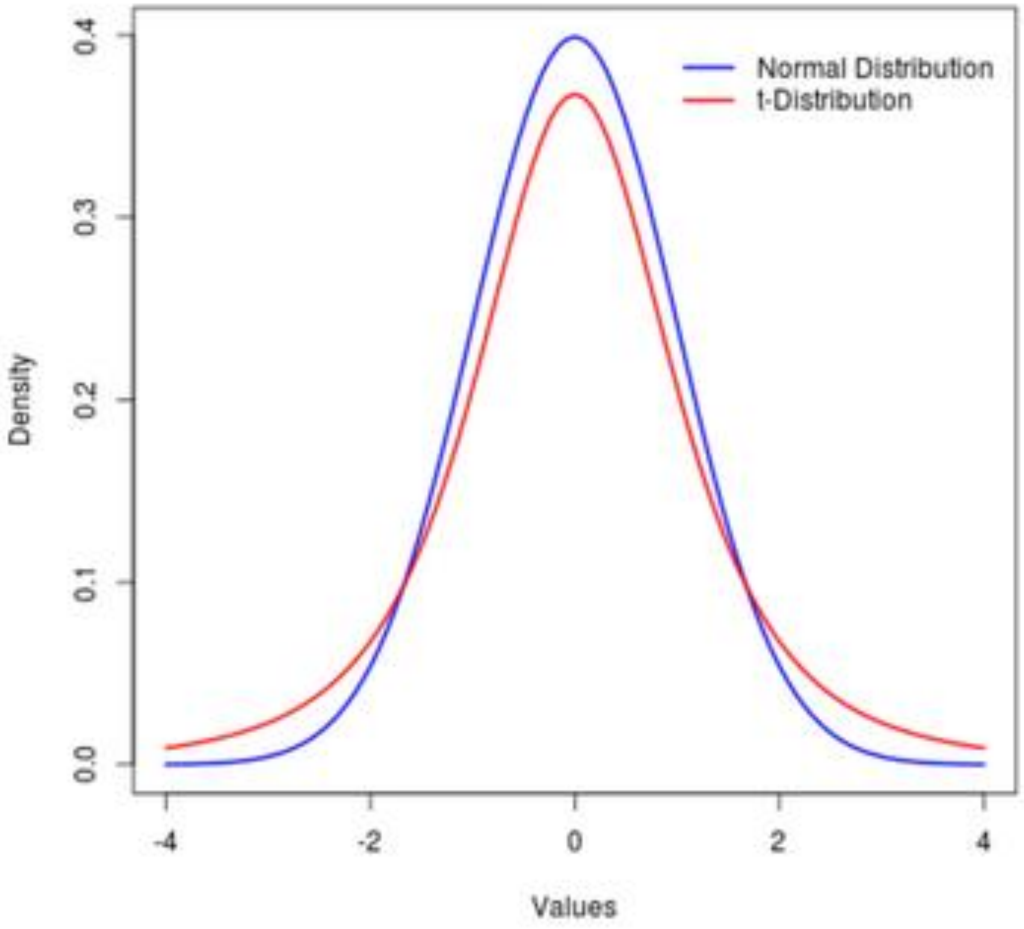
En la mayoría de los casos de muestreo no se conoce la desviación estándar de la población

Se utiliza la desviación estándar de la muestra $\boldsymbol{s}$ para estimar la desviación estándar poblacional $\boldsymbol{\sigma}$.

Como **no se conoce $\boldsymbol{\sigma}$, no se** puede utilizar la distribución z. Pero se puede utilizar la desviación estándar de la media y **sustituir** la distribución z con la distribución $\boldsymbol{t}$.

T -student

La distribución t es una distribución de probabilidad continua similar a las de la distribución z.



$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Hay discrepancias entre s y σ cuando s se calculaba a partir de una muestra muy pequeña.

La distribución t es más plana y se extiende más que la distribución normal estándar.

La desviación estándar de la distribución t es mayor que la distribución normal estándar

Propiedades de la t-Student

Estas características de t tiene el supuesto de que la población de interés es de naturaleza normal, o casi normal.

Como en el caso de la distribución z, es una distribución continua.

Como en el caso de la distribución z, tiene forma de campana y es simétrica.

No existe una distribución t, sino una familia de distribuciones t. Todas tienen una media 0, y sus s difieren de acuerdo con el tamaño de la muestra, n.

A mas observaciones mayor desviación.

En La distribución t conforme se incrementa el tamaño de la muestra, esta se aproxima a la distribución normal estándar, pues los errores que se cometen al utilizar s para estimar σ disminuyen con muestras más grandes.

Intervalo de confianza con t-Student

Para crear un intervalo de confianza de la media poblacional con una desviación estándar desconocida:

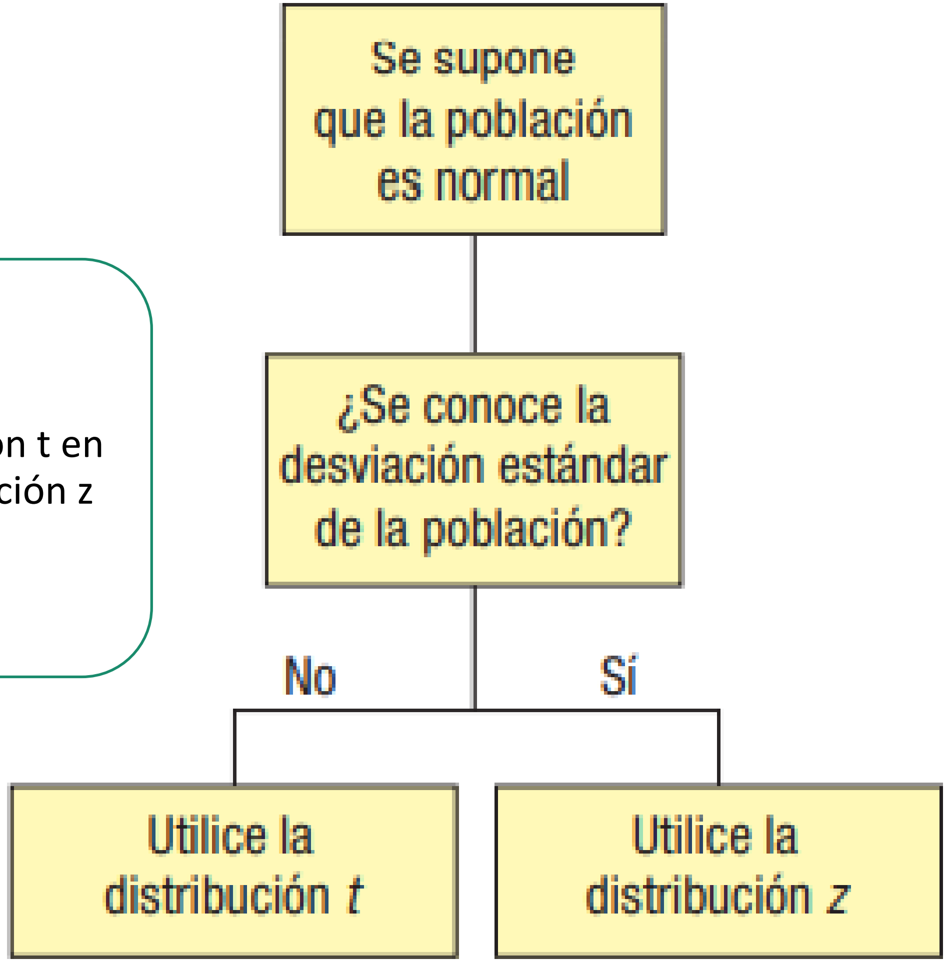
Suponga muestra es aproximadamente normal.

Es cuestionable en el caso de muestras pequeñas, y es más válida en el de muestras más grandes.

Estime la desviación estándar de la población (σ) con la desviación estándar de la muestra (s).

Utilice la distribución t en lugar de la distribución z

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$



Ejemplo

Un fabricante de llantas desea investigar la durabilidad de sus productos. Una muestra de 10 llantas que recorrieron 50 000 kilómetros reveló una media muestral de 0.32 cm de huella restante con una desviación estándar de 0.09 cm. Construya un intervalo de confianza de 95% de la media poblacional.

¿Sería razonable que el fabricante concluyera que después de 50 000 Km la cantidad media poblacional de huella restante es de 0.30 cm?

Solución

No se conoce la desviación estándar de la población, pero sí la desviación estándar de la muestra, que es de 0.09 cm.

$$\bar{x} \pm t\frac{s}{\sqrt{n}} = 0.32 \pm 2.262\frac{0.09}{\sqrt{10}} = 0.32 \pm 0.064$$

Como el valor de 0.30 se encuentra en este intervalo, es posible que la media de la población sea de 0.30 cm

Intervalo de confianza de una proporción

PROPORCIÓN

Fracción, razón o porcentaje que indica la parte de la muestra de la población que posee un rasgo de interés particular.

Una encuesta reciente indicó que 92 de cada 100 entrevistados estaban de acuerdo con el horario de verano para ahorrar energía.

- *La proporción de la muestra es de 92/100, o 0.92, o 92%.*
- *Si $\hat{p}$ representa la proporción de la muestra, X el número de éxitos y n el número de elementos de la muestra.*

PROPORCIÓN MUESTRAL

$$\hat{p} = \frac{x}{n}$$

Intervalo de confianza de una proporción

Sea $\hat{p}$ la proporción observada de «éxitos» en una muestra aleatoria de n observaciones, una población que tiene una proporción de éxitos P.

Se obtiene un intervalo de confianza al $100(1-\alpha)\%$ de la proporción de la población de la siguiente manera:

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

$$\hat{p} \pm ME$$

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

Ejemplo intervalo para proporciones (muestras grandes)

El sindicato que representa a Bottle Blowers of America (BBA) considera la propuesta de fusión con Teamsters Union. De acuerdo con el reglamento del sindicato de BBA, por lo menos tres cuartas partes de los miembros del sindicato deben aprobar cualquier fusión.

Una muestra aleatoria de 2 000 miembros actuales de BBA revela que 1 600 planean votar por la propuesta.

- Determine el intervalo de confianza de 95% de la proporción poblacional. Fundamente su decisión en esta información de la muestra:
- ¿puede concluir que la proporción necesaria de miembros del BBA favorece la fusión? ¿Por qué?

Solución

Primero calculamos la proporción de la muestra:

$$p = \frac{x}{n} = \frac{1600}{2000} = 0.80$$

Luego, se calcula que 80% de la población favorece la propuesta de fusión

$$p \pm z\sqrt{\frac{p(1-p)}{n}} = 0.80 \pm 1.96\sqrt{\frac{0.80(1-.80)}{2000}}$$
$$= 0.80 \pm 0.018$$

El valor z correspondiente al nivel de confianza de 95% es de 1.96.

Los puntos extremos del intervalo de confianza son 0.782 y 0.818. El valor inferior es mayor que 0.75. A es probable que se apruebe la propuesta de fusión, ya que se incluyen valores superiores a 75% de los miembros del sindicato.

Elección del tamaño de muestra adecuado

El tamaño de la muestra, en la práctica, es una decisión que se toma para que la estimación del parámetro de población sea bueno. Esta decisión se basa en tres variables:

1. *El margen de error que tolerará el investigador.*
2. *El nivel de confianza deseado.*
3. *La variabilidad o dispersión de la población que se estudia*

Tamaño de la muestra para calcular una media poblacional

$$n = \left(\frac{z\sigma}{E}\right)^2$$

El resultado de este cálculo no siempre es un número entero. Por tanto debe redondearse

donde:

n *es el tamaño de la muestra.*

z *es el valor normal estándar correspondiente al nivel de confianza deseado.*

σ *es la desviación estándar de la población*

E *es el error máximo admisible.*

Ejemplo tamaño de muestra

Un estudiante de administración pública desea determinar la cantidad media que ganan al mes los miembros de los consejos ciudadanos de las grandes ciudades. El error al calcular la media debe ser inferior a \$100, con un nivel de confianza de 95%. El estudiante encontró un informe del Departamento del Trabajo en el que la desviación estándar es de \$1 000.

¿Cuál es el tamaño de la muestra que se requiere?

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{1.96 \times \$1000}{\$100}\right)^2 = 19.6^2 = 384.16$$

Tamaño de la muestra para calcular la proporción de una población

$$n = P(1 - P)\left(\frac{z}{E}\right)^2$$

Donde:

n es el tamaño de la muestra.

z es el valor normal estándar correspondiente al nivel de confianza deseado.

P es la proporción de la población.

E es el máximo error tolerable.

Ejemplo

El país desea hacer un estudio sobre las ciudades con recolectores privados. Se desea que el margen de error se encuentre a 0.10 de la proporción de la población; el nivel de confianza deseado es de 90%; sin embargo, no se encuentra disponible ningún estimador de la proporción de la población.

¿Cuál es el tamaño de la muestra que se requiere?

Solución

El estimador de la proporción de la población se encuentra a 0.10, por lo que E = 0.10. El nivel de confianza deseado es de 0.90, que corresponde a un valor z de 1.65. Como no se encuentra disponible ningún estimador de la población, se utiliza 0.50. El número de observaciones que se sugiere es

$$n = \left(0.5 \times (1 - 0.5)\right)\left(\frac{1.65}{0.10}\right)^2 = 68.0625$$

Intervalos de confianza de la diferencia entre las medias de dos poblaciones normales: muestras dependientes

Las muestras **son dependientes o pareadas** si los valores de una de las muestras influyen en los de la otra. O cuando los mismos individuos u objetos son contrastados dos veces.

El muestreo dependiente también puede referirse a **dos mediciones realizadas sobre la misma persona** u objeto. Por ejemplo una evaluación a priori y luego una a posteriori

¿Qué hacer si se tienen dos muestras aleatorias para estudiar el mismo fenómeno, es decir, con muestras dependientes?

Intervalos de confianza para la diferencia de dos medias

Supongamos que tenemos una muestra aleatoria de ***n pares*** de observaciones enlazadas procedentes de ***distribuciones normales*** de medias μ_x $\boldsymbol{y}$ μ_y. Es decir, sean $x_1, x_2, \ldots, x_n$ los valores de las observaciones μ_x *; e* $y_1, y_2, \ldots, y_n$ de la población que tiene la media los valores correspondientes de la población que tiene la media μ_y. Sean $\boldsymbol{\bar{d}}$ ***y*** $\boldsymbol{s_d}$ ***la media y la desviación típica muestrales*** observadas de las *n* **diferencias** $\boldsymbol{d_i = x_i - y_i}$. Si se supone que la distribución poblacional de las diferencias es normal, entonces se obtiene un intervalo de confianza al $100(1-\alpha)\%$ de la diferencia entre dos medias ***(***$\boldsymbol{\mu_d} = \mu_x$**-** μ_y***)***

Intervalos de confianza para la diferencia de dos medias

Las muestras dependientes, de la forma siguiente:

$$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}} \qquad \bar{d} \pm ME$$

La desviación estándar de las diferencias y el Margen de Error son:

$$s_d = \sqrt{\frac{\sum\left(d_i - \bar{d}\right)^2}{n-1}} \qquad ME = t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

La variable aleatoria $t_{n-1,\alpha/2}$ es el numero para:

$$P\left(t_{n-1} > t_{n-1,\alpha/2}\right) = \frac{\alpha}{2}$$

La variable aleatoria, $t_{n-1,}$, tiene una distribución t de Student con $(n-1)$ grados de libertad.

Ejemplo:

Par	Medicamento X	Medicamento Y	Diferencia $d_i = x_i - y_i$
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	30		
6	32	30	2
7	29	26	3
8	31	33	-2
9	30	36	-6

La tabla muestra el número de puntos en que se ha reducido el nivel de colesterol de cada persona, así como las diferencias, $d_i = x_i - y_i$, correspondientes a cada par.

Estime con un nivel de confianza del 99% la diferencia media de eficacia entre los dos medicamentos, X e Y, para reducir el colesterol

Solución:

En las respuestas a encuestas, los ensayos clínicos y otras investigaciones es frecuente que falten valores por tanto se deben eliminar los registros con datos perdidos.

Calculamos la media muestral y la desviación típica muestral observada de las diferencias :

$$\bar{d} = 1.625 \qquad s_d = 3.777$$

$$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

De la distribución t de Student y obtenemos el intervalo de confianza al 99 %

$$1.625 \pm 3.499 \frac{3.777}{\sqrt{8}}$$

El tamaño de nuestra muestra se reduce de nueve datos pareados a 8

El límite inferior de confianza es un número negativo (- 3,05), mientras que el límite superior de confianza es un número positivo (6,30)

Interpretación:

Como el intervalo de confianza contiene el valor de cero, hay tres posibilidades:

$\boldsymbol{\mu_d} = \mu_x - \mu_y$ podría ser positivo, lo que sugeriría que el medicamento X es más eficaz

$\boldsymbol{\mu_d} = \mu_x - \mu_y$ podría ser negativo, lo que sugeriría que el medicamento Y es más eficaz;

$\boldsymbol{\mu_d} = \mu_x - \mu_y$ podría ser cero, lo que sugeriría que el medicamento X y el Y son igual de eficaces.

Intervalos de confianza de la diferencia entre las medias de dos poblaciones normales: muestras independientes

Calculamos la estimación de intervalos de confianza cuando se extraen dos muestras independientemente de dos poblaciones que siguen una distribución normal.

Examinamos tres situaciones:

1. Ambas varianzas poblacionales son conocidas.
2. Ambas varianzas poblacionales son desconocidas, pero se puede considerar que son iguales.
3. Ambas varianzas poblacionales son desconocidas, pero no se considera que sean iguales.

Dos medias, muestras independientes y varianzas poblacionales conocidas

Aquí se extraen muestras *independientemente* de las dos poblaciones que siguen una distribución normal, por lo que la pertenencia a una de las muestras no influye en la pertenencia a la otra. También conocemos las varianzas poblacionales de las dos poblaciones.

Supongamos que tenemos dos muestras aleatorias independientes de n_x y n_y observaciones procedentes de poblaciones que siguen una distribución normal de medias μ_x y μ_y y varianzas σ_x^2 y σ_y^2 Si las medias muestrales observadas son $\bar{x}$ e $\bar{y}$, entonces obtenemos un intervalo de confianza al $100(1-\alpha)\,\%$ de la diferencia entre dos medias, muestras independientes y varianzas poblacionales conocidas de la forma siguiente:

$$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

En algunas aplicaciones, pueden utilizarse las varianzas históricas de estudios similares como las verdaderas varianzas poblacional

Ejemplo

En una gran universidad, se extrajeron muestras aleatorias independientes de **120 estudiantes de marketing** y de **90 de economía financiera**. Se observó que la calificación **media de la muestra** aleatoria de estudiantes de marketing era de **3,08** y la de la muestra aleatoria de estudiantes de **economía financiera era de 2,88**.

Basándose en estudios **similares anteriores**, se supone que la *desviación típica poblacional* de los estudiantes de marketing **es 0,42** y que la desviación típica poblacional de los estudiantes de **economía financiera es 0,64**.

Representando la media poblacional de los estudiantes de marketing por medio de μ_x y la de los estudiantes de economía financiera por medio de μ_y, halle un intervalo de confianza al 95 % de $(\mu_{\boldsymbol{x}} - \mu_{\boldsymbol{y}})$

Solución

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

$$z_{\alpha/2} = z_{0.025} = 1.96$$

$$n_x = 120$$

$$n_y = 90$$

$$\bar{x} = 3.08$$

$$\bar{y} = 2.88$$

$$\sigma_x = 0.42$$

$$\sigma_y = 0.64$$

$$(3.08 - 2.88) \pm 1.96\sqrt{\frac{0.42^2}{120} + \frac{0.64^2}{90}}$$

$$0.20 \pm 0.1521$$

Dos medias, muestras independientes y varianzas poblacionales desconocidas que se supone que son iguales

Supongamos que tenemos dos muestras aleatorias independientes de n_x y n_x observaciones procedentes de poblaciones que siguen una distribución normal de medias μ_x y μ_y y una varianza poblacional común, pero desconocida. Si las medias muestrales observadas son $\bar{x}$ e $\bar{y}$ y las varianzas muestrales observadas son s_x^2 y s_y^2, entonces se obtiene un intervalo de confianza al $100(1-\alpha)$ % de la diferencia entre dos medias, muestras independientes y varianzas poblacionales desconocidas que se supone que son iguales de la forma siguiente:

$$(\bar{x}-\bar{y}) \pm t_{n_x+n_y-2\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

$$s_p^2=\frac{(n_x-1)s_x^2+\left(n_y-1\right)s_y^2}{n_x+n_y-2}$$

$$P\left(t_{n_x+n_y-2}>t_{n_x+n_y-2\alpha/2}\right)=\frac{\alpha}{2}$$

Ejemplo

Los residentes de St. Paul (Minnesota) se quejan de que las multas de tráfico por exceso de velocidad son más altas en su ciudad que las que se imponen en la vecina Minneapolis. Se obtuvieron muestras aleatorias independientes de las multas pagadas por los residentes de cada una de las dos ciudades durante tres meses. Las cuantías de estas multas eran:

St. Paul	100	125	135	128	140	142	128	137	156	142
Minnesota	95	87	100	75	110	105	85	95		

Suponiendo que las varianzas poblacionales son iguales, halle un intervalo de confianza al 95 % de la diferencia entre los costes medios de las multas de estas dos ciudades.

Solución

Suponiendo de las muestras siguen una distribución normal, se calculan la media y la varianza de cada muestra

$n_x = 10$ $n_y = 8$

$s_x^2 = 218.0111$ $s_y^2 = 129.4286$

$\bar{x} = 133.30$ $\bar{y} = 94$

$$s_p^2 = \frac{(n_x - 1)s_y^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$
$$= \frac{(10-1) \times 218.0111 + (8-1) \times 129.4286}{10+8-2}$$
$$= 179.2563$$

Grados de libertad: $n_x + n_y - 2 = 16$

$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$ $t_{16,0.025=2.12}$

$$39.3 \pm 2.12\sqrt{\frac{179.2563}{10} + \frac{179.2563}{8}}$$

$$\boxed{(\bar{x} - \bar{y}) = (133.30 - 94) = 39.30}$$

Intervalos de confianza de dos medias: varianzas poblacionales desconocidas, no se supone que sean iguales

Supongamos que tenemos dos muestras aleatorias ***independientes*** *de* n_x *y* n_y *observaciones procedentes de poblaciones que siguen una* ***distribución normal*** *de medias* μ_x *y* μ_y *y supongamos que las varianzas poblacionales no son iguales. Si las medias y las varianzas muestrales observadas son* $\bar{x}$ *e* $\bar{y}$ *y* s_x^2 *y* s_y^2*, entonces se obtiene un intervalo de confianza al* $100(1-\alpha)\%$ *de la diferencia entre dos medias, muestras independientes y varianzas poblacionales desconocidas que no se supone que sean iguales de la forma siguiente:*

Intervalos de confianza de dos medias: varianzas poblacionales desconocidas, no se supone que sean iguales

El intervalo de confianza y el margen de error:

$$(\bar{x} - \bar{y}) \pm t_{v,2\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

$$ME = t_{v,2\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

Los grados de libertad:

$$v = \frac{\left[\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2 \Big/ (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 \Big/ (n_y - 1)}$$

Si las muestras son del mismo tamaño. Los grados de libertad:

$$v = \left(1 + \frac{2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}}\right) \times (n - 1)$$

Ejemplo

Una empresa de auditoría tomó una muestra aleatoria de facturas pendientes de pago de las oficinas este y oeste de uno de sus clientes. Quería estimar con estas dos muestras independientes la diferencia entre los valores medios poblacionales de las facturas pendientes de pago. Los estadísticos muestrales obtenidos fueron los siguientes:

Parámetro	**Oficina Este (población X)**	**Oficina Oeste (población Y)**
Media muestral	$290	$250
Tamaño de la muestra	16	11
Desviación típica muestral	15	50

No suponemos que las varianzas poblacionales desconocidas son iguales. Estime la diferencia entre los valores medios de las facturas pendientes de pago de las dos oficinas. Utilice un nivel de confianza del 95 %.

Solución

Primero calculamos los grados de libertad

$$v = \frac{\left[\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2 \Big/ (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 \Big/ (n_y - 1)} = \frac{[225/16 + 2500/11]^2}{\left(\frac{225}{16}\right)^2/15 + \left(\frac{2500}{11}\right)^2/10} \approx 11$$

Ahora hallamos el margen de error

$$ME = t_{v,2\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y} = t_{11,0.025}\sqrt{\frac{225}{16} + \frac{2500}{11}} = 2.201 \times 15.53497 = 34.19}$$

Intervalo de confianza al 95%

$$(290 - 250) \pm 34.19$$

Estimación de intervalos de confianza de la diferencia entre dos proporciones poblacionales (grandes muestras)

Con frecuencia interesa comparar dos proporciones poblacionales.

Sea p_x la proporción observada de éxitos en una muestra aleatoria de n_x observaciones procedentes de una población que tiene una proporción p_x de éxitos y sea $\hat{p}_y$ y la proporción de éxitos observada en una muestra aleatoria independiente de n_y observaciones procedentes de una población que tiene una proporción p_y de éxitos.

En ese caso, si las muestras son de gran tamaño (generalmente al menos 40 observaciones en cada una), se obtiene un intervalo de confianza al $100(1-\alpha\,)\%$ de la diferencia entre proporciones poblacionales (grandes muestras), (p_x - p_y), de la forma siguiente:

$$\left(\hat{p}_x - \hat{p}_y\right) \pm ME$$

$$ME = \ z\alpha/_2\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

Ejemplo

Durante un año de elecciones generales, se realizan muchos pronósticos para averiguar cómo perciben los votantes a un determinado candidato. En una muestra aleatoria de 120 votantes censados del distrito X, 107 declararon que apoyaban al candidato en cuestión. En una muestra aleatoria independiente de 141 votantes censados del distrito Y, solo 73 declararon que apoyaban a ese candidato.

Las proporciones poblacionales respectivas se representan por medio de p_x y p_y.

Halle un intervalo de confianza al 95 % de la diferencia poblacional, (p_x - p_y).

Solución

Dada la información muestral:

n_x=120 $n_y = 141$ $\hat{p}_x = \frac{107}{120} = 0{,}892$ $\hat{p}_y = \frac{73}{141} = 0{,}518$

El intervalo al 95%:

$$(\hat{p}_x - \hat{p}_y) \pm ME = (0{,}92 - 0{,}518) \pm 1{,}96\sqrt{\frac{0{,}829(1-0{,}829)}{120} + \frac{0{,}518(1-0{,}518)}{141}} = 0{,}374 \pm 0{,}099$$

Se deduce que el intervalo al 95 % de la diferencia entre las proporciones poblacionales de votantes encuestados del distrito X y del distrito Y va de 0,274 a 0,473.