Inferencia estadística

CURSO	TEMA	WWW.DANIPARTAL.NET
1ºBach	Estadística 05	Colegio Marista "La Inmaculada" de Granada

INFORMACIÓN GENERAL

Población y muestreo. Tipos de muestreo. Distribución muestral de medias y de proporciones. Estimación puntual y por intervalos. Tamaño de la muestra

POBLACIÓN Y MUESTREO

El conjunto de todos los individuos sobre los que deseamos realizar un estudio estadístico se llama población. Cada miembro de la población es un individuo. El número total de individuos de la población suele representarse como "N".

Un subconjunto de la población es una muestra. El tamaño de la muestra lo determina el número de individuos que forman parte de la muestra. Por norma general, el tamaño de la muestra viene representado por el valor "n".

Un **muestreo aleatorio simple** significa que los n individuos de la muestra se eligen al azar, con reemplazamiento. De esta forma, todos los individuos tienen la misma probabilidad de ser elegidos, independientemente de qué individuo se elija el primero o el último.

Un **muestreo estratificado** ocurre cuando los N individuos de la población se organizan en diversos grupos (estratos) A_1, A_2, A_3,…, A_k. En cada estrato contamos con N_1, N_2, N_3,…, N_k individuos. De tal forma que se cumple:

$$N = N_1 + N_2 + N_3 + \ldots + N_k$$

Si elegimos de la población N una muestra de tamaño n, el número de individuos n_1, n_2, n_3,…, n_k de cada estrato se deben elegir de manera proporcional. Es decir:

$$\frac{N}{n} = \frac{N_1}{n_1} = \frac{N_2}{n_2} = \frac{N_3}{n_3} = \ldots = \frac{N_k}{n_k}$$

DISTRIBUCIÓN MUESTRAL DE MEDIAS

Sea una población de la que sabemos que un estadístico tiene una media μ y una desviación típica σ. Por ejemplo: la altura de todos los españoles, el número de accidentes de tráfico diarios a lo largo de un año o la cantidad de lluvia anual recogida en una estación meteorológica en el último siglo.

Elijamos muestras de tamaño n (por ejemplo, 2.000 españoles sobre los que estudiar su altura). Cada muestra tendrá su correspondiente media y su correspondiente desviación. La media de las medias de todas las muestras se llama media muestral $\mu_{\bar{X}}$. Y la media de las desviaciones de todas las muestras se llama desviación muestral $\sigma_{\bar{X}}$.

¿Qué relación hay entre la media muestral y la media de la población? ¿Y entre la desviación muestral y la desviación de la población?

$$\mu_{\bar{X}} = \mu$$

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

Es decir: **la media muestral coincide con la media de la población. La desviación muestral coincide con la desviación de la población dividida por la raíz cuadrada del tamaño de las muestras.**

Cuando mayor sea el tamaño de la muestra n, menor será el grado de dispersión de la media muestral.

Si el muestreo ha sido aleatorio simple y la población de partida sigue una distribución normal $N(\mu,\sigma)$, o bien el tamaño de la muestra es suficientemente grande ($n \geq 30$), podemos afirmar que la distribución muestral de medias sigue también una distribución normal $N(\mu_{\bar{X}},\sigma_{\bar{X}})$.

Esta conclusión facilita muchos los cálculos en ejemplos como este:

Sea una población formada por los elementos 3, 4, 5 y 8. Se pretende seleccionar una muestra de tamaño 2 con reemplazamiento.

a) Calcule todas las muestras posibles.
b) Calcule la media y la varianza de la población.
c) Calcule la media y la varianza de las medias muestrales.

Conjunto de todas las muestras:

{3,3},{3,4},{3,5},{3,8}

{4,3},{4,4},{4,5},{4,8}

{5,3},{5,4},{5,5},{5,8}

{8,3},{8,4},{8,5},{8,8}

La varianza de la población se obtiene considerando de manera separada los elementos que forman la población. La media de esos cuatro elementos resulta:

$$\mu = \frac{3+4+5+8}{4} = 5$$

Siendo N=4 el tamaño de la población.

La varianza se obtiene con la expresión:

$$s^2 = \frac{n_1(x_1-\overline{x})^2 + n_2(x_2-\overline{x})^2 + \ldots + n_n(n_n-\overline{x})^2}{n_1+n_2+\ldots+n_n} = \frac{\sum_{i=1}^{n} n_i(x_i-\overline{x})^2}{N} = \sum_{i=1}^{n} f_i(x_i-\overline{x})^2$$

Por lo que podemos formar la siguiente tabla para poder obtener la varianza.

x	n	n·(x-μ)²
3	1	4
4	1	1
5	1	0
8	1	9
Acumulado	**4**	**14**

Por lo tanto:

$$\sigma^2 = \frac{14}{4} = 3{,}5$$

Para calcular la media y la varianza muestrales podemos razonar así: **la media muestral coincide con la media de la población. La desviación muestral coincide con la desviación de la población dividida por la raíz cuadrada del tamaño de las muestras.**

$$\mu_{\bar{X}} = 5$$

$$\sigma_{\bar{X}} = \frac{\sqrt{3{,}5}}{\sqrt{2}}$$

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

Imagina una población con variable aleatoria que sigue una distribución binomial y sabemos la proporción "p" (porcentaje) de esa población que cumple con éxito el

experimento aleatorio. ¡Importante! Ahora hablamos de porcentaje de la población que cumple con éxito el experimento aleatorio.

Tomamos muestras de tamaño "n". Cada muestra tendrá su correspondiente porcentaje de éxito.

Se demuestra que la media de los porcentajes de éxito en muestras de tamaño n (distribución muestral de proporciones) sigue una distribución normal con las condiciones siguientes:

- La media de la distribución normal es igual al valor "p" de la proporción de la población:

$$\mu_{\bar{p}} = p$$

- La desviación típica de la distribución normal se aproxima a:

$$\sigma_{\bar{p}} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

Esta aproximación a $N(p, \sqrt{\frac{p\cdot(1-p)}{n}})$ es mejor cuanto mayor sea el tamaño de n y cuanto más cerca esté la proporción "p" del factor 0,5. Nuevamente esta aproximación es válida si:

$n \cdot p \geq 5$

$\mathrm{n} \cdot (1 - \mathrm{p}) \geq 5$

Ejemplo:

El porcentaje de familias españolas con un solo hijo es del 20%.

Tomamos una muestra de 1.000 familias. ¿Cuál es la probabilidad de que al menos el 21% de las familias de la muestra tengan un solo hijo?

Contamos con variable binomial (o tiene un solo hijo, o no tiene un solo hijo). La proporción de la población es p=20% (0,20 en tanto por uno). El tamaño de la muestra es n=1.000.

El porcentaje deseado para la muestra es del 21% (0,21 en tanto por uno).

La proporción muestral podemos aproximarla a la normal:

$$N(p, \sqrt{\frac{p \cdot (1-p)}{n}}) = N(0{,}20;\ 0{,}0126)$$

Necesitamos que esta proporción sea al menos del 21%. Es decir, la variable aleatoria debe ser mayor o igual que 0,21.

$$P(\bar{P} \geq 0{,}21)$$

Tipificamos como de costumbre:

$$P(\bar{P} \geq 0{,}21) = P(Z \geq \frac{0{,}21 - 0{,}20}{0{,}0126}) = P(Z \geq 0{,}79)$$

Recordamos que la tabla tipificada que utilizamos nos da valores acumulados a la izquierda. Por lo tanto:

$$P(Z \geq 0{,}79) = 1 - P(Z \leq 0{,}79) = 0{,}2148$$

Es decir, la probabilidad de que al menos el 21% de las 1.000 familias de la muestra tengan un solo hijo es del 21,48%. Y la aproximación es coherente porque:

$$n \cdot p = 200 \geq 5$$

$$\mathrm{n} \cdot (1 - \mathrm{p}) = 800 \geq 5$$

INTERVALOS DE CONFIANZA

Hasta ahora hemos obtenido información de los parámetros de la media a partir de los datos de los parámetros de la población.

¿Sería imposible invertir el proceso? Es decir, inferir desde la muestra (caso particular) información relevante sobre la población (caso general).

A esto se dedica la inferencia matemática. Por ejemplo: obtener la media de una población a partir de la media y la desviación de una muestra. O bien la proporción de una población a partir de la proporción y la desviación de una muestra.

Si el resultado de nuestro razonamiento nos da un valor concreto, hablaremos de estimación puntual. Y si nos da un intervalo donde encontrar ese valor con un porcentaje concreto de fiabilidad, hablaremos de estimación por intervalo de confianza.

Vamos a centrarnos en esto último, en los intervalos de confianza.

Llamaremos $\boldsymbol{\alpha}$ **al riesgo** de no encontrar el valor deseado dentro de nuestro intervalo de confianza. Por lo tanto, el factor $\boldsymbol{1-\alpha}$ **será el nivel de confianza** de encontrar el valor dentro de nuestro intervalo.

De esta forma, la probabilidad de que el valor que estamos estimando se encuentre dentro de un intervalo (a,b) será:

$$P(a \leq parámetro \leq b) = 1-\alpha$$

Al valor de α también se le denomina nivel de significación.

Si el parámetro a estimar de la población sigue una distribución normal y deseamos aglutinar alrededor de su valor medio el $(1-\alpha)\%$ por ciento de las observaciones, significa que dejaremos a un lado y a otro del intervalo (a,b) un $(\alpha/2)\%$ de observaciones (la suma de $\alpha/2$ más $\alpha/2$ da lugar al valor total α).

Por ejemplo:

Si la distribución del parámetro de la población ya estuviese tipificada a la normal N(0,1) y el nivel de significación fuese del 5%, significaría que $\alpha = 0{,}05$. Y si el intervalo (a,b) fuese simétrico respecto de la media, estaría formado por los valores $(-z_{\alpha/2}, z_{\alpha/2})$. Donde $z_{\alpha/2}$ es el valor de la variable que deja a la derecha un porcentaje de observaciones igual a:

$$\alpha/2 \;=\; 0{,}05/2 \;=\; 0{,}025$$

Como nuestra tabla tipificada nos da los valores acumulados a la izquierda, razonaríamos como de costumbre:

$$P(Z \geq z_{0{,}05/2}) = 0{,}025$$

$$P(Z \geq z_{0{,}05/2}) = 1 - P(Z \leq z_{0{,}05/2})$$

$$0{,}025 = 1 - P(Z \leq z_{0{,}05/2})$$

$$P(Z \leq z_{0{,}05/2}) = 0{,}975$$

$$z_{0{,}05/2} = 1{,}96$$

Es decir, existe un 95% de probabilidad (95=100-5) de que el valor estimado para la población estuviese comprendido en el intervalo (-1,96, 1,96). Repito, en el caso en que la distribución ya estuviese tipificada.

¿Qué casos vamos a estudiar para aplicar la estimación por intervalo?

- Intervalo de confianza para obtener la media población.
- Intervalo de confianza para la proporción poblacional.

Para ello, partiremos de las distribuciones normales estudiadas al inicio de este PDF para la media muestral y la proporción muestral.

Y obtendremos una expresión para obtener el intervalo de confianza para un nivel de significación dado, el error máximo cometido por el intervalo y el tamaño de la muestra que cumpla las condiciones de un intervalo conocido.

Intervalo de confianza para la media poblacional: hemos visto que la distribución muestral de medias $\bar{X}$, sigue una distribución normal $N(\mu_{\bar{X}}, \sigma_{\bar{X}}) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Tipificándola, estudiando la probabilidad y despejando, nos queda el intervalo:

$$\left(\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Intervalo de confianza para la proporción: La distribución muestral de proporciones, se aproxima a una normal: $N\left(p, \sqrt{\frac{p \cdot (1-p)}{n}}\right)$ y razonando como antes, tenemos el intervalo:

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

Tamaño de la muestra: En esos intervalos de confianza, el error máximo que se puede cometer es $E = Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, en la media. Si lo predeterminamos o fijamos, podemos despejar el tamaño de la muestra, tanto en la media como en la proporción:

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right)^2 \qquad n = \hat{p} \cdot (1-\hat{p}) \left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2$$

EJEMPLO RESUELTO 1

Se ha obtenido una muestra de 25 alumnos de una Facultad. Se desea estudiar la calificación media de los expedientes de los alumnos de la facultad. Se sabe, de cursos anteriores, que la desviación típica de toda la Facultad es 2,01.

La media de la muestra fue de 4,9.

Obtener la media de la población en el intervalo de confianza del 90% y al 99%.

Conocemos la desviación de la población:

$$\sigma = 2{,}01$$

El tamaño de la muestra es n = 25.

Conocemos la media de la muestra:

$$\mu_{\bar{X}} = \bar{x} = 4{,}9$$

Si el nivel de significación es:

$$1 - 0{,}90 = 0{,}10$$

El intervalo de confianza de la media de la población resulta:

$$\left(4{,}9 - z_{0{,}05} \cdot \frac{2{,}01}{\sqrt{25}}, 4{,}9 + z_{0{,}05} \cdot \frac{2{,}01}{\sqrt{25}}\right)$$

Sabemos que:

$$P(Z \geq z_{0{,}05}) = 0{,}05$$

$$P(Z \leq z_{0{,}05}) = 1 - 0{,}05 = 0{,}95$$

$$z_{0,05} = 1,65$$

Por lo que el intervalo de confianza para la media de la población será:

$$\left(4,9 - 1,65 \cdot \frac{2,01}{\sqrt{25}}, 4,9 + 1,65 \cdot \frac{2,01}{\sqrt{25}}\right) = (4,24, 5,56)$$

Repetimos el razonamiento para el nuevo nivel de significación:

$$1 - 0,99 = 0,01$$

El intervalo de confianza de la media de la población resulta:

$$\left(4,9 - z_{0,005} \cdot \frac{2,01}{\sqrt{25}}, 4,9 + z_{0,005} \cdot \frac{2,01}{\sqrt{25}}\right)$$

Sabemos que:

$$P(Z \geq z_{0,005}) = 0,005$$

$$P(Z \leq z_{0,005}) = 1 - 0,005 = 0,995$$

$$z_{0,005} = 2,58$$

Por lo que el intervalo de confianza para la media de la población será:

$$\left(4,9 - 2,58 \cdot \frac{2,01}{\sqrt{25}}, 4,9 + 2,58 \cdot \frac{2,01}{\sqrt{25}}\right) = (3,86, 5,94)$$

EJEMPLO RESUELTO 2

El peso de los individuos de una población se distribuye según una ley Normal de desviación típica 6 kg. Calcule el tamaño mínimo de la muestra para estimar, con un nivel de confianza del 95%, el peso medio en la población con un error no superior a 1 kg.

Conocemos la desviación de la población:

$$\sigma = 6$$

El nivel de confianza es del 95%. Por lo que el error será del 5% y el nivel de significación 0,05.

$$\alpha = 0,05$$

El error máximo es 1. La expresión para calcular ese error máximo resulta:

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

El tamaño de la muestra es "n" y es el valor que debemos calcular.

$$1 = z_{0,025} \cdot \frac{6}{\sqrt{n}}$$

Como calculamos en un ejemplo anterior de este PDF:

$$z_{0,025} = 1,96$$

Sustituyendo:

$$1 = 1,96 \cdot \frac{6}{\sqrt{n}}$$

$$\sqrt{n} = 11,76$$

$$n = 138,30$$

Como la variable es discreta, debemos aproximar a un número entero. ¿Elegimos 138 o 139?

Si el tamaño de la muestra disminuye, el error máximo aumenta. Como no queremos superar un el valor 1 en el error, redondeamos al alza. Por lo que el tamaño mínimo de la muestra debe ser 139.