BLOQUE 3 – TEMA 11

ESTIMACIÓN DE PARÁMETROS. ERRORES DE ESTIMACIÓN

Aproximación intutitiva a la inferencia estadística

La Estadística es la ciencia que se ocupa de la ordenación y análisis de datos procedentes de muestras, y de la realización de inferencias acerca de las poblaciones de las que éstas proceden.

POBLACIÓN: conjunto de todos los elementos que cumplen una o varias características o propiedades. Los valores numéricos que describen a la población se llaman **parámetros**. Normalmente, los valores de los parámetros son desconocidos porque resultaría enormemente costoso y complicado obtener datos de todos los sujetos de la población. Por esta razón se trabaja con muestras a partir de las cuales, si procede, se trata de estimar los valores de los parámetros.

MUESTRA: es un subconjunto de los elementos de una población. Los índices numéricos que describen a las muestras se denominan **estadísticos**. La técnica para seleccionar a los sujetos que entrarán a formar parte de la muestra se denomina **muestreo**. Siempre que sea posible debe utilizarse el muestreo **aleatorio** porque nos da mayores garantías de que la muestra sea **representativa de la población**. Otra de las características a tener en cuenta en el muestreo, es que la muestra sea **suficiente en número**, de forma que a mayor tamaño de la muestra, mayores garantías de representatividad, si bien debemos matizar que a partir de cierto tamaño que el investigador debe establecer, su aumento no aporta mejoras sustanciales a la representatividad.

MEDICIÓN: es uno de los procesos más peliagudos a los que nos enfrentamos en educación y Psicología. En muchas ocasiones nos enfrentamos a la medición de **constructos**, es decir, características del ser humano que no son directamente mensurables. Lo que medimos son las manifestaciones observables que atribuimos a esos constructos. Una vez elaborados los instrumentos de medida, se aplican a la muestra, se recogen los datos y se elabora la matriz de datos. A partir de aquí volveríamos a lo dicho sobre ordenación y categorización de datos y análisis descriptivo.

ESTADÍSTICA INFERENCIAL: pretende sacar conclusiones sobre el conjunto de datos a través de observaciones de una parte de esos datos. Mediante la Estadística Inferencial se pueden estimar parámetros y realizar contrastes de hipótesis (ver fig. 11.1 pág. 226).

Dentro del marco de la Estadística Inferencial suelen distinguirse dos tipos de estudios: **la estimación de parámetros** y el **contraste de hipótesis**. *Un parámetro se estima siempre a partir de un estadístico calculado en una muestra.* Se habla de dos tipos de estimación: **estimación puntual** y **estimación por intervalos**. Básicamente, en la estimación puntual se hace coincidir el estadístico con el parámetro; mientras que en la estimación por intervalos se ofrece un intervalo de puntuaciones en el que es más probable que se encuentre el valor del parámetro.

Propiedades de los estimadores

La estimación consiste en la técnica que permite conocer el valor aproximado de un parámetro de un población con una determinada probabilidad a partir de los datos proporcionados por una muestra.

Estas son las características que debe poseer un buen estimador:

Ⓐ ***Carencia de sesgo***: un estimador insesgado es aquel que tiene sesgo igual a cero. La propiedad del insesgamiento nos garantiza que las estimaciones que hagamos con el estimador se encuentran alrededor del parámetro en cuestión, de forma simétrica, es decir, que el promedio de los estadísticos coincide con el verdadero valor del parámetro.

Ⓑ ***Eficiencia***: un estimador es tanto más eficiente cuanto menor es su desviación típica. Si pensamos en una distribución muestral, cuanto más estrecha sea, cuanto menor sea su error típico, más cercanos estarán los estadísticos al valor del parámetro.

Ⓒ ***Consistencia***: un estimador es consistente si a medida que aumenta el tamaño de la muestra, la probabilidad de que el valor del estadístico se acerque al valor del parámetro va siendo mayor. Si el valor de N tiende a ∞, un estimador consistente es, al mismo tiempo, insesgado.

Ⓓ ***Suficiencia***: un estimador es suficiente cuando es capaz de obtener de la muestra toda la información que ésta contenga acerca del parámetro.

Distribución muestral, error muestral y error típico: estimación del parámetro.

La distribución muestral se define como la distribución de un estadístico en el muestreo. Está formada por los ∞ valores del estadístico obtenidos en ∞ muestras aleatorias del mismo tamaño extraídas de la misma población.

Si nos remitimos a la distribución muestral de la μ (media poblacional), está demostrado según el teorema central del límite y para muestras grandes (N > 30), que la distribución se asemeja a una distribución normal. Si el parámetro μ fuese conocido, comprobaríamos como la mayoría de las $\bar{x}$ de las muestras se encontrarían cerca del valor del parámetro, pero precisamente por ser muestras aleatorias, algunas se alejan un poco y otras, muy pocas, se alejan mucho de ese valor.

Por tanto, confiamos en que la muestra elegida sea una de las que el valor de su $\bar{x}$ esté cercano al valor de μ, pero no lo podemos asegurar. Por esta razón, en inferencia siempre hablamos de **nivel de confianza** (*porcentaje de confianza al hacer la estimación*; también puede darse en probabilidad y se denomina 1-α) y del **nivel de significación** α (*probabilidad de error que estamos dispuestos a asumir en la estimación*). Obviamente, se trata de conceptos complementarios que se refieren a lo mismo.

Para hallar el **intervalo confidencial**, es decir, *los valores entre los cuales es más probable que se encuentre el verdadero valor del parámetro*, necesitaremos calcular la $\bar{x}$, a la que sumaremos y restaremos el **error muestral**; es decir, la diferencia más probable entre el estadístico y el parámetro.

$$IC = \bar{x} \pm EM$$

El error muestral nos da una idea de la precisión de nuestra inferencia estadística. Cuanto más grande sea el error muestral, menor será la precisión en la estimación y menor será la utilidad de la estimación.

Una distribución muestral tiene variabilidad, por tanto tendrá su propia **desviación típica** σ, que recibe el nombre de **error típico**. Es un medida de dispersión con respecto al parámetro, es decir, nos indica la dispersión de las $\bar{x}$ de las ∞ muestras aleatorias extraídas respecto a la μ.

Cómo podemos conocer este dato?

Nuevamente lo estimaremos y nos basaremos en los datos de la muestra, en este caso Sx. Para la distribución muestral de medias, la fórmula es:

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{N-1}}$$

pondremos N-1 en el denominador cuando se haya calculado en la muestra la S_x^2 insesgada, o solo N cuando se haya calculado la S_x^2 sesgada (cuasivarianza o varianza, respectivamente).

El error típico no solo depende de la Sx, sino en gran medida del tamaño de la muestra. Cuanto más grande es N, más pequeño el cociente, más pequeño el error típico y más precisión en la estimación del parámetro.

Cómo se calcula el error muestral?

Previamente hemos tenido que definir el **nivel de significación** α con el que vamos a realizar la estimación, **calcular la puntuación** z correspondiente a ese nivel (porque la distribución es normal) y aplicamos la fórmula:

$$EM = z\alpha_{/2} . \sigma_{\bar{x}}$$

ver fig. 11.2 pág. 230. **Un ejemplo**: sean N = 1.000; $\bar{x}$ = 105 y Sx = 10. Estimar el valor del parámetro μ (intervalo confidencial) con un nivel de confianza del 99%.

IC = $\bar{x} \pm$ EM

nivel de confianza → 1-α = 99% → 0,99

nivel de significación → α = 0,01

Por tanto, α/2 = 0,005. El área bajo la curva normal de valor 0,005, corresponde a una z = - 2,575 (columna C)

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{N-1}} = \frac{10}{\sqrt{999}} = 0{,}316$$

$$IC = 105 \pm z_{\alpha/2} . \sigma_{\bar{x}} = 105 \pm (-2{,}575)\ 0{,}316 \begin{cases} 104{,}18 \\ 105{,}81 \end{cases}$$

podemos afirmar con un nivel de confianza del 99%, que el valor de μ se encuentra en el intervalo siguiente: (104,18 ≤ μ ≤ 105,81)

Estimación de μ para pequeñas muestras

Cuando N ≤ 30, la distribución muestral de la media sigue la distribución **t** de Student. Esta distribución varía en función del número de sujetos, es decir, de sus grados de libertad, aunque se trata también de una distribución simétrica y asintótica. **Cuando N tiende a ∞, la distribución t tiende a la distribución z.**

Solo nos cambia en este caso el estadístico para calcular el error muestral. En lugar de trabajar con $z_{\alpha/2}$, lo haremos con $t_{\alpha/2}$ (ver tablas en pág. 347).

Ejemplo: los mismos datos del ejemplo anterior pero ahora con N = 25.

IC = 105 ± EM

$\alpha/2$ = 0,005. Se distribuye según **t** con N-1 grados de libertad. Buscando en tablas, corresponde a un valor de $t_{\alpha/2}$ = 2,797

$\sigma_{\bar{X}} = \frac{10}{\sqrt{24}}$ = 2.04

IC = 105 ± $t_{\alpha/2}$. $\sigma_{\bar{x}}$ = 105 ± (2,797) 2,04 → 99,29 / 110,70

como puede verse, hemos perdido una gran cantidad de precisión al disminuir el tamaño de la muestra, puesto que el intervalo de confianza es ahora mucho mayor.

Estimación del parámetro proporción (π)

Es un caso particular del anterior, en el que la media oscila entre 0 y 1. En las variables dicotómicas ya vimos que **p** corresponde a la proporción de unos y **q** a la proporción de ceros (q = 1-p). En este caso, la σ_p viene dada por la fórmula:

$$\sigma_p = \frac{p.q}{\sqrt{N-1}}$$

El intervalo de confianza se establece igual que en el caso de la μ, pero ahora partiendo de una proporción. Se emplea el valor de $z_{\alpha/2}$ al nivel de confianza correspondiente.

Estimación de la puntuación verdadera en una prueba

Queremos realizar la estimación de la puntuación verdadera de un sujeto en un instrumento, o dicho de otra forma, *entre qué puntuaciones es más probable que se encuentre la verdadera puntuación*, ya que toda medida tiene algún margen de error.

De nuevo se trata de hallar el **intervalo de confianza** en el que es probable que se encuentre la verdadera puntuación del sujeto en la prueba. Sabiendo que la distribución muestral es normal, **necesitamos conocer el error típico de medida:**

$$\sigma_e = S_t.\sqrt{1-r_{xx}}$$

$S_t \rightarrow$ desviación típica total en el instrumento de medida

$r_{xx} \rightarrow$ coeficiente de fiabilidad

IC = X_i ± EM

$X_i \rightarrow$ puntuación obtenida en la prueba

Intervalo de confianza de la puntuación estimada en la regresión lineal simple.

Queremos estimar la puntuación de un sujeto en una variable denominada criterio (Y) a partir de las puntuaciones conocidas en otra variable denominada predictora (X). Sabiendo que la distribución muestral de Y´ es la normal, debemos conocer el **error típico de estimaciones**, que para muestras grandes es:

$$\sigma_{est} = S_y.\sqrt{1-r_{xy}^2}$$

IC = Y´ ± EM

Estimación del parámetro correlación de Pearson.

Introducción al concepto de significatividad estadística.

En muchas ocasiones, cuando calculamos una correlación, en realidad estamos más interesados en la relación que existe entre esas dos variables en la población, que en la muestra. **Se trata de calcular el**

intervalo de confianza para la correlación de Pearson, de modo que podamos estimar entre qué valores se encuentra dicho coeficiente entre la población.

Sabemos que la distribución muestral de la correlación de Pearson se asemeja a la distribución normal con el siguiente **error típico**:

* para muestras grandes (N > 100) $\rightarrow$ $\sigma_{r_{xx}} = \frac{1}{\sqrt{N-1}}$

* y para muestras pequeñas ----------$\rightarrow$ $\sigma_{r_{xx}} = \frac{1-r_{xx}^2}{\sqrt{N-1}}$

y como ya es habitual $IC = r_{xx} \pm EM$; donde $EM = Z_{\alpha/2} . \sigma_{r_{xx}}$

Ejemplo: Sea N = 20 y r_{xx} = 0,35. Cómo será la correlación en la población con un nivel de confianza del 95%?.

$\sigma_{r_{xx}} = \frac{1-(0,35)^2}{\sqrt{20-1}} = \frac{0,8775}{4,358} = 0,201$

α = 0,05 $\rightarrow \alpha/2$ = 0,025. En tablas, un área bajo la curva de 0,025 (columna C) corresponde a una z = -1,96

IC = 0,35 $\pm$ (-1,96).0,201 $\begin{cases} -0,04 \\ 0,74 \end{cases}$

El intervalo tan amplio es debido al pequeño tamaño de la muestra. Cuanto más pequeña sea la muestra, más imprecisa será la estimación.

<u>Significación o significatividad estadística</u>

Hablamos de significación estadística cuando nos referimos al significado de la diferencia entre dos medidas. Decimos también que **un coeficiente de correlación es estadísticamente significativo cuando es distinto de cero**, es decir, cuando el coeficiente de correlación (*<u>estadístico</u>*) es lo suficientemente grande como para decir que la correlación en la población (*<u>parámetro</u>*) de referencia es distinta de cero. Si queremos conocer cuál es su magnitud en la población, entonces calcularemos el intervalo de confianza.

Cuando estimamos **el intervalo de confianza** en el cual es probable que se encuentre la verdadera correlación en la población, **y este intervalo contiene el valor cero** (ausencia absoluta de correlación) diremos que dicha correlación NO es estadísticamente significativa.

<u>Estimación del parámetro diferencia de medias</u> ($\mu_1 - \mu_2$)

Si una universidad encuentra que el CI medio de sus estudiantes es de 110, y otra universidad lo tiene de 105, cómo es la diferencia entre estas dos puntuaciones en la población de referencia?. La diferencia entre estas dos puntuaciones es estadísticamente igual a cero?. Si dicha diferencia es compatible con una diferencia igual a cero, concluimos que ambas muestras tienen el mismo CI, al nivel de confianza fijado, y que la diferencia encontrada es aleatoria.

Si establecemos **el intervalo de confianza** a partir del estadístico diferencia de medias, obtendremos los límites confidenciales entre los cuales es más probable que se encuentre la diferencia de medias en la población.

Si este intervalo incluye la puntuación cero, entonces dicha diferencia es compatible con una diferencia de medias igual a cero, y en consecuencia, podremos interpretar que dicha diferencia es estadísticamente igual a cero.

$IC = (\bar{x}_1 - \bar{x}_2) \pm EM$ $\qquad$ $EM = Z_{\alpha/2} . \sigma_{(\bar{x}_1 - \bar{x}_2)}$

$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{S_1^2}{N_1 - 1} + \frac{S_2^2}{N_2 - 1}}$ para muestras grandes e independientes (N > 100)

$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$ para muestras pequeñas e independientes (N ≤ 100)

Ejemplo:

Universidad 1	Universidad 2
$\bar{x}$ = 110	$\bar{x}$ =105
S_1 = 10	S_2 = 12
N_1 = 90	N_2 = 120

Calculamos primero el error típico de la diferencia de medias.

$\sigma_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{90.100+120.144}{90+120-2}\left(\frac{1}{90}+\frac{1}{120}\right)} = \sqrt{\frac{183960}{74880}} = \sqrt{2,4567} = 1,567$

La $z_{\alpha/2}$ al 95% de nivel de confianza, ya sabemos que corresponde a una z = - 1,96

EM = (- 1,96) 1,567 = - 3,071 $\rightarrow$ por tanto:

IC = (110 – 105) $\pm$ (- 3,071) $\begin{cases} 1,929 \\ 8,071 \end{cases}$

Al ser el intervalo incompatible con el valor cero, concluimos que la diferencia marcada es estadísticamente significativa, es decir, que las diferencias de medias en CI entre ambos grupos no son aleatorias. (Ver fig. 11.4 pág. 238)

Estimación del parámetro diferencia de proporciones ($\pi_1 - \pi_2$)

Lo único que varía en este caso es el error típico. Al igual que en el contraste de medias, lo más interesante suele ser la diferencia entre dos proporciones.

El **error típico de la diferencia de proporciones** es:

$$\sigma_{(p_1-p_2)} = \sqrt{pq.\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

y la estimación del intervalo de confianza y su interpretación será la misma que en el caso anterior, pero con medias en términos de proporción, que frecuentemente se multiplican por 100 para convertirlos en porcentajes.

Estimación de parámetros y contraste de hipótesis

Todo se reduce al contraste de una hipótesis estadística, denominada **hipótesis nula (Ho)**, según la cual NO existen diferencias estadísticamente significativas. Se afirma en ella que no hay efecto presente de la VI sobre la VD.

Ver fig. 11.5 pág. 240

Se establece una zona central alrededor de la media en la curva normal, en la que es más probable que las diferencias encontradas entre las medias de las muestras se deban efectivamente a los efectos del azar. Es la **zona de aceptación** (ó de no rechazo) **de Ho**. Por tanto, al no poder rechazar la Ho, decimos que no es falsa.

Queda de forma inmediata **otra zona** (bilateral o unilateralmente) en la que las diferencias entre las medias resulta muy improbable (al nivel de confianza establecido) que sean aleatorias, por lo que dichas diferencias se atribuyen a los efectos de la VI.

La Ho se expresa así:

Hipótesis nula $\rightarrow$ Ho : $\mu 1 - \mu 2 = 0$

las diferencias entre las $\bar{x}$ de los grupos o muestras son estadísticamente iguales a cero;

o bien, que las diferencias empíricas que existen entre las $\bar{x}$ de las muestras se deben al azar;

o bien, que los valores paramétricos son iguales;

o bien, que ambas muestras pertenecen a la misma población.

Las fórmulas para hallar **el intervalo de confianza** son las mismas que el epígrafe anterior, para muestras grandes y pequeñas.

Si resulta que nuestra diferencia empírica de medias se encuentra dentro del intervalo de confianza de la distribución muestral conforme a Ho, diremos que nuestra diferencia de medias es compatible con una diferencia de medias igual a cero, y por tanto, que se trata de una diferencia estadísticamente no significativa o igual a cero.

Lo que haremos después será **calcular un estadístico** (**t**, z, etc.) que nos dirá cuántas desviaciones típicas (errores típicos) se aleja nuestra diferencia de medias de una diferencia de medias igual a cero.