ALGUNAS CUESTIONES DESTACABLES EN INFERENCIA ESTADÍSTICA

Las encuestas sociológicas suelen trabajar con muestras. Sería demasiado costoso entrevistar al total de la población española adulta, que suele ser muchas veces el marco poblacional sobre el que los sociólogos realizan sus estudios.

Así observamos cómo de la información obtenida a partir de una muestra de unos pocos individuos podemos producir información sobre el conjunto de la población española.

Cuando aquí hablamos de 'información' estamos hablando de cómo se distribuye una variable en un determinado conjunto de individuos o, lo que es lo mismo, cómo se distribuye una población entre un conjunto de grupos (generados por los valores de una variable).

Y para ver cómo se distribuye una variable en un conjunto de individuos teníamos unas medidas de resumen. Esto se veía en el tema 5.

Si la variable era *nominal*, el dato que resumía la composición de una población era la *proporción* del valor de la variable que nos interesaba (esto es, el peso relativo de un determinado grupo en el conjunto de la población).

Si la variable era numérica o *de intervalo*, el valor que resumía al conjunto de la población era el *valor medio*.

Al número de individuos de la muestra lo denominamos aquí «**n**».
Al número de individuos de la población total lo denominamos «**N**».

Tenemos que tener en cuenta que en temas anteriores la «n» y la «N» tenían otro significado.

Si recordamos el tema 4, al número de individuos de cada uno de los grupos que constituían una variable lo llamábamos «**n**».

Así, si el grupo (o valor) 1 de una variable en una determinada población contaba con 39 individuos, escribíamos que $\mathbf{n_1} = 39$.

Si miráis la tabla 4.2 (en la que observábamos cómo se distribuía la «edad» en la población juvenil o, por decirlo de otra forma, cómo se distribuía la población muestral de la juventud española por edades) [página 185], cada una de las edades contaba con unos determinados casos.

Como «edad» es una variable la llamamos «x». Y está compuesto de 15 grupos (o valores) (k=15) [páginas 226-227].

Si x_1=15 años, n_1=238; si x_2=16, n_2=267, …etc. El conjunto de la población muestral lo llamábamos «**N**». De tal manera que $N = n_1 + n_2 + n_3 + \ldots + n_{15}$.

$N = 238 + 267 + 284 + \ldots + 441 = 5.014$.

¿Cuál es la razón para que hayamos designado de igual manera a cosas en cierto modo distintas?

La única razón que nos ha llevado a ello es que aunque en la mayoría de los manuales de estadística no llaman «**n**» al número de individuos de un grupo, sino que a este lo llaman «**f**» (de **f**recuencias de un determinado valor en una población), a nosotros nos pareció que era más fácil hablar de «**n**» para hablar del tamaño de un grupo (del número de individuos que componen un grupo de una determinada población). Sin embargo, en estadística inferencial, aquella que trata de deducir un valor en una población a partir de lo que se ha obtenido en una muestra, al tamaño de esta muestra (al número de individuos que la componen) se le suele llamar «**n**».

Para distinguir una «n» (el número de individuos de un grupo de una población) de otra «n» (el número de individuos que componen una muestra), bastará con saber si nos situamos en el campo de la estadística descriptiva (primer caso) o en el de la estadística inferencial (segundo caso).

En estadística descriptiva:

«**n**» es el número de individuos de un grupo de una población (de un grupo, se entiende, dentro de los grupos que generan los valores de una variable).

«**N**» es el número total de los individuos de una población y se obtiene sumando los de los distintos grupos que componen una población.

[Es lo que vemos en las páginas 226-227, cuando se habla del valor medio de una población.]

En estadística inferencial:

«**n**» es el número de individuos de una muestra de una población (el tamaño de la muestra utilizada para realizar inferencias).

«**N**» es el número total de los individuos de la población de la que extraemos la muestra.

[Es lo que vemos en las páginas 300 y siguientes, cuando se habla de poblaciones y muestras.]

Aquí, en esta asignatura, sólo vamos a tratar de inferir proporciones (si trabajamos con variables nominales) y valores medios (si trabajamos con variables de intervalo).

A la proporción o al valor medio, en su caso, obtenido en la muestra le vamos a llamar "estadístico". A la proporción, o valor medio, en su caso, que corresponde a la población total (de la que se extrae la muestra) le llamamos "parámetro".

Estimación de proporciones

En el caso de que queramos inferir la proporción (es decir, el peso relativo) de un determinado grupo (generado por un valor de una variable) en una población a partir de la proporción que hemos obtenido para ese grupo en una muestra, hemos de saber que la proporción en la población (que vamos a escribir **P**, con mayúscula) se moverá entre unos determinados valores alrededor de la proporción obtenida en la muestra (que escribimos con **p** minúscula. Estos valores vendrán determinados por un error, que se llama ***error muestral*** (que escribimos **e**). De tal manera que:

$$\mathbf{P = p \pm e}$$

Esto quiere decir que la proporción en la población (**P**) estará alrededor de la proporción obtenida en la muestra (**p**) comprendida entre **p - e** y **p + e**.

$$\mathbf{p - e \leq P \leq p + e}.$$

Esos 'valores' (no hablamos aquí de valores de una variable) marcan los límites del ***intervalo de estimación***.

Sólo podemos estimar la proporción de un valor (aquí sí hablamos de valor de una variable) de manera aproximada: sólo podemos decir que se encontrará en un intervalo.

Y además sólo podemos realizar una estimación aproximada en términos probabilísticos, nunca con absoluta certeza. Pero, si el muestreo es probabilístico, siempre podremos conocer esta probabilidad: la probabilidad de que el parámetro poblacional se encuentre en el intervalo de estimación. Esta probabilidad la conocemos como ***nivel de confianza***. A cada nivel de confianza le corresponde un determinado valor Z. Si trabajamos con un nivel de confianza del 95% (esto es si queremos tener una probabilidad de acertar del 95% al realizar la estimación), Z será igual a 1,96.

El error muestral (**e**) depende del nivel de confianza, de la dispersión de la variable en donde se inscribe la proporción del valor que intentamos estimar y del tamaño de la muestra.

El error muestral es directamente proporcional al nivel de confianza: a mayor nivel de confianza mayor error. Está claro que si queremos incrementar nuestra seguridad de acertar (en términos probabilísticos) (es decir, si queremos reducir nuestra probabilidad de equivocarnos al realizar la estimación) el intervalo de estimación tendremos que hacerlo mayor (y esto sólo se consigue incrementando el error muestral). Mientras más grande sea este intervalo más probabilidades tendremos de que se encuentre en él el parámetro poblacional (la proporción a estimar).

El error también es directamente proporcional a la dispersión de la variable en donde se sitúa el valor cuya proporción queremos estimar. Mientras más dispersión presente esta variable mayor será el error, es decir, más grande tendremos que hacer el intervalo de estimación para poder encontrar el parámetro poblacional en él.

Pero el error es inversamente proporcional al tamaño de la muestra. Esto quiere decir que una muestra pequeña presenta un error grande. Y que con una muestra grande tendremos un error pequeño.

La fórmula del error, que hay que saber de memoria, en este caso, en que queremos estimar la proporción de un valor, es:

$$e = \frac{Z \cdot \sqrt{p \cdot q}}{\sqrt{n}}$$

No necesitamos conocer de dónde sale esta fórmula. Nos basta con saber lo que ya hemos apuntado: el error es directamente proporcional al nivel de confianza (que se expresa en Z), a la dispersión o variabilidad de la población y es inversamente proporcional al tamaño de la muestra. Todo en los términos en los que se señala en la fórmula anterior, que podemos simplificar así (esta sería la que habría que aprender):

$$\mathbf{e = Z \cdot \sqrt{\frac{p \cdot q}{n}}}$$

Aquí, hay que recordar que la varianza de una distribución nominal (dicotomizada: binomial) [esto es del tema 5] se expresaba en el producto de «p · q».

Si recordáis, la variable nominal se reducía a dos únicos valores, el que nos interesaba y el resto. El primer valor representaba en el conjunto de la población una proporción «p». El valor restante representaba una proporción «1 - p», ya que los dos valores contenían el total de la población (y este total en términos de proporción siempre es igual a 1). A la proporción del valor restante la llamábamos «q», sabiendo que q = 1 - p.

Por ejemplo, si en una población tenemos la variable “estado civil” y nos interesa en esta fijarnos en los solteros, reducimos la variable a dos valores: ‘solteros’ y ‘no solteros’ (donde agrupamos los estados civiles restantes).

Si la proporción de 'solteros', que llamamos «p», es de 0,432 (es decir, tenemos un 43,2% de solteros), la proporción de no solteros, que llamamos «q», será igual a «1 - p»: 1 – 0.432= 0.568 (es decir, tenemos un 56,8% de no solteros).

La dispersión del estado civil (según esta división de él) se mide por «p · q», que es un producto que nos da la variabilidad de la población según la división dicotómica establecida y que es equivalente a la "varianza". Lo que utilizamos en la fórmula del error es la raíz cuadrada de la varianza, por lo tanto,« $\sqrt{p \cdot q}$».

De la fórmula que utilizamos para calcular el error muestral 'deriva' la fórmula del **tamaño muestral**.

$$e = \frac{Z \cdot \sqrt{p \cdot q}}{\sqrt{n}}$$

$$e^2 = \frac{Z^2 \cdot p \cdot q}{n}$$

$$n \cdot e^2 = Z^2 \cdot p \cdot q$$

$$\boxed{\mathbf{n = \frac{Z^2 \cdot p \cdot q}{e^2}}}$$

<u>En estimación de proporciones, el error se expresa siempre en términos de proporción</u>.

Hay que tener en cuenta que muchas veces lo que estimamos son porcentajes, por lo que tendríamos que convertir los porcentajes en proporciones para aplicar la fórmula del error. Y si el error está en porcentaje tendremos que convertirlo (en la fórmula) en proporción. [Recordad el tema 4.]

Estimación de medias

También, decíamos más arriba, que si estamos trabajando con una variable cuantitativa, de carácter numérico, como puede ser la "edad", los "ingresos", el "número de hijos", el "número de miembros de un hogar", el "tamaño del municipio de residencia" o la "distancia entre el domicilio y el trabajo", por ejemplo, en estos casos, el estadístico que nos interesa, el dato que resume cada una de estas características en la población, es el valor medio.

Así, si queremos estimar cuál sería el valor medio de una población (parámetro) a partir del valor medio obtenido en una muestra de dicha población (estadístico), contaríamos, de igual modo que con las proporciones, con un ***intervalo de estimación***.

El valor medio de la población (al parámetro a estimar) lo llamamos con la letra griega 'mu' (**μ**), para distinguirlo del valor medio obtenido en una muestra, que lo llamamos 'equis con una rayita encima' ($\overline{x}$).

El valor medio en la población de la variable que nos interesa se encontrará alrededor del valor medio obtenido en la muestra:

$$\mu = \overline{x} \pm e$$

Esto quiere decir que el valor medio en la población (**μ**) estará alrededor del valor medio obtenido en la muestra ($\overline{x}$), comprendido entre los valores $\overline{x} - e$ y $\overline{x} + e$.

$$\overline{x} - e \leq \mu \leq \overline{x} + e.$$

Esos valores marcan los límites del ***intervalo de estimación***.

El error muestral (**e**) depende del nivel de confianza, de la dispersión de la población en torno al valor medio y del tamaño de la muestra.

El error será directamente proporcional al nivel de confianza y a la dispersión de los datos en la población. E inversamente proporcional al tamaño de la muestra.

En esta ocación, cuando pretendemos estimar un valor medio, la dispersión de los datos de la población la medimos por la "varianza" (s^2). La varianza es, si recordáis, el cuadrado de la desviación típica (**s**). [También la varianza en una variable cualquiera ("X") la podemos anotar como s_x^2; y la desviación típica como s_x.]

Y el error muestral en este caso sería:

$$\boxed{e = \frac{Z \cdot s}{\sqrt{n}}}$$

Y la fórmula del tamaño de la muestra quedaría:

$$e = \frac{Z \cdot s}{\sqrt{n}}$$

$$e \cdot \sqrt{n} = Z \cdot s$$

$$\sqrt{n} = \frac{Z \cdot s}{e}$$

$$\boxed{\mathbf{n = \frac{Z^2 \cdot s^2}{e^2}}}$$

En estimaciones de valores medios, el error se mide en las unidades de la variable. Si, por ejemplo, estimamos una "edad" media, el error se dará en 'años'; si estimamos un "número (medio) de hijos", el error se dará en 'hijos'; y si estimamos un "salario" medio, el error se dará en euros.