Inferencia Estadística. Conceptos Generales

Inferencia Estadística

De forma muy sintética podemos decir que la Inferencia Estadística es la Teoría matemática que proporciona los métodos para obtener conocimiento probable acerca de la distribución de una variable sobre un conjunto de objetos (*Población)*, a partir de cierto número de observaciones de esa variable (*Muestra)*.

De los conceptos que intervienen en la anterior expresión de Inferencia Estadística, pueden darse las siguientes definiciones:

Población

Es cualquier conjunto de individuos o elementos que son el objeto de nuestro estudio. O dicho de otra forma, es el conjunto de elementos sobre los cuales queremos ampliar nuestro conocimiento.

Una aclaración que conviene hacer es la distinción entre población y universo. La estadística en sus procedimientos realiza una serie de simplificaciones y síntesis en orden a hacernos comprensible una realidad compleja y variable, una de estas simplificaciones es la de delimitar uno o varios caracteres sobre los individuos u objetos que se quieren estudiar. La medición de esos caracteres nos proporciona para cada individuo-objeto un valor, el conjunto de esos valores constituye la distribución de la variable que es el elemento sobre el cual podemos adquirir conocimiento.

Universo es el conjunto de elementos observables que se estudian. Población es el conjunto de valores numéricos que se obtienen al medir una característica sobre los individuos del universo. De esta forma vemos que un mismo universo puede dar origen a distintas poblaciones si se miden características diferentes. En lo sucesivo, hablaremos de los individuos de la población o de los individuos de la muestra, pero debe entenderse que nos referimos a los valores que arroja la medición de la variable correspondiente sobre un individuo.

Muestra

Es el subconjunto de la población que se observa en orden a ampliar nuestro conocimiento acerca de la población. Debemos recalcar la idea de que una muestra no se extrae para estudiar a los sujetos particulares de esa muestra sino en tanto en cuanto nos pueden proporcionar información acerca de la totalidad de la población.

Las anteriores definiciones pueden proyectar la idea de que la estadística inferencial parte de una población existente y selecciona una porción de la misma. Esto no es estrictamente cierto, se parte de la realización de un experimento aleatorio un cierto número de veces y los resultados de estas repeticiones, o una variable aleatoria función de los sucesos ocurridos, constituyen la muestra. El número de veces que se repite el experimento aleatorio y por consiguiente el número de resultados se denomina tamaño de la muestra.

La repetición indefinida del experimento aleatorio daría lugar a un conjunto que constituye la población potencial. Esta población puede coincidir con una población realmente existente, en cuyo caso la imagen anterior de la muestra como una porción extraída de una población tendría sentido, pero otras veces la población a que hacemos referencia es una población hipotética que no tiene existencia y a la cual no podemos acceder.

Por ejemplo si seleccionamos a un grupo de alumnos de la asignatura de Análisis de Datos y mediante el test adecuado determinamos su C.I. podemos ver estos datos como una muestra de la población constituida por los C.I. de todos los alumnos y plantearnos si la distribución del C.I. la debemos estudiar a partir de esa muestra o si sería mas conveniente estudiarla directamente a partir de los C.I. de todos los alumnos matriculados. Por el contrario, si a un grupo de recién nacidos los incorporamos a un programa de estimulación precoz y posteriormente determinamos su C.I. los valores que obtengamos los podemos considerar como una muestra de una población hipotética que estaría constituida por los valores del C.I. que obtendríamos si todos los niños fuesen sometidos a ese programa de estimulación precoz, obviamente esa población no existe como tal en la realidad y tendremos que utilizar siempre procedimientos de inferencia estadística para estudiarla.

Distribuciones Poblacional y Muestral. Función de Verosimilitud.

Distribución Poblacional

Si consideramos la población como el conjunto de valores asociados a la repetición indefinida del experimento aleatorio en cuestión, tendremos que la variable aleatoria ligada a los sucesos que ocurren tomará todos los valores posibles y que la frecuencia relativa de estos valores irá tendiendo a un límite que constituye la definición de probabilidad de esos valores. Tendremos así definidos los valores de la variable junto con sus probabilidades de ocurrencia lo cual constituye la distribución de la variable aleatoria en la población. Sus correspondientes Funciones de Distribución y Densidad $F(x ; \theta)$ y $f(x ; \theta)$ serán la Función de Distribución y la Función de Densidad de la Población.

Es obvio que esta distribución permanecerá desconocida para nosotros, al menos en parte, ya que si la conociésemos en su totalidad carecería de sentido el plantearnos la obtención de muestras o cualquier otro estudio. A lo más, basándonos en las características del experimento aleatorio de que se trate podremos suponer el modelo teórico que seguirá la variable aleatoria pero permaneciendo desconocidos los parámetros de la distribución. Precisamente esos parámetros de la distribución poblacional centrarán gran parte del interés de la inferencia estadística.

Distribución de frecuencias de una muestra

Si realizamos n veces un experimento aleatorio y anotamos los resultados de la variable aleatoria que estemos considerando, tendremos n valores que constituyen una muestra de tamaño n. El conjunto de los diferentes valores que se presentan, junto con las frecuencias de esos valores constituyen la distribución de frecuencias de la muestra o distribución empírica.

Habitualmente, la distribución empírica de la muestra se expresa en términos de frecuencias relativas para poder compararla más fácilmente con la distribución de la población que viene dada en términos de probabilidad. Si la variable fuese continua, la distribución muestral se determina considerando las frecuencias relativas de intervalos de valores.

Por ejemplo, si consideramos la variable aleatoria "número de ensayos que necesita una rata para completar una tarea de laberinto" y determinamos esta variable en una muestra de 50 ratas obteniendo los siguientes valores:

7, 6, 2, 5, 6, 3, 6, 5, 2, 4, 6, 3, 5, 1, 3, 5, 6, 4, 3, 2, 6, 3, 7, 4, 1, 7, 5, 6, 4, 6, 4, 4, 7, 4, 6, 4, 2, 5, 7, 5, 2, 5, 4, 5, 6, 5, 3, 6, 2, 3.

Tendremos la distribución:

x_i	1	2	3	4	5	6	7	total
n_i	2	6	7	9	10	11	5	50
f_i	0,04	0,12	0,14	0,18	0,2	0,22	0,1	1

Distribución de la muestra

Si previamente a la realización de los experimentos nos planteamos cuales son los posibles valores que pueden aparecer, tendremos que cada componente de la muestra constituye una variable aleatoria que puede tomar distintos valores con diferentes probabilidades. De esta forma una muestra genérica de tamaño n constituye una variable aleatoria n dimensional, es decir una variable con n componentes X_1, X_2, ... X_n donde cada componente es una variable aleatoria, la distribución de esta variable aleatoria n-dimensional constituye la distribución de la muestra y sus correspondientes Funciones de Distribución y de Densidad que notaremos:

$$F_n(x_1, x_2, \ldots x_n / \theta) \quad f_n(x_1, x_2, \ldots x_n / \theta)$$

serán la Función de Distribución y la Función de Densidad de la muestra.

Función de Verosimilitud

No es casual que hayamos designado a los parámetros de la distribución de la población y de la muestra con la misma letra θ. En general, la distribución de la muestra dependerá de los valores de los parámetros de la población . Dicho de una forma simplista, será más probable encontrar estaturas de 1,95, 1,97, 2,02 en una muestra de jugadores de baloncesto, cuya distribución de estaturas tiene una media de 1,90 que en una muestra de la población general cuya estatura media es de 1,70. Es precisamente este hecho de que la distribución de la muestra dependa de los parámetros de la distribución de la población, lo que nos permite obtener información acerca de la distribución de la población a partir de los datos de la muestra.

Dicho lo anterior, si consideramos fijo el valor de θ, la función de densidad de la muestra $f_n(x_1, x_2, \ldots x_n / \theta)$ nos proporciona la densidad de probabilidad de las distintas muestras. Ahora bien, si consideramos fijos los valores $x_1, x_2, \ldots x_n$ y variable

θ tendremos que la anterior función nos indica como varía la densidad de probabilidad de una muestra determinada en función del valor del parámetro de la población de que proceda.

Concretando, la *Función de verosimilitud es la función de densidad de la muestra considerada como función del parámetro.*

Métodos de Muestreo. Muestreo aleatorio simple.

Cómo ya hemos dicho una muestra no se extrae para aprender algo en concreto de los sujetos de la muestra. Sino en tanto en cuanto me proporciona datos generalizables a toda la población. Para ello la muestra que utilicemos, ha de ser representativa de la población que queremos estudiar. Una muestra será representativa cuando su distribución de frecuencias sea semejante a la distribución de probabilidad de la población, constituyendo una especie de representación a escala reducida de la misma. Dicho de otro modo, cuando la muestra presente los mismos valores de la población y aproximadamente en la misma proporción.

Un procedimiento de muestreo es un mecanismo por medio del cual se determina los individuos u observaciones que van a formar parte de la muestra. Nos interesará buscar procedimientos que nos aseguren que las muestras obtenidas a través de ellos sean representativas. Nosotros en los sucesivo y por razones de sencillez en el desarrollo teórico nos limitaremos a considerar el muestreo aleatorio simple.

Muestreo aleatorio

Es aquél procedimiento en que los individuos de la población que pertenecen a la muestra han sido determinados mediante un mecanismo aleatorio.

Muestreo Probabilístico

Es aquél muestreo aleatorio que nos permite calcular la probabilidad de que aparezca una muestra concreta.

Muestreo aleatorio simple

Es aquél muestreo probabilístico en el que la función de densidad de muestra es igual al producto n veces de la función de densidad de población, es decir:

$$f_n(x_1, x_2, \ldots x_n / \theta) \quad = \quad \prod_{i=1}^{n} f(x_i ; \theta)$$

Una muestra en la que se verifica la igualdad anterior se dice que es una muestra aleatoria simple. Una forma de garantizar que un procedimiento de muestreo sea aleatorio simple es que todos los elementos de la población tengan la misma probabilidad de ser seleccionados en cada extracción y que cada extracción sea independiente de las restantes.

Ventajas de los procedimientos muestrales

Es obvio que hay situaciones en que las investigaciones basadas en muestras son las únicas posibles, este es el caso cuando el procedimiento de determinación de la variable es destructivo o el proceso de medida deteriora seriamente el objeto. También tendremos que recurrir a los estudios muestrales cuando la población sea infinita o tan numerosa que no exista posibilidad de investigarla exhaustivamente. Pero, incluso en aquellos casos en que la totalidad de la población podría ser investigada y determinado en todos los individuos el valor de la variable bajo estudio, puede ser conveniente recurrir al estudio de una muestra en virtud de las siguientes ventajas que poseen los procedimientos de muestreo:

Economía

Es evidente que los procedimientos muestrales al tener que operar con menos individuos son menos costosos que los estudios exhaustivos, esto hace que investigaciones cuyo presupuesto sería prohibitivo en el caso de estudiar toda la población, puedan ser llevadas a cabo. Aún en el caso de contar con las asignaciones económicas suficientes para un estudio de este tipo, debemos de plantearnos su conveniencia, pues la naturaleza de los procedimientos muestrales es tal que a partir de un determinado tamaño de muestra las ganancias en precisión que se consiguen aumentado éste son prácticamente despreciables. Un criterio de eficiencia nos conducirá a considerar el tamaño óptimo de la muestra y a emplear el resto de los fondos en otros estudios o en perfeccionar aspectos diferentes del estudio.

Calidad

Es bastante fácil de comprender que el estudiar una población a partir de una parte de ella, por muy perfeccionado que sea el procedimiento de selección, conlleva un error o por lo menos el riesgo de cometerlo, este tipo de error se denomina error de muestreo, pero también es cierto que este de riesgo de error puede ser controlado y mantenido dentro de unos límites que además pueden ser fijados por el investigador.

Por otra parte, en cualquier estudio, sea muestral o exhaustivo, existen errores ajenos al muestreo. Estos errores proceden de errores materiales en las distintas fases del proceso de datos como cumplimentación, codificación, grabación, etc. También de deficiencias en las instrucciones impartidas a los entrevistadores y de la carga subjetiva que estos y sus supervisores inevitablemente introducen en el estudio. De deficiencias en la información previa referente a los límites de la población que se estudia y en cuanto a los individuos que la componen. Se ha comprobado que estos errores ajenos al muestreo, suelen llegar a ser más importantes que los errores de muestreo y que además crecen con el número de individuos a investigar, por otra parte este tipo de errores no puede ser controlado ni mantenerse acotado. Como consecuencia de lo anterior, puede producirse la aparente paradoja de que un estudio basado en una muestra de una población sea más preciso que uno realizado sobre la totalidad de los individuos de la misma.