Inferencia

Estadística

(Teoría y problemas)

I. Espejo Miranda
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
A. M. Rodríguez Chía
A. Sánchez Navas
C. Valero Franco

I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C. Valero Franco

Edita: Servicio de Publicaciones de la Universidad de Cádiz
c/ Doctor Marañón, 3. 11002 Cádiz (España)
www.uca.es/publicaciones

ISBN: 978-84-9828-131-6

Inferencia Estadística (Revisión: Marzo 2007)
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C. Valero Franco

http://www.uca.es/teloydisren

Capítulo 1

La Inferencia Estadística

1. Introducción

En un sentido amplio, se entiende por Inferencia a la parte de la Estadística que estudia grandes colectivos a partir de una pequeña parte de éstos. El conjunto de individuos que se pretende analizar se denomina *población*, mientras que la parte que sirve de apoyo para realizar dicho análisis se llama *muestra*. Técnicamente la Inferencia consiste en, una vez estudiada la muestra, proyectar las conclusiones obtenidas al conjunto de la población. Por motivos obvios, la calidad del estudio que se realice depende, por una parte, de la calidad de la muestra y, por otra, del uso que de ella se haga. La primera de las cuestiones se resuelve a través de la *Teoría de Muestras*, mientras que en la segunda se utilizan las herramientas suministradas por la Estadística Descriptiva y el Cálculo de Probabilidades.

A continuación se dan unas pinceladas que ayudan a comprender algunos de los aspectos de la Teoría de Muestras. Su análisis en profundidad escapa a nuestros objetivos, y resulta fuera de lugar debido a su gran extensión y complejidad.

1. Ante todo, una muestra debe ser suficientemente representativa de la población de la cual ha sido extraída, tratando de reflejar lo

mejor posible las particularidades de ésta. Las partes de la citada población que no estén debidamente representadas en la muestra llevan a la aparición de sesgos o errores sistemáticos que viciarán el proceso de la Inferencia desde el origen. Para alcanzar buenos niveles de representatividad existen distintos tipos de muestreo que, de forma suscinta, se repasan posteriormente.

2. La segunda de las condiciones que se pide a una muestra, es que a través de ella se alcancen unos objetivos de precisión fijados de antemano. Esta condición tiene que ver con el hecho de que, al no hacerse un estudio exhaustivo, existen márgenes de error en el cálculo de las características de la población, en la determinación de la estructura probabilística de ésta, etc. Mayores niveles de precisión exigirán una mayor información sobre la población, es decir, un mayor *tamaño muestral* (número de elementos que componen la muestra).

2. Clasificación de los procedimientos inferenciales

En primer lugar, se ha de hacer notar que la población va a venir representada por una variable aleatoria con una determinada distribución de probabilidad. Dependiendo del grado de conocimiento de ésta se distinguen dos métodos para realizar el proceso inferencial:

1. *Inferencia paramétrica.* Es aquella en la que se admite que la distribución de la población pertenece a una cierta familia paramétrica de distribuciones, siendo necesario únicamente precisar el valor de los parámetros para determinar la distribución poblacional.

2. *Inferencia no paramétrica.* No supone ninguna distribución de probabilidad de la población, exigiendo sólo hipótesis muy generales, como puede ser la de simetría. A su vez los procedimientos no paramétricos se pueden clasificar en:

   a) Procedimientos de localización, que estudian los parámetros de localización de la distribución.

b) Procedimientos de estructura, que analizan las condiciones que se dan en la distribución de la variable.

c) Procedimientos sobre las condiciones de la muestra, que comprueban si se verifican las hipótesis exigibles a los valores muestrales, como la independencia, ausencia de valores atípicos, etc.

Por su parte, la inferencia paramétrica puede ser estudiada desde dos enfoques diferentes:

1. *Enfoque clásico*. En el cual los parámetros de la distribución de probabilidad de la población se consideran constantes.
2. *Enfoque bayesiano*. Considera a los parámetros como variables aleatorias, permitiendo introducir información sobre ellos a través de la distribución a priori.

3. Naturaleza de la información extraída de la población

La introducción se ha centrado en lo que se conoce como Teoría de Muestras. Sin embargo, con el objeto de obtener una visión más global del proceso inferencial, se distinguen dos procedimientos para la obtención de información. En el primero, dicha información se obtiene de forma aséptica, con el sólo propósito de observar las unidades muestrales, y en el segundo, se establecen las condiciones en las cuales se procederá a la medición de lo que se conoce como unidades experimentales. Formalmente, dicha distinción implica dos categorías, la primera de ellas, como ya ha quedado de manifiesto, supone el encontrarse dentro de la Teoría de Muestras, mientras que la segunda se conoce como *Diseño de Experimentos*.

La Teoría de Muestras, en primer lugar necesita establecer los protocolos que se deben respetar para alcanzar los niveles de representatividad y precisión prefijados; a ésto se le llama *diseño muestral*, que conduce a una *muestra potencial*. Una vez realizado dicho diseño, se procede a la obtención de una o varias *muestras* mediante la observación,

la medición o la encuestación. Estas alternativas están directamente relacionadas con la naturaleza de los datos: atributos, variables continuas, discretas o de clase, ordenadas o no.

El Diseño de Experimentos, por su parte, fue creado por Fisher en la década de 1920 y en sus orígenes tuvo una clara aplicación al mundo agrícola, relacionando las condiciones en las que se realizaban los cultivos, que constituyen los denominados *factores*, con la producción obtenida, *variable dependiente*. El campo de aplicación se ha ido extendiendo con el paso de los años, teniendo en la actualidad una aplicación generalizada en la mayoría de los campos científicos. En cualquier caso, en lo que sigue no se considera como objeto de análisis, con lo cual los estudios que a continuación se llevan a cabo se restringen a la Teoría de Muestras.

Antes de continuar, es necesario aclarar algunas cuestiones de vital importancia para entender el desarrollo teórico que aquí se presenta.

- Cuando se plantea realizar un estudio inferencial se debe realizar un diseño muestral. Esto implica que cada elemento de la muestra potencial es una variable aleatoria unidimensional, mientras que la muestra es un vector aleatorio de dimensión el tamaño de ésta. Además, no debe confundirse el individuo físico con la característica o características que se desean estudiar de éste.
- Cuando de cada individuo se estudia una única característica se habla de análisis univariable o univariante, cuando se estudian dos, bivariable o bivariante y cuando se consideran más de dos, multivariable o multivariante. En lo que sigue, se considerará un análisis univariable.
- Una muestra de tamaño n será denotada por $(X_1, \ldots, X_n)$ o bien, por $\underline{X}$. Cada X_i, con $i = 1, \ldots, n$, es una variable aleatoria que representa la característica bajo estudio del elemento i-ésimo de la muestra. Cuando las mediciones se hayan llevado a cabo, es decir, una vez realizado el muestreo los resultados obtenidos se denotan por $(x_1, \ldots, x_n)$ o bien, por $\underline{x}$.

4. Razones que justifican un estudio inferencial

La realización de un estudio inferencial se justifica por distintas circunstancias, entre las que destacan las siguientes:

- Por motivos presupuestarios. La realización de un estudio a través de muestras supone un ahorro tanto de dinero como de tiempo. Imagínese el tiempo y dinero que supondría realizar un estudio sobre la altura media de la población de Andalucía.
- A veces no todos los elementos de una población están localizables. En el ejemplo anterior puede ocurrir que haya personas nacidas en Andalucía que vivan en otras comunidades.
- En ocasiones la población tiene un gran número de elementos, pudiendo ser ésta potencialmente infinita. Considérense, por ejemplo, poblaciones cuyos elementos se obtienen a partir de la realización de un experimento aleatorio, como la tirada de un dado o la contabilización del número de clientes que utilizan un cierto servicio en un tiempo fijo.
- Existen situaciones en las que cuando se analiza un elemento éste queda inutilizable o destruido. Si se quiere comprobar la calidad del vino de una cierta cosecha, un análisis completo llevaría a la desaparición de la población. Bastaría tomar una medición en cada tonel o conjunto de éstos.
- Por motivos de precisión. Aunque parezca contradictorio, a veces un análisis total, implica el que se cometan errores graves en la medición, codificación, resumen, etc., cuestiones que pueden ser mucho mejor controladas utilizando un estudio a partir de una muestra. Por otro lado, es mucho más fácil formar y controlar a un pequeño número de medidores–observadores–encuestadores, que a un gran número de éstos.

5. Tipos de muestreo

A continuación se establece una primera clasificación de los tipos de muestreo que es comúnmente aceptada en la Estadística:

1. Muestreo probabilístico o aleatorio. Es aquel en el que a priori se conoce la probabilidad de que cada uno de los elementos de la población pertenezca a la muestra.

2. Muestreo opinático. Es aquel en el que el muestrador decide subjetivamente los individuos que compondrán la muestra.

3. Muestreo sin norma. Es aquel en el que se toma como muestra un trozo de la población por razones, en general, de comodidad.

La ventaja del muestreo aleatorio es que pueden determinarse los errores que se cometerán en el proceso inferencial, siendo el único que interesa desde el punto de vista estadístico. El muestreo opinático se justifica en función del conocimiento que se tenga de la población bajo estudio. Finalmente, el muestreo sin norma puede utilizarse como una primera aproximación a una población de la que no se dispone de información alguna.

A continuación, el estudio se centra en el muestreo probabilístico. Se distinguen, de forma poco exhaustiva, las siguientes variedades:

1. Muestreo aleatorio simple con reemplazamiento. Es aquel en el que todas las unidades poblacionales tienen la misma probabilidad de pertenecer a la muestra, pudiendo medirse varias veces el mismo individuo. Las variables aleatorias que componen una muestra obtenida a través de este procedimiento son independientes e idénticamente distribuidas.

   Ejemplo 1.1 *En una urna se tienen* 100 *bolas:* 60 *bolas rojas,* 25 *bolas blancas y* 15 *bolas amarillas. Se extraen de la misma (con reemplazamiento) dos de ellas. Para averiguar cuál es la probabilidad de que la primera bola sea blanca y la segunda roja, se definen los sucesos*

   $$B_1 = \{\textit{sacar la primera bola blanca}\}$$
   $$R_2 = \{\textit{sacar la segunda bola roja}\}.$$

   Puesto que hay reemplazamiento, sacar bola

blanca y sacar bola roja son sucesos independientes, con lo cual,

$$P(B_1 \cap R_2) = P(B_1)P(R_2) = \frac{25}{100}\frac{60}{100}.$$

2. Muestreo aleatorio simple sin reemplazamiento. Igual que en el caso anterior todos los individuos tienen idéntica probabilidad de pertenecer a la muestra, pero los individuos no pueden seleccionarse varias veces. En este caso, las variables aleatorias que componen la muestra no son independientes.

   Ejemplo 1.2 *En el caso del ejemplo anterior, si se extraen de nuevo dos bolas de la urna pero esta vez sin reemplazamiento, la probabilidad de extraer primero una bola blanca y luego una roja es*

   $$P(B_1 \cap R_2) = P(B_1)P(R_2/B_1) = \frac{25}{100}\frac{60}{99}.$$

3. Muestreo estratificado. Este tipo de muestreo se basa en la especificación de subpoblaciones o *estratos* conteniendo elementos parecidos entre sí. La composición de la muestra se distribuye entre los distintos estratos mediante un procedimiento que recibe el nombre de afijación. Existen principalmente tres tipos de afijación:

   a) Uniforme. En la muestra habrá el mismo número de representantes de cada estrato. Es decir, si existen k estratos y el tamaño de la muestra es n, se extraerán, aproximadamente $\frac{n}{k}$ elementos de cada estrato.

   Ejemplo 1.3 *En una empresa hay seis categorías diferentes de trabajadores, cada una con un número similar de empleados y con varianzas parecidas para la variable salario. Si se quiere tomar una muestra de* 60 *individuos para estudiar el salario medio de los trabajadores, habría que tomar de cada categoría* $\frac{60}{6} = 10$ *trabajadores.*

   b) Proporcional. En la muestra habrá un número de representantes de cada estrato proporcional a su tamaño. Es decir, si

un estrato, i, contiene N_i elementos de los N de la población, le corresponderá un total de $\frac{N_i}{N}n$ elementos muestrales.

Ejemplo 1.4 *Para realizar un estudio sobre una característica de una población de* 1000 *habitantes, donde* 600 *son hombres y* 400 *mujeres, suponiendo que la varianza de dicha característica sea similar para ambos sexos, se debería tomar la muestra de manera que se mantuviera esa proporción, es decir, que el* 60 % *de la muestra fuesen hombres y el* 40 % *fuesen mujeres.*

c) Óptima. La asignación de unidades muestrales se hace teniendo en cuenta tanto el tamaño de los estratos como su variabilidad, de forma que, un estrato más heterogéneo necesita de más unidades muestrales, mientras que uno más homogéneo se explica con un menor número relativo de elementos de la muestra. Así, si σ_i representa la desviación típica del estrato i-ésimo, la asignación de unidades muestrales para dicho estrato vendrá dada por

$$n_i = n\frac{\sigma_i N_i}{\sum_{j=1}^{k} \sigma_j N_j}.$$

Ejemplo 1.5 *Se quiere realizar un estudio sobre el tiempo dedicado a la lectura a la semana en una población de* 1000 *habitantes. La siguiente tabla refleja los porcentajes y desviaciones típicas de los grupos en los que se divide la población.*

Grupo	*Edades*	f_i	σ_i
1	< 18	$0'25$	$0'1$
2	$19 - 35$	$0'40$	$0'3$
3	$36 - 55$	$0'20$	$0'5$
4	> 55	$0'15$	$0'1$

Se decide tomar una muestra de 600 habitantes, de manera que de cada grupo,

dado que $\sum_{i=1}^{4} \sigma_i N_i = 260$, *habrá que tomar:*

$$\begin{aligned} n_1 &= 600\frac{250\cdot 0'1}{260} = 60 \\ n_2 &= 600\frac{400\cdot 0'3}{260} = 276 \\ n_3 &= 600\frac{200\cdot 0'5}{260} = 228 \\ n_4 &= 600\frac{150\cdot 0'1}{260} = 36. \end{aligned}$$

Como propiedad a destacar, hay que señalar que el muestreo estratificado permite un estudio diferenciado para cada estrato.

4. Muestreo por áreas o conglomerados. En este caso se trata de establecer grupos de elementos físicamente próximos entre ellos, frecuentemente constituidos por una partición geográfica de la población.

   Se puede observar que las ideas que subyacen en el muestreo estratificado y por conglomerados son opuestas, ya que los elementos de la población que pertenecen al mismo estrato son homogéneos entre sí y heterogéneos con el resto de los estratos, sin embargo, los conglomerados son homogéneos entre ellos y heterogéneos internamente.

   La característica principal del muestreo por áreas es que permite limitar la toma de muestras a un conjunto de áreas que representen al resto.

   Ejemplo 1.6 *Se quiere realizar un estudio sobre cuánto gastan las familias españolas al año. Para simplificar el problema que supone obtener las listas de toda la población, se eligen aleatoriamente algunas provincias como representantes del conjunto de ellas, de las cuales se obtendrá la muestra deseada.*

Además de los anteriores que tienen un uso generalizado, existen una gran cantidad de procedimientos para muestrear que pretenden adaptarse de la mejor manera a las circunstancias de la población bajo estudio. A modo ilustrativo destacan los muestreos: *sistemático*, en el que selecciona una de cada k unidades ordenadas; *polietápico*, que es una generalización del de áreas; *bifásico*, en el que se parte de una muestra

grande que permita reconocer las características más acentuadas de la población, al objeto de poder definir un diseño más fino, etc.

Tanto en el caso de muestreo estratificado como en el de áreas y en cualquier otro muestreo probabilístico, la última etapa del muestreo implica la realización de un muestreo aleatorio simple; ello justifica el hecho de que en lo que sigue sólo se consideren muestras aleatorias simples, (m.a.s.). En concreto, el tipo de muestras aleatorias simples que van a ser analizadas a partir de ahora son aquellas obtenidas de una población infinita o de un muestreo aleatorio simple con reemplazamiento. Esto supone que dada una muestra $(X_1, \ldots, X_n)$ se tiene que:

- Las variables X_i, con $i = 1, \ldots, n$, tienen igual distribución de probabilidad que la población de la cual se ha extraído la muestra. Es decir, si F es la función de distribución de la población entonces

$$F_{X_i} = F, \quad i = 1, \ldots, n.$$

- Las variables X_i, con $i = 1, \ldots, n$, son independientes. Por tanto, si $F_{X_1,\ldots,X_n}$ es la función de distribución conjunta de la muestra, entonces

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) = \prod_{i=1}^{n} F(x_i).$$