Chapter 2

Introducción a la Inferencia Estadística

2.1 Poblaciones finitas

Frecuentemente en los problemas de las diferentes disciplinas se estudia el comportamiento de varias variables definidas sobre un conjunto de objetos. El conjunto de objetos será denominado *población* y será representado por $\mathcal{P} = \{a_1, a_2, \ldots, a_n\}$; $a_1, a_2, \ldots, a_n$ serán denominados los *elementos* de la población $\mathcal{P}$. Sobre esos elementos se observan variables, indicadas $X_1, X_2, \ldots, X_k$, que son características que cambian de individuo a individuo. Luego para cada elemento a en $\mathcal{P}$, estará definido $X_1(a), X_2(a), \ldots, X_k(a)$.

Ejemplo 1: Consideremos una población $\mathcal{P}$ formada por un conjunto de 1000 parcelas que constituyen una explotación agrícola y donde se cultiva solamente trigo. Sea $X(a)$ la cosecha en la parcela a durante un determinado año medida en kilogramos.

Ejemplo 2: Consideremos el conjunto $\mathcal{P}$ de votantes en una determinada elección donde se presentan 3 candidatos, que denominamos 1, 2 y 3. Definimos $X(a)$ como el número del candidato votado por a.

Ejemplo 3: Supongamos que la población $\mathcal{P}$ consiste de todos los pájaros de una especie determinada que habitan en una región determinada. Para

cada pájaro se define $X(a)$ como el largo del pájaro y $Y(a)$ el área de las alas.

Distribución de una variable en la población. Llamaremos *distribución de una variable X en la población $\mathcal{P}$* a la distribución que se obtiene cuando se elige al azar un elemento de la población, es decir, cuando se le da a todo elemento de la población la misma probabilidad. Luego se tiene

$$F_X(x) = \frac{\#\{a \in \mathcal{P}, \quad X(a) \leq x\}}{\#\mathcal{P}}$$

donde $\#A$ indica el número de elementos de A. Del mismo modo se define distribución conjunta de dos o más variables en la población $\mathcal{P}$. Luego si X e Y son variables definidas sobre la población $\mathcal{P}$ será

$$F_{XY}(x, y) = \frac{\#\{a \in \mathcal{P} \ : \ X(a) \leq x, \ Y(a) \leq y\}}{\#\mathcal{P}}$$

Obsérvese que la distribución de una variable definida en una población finita es necesariamente discreta, ya que la variable correspondiente toma sólo un número finito de valores.

2.2 Poblaciones infinitas

En muchos problemas interesa la distribución de una variable aleatoria X (o de varias variables $X_1, X_2, \ldots, X_k$) que se observan cada vez que se repite un mismo experimento perfectamente definido. En estos casos, cada elemento a estudiar corresponde al resultado de un experimento, pero no existe un conjunto finito fijo de experimentos definido de antemano, ya que al menos teóricamente se puede repetir el experimento tantas veces como se quiera. Se puede pensar entonces en una *población infinita* compuesta por los infinitos posibles experimentos que teóricamente se pueden realizar, aunque tal población no tiene existencia real.

Ejemplo 1: El experimento consiste en tirar una moneda y X vale 0 ó 1 según caiga ceca o cara.

Ejemplo 2: El experimento consiste en repartir 10 cartas elegidas al azar de un mazo de 52. X es el número de corazones, e Y el número de sietes.

Ejemplo 3: El experimento consiste en fabricar y probar una lámpara; X es el tiempo de duración de la misma.

Ejemplo 4: Se desea medir una magnitud física, cuyo valor verdadero μ es desconocido. Cada medición está afectada de un error aleatorio. Luego lo que se observa al hacer una medición es una variable $X = \mu + \varepsilon$, donde ε es el error. La medición se puede repetir tantas veces como se quiera.

Lo que hace que una población sea infinita es que el experimento pueda repetirse infinitas veces y no el número de posibles resultados que puede ser finito como puede verse en los ejemplos 1 y 2.

Distribución de una variable en una población infinita. En el caso de *población infinita* se puede suponer que cada vez que se repite el experimento se observa una variable aleatoria X (o varias variables $X_1, X_2, \ldots, X_k$) con una cierta distribución $F(x)$ (o distribución conjunta $F(x_1, x_2, \ldots, x_k)$), y que a diferentes experimentos corresponden variables aleatorias independientes. De acuerdo a la ley de los grandes números, $F(x)$ puede verse como el límite en casi todo punto de la distribución empírica asociada a n repeticiones independientes del experimento. Es decir, si se realiza una sucesión de experimentos y los valores observados son $x_1, x_2, \ldots, x_n, \ldots$, entonces si $F_n(x) = \#\{x_i : x_i \leq x,\ 1 \leq i \leq n\} \ / \ n$ se tendrá $F_n(x) \to F(x)$ en c.t.p. La distribución $F(x)$ será denominada *distribución de la variable X en la población infinita considerada.*

2.3 Modelos para la distribución de una variable en una población

Tanto en el caso de poblaciones finitas como en el de poblaciones infinitas, la distribución F puede ser muy complicada e irregular. Sin embargo, frecuentemente puede ser aproximada por una distribución de forma relativamente sencilla. Consideremos el ejemplo 1 de 2.1. Como la población es finita, la distribución real de X es discreta. Sin embargo, como el número de parcelas es muy grande, 1000, y como es muy probable que los valores $X(a_i)$ sean todos diferentes (pueden diferir muy poco, pero es muy difícil que haya 2 exactamente iguales), resulta que la probabilidad de cada uno de los valores es muy pequeña (1/1000). Por lo tanto, se puede pensar que la distribución real puede aproximarse por una distribución continua de forma

sencilla, por ejemplo una distribución normal. Esto sugiere la introducción del concepto de *modelo.*

Llamaremos *modelo de la distribución de una variable en una población* a un conjunto de hipótesis que se suponen válidas para la distribución de una variable en una población. Más formalmente, supongamos que la variable tiene distribución F perteneciente a una familia $\mathcal{F}$. Al fijar el modelo, se establecen hipótesis sobre la familia $\mathcal{F}$ que, en general, se cumplirán en forma aproximada. La *bondad* de un modelo para describir la distribución de una población estará dada por el grado de aproximación que tengan las hipótesis del modelo con la distribución real.

Por lo tanto, de acuerdo a lo que dijimos anteriormente, se podría usar un modelo continuo para la distribución de variables en poblaciones finitas.

Clasificaremos los modelos en *paramétricos* y *no paramétricos.*

Modelos paramétricos: Consisten en suponer que la distribución $F(x)$ de la variable en la población pertenece a una familia de distribuciones que depende de un número finito de parámetros reales. Así, ejemplos de modelos paramétricos son los siguientes:

(a) $F(x)$ pertenece a la familia $N(\mu, \sigma^2)$,

(b) $F(x)$ pertenece a la familia $B_i(\theta, n)$,

(c) $F(x)$ pertenece a la familia $P(\lambda)$,

(d) $F(x)$ pertenece a la familia $\varepsilon(\lambda)$,

(e) Si $F(x, y)$ es la distribución de dos variables, un modelo puede ser $F(x, y)$ pertenece a la familia $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$,

(f) Si $F(x_1, x_2, \ldots, x_k)$ es la distribución de k variables un modelo puede ser $F(x_1, \ldots, x_k)$ pertenece a la familia $M(\theta_1, \theta_2, \ldots, \theta_k, n)$.

En general, un modelo paramétrico tendrá la siguiente forma. Si $F(x)$ es la distribución de una variable X, entonces $F(x)$ pertenece a la familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_k) \quad \boldsymbol{\theta} \in \Theta\}$, donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ es el vector de parámetros que toma valores en un conjunto $\Theta \subset R^k$. Esto significa que existe algún valor $\boldsymbol{\theta} \in \Theta$, digamos $\boldsymbol{\theta}_0$ tal que $F(x, \boldsymbol{\theta}_0)$ coincide con la distribución $F(x)$ (aunque en la realidad no coincidirá, sino que resultará parecida).

Ejemplo 1: Para el ejemplo 1 de 2.1, podemos usar el modelo definido por la familia de distribuciones $N(\mu, \sigma^2)$.

Ejemplo 2: Para el ejemplo 2 de 2.1 , podemos usar el modelo $M(\theta_1, \theta_2, \theta_3, 1)$. En este caso, el modelo será exacto con

$$\theta_i = \frac{\#\{a \in P; \quad X(a) = i\}}{\#P}, \quad i = 1, 2, 3.$$

Ejemplo 3: Para el ejemplo 3 de 2.1, podemos usar para la distribución $F(x, y)$ el modelo $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Ejemplo 4: Para el ejemplo 3 de 2.2 podemos usar el modelo $\varepsilon(\lambda)$.

Ejemplo 5: Para el ejemplo 4 de 2.2 se puede usar el modelo $N(\mu, \sigma^2)$.

Modelos no paramétricos: En los modelos no paramétricos se supone que la distribución $F(x)$ de la variable (o de las variables si hay más de una) en la población, pertenece a una familia $\mathcal{F}$, pero esta familia no puede ser indicada con un número finito de parámetros reales.

Ejemplo 6: Consideremos nuevamente el ejemplo 4 de 2.2. Un modelo no paramétrico razonable sería el siguiente. Sea μ el valor verdadero que se quiere medir, luego la distribución de X (el valor observado en una medición pertenece a la familia $\mathcal{F}$ de todas las distribuciones tales que:

(i) Son continuas con densidad $f(x)$,

(ii) $f(\mu + x) = f(\mu - x)$ es decir son simétricas alrededor del verdadero valor μ, por lo tanto la "probabilidad" de un error positivo es la misma que de uno de igual valor absoluto pero negativo.

(iii) Si $\mu > x > x'$, entonces $f(x') < f(x) < f(\mu)$. Es decir, a medida que se alejan del verdadero valor los posibles resultados tiene menor "probabilidad".

Esta familia de distribuciones $\mathcal{F}$ descripta por (i), (ii) y (iii) no puede ser indicada por un número finito de parámetros.

Ventajas relativas de los modelos paramétricos y no paramétricos

La ventaja fundamental de los modelos paramétricos, consiste en que la distribución que se elige para representar a la distribución de la variable en la población puede ser descripta por un número finito de parámetros. Esto permite inclusive la posibilidad de tabulación. Por ejemplo en el caso de la familia $N(\mu, \sigma^2)$ basta tabular la distribución $N(0, 1)$. Para obtener otra distribución de la familia basta con realizar una transformación lineal. En el caso de la familia $P(\lambda)$ basta tabularla para algunos valores de λ. Por ejemplo, para valores de λ escalonados de 0.1 en 0.1. Para otros valores de λ, la distribución se puede obtener por interpolación.

Además, como la descripción del modelo tiene una formulación analítica relativamente simple, su tratamiento matemático es más sencillo y las conclusiones a las que se pueden arribar más fuertes.

Los modelos no paramétricos carecen de estas ventajas, pero en recompensa tienen mucha mayor flexibilidad. Esto se debe a que la familia de posibles distribuciones para la población es más numerosa y por lo tanto mayor es la posibilidad que haya en esta familia una distribución muy próxima a la real.

Por ejemplo, en el caso del ejemplo 6 de 2.3 μ ya no representa el valor esperado de la variable X, que podría no existir. Por lo tanto, su valor aproximado no podría conocerse promediando los valores observados como en el caso paramétrico, en el que se supone, por ejemplo, que X tiene distribución $N(\mu, \sigma^2)$.

Elección del modelo: La elección del modelo puede ser hecha en base a consideraciones teóricas, o porque la experiencia indica que ajusta bien. Por ejemplo, si F es la distribución del tiempo de espera hasta que un determinado mecanismo falle, y por consideraciones teóricas podemos suponer que el mecanismo tiene "falta de desgaste", podemos suponer como modelo para F la familia exponencial $\varepsilon(\lambda)$. En otros problemas puede suceder que no se pueda elegir el modelo en base a consideraciones teóricas, pero si la experiencia indica a través de estudios anteriores, por ejemplo, que puede ser bien aproximada por una distribución normal, entonces se usaría como modelo la familia $N(\mu, \sigma^2)$.

Veremos en el transcurso del curso, métodos para poner a prueba el modelo elegido, es decir métodos para determinar si el modelo elegido puede describir dentro de una aproximación aceptable la distribución de la variable (o variables) en la población. Esto se hará en el capítulo 6.

2.4 Muestra de una distribución. Inferencia estadística

Supongamos que hemos definido un modelo para la distribución F de una variable en una población, y para fijar ideas supongamos que hemos elegido un modelo paramétrico $F(x, \boldsymbol{\theta})$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$, donde $\Theta \in R^k$. En general, va a interesar saber sobre F algo más que el hecho de pertenecer a la familia $F(x, \boldsymbol{\theta})$. Puede interesar conocer totalmente la distribución, es decir, el valor de $\boldsymbol{\theta}$, o algunas características de la misma.

Ejemplo 1: Volvamos al ejemplo 1 de 2.1 y supongamos que hemos elegido para la distribución de X en la población la familia $N(\mu, \sigma^2)$. Consideremos tres problemas diferentes.

(a) Interesa conocer la distribución F completamente. En este caso hace falta conocer los valores de ambos parámetros, μ y σ^2.

(b) Se requiere sólo el conocimiento de la producción total. Como hay 1000 parcelas la producción total sería 1000 μ y por lo tanto bastaría con conocer μ.

(c) Se ha fijado una meta de producir al menos 200 toneladas de trigo y lo único que interesa es saber si se cumple o no la meta. Luego en este caso lo único que interesa es saber si $\mu < 200$ o $\mu \geq 200$, aunque no interesa el valor exacto de μ.

Volvamos al problema general, la característica numérica que interesa de la distribución puede ser expresada como $q(\theta_1, \theta_2, \ldots, \theta_k)$, donde $q(\theta_1, \theta_2, \ldots, \theta_k)$ es una función de Θ en R si interesa una sola característica numérica, o en R^h si interesan h características. En el ejemplo 1, tendríamos para (a) $q(\mu, \sigma^2) = (\mu, \sigma^2)$; para (b) $q(\mu, \sigma^2) = 1000\mu$ y para (c)

$$q(\mu, \sigma^2) = \begin{cases} 0, & \text{si} \quad \mu < 200 \\ 1, & \text{si} \quad \mu \geq 200 \end{cases} .$$

Así, en este último caso $q(\mu, \sigma^2) = 0$ nos indica que no se cumplió la meta y $q(\mu, \sigma^2) = 1$ indica que se cumplió.

Para conocer el valor de $q(\theta_1, \theta_2, \ldots, \theta_k)$ exactamente, deberíamos conocer el valor de la variable X en toda la población. Así, en el ejemplo 1,

deberíamos conocer la producción de todas las parcelas. Observar el valor de la variable para todos los elementos de la población puede ser muy costoso, o aún imposible, como en el caso de poblaciones infinitas. Inclusive en el caso de poblaciones finitas puede ser imposible si se quiere la información con cierta premura. En el ejemplo 1, si se pueden cosechar sólo 20 parcelas por día, se necesitarían 50 días para conocer cuál es la producción de cada una de las 1000 parcelas. Si se quisiera el primer día de la cosecha hacer una estimación de la producción total, ésta debería hacerse en base a los resultados de las 20 parcelas cosechadas ese día.

Se puede definir la *Estadística* como la ciencia que estudia los procedimientos para determinar el valor de una o varias características $q(\theta_1, \ldots, \theta_k)$ de una distribución de una variable en una población que se supone pertenece a una familia $F(x, \theta_1, \theta_2, \ldots, \theta_k)$ observando sólo unos pocos elementos si se trata de una población finita o realizando unos pocos experimentos en el caso de una población infinita. Al conjunto de estas pocas observaciones en base a las cuales se determinará $q(\theta_1, \theta_2, \ldots, \theta_k)$ se denomina *muestra.* Si el modelo es no paramétrico esta formulación cambiará ligeramente, como se verá más adelante.

Los procedimientos estadísticos pueden clasificarse en dos grandes tipos: procedimientos de diseño y procedimientos de inferencia.

Procedimientos de diseño: Son los procedimientos para elegir las observaciones que componen la muestra, de manera que con pocas observaciones se pueda obtener la mayor información posible sobre $q(\theta_1, \theta_2, \ldots, \theta_k)$.

Procedimientos de inferencia: Son los procedimientos que permiten a partir de la muestra inferir la característica de la distribución de la variable en la población que interesa, es decir $q(\theta_1, \theta_2, \ldots, \theta_k)$.

Para ejemplificar, volvemos nuevamente al Ejemplo 1. En este caso un posible diseño, no necesariamente el óptimo, para la selección de la muestra de 20 observaciones puede ser el siguiente. Se elige la primera parcela al azar. El rendimiento de esta parcela será una variable aleatoria que llamaremos X_1 y que tendrá distribución $N(\mu, \sigma^2)$. La segunda parcela se elige al azar entre todas las que quedan. El rendimiento de esta parcela será una variable aleatoria que llamaremos X_2. Como la población de parcelas es grande (hay 1000 parcelas), la distribución de la variable X prácticamente no se modificará después de la extracción de la primera parcela, por lo tanto a los efectos prácticos, X_2 puede ser considerada como una variable aleatoria

independiente de X_1 y con la misma distribución $N(\mu, \sigma^2)$. Repitiendo este procedimiento tendremos variables aleatorias $X_1, X_2, \ldots, X_{20}$ que podemos considerar independientes y cada una con una distribución $N(\mu, \sigma^2)$. Denominaremos a $X_1, X_2, \ldots, X_{20}$ *muestra aleatoria* de tamaño 20 de la distribución $N(\mu, \sigma^2)$.

En general, se dirá que $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ es una *muestra aleatoria de tamaño n de una distribución* $F(\mathbf{x})$ si $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ son variables aleatorias (o vectores aleatorios) independientes e idénticamente distribuídas con distribución $F(\mathbf{x})$. Es decir si

$$F_{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = F(\mathbf{x}_1)\, F(\mathbf{x}_2) \ldots F(\mathbf{x}_n) \tag{2.1}$$

y en el caso que $F(x)$ sea una distribución discreta o continua con función de frecuencia o de probabilidad p, (2.1) será equivalente a

$$p_{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = p(\mathbf{x}_1)\, p(\mathbf{x}_2) \ldots p(\mathbf{x}_n)$$

En el caso de poblaciones finitas, una muestra aleatoria de tamaño n se obtendrá observando n elementos de la población elegidos al azar. Para que las variables fuesen estrictamente independientes los elementos deberían elegirse uno a uno y ser restituídos en la población antes de elegir el próximo. Sin embargo si el tamaño de la muestra es relativamente pequeño respecto al total de la población, aunque no se haga la restitución las variables observadas serán aproximadamente independientes, y a los fines prácticos podemos considerarla una muestra aleatoria.

En el caso de poblaciones infinitas, la muestra aleatoria se obtendrá simplemente repitiendo el experimento n veces y observando cada vez el vector de variables correspondiente.

Consideremos ahora cómo a partir de la muestra $X_1, X_2, \ldots, X_{20}$ que hemos obtenido, utilizando procedimientos de inferencia resolvemos los problemas (a), (b) y (c) que hemos planteado.

El problema (a) consistía en encontrar aproximadamente la distribución de la variable X en la población, es decir, estimar μ y σ^2.

Definamos $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$; luego para estimar μ se puede usar $\overline{X}_{20}$. Es de esperar que $\overline{X}_{20}$ se aproxima a μ ya que de acuerdo a la ley de los grandes números $\lim_{n\to\infty} \overline{X}_n = \mu$ c.t.p.

El procedimiento estadístico para estimar μ a partir de la muestra, es formar el promedio de los valores que la componen; es decir $\overline{X}_{20}$. Esto es un *procedimiento de inferencia estadística*, ya que a partir de una muestra

de 20 observaciones, *inferimos* el valor μ característico de la distribución de la variable en la población.

Similarmente se puede estimar σ^2. Partimos de $\sigma^2 = \text{Var } X_i = E(X_i^2) - (E(X_i))^2$. Dado que $E(X_i^2)$ puede estimarse por $(1/20)\sum_{i=1}^{20} X_i^2$, σ^2 puede estimarse por

$$\widehat{\sigma}^2_{20} = \frac{1}{20}\sum_{i=1}^{20} X_i^2 - \overline{X}^2_{20}$$

Haciendo manipulaciones algebraicas, se obtiene

$$\widehat{\sigma}^2_{20} = \frac{1}{20}\sum_{i=1}^{20} (X_i - \overline{X}_{20})^2$$

En general, si se tuviese una muestra aleatoria de tamaño n, σ^2 podría estimarse por

$$\widehat{\sigma}^2_{n} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_{n})^2$$

En el problema (b), cuando se quiere conocer la producción total, es decir $q(\mu, \sigma^2) = 1000\mu$, podemos usar para esta estimación 1000 $\overline{X}_{20}$. Es decir, el procedimiento de inferencia sería el siguiente. Se hace el promedio de las observaciones que componen la muestra, y se lo multiplica por 1000.

En el problema (c), es decir el problema de decidir si $\mu < 200$ o $\mu \geq 200$, el procedimiento de inferencia puede ser el siguiente: se decidirá que $\mu < 200$ si $\overline{X}_{20} < 200$ y se decidirá que $\mu \geq 200$ si $\overline{X}_{20} \geq 200$.

Los problemas (a) y (b) son los que se denominan de *estimación puntual,* mientras que el problema (c) es un problema de *test de hipótesis,* ya que en base a la muestra se desea decidir entre dos opciones y determinar las probabilidades de error. Como veremos más adelante, las dos hipótesis no se considerarán en forma simétrica y se determinará cuál de los dos errores a cometer es más grave, para poder controlar su probabilidad.

Los procedimientos que hemos propuesto no son los únicos posibles, ni necesariamente los mejores; solamente fueron introducidos para ejemplificar la naturaleza de los procedimientos estadísticos. Podemos formular una *primera* generalización de la situación descripta en el Ejemplo 1 diciendo que un *problema de inferencia estadística paramétrica* consistirá en: dada una muestra aleatoria de tamaño n, $X_1, X_2, \ldots, X_n$ de la distribución de una variable en una población de la cual se conoce solamente que pertenece a una familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_k)$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta\}$, donde

$\Theta \subseteq R^k$, se quiere inferir conocimiento de algunas características de esta distribución, definidas por una función $q(\boldsymbol{\theta})$ que va de Θ en R^h, siendo h el número de características en las que se está interesado.

Ejemplo 2: Volvamos al ejemplo 6 de 2.3. Supongamos que se quiere conocer μ. Observemos que si F es la distribución de la variable X, entonces de acuerdo con las hipótesis del modelo para toda $F \in \mathcal{F}$ se tiene que μ es la esperanza correspondiente a la distribución F, si es que esta existe (puede no existir) y también μ es la mediana correspondiente a F (la mediana siempre existe). Luego μ es una cierta función de F, digamos $\mu = q(F)$. Si queremos estimar μ, debemos tomar una muestra aleatoria de F, digamos de tamaño n; $X_1, X_2, \ldots, X_n$. Esto se logrará repitiendo n veces la medición de μ. Consideremos ahora el procedimiento para inferir μ. Si estuviésemos seguros que F tiene esperanza podríamos usar para estimar μ, $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$, ya que de acuerdo a la ley de los grandes números debería converger a $E(X_i) = \mu$. Sin embargo la existencia de esperanza no es una hipótesis que hemos requerido para que $F \in \mathcal{F}$. En caso que F no tenga esperanza, se puede mostrar que $\overline{X}_n$ no converge a μ y por lo tanto no será un buen estimador.

En este caso, podemos usar el siguiente procedimiento: ordenamos las X_i, obteniendo $X^{(1)} < X^{(2)} < X^{(3)} < \cdots < X^{(n)}$, donde $X^{(1)}$ es la menor de las X_i, $X^{(2)}$ la siguiente, hasta llegar a $X^{(n)}$, que sería la mayor de todas. Supongamos que $n = 2p + 1$, luego estimamos μ por $\widehat{\mu} = X^{(p+1)}$, es decir por la observación central. Si $n = 2p$ podemos tomar como $\widehat{\mu} = (X^{(p)} + X^{(p+1)})/2$. Por ejemplo, si tuviésemos 7 mediciones y estas resultasen 6.22; 6.25; 6.1; 6.23; 6.18; 6.15; 6.29, se tendría $X^{(1)} = 6.1$; $X^{(2)} = 6.15$; $X^{(3)} = 6.18$; $X^{(4)} = 6.22$; $X^{(5)} = 6.23$; $X^{(6)} = 6.25$ y $X^{(7)} = 6.29$. Estimaríamos μ por $\widehat{\mu} = X^{(4)} = 6.22$. Se puede mostrar que este procedimiento da resultados razonables para una familia $\mathcal{F}$ como la estudiada.

El ejemplo 2 nos sugiere la siguiente formulación del *problema de inferencia estadística no paramétrica:* Dada una muestra aleatoria de tamaño n, $X_1, \ldots, X_n$ de la distribución F de una variable en una población, y de la cual se sabe solamente que pertenece a una familia $\mathcal{F}$ que no puede ser indicada por un número finito de parámetros reales, interesa conocer algunas características de F expresadas como una función $q(F)$ que va de $\mathcal{F}$ a R^h, siendo h el número de características que interesan.

El siguiente ejemplo nos permitirá formular un tipo de problemas de inferencia estadística más general que el estudiado hasta ahora.

Ejemplo 3: Supongamos que el rendimiento por hectárea de un cierto cultivo depende de la cantidad de fertilizante que se usa y que la relación es de la forma

$$X = aG + b + \varepsilon$$

donde G es la cantidad de fertilizante usado por hectárea, X el rendimiento por hectárea y ε un término aleatorio que tiene en cuenta todos los otros factores que intervienen en la determinación de los rendimientos, a y b son parámetros desconocidos.

Supongamos que se cultivan n parcelas usando respectivamente $G_1, G_2, \ldots, G_n$ cantidad de fertilizante por hectárea y sean los rendimientos respectivos observados $X_1, X_2, \ldots, X_n$. Luego se tendrá:

$$X_i = aG_i + b + \varepsilon_i \qquad\qquad 1 \leq i \leq n$$

Supongamos que las ε_i son variables aleatorias independientes igualmente distribuídas con distribución $N(0, \sigma^2)$, donde σ^2 es desconocido. Los valores $G_1, G_2, \ldots, G_n$ son valores numéricos conocidos (no variables aleatorias).

Luego en este caso las variables aleatorias $X_i, \quad 1 \leq i \leq n$, serán independientes con distribución $N(aG_i + b, \sigma^2)$ y por lo tanto no son igualmente distribuídas. En este caso estamos interesados en conocer los parámetros a y b que establecen la relación entre G y X quizás también en σ^2 que establece la varianza de ε, es decir del término residual.

Estos parámetros deben ser estimados a partir del vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Sin embargo, el vector $\mathbf{X}$ tiene componentes con diferentes distribuciones. Se podrían dar ejemplos donde las variables no sean tampoco independientes.

Esto nos sugiere un concepto más amplio de problema estadístico que los vistos anteriormente.

Un *problema de inferencia estadística paramétrica general* consistirá en: dado un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de cuya *distribución conjunta* se conoce solamente que pertenece a una familia $\mathcal{F} = \{F(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_k) \quad \text{con } \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subset R^k\}$, inferir conocimiento sobre una función $q(\boldsymbol{\theta})$ de Θ en R^h.

En el ejemplo 3, $\boldsymbol{\theta} = (a, b, \sigma^2)$ y la densidad correspondiente a la distribución es

$$p(x_1, x_2, \ldots, x_n; a, b, \sigma^2) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(x_i - a\,G_i - b)^2}$$

La función $q(\boldsymbol{\theta})$ dependerá del problema que interesa. Si se quiere conocer la relación entre G y X lo que interesará será $q(\boldsymbol{\theta}) = (a, b)$. Si interesa saber

cuál es el rendimiento promedio cuando se utilizan 200 kg por hectárea, lo que interesará conocer será $q(\boldsymbol{\theta}) = 200\,a + b$. Si interesa saber solamente si el fertilizante tiene un efecto positivo, la función $q(\boldsymbol{\theta})$ estará dada por $q(\boldsymbol{\theta}) = \begin{cases} 0 & \text{si} \quad a \leq 0 \\ 1 & \text{si} \quad a > 0 \end{cases}$.

Un procedimiento de inferencia estadística para este problema se verá en el ejemplo 1 de la sección 3.4. Una teoría general que abarca este problema se verá en el capítulo 7.

De la misma forma se podría formular el concepto de *problema de inferencia estadística no paramétrica general.*

Concepto de estadístico

Supongamos dado un problema de inferencia estadística donde se observa un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ con distribución en la familia $F(x_1, x_2, \ldots, x_n; \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$ y donde se quiera inferir acerca de $q(\boldsymbol{\theta})$. Esta inferencia se tendrá que hacer a partir de $\mathbf{X}$, es decir, por funciones de $\mathbf{X}$. Luego se define como *estadístico* a cualquier función medible que tenga como argumento a $\mathbf{X}$ y que tome valores en un espacio euclideo de dimensión finita. En el ejemplo 1, hemos visto que la estimación de μ y σ^2 se hacía mediante el estadístico

$$\mathbf{T} = r(\mathbf{X}) = \left(\sum_{i=1}^{n} \frac{X_i}{n} \,, \quad \sum_{i=1}^{n} \frac{(X_i - \overline{X}_n)^2}{n} \right)$$

En el ejemplo 3, se usó el estadístico $\mathbf{T} = r(\mathbf{X}) = X^{(p+1)}$.

Hasta ahora, hemos supuesto que el parámetro de existir es fijo. Existe otra aproximación, en la cual, el parámetro es una variable aleatoria. Los procedimientos estadísticos bayesianos suponen que $\boldsymbol{\theta}$ es una variable aleatoria no observable, a valores en un espacio Θ con distribución τ. La distribución *a priori* τ establecida antes de tomar la muestra, se modifica en base a los datos para determinar la distribución *a posteriori,* que resume lo que se puede decir del parámetro $\boldsymbol{\theta}$ en base a las suposiciones hechas y a los datos.

Los métodos estadísticos, que van desde el análisis de datos hasta el análisis bayesiano, permiten sacar en forma creciente conclusiones cada vez más fuertes, pero lo hacen al precio de hipótesis cada vez más exigentes y, por lo tanto, menos verificables.