Inferencia estadística

Matemáticas en LaTeX

3 de noviembre de 2024

Índice general

Capítulo 1

Introducción a la inferencia

La estadística inferencial se ocupa de predecir y sacar conclusiones para una población tomando como base una muestra de dicha población. Como todas las predicciones, siempre han de hacerse bajo un cierto grado de fiabilidad o confianza.

En un sentido amplio, la inferencia es la parte de la estadística que estudia grandes poblaciones a partir de una pequeña parte de estas.

Existen diversos problemas de inferencia estadística según el tipo de conclusiones que se quieran establecer sobre la situación aleatoria.

- **Estimación puntual.** Se pretende obtener un pronóstico numérico único acerca de un determinado parámetro de la distribución.
- **Estimación por intervalos.** El objetivo es proporcionar un margen de variación para un determinado parámetro de la distribución.
- **Contrastes de hipótesis.** Se trata de corroborar o invalidar una determinada afirmación acerca de la distribución.

Dependiendo del grado de conocimiento de esta distribución, se distinguen dos métodos para realizar procesos de inferencia.

- **Inferencia paramétrica.** 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.

  Usaremos el enfoque clásico, en el que los parámetros de la distribución de la población se consideran constantes. También existe el enfoque bayesiano, que considera los parámetros de la distribución como variables aleatorias y permite introducir información sobre ellos.
- **Inferencia no paramétrica.** No supone ninguna distribución de probabilidad de la población y, en su lugar, exige solo hipótesis muy generales.

1.1. Elementos de la inferencia estadística

Todo problema de inferencia estadística está motivado por un cierto grado de desconocimiento de la ley de probabilidad que rige cierto fenómeno aleatorio. En los casos más simples, se estudia un fenómeno aleatorio que se rige según una variable aleatoria X con función de distribución F, que se llama función de distribución teórica.

Los elementos de la inferencia estadística son:

- **Población.** Conjunto de elementos sobre los que se observa una característica común.

- **Muestra.** Conjunto de unidades de una población.
- **Parámetros poblacionales.** Índices centrales y de dispersión que definen a una población.

En los modelos de inferencia estadística, el grado de desconocimiento acerca de la distribución teórica F se refleja mediante una familia $\mathcal{F}$ de distribuciones. La situación más sencilla es aquella en la que la familia $\mathcal{F}$ está compuesta por distribuciones que tienen una forma fija y dependen de un parámetro θ que varía en un subconjunto $\Theta \subset \mathbb{R}^k$, llamado espacio paramétrico. Es decir,

$$\mathcal{F} = \{F_\theta : \theta \in \Theta \in \mathbb{R}^k\}.$$

Ejemplo. Si X es el número de éxitos en 30 pruebas de $\text{Ber}(\theta)$, escribimos

$$F_X \equiv F_\theta \in \mathcal{F}, \quad \mathcal{F} = \{\text{Bi}(n = 30, \theta) : \theta \in (0, 1)\}.$$

1.2. Muestreo aleatorio

Para que la muestra sea representativa debe reflejar las similitudes y diferencia encontradas en la población. Los errores más comunes que se pueden cometer son:

- **Error de muestreo.** Hacer conclusiones muy generales a partir de la observación de solo una parte de la población.
- **Error de inferencia.** Hacer conclusiones hacia una población de mayor tamaño que de la que se tomó la muestra.

Los métodos para muestreo se clasifican en probabilísticos y no probabilísticos.

1.2.1. Muestreo probabilístico

Los muestreos probabilísticos se basan en el principio de equiprobabilidad, que establece que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra.

- **Muestreo aleatoria simple.** Todas las unidades muestrales tienen la misma probabilidad de ser elegidas. Puede ser con o sin reemplazamiento.
- **Muestreo estratificado.** La población está dividida en estratos que contienen elementos parecidos entre sí. La composición de la muestra se distribuye entre los distintos estratos mediante un procedimiento que se llama afijación. Existen dos tipos:
  - **Afijación uniforme.** En la muestra hay el mismo número de representantes por cada estrato.
  - **Afijación proporcional.** En la muestra hay un número de representantes de cada estrato proporcional a su tamaño.
- **Muestreo por conglomerados.** Se establecen grupos de elementos físicamente próximos entre ellos, frecuentemente constituidos por una partición geográfica de la población.
- **Muestreo sistemático.** Se elige un individuo al azar y, a partir de él, se eligen los demás a intervalos constantes hasta completar la muestra.

1.2.2. Muestreo no probabilístico

Cuando el muestreo probabilístico resulta costoso o no se tiene asegurado que cualquier elemento de la población pueda pertenecer a una muestra, se consideran muestreos no probabilísticos. Sin embargo, estos métodos no sirven para realizar generalizaciones debido a que no se tiene la certeza de que la muestra sea representativa.

- **Muestreo por cuotas.** Se fijan cuotas que consisten en un número de individuos que reúnen determinadas condiciones.
- **Muestreo de conveniencia.** Se intentan obtener muestras representativas mediante la inclusión en la muestra de grupos típicos.
- **Bola de nieve.** Se localiza a algunos individuos que a su vez conducen a otros, y así sucesivamente hasta conseguir una muestra suficiente. Se usa cuando se hacen estudios en poblaciones marginales, delincuentes, sectas, determinados tipos de enfermos, etc.

1.3. Muestra aleatoria simple

Definición 1.1. Una muestra aleatoria simple de tamaño n de una variable aleatoria X con distribución teórica F_X es un conjunto de n variables aleatorias independientes $(X_1, \ldots, X_n)$ e igualmente distribuidas con distribución común F_X.

Cada X_i es una variable aleatoria que representa la característica bajo estudio del elemento i-ésimo de la muestra. Una vez realizado el muestreo, los resultados obtenidos $(x_1, \ldots, x_n)$ se denominan realización de la muestra.

Como las variables aleatorias son independientes, la función de distribución conjunta de la muestra aleatoria simple es

$$F(x_1, \ldots, x_n) = F_X(x_1) \ldots F_X(x_n).$$

1.4. Concepto de estadístico y su distribución

Definición 1.2. Llamamos estadístico a cualquier función medible de las variables aleatorias observables de la muestra aleatoria simple que no depende de ningún parámetro desconocido en el estudio.

En general, un estadístico es una función medible $T : \mathbb{R}^n \to \mathbb{R}^k$ aplicada a la muestra aleatoria simple.

Un estadístico T induce una nueva varible aleatoria que tendrá una distribución asociada llamada distribución muestral con una función de distribución llamada función de distribución empírica.

Los estadísticos más utilizados son los siguientes:

- **Media muestral.** Estima la media.

$$T(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- **Varianza muestral.** Estima la varianza.

$$T(X_1, X_2, \ldots, X_n) = \mathrm{Var}_n(\vec{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- **Cuasivarianza muestral.** Estima la varianza. En general, es una mejor aproximación que la varianza muestral.

$$T(X_1, X_2, \ldots, X_n) = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} \mathrm{Var}_n(\vec{X}).$$

- **Máximo y mínimo.**

$$T(X_1, X_2, \ldots, X_n) = \mathrm{máx}(X_1, X_2, \ldots, X_n) = X_{(n)},$$
$$T(X_1, X_2, \ldots, X_n) = \mathrm{mín}(X_1, X_2, \ldots, X_n) = X_{(1)}.$$

Nota. Todos estos estadísticos son variables aleatorias.

Ejemplo. Sea $X \sim \{F_\theta : \theta \in \Theta\}$, con $E(X) = \mu$ y $V(X) = \sigma^2$. Consideramos $(X_1, X_2, X_3, X_4, X_5)$ muestra aleatorias simple, con $X_i \sim F_\theta$ para todo i. Al realizar la muestra, se obtiene $(3, 8, 4, 5, 5)$.

Para estimar la media, consideramos la variable aleatoria

$$\overline{X} = \frac{1}{5}\sum_{i=1}^{5} X_i.$$

Podemos aproximar

$$\hat{\mu} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{3 + +8 + 4 + 5 + 5}{5} = 5.$$

Para estimar la varianza, consideramos la variable aleatoria

$$\mathrm{Var}_5(\vec{X}) = \frac{1}{5}\sum_{i=1}^{5}(X_i - \overline{X})^2 = \frac{\sum_{i=1}^{5} X_i^2 - 5\overline{X}^2}{5}.$$

Obtenemos la estimación

$$\hat{\sigma}^2 = \frac{(3^2 + 8^2 + 4^2 + 5^2 + 5^2) - 5 \cdot 5^2}{5} = 2{,}8.$$

También podemos considerar la cuasivarianza.

$$S^2 = \frac{1}{4}\sum_{i=1}^{5}(X_i - \overline{X})^2 = \frac{\sum_{i=1}^{5} X_i^2 - n\overline{X}^2}{4} = \frac{5}{4}\mathrm{Var}_5(\vec{X}).$$

De esta forma, $\hat{\sigma}^2 = \frac{5}{4} \cdot 2{,}8 = 3{,}5$.

Calculamos el máximo.

$$T(X_1, X_2, \ldots, X_5) = \text{máx}(X_1, X_2, \ldots, X_5), \quad T(3, 8, 4, 5, 5) = 8.$$

Calculamos el mínimo.

$$T^*(X_1, X_2, \ldots, X_5) = \text{mín}(X_1, X_2, \ldots, X_5), \quad T^*(3, 8, 4, 5, 5) = 3.$$

1.4.1. Función de distribución muestral

Sea $X \sim F_\theta$, con $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$. Identificamos $F_\theta \equiv F$.

Tomamos una muestra aleatoria simple $(X_1, X_2, \ldots, X_n)$, con $X_i \sim F$ para todo i e independientes. Definimos la función de distribución muestral:

$$F^*_{(X_1, X_2, \ldots, X_n)}(x) = \frac{\text{número de variables tales que } X_i \leq x}{n}, \quad x \in \mathbb{R}.$$

Es una nueva variable aleatoria.

Observamos que podemos escribir dicha variable aleatoria de la siguiente forma:

$$F^*_{(X_1, X_2, \ldots, X_n)}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x]}(X_i).$$

Si tomamos

$$Y_i = I_{(-\infty, x]}(X_i) = \begin{cases} 1, & \text{si } X_i \leq x, \\ 0, & \text{si } X_i > x, \end{cases}$$

tenemos que $Y_i \sim \text{Ber}(p)$ con

$$p = P(Y_i = 1) = P(X_i \leq x) = F_{X_i}(x) = F(x).$$

Entonces

$$F^*_{(X_1,X_2,\ldots,X_n)}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

Como Y_i son variables aleatorias independientes con $Y_i \sim \text{Ber}(F(x))$, entonces

$$nF^*_{(X_1,X_2,\ldots,X_n)}(x) = \sum_{i=1}^{n} Y_i \sim \text{Bi}(n, F(x)).$$

De esta forma, podemos calcular:

$$E(F^*_{(X_1,X_2,\ldots,X_n)}(x)) = \frac{1}{n}nF(x) = F(x),$$
$$V(F^*_{(X_1,X_2,\ldots,X_n)}(x)) = \frac{1}{n^2}nF(x)(1-F(x)) = \frac{1}{n}F(x)(1-F(X)).$$

Usando el teorema central del límite, podemos aproximarla como

$$F^*_{(X_1,X_2,\ldots,X_n)}(x) \sim N\left(F(x), \sqrt{\frac{F(x)(1-F(x))}{n}}\right).$$

Ejemplo. Obtengamos la función de distribución muestral para la realización de la muestra aleatoria simple $(3, 8, 4, 5, 5)$.

$$F^*_{(3,8,4,5,5)}(x) = \begin{cases} 0, & \text{si } x < 3, \\ \frac{1}{5}, & \text{si } 3 \leq x < 4, \\ \frac{2}{5}, & \text{si } 4 \leq x < 5, \\ \frac{4}{5}, & \text{si } 5 \leq x < 8, \\ 1, & \text{si } x \geq 8. \end{cases}$$

1.4.2. Momentos muestrales centrales y no centrales

Sea $X \sim F_X$ y sea $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria simple, con $X_i \sim F$ para todo i e independientes.

- El momento ordinario de orden k respecto del origen es

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- El momento central de orden k viene dado por

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

Nota. Estos momentos son variables aleatorias.

Capítulo 2

Distribución normal

Definición 2.1. Sea $\vec{X} = (X_1, X_2, \ldots, X_n)^t$ un vector aleatorio y sea $\vec{\mu} = (E(X_1), E(X_2), \ldots, E(X_n))^t$ el vector de medias. Se define la matriz de varianzas y covarianzas Σ como:

$$\Sigma = E((\vec{X} - \vec{\mu}) \cdot (\vec{X} - \vec{\mu})^t) = \begin{pmatrix} V(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & V(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & V(X_n) \end{pmatrix}.$$

2.1. Distribución chi cuadrada

Definición 2.2. Una variable aleatoria X se distribuye según una $\chi^2_{(n)}$ si $X \sim \mathrm{Ga}\left(\alpha = \frac{n}{2}, \beta = \frac{1}{2}\right)$, con $n \in \mathbb{N}$. La función de densidad es de la forma

$$f(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} e^{-\frac{1}{2}x} x^{\frac{n}{2}-1}, \quad x > 0.$$

- **Génesis de la distribución.** Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes con $X_i \sim N(0,1)$. Si consideramos la variable aleatoria $Y = \sum_{i=1}^n X_i^2$, entonces $Y \sim \chi^2_{(n)}$.

  El parámetro n se conoce como grados de libertad y hace referencia al número de sumandos que aportan variabilidad a la suma.

- **Características numéricas.** Si $Y \sim \chi^2_{(n)}$, entonces $E(Y) = n$ y $V(Y) = 2n$.

- **Aproximación por el teorema central del límite.** Si $Y \sim \chi^2_{(n)}$, mediante el teorema central del límite podemos aproximar:

$$\frac{Y - n}{\sqrt{2n}} \longrightarrow N(0,1) \Rightarrow \sqrt{2Y} - \sqrt{2n-1} \longrightarrow N(0,1).$$

- **Reproductividad.** Si $T \sim \chi^2_{(n)}$ y $W \sim \chi^2_{(m)}$ son variables aleatorias independientes, entonces $T + W \sim \chi^2_{(n+m)}$.

Teorema 2.1 (Teorema de Fisher)**.** *Si $(X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple de una población $N(0,1)$, entonces*

1. *$(n-1)S^2$ y $\overline{X}$ son variables aleatorias independientes.*
2. *La distribución del muestreo es:*

$$(n-1)S^2 \sim \chi^2_{(n-1)}, \quad \overline{X} \sim N\left(0, \frac{1}{\sqrt{n}}\right).$$

Teorema 2.2 (Teorema de Fisher generalizado)**.** *Si $(X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple de una población $N(\mu, \sigma)$, entonces:*

1. *$\frac{(n-1)S^2}{\sigma^2}$ y $\overline{X}$ son variables aleatorias independientes.*
2. *La distribución del muestreo es:*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}, \quad \overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

2.2. Distribución t de Student

Si $(X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple de una población $N(\mu, \sigma)$, entonces la distribución en el muestreo de $\overline{X}$ es $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ o, equivalentemente,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

Si sustituimos σ por S, donde S^2 es la cuasivarianza muestral, entonces tenemos el estadístico

$$t = \sqrt{n-1}\frac{\overline{X} - \mu}{S}.$$

Definición 2.3. Sean $X, X_1, X_2, \ldots, X_n$ $n+1$ variables aleatorias independientes con distribución $N(0, \sigma)$. Sea $Y = \sum_{i=1}^n X_i^2$ y $U = \sqrt{Y/n}$. Entonces

$$t = \frac{X}{U} = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}}.$$

La distribución de t se denomina distribución t-Student con n grados de libertad y la notaremos $t \sim t_n$.

- **Características numéricas.** Para $n = 1$, la distribución t_1 es la distribución de Cauchy con densidad

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}.$$

  Esta distribución no posee momentos de primer orden y por consiguiente carece de varianza. Si $n > 1$, la media es finita y vale cero. Para $n > 2$ la varianza existe y es $\frac{n}{n-2}$.

- **Aproximación por el teorema central del límite.** Mediante el teorema central del límite podemos aproximar la distribución t_n por una distribución $N(0, 1)$.

2.3. Distribución F de Snedecor

Consideremos dos poblaciones normales independientes $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ que queremos comparar. Utilizaremos el estadístico $\frac{S_1^2}{S_2^2}$, cuya distribución en el muestreo se puede calcular de forma explícita.

Definición 2.4. Si $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ son variables aleatorias independientes con distribución $N(0, 1)$, entonces la distribución del estadístico

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i^2}{\frac{1}{m}\sum_{j=1}^m Y_j^2}$$

es una distribución F-Snedecor con n y m grados de libertad y la notaremos $F_{n,m}$.

Una distribución $F_{n,m}$ es la distribución del cociente de dos distribuciones χ^2 independientes, de n y m grados de libertad respectivamente, divididas cada una de ellas por sus grados de libertad.

- **Características numéricas.** La media de esta distribución es $\frac{m}{m-2}$ si $m > 2$ y su varianza existe si $m > 4$.
- **Propiedad.** Si $\alpha \in (0,1)$, entonces el percentil α para $F_{n,m}$ es el inverso del percentil $1-\alpha$ para $F_{m,n}$. Es decir,

$$F_{n,m;\alpha} = \frac{1}{F_{m,n;1-\alpha}}, \quad \alpha \in (0,1).$$

Estas tres distribuciones están tabuladas. Las tablas, en general, dan percentiles de la distribución.

Ejemplo. Sea $(X_1, X_2) \sim N_2((\mu_1, \mu_2)^t, \Sigma)$, con

$$\begin{cases} E(X_1) = \mu_1, \\ E(X_2) = \mu_2, \end{cases} \qquad \Sigma = \begin{pmatrix} V(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & V(X_2) \end{pmatrix}.$$

Si X_1 y X_2 son independientes, entonces

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{V(X_1) \cdot V(X_2)}} = 0.$$

En general, $\rho = 0$ no implica que X_1 y X_2 sean independientes. Sin embargo, si $(X_1, X_2) \sim N_2(\vec{\mu}, \Sigma)$, entonces $\rho = 0 \Leftrightarrow X_1$ y X_2 son independientes.

Podemos escribir:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$