Inferencia Estadísitica: Teoría

Daniel Monjas Miguélez

16 de enero de 2022

Índice

1. Introducción a la Inferencia Estadística. Estadísticos muestrales.

1.1. Muestra Aleatoria Simple

Defnción: Una muestra aleatoria simple (de tamaño n) de una variable X es un vector $(X_1, \ldots, X_n)$ formado por n variables aleatorias independientes, todas con la misma distribución que X.

- **Realización muestral:** Cada valor $(x_1, \ldots, x_n)$ obtenido al observar $(X_1, \ldots, X_n)$.
- **Espacio muestral (χ^n):** Conjunto de todas las posibles realizaciones muestrales según nuestro conocimiento de P_X (valores de $(X_1, \ldots, X_n)$ bajo alguna distribución de la familia):

$$\chi_\theta = \{x / f_\theta(x) > 0\} \rightarrow \textit{valores de } X \textit{ bajo } P_\theta$$
$$\chi = \bigcup_{\theta \in \Theta} \chi_\theta \rightarrow \textit{valores de } X \textit{ bajo } \mathcal{P}$$
$$\chi^n = \chi \times \chi \times \cdots \times \chi = \{(x_1, \ldots, x_n);\ x_1, \ldots, x_n \in \chi\}$$

1.2. Función de distribución muestral

Definición: Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X con función de distribución F_X. Se define la función de distribución muestral, $F^*_{X_1,\ldots,X_n}$, como una función sobre $\mathbb{R}$ dada por,

$$F^*_{X_1,\ldots,X_n}(x) = \frac{n \textit{ de variables } X_i \leq x}{n} = \frac{\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)}{n}, \quad \forall x \in \mathbb{R}$$

Propiedades de la función de distribución muestral:

1. $\forall x \in \mathbb{R}$, $F^*_{X_1,\ldots,X_n}(x)$ es una variable aleatoria tal que $nF^*_{X_1,\ldots,X_n}(x) \rightsquigarrow B(n, F(x))$:

$$E[F^*_{X_1,\ldots,X_n}(x)] = F(x), \qquad Var[F^*_{X_1,\ldots,X_n}(x)] = \frac{F(x)(1-F(x))}{n}$$

2. Si n es grande, y la distribución binomial poco manejable, puede aproximarse por una normal según el Teorema límite de Lèvy:

$$F^*_{X_1,\ldots,X_n}(x) \rightsquigarrow \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Demostración:

1. Se tiene que:

$$I_{(-\infty,x]}(X_i) \rightsquigarrow B(1, F(x)) \Rightarrow \sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \rightsquigarrow B(n, F(x)) \Rightarrow$$
$$\Rightarrow nF^*_{X_1,\ldots,X_n}(x) \rightsquigarrow B(n, F(x))$$

   Teniendo en cuenta la linealidad de la esperanza:

$$E[nF^*_{X_1,\ldots,X_n}(x)] = nF(x) \Rightarrow nE[F^*_{X_1,\ldots,X_n}(x)] = nF(x) \Rightarrow$$
$$\Rightarrow E[F^*_{X_1,\ldots,X_n}(x)] = F(x)$$

   Por otro lado, puesto que $Var[nX] = n^2 Var[X]$:

$$Var[nF^*_{X_1,\ldots,X_n}(x)] = n^2 Var[F^*_{X_1,\ldots,X_n}(x)] = nF(x)(1-F(x)) \Rightarrow$$
$$\Rightarrow Var[F^*_{X_1,\ldots,X_n}(x)] = \frac{F(x)(1-F(x))}{n}$$

2. Como $nF^*_{X_1,\ldots,X_n}(x)$ es usma de variables aleatorias independientes e idénticamente distribuidas, el Teorema Central del Límite permite afirmar que, cuando $n \to +\infty$,

$$\frac{nF^*_{X_1,\ldots,X_n}(x) - nF(x)}{\sqrt{nF(x)(1-F(x))}} = \sqrt{n}\frac{F^*_{X_1,\ldots,X_n}(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \rightsquigarrow \mathcal{N}(0,1)$$

   de manera que, cuando el tamaño muestral es grande, se tiene

$$F^*_{X_1,\ldots,X_n}(x) \rightsquigarrow \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Teorema de Glivenko-Cantelli: Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con función de distribución F_X. Si $F^*_{X_1,\ldots,X_n}(x)$ es la función de distribución muestral asociada a la muestra aleatoria simple $(X_1,\ldots,X_n)$ de $X \rightsquigarrow P$, se verifica que $F^*_{X_1,\ldots,X_n}(x)$ converge casi seguramente y uniformemente a la función de distribución de X, F_X.

$$P\left[\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} |F^*_{X_1,\ldots,X_n}(x) - F(x)| = 0\right] = 1$$

Es decir, con probabilidad 1, al tomar sucesivas observaciones independientes de la variable y considerar las correspondientes funciones de distribución muestrales, se verifica:

$$\varepsilon > 0, \exists n_\varepsilon / n > n_\varepsilon \Rightarrow F(x) \in (F^*_{X_1,\ldots,X_n}(x) - \varepsilon, F^*_{X_1,\ldots,X_n}(x) + \varepsilon)$$

1.3. Estadístico muestral

Definición: Dada una muestra aleatoria simple $(X_1, \ldots, X_n)$ de X, un estadístico muestral es una función de ella, medible e independiente de cualquier parámetro desconocido; esto es, una transformación $T(X_1, \ldots, X_n)$, tal que $T : (\mathbb{R}^n, \mathbf{B}^n) \to (\mathbb{R}^k, \mathcal{B}^k)$ es medible $(T^{-1}(B) \in \mathcal{B}^n,\ \forall B \in \mathcal{B}^k)$.

Definición: Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X. La distribución en el muestreo de un estadístico T definido en el espacio muestral $(\chi^n, \mathcal{B}^n)$ es la distribución de la variable aleatoria $T(X_1, \ldots, X_n)$.

Proposición: La función generatriz de momentos del estadístico media muestral viene dada por:

$$M_{\overline{X}}(t) = (M_x(t/n))^n$$

Demostración: Recordemos que la función generatriz de momentos de una variable aleatoria es aditiva-multiplicativa en el sentido de que la función generatriz de momentos de la suma de n variables aleatorias independientes coincide con el producto de las n funciones generatrices de momentos.

Sea X una variable aleatoria, y $(X_1, \ldots, X_n)$ una muestra aleatoria simple de X de tamaño n.

$$\begin{aligned} M_{\overline{X}}(t) &= E\left[e^{t\overline{X}}\right] = E\left[e^{t\frac{\sum X_i}{n}}\right] = E\left[\prod e^{\frac{tX_i}{n}}\right] = \\ &= \prod_{i=1}^{n} E\left[e^{\frac{t}{n}X_i}\right] = \prod_{i=1}^{n} M_{X_i}(t/n) = (M_X(t/n))^n \end{aligned}$$

Proposición: La distribución muestral general de los estadísiticos de orden, según la variable sea discreta o continua, es:

- X discreta:

$$\begin{aligned} P\left[X_{(r)} \leq x\right] &= P\left[\textit{al menos } r \textit{ elementos muestrales sean} \leq x\right] = \\ &= \sum_{i=1}^{n} \binom{n}{i} (P[X \leq x])^i (P[X \geq x])^{n-i} \end{aligned}$$

- X continua:

$$g_r(x_{(r)}) = \frac{n!}{(r-1)!(n-r)!}[F(x_{(r)})]^{r-1}[1-F(x_{(r)})]^{n-r} f(x_{(r)})$$

Definición (Momentos muestrales centrados y no centrados): Se definen los momentos centrados y no centrados de una muestra aleatoria simple de tamaño n como:

- Momento no centrado de orden $k \in \mathbb{N} : A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

  En particular, se tiene $A_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} =$ media muestral.

- Momentos centrado de orden $k \in \mathbb{N} : B_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k$.

  En particular, se tiene $B_1 = 0$ (por las propiedades de la media de una distribución), y definimos $B_2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = Var[X] =$ varianza muestral.

Proposición: Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X.

1. Para los momentos no centrados, se tiene:

$$E[A_k] = E[X^k] \qquad Var[A_k] = \frac{1}{n}(E[X^{2k}] - E[X^k]^2)$$

   En particular, $E[A_1] = \mu$, $Var[A_1] = \frac{\sigma^2}{n}$, donde μ y σ^2 son, respectivamente, la media y la varianza poblacional.

2. Para el momento centrado de orden 2, se tiene

$$E[B_2] = \frac{(n-1)\sigma^2}{n}$$

   En particular, se verifica $E[S^2] = \sigma^2$, donde $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

A S^2 así definida le llamamos cuasivarianza muestral.

Demostración: Mirar apuntes.

Definición (Cuantil Muestral): Para cada $p \in (0, 1)$, el cuantil de orden p, c_p, es un valor real tal que

$$F_n^*(c_p) \geq p \quad y \quad F_n^*(c_p^-) \leq p$$

Se puede expresar de la siguiente forma en función de los elementos de la muestra ordenada:

- Si $np \in \mathbb{N}$, $c_p = \frac{X_{(np)} + X_{(np+1)}}{2}$
- En otro caso, sea $[np]$ la parte entera de np, entonces $cp = X_{([np]+1)}$

Defínción (Función generatriz de momentos muestral): Se define la función generatriz de momentos muestral como

$$M^*(t) = \frac{1}{n} \sum_{i=1}^{n} e^{tX_i}$$

Proposción: La función generatriz de momentos muestral se utiliza para obtener los momentos no centrados, pues se verifica

$$\frac{\partial^k M^*(t)}{\partial t^k}_{|t=0} = A_k$$

2. Distribuciones en el muestreo de poblaciones normales.

2.1. Distribuciones χ^2 de Pearson, t de Student y F de Snedecor.

Proposición (Distribución de muestreo de la meida muestral): Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, y $(X_1, \ldots, X_n)$ muestra aleatoria simple de X, entonces

$$\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

2.1.1. Distribución χ^2 de Pearson

Definición (Distribución χ^2 de Pearson): Se dice que una variable aleatoria X tiene una distribución χ^2 de Pearson con n grado de libertad ($n \in \mathbb{N}$), y se denota $X \rightsquigarrow \chi^2(n)$, si su función de densidad de probabilidad es

$$f_X(x) = \frac{1}{\Gamma(n/2) \cdot 2^{n/2}} x^{n/2-1} e^{-x/2}, \qquad x > 0$$

La distribución χ^2 de Pearson es un caso particular de la distribución Gamma, $\Gamma(p, a)$, cuya función de densidad es

$$f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}, \qquad x > 0$$

Propiedades (Distribución χ^2 de Pearson):

1. La función generatriz de momentos de una variable aleatoria $X \rightsquigarrow \chi^2(n)$ es
$$M_X(t) = \frac{1}{(1-2t)^{n/2}}, \qquad t < \frac{1}{2}$$

2. Los momentos no centrados de la distribución $\chi^2(n)$ vienen dados por
$$E[X^k] = 2^k \frac{\Gamma(\frac{n}{2} + k)}{\Gamma(\frac{n}{2})}$$
En particular, $E[X] = n$, $E[X^2] = n^2 + 2n \Rightarrow Var[X] = E[X^2] - E[X]^2 = 2n$.

3. **Reproductividad.** Si $X_1, \ldots, X_n$ son variables aleatorias independientes con $X_i \rightsquigarrow \chi^2(k_i)$, entonces
$$\sum_{i=1}^{n} X_i \rightsquigarrow \chi^2\left(\sum_{i=1}^{n} k_i\right)$$

4. **Relación entre las distribuciones χ^2 y $\mathcal{N}(0,1)$.** Si $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas con distribución común $\mathcal{N}(0,1)$, entonces:

$$\sum_{i=1}^{n} X_i^2 \rightsquigarrow \chi^2(n)$$

Observación importante: Para valores de n grandes ($n > 50$, normalmente), se utiliza que su distribución se puede aproximar por una $\mathcal{N}(n, 2n)$, por el teorema central del límite de Lèvy.

Gráfica de la función de densidad de $\chi^2(n)$: Asimétrica a la derecha y unimodal.

2.1.2. Distribución t de Student

Definición: Sean X e Y variables aleatorias independientes con distribuciones $X \rightsquigarrow \mathcal{N}(0,1)$ e $Y \rightsquigarrow \chi^2(n)$. Entonces la variable aleatoria,

$$\frac{X}{\sqrt{Y/n}}$$

se dice que tiene una distribución t de Student de n grados de libertad, y se denota $T \rightsquigarrow t(n)$.

Propiedades de la distribución t de Student:

1. La función de densidad de probabilidad de una distribución $t(n)$ viene dada por

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \qquad t \in \mathbb{R}.$$

2. Sea T una variable aleatoria con distribución $t(n)$, con $n > 1$. Entonces, se tiene que existen los momentos no centrados $E[T^r]$ para $r < n$, y se verifica que:

   - Si r es impar, entonces $E[T^r] = 0$.
   - Si r es par, entonces

   $$E[T^r] = \frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{n-r}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

   En particular, para $n > 2$, existen los momentos de primer y segundo orden, y se tiene $E[T] = 0$, $E[T^2] = Var[T] = \frac{n}{n-2}$

Observación importante: Para valores de n grande, se utiliza que la distribución $t(n)$ se puede aproximar por $\mathcal{N}(0,1)$.

Gráfica de la función de densidad de $t(n)$: La función de densidad de la distribución $t(n)$ cumple que es simétrica y unimodal.

2.1.3. Distribución F de Snedecor

Definición (Distribución F de Snedecor): Sean X e Y variables aleatorias independientes con distribuciones $X \rightsquigarrow \chi^2(m)$ e $Y \rightsquigarrow \chi^2(n)$. Entonces, la variables aleatoria

$$F = \frac{X/m}{Y/n}$$

se dice que tiene una distribución F de Snedecor con (m,n) grados de libertad, y se denota por $F \rightsquigarrow F(m,n)$.

Propiedades de la distribución F de Snedecor:

1. La función de densidad de probabilidad de una distribución $F(m,n)$ viene dada por

$$g(f) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{m/2} f^{m/2-1}\left(1+\frac{m}{n}f\right)^{-\frac{m+n}{2}}, \qquad f>0$$

2. Sea F una variable aleatoria con distribución $F(m,n)$. Entonces, se verifica que

$$E[F^r] = \left(\frac{n}{m}\right)^r \frac{\Gamma\left(\frac{m}{2}+r\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}, \qquad 0<r<\frac{n}{2}$$

   En particular, $E[F] = \frac{n}{n-2}$ si $n>2$, $E[F^2] = \frac{n^2(m+2)}{m(n-4)(n-2)}$ si $n>4$, luego se tiene que $Var[F] = \frac{n^2(2m+2n-4)}{m(n-2)^2(n-4)}$ si $n>4$.

3. Se verifican las siguientes propiedades:
   - $F \rightsquigarrow F(m,n) \Leftrightarrow F^{-1} \rightsquigarrow F(n,m)$
   - $T \rightsquigarrow t(n) \Leftrightarrow T^2 \rightsquigarrow F(1,n)$

Gráfica de la función de densidad de $F(m,n)$: Es asimétrica a la derecha y unimodal.

2.2. Muestreo en poblaciones normales

2.2.1. Muestreo en una pobalación normal unidimensional

Teorema: Sea $(X_1,\dots,X_n)$ una muestra aleatoria simple de una variable aleatoria $X \rightsquigarrow \mathcal{N}(\mu,\sigma^2)$. Entonces, $\overline{X}$ y $(X_1-\overline{X},\dots,X_n-\overline{X})$ son independientes.

Corolario: Bajo las mismas condiciones del teorema anterior, se tiene:

1. $\overline{X} \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n})$
2. **Lema de Fisher**. $\overline{X}$ y S^2 son independientes.
3. $\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$
4. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \rightsquigarrow t(n-1)$

Estadísticos para hacer inferencia sobre los parámetros de una población unidimensional en diferentes situaciones:

- Inferencia sobre μ:

$$\begin{array}{ll} \sigma_0^2 \quad \textit{conocida}: & \dfrac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1) \\ \sigma^2 \quad \textit{desconocida}: & \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \rightsquigarrow t(n-1) \end{array}$$

- Inferencia sobre σ^2:

$$\begin{array}{ll} \mu_0 \quad \textit{conocida}: & \dfrac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sigma^2} \rightsquigarrow \chi^2(n) \\ \mu \quad \textit{desconocida}: & \dfrac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1) \end{array}$$

2.2.2. Muestreo en dos poblaciones normales unidimensionales

Proposición:

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0,1)$$

Lema extendido de Fisher: Los vectores $(\overline{X}, \overline{Y})$ y (S_1^2, S_2^2) son independientes.

Estadísticos para hacer inferencia sobre los parámetros de dos poblaciones normales unidimensionles en los siguientes casos:

- Inferencia sobre $\mu_1 - \mu_2$:

$$\begin{array}{ll} \sigma_1^2, \sigma_2^2 \quad \textit{conocidas}: & \dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0,1) \\ \sigma_1^2=\sigma_2^2=\sigma^2 \quad \textit{desconocidas}: & \dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}}\sqrt{\dfrac{n_1+n_2-2}{\frac{1}{n_1}+\frac{1}{n_2}}} \rightsquigarrow t(n_1+n_2-2) \end{array}$$

- Inferencia sobre $\frac{\sigma_2^2}{\sigma_1^2}$:

$$\mu_1, \mu_2 \quad conocidas: \qquad \frac{n_2}{n_1}\frac{\sigma_2^2}{\sigma_1^2}\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2} \rightsquigarrow F(n_1, n_2)$$

$$\mu_1, \mu_2, \quad desconocidas: \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_1^2} \rightsquigarrow F(n_1-1, n_2-1)$$