

Tema 3. Conceptos relacionados con la Distribución Empírica

basado en

B. Efron, R. Tibshirani (1993). An Introduction to the bootstrap.

O. Kirchkamp (2019). Resampling methods.

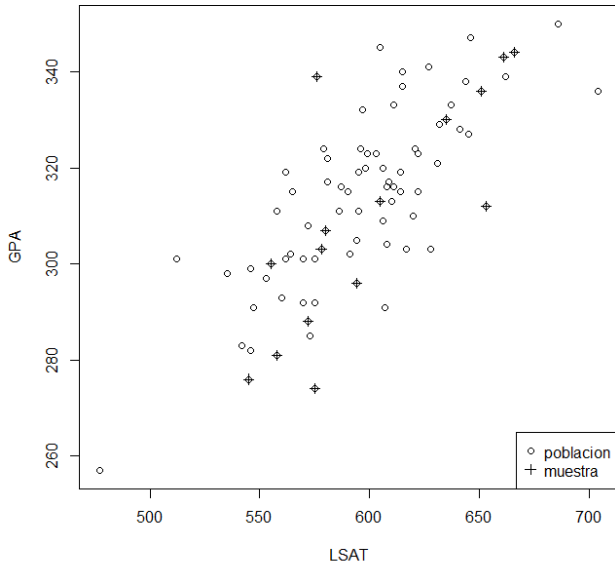
Curso 2022/2023

Parámetros, distribuciones y el principio de plug-in

- ▶ El mejor modo de visualizar una muestra aleatoria es en términos de una **población finita**: un universo U de unidades individuales U_1, U_2, \dots, U_N cada una de las cuales tiene la misma probabilidad de ser seleccionada.
- ▶ En cada unidad U_i se mide una variable de interés X_i de modo que se obtiene un *censo* X_1, X_2, \dots, X_N , es decir, \mathbb{X} .
- ▶ Una muestra aleatoria de tamaño n es una colección de n unidades u_1, u_2, \dots, u_n seleccionadas al azar del universo U .
- ▶ En cada unidad seleccionada u_i se toma una medida de interés x_i de modo que una muestra se denota como \mathbf{x} .
- ▶ Los problemas en estadística en general se refieren a estimar algún aspecto de la distribución de probabilidad F en base a una muestra.

Ejemplo sobre el principio de *plug-in*

- ▶ Se toma el ejemplo de las universidades con máster en leyes que está incluido en el libro de Efron y Tibshirani.
- ▶ Se trata de una población compuesta por 82 universidades en relación un máster en *Leyes* (está en la librería **bootstrap** de R con el nombre **law82**), donde **GPA** es la puntuación media en los cursos de grado, y **LSAT** es la calificación de admisión.
- ▶ La muestra contiene 15 observaciones.



Ejemplo sobre el principio de *plug-in*

- ▶ Interesa calcular la correlación entre GPA (la puntuación media en los cursos de grado) y LSAT (calificación de admisión).
- ▶ Con la verdadera puntuación poblacional:

```
with(law82, cor(GPA, LSAT))
```

```
[1] 0.7599979
```

- ▶ El estimador *plug-in* es

```
with(law, cor(GPA, LSAT))
```

```
[1] 0.7763745
```

Función de distribución empírica

- ▶ La distribución empírica denominada \hat{F} es un estimador simple de la función de distribución teórica F .
- ▶ El principio de *plug-in* consiste en estimar algún aspecto de F como la media, mediana etc. mediante \hat{F} .
- ▶ El bootstrap es una aplicación directa de este principio.
- ▶ Supongamos que se observa una muestra aleatoria de tamaño n con función de distribución F

$$F \rightarrow (x_1, x_2, \dots, x_n)$$

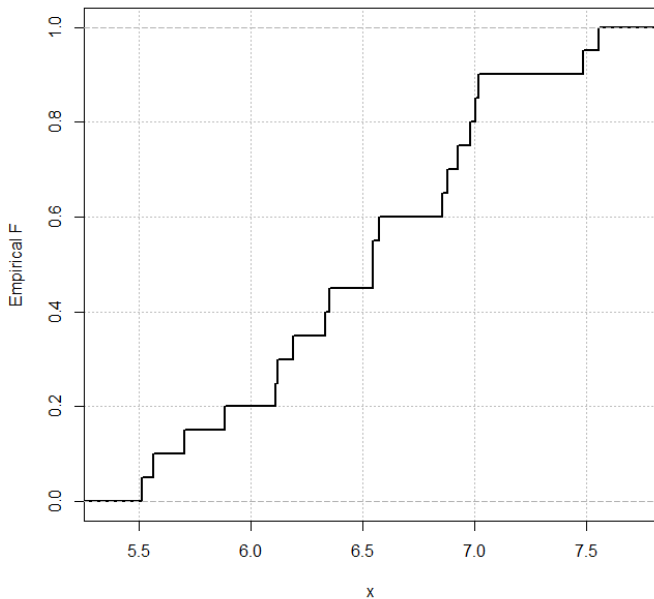
Función de distribución empírica

- ▶ La función de distribución empírica \hat{F} se define como la distribución discreta que asigna probabilidad $\frac{1}{n}$ a cada valor x_i tal que $i = 1, 2, \dots, n$
- ▶ De este modo \hat{F} asigna a un conjunto A del espacio muestral de x la probabilidad empírica

$$\hat{P}(A) = \frac{\#\{x_i \in A\}}{n}$$

- ▶ Esa es la proporción de la muestra observada x que ocurre en A .
- ▶ Se simulan datos de calificaciones según una normal y se dibuja la correspondiente función de distribución empírica.

EDF de Calificaciones



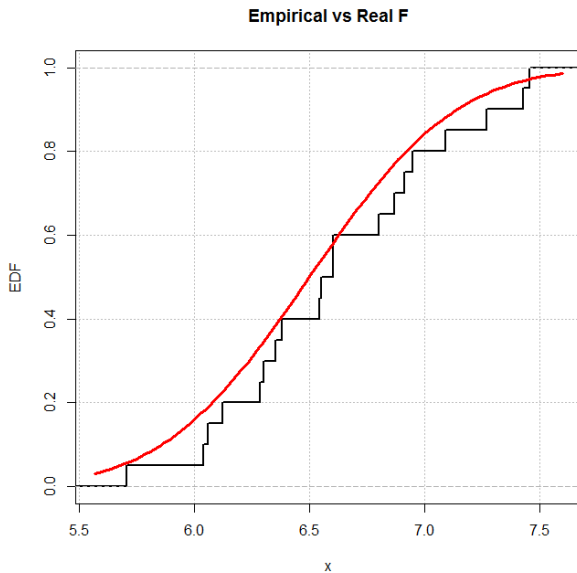


Figura: Distribución empírica y función de distribución real

Función de distribución empírica

- ▶ Se define la función de distribución empírica como

$$\hat{F}_n(x) = \frac{\text{Número de elementos de la muestra } \leq x}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{x_i \leq x\}$$

donde $\mathbf{I}\{A\}$ es la función indicatriz del suceso A .

- ▶ En general, $\hat{F}_n(x)$ se puede considerar como una función de distribución discreta que asigna probabilidad igual a $\frac{1}{n}$ a cada uno de los n valores x_1, \dots, x_n .
- ▶ Así $\hat{F}_n(x)$ es una función escalón con un salto de tamaño $\frac{1}{n}$ en cada punto x_i para $i = 1, \dots, n$.

Función de distribución empírica

- ▶ Si se ordenan los valores de la muestra de menor a mayor $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ entonces $\hat{F}_n(x) = 0$ para $x < x_{(1)}$
- ▶ $\hat{F}_n(x)$ salta al valor $1/n$ en $x = x_{(1)}$ y se mantiene igual a $1/n$ para $x_{(1)} \leq x < x_{(2)}$
- ▶ $\hat{F}_n(x)$ salta al valor $2/n$ en $x = x_{(2)}$ y se mantiene igual a $2/n$ para $x_{(2)} \leq x < x_{(3)}$ y así sucesivamente
- ▶ Si se fija el valor de x entonces la variable aleatoria $\mathbf{I}\{x_i \leq x\}$ es una v.a. Bernoulli de parámetro $p = F(x)$
- ▶ Entonces $n\hat{F}_n(x)$ es una v.a. **binomial** de media $nF(x)$ y varianza $nF(x)(1 - F(x))$.

Propiedades de la función de distribución empírica

- ▶ Así,

$$E \left[n\hat{F}(x) \right] = nF(x) \Rightarrow E \left[\hat{F}(x) \right] = F(x)$$

$$\text{Var} \left[n\hat{F}(x) \right] = nF(x)(1 - F(x)) \Rightarrow \text{Var} \left[\hat{F}(x) \right] = \frac{1}{n}F(x)(1 - F(x))$$

de modo que $\hat{F}(x)$ es un estimador insesgado de $F(x)$.

- ▶ Si se denota como $F(x)$ la función de distribución de la v.a. de la que procede la muestra entonces, para todo número $(-\infty < x < \infty)$, la probabilidad de que una observación dada X_i sea menor o igual que x es $F(x)$.
- ▶ Por tanto, por la *ley de los grandes números*, cuando $n \rightarrow \infty$, la proporción $\hat{F}_n(x)$ de observaciones en la muestra que son menores o iguales que x convergen en probabilidad a $F(x)$.

$$\hat{F}_n(x) \xrightarrow{P} F(x) \quad -\infty < x < \infty$$

Propiedades de la función de distribución empírica

- ▶ Pero se tiene un resultado más potente: $\hat{F}_n(x)$ converge a $F(x)$ de manera **uniforme** para todos los valores de x
- ▶ **Lema de Glivenko-Cantelli** Sea $\hat{F}_n(x)$ una función de distribución empírica de una m.a.s X_1, \dots, X_n de una función de distribución F entonces,

$$D_n = \sup_{-\infty < x < \infty} |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0$$

- ▶ **Nota:** Antes de que los valores X_1, \dots, X_n hayan sido observados D_n es una v.a.
- ▶ Cuando el tamaño muestral n es grande la función de distribución empírica $\hat{F}_n(x)$ está muy próxima a $F(x)$ sobre la recta real.
- ▶ Así, cuando se desconoce la función de distribución $F(x)$ se puede considerar que $\hat{F}_n(x)$ es un estimador muy eficiente de $F(x)$.

Propiedades de la función de distribución empírica

- ▶ Supongamos que se tiene un conjunto de observaciones X_1, \dots, X_n procedente de una m.a.s. de una población con función de distribución F .
- ▶ Dado un estadístico de interés $T_n = T_n(X_1, \dots, X_n)$, se trata de estimar la distribución de una función de F y T_n , digamos $R_n(T_n, F)$:

$$H_n(x) = P\{R_n \leq x\},$$

- ▶ Por ejemplo, R_n podría ser la cantidad pivotal que se usa para construir los intervalos para la media asumiendo normalidad:

$$R_n = \frac{\bar{X}_n - \mu}{\sigma_0/\sqrt{n}}$$

Propiedades de la función de distribución empírica

- ▶ Entonces, se obtienen muestras bootstrap X_1^*, \dots, X_n^* de la distribución empírica basada en X_1, \dots, X_n de modo que se puede definir $T_n^* = T_n(X_1^*, \dots, X_n^*)$ y $R_n^* = R_n(T_n^*, F_n)$, donde F_n es la función de distribución empírica.
- ▶ El estimador bootstrap de H_n se calcula como

$$\hat{H}_n(x) = P_* \{R_n^* \leq x\},$$

donde P_* es la distribución basada en las muestras bootstrap.

- ▶ Se observa que $\hat{H}_n(x)$ depende de la distribución empírica F_n . Así, $\hat{H}_n(x)$ cambia cuando los datos $\{x_1, \dots, x_n\}$ varían.

Propiedades de la función de distribución empírica

- ▶ Supongamos que se trata de estimar la varianza de un estimador $\hat{\theta} = \theta(x_1, \dots, x_n)$.
- ▶ La varianza teórica viene dada por

$$\text{Var}(\hat{\theta}) = \int \cdots \int \left\{ \theta(x_1, \dots, x_n) - E(\hat{\theta}) \right\}^2 dF(x_1) \cdots dF(x_n)$$

donde

$$E(\hat{\theta}) = \int \cdots \int \theta(x_1, \dots, x_n) dF(x_1) \cdots dF(x_n)$$

Propiedades de la función de distribución empírica

- La solución natural es usar como estimador *plug-in* la distribución empírica \hat{F}

$$\widehat{Var}(\hat{\theta}) = \int \cdots \int \left\{ \theta(x_1, \dots, x_n) - \hat{E}(\hat{\theta}) \right\}^2 d\hat{F}(x_1) \cdots d\hat{F}(x_n)$$

- es decir, al ser \hat{F} una distribución discreta

$$\widehat{Var}(\hat{\theta}) = \frac{1}{n^n} \sum_j \left\{ \theta(x_1^j, \dots, x_n^j) - \hat{E}(\hat{\theta}) \right\}^2$$

donde (x_1^j, \dots, x_n^j) varía entre todas las posibles combinaciones de los datos muestrales.

Propiedades de la función de distribución empírica

- ▶ A no ser que n sea pequeño, los cálculos anteriores pueden llevar bastante tiempo computacional.
- ▶ Sin embargo, se pueden aproximar las expresiones mediante integración Monte Carlo.
- ▶ Se aproxima la integral tomando muestras aleatorias de tamaño n de la función \hat{F} y calculando la media muestral del integrando.
- ▶ Por la *Ley de los Grandes Números* esta aproximación converge al verdadero valor de la integral cuando $n \rightarrow \infty$.

Propiedades de la función de distribución empírica

- ▶ ¿A qué se parece una muestra aleatoria tomada de \hat{F} ?
- ▶ Como \hat{F} asigna igual masa de probabilidad a cada valor observado x_i , el hecho de tomar una muestra aleatoria de \hat{F} es equivalente a tomar n valores con reemplazamiento de x_1, \dots, x_n .
- ▶ De hecho, a la técnica de tomar nuevas muestras a partir de la muestra original es a lo que realmente se denomina **remuestreo**.

Intervalos de confianza basados en la función de distribución empírica

► **Teorema de Dvoretzky-Kiefer-Wolfowitz (DKW):**

Sea X_1, \dots, X_n una muestra aleatoria de una v.a. con función de distribución F . Entonces, para todo $\varepsilon > 0$

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

- Se pueden construir así los intervalos de confianza para \hat{F} .

IC basados en la función de distribución empírica

- ▶ Se define

$$L(x) = \max \left\{ \hat{F}_n(x) - \varepsilon_n, 0 \right\}$$

$$U(x) = \min \left\{ \hat{F}_n(x) + \varepsilon_n, 1 \right\}$$

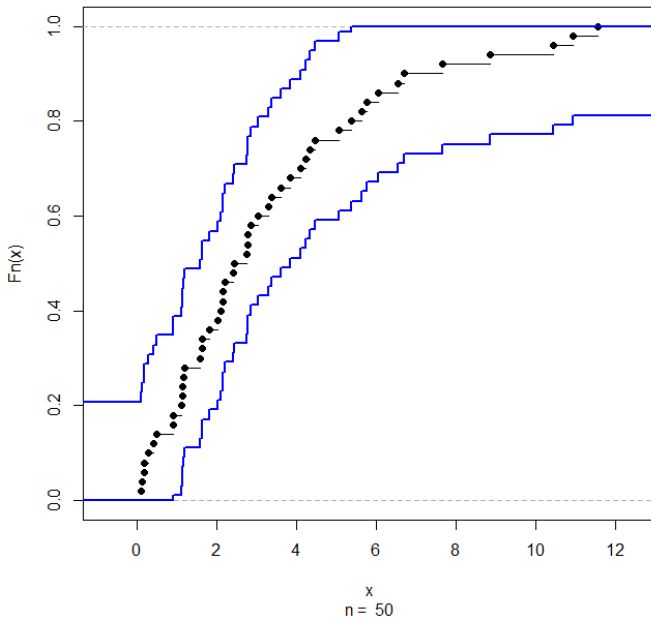
donde $\varepsilon_n = \sqrt{\frac{1}{2n} \log \left(\frac{2}{\alpha} \right)}$

- ▶ de este modo, para toda función de distribución F y para todo x :

$$P(L(x) \leq F(x) \leq U(x)) \geq 1 - \alpha$$

- ▶ Se pueden programar fácilmente las cotas, pero los intervalos son amplios para tamaños muestrales pequeños.

ecdf(x) + 95% K.S. bands



ecdf(datos) + 95% K.S. bands

