Histograma y Función empírica

1. Histograma

La función de densidad es muy útil para caracterizar la distribución de una variable aleatoria X. En general en la práctica, $f(x)$ es desconocida y buscamos estimarla a partir de una muestra de variables aleatorias independientes y todas con distribución $f(x)$.
El histograma es la forma no paramétrica mas común para estimar $f(x)$. La construcción es bastante simple. Dada una muestra de datos $x_1, x_2, \ldots, x_n$, que se asumen realizaciónes de las variables aleatorias $X_1, X_2, \ldots, X_n$, todas con distribución $f(x)$ e independientes, se realizan los siguientes pasos:

- Se selecciona un origen x_0 y se divide la recta real en intervalos de longitud h
$$B_j = [x_0 + (j-1)h, x_0 + jh], j \in \mathbb{N}$$
No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Esto facilita la lectura.

- Se cuenta cuantas observaciones caen en cada intervalo armando una tabla de frecuencias. Denotamos a la cantidad de observaciones que caen en el intervalo j como n_j

- Para cada intervalo, se divide la frecuencia absoluta por la cantidad total de la muestra n (para convertirlas en frecuencias relativas, análogo a como se hace con las probabilidades) y por la longitud h (para asegurarse que el area debajo del histograma sea igual a 1):
$$f_j = \frac{n_j}{nh}$$

- Se grafica el histograma realizando una barra vertical sobre cada intervalo con altura f_j y ancho h

Formalmente, el histograma está dado por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} \mathbf{1}(x_i \in B_j)\mathbf{1}(x \in B_j)$$

Si m_j es el centro del intervalo j, podemos escribir $B_j = [m_j - \frac{h}{2}, m_j + \frac{h}{2})$.

Se puede verificar facilmente que el area del histograma es igual a 1, propiedad que se requiere para cualquier estimador razonable de una función de densidad.

1.1. Ejemplo

Los siguientes valores se refiere al peso de 30 saquitos de té en gramos.

1,50	1,35	0,91	1,09	1,25	1,17	1,03	0,99	1,20	1,11
1,35	1,10	1,05	1,00	1,30	0,98	1,28	1,02	1,19	1,15
1,14	1,08	0,99	0,92	0,95	1,20	1,07	0,93	1,07	1,19

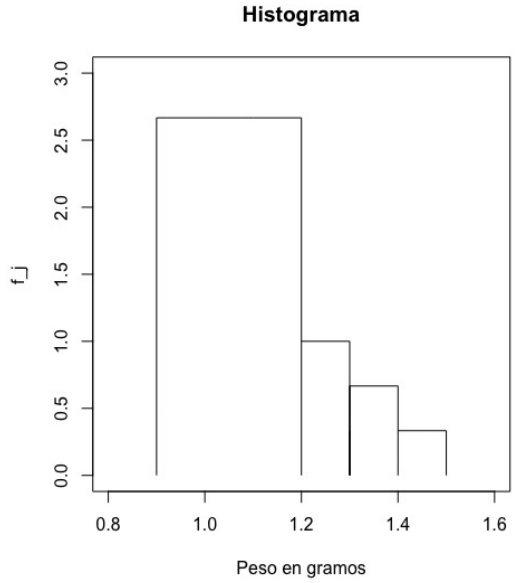
1. Construir la tabla de frecuencias usando $h = 0{,}1$
2. Realizar el histograma correspondiente.

Armamos la tabla de frecuencias indicando primero el intervalo tomando como x_0 al valor observado más pequeño, luego la frecuencia absoluta, y por último el valor de la función histograma:

B_j	n_j	f_j
[0.9,1.1)	8	8/3
[1,1.1)	8	8/3
[1.1,1.2)	8	8/3
[1.2,1.3)	3	1
[1.3,1.4)	2	2/3
[1.4,1.5)	1	1/3

Una vez costruida la tabla podemos entonces expresar y graficar la función histograma:

$$\hat{f}_h(x) = \tfrac{8}{3}\mathbf{1}(0{,}9 \leq x < 1) + \tfrac{8}{3}\mathbf{1}(1 \leq x < 1{,}1) + \tfrac{8}{3}\mathbf{1}(1{,}1 \leq x < 1{,}2) \\ +\mathbf{1}(1{,}2 \leq x < 1{,}3) + \tfrac{2}{3}\mathbf{1}(1{,}3 \leq x < 1{,}4) + \tfrac{1}{3}\mathbf{1}(1{,}4 \leq x < 1{,}5)$$



2. Función empírica

La Función empírica es una estimación de la función de distribución acumulada de una variable aleatoria. Lo mínimo que puede pedirse a esta función es que cumpla con las condiciones que debe tener una función de distribución:

1. $\hat{F}_X(x) \in [0,1], \forall x \in \mathbb{R}$
2. $\hat{F}_X(x)$ es monótona no decreciente
3. $\hat{F}_X(x)$ es continua a derecha
4. $lim_{x\to-\infty}\hat{F}_X(x) = 0$ y $lim_{x\to\infty}\hat{F}_X(x) = 1$

Se define la Función empírica como

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(x_i \leq x)$$

Donde $x_1, x_2, \ldots, x_n$ se asumen realizaciónes de las variables aleatorias $X_1, X_2, \ldots, X_n$, todas con distribución $F_X(x)$ e independientes

2.1. Ejemplo

De una variable aleatoria X se ha obtenido la muestra: {2,5; 2,2; 2,4; 2,2; 2,1}. Hallar y graficar la función de distribución empírica asociada a la muestra y

usarla para estimar la probabilidad $\mathbf{P}(X \leq 2{,}3)$

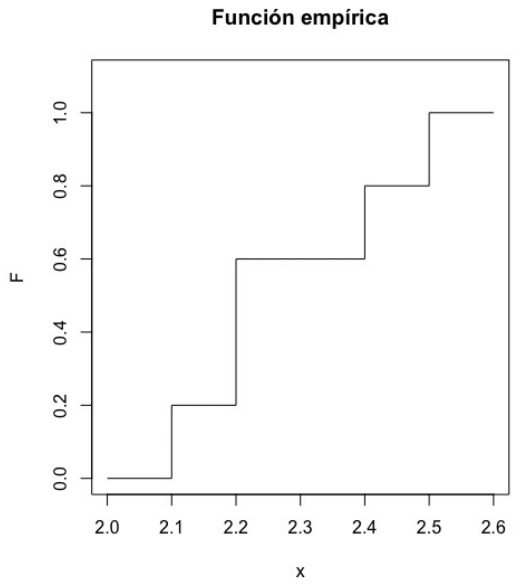
Según la definición:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x)$$

En este caso, n=5, por lo tanto:

$$\begin{aligned}
\hat{F}(x) &= \frac{1}{5}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x) \\
&= \frac{1}{5}\mathbf{1}(2,1 \leq x) + \frac{1}{5}\mathbf{1}(2{,}2 \leq x) + \frac{1}{5}\mathbf{1}(2,2 \leq x) + \frac{1}{5}\mathbf{1}(2,4 \leq x) + \frac{1}{5}\mathbf{1}(2,5 \leq x) \\
&= \frac{1}{5}\mathbf{1}(2,1 \leq x) + \frac{2}{5}\mathbf{1}(2{,}2 \leq x) + \frac{1}{5}\mathbf{1}(2,4 \leq x) + \frac{1}{5}\mathbf{1}(2,5 \leq x) \\
&= \frac{1}{5}\mathbf{1}(2,1 \leq x < 2{,}2) + \frac{3}{5}\mathbf{1}(2{,}2 \leq x < 2{,}4) + \frac{4}{5}\mathbf{1}(2{,}4 \leq x < 2{,}5) + \mathbf{1}(2,5 \leq x)
\end{aligned}$$

El gráfico de la Función empírica tendra forma de escalera:



Para estimar la probabilidad:

$$\mathbf{P}(X \leq 2{,}3) = F_X(2{,}3) \approx \hat{F}(2{,}3) = \frac{3}{5}$$