Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued

Esperanza, Varianza, Sumas y .... Figuritas!

Pregunta de Cuestionario para Químicos

Pregunta: ¿Qué interpretás si te decimos que la probabilidad de que una lampara producida por cierta fábrica dure a lo sumo un año es 0.3?

- Interpreto que 3 de cada 10 lamparas durarían a lo sumo un año.
- que de 10 lamparas 3 duran a lo sumo 1 año
- que de 10 lamparas en general 3 duran a lo sumo un año
- Que de todas las l ámparas que produce esa f ábrica un 30 % durar á a lo sumo 1 año.

Frecuencias relativas

Frecuencia relativa con la que se observa el evento de interés.

- n: número de veces que repetimos el experimento,
- n_A: número de veces que el resultado del experimento pertenece al evento A en las n repeticiones,
- $\frac{n_A}{n}$: frecuencia relativa del evento A en n repeticiones.
- Se observa *empíricamente* que las frecuencias relativas se estabilizan (convergen a cierto valor).

Frecuencias relativas - con m repeticiones

Frecuencia relativa con la que se observa el evento de interés.

- m: número de veces que repetimos el experimento,
- m_A: número de veces que el resultado del experimento pertenece al evento A en las m repeticiones,
- $\frac{m_A}{m}$: frecuencia relativa del evento A en m repeticiones.
- Se observa *empíricamente* que las frecuencias relativas se estabilizan (convergen a cierto valor).

Probabilidad: Motivación Frecuentista

- $\mathbb{P}(A)$ *procura representar el límite de las frecuencias relativas: porcentaje de veces que esperamos que A ocurra en* ***infinitas*** *repeticiones*
- Kolmogorov propone una Teoría de la Probabilidad donde esta propiedad empírica RESULTA un teorema (en un par de clases llega).
- Exito de la ciencia y el modelado matemático.

Motvación - Inspiración

- *Pretendemos construir una teoría donde $\mathbb{P}(A)$ represente el límite de las frecuencias relativas. porcentaje de veces que esperamos que A ocurra en* ***infinitas*** *repeticiones*
- Inspirados en este hecho, hacemos una propuesta y vemos como funciona.
- La propuesta de Kolmogorov funcionó!!!

Definición: Función de Probabilidad - Teoría axiomática de Kolmogorov.

1- $0 \leq \mathbb{P}(A) \leq 1$

2- $\mathbb{P}(\mathcal{S}) = 1$.

- Si $A_1 \cap A_2 = \emptyset$, entonces $\mathbb{P}(A_1 \cup A_2) \;=\; \mathbb{P}(A_1) + \mathbb{P}(A_2)\;^{(*)}$

3- $^{(*)}$ $A_1, A_2, A_3, \ldots$ disjuntos dos a ($A_i \cap A_j = \emptyset$ para $i \neq j$),

$$\mathbb{P}\left(A_1 \cup A_2 \cup A_3 \cup \ldots\ldots\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \ldots$$

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\ .$$

Esperanza- Motivación Frecuentista - Ejemplo

X: Ganancia	-1	35
$p_X(x)$: Probabilidad Puntual	36/37	1/37

- m_{-1} cantidad de veces que pierdo en m jugadas.
- m_{35} cantidad de veces que gano en m jugadas.

$$\frac{m_{-1}}{m} \longrightarrow p_X(-1)$$

$$\frac{m_{35}}{m} \longrightarrow p_X(35)$$

Una larga noche en el casino...

[1] −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[20] −1 −1 −1 −1 −1 −1 −1 −1 35 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[39] −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[58] −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[77] −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[96] −1 −1 −1 −1 35 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[115] −1 −1 −1 −1 −1 −1 35 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[134] −1 −1 −1 −1 −1 −1 −1 35 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

[153] −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1 −1

Ganancia media en $m = 29$ partidas

- Suma de los ganado: $r_1 + r_2 + \cdots + r_{29} = (-1) \times 28 + 35 \times 1$
- Ganancia promedio:

$$\frac{1}{29}(r_1 + r_2 + \cdots + r_{29)} = (-1) \times \frac{28}{29} + 35 \times \frac{1}{29}$$

- $m = 29,\ m_{-1} = 28,\ m_{35} = 1$

$$(-1) \times \frac{m_{-1}}{m} + 35 \times \frac{m_{35}}{m} \longrightarrow (-1) \times p_X(-1) + 35 \times p_X(35)$$

Esperanza - Definición

- X variable aleatoria: valor del experimento (X : Ganancia)
- x_i posibles valores que toma X. ($\{-1, 35\}$)
- $p(x_i)$ probabilidad de obtener el valor x_i
- $p(x_i) \geq 0$, $p(x_1) + p(x_2) + \cdots = \sum p(x_i) = 1$

Posibles valores	x_1	x_2	x_3	...
Probabilidad	$p_X(x_1)$	$p_X(x_2)$	$p_X(x_3)$	...

$$\begin{aligned} \mathbb{E}(X) &:= x_1 p_X(x_1) + x_2 p_X(x_2) + \ldots \\ &= \sum_{i \geq 1} x_i p_X(x_i) \end{aligned}$$

Vocabulario

$$\mathbb{E}(X) = \mu_X = \sum_{i \geq 1} x_i p(x_i) \quad (\mu)$$

- Esperanza
- Media
- *Valor esperado* – $\mathbb{E}(X)$ puede no estar en el rango de X.

Esperanza- Definición

Definición:

$$\mathbb{E}(X) \;=\; \sum_{i\geq 1} x_i \, p_X(x_i) \;,$$

siempre que $\sum_{i\geq 1} |x_i| \, p_X(x_i) < \infty$.

Lema (The Rule of the Lazy Statistician, L. W.)

$$\mathbb{E}[g(X)] = \sum_{i=i}^{\infty} g(x_i) P(X = x_i) \;.$$

Esperanza - Otra motivación:

- Perdida cuadrática.

$$H(a) \;=\; \sum_{i=1}^{k} (x_i - a)^2 \, p_X(x_i) \; .$$

- H se minimiza en

$$a \;=\; \sum_{i=1}^{k} \; x_i \; p_X(x_i) = \mathbb{E}(X) = mu_X \; .$$

- ¿Cuánto se paga por reemplazar a X por $\mu_X = \mathbb{E}(X)$?

$$H(\mu_X) \;=\; \sum_{i=1}^{k} (x_i - \mu_X)^2 \, p_X(x_i) \; .$$

$$H(\mu_X) \;=\; \mathbb{E}[(X - \mu_X)^2] \text{ VARIANZA}$$

Esperanza y Varianza de algunas v.a.

X	$\mathbb{E}(X)$	$\mathbb{V}(X)$
$\mathcal{B}(1,p)$	p	$p(1-p)$
$\mathcal{B}(n,p)$	np	$np(1-p)$
$\mathcal{P}(\lambda)$	λ	λ
$\mathcal{U}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
${\chi^2}_k$	k	$2k$

Se acuerdan de la Guía 5? Sigue tirando, sigue tirando

- Implementar la función `perseverancia_exito(p)` que dado un valor p emule el número de repeticiones necesarias hasta observar el primer éxito, siendo p la probabilidad de éxito en cada repetición.
- X: número de repeticiones necesarias hasta observar el primer éxito, siendo p la probabilidad de éxito en cada repetición.

$$X \sim ? \ , \quad \mathbb{E}(X) = ?$$

Se acuerdan de la Guía 5? Sigue tirando, sigue tirando

- Implementar la función `perseverancia_exito(p)` que dado un valor p emule el número de repeticiones necesarias hasta observar el primer éxito, siendo p la probabilidad de éxito en cada repetición.
- X: número de repeticiones necesarias hasta observar el primer éxito, siendo p la probabilidad de éxito en cada repetición.

$$X \sim ? \, , \quad \mathbb{E}(X) = ?$$

$$\texttt{perseverancia\_exito(p)} \equiv \texttt{rgeom(n=1,prob=p)}$$

$$p_X(k) = (1-p)^{k-1} p \, , \quad k \geq 1$$

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \, p_X(k) = \sum_{k=1}^{\infty} k \, (1-p)^{k-1} p = \frac{1}{p}$$

Se acuerdan de la Guía 5? Sigue tirando, sigue tirando

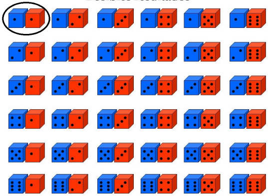
- Graficar p (en la grilla) vs el promedio de `perseverancia_exito(p)` en $Nrep = 1000$. Proponga alguna curva para modelar este fenómeno.
- X: número de repeticiones necesarias hasta observar el primer éxito, siendo p la probabilidad de éxito en cada repetición.

$$X \sim \mathcal{G}(p) \;, \quad \mathbb{E}(X) = 1/p$$

Suma de Variables Aleatorias

Suma de dados

$$X_i = \text{resultado del } i\text{-ésimo dado}, \; S = X_1 + X_2$$



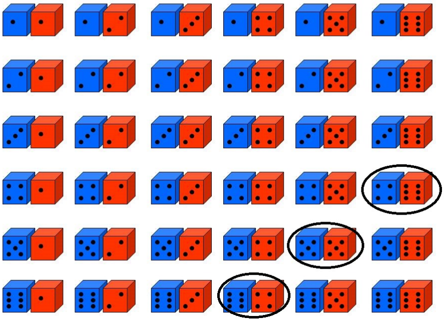
Los valores posibles de S son $\{2,3,4,5,6,7,8,9,10,11,12\}$.

$$\mathbb{P}(S=2) = \mathbb{P}(X_1 = 1 \cap X_2 = 1) \overset{indep}{=} \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}$$

$$\mathbb{P}(S=3) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 2) + \mathbb{P}(X_1 = 2)\mathbb{P}(X_2 = 1) = 2\frac{1}{6}\frac{1}{6} = \frac{2}{36}$$

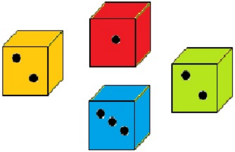
Suma de dados

Posibles resultados



s	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(S = s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

¿Cuán fácil sería resolver un problema análogo que involucrase la variable suma si tuviéramos 4 dados?



$$X_i = \text{resultado del } i\text{-éstimo lanzamiento}$$

$$X = \text{resultado del lanzamiento de un dado}$$

x	1	2	3	4	5	6
p_X	1/6	1/6	1/6	1/6	1/6	1/6

Revisitamos Guía vieja...

Crear una función **suma.cubilete** que simule arrojar el cubilete y dé por resultado la suma de las 5 caras obtenidas.

..... una aproximación vía simulación

Propuesta: simulemos la distribución de S

1. Simulemos el lanzamiento de 5 dados equilibrados.
2. Consideremos $S =$ la suma de las caras obtenidas.
3. Repitamos $Nrep = 10000$ y grafiquemos el histograma para los valores de S obtenidos.

..... una aproximación vía simulación

```r
dado=c(1,2,3,4,5,6)
proba=c(1/6,1/6,1/6,1/6,1/6,1/6)

sum(sample(dado,5,replace=T,prob=proba))

nrep=10000

set.seed(999)
suma=rep(0,nrep)


for(i in 1:nrep){
    suma[i]=sum(sample(dado,5,replace=T,prob=proba))
}

hist(suma,freq=F,main="Suma 5 dados")

mean(suma==18)
```

Suma de v.a.: algunos casos conocidos

Sean X e Y v. a. independientes y $S = X + Y$, entonces:

1. $X \sim \mathcal{B}(n, p)$ e $Y \sim \mathcal{B}(m, p) \Rightarrow S \sim \mathcal{B}(n + m, p)$.
2. $X \sim \mathcal{P}(\lambda_1)$ e $Y \sim \mathcal{P}(\lambda_2) \Rightarrow S \sim \mathcal{P}(\lambda_1 + \lambda_2)$.
3. $X \sim G(p)$ e $Y \sim G(p) \Rightarrow S \sim BN(2, p)$.
4. $X \sim BN(k_1, p)$ e $Y \sim BN(k_2, p) \Rightarrow S \sim BN(k_1 + k_2, p)$.

Propiedades

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
- X e Y indep. $\Rightarrow V(X+Y) = V(X) + V(Y)$
- X e Y indep. $\Rightarrow V(X-Y) = V(X) + V(Y)$

Generalizando

- $X_1, \ldots, X_n$, variables aleatorias, entonces

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}\left(X_i\right)$$

- $X_1, \ldots, X_n$ independientes, entonces

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V\left(X_i\right)$$