Simulaciones

Probabilidades

2023

Antes de empezar, los invitamos a visitar el sitio Point of Significance, una publicación de Nature dedicada a la divulgación de la estadística dentro de las ciencias naturales. En particular, los invitamos a que miren el trabajo Importance of being uncertain, considerando que vamos a querer replicar parte de los resultados presentados en la Figura 3.

A lo largo de esta Guía estudiaremos empíricamente la distribución del promedio de variables aleatorias independientes e idénticamente distribuidas. A través de los histogramas correspondientes, analizaremos el comportamiento de la distribución del promedio a medida que aumentamos n, la cantidad de variables a promediar.

Para ello generaremos un conjunto de n datos con una distribución dada y luego calcularemos su promedio. Replicaremos ésto mil veces, es decir, generaremos $Nrep = 1000$ realizaciones de la variable aleatoria $\overline{X}_n$, para diferentes valores de n. Observemos que, en principio, desconocemos la distribución de $\overline{X}_n$. Utilizando las $Nrep = 1000$ realizaciones del promedio realizaremos un histograma de los promedios generados para obtener una aproximación de la densidad o la función de probabilidad de $\overline{X}_n$.

1 Teorema: Ley de los Grandes Números

Ley de los Grandes Números establece que

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i \longrightarrow \mathbb{E}(W) \quad \text{en probabilidad }, \tag{1}$$

donde $(W_i)_{i\geq 1}$ son variables aleatorias i.i.d. En particular, si las variables tienen distribución Bernoulli, obtenemos que la frecuencia relativa de éxitos converge a su probabilidad. Más generalmente, si $(X_i)_{i\geq 1}$ son variables i.i.d. y A es un conjunto de número reales, podemos definir

$$Y_i = I_{X_i \in A} = I_A(X_i)$$

Es decir, Y_i vale 1 si X_i pertenece al conjunto A y 0 caso contrario. En tal caso, las variables Y_i tienen distribución Bernoulli, con $\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \in A)$, y por consiguiente, $\mathbb{E}(Y) = \mathbb{P}(X \in A)$. Invocando la Ley de los Grandes Números, utilizando Y_i en lugar de W_i, tenemos que

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \longrightarrow \mathbb{E}(Y) \quad \text{en probabilidad}$$

Es decir,

$$\frac{1}{n}\sum_{i=1}^{n} I_A(X_i) \longrightarrow \mathbb{P}(X \in A) \quad \text{en probabilidad} \tag{2}$$

Es decir, la frecuencia relativa converge a la probabilidad.

A lo largo de esta propuesta vamos a verificar que en cada uno de los casos considerados, el valor límite propuesto se condice con el indicado por la Ley de los Grandes Números.

2 Teorema Central del Límite

Antes de empezar con nuestras simulaciones, recordemos el Teorema Central del Límite. Sea $(W_i)_{i\geq 1}$ una sucesión de v.a.i.i.d. con $\mathbb{E}(W_i) = \mu$ y $\mathbb{V}(X_i) = \sigma^2$.

Una de las maneras de ver el TCL en términos del promedio es la siguiente:

$$\overline{W}_n \overset{a}{\approx} \mathcal{N}(\mu, \sqrt{\sigma^2/n}) \ . \tag{3}$$

Si estandarizamos al promedio, obtenemos esta otra posible presentación:

$$\frac{\overline{W}_n - \mu}{\sqrt{\sigma^2/n}} \overset{a}{\approx} \mathcal{N}(0,1) \ . \tag{4}$$

El obetivo es estudiar empiricamente la distribución de diferentes sumas y promedios.

2.1 La propuesta

A lo largo de esta guía contemplaremos las siguientes distribuciones: Bernoulli, Uniforme, Exponencial, Normal, y las denotaremos con `ber`, `unif`,`exp` y `norm`, respectivamente. Utilizaremos `tita` para p, (a, b), λ y (μ, σ), según corresponda. A lo largo de las simulaciones, utilizaremos los siguientes valores para el parametro `tita` en cada caso: `tita=0.2`, `tita=(67,73)` `tita=1/70` y `tita=(70,3)`.

1. Implementar una función llamada `promedios(N_infty,distribucion, tita)` que tenga por parámetros `N_infty`, representando la cantidad máxima de datos que se van a generar, una distribución con la que generar los datos y un vector `tita` con los parámetros asociados a la distribución indicada. Debe devolver un vector con promedios utilizando 1,2,..., `N_infty` datos generados bajo la distribución indicada, de forma tal que cada promedio se obtiene agregando un dato nuevo a los ya generados.

2. Invocando la funcion `promedios`, graficar el valor del promedio en función de la cantidad de datos utilizados para su computo. ¿Qué observa? ¿Se condice con lo que predice la teoría?

3. Implementar una función llamada `muchos_promedios` que agregue un parámetro `N_gente`, y devuelva una matriz donde cada fila tenga la salida de la función `promedios`.

4. Construir una matriz correspondiente a promedios de muchas personas. Realizar histogramas de las columnas y graficos de dispersion para diferentes filas. ¿Qué se observa?

5. Realizar ahora histogramas para algunas columnas, pero estandarizando los promedios, como se indica en la fórmula (4)