Trabajo Práctico 1 en `R`

Ley de los Grandes Números y Teorema Central del Límite

Segunda Parte: Teorema Central del Límite

Introducción

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*, dado que vamos a querer replicar parte de los resultados presentados en la Figura 3.

Más precisamente, en esta última sección 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 esto diez mil veces, es decir, generaremos $N_{rep} = 10000$ 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 $N_{rep} = 10000$ realizaciones del promedio realizaremos un histograma de los promedios generados para obtener una aproximación de la densidad o de la f.p.p. de $\overline{X}_n$.

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

Recordemos que el TCL establece que

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

donde $\approx$ significa “se distribuye aproximadamente como”.

Si despejamos el promedio, tenemos que

$$\overline{W}_n = Z_n \sqrt{\frac{\sigma^2}{n}} + \mu$$

de donde se deduce esta otra posible representación:

$$\overline{W}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) . \tag{2}$$

Actividad 3: Simulaciones para observar la distribución empírica de los promedios según diferentes distribuciones

Para lo que sigue, utilizaremos X para denotar de manera genérica una variable con distribución F. Es decir, $X \sim F$. Generaremos n datos con la distribución F, promediaremos y repetiremos $N_{rep} = 10000$ veces para luego observar la distribución empírica de dichos promedios, como postula la ecuación (2) y de los promedios estandarizados, como postula la ecuación (1).

Distribución Bernoulli

Sea $X \sim \mathcal{B}e(p)$ siendo $p = 0,2$.

- Indicar el valor de $\mathbb{E}(X)$ y de $\mathbb{V}(X)$.
- Guardar $N_{rep} = 10000$ realizaciones de X en el vector `ber_N_infty` con el comando `rbinom()` (prestar atención a los parámetros que utiliza esta función). Luego, obtener la frecuencia relativa de cada posible valor del rango de X. Se puede utilizar el comando `table()`. ¿Qué valores espera observar en esas frecuencias relativas? ¿Por qué?
- Calcular el promedio y la varianza muestral con los datos del vector `ber_N_infty` y comparar con los obtenidos en el primer ítem. ¿Qué observa?
- Realizar un histograma de los datos. ¿Qué observa? ¿Es consistente con la distribución a partir de la cual fueron generados los datos?

La distribución empírica del promedio

- Guardar en el vector `promedios_bernoullies`, $N_{rep} = 10000$ promedios utilizando $n = 5$ datos con distribución X.
- Con los promedios guardados en `promedios_bernoullies`, realizar un histograma de densidad. Para el histograma de densidad, usar el argumento `probability = TRUE` en el comando `hist()`. Además, como $n = 5$ puede ser pequeño, al usar `hist()` puede ajustar el parámetro `breaks`. Observar qué ocurre al poner (o no) ese parámetro. Luego, superponer el gráfico de la curva de la densidad normal con la media y desvío que considere razonables.
- Repetir los ítems 2 y 3 con $n = 30$ y con $n = 100$. Comparar los tres histogramas obtenidos y explicar qué observa en cada caso y a qué cree que se debe. Tener cuidado con la diferencia en las escalas.

La distribución empírica del promedio *estandarizado*

- Realizar ahora histogramas de densidad con los datos de `promedios_bernoullies_est` donde a los valores guardados en `promedios_bernoullies` se les resta μ y se los divide por $\sqrt{\sigma^2/n}$. Utilizar $n \in \{5, 30, 100\}$. ¿Qué observa? Explicar.
- A los histogramas de densidad del ítem anterior, superponer las curvas de la densidad normal con la media y desvío que considere adecuados. ¿Qué observa? Explicar.
- Repetir todos los ítems anteriores con $X \sim \mathcal{B}e(p)$ siendo $p = 0,5$. ¿Hay alguna diferencia? Comentar.

Distribución Uniforme

Sea $X \sim \mathcal{U}(a, b)$ siendo $a = 67$, $b = 73$.

- Indicar el valor de $\mathbb{E}(X)$ y de $\mathbb{V}(X)$.
- Guardar $N_{rep} = 10000$ realizaciones de X en el vector `unif_N_infty` con el comando `runif()`.
- Calcular el promedio y la varianza muestral con los datos del vector `unif_N_infty` y comparar con los obtenidos en el primer ítem. ¿Qué observa?
- Realizar un histograma de los datos. ¿Qué observa? ¿Es consistente con la distribución a partir de la cual fueron generados los datos?

La distribución empírica del promedio

- Guardar en el vector `promedios_uniformes`, $N_{rep} = 10000$ promedios utilizando $n = 5$ datos con distribución X.
- Con los promedios guardados en `promedios_uniformes`, realizar un histograma de densidad. Para el histograma de densidad, usar el argumento `probability = TRUE` en el comando `hist()`. Luego, superponer el gráfico de la curva de la densidad normal con la media y desvío que considere razonables.
- Repetir los ítems 2 y 3 con $n = 30$ y con $n = 100$. Comparar los tres histogramas obtenidos y explicar qué observa en cada caso y a qué cree que se debe. Tener cuidado con la diferencia en las escalas.

La distribución empírica del promedio *estandarizado*

- Realizar ahora histogramas de densidad con los datos de `promedios_uniformes_est` donde a los valores guardados en `promedios_uniformes` se les resta μ y se los divide por $\sqrt{\sigma^2/n}$. Utilizar $n \in \{5, 30, 100\}$. ¿Qué observa? Explicar.
- A los histogramas de densidad del ítem anterior, superponer las curvas de la densidad normal con la media y desvío que considere adecuados. ¿Qué observa? Explicar.

Distribución Exponencial

Sea $X \sim \mathcal{E}(\lambda)$ siendo $\lambda = 1/10$.

- Indicar el valor de $\mathbb{E}(X)$ y de $\mathbb{V}(X)$.
- Guardar $N_{rep} = 10000$ realizaciones de X en el vector `exp_N_infty` con el comando `rexp()`.
- Calcular el promedio y la varianza muestral con los datos del vector `exp_N_infty` y comparar con los obtenidos en el primer ítem. ¿Qué observa?
- Realizar un histograma de los datos. ¿Qué observa? ¿Es consistente con la distribución a partir de la cual fueron generados los datos?

La distribución empírica del promedio

- Guardar en el vector `promedios_exponenciales`, $N_{rep} = 10000$ promedios utilizando $n = 5$ datos con distribución X.
- Con los promedios guardados en `promedios_exponenciales`, realizar un histograma de densidad. Para el histograma de densidad, usar el argumento `probability = TRUE` en el comando `hist()`. Luego, superponer el gráfico de la curva de la densidad normal con la media y desvío que considere razonables.
- Repetir los ítems 2 y 3 con $n = 30$ y con $n = 100$. Comparar los tres histogramas obtenidos y explicar qué observa en cada caso y a qué cree que se debe. Tener cuidado con la diferencia en las escalas.

La distribución empírica del promedio *estandarizado*

- Realizar ahora histogramas de densidad con los datos de `promedios_exponenciales_est` donde a los valores guardados en `promedios_exponenciales` se les resta μ y se los divide por $\sqrt{\sigma^2/n}$. Utilizar $n \in \{5, 30, 100\}$. ¿Qué observa? Explicar.
- A los histogramas de densidad del ítem anterior, superponer las curvas de la densidad normal con la media y desvío que considere adecuados. ¿Qué observa? Explicar.

Distribución normal: un caso particular

Sean $X \sim \mathcal{N}(\mu, \sigma^2)$ siendo $\mu = 70$ y $\sigma^2 = 1{,}2$.

- Indicar el valor de $\mathbb{E}(X)$ y de $\mathbb{V}(X)$.
- Guardar $N_{rep} = 10000$ realizaciones de X en el vector `norm_N_infty` con el comando `rnorm()`.

- Calcular el promedio y la varianza muestral con los datos del vector `norm_N_infty` y comparar con los obtenidos en el primer ítem. ¿Qué observa?
- Realizar un histograma de los datos. ¿Qué observa? ¿Es consistente con la distribución a partir de la cual fueron generados los datos?

La distribución empírica del promedio

- Guardar en el vector `promedios_normales`, $N_{rep} = 10000$ promedios utilizando $n = 5$ datos con distribución X.
- Con los promedios guardados en `promedios_normales`, realizar un histograma de densidad. Para el histograma de densidad, usar el argumento `probability = TRUE` en el comando `hist()`. Luego, superponer el gráfico de la curva de la densidad normal con la media y desvío que considere razonables.
- Repetir los ítems 2 y 3 con $n = 30$ y con $n = 100$. Comparar los tres histogramas obtenidos y explicar qué observa en cada caso y a qué cree que se debe. Tener cuidado con la diferencia en las escalas.
- Explicar las diferencias que puede haber en las conclusiones para este apartado (distribución normal) y las de las distribuciones anteriores (Bernoulli, uniforme y exponencial).

La distribución empírica del promedio *estandarizado*

- Realizar ahora histogramas de densidad con los datos de `promedios_normales_est` donde a los valores guardados en `promedios_normales` se les resta μ y se los divide por $\sqrt{\sigma^2/n}$. Utilizar $n \in \{5, 30, 100\}$. ¿Qué observa? Explicar.
- A los histogramas de densidad del ítem anterior, superponer las curvas de la densidad normal con la media y desvío que considere adecuados. ¿Qué observa? Explicar.
- Explicar las diferencias que puede haber en las conclusiones para este apartado (distribución normal) y las de las distribuciones anteriores (Bernoulli, uniforme y exponencial).