1. Comprobación empírica de la LGN: generaremos muestras aleatorias de una distribución dada, de tamaño cada vez mayor, y calcularemos las medias muestrales. Veremos que estos promedios muestrales se acercan al valor verdadero de la media cuando el tamaño muestral crece.

   (a) Probamos primero con una distribución $\mathcal{U}(0,1)$. Indique cuál es el valor verdadero de la media μ .

   (b) Genere una muestra de tamaño $N = 1000$ y para cada k entre 1 y 1000 calculemos el promedio de las primeras k observaciones. Realizamos un scatterplot de k vs. los promedios obtenidos. Incluya una linea horizontal en el valor de y correspondiente a la verdadera media μ.

   (c) Repetimos b) comenzando con otra semilla y superpongamos el nuevo gráfico utilizando un color diferente. Comparamos los gráficos obtenidos. ¿ Qué podemos concluir?

   Generamos muestras de tamaño k cada vez mayor k entre 1 y 1000, calculamos sucesivamente los promedios y realizamos un scatter plot de k vs. los promedios obtenidos.

   (d) Probamos ahora con una distribución normal $\mathcal{N}(3,4)$. Generamos muestras de tamaño 10^k cada vez mayor, $k = 0$ a 7, calculamos sucesivamente los promedios y realizamos un scatter plot de k vs. los promedios obtenidos.

   (e) Repetimos el item anterior comenzando con otra semilla, superponiendo el gráfico con otro color, y comparamos los gráficos obtenidos. ¿ Qué podemos concluir?

   (f) Generamos ahora datos con una distribución de Cauchy. Recordemos que la distribución de Cauchy corresponde a una distribución t de student con un grado de libertad, con densidad dada por

   $$f_X(x) = \frac{1}{\pi(1+x^2)},$$

   que es una densidad simétrica alrededor del cero, con colas que acumulan más probabilidad que la normal estándar, y que no tiene esperanza ni varianza finitas. Generamos muestras de tamaño cada vez mayor $10^k, k = 0$ a 7, calculamos sucesivamente los promedios y realizamos un scatter plot de k vs. los promedios obtenidos. Comparemos con los resultados obtenidos en los ítems anteriores.

2. En este ejercicio estudiaremos la distribución del promedio $\bar{X}_n$ de variables $X_1, \ldots, X_n$ independientes e idénticamente distribuidas (i.i.d.), definido por

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A través de los correspondientes histogramas analizaremos el comportamiento de la distribución del promedio $\bar{X}_n$, a medida que promediamos un número creciente de variables aleatorias (n aumentando). Es decir, trataremos de validar empíricamente los resultados de la Ley de los Grandes Números y el Teorema Central del Límite.

Para ello, fijado n, generaremos datos correspondientes a una muestra $X_1, \ldots, X_n$ i.i.d. de variables aleatorias distribuídas como X, con una distribución dada y luego calcularemos el promedio de cada conjunto de datos. Repetimos este procedimiento $Nrep = 1000$ veces. A partir de las $Nrep = 1000$ replicaciones realizaremos un histograma con los promedios generados, para obtener una aproximación de la densidad o la función de probabilidad de $\bar{X}_n$.

(a) Consideremos $n = 1$: la variable coincide con el promedio. Generamos entonces $Nrep = 1000$ datos correspondientes a $X \sim \mathcal{U}(0; 1)$ y luego hacemos un histograma.

¿A qué densidad se parece el histograma obtenido?

(b) Consideramos $n = 2$ variables aleatorias X_1 y X_2 independientes con distribución $\mathcal{U}(0; 1)$ y el promedio de ambas, es decir,

$$\bar{X}_2 = \frac{X_1 + X_2}{2} .$$

Genermos $n = 2$ datos (independientes) correspondientes a una variable aleatorias con distribución $\mathcal{U}(0; 1)$ y computamos el promedio. Replicamos $Nrep = 1000$ veces y realizamos un histograma con los $Nrep = 1000$ promedios obtenidos.

¿Qué características tiene este histograma?

(c) Aumentamos a $n = 5$ las variables promediadas. Consideramos ahora 5 variables aleatorias uniformes independientes,es decir $X_1, X_2, \ldots, X_5$ i.i.d. con $X_i \sim \mathcal{U}(0; 1)$ y definimos elpromedio de las mismas

$$\bar{X}_5 = \frac{X_1 + X_2 \cdots + X_5}{5} .$$

Generamos datos de 5 variables aleatorias con distribución $\mathcal{U}(0; 1)$ computamos el promedio. Repetimos $Nrep = 1000$ veces y realizamos un histograma para los valores obtenidos. Comparamos con el histograma anterior.

¿Qué se observa?

(d) Aumentamos aún más la cantidad de variables promediadas. Generando $n = 30$ datos con distribución $\mathcal{U}(0; 1)$ repetimos el ítem anterior.

¿ Qué se observa?

(e) Idem anterior generando $n = 500$ datos.

¿ Qué pasa si aumentamos el tamaño de la muestra?

Observemos que para poder comparar los histogramas de los distintos conjuntos de datos será necesario tenerlos dibujados en la misma escala tanto para el eje horizontal como para el vertical. Por eso, en general es más cómodo hacer boxplots para comparar distintos conjuntos de datos.

(f) Finalmente, lo hacemos también para $n = 1200$, y graficamos un boxplot de los 6 conjuntos de datos en el mismo gráfico. En este gráfico se verá que a medida que aumenta el n, o sea el tamaño muestral, los valores de los promedios tienden a concentrarse, ¿alrededor de qué valor?

Calculamos media y varianza muestral para cada conjunto de datos.

¿A qué valores teóricos deberían parecerse?

(g) El Teorema Central del Límite nos dice que cuando hacemos la siguiente transformación con los promedios,

$$\frac{\bar{X}_n - E(X_1)}{\sqrt{Var(\bar{X}_n)}} = \frac{\bar{X}_n - E(X_1)}{\sqrt{Var(X_1)/n}}$$

la distribución de estas variables aleatorias se aproxima a la de la normal estándar, cuando n es suficientemente grande.

Para comprobarlo empíricamente, hagamos esta transformación en los 6 conjuntos de datos (es razonable hacerlo para valores de n suficientemente grandes, lo realizaremos en todos los casos para comparar) y luego comparemos los datos transformados mediante histogramas y boxplots.

(h) Repitamos los ítems anteriores generando ahora variables con distribución Cauchy C(0; 1): Comparemos los resultados obtenidos.