Estadística (Q)
Ejercicio para las clases prácticas de la tercer semana (C), teorema central del límite

---

En este ejercicio estudiaremos la distribución del promedio de variables independientes e idénticamente distribuidas y a través de los histogramas correspondientes analizaremos el comportamiento de estas distribuciones a medida que promediamos un número creciente de variables aleatorias. Es decir, trataremos de validar empíricamente los resultados de la Ley de los Grandes Números y el Teorema Central del Límite. Acompaña el archivo con instrucciones: `script_tcl.R`

Para ello generaremos una muestra de variables aleatorias con una distribución dada y luego calcularemos el promedio de cada muestra. Replicaremos ésto mil veces, es decir, generaremos una muestra aleatoria de la variable $\overline{X}_n$ de tamaño $B = 1000$. Observe que, en principio, desconocemos la distribución de $\overline{X}_n$. A partir de todas las replicaciones realizaremos un histograma para los promedios generados para obtener una aproximación de la densidad o la función de probabilidad de $\overline{X}_n$.

a) Comencemos por tomar un primer conjunto de datos de variables aleatorias $X_1, \ldots, X_{1000}$ independientes con distribución $U(0, 1)$. Le pedimos al `R` que nos genere una muestra de ellas y luego hacemos un histograma. ¿A qué densidad se parece el histograma obtenido?

b) Considerar dos variables aleatorias X_1 y X_2 independientes con distribución $U(0, 1)$ y el promedio de ambas, es decir,

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

Generando una muestra de dos variables aleatorias con distribución $U(0, 1)$ computar la variable promedio. Replicar $B = 1000$ veces y a partir de los valores replicados realizar un histograma. ¿Qué características tiene este histograma?

c) Aumentemos a cinco las variables promediadas. Considerar ahora 5 variables aleatorias uniformes independientes,es decir $X_1, X_2, \ldots, X_5$ i.i.d. con $X_i \sim U(0, 1)$ y definir

$$\overline{X}_5 = \frac{1}{5}\sum_{i=1}^{5} X_i\,.$$

Generando muestras de cinco variables aleatorias con distribución $U(0, 1)$ computar la variable promedio. Repetir $B = 1000$ veces y realizar un histograma para los valores obtenidos. Comparar con el histograma anterior. ¿Qué se observa?

d) Aumentemos aún más la cantidad de variables promediadas. Generando muestras de $n = 30$ variables aleatorias con distribución $U(0, 1)$ repetir el ítem anterior. ¿Qué se observa?

e) Ídem anterior generando muestras de $n = 500$ variables aleatorias. ¿Qué pasa si se aumenta el tamaño de la muestra? Observar 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 hacerlo también para 1200, y hacer un boxplot de los 6 conjuntos de datos en el mismo gráfico. En este gráfico se verá que a medida que aumenta el n los valores de los promedios tienden a concentrarse, ¿alrededor de qué valor? Calcule media y varianza muestral para cada conjunto de datos. ¿Puede dar los valores teóricos a los que deberían parecerse? Realice un qqplot para cada uno de los 6 conjuntos de datos? ¿Son esperables los resultados?

g) El teorema central del límite (TCL) nos dice que cuando hacemos la siguiente transformación con los promedios, $\frac{\overline{X}_n - E(X_1)}{\sqrt{\frac{Var(X_1)}{n}}}$, 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.

Repetir los ítems anteriores generando ahora variables con distribución t de Student con un grado de libertad (t_1). Comparar los resultados obtenidos. La densidad de una t_1 es simétrica alrededor del cero, con colas que acumulan más probabilidad que la normal estándar, y que no tiene esperanza ni varianza finitas. Su densidad está dada por

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