1. Se compararon tres dietas respecto al control de azúcar en la sangre en pacientes diabéticos. En el archivo `estad descriptiva.txt` se encuentran los valores de glucosa para las tres dietas consideradas (A, B, C), que contienen las lecturas de glucosa en la sangre de los pacientes. Es deseable que el paciente tenga valores entre 80 — 110 mg/dl.

   (a) Cargue los datos al R.

   (b) Para cada una de las tres dietas calcule medidas de centralidad: la media, la mediana, la media α-podada para $\alpha = 0.1, 0.2$. Para cada dieta compare los valores obtenidos de las cuatro medidas de posición, si observa una notable diferencia ¿a que podría deberse?

   (c) Calcule medidas de dispersión: el desvío estándar, la distancia intercuartil (o intercuartos) y la MAD en cada una de las dietas. Compare los valores de dispersión obtenidos, si observa una notable diferencia ¿a que podría deberse? ¿Cuál de las dietas parece ser la más estable?

   (d) Obtenga los percentiles 10, 25, 50, 75 y 90. Compare los valores de los percentiles obtenidos entre las distintas dietas.

   (e) Construya histogramas que permitan visualizar los valores de glucosa para cada dieta. Compare la distribución de glucosa. ¿Alguna de ellas parece bimodal? ¿En alguna de ellas parece haber valores alejados? ¿Las dietas mantienen a los pacientes en los valores deseados? ¿La distribución de glucosa es asimétrica en alguno de los grupos? ¿En algún caso el ajuste normal parece razonable? Realice los diagramas de tallo-hoja correspondientes.

   (f) Grafique los box-plots correspondientes. ¿Cómo se compara la información que dan estos gráficos con la obtenida con los histogramas? En base a los gráficos obtenidos, discuta simetría, presencia de outliers y compare dispersiones nuevamente.

   (g) Grafique los qqplots correspondientes. ¿En algún caso el ajuste normal parece razonable?

   (h) ¿En base al análisis anterior, cuál le parece la dieta más aconsejable?

2. 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.

   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 $\overline{X}$ de tamaño 1000. Observe que, en principio, desconocemos la distribución de $\overline{X}$. 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 $\overline{X}$.

   (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,

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

   Generando una muestra de dos variables aleatorias con distribución $U(0, 1)$ computar la variable promedio. Replicar 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

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

   Generando muestras de cinco variables aleatorias con distribución $U(0, 1)$ computar la variable promedio. Repetir 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 30 variables aleatorias con distribución $U(0,1)$ repetir el ítem anterior. ¿Qué se observa?

(e) Ídem anterior generando muestras de 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 1000, 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 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.

(h) Repetir los ítems anteriores generando ahora variables con distribución $\mathcal{C}(0,1)$. Comparar los resultados obtenidos. Recordar que la densidad de una Cauchy es

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

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.