Probabilidad y Estadística (C)

Trabajo en laboratorio

#PARA MANDAR TP NAHUEL.I.ARCA@GMAIL.COM

Enunciado 1. A orillas del río Reconquista yacen numerosas industrias, de las cuales el %70 no cumple con alguna de las normas establecidas para arrojar residuos al río. Un inspector visita 30 de ellas.

(a) Estimar la probabilidad de que más de 18 estén en infracción. Estimar cuántas industrias debe visitar el inspector para que tal probabilidad sea mayor a 0.95.

La probabilidad de que más de 18 estén en infracción es de 0.84 aproximadamente.

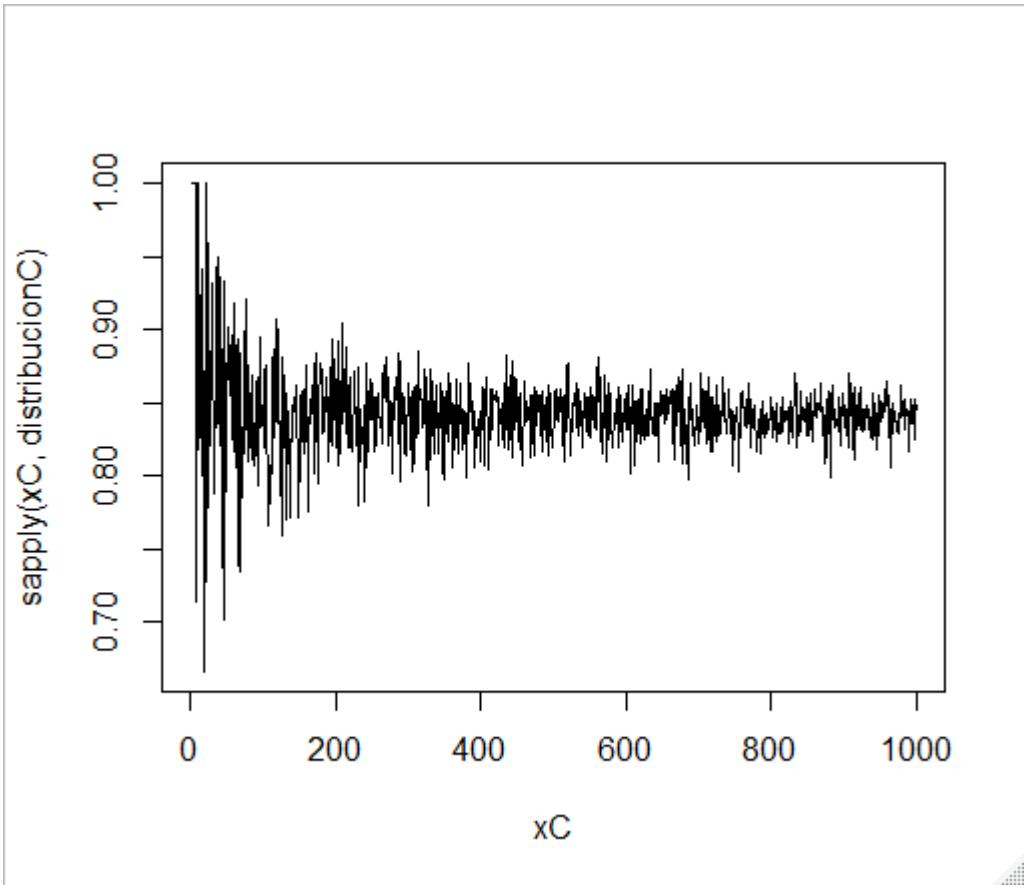
Para que tal probabilidad sea mayor a 0.95, el inspector debería visitar al menos 33 industrias.

(b) Resolver el ítem anterior usando R y calcularlo de manera exacta.

Según R, la probabilidad exacta de que más de 18 estén en infracción es de 0.8406782.

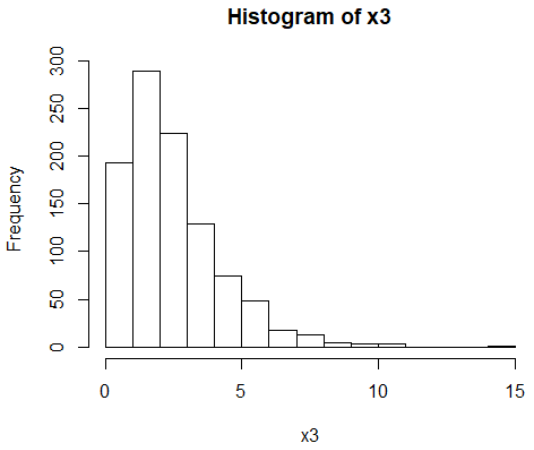
Para asegurarse de que haya 18 industrias en infracción con un 95% de probabilidad, el inspector necesita visitar al menos 21 industrias.

(c) Graficar la convergencia de la primera estimación.

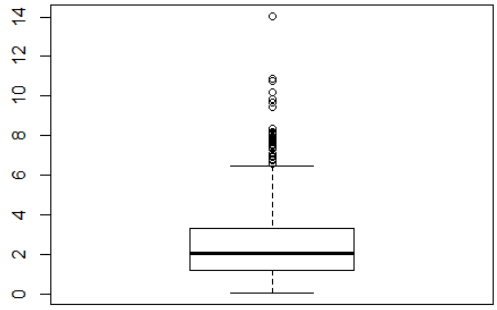


Enunciado 2. Para apreciar aún un poco más la Ley de los Grandes Números, realizar el siguiente experimento:

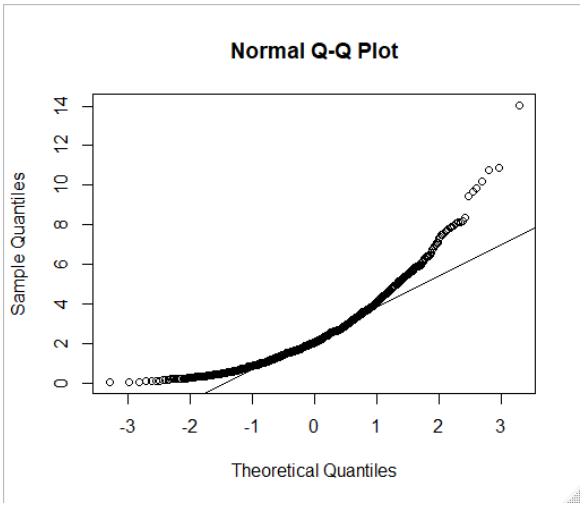
(a) Considerar dos observaciones x1 y x2 de variables aleatorias X1 y X2 independientes con distribución E(λ) (para algún λ a elección) y guardar el promedio de ambas, es decir, x2. Repetir 1000 veces y a partir de los valores obtenidos realizar un histograma, un box-plot y un QQ-plot. ¿Qué características tienen?



El histograma presenta una concentración de valores en el lado izquierdo, acercándose al cero, con una cola hacia derecha.



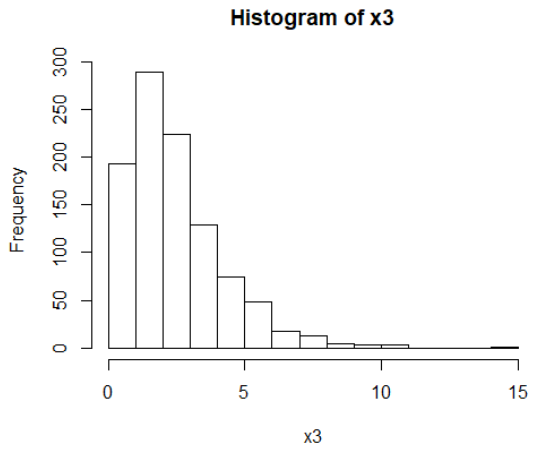
El box-plot, de manera similar, muestra una concentración hacia los valores cercanos al cero (estando siempre positivo), pero con una media más cercana al dos. Asimismo, el tercer cuartil se halla más alejado de la media que el primer cuartil.

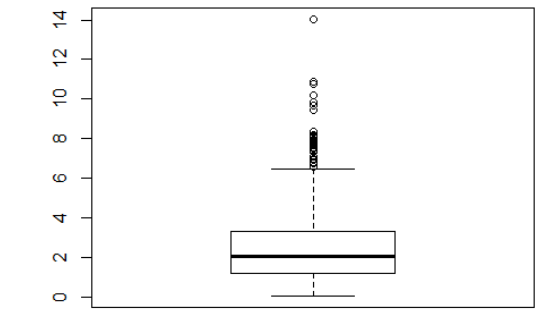


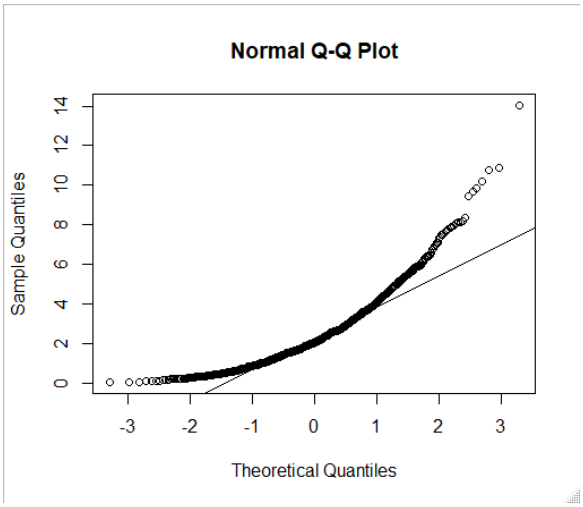
Para terminar con las observaciones, al comparar nuestra curva con la recta de una normal, la misma presenta una separación hacia arriba de ambos lados de la cola, lo cual se extrapola en una concentración hacia izquierda, y una descentralización a derecha.

(b) Aumentar a cinco las variables promediadas, es decir, considerar ahora n = 5 observaciones de variables aleatorias independientes con la misma distribución del ítem anterior y guardar x5. Repetir 1000 veces y realizar un histograma, un box-plot y un QQ-plot para los valores obtenidos. Comparar con los obtenidos en el ítem anterior. ¿Qué se observa?

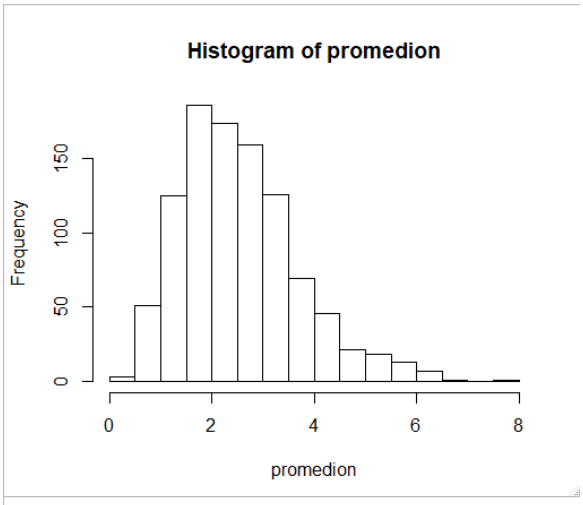
Con n = 2

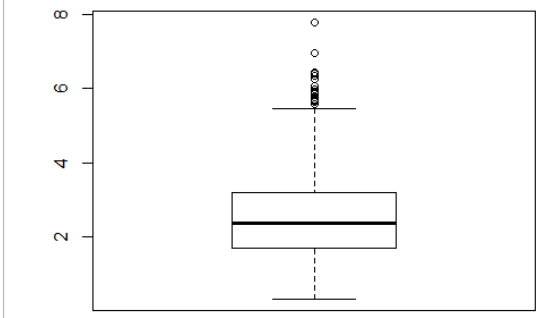


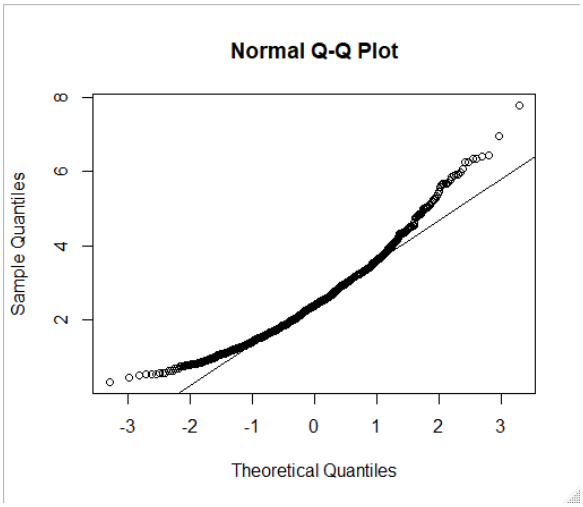




Con n = 5



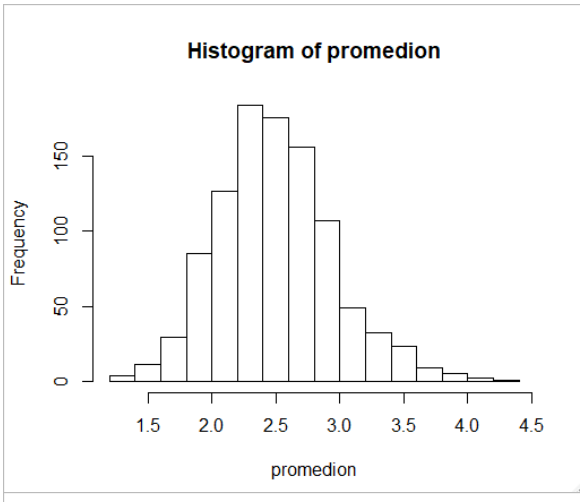


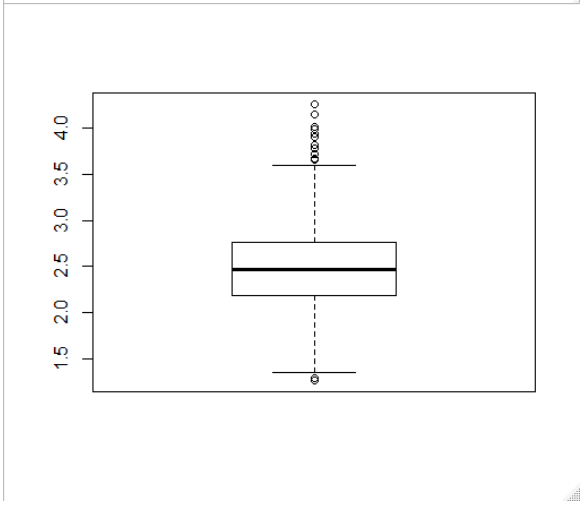


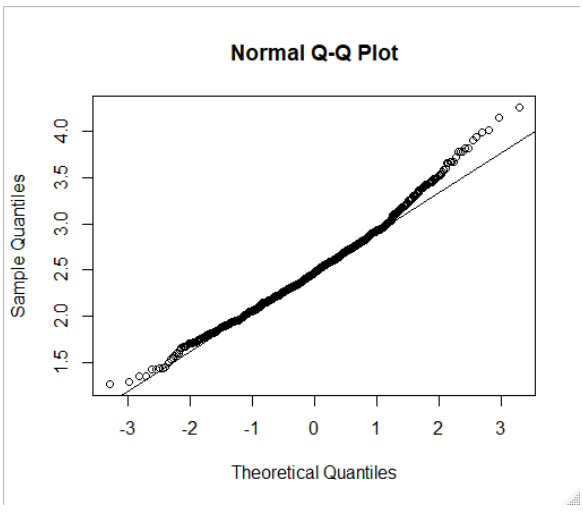
Creo que lo que más hay que observar son los valores que acompañan las gráficas, pues muestran que la concentración de valores se volvió mas pronunciada. Asimismo, se presentan menor cantidad de excepciones y, conjuntamente, los mismos tienen valores mas cercanos a la media.

(c) Aumentar a n = 30 el número de observaciones de v.a.i.i.d. y repetir el ítem anterior. Repetir con n = 500.

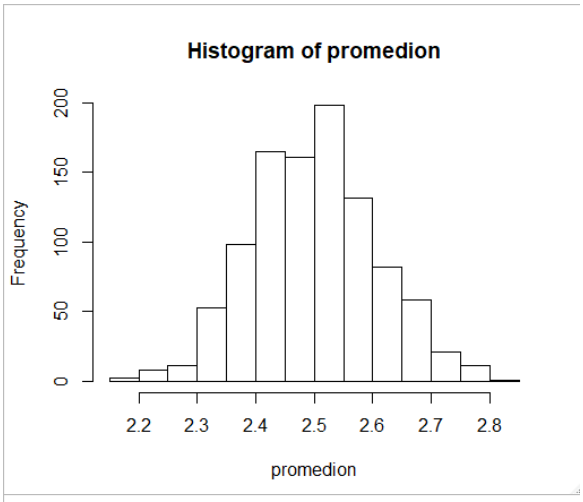
n = 30

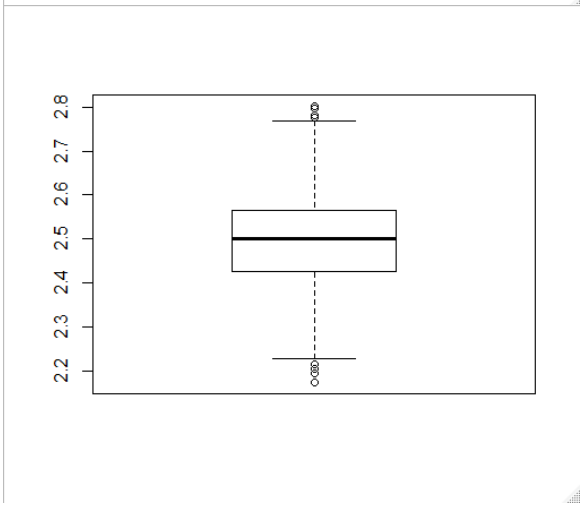


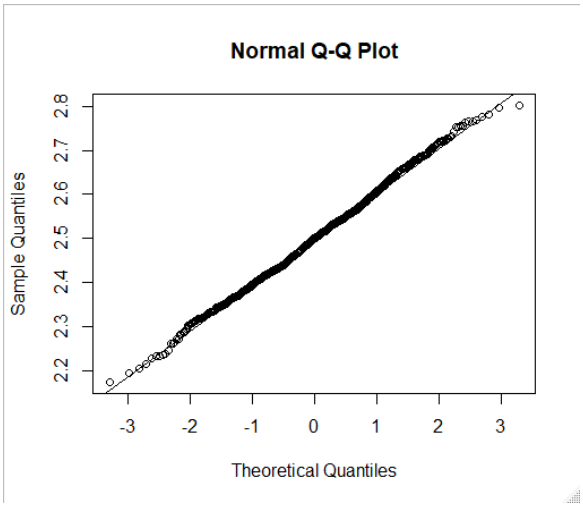




n = 500



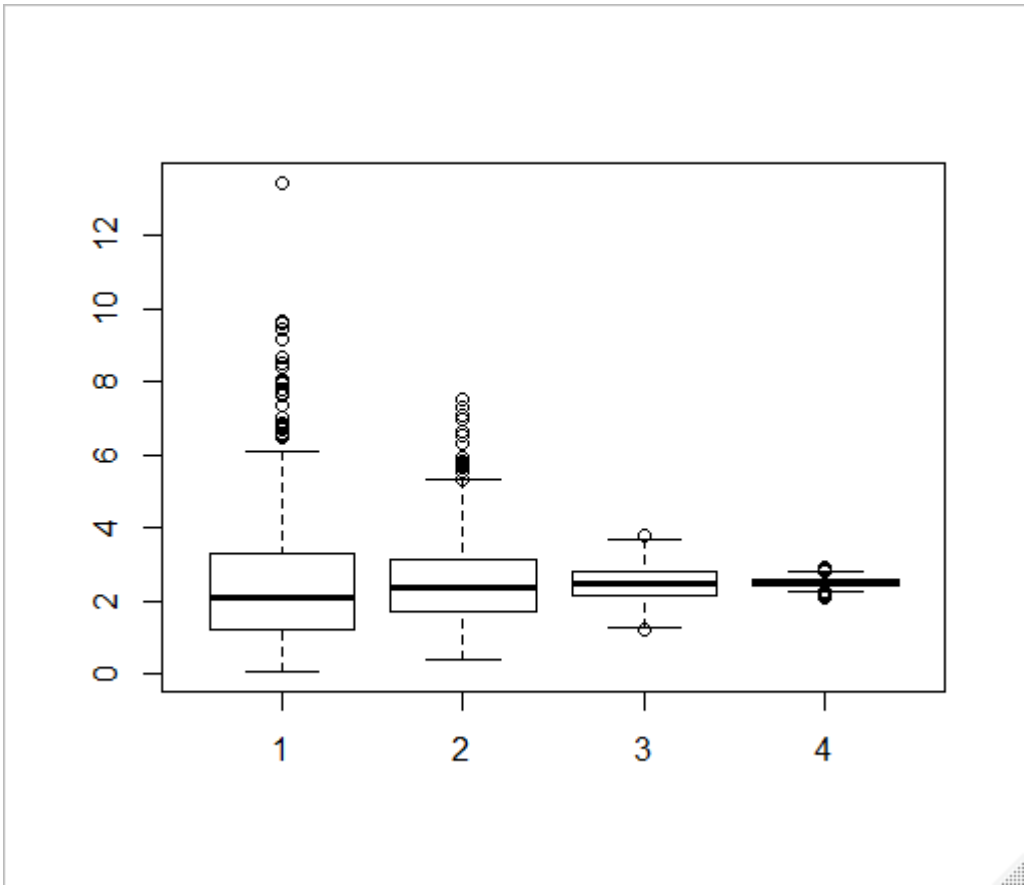




(d) ¿Qué pasaría si se siguiera aumentando el tamaño de la muestra?

Los gráficos presentarían dos características: Los valores se irían concentrando mas alrededor de la media, y las gráficas en si se irían pareciendo más al de una normal.

(e) Por último, hacer un box-plot de los 4 conjuntos de datos en el mismo gráfico (es decir, "box-plots paralelos").



Enunciado 3. El teorema central del límite nos dice que cuando hacemos la siguiente transformación con los promedios:

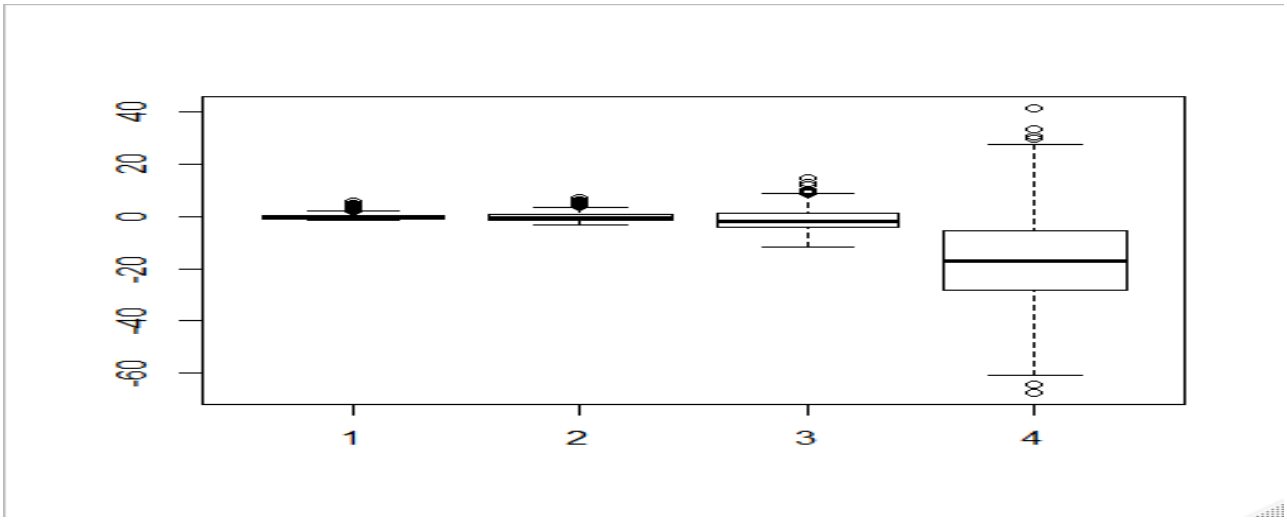
$$\frac{\overline{X_n} - E(X_1)}{\sqrt{\frac{\mathrm{Var}(X_1)}{n}}}$$

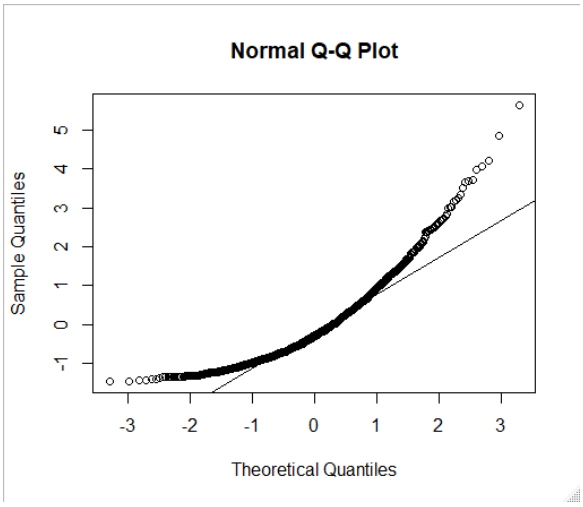
la distribución de esta variable aleatoria se aproxima a la de la normal estándar si n es suficientemente grande. Comprobaremos mediante una simulación este resultado.

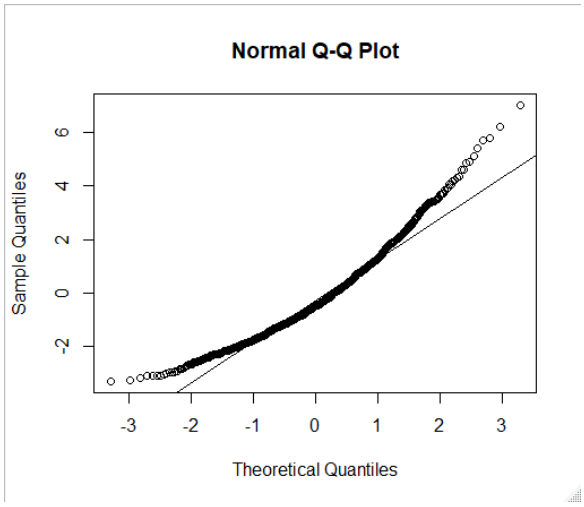
(a) Calcular la esperanza y varianza de X1 donde X1 es la misma distribución que en el ejercicio 2.

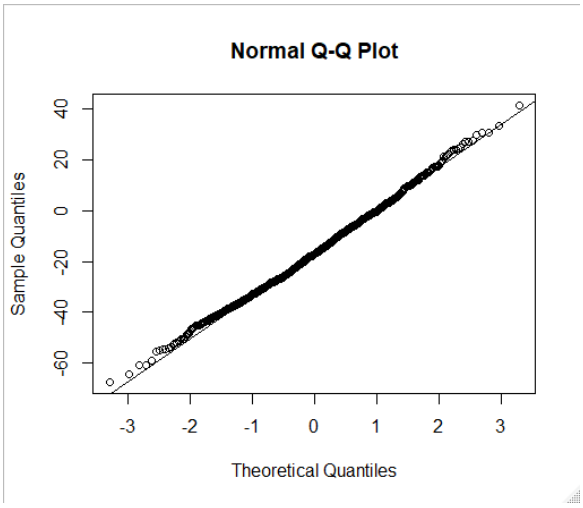
La esperanza me da 2.412958, mientras que la varianza me da 5.795832 (calculé la varianza y no la cuasi-varianza).

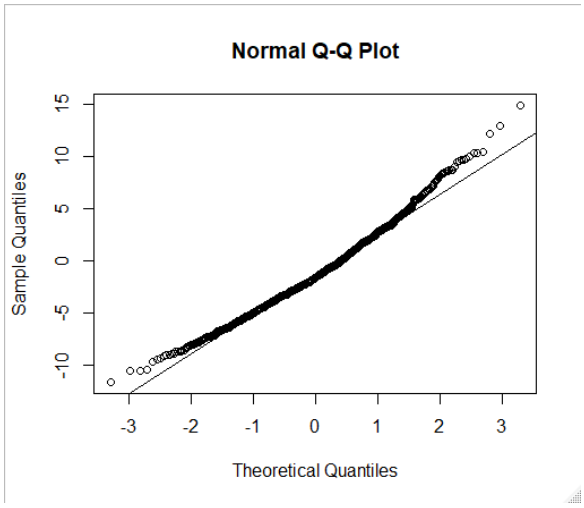
(b) Realizar la transformación mencionada en los 4 conjuntos de datos del ítem 2 y graficar box-plots paralelos y QQ-plots.



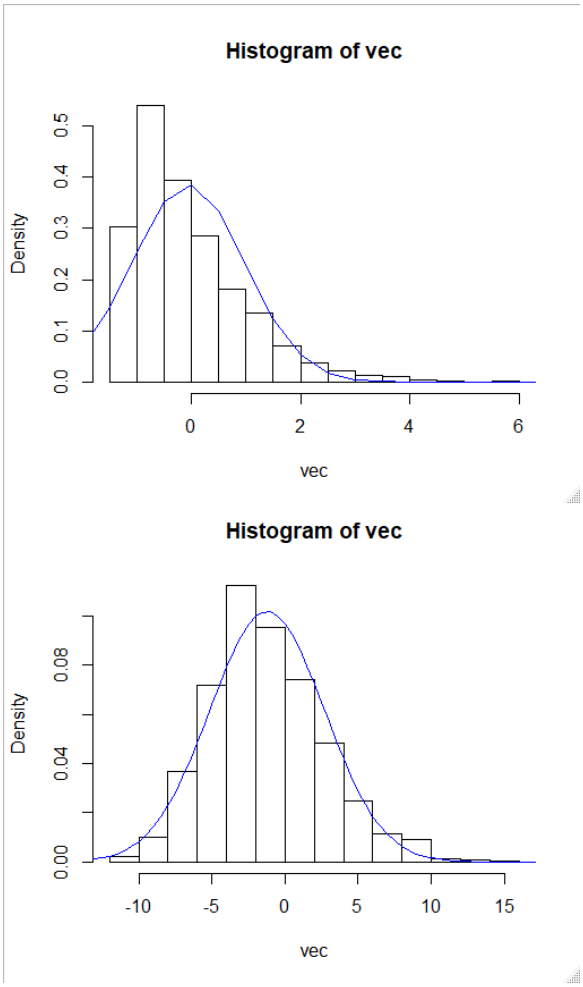








(c) Realizar 4 histogramas y a cada uno de ellos superponerle la densidad de la normal estándar.



(d) Explicar los resultados obtenidos.

A medida que aumenta el n, las graficas se van asemejando más a la densidad de una normal estándar.

Enunciado 4. Sea (Ui)i∈N una sucesión de variables aleatorias uniformes en [0, 1]. Definimos N = ínf{n ∈ N : SUM[i=1;n](Ui) ≥ 1}. Realizar simulaciones de la variable aleatoria N y estimar E(N).

Mi estimación de E(N) = `2.569.`

Enunciado 5. 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 α = 0.1, 0.2. Para cada dieta compare los valores obtenidos de las cuatro medidas de posición, si observa una notable diferencia ¿a qué podría deberse?

De la dieta-a:
- media: 98.68
- mediana: 99
- media 0.1-podada: 98.525
- media 0.2-podada: 98.5333

De la dieta-b:
- media: 94.46
- mediana: 94.5
- media 0.1-podada: 94.5
- media 0.2-podada: 94.4

De la dieta-c:
- media: 77.29
- mediana: 73
- media 0.1-podada: 75.3
- media 0.2-podada: 74.5

Si bien entra las dos primeras dietas no hay una notable diferencia, no es así al considerar la dieta-c. Es posible que esto se deba a una distribución dispareja de sus valores, lo cual se puede observar por la diferencia entre la media y la mediana de la misma, que presentan una diferencia bastante mas grande que las vistas en las otras dietas. Esto probablemente se deba a la presencia de valores extremos.

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

De la dieta-a:
- desvío estándar: 10.18404
- distancia inter-cuartil: 14.5
- MAD: -6.82121*e^-15

De la dieta-b:
- desvío estándar: 16.74552
- distancia inter-cuartil: 31.25
- MAD: 6.252776*e^-15

De la dieta-c:
- desvío estándar: 10.58329
- distancia inter-cuartil: 11
- MAD: -6.252776*e^-15

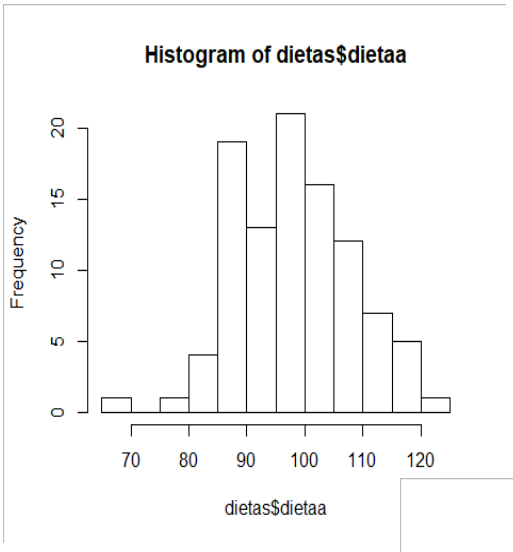
La mayor diferencia se nota en los valores de la dieta-b, que tiene un mayor desvío estándar.
La más estable parecería ser la dieta-a, dado que presenta menor desvío estándar. La dieta-c parecería estar más concentrado en valores (por su poca distancia inter-cuartil).
Las MAD son muy similares.

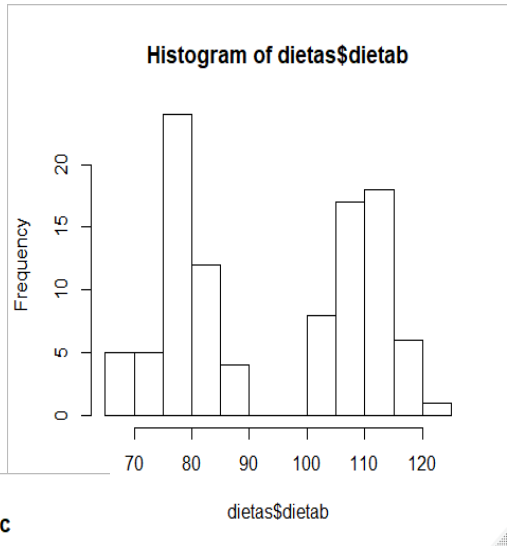
(d) Obtenga los percentiles 10, 25, 50, 75 y 90. Compare los valores de los percentiles obtenidos entre las distintas dietas.

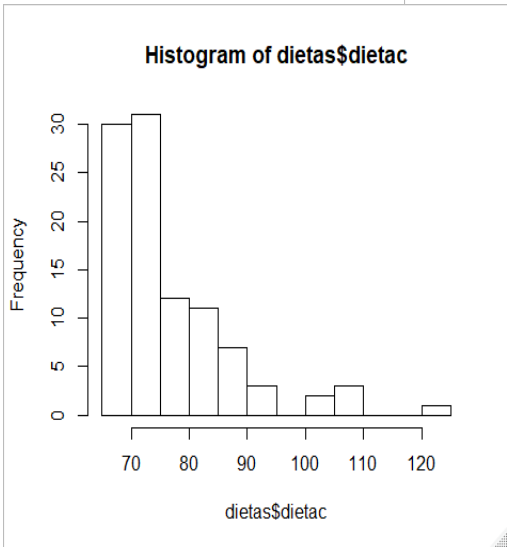
De la dieta-a:	De la dieta-b:	De la dieta-c:
10-percentil: 86	10-percentil: 75.8	10-percentil: 68
25-percentil: 90.75	25-percentil: 79	25-percentil: 70
50-percentil: 99	50-percentil: 94.5	50-percentil: 73
75-percentil: 105.25	75-percentil: 110.25	75-percentil: 81
90-percentil: 111.1	90-percentil: 115	90-percentil: 90

Los percentiles muestran más claramente lo expuesto anteriormente en las conclusiones de los ejercicios, como la poca variación en las dietas a y c, al 'podarle' sus extremos.

(e) Construya histogramas que permitan visualizar los valores de glucosa para cada dieta. Compare la distribución de glucosa.







En alguna de ellas ¿parece haber valores alejados?

Si. En realidad, en todas. (se cuentan: valores por encima de los 120 y aquellos menores a 70).

¿Las dietas mantienen a los pacientes en los valores deseados?

Solo la dieta A parece mantener notablemente los valores entre 80 y 110.

¿La distribución de glucosa es asimétrica en alguno de los grupos?

Si. La misma se ve muy tirada a izquierda para la dieta c.

¿En algún caso el ajuste normal parece razonable? Realice los diagramas de tallo-hoja correspondientes.

Si; los valores de glucosa de aquellos que siguen la dieta A parecería estar siguiendo una distribución normal.

Dieta-A

```
  6 | 8
  7 |
  8 | 011456666667777888899
  9 | 000112233455555566667777788999999
 10 | 000001122222244445555668888999
 11 | 000111234566899
 12 | 1
```

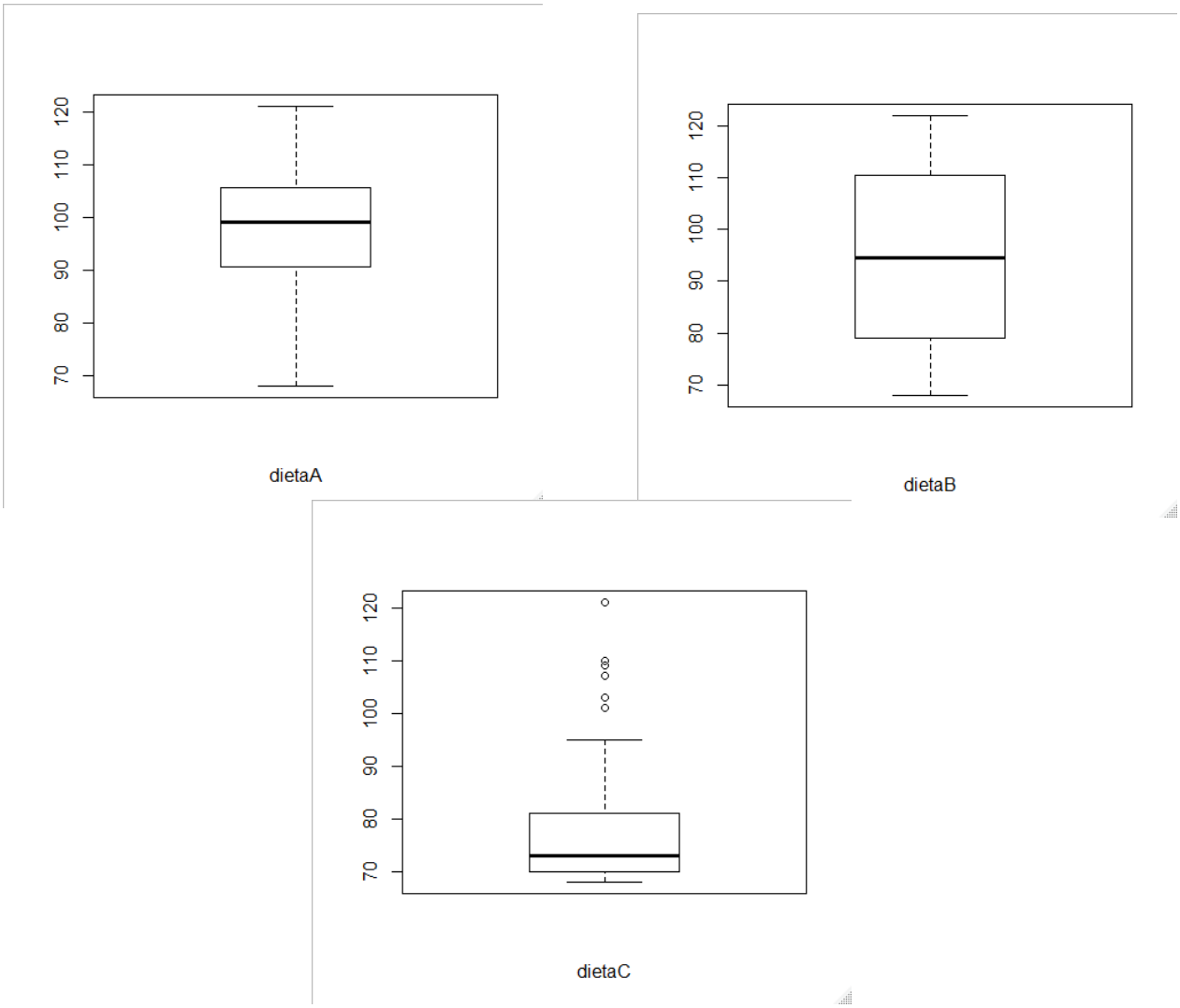
Dieta-B

```
  6 | 889
  7 | 002334466666777788899999
  8 | 00000001222233333346678
  9 |
 10 | 11244455666666678899999
 11 | 000111111222234455555778899
 12 | 2
```

Dieta-C

```
  6 | 88888888888999999999
  7 | 000000000011111112222222233333333444455556788999 9
  8 | 0000111222344556679
  9 | 000455
 10 | 1379
 11 | 0
 12 | 1
```

(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.

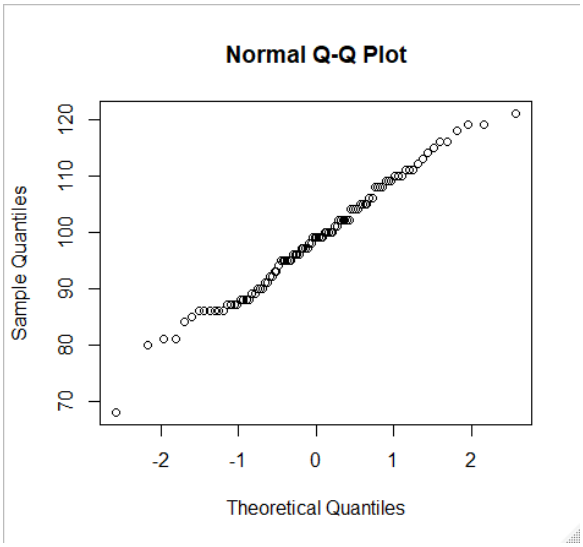
Nuevamente se ve que la dieta-a presenta similitudes con una distribución normal.

De la dieta-b, se sigue viendo la simetría, aunque es más fácil ver en el histograma la 'trampa' que tienen sus datos, escondiendo el hecho de que hay tan pocos valores en su centro.

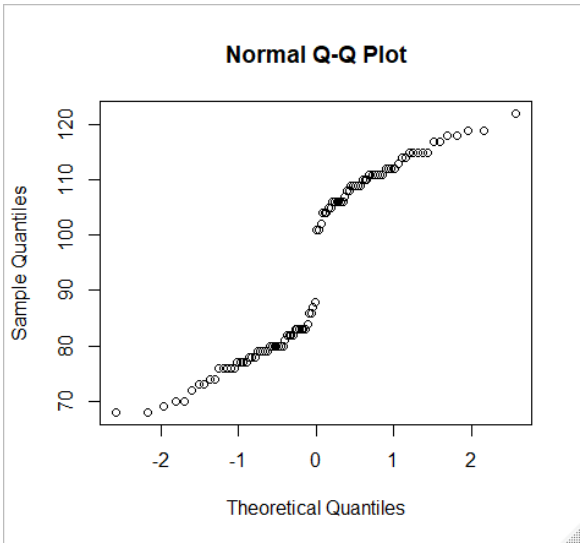
En el box-plot de la dieta-c se ve muy bien la asimetría presente, así como la presencia de valores muy alejados en los extremos.

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

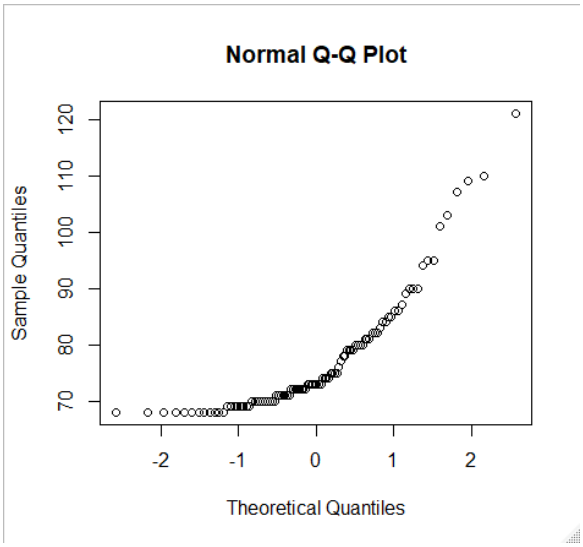
Dieta-A



Dieta-B



Dieta-C



Nuevamente, en el qq-plot normal~dieta-A se ven las similitudes, mientras que no tanto asi en los otros dos qq-plots.

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

Basado en todo lo visto, la dieta mas recomendable parecería ser la dieta a.