Estadística descriptiva

Nicolás Villagrán dos Santos

29 de octubre de 2020

Estadística descriptiva: buscamos describir un conjunto de datos ("muestra") para estudiar sus propiedades. La muestra es tomada de entre toda la población, de la que no conocemos la distribución de las variables de interés, ni sus parámetros correspondientes.

Variables: aquella propiedad que medimos (consumo de carne por habitante por año, si una persona fuma o no, sueldo, edad, altura, a quién va a votar): "lo que medimos".

Datos: los valores que se obtienen al estudiar estas variables sobre la muestra: "lo que da la medición".

Métodos gráficos: buscar visualizar la información asociada a la muestra.

Métodos resumen: dar un resumen numérico (parámetros, si son poblacionales, o estadísticos, si son muestrales).

Podemos tener datos categóricos o numéricos. Hoy trabajaremos con numéricos.

Los estadísticos que usaremos nos darán una idea de la posición o la centralidad de los datos (media, mediana, percentiles, etc.) y la dispersión de los datos (desvío muestral, IQR, MAD). Hay otros que se pueden usar para estudiar si las colas de la distribución son pesadas o no tanto (curtosis), y para evaluar asimetría estadística ("skewness").

En cuanto a métodos gráficos, haremos histogramas, boxplots y qqplots. No haremos diagramas de hoja y tallo.

En bioinformática, se emplea una técnica de simulación llamada "docking" para evaluar fácilmente si moléculas de interés se podrían unir o no a proteínas u otras moléculas "blanco". Esto sirve para seleccionar de entre miles de candidatos posibles para el desarrollo de nuevos fármacos, por ejemplo, sólo aquellas moléculas que sean más promisorias.

Ver por ejemplo: Arcon et al, 2017, https://pubs.acs.org/doi/10.1021/acs.jcim.6b00678.

Para evaluar esto, se emplean funciones de "scoring", que intentan predecir la afinidad de estas moléculas por una sección blanco en una proteína de interés. Para evaluar si una función de scoring es adecuada o no, se deben calibrar con moléculas de las que ya se conozcan los parámetros adecuados. En este caso, se comparan tres funciones de scoring, evaluadas sobre una base de datos con 100 moléculas ya ensayadas. Es deseable que:

- en general, para las moléculas de interés, estas funciones tengan un valor que no diste de 1 en más de $\pm 0,1$ (1 representa una afinidad de referencia).
- los valores deben tener una distribución aproximadamente normal (que apoyaría la hipótesis de que los valores de scoring sólo tienen un error aleatorio, y no hay ningún sesgo).

Evaluaremos cuál de las tres funciones cumple mejor con estos requerimientos.

1) *Calcular medidas de centralidad para los valores de cada una de estas funciones de scoring ensayadas sobre el banco de datos de referencia: - media, - mediana, - media α-podada para $\alpha = 0.1, 0.2$.*

Comparar los valores obtenidos para cada función. ¿Qué diferencias observa? ¿Hay alguna función de scoring que podríamos considerar peor que las demás?

```
score <- read.table("scoring.txt", header = TRUE)
score
```

```
##     func1 func2 func3
## 1    1.05  0.83  0.72
## 2    0.86  0.80  0.72
## 3    1.09  0.69  0.69
## 4    0.87  0.77  0.95
## 5    0.90  0.76  0.70
## 6    1.04  0.78  0.68
## 7    0.68  0.76  0.72
## 8    1.04  0.70  0.72
## 9    1.10  0.82  0.75
## 10   1.12  0.80  0.70
## 11   0.90  0.76  0.86
## 12   0.88  0.79  0.75
## 13   0.99  0.77  0.71
## 14   1.08  0.86  0.71
## 15   0.95  0.83  0.70
## 16   1.04  0.76  0.90
## 17   1.08  0.80  0.74
## 18   0.81  0.79  0.69
## 19   0.84  0.72  0.68
## 20   1.00  0.79  0.79
## 21   0.97  0.80  0.70
## 22   0.97  0.80  0.95
## 23   0.99  0.79  0.77
## 24   0.88  0.83  0.72
## 25   1.11  0.80  0.79
## 26   1.05  0.78  0.71
## 27   1.02  0.84  0.73
## 28   0.91  0.83  0.69
## 29   1.19  0.81  0.86
## 30   0.96  0.82  0.73
## 31   0.92  0.74  0.89
## 32   0.81  0.78  1.03
## 33   0.86  0.83  0.74
```

```
## 34   0.95  0.77  0.84
## 35   0.93  0.70  0.78
## 36   0.87  0.86  0.71
## 37   1.01  0.79  0.74
## 38   1.13  0.76  0.90
## 39   0.93  0.74  0.81
## 40   0.87  0.77  0.72
## 41   0.91  0.73  0.72
## 42   0.95  0.82  0.78
## 43   1.18  0.68  0.76
## 44   1.09  0.73  0.68
## 45   0.86  0.87  0.71
## 46   0.99  0.82  0.68
## 47   1.08  0.88  0.83
## 48   1.04  0.68  0.68
## 49   1.11  0.80  0.68
## 50   0.86  0.83  0.85
## 51   1.05  1.11  1.21
## 52   0.94  1.02  0.70
## 53   1.15  1.15  0.69
## 54   0.80  1.17  0.80
## 55   0.89  1.11  0.82
## 56   0.85  1.04  0.70
## 57   1.09  1.08  0.82
## 58   1.00  1.11  0.70
## 59   1.10  1.08  0.73
## 60   0.96  1.19  0.70
## 61   0.96  1.17  0.68
## 62   1.02  1.10  0.68
## 63   1.02  1.11  0.68
## 64   1.00  1.10  0.73
## 65   1.08  1.14  1.01
## 66   0.95  1.10  0.71
## 67   0.96  1.11  0.71
## 68   0.99  1.19  0.81
## 69   0.95  1.06  0.68
## 70   0.95  1.05  0.80
## 71   1.02  1.06  1.09
## 72   0.90  1.15  0.74
## 73   0.92  1.18  0.79
## 74   1.11  1.09  0.94
## 75   0.88  1.06  0.69
## 76   1.19  1.13  0.75
## 77   0.99  1.15  0.73
## 78   1.02  1.14  0.69
## 79   1.06  1.12  0.85
## 80   1.21  1.11  0.68
## 81   1.05  1.01  0.84
## 82   0.89  1.12  0.68
## 83   1.14  1.06  0.80
## 84   0.86  1.04  0.69
## 85   0.86  1.01  0.90
## 86   1.16  1.15  0.87
## 87   1.10  1.12  0.81
```

```
## 88   0.97  1.09  0.70
## 89   1.02  1.15  0.73
## 90   1.01  1.07  0.75
## 91   0.98  1.09  0.70
## 92   0.87  1.09  0.82
## 93   0.88  1.06  0.72
## 94   0.98  1.05  0.79
## 95   1.06  1.06  0.73
## 96   1.00  1.12  1.07
## 97   1.16  1.04  0.69
## 98   0.99  1.22  0.73
## 99   0.97  1.09  0.80
## 100  1.00  1.18  1.10
```

Estos son los valores directamente obtenidos. La tabla es demasiado extensa para trabajar con todos los datos cómodamente, y por eso usaremos medidas resumen. Calculamos las medias:

```
media_1 <- mean(score[,1])
media_1
```

```
## [1] 0.9868
```

```
#Podemos elegir también según los nombres de las columnas

media_1 <- mean(score$func1)
media_1
```

```
## [1] 0.9868
```

```
media_2 <- mean(score$func2)
media_2
```

```
## [1] 0.9446
```

```
media_3 <- mean(score$func3)
media_3
```

```
## [1] 0.7729
```

La función colMeans calcula todo de una. También podemos aplicar las funciones mean y median (el argumento 2 indica que aplica la función por columnas).

```
colMeans(score)
```

```
##  func1  func2  func3
## 0.9868 0.9446 0.7729
```

```
mediafunc <- apply(score, 2, FUN = "mean")
mediafunc
```

```
##  func1  func2  func3
## 0.9868 0.9446 0.7729
```

```
medianafunc <- apply(score, 2, FUN = "median")
medianafunc
```

```
## func1 func2 func3
## 0.990 0.945 0.730
```

La media α-podada quita el $100\alpha\%$ de los valores a cada extremo. Hacemos la media α-podada para $\alpha = 0, 1$, el caso 0,2 queda para hacer en casa.

```r
media_01 <- rep(0,3)
media_1_01 <- mean(score$func1, 0.1)
media_1_01
```

```
## [1] 0.98525
```

```r
media_01[1] <- media_1_01
media_2_01 <- mean(score$func2, 0.1)
media_2_01
```

```
## [1] 0.945
```

```r
media_01[2] <- media_2_01
media_3_01 <- mean(score$func3, 0.1)
media_3_01
```

```
## [1] 0.753
```

```r
media_01[3] <- media_3_01
```

Comparamos los valores de estas medidas de centralidad. ¿Qué podemos observar?

```r
centralidadfunc <- data.frame(
                          Media = c("Media", mediafunc),
                          Mediana = c("Mediana", medianafunc),
                          Podada01 = c("Media01", media_01))
centralidadfunc
```

```
##        Media Mediana Podada01
##        Media Mediana  Media01
## func1 0.9868    0.99  0.98525
## func2 0.9446   0.945    0.945
## func3 0.7729    0.73    0.753
```

Probablemente, los datos para la función 3 sean asimétricos, dado que la mediana es menor a la media (esto se llama asimetría positiva o a derecha). En los demás casos, media y mediana son bastante similares.

2) *Obtener los percentiles 10, 25, 50, 75 y 90 y los valores máximos y mínimos, para cada una de las funciones de scoring. Comparar los valores obtenidos.*

En general, $x_{P\%}$ es el valor tal que el P % de los datos medidos son menores a $x_{P\%}$ (o tal que $100 - P$ % son mayores a él). Percentiles famosos:

$x_{50\%}$ es la mediana.

$x_{25\%} \to Q_1$, primer cuartil.

$x_{75\%} \to Q_3$, tercer cuartil.

```r
quantile(score$func1, c(0.1,0.25,0.5,0.75,0.9))
```

```
##    10%    25%    50%    75%    90%
## 0.8600 0.9075 0.9900 1.0525 1.1110
```

```r
quantile(score$func2, c(0.1,0.25,0.5,0.75,0.9))
```

```
##    10%    25%    50%    75%    90%
## 0.7580 0.7900 0.9450 1.1025 1.1500
```

```r
quantile(score$func3, c(0.1,0.25,0.5,0.75,0.9))
```

```
##  10%  25%  50%  75%  90%
## 0.68 0.70 0.73 0.81 0.90
```

Una forma rápida de calcular algunos de estos valores es pedirle a R que haga un summary de los datos:

```r
summary(score$func1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6800  0.9075  0.9900  0.9868  1.0525  1.2100
```

```r
summary(score$func2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6800  0.7900  0.9450  0.9446  1.1025  1.2200
```

```r
summary(score$func3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6800  0.7000  0.7300  0.7729  0.8100  1.2100
```

Los valores informados por el summary son el mínimo, Q_1, la mediana, la media, Q_3 y el máximo. Sirven para tener a primera vista información sobre el rango y la centralidad de los datos, es recomendable correrlo cuando se empieza un análisis exploratorio de datos.

3) *Calcular medidas de dispersión para estos tres conjuntos de datos: - desvío estándar, - rango intercuartil o intercuartílico (IQR), - MAD (mediana de la desviación absoluta).*

Comparar los valores de dispersión obtenidos. ¿Cuál de las funciones parece tener valores menos dispersos?

Recordemos las definiciones:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

El desvío estandar muestral es $s = \sqrt{s}^2$

El rango intercuartil, IQR, se define como $Q_3 - Q_1$, donde Q_1 y Q_3 son el primer y tercer cuartil, respectivamente.

La MAD:

$$\text{mediana}|x_i - \tilde{x}|$$

(ver ejs. 5.)

```r
sd(score$func1)
```

```
## [1] 0.1018404
```

```r
IQR(score$func2)
```

```
## [1] 0.3125
```

```r
quantile(score$func1,0.75)-quantile(score$func1,0.25)
```

```
##   75%
## 0.145
```

```r
#Calculamos todo a la vez

apply(score, 2, "sd")
```

```
##     func1     func2     func3
## 0.1018404 0.1674552 0.1058329
```

```r
apply(score, 2, "IQR")
```

```
##  func1  func2  func3
## 0.1450 0.3125 0.1100
```

```r
apply(score, 2, "mad")
```

```
##    func1    func2    func3
## 0.111195 0.229803 0.059304
```

```r
dispersionfunc <- data.frame(
                      SD = c(round(apply(score, 2, "sd"),3)),
                      IQR = c(round(apply(score, 2, "IQR"),3)),
                      MAD = c(round(apply(score, 2, "mad"),3)))

dispersionfunc
```

```
##          SD   IQR   MAD
## func1 0.102 0.145 0.111
## func2 0.167 0.312 0.230
## func3 0.106 0.110 0.059
```

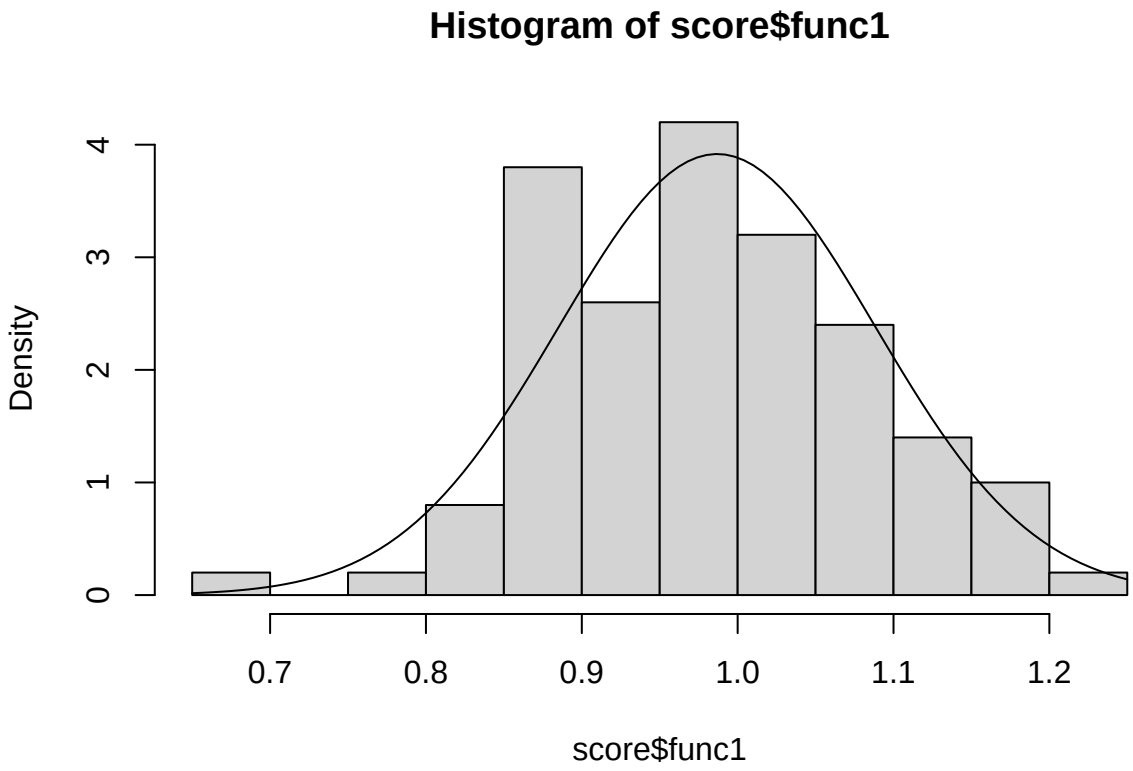
La función 3 parece tener la menor dispersión de datos, seguida de cerca por la función 1.

4) *Construir histogramas que permitan visualizar los valores de scoring para cada función. ¿Qué observaciones haría sobre la distribución de estos valores?. ¿Alguna de ellas parece bimodal? ¿En alguna de ellas parece haber valores atípicos o outliers?*

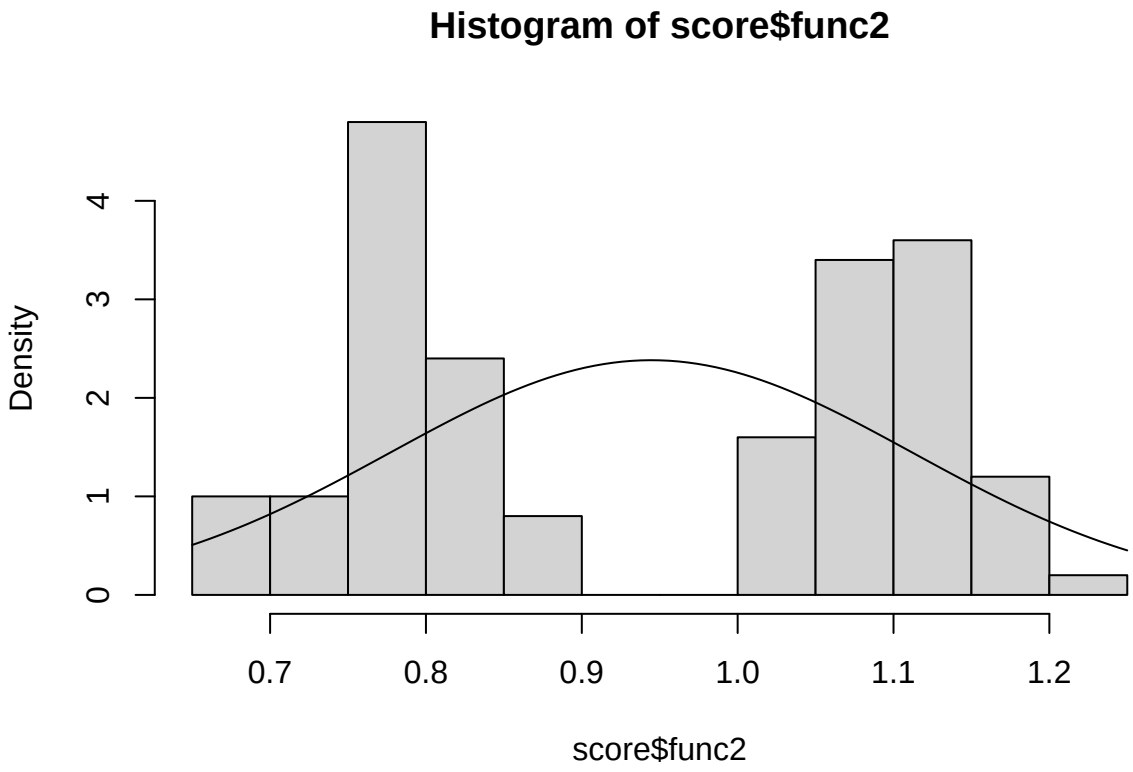
¿Los valores de scoring se hallan en el rango deseado? ¿Hay alguna asimetría en la distribución de los valores de una función? ¿En algún caso el ajuste normal parece razonable?

```r
media_1<-mean(score$func1);desvio_1<-sd(score$func1)
media_2<-mean(score$func2);desvio_2<-sd(score$func2)
media_3<-mean(score$func3);desvio_3<-sd(score$func3)
```

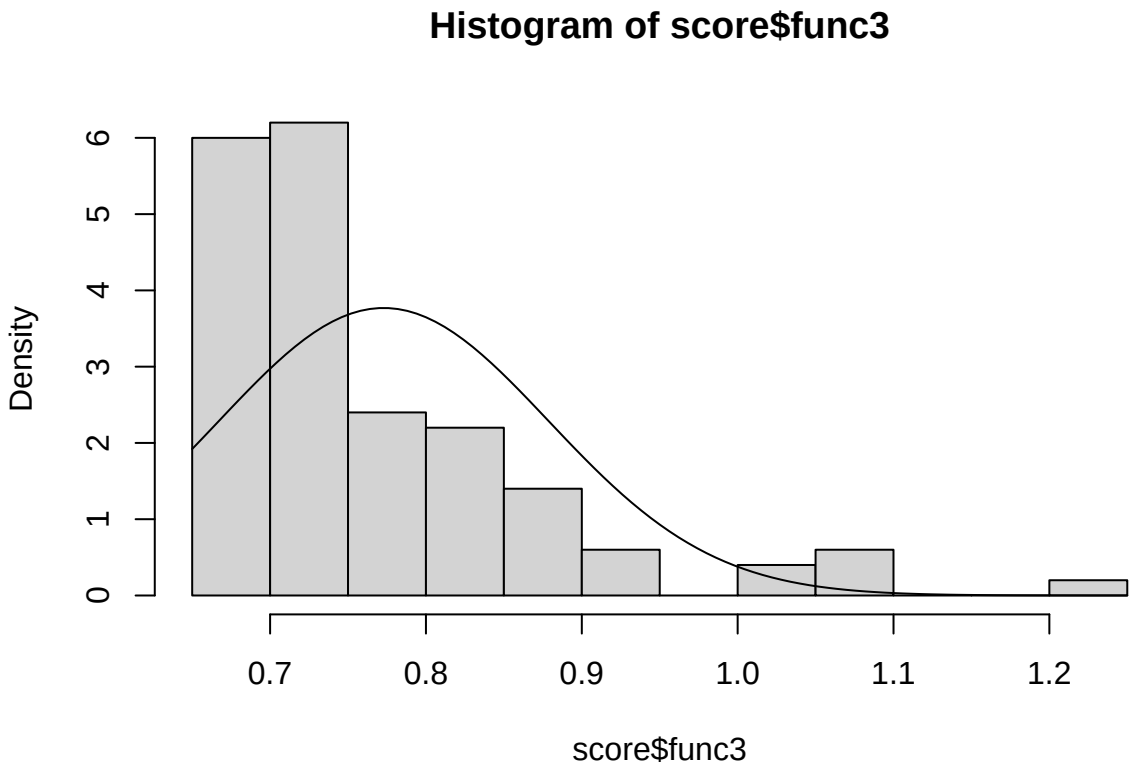
```r
#Calculo medias y desvíos para representar las curvas normales.

hist(score$func1, prob=TRUE)
curve(dnorm(x, mean = media_1, sd= desvio_1), add=TRUE)
```




```r
hist(score$func2, prob=TRUE)
curve(dnorm(x, mean = media_2, sd= desvio_2), add=TRUE)
```




```
hist(score$func3, prob=TRUE)
curve(dnorm(x, mean = media_3, sd= desvio_3), add=TRUE)
```



5) *Graficar 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, discutir simetría, presencia de outliers y comparar dispersiones.*

Un box-plot (o gráfico/diagrama de caja) es un gráfico que rápidamente nos permite representar conjuntos de datos a partir de la información brindada por sus cuartiles. En la caja se toman como borde izquierdo (o inferior) el valor del primer cuartil, Q_1, y como borde derecho (o superior), el tercer cuartil, Q_3. Una línea en el medio de la caja representa la mediana.

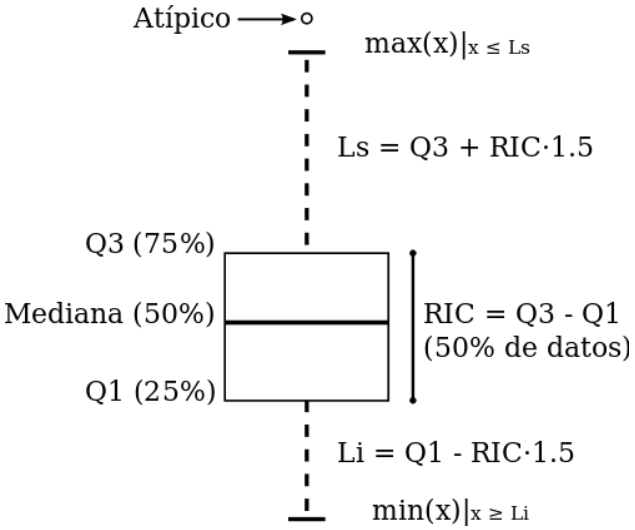
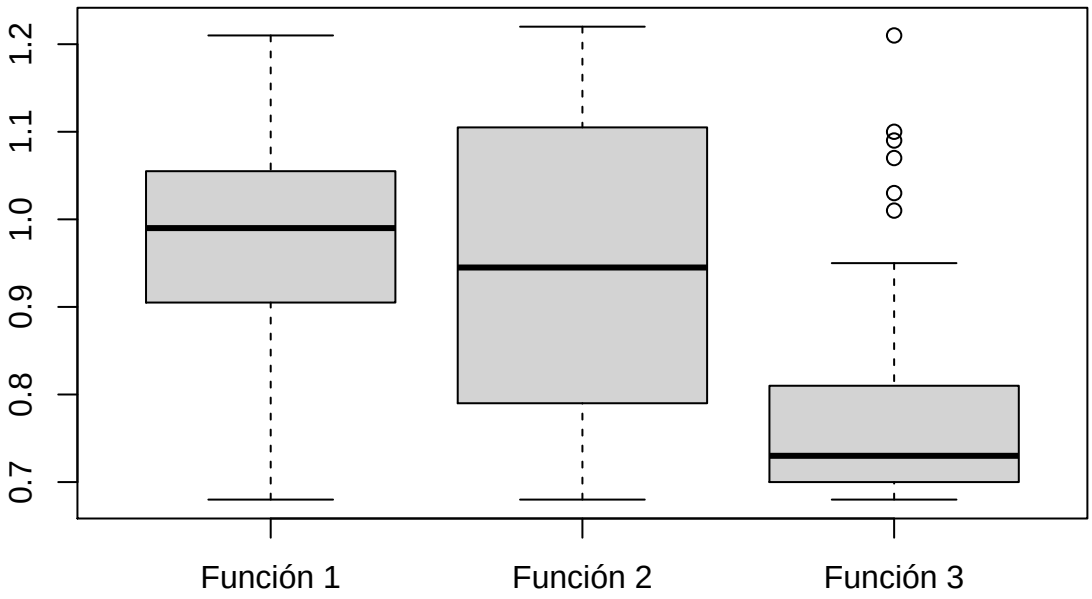


Figure 1: Elementos de un boxplot (Wikipedia)

Por fuera de la caja (que tendrá longitud igual a IQR, el rango intercuartil) se dibujan los "bigotes": tienen longitud $1,5 \times IQR$ (en R), y se extienden a ambos lados. Si se alcanzan el máximo o el mínimo antes de $1,5 \times IQR$, se cortan ahí los bigotes. Si hay valores por fuera de ese rango, se los representa con un símbolo especial, y se los llama valores atípicos o "outliers".

Figure 2: Maldito seas, mi diagrama es una caja

```
#Se puede hacer de dos formas:
#with(data=score, boxplot(func1, func2, func3))
#O así, y le ponemos los nombres
boxplot(score$func1, score$func2, score$func3, names=c("Función 1","Función 2", "Función 3"))
```



Comparando las longitudes de las cajas se puede visualizar la dispersión de los datos, mientras que la posición de la mediana da una indicación de centralidad. La posición de la mediana dentro de la caja muestra la asimetría en la distribución (si no está centrada), así como también lo hace la longitud de los bigotes y la presencia de outliers.

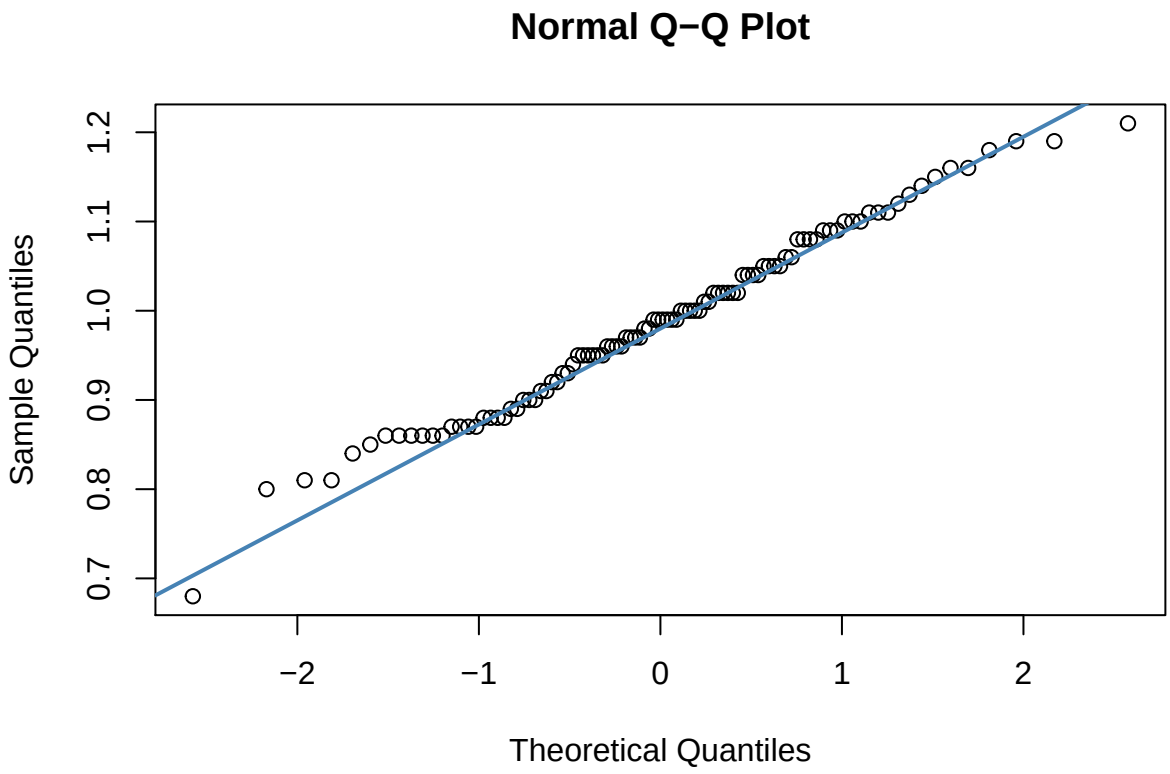
6) *Graficar los qqplots correspondientes. ¿En algún caso el ajuste normal parece razonable?*

Un qqplot (del inglés quantile-quantile, gráfico cuantil-cuantil) es un gráfico que sirve para comparar visualmente las diferencias en las distribuciones de dos conjuntos de datos (o, como haremos aquí, comparar un conjunto de datos contra una variable aleatoria de distribución conocida).

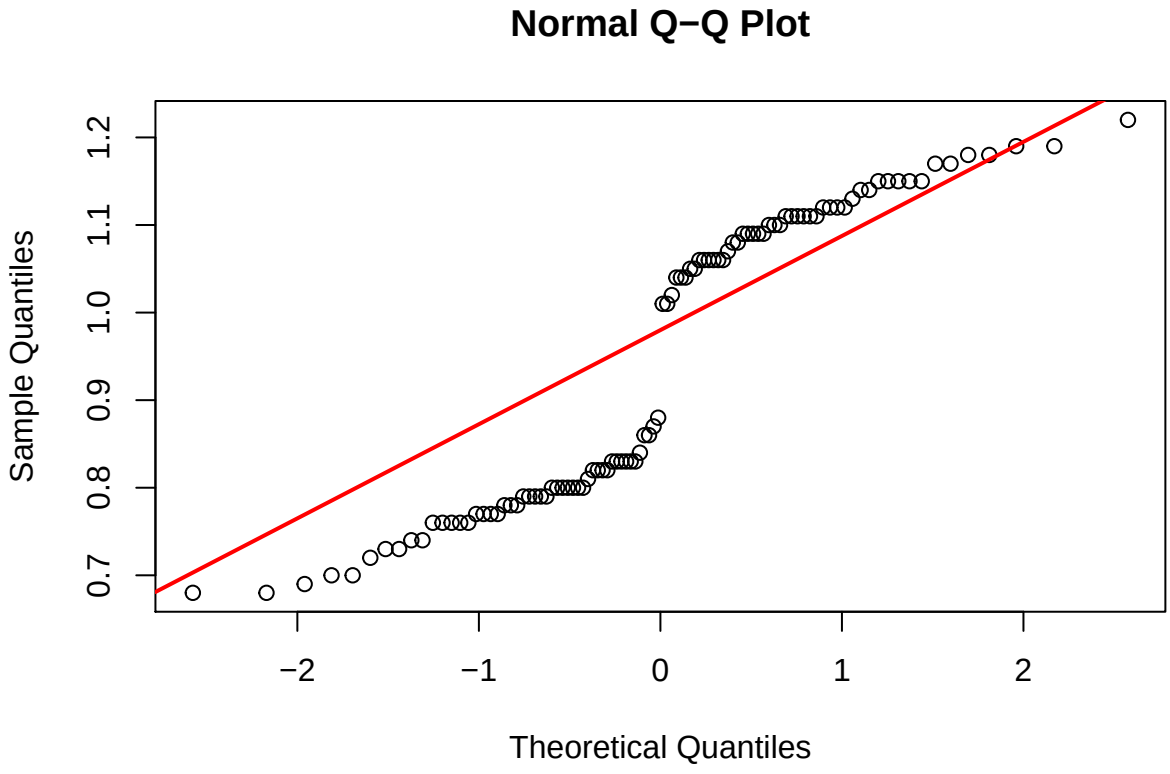
Los gráficos que haremos aquí comparan los cuantiles de los datos contra los cuantiles de una normal de igual media y desvío. Como en este caso tenemos 100 datos, el método de comparación es: primer valor de los datos contra primer percentil normal, segundo valor contra segundo percentil, y así.

Si estos valores son iguales, se hallarán sobre la recta $y = x$. Si hay diferencias en la distribución de los datos, la posición de los puntos (x, y) se alejará de la recta.

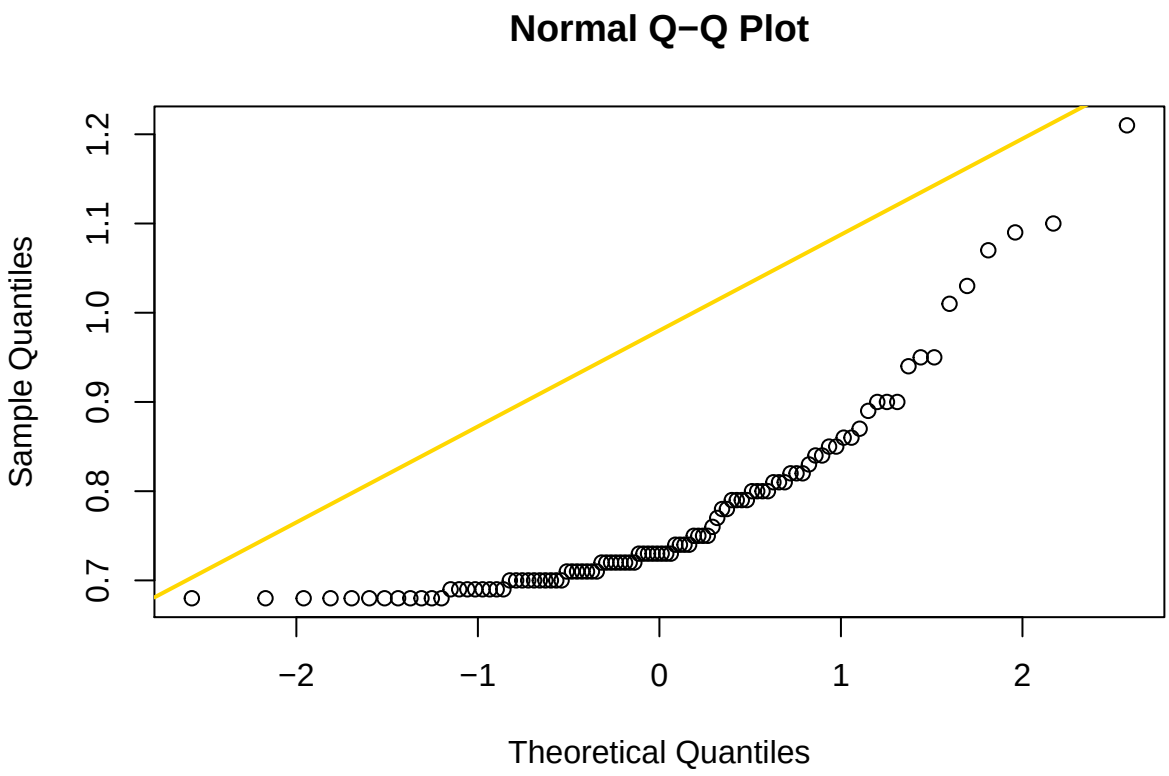
```
with(data=score, qqnorm(func1))
qqline(score$func1, col = "steelblue", lwd = 2)
```



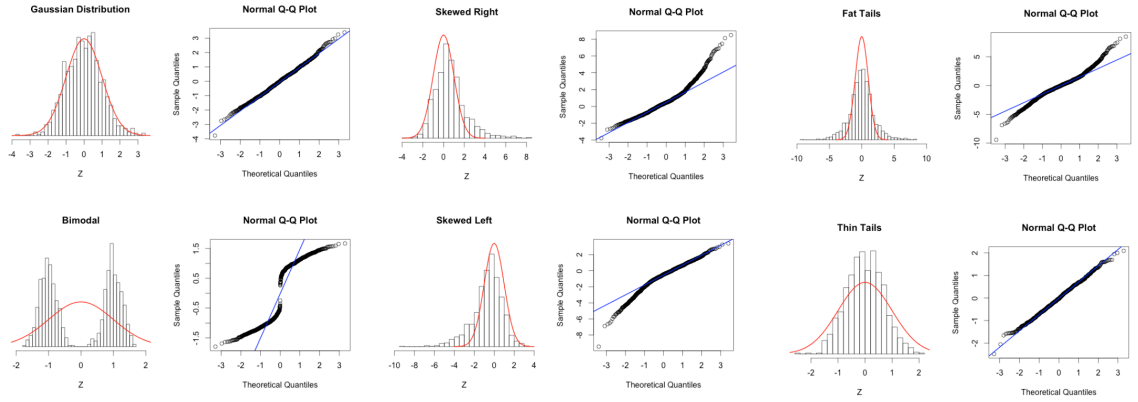

```
with(data=score, qqnorm(func2))
qqline(score$func1, col = "red", lwd = 2)
```




```
with(data=score, qqnorm(func3))
qqline(score$func1, col = "gold", lwd = 2)
```



Dependiendo de cómo se desvíen los valores de los cuantiles de la recta $y = x$, podemos deducir algunas características de la distribución estudiada.



En este caso, podemos observar bien el comportamiento bimodal de los datos de la función 2 (reconocemos el salto que pegan en el medio).

Los datos de la función 1 ajustan bastante bien a una distribución normal, aunque para las colas el ajuste es un poco peor. El comportamiento de las colas, donde parece que van bajando, indicaría que son un poco más livianas que las de la normal.

Para la función 3, se observa que tiene una asimetría a derecha, dado que la curva que hacen los datos se va por debajo de la línea de referencia del qqplot.

7) *En base a todo el análisis anterior, ¿cuál sería la función de scoring que más se ajusta a los requerimientos pedidos?*

La función 1 es la que más se ajusta a lo pedido:

- su media (mediana, media α-podada) es la más cercana a 1, aunque la función 2 también cumpliría lo pedido.
- los datos están concentrados en el rango $1 \pm 0,1$, como se observa a partir de las medidas de dispersión y del histograma y el boxplot. La función 3 tiene una dispersión similar, pero no es en el rango correcto.
- cumple los requerimientos de ajustar aproximadamente a una normal. La función 2 no los cumple por ser una distribución bimodal, y la función 3 tiene una asimetría a derecha, con una cola derecha muy larga con outliers, y una cola a izquierda muy corta. Con eso puedo descartar 2 y 3. Pero además, a partir del qqplot, se observa que la distribución es similar a una normal.
- sólo comprobamos que visualmente 1 tiene distribución similar a una normal. Para terminar de verificar (en realidad, fallar en descartar) el supuesto de normalidad de la función 1, podríamos hacer un test de hipótesis (guía 9).