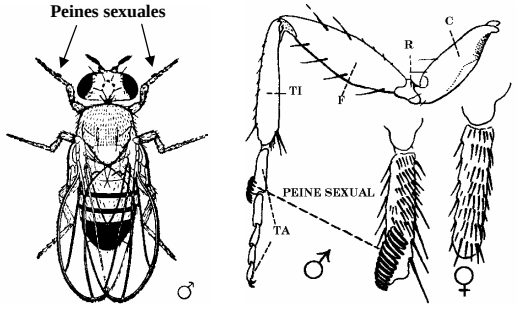
ESTADÍSTICA
Primer Curso del Grado en Bioquímica (2018/19)

Tema 1: ESTADÍSTICA DESCRIPTIVA

Observación: En este tema se debe intentar resolver todos los ejercicios posibles tanto "a mano" (ayudándose de calculadora) como con con el programa R.

1.1. A continuación reproducimos el número de dientes en el peine sexual de las patas derecha (X) e izquierda (Y) en 20 machos de *Drosophila*[1]:



X	6	6	5	6	7	5	6	7	8	6	7	7	7	6	8	8	7	6	8	7
Y	5	6	7	6	6	8	7	6	6	8	7	7	7	8	6	7	8	9	8	9

a) Calcular la media, mediana, desviación típica, coeficiente de variación, mínimo y máximo para pata derecha e izquierda respectivamente. Utilizar estas medidas para comparar las patas entre sí.

b) Determinar el coeficiente de correlación entre X e Y. Calcular la recta de regresión de Y sobre X y de X sobre Y. ¿Son la misma recta?

c) Dibujar el diagrama de dispersión de Y en términos de X y superponer la recta de regresión correspondiente. Evaluar la bondad del ajuste.

1.2. Se mide la variable X, nivel plasmático de cortisol (en mg/dl), en 20 perros que padecen hiperadrenocorticismo hipofisario, obteniéndose la siguiente muestra:

0.8 2.2 2.2 2.3 2.4 2.6 2.9 2.9 2.9 3.1
3.1 3.2 3.7 3.8 3.8 4.2 4.2 4.4 5.0 5.3.

Además sabemos que $\sum_{i=1}^{20} x_i = 65$ y que $\sum_{i=1}^{20} x_i^2 = 233.12$.

a) Calcular la media, la varianza, la desviación típica, el coeficiente de variación, la mediana, los cuartiles, el rango y el rango intercuartílico de los datos.

b) Dibuja un diagrama de caja de los datos, especificando claramente qué es o cómo se determina cada parte del diagrama.

1.3. Se ha efectuado un análisis químico de vinos de la misma región de Italia[2]. Reproducimos algunas medidas de ácido málico obtenidas:

1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35
2.16 1.48 1.73 1.73 1.87 1.81 1.92 1.57 1.59 1.63.

Realizar un análisis descriptivo de los datos, incluyendo su representación gráfica.

[1] Fuente de los datos: Griffiths *et al.* (2008). *Genética*. McGrawHill.
[2] Fuente de los datos: UCI Machine Learning Repository.

1.4. Meloun *et al.* (2001), *Clin. Chem. Lab. Med.*, estudian la intensidad del acné en mujeres según los niveles de la globulina fijadora de hormonas sexuales SHBG (del inglés *sex hormone binding globulin*). Se ha medido el nivel de SHBG en un grupo de mujeres con poco acné y en un grupo de mujeres con acné severo obteniéndose las siguientes observaciones:

Poco acné										
58.6	62.2	36.2	22.2	37.4	81.8	95.6	42.7	53.1	107.3	25.3
47.6	30.5	114.0	31.1	69.4	68.5	35.0	146.1	167.0	84.5	39.1
96.7	80.7	39.5	42.9	96.3	57.3	43.4	56.8	56.6	14.4	14.1

Acné severo									
200.0	42.1	97.5	189.0	28.0	59.6	91.8	165.2	173.5	71.4
179.2	48.7	90.3	83.7	95.2	71.0	83.1	65.6	62.6	47.9
70.3	29.6	11.5	59.0	110.7	40.0	67.4	78.2	194.1	49.6

Los datos están en la web de la asignatura en dos ficheros de texto. Se pueden cargar en R mediante los comandos:

```
Poco = scan("AcnePocoSHBG.txt")
Severo = scan("AcneSeveroSHBG.txt")
```

La función `scan` se suele utilizar para cargar en R (como un vector) ficheros de datos que sólo contienen observaciones de una única variable (en una fila o columna). Para cargar una tabla de datos (que tiene más de una fila y más de una columna) utilizaremos la función `read.table`.

a) Calcular media, mediana, desviación típica, varianza, coeficiente de variación y distancia intercuartílica en cada conjunto de datos.

b) Calcular los percentiles 90, 75, 50, 25 y 10.

c) Representar los datos con diagramas de caja en paralelo.

d) Utilizando la información obtenida en los apartados anteriores comparar los niveles de SHBG en los dos grupos de mujeres.

1.5. En las siguientes tablas aparecen el número de hojas por planta en generaciones F_1 y F_2 del cruce de dos variedades de tabaco cultivado[3].

F_1				
18	15	16	18	15
16	14	16	18	17
16	13	16	14	16
15	16	15	15	16
15	16	16	15	16

F_2				
16	20	19	17	14
16	14	14	15	17
20	13	12	15	16
21	18	15	14	18
14	17	13	15	13

Realizar un análisis descriptivo de los datos para saber si hay diferencias entre ambas generaciones: comparar gráficos, medidas de localización y de dispersión.

1.6. La distrofia muscular de Duchenne (DMD) es una enfermedad transmitida genéticamente por la madre a los hijos. Las mujeres afectadas normalmente no sufren síntomas visibles y pueden ser portadoras de la enfermedad sin saberlo. La DMD se manifiesta primariamente en los varones debido a que el gen de la enfermedad se encuentra en el cromosoma X. El conjunto de datos `dmd.dat` proviene de Percy *et al.* (1981) y contiene observaciones de 194 mujeres, familiares de niños con DMD. La muestra estaba constituida por 67 portadoras de la DMD y 127 no portadoras. Esta muestra formó parte de un programa canadiense cuyo objetivo era informar a las mujeres de su probabilidad de ser portadoras, basándose en marcadores séricos, así como en su historial familiar. Los dos primeros marcadores, creatina quinasa y hemopexina, son baratos de medir, mientras que la piruvato quinasa o la lactato deshidrogenasa son costosas. Las variables del estudio son

[3]Fuente de los datos: Falconer & Mackay (1996). *Introduction to Quantitative Genetics*. Longman.

Columna	Variable
1	Edad de la mujer
2	Nivel de creatina quinasa
3	Nivel de hemopexina
4	Nivel de piruvato quinasa
5	Nivel de lactato deshidrogenasa
6	Indicador de si una mujer es portadora (1) de DMD o no (0)

El siguiente código permite cargar los datos, extraer variables de la tabla y separar sus valores en los grupos de portadoras y no portadoras. Ejecuta el código paso a paso para entender qué es lo que hace, mirando la ventana de comandos (`console`) y las variables creadas en el `environment`.

```
Datos = read.table("dmd.dat")
C = Datos$V6 # Indicador de si es portadora (1) o no (0)
portadora = (C==1) # TRUE si son portadoras, FALSE si no lo son
CQ = Datos$V2 # Nivel de creatina quinasa
CQportadoras = CQ[portadora] # Creatina quinasa en portadoras
CQnoportadoras = CQ[!portadora] # Creatina quinasa en no portadoras
PQ = Datos$V4 # Nivel de piruvato quinasa
PQportadoras = PQ[portadora] # Piruvato quinasa en portadoras
PQnoportadoras = PQ[!portadora] # Piruvato quinasa en no portadoras
```

a) Examinar y ejecutar el siguiente código y explicar qué realiza. Interpretar los resultados.

```
summary(CQportadoras)
summary(CQnoportadoras)
boxplot(CQportadoras,CQnoportadoras,names=c("portadoras","no portadoras"))
boxplot(log(CQportadoras),log(CQnoportadoras),
        names=c("portadoras","no portadoras"))
```

b) Examinar y ejecutar el siguiente código y explicar qué realiza. Interpretar los resultados.

```
plot(log(CQportadoras),log(PQportadoras),type="p",
     xlab="log(CQportadoras)",ylab="log(PQportadoras)")
cor(log(CQportadoras),log(PQportadoras))
```

1.7. Como parte de un estudio de la relación entre el tamaño del cerebro y la capacidad para resolver "problemas", se midió el peso corporal (en kg) y el volumen cerebral (en ml) de 39 especies carnívoras alojadas en zoos de Norteamérica[4]. Los datos están en el fichero `BrainSizeBodyMass.txt`, que se puede descargar de la web de la asignatura.
Tomar logaritmos (por defecto siempre son neperianos) de ambas variables. Determinar la recta de regresión de $Y = \log(\text{Volumen cerebro})$ sobre $X = \log(\text{Peso cuerpo})$, y el coeficiente de correlación. Representar la recta de regresión sobre el diagrama de dispersión. ¿Es bueno el ajuste?

1.8. En un experimento se expuso a bacterias oceánicas a rayos X durante 15 intervalos de 6 minutos cada uno. A continuación aparece el número de bacterias supervivientes al final de cada intervalo.

No. de bac.	355	211	197	166	142	106	104	60	56	38	36	32	21	19	15
Intervalo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

a) ¿Qué aspecto tiene la relación entre el número de bacterias supervivientes y el tiempo de exposición a la radiación? ¿Te parece razonable utilizar una regresión lineal?

b) Intenta transformar los datos de manera que una recta se ajuste bien a los datos transformados.

c) Estima la cantidad inicial de bacterias antes de comenzar a radiarlas.

[4]Fuente de los datos: Benson-Amram, S. *et al.* (2016). Brain size predicts problem-solving ability in mammalian carnivores. *PNAS*.

1.9. La esteatosis (acumulación anormal de grasa en las células) se da en más de la mitad de los pacientes con infección crónica del virus de la hepatitics C (VHC). Hickman *et al.* (2002)[5] conjeturaron que una reducción de peso en estos pacientes produciría una disminución en el grado de la esteatosis. Para comprobarlo, se sometió a 10 sujetos con VHC crónico a un programa de reducción de peso de tres meses de duración. En cada sujeto se biopsió el hígado antes y después del tratamiento. Entre otros marcadores, en cada biopsia se midió la expresión (en células/mm^2) de alfa actina de músculo liso (ACTA-2) en el tracto portal, obteniéndose los siguientes resultados:

ACTA-2

Antes 347	412	488	219	168	1273	410	319	209	225
Después 60	363	12	156	199	691	123	239	141	49

a) En la consola de R escribimos

```
Datos = read.table("ACTA2portal.txt",header=TRUE)
X = Datos$Antes
Y = Datos$Despues
```

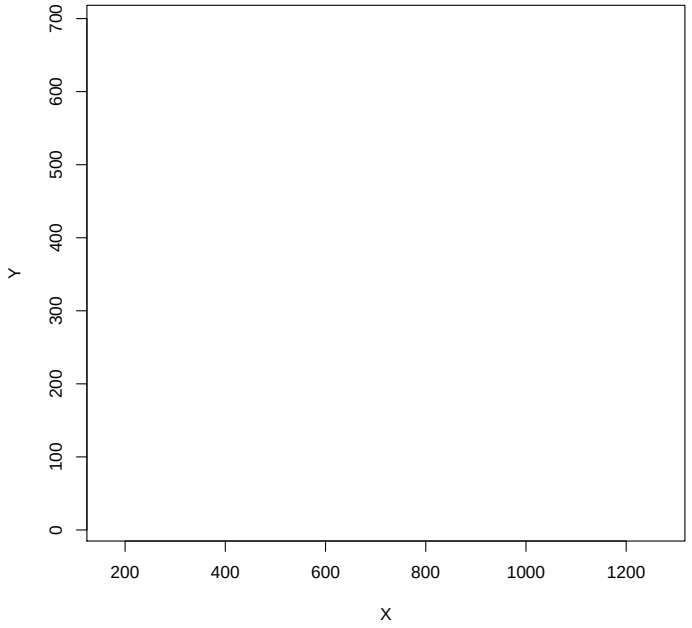
Explica qué se hace con el siguiente código de R y lo que se obtiene. Dibuja (aproximadamente) el gráfico resultante en el recuadro.

```
L = lm(Y~X)
L

    Call:
    lm(formula = Y ~ X)

    Coefficients:
    (Intercept)            X
         1.1698       0.4966

plot(X,Y)
abline(L)
```



b) Obtén el coeficiente de correlación entre la expresión de ACTA-2 antes y después de la dieta a partir de la siguiente información:

```
var(X)
[1] 103632
var(Y)
[1] 39794.9
cov(X,Y)
[1] 51467.22
```

Interpreta el valor obtenido de la correlación ayudándote del gráfico que hayas dibujado en (a).

c) Sabiendo que $\sum_{i=1}^{10} x_i = 4070$, calcular $\sum_{i=1}^{10} x_i^2$ y $\sum_{i=1}^{10} x_i y_i$ a partir de los resúmenes de los datos dados en (a) y (b).

[5] I J Hickman, A D Clouston, G A Macdonald, D M Purdie, J B Prins, S Ash, J R Jonsson, E E Powell (2002). Effect of weight reduction on liver histology and biochemistry in patients with chronic hepatitis C. *Gut*, 51, 89–94.