Índice

1. Arqueología y Paleontología **11**
1.1. Introducción 11
1.2. Problemas Básicos Resueltos 11
1.3. Referencias 57

2. Biología y Ciencias Ambientales **61**
2.1. Introducción 61
2.2. Problemas Básicos Resueltos 61
2.3. Referencias 172

3. Ciencias de la Salud **175**
3.1. Introducción 175
3.2. Problemas Básicos Resueltos 175
3.3. Referencias 318

4. Ingeniería **321**
4.1. Introducción 321
4.2. Problemas Básicos Resueltos 321
4.3. Referencias 364

5. Economía y Ciencias Sociales **365**
5.1. Introducción 365
5.2. Problemas Básicos Resueltos 365
5.3. Referencias 498

Capítulo 1

Arqueología y Paleontología

1.1. Introducción

En este capítulo se resuelven muchos problemas elementales de Estadística, por lo que en casi todos ellos se hará referencia al capítulo en el que se desarrolló la técnica correspondiente estudiada en el texto denominado CB, como dijimos en el Prólogo.

Por otro lado, como muchos problemas se resolverán utilizando el paquete estadístico R, el texto EBR aparecerá en muchas ocasiones como la referencia.

Aunque en este capítulo hay algún problema sobre Regresión Lineal, no es una técnica que se utilice mucho en Arqueología debido a que no está claro qué variable debe considerarse como dependiente y cuáles como independientes. Por otro lado, suele haber dependencia entre las posibles variables teóricamente independientes (problema de colinealidad), lo que invalidarían las conclusiones obtenidas con esta técnica.

1.2. Problemas Básicos Resueltos

Problema 1.1

Los siguientes datos, obtenidos de Wainwright (1979), corresponden a cantidades de fragmentos de distintos tipos de cerámica del yacimiento de Mount Pleasant en Dorset (Inglaterra):

Cuenco Neolítico liso	**391**
Cerámica Acanalada	**657**
Campaniformes	**1695**
Cerámica de Petersborough	**6**
Edad del Bronce	**591**

Realizar representaciones gráficas de estos datos.

Como se trata de datos de una variable cualitativa, las dos representaciones gráficas sugeridas en EBR-sección 2.3.1 ó CB-sección 2.3.1, son el Diagrama de sectores y el Diagrama de rectángulos (también denominado por algunos Diagrama de barras aunque nosotros hemos reservado esta última denominación para el caso de datos cuantitativos sin agrupar). Realizaremos los gráficos utilizando el software R.

Para obtener el Diagrama de sectores, primero introducimos los datos creando el vector de datos en `(1)`, el vector de etiquetas que queremos aparezca en el dibujo en `(2)` y el vector de colores, de cada uno de los sectores, en `(3)`,

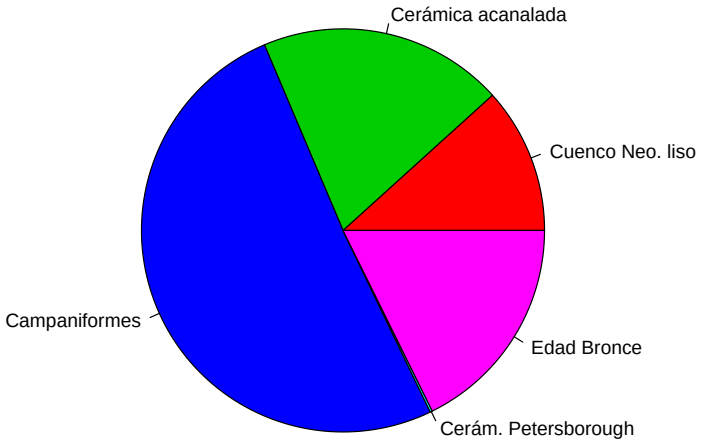

Figura 1.1 : Diagrama de sectores

```
> x<-c(391,657,1695,6,591)                                          (1)
> eti<-c("Cuenco Neo. liso","Cerámica acanalada","Campaniformes",
+ "Cerám. Petersborough","Edad Bronce")                             (2)
> c1<-c(2,3,4,5,6)                                                  (3)
```

Ahora ya podemos obtener el gráfico buscado mediante la función `pie` ejecutando `(4)` con lo que obtenemos el gráfico en la Figura 1.1

```
> pie(x,labels=eti,col=c1)                                          (4)
```

Para obtener el Diagrama de rectángulos utilizaremos la función `barplot`, en la que observamos el argumento `names` en lugar de `labels` que utilizábamos en la función anterior. (Realmente hemos definido un nuevo vector de etiquetas con nombres abreviados para que éstos quepan en la base del gráfico).

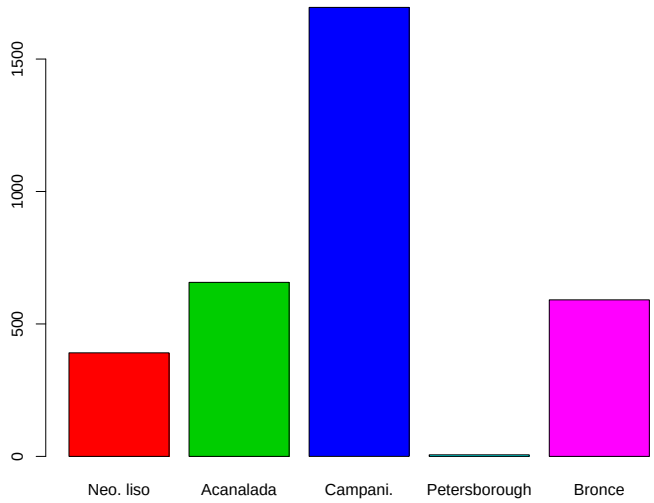

Figura 1.2 : Diagrama de rectángulos

El gráfico buscado lo obtenemos en la Figura 1.2 ejecutando `(5)`

```
> barplot(x,names=eti,col=c1)                                    (5)
```

Problema 1.2

Los tamaños (en hectáreas) de 25 asentamientos del Uruk tardío en la antigua Mesopotamia son, según Johnson (1973),

45	**37**	**34′8**	**52**	**75**	**86**	**59′7**	**74**	**32**	**57′7**
65	**86**	**37**	**38′4**	**90′5**	**45**	**67**	**50**	**33**	**30**
43′2	**32**	**35′2**	**54′5**	**43′1**					

Realizar un análisis descriptivo de estos datos.

Primero haremos una representación gráfica de los datos del enunciado. Aunque éstos no vienen clasificados por intervalos, utilizaremos un histograma para representarlos. Para ello, después de introducir los datos en `(1)`, obtenemos el histograma de la Figura 1.3 ejecutando en `(2)` la función `hist`, en donde destacamos dos cosas: una, que con el argumento `main` podemos poner un título (esto también se podía haber hecho en los gráficos del ejercicio anterior). Y dos, que como hemos utilizado el mismo vector de colores que

teníamos en el ejercicio anterior, con cinco colores y tenemos siete intervalos, éstos se empiezan a repetir.

Podíamos haber forzado a la función `hist` para que hiciera los intervalos de corte en donde quisiéramos, utilizando el argumento `breaks`, con un vector de puntos de corte, o con un valor que indique con cuántos intervalos queremos que haga la representación gráfica. No lo hemos hecho así y hemos dejado al ordenador que los eligiera.

```
> x<-c(45,37,34.8,52,75,86,59.7,74,32,57.7,65,86,
+ 37,38.4,90.5,45,67,50,33,30,43.2,32,35.2,54.5,43.1)                    (1)
> c1<-c(2,3,4,5,6)
> hist(x,col=c1,main="Tamaño de asentamientos")                           (2)
```

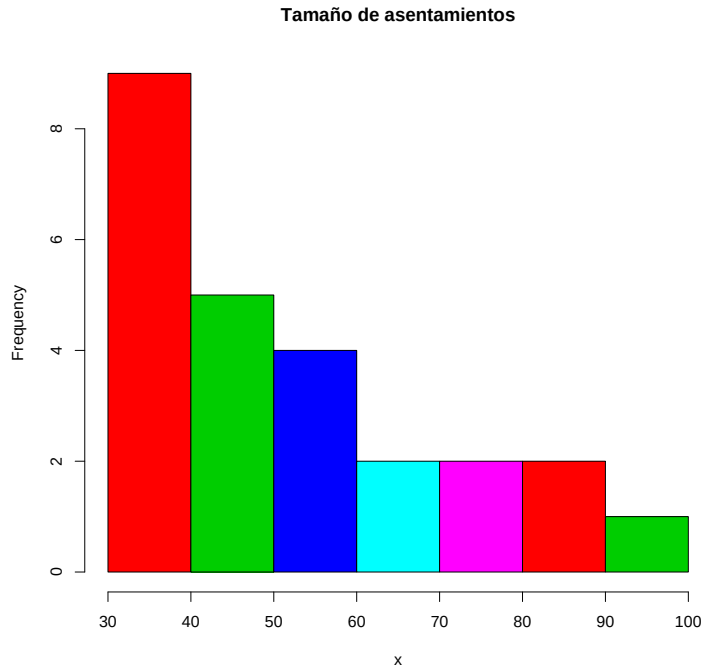

Figura 1.3 : Histograma

Si quisiéramos que en el eje de ordenadas pusiera *Frecuencias absolutas* en lugar de *Frequency* teclearíamos

```
> hist(x,col=c1,main="Tamaño de asentamientos",ylab="Frecuencias absolutas")
```

utilizando un argumento más de la función `hist`. Análogamente se podría hacer con el eje la abscisas. De nuevo estos argumentos se pueden utilizar en las otras representaciones gráficas.

Ahora vamos a calcular algunas medidas de posición como la media muestral (ejecutando `(3)`), la mediana muestral (ejecutando (4)), y alguna medida de dispersión como la cuasivarianza muestral (ejecutando (5)) y la desviación típica muestral (ejecutando (6)). Como se ve, la función `var` nos proporciona la cuasivarianza muestral y no la varianza muestral, por lo que, para obtener la desviación típica muestral debemos utilizar el tamaño $n = 25$ de la muestra en `(6)` y la función raíz cuadrada `sqrt`.

```
> mean(x)                                                    (3)
[1] 52.124
> median(x)                                                  (4)
[1] 45
> var(x)                                                     (5)
[1] 350.6494
> sqrt(24*var(x)/25)                                         (6)
[1] 18.3473
```

Problema 1.3

En dos cuevas del Pleistoceno de una zona meridional de Inglaterra, se obtuvieron (Boyle, 1983) los siguientes datos del número de fragmentos de hueso, clasificados por sus longitudes en mm.,

	Longitudes					
	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)
Cueva 1	**1**	**21**	**15**	**5**	**7**	**1**
Cueva 2	**0**	**6**	**11**	**11**	**6**	**6**

	Longitudes				
	[60,70)	[70,80)	[80,90)	[90,100)	[100,110)
Cueva 1	**2**	**3**	**3**	**0**	**2**
Cueva 2	**6**	**4**	**2**	**5**	**9**

¿Se puede concluir que las distribuciones de las longitudes de huesos son distintas en ambas cuevas?

Se trata de contrastar si existen diferencias significativas entre las dos poblaciones.

Observando los histogramas de las Figuras 1.4 y 1.5 ya se aprecia que no pueden considerarse los datos observados como procedentes de poblaciones normales por lo que debemos utilizar algún test no paramétrico.

De los posibles contrastes de comparación de dos poblaciones, podemos realizar el test de homogeneidad de varias muestras (dos muestras en nuestro caso) pues tenemos recuentos de observaciones en clases previamente especificadas (los intervalos), como se indica en EBR-sección 8.2.3 ó CB-sección 12.3. Alternativamente podríamos considerar las marcas de clase de los intervalos y ejecutar alguno de los tres contrastes expuestos en CB-sección 13.4. Dado que,

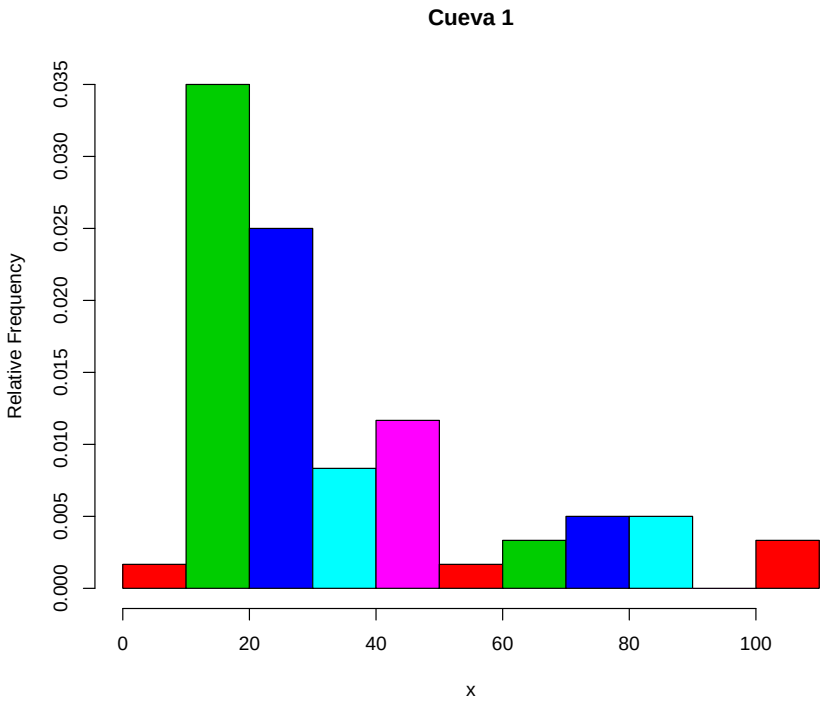

Figura 1.4 : Histograma

de esta manera, resumimos toda la información del intervalo en un solo valor (por esta razón no se utilizan datos clasificados en intervalos), de los tres tests allí propuestos, el test de Wilcoxon-Mann-Whitney no va a ser adecuado. De los otros dos tests, ejecutaremos el de Kolmogorov-Smirnov para dos muestras, cuyo desarrollo se estudia en CB-sección 13.4.2.

Aunque las funciones de distribución empíricas $F_m^*(x)$ y $G_n^*(x)$ de ambas muestras de tamaños $m = 60$ y $n = 66$, respectivamente, deberían saltar en cada valor observado, dado que los datos muestrales los tenemos agrupados por intervalos, los saltos se producirán en los extremos de éstos.

En la tabla que sigue tenemos las frecuencias absolutas acumuladas N_i y los valores de cada una de los dos funciones de distribución empíricas $F_m^*(x)$ y $G_n^*(x)$ respectivamente. El valor de éstas que aparece allí es su valor en todo el intervalo considerado (y justo en el extremo inferior puesto que son continuas por la derecha).

Así, $F_m^*(x)$ es cero hasta $x < 0$, en el intervalo $[0, 10)$ es igual $0'02$, etc., hasta el intervalo $[100, 110)$ que salta a 1 al comienzo de dicho intervalo y ya es constantemente igual a 1. Formalmente, será

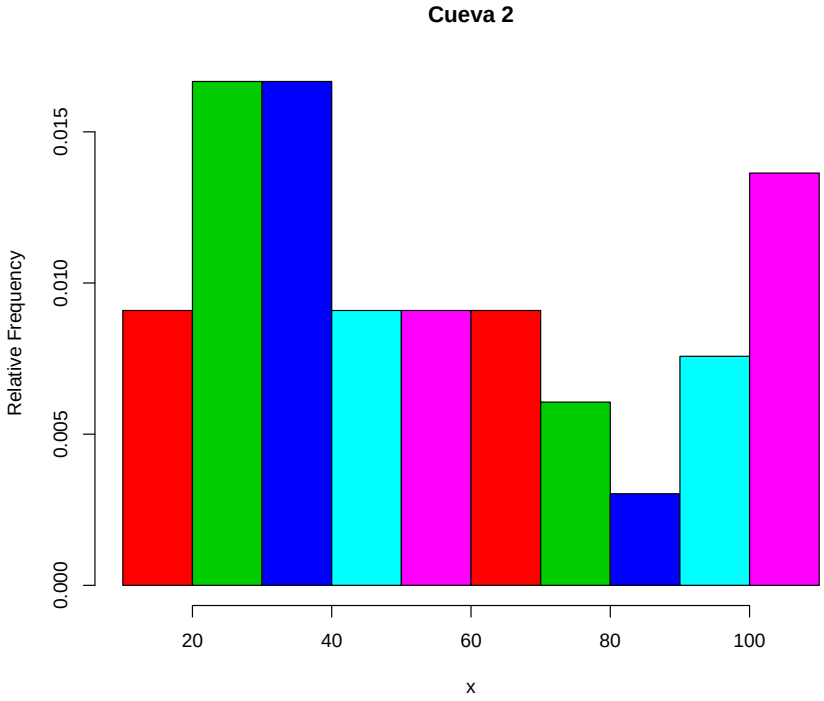

Figura 1.5 : Histograma

$$F_m^*(x) = \begin{cases} 0 & & x < 0 \\ 0'02 & & 0 \leq x < 10 \\ 0'37 & & 10 \leq x < 20 \\ \dots & \dots & \dots \\ 0'97 & & 90 \leq x < 100 \\ 1 & & x \geq 100 \end{cases}$$

Análogamente,

$$G_n^*(x) = \begin{cases} 0 & & x < 0 \\ 0 & & 0 \leq x < 10 \\ 0'09 & & 10 \leq x < 20 \\ \dots & \dots & \dots \\ 0'86 & & 90 \leq x < 100 \\ 1 & & x \geq 100 \end{cases}$$

Intervalo	N_i Cueva 1	N_i Cueva 2	$F_m^*(x)$	$G_n^*(x)$	$\lvert F_m^*(x) - G_n^*(x)\rvert$
[0,10)	1	0	0′02	0	0′02
[10,20)	22	6	0′37	0′09	0′28
[20,30)	37	17	0′62	0′26	0′36
[30,40)	42	28	0′70	0′42	0′28
[40,50)	49	34	0′82	0′52	0′30
[50,60)	50	40	0′83	0′61	0′22
[60,70)	52	46	0′87	0′70	0′17
[70,80)	55	50	0′92	0′76	0′16
[80,90)	58	52	0′97	0′79	0′18
[90,100)	58	57	0′97	0′86	0′11
[100,110)	60	66	1	1	0

De la tabla anterior se deduce que la máxima diferencia, es decir, el valor del estadístico de contraste es

$$D_{m,n} = D_{60,66} = \sup_x |F_m^*(x) - G_n^*(x)| = 0'36.$$

Por otro lado, de la tabla de probabilidades cola del estadístico de Kolmogorov-Smirnov para dos muestras, ADD-tabla 8, vemos (puesto que m y n son grandes) que

$$P\left\{D_{60,66} > 1'63\sqrt{\frac{60+66}{60\cdot 66}}\right\} = P\{D_{60,66} > 0'2908\} \leq 0'01$$

por lo que el p-valor del test de la hipótesis nula H_0:*las dos distribuciones son iguales*, frente a la alternativa H_1:*las dos distribuciones son distintas*, será

$$P\{D_{60,66} > 0'36\} < P\{D_{60,66} > 0'2908\} \leq 0'01.$$

Al ser el p-valor menor que 0′01, se concluye que no se puede aceptar la igualdad de las distribuciones de longitudes de huesos en ambas cuevas.

Problema 1.4

Se llevó a cabo un estudio en unas regiones de la antigua Mesopotamia (Johnson, 1973) en donde se realizaron intercambios locales, analizándose un tipo específico de cerámica. Un examen inicial de éstas sugiere que la anchura de las líneas representadas en ellas es distinta según las zonas de producción, clasificando las 90 cerámicas analizadas, según la siguiente tabla de doble entrada,

	Línea Gruesa	**Línea Fina**
Zona Oriental	**42**	**17**
Zona Occidental	**10**	**21**

¿Confirman estos datos las diferencias de los grosores de líneas en ambas zonas?

Puede plantearse como un contraste de independencia de caracteres (CB-sección 12.4 ó EBR-sección 8.2.4), *Zona* y *Ancho de la Línea*, o como un contraste de homogeneidad de varias muestras (CB-sección 12.3 ó EBR-sección 8.2.3), en concreto de la dos Zonas en estudio. Matemáticamente es lo mismo puesto que el estadístico y su distribución son idénticas. No obstante, la diferencia entre ambos tests estriba en cómo se recogieron las muestras. Si se recogió un número determinado de cerámicas y se las clasificó según las características Zona y Tamaño de la Línea, sería un contraste de Independencia de ambos caracteres. Si, como parece, se eligieron unas cuantas cerámicas de la Zona Oriental, 59, y se las clasificó según el tamaño de la Línea dibujada en ellas y lo mismo se hizo con otras cerámicas elegidas al azar de la Zona Occidental, 31, se trataría de un contraste de la hipótesis nula de igualdad de las dos zonas respecto a la característica Tamaño de la Línea.

En ambos casos el estadístico de Pearson toma el mismo valor

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n}$$

siendo r el número de poblaciones a comparar (número de filas de la tabla anterior), s el número de clases en las que se han clasificado las poblaciones a comparar (número de columnas de la tabla), n_{ij} las frecuencias observadas de cada celdilla de la tabla, n_i el tamaño de la muestra extraída de la población i-ésima y m_j la frecuencia absoluta de la clase j-ésima.

Para calcular el valor del estadístico de Pearson utilizaremos el software R. Primero introducimos los datos según se indica en `(1)`; después se ejecuta `(2)` y, finalmente obtenemos el p-valor del test en `(3)`.

Observamos que, como la tabla es 2×2, si no decimos nada, como ocurre en `(4)`, R calcula el valor del estadístico con la corrección de Yates, como obtenemos en `(5)`.

```
> X<-matrix(c(42,10,17,21),ncol=2)                                    (1)
> chisq.test(X,correct=FALSE)                                         (2)

        Pearson's Chi-square test

data:  X
X-squared = 12.6241, df = 1, p-value = 0.0003808                      (3)

> chisq.test(X)                                                       (4)

        Pearson's Chi-square test with Yates' continuity correction

data:  X
```

```
X-squared = 11.0788, df = 1, p-value = 0.0008732                    (5)
```

Un p-valor tan bajo sugiere rechazar claramente la hipótesis nula de homogeneidad de ambas poblaciones y concluir que sí había diferencias significativas entre ambas Zonas a la hora de hacer ese tipo de representaciones gráficas.

Problema 1.5

Se cree que el tamaño X de los asentamientos prehistóricos puede servir para predecir el tamaño de la población Y del lugar donde aquéllos se produjeron. Por ello se quiere determinar la recta de regresión basándose en datos actuales y, con ella, hacer estimaciones de tiempos pasados. Con este propósito, Carothers y McDonald (1979) obtuvieron los siguientes 20 datos de Tamaño de Asentamientos en hectáreas, X y Número de habitantes, Y, de los pueblos actuales del área en estudio:

X	0′6	1′0	1′1	1′2	1′6	1′9	2′3	3′0	3′1	3′3
Y	20	70	100	130	120	170	195	190	210	360

X	3′7	4′0	4′5	5′4	5′9	6′1	6′4	8′9	10′0	12′0
Y	300	250	500	270	190	630	650	310	730	850

Determinar la recta de regresión y analizar si es significativa.

Vamos a resolver este ejemplo con R (EBR-sección 10.4). Para ello, primero introducimos los datos como sigue,

```
> x<-c(0.6,1,1.1,1.2,1.6,1.9,2.3,3,3.1,3.3,3.7,4,4.5,5.4,5.9,6.1,6.4,8.9,10,12)
> y<-c(20,70,100,130,120,170,195,190,210,360,300,250,500,270,190,630,650,310,
+ 730,850)
```

Luego obtenemos la recta de regresión ejecutando `(1)` y analizamos los resultados ejecutando `(2)`,

```
> recta<-lm(y~x)                                                    (1)
> recta
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      39.79        63.36


> summary(recta)                                                    (2)
Call:
lm(formula = y ~ x)

Residuals:
       Min         1Q     Median         3Q        Max
```

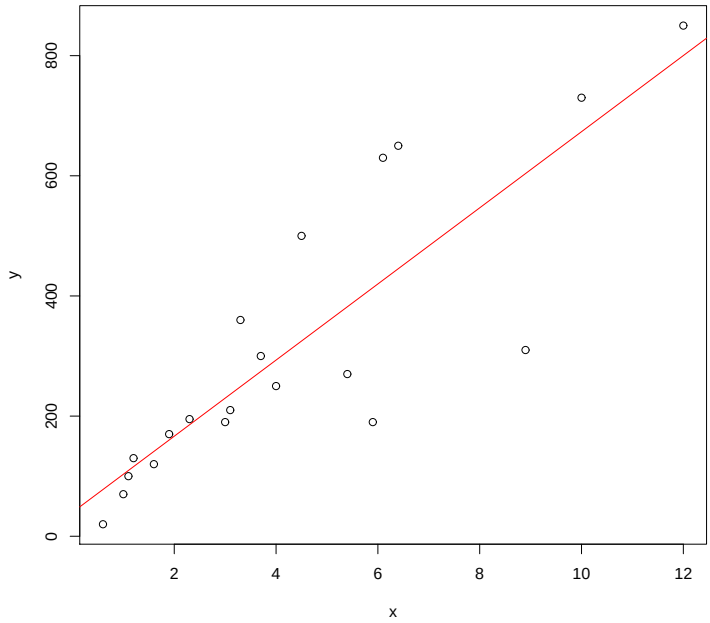

Figura 1.6 : Nube de puntos y recta de regresión

```
-2.937e+02 -4.072e+01 -7.797e-03  5.154e+01  2.047e+02

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.792     48.768   0.816    0.425                          (3)
x             63.362      9.214   6.877 1.97e-06 ***                      (4)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 127.2 on 18 degrees of freedom
Multiple R-Squared: 0.7243,     Adjusted R-squared: 0.709
F-statistic: 47.29 on 1 and 18 degrees of freedom, p-value: 1.969e-006
                                                                         (5)
```

La recta de regresión (o mínimos cuadrados) obtenida con la primera sentencia es

$$y_t = 39'79 + 63'36\, x$$

En la segunda salida vemos un contraste sin interés, el `(3)`, sobre la ordenada en el origen (*Intercept*) y un contraste, el `(4)`, sobre el coeficiente de regresión, es decir, el de la hipótesis nula $H_0 : \beta = 0$ frente a la alternativa $H_1 : \beta \neq 0$. Este contraste indica claramente rechazar H_0, es decir, rechazar

que la covariable independiente X no sea válida para explicar a la variable dependiente Y (es decir, rechazar claramente que la recta de regresión no sea válida), al ser el p-valor muy bajo.

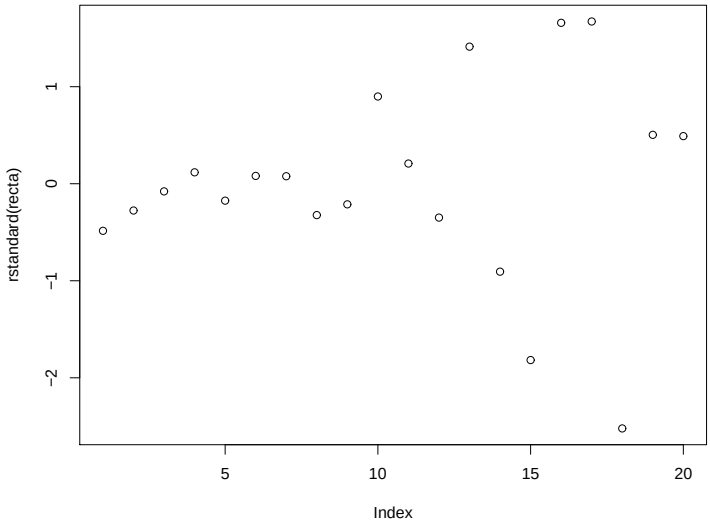

Figura 1.7 : Gráfico de Residuos Estandarizados

Como vemos en la Figura 1.6, obtenida ejecutando las sentencias

```
> plot(x,y)
> abline(recta,col=2)
```

el ajuste de la recta es bastante bueno.

Obsérvese que este contraste es exactamente el mismo que el obtenido en (5) basado en la F de Snedecor, por tratarse de una Regresión Lineal Simple. Para la Regresión Múltiple, el de la F de Snedecor será un contraste global del hiperplano ajustado, sin mucho interés, mientras que los contrastes de interés serán los tests sobre los coeficientes de regresión. Por tanto, para estos datos vemos que la recta de regresión se ajusta bien, lo que permitiría realizar predicciones sobre tiempos pasados, supuesto que en ellos se daba la misma relación de dependencia entre el tamaño de los asentamientos y la población.

Una cuestión adicional pocas veces considerada es el análisis de los residuos, es decir, de las diferencias entre los valores observados y_i y los teóricos obtenidos con las recta ajustada y_{t_i}. Habitualmente se utilizan, para detectar los datos anómalos, los residuos estandarizados (estimados) $(y_i - y_{t_i})/\hat{\sigma}$, que podemos representar ejecutando

```
> plot(rstandard(recta))
```

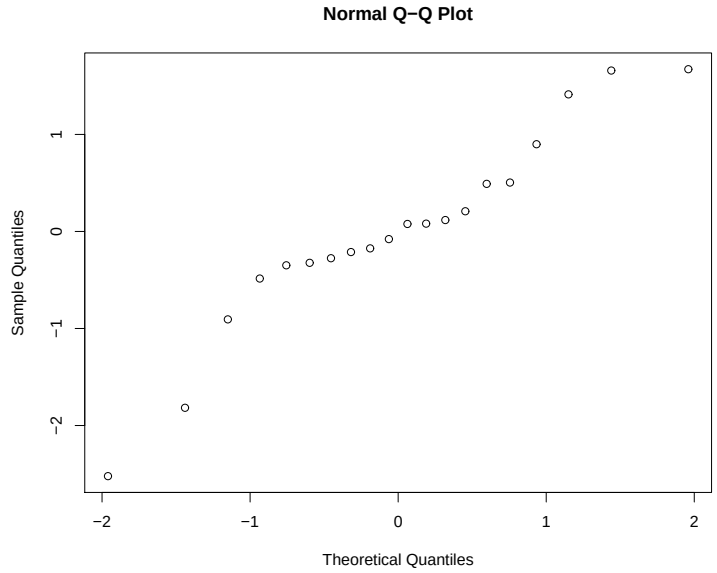

Figura 1.8 : Gráfico de Normalidad

obteniendo la Figura 1.7. En este gráfico se aprecia que el antepenúltimo dato, el $(8'9, 310)$, es un outlier, puesto que su residuo está más allá del valor $(-1'96, 1'96)$, que es un intervalo de confianza al 95 % de una normal $N(0, 1)$, distribución de donde teóricamente vienen los residuos, suposición también un tanto dudosa observando la Figura 1.8 que es el qq-plot de estos residuos, figura obtenida ejecutando la siguiente sentencia,

```
> qqplot(rstandard(recta))
```

debido a que los puntos no parecen estar en la diagonal del dibujo (CB-sección 14.4.1).

Estos gráficos parecen indicar que debemos utilizar alguna recta de regresión robusta, (EARR-capítulo 7) por ejemplo de la librería MASS; en concreto la recta de Huber, obtenida ejecutando `(6)`, la LMS ejecutando `(7)` y la recta LTS ejecutando `(8)`. Estas rectas se han añadido a la Figura 1.6 con las sentencias que siguen a su obtención, consiguiendo finalmente la Figura 1.9 en donde apenas si se aprecian diferencias entre las rectas robustas:

Regresión de Huber : $y_t = 19'864 + 69'03\, x$
Least Median of Squares, LMS : $y_t = 23'41 + 68'18\, x$
Least Trimmed Squares, LTS : $y_t = 17'65 + 70'19\, x$

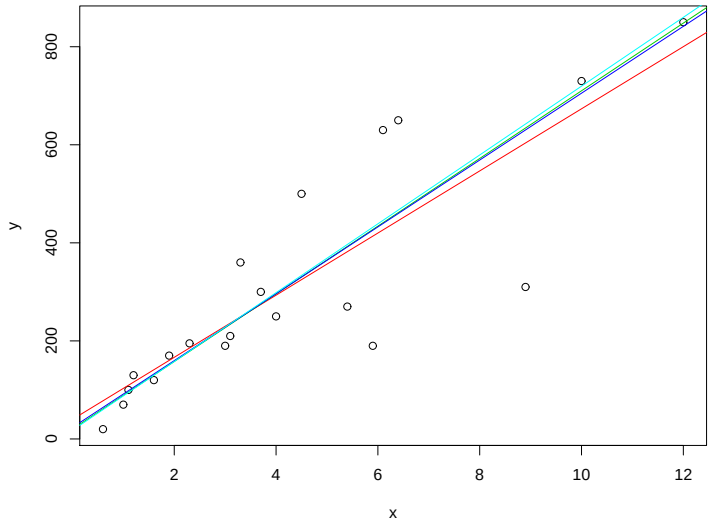

Figura 1.9 : Nube de puntos y rectas ajustadas

y sí entre éstas y la de mínimos cuadrados, la cual es *atraída* por el dato anómalo.

```
> library(MASS)
> recta2<-rlm(y~x)                                    (6)
> recta3<-lqs(y~x,method="lms")                       (7)
> recta4<-lqs(y~x)                                    (8)
> abline(recta2,col=3)
> abline(recta3,col=4)
> abline(recta4,col=5)
```

Problema 1.6

La Obsidiana es un mineral de origen volcánico al que los pobladores de Mesoamérica, en la época prehispánica, atribuían propiedades mitológicas (a causa de la leyenda del guerrero Obsid) y era muy utilizado en la fabricación de elementos de caza y defensa (tales como puntas de flecha, de lanza, raspadores, cuchillos) así como objetos rituales. En Arqueología se cree que los dos factores que influyen en la Densidad en gramos, Y, de hallazgos de este mineral en los yacimientos arqueológicos son, la Distancia en kilómetros, X_1, a la cual se hallaba la fuente de donde se extraía el mineral y el Tamaño en hectáreas, X_2, del asentamiento.

a) Se hizo primero un ensayo en el que se examinaron cinco asentamientos y se obtuvieron los siguientes datos: