Prácticas de Estadística Descriptiva y Análisis de Datos

Diplomatura en Estadística
Universidad Carlos III de Madrid
Curso 2007/08

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid

Práctica 1. Introducción a Statgraphics

1. Los datos de la siguiente tabla representan el sueldo actual y el sueldo inicial, en euros, de 10 empleados de una empresa.

Sexo	Sueldo actual	Sueldo inicial
H	57000	27000
H	40200	18750
M	21450	12000
M	21900	13200
H	45000	21000
H	32100	13500
H	36000	18750
M	21900	9750
M	27900	12750
M	24000	13500

a) Cread un fichero con los datos anteriores.
b) Calculad una nueva variable DIFERENCIA que represente la diferencia del sueldo actual menos el sueldo inicial.
c) Ordenad el fichero respecto del sueldo inicial.
d) Generad una variable categórica (AUMENTO) para clasificar a los individuos en tres categorías en función del aumento del sueldo. Definid las tres categorías como sigue:

1. Aumento inferior a 10.000 €.
2. Aumento superior o igual a 10.000 €, pero inferior a 20.000 €.
3. Aumento superior o igual a 20.000 €.

e) Calculad el promedio del sueldo actual.
f) Calculad el promedio del sueldo actual, pero sólo para las personas que han tenido un aumento de sueldo superior a 20.000 €.
g) Calculad el promedio de la variable DIFERENCIA pero de forma separada para el grupo de hombres y para el grupo de mujeres.

Práctica 2. Descriptiva univariante

1. Un corrector de textos contabiliza el número de erratas que encuentra en cada página. Después de pasar este corrector por un texto de 50 páginas, se obtiene el siguiente número de erratas por página:

2	3	5	0	1	4	0	6	2	1
1	0	2	4	5	3	1	2	3	2
2	5	4	1	3	2	6	8	2	0
1	0	2	3	1	5	10	2	1	3
3	1	2	4	4	6	2	0	1	3

a) A partir del enunciado del problema, introducid una variable estadística y decidid de qué tipo es.
b) Construid la tabla de frecuencias correspondiente.

	Frec. absoluta	Frec. relativa	Frec. relativa acumulada
0			
1			
2			
3			
4			
5			
6			
8			
10			
Total	50	1	

c) ¿Cómo sería el diagrama de barras de las frecuencias absolutas? Solamente mirando este gráfico, ¿sabríais decir cuánto vale la moda? ¿Cómo seria el polígono de frecuencias de esta distribución?
d) ¿Qué porcentaje de páginas, respecto del total de las que se han corregido, tienen 2 erratas?
e) ¿Qué porcentaje de páginas respecto del total tienen menos de 6 erratas? ¿Y 6 erratas o más?
f) ¿Qué porcentaje de páginas respecto del total tienen como mínimo 5 erratas?
g) Calculad los siguientes estadísticos descriptivos: media aritmética, mediana, moda, varianza y cuartiles.
h) Si comparáis la media aritmética y la mediana, ¿qué podéis decir de la simetría de la distribución? ¿Entre qué valores se encuentran el 50% central de los datos?
i) Construid un diagrama de tallo y hojas y un diagrama de caja. ¿Hay algún valor anómalo? ¿Cuál es? Fijaos en la asimetría de la distribución: ¿presenta una cola hacia los valores grandes o hacia los valores pequeños?

2. Las siguientes medidas corresponden a las alturas de cincuenta niños y niñas.

Estatura (en metros)				
1.56	1.59	1.63	1.62	1.65
1.61	1.59	1.51	1.62	1.62
1.53	1.49	1.57	1.54	1.53
1.59	1.58	1.57	1.47	1.64
1.55	1.59	1.53	1.56	1.53
1.47	1.57	1.60	1.54	1.56
1.50	1.62	1.59	1.62	1.54
1.68	1.52	1.62	1.59	1.49
1.65	1.53	1.59	1.56	1.54
1.58	1.52	1.63	1.56	1.62

a) Construid las tablas de frecuencias absolutas y relativas.
b) Obtened también las frecuencias acumuladas.
c) Representad las distribuciones de frecuencias anteriores mediante histogramas.
d) Dibujad los correspondientes polígonos de frecuencias.

e) Hallad, a partir del polígono de frecuencias acumuladas, la proporción de observaciones entre 1.59 y 1.62, ambas inclusive.
f) ¿Qué conclusiones pueden sacarse?

3. Los siguientes datos corresponden al *gasto en vestido y calzado* (en pesetas) de 75 familias españolas (datos de 1973).

6880	2620	1000	7980	8080	43100	19976
18832	28400	31141	49760	15076	15220	0
8144	19941	24072	11804	28236	18160	5900
5600	11980	128640	0	7000	1200	3040
46600	27480	408	2220	61000	6480	10080
9660	11080	10000	24656	0	8000	22400
4476	0	8000	22400	4476	6000	16480
9414	60940	63600	10400	0	2040	39868
18360	0	9301	42280	13500	3480	9780
17972	3760	20224	728	9200	1200	113200
30392	12172	21336	22840	6360	14360	78161
3840	24527					

a) Obtened un histograma con clases iguales.
b) Señalad en el histograma anterior el área correspondiente a valores del *gasto* superiores a 30000 pesetas. ¿Qué porcentaje del total de las observaciones corresponde a estos datos?
c) Haced lo mismo con los valores entre 20000 y 60000 pesetas. ¿Qué proporción del total de los datos representan?
d) ¿Es simétrica la distribución de la variable?
e) Hallad la mediana y el rango intercuartílico.
f) Dibujad el diagrama de caja. ¿Qué tipo se asimetría presenta la variable?
g) ¿Existen datos atípicos? En caso afirmativo, ¿cuáles son?
h) ¿Hay datos atípicos extremos? ¿Cuáles son?
i) Obtener la media de las observaciones. ¿Cómo es la posición relativa de la media y la mediana? ¿Por qué?
j) Interpretad los resultados.

Práctica 3. Descriptiva por subgrupos.

1. Una refinería de petróleo de San Francisco midió durante 31 días los niveles de monóxido de carbono que emitía una de sus chimeneas entre el 16 de abril y el 16 de mayo de 1993. Las medidas fueron enviadas como prueba para establecer una línea de actuación por parte del *Bay Area Air Quality Management District* (BAAQMD). Independientemente, el personal del BAAQMD hizo 9 medidas de los niveles de monóxido de carbono emitidos por la misma chimenea, pero en el período de tiempo del 11 de septiembre de 1990 hasta el 30 de marzo de 1993.
En este caso la refinería tenía un incentivo para sobreestimar las emisiones de monóxido de carbono. Encontraréis los daros en el fichero **airpolut.xls**.
Nombre de las variables:

1. ***CO***: monóxido de carbono (en partes por millón)
2. ***source*** : equipo que realizó la medida (*refinery, BAAQMD*)
3. ***date***: fecha en que se realizó la medida

a) Calculad la media, la mediana y la moda de la variable *CO* para cada uno de los grupos que define la variable *source*. Comparadlas.
b) Comparad las dispersiones de ambos grupos de medidas.
c) Haced un diagrama de caja para cada uno de los grupos y comparadlos.

2. La siguiente tabla contiene valores de indicadores del nivel de vida de algunos estados de Europa (datos de 1970-71): **renta** por habitante en dólares, consumo de **energía** en Kwh., número de **vehículo**s por cada 100 habitantes (V), número de **teléfono**s por cada 100 habitantes y número de habitantes por cada cama de **hospital**.

país	renta	energía	vehículo	teléfono	hospital
Alemania	3168	5112	33.3	22.5	87
Bélgica	2726	5929	25.0	21.1	125
Austria	1993	3419	20.0	19.3	92
Dinamarca	3192	5924	33.3	33.9	112
España	998	1495	12.5	13.5	218
Finlandia	2178	4150	20.0	24.9	72
Francia	2606	5794	33.3	17.2	113
Gran Bretaña	2249	5112	25.0	26.7	106
Grecia	998	1259	3.8	12.0	164
Italia	1723	2681	25.0	17.1	99
Irlanda	1442	2993	14.3	10.4	75
Noruega	2553	4814	25.0	29.4	109
Países Bajos	2877	5073	25.0	26.0	192
Portugal	673	687	7.1	7.8	177
Suecia	4032	6311	33.3	53.7	67
Suiza	2859	3353	25.0	48.2	88

Cread un fichero de nombre **europa.sf3** que contenga la información de la tabla anterior.

a) Cread una nueva variable que se llame *situación* a la derecha de la variable *país.* Llenad los valores para cada individuo teniendo en cuenta que esta nueva variable sólo puede tomar 3 valores diferentes (norte, centro y sur) y debe reflejar la situación sobre el mapa de Europa donde se encuentra cada país (por ejemplo, para Dinamarca vale norte, para Portugal vale sur, etc.).
b) Recodificad la variable *renta* en una nueva variable de nombre *rico* que indique si un país es pobre o rico. Definid los intervalos de la siguiente manera: si la renta pertenece al intervalo [500, 1500) diremos que el país es pobre, si la renta toma un valor del intervalo [1500, 3000) diremos que el país es normal y si la renta toma un valor del intervalo [3000, 4500] diremos que el país es rico.
c) Comparad las medias de las variables *renta* y *hospital* según los grupos que define la variable *situación*. Repetid lo mismo, pero ahora según los grupos que define la variable *rico*. ¿Qué conclusiones podéis sacar?

d) Haced un gráfico de barras que represente la media de la variable *hospital* según la variable *situación*. Repetid lo mismo, pero ahora según los grupos que define la variable *rico*.
e) Haced diagramas de caja para la variable *energía* según los grupos que define la variable *situación*. ¿Qué grupo presenta más dispersión?

3: El control del crecimiento de la población de mustang en las tierras federales no está libre de controversia. Un método para controlar la superpoblación consiste en esterilizar al macho dominante de cada grupo. Eagle, Asa, and Garrott et al. (1993) llevaron a cabo un experimento para evaluar la efectividad de la esterilización de los machos dominantes con el fin de reducir el nacimiento de potros durante dos o más años.
Para realizar este estudio, los investigadores escogieron dos zonas donde había las manadas de caballos salvajes: Flanigan, en el noroeste de Nevada y Beauty Butte, en el sureste de Oregon. En diciembre de 1985, acorralaron a los caballos en grupos y los contaron, determinaron su sexo, y estimaron sus edades mirándoles sus dentaduras. Hicieron fotos a todos los caballos de tres o más años y los identificaron mediante collares para poder hacer un seguimiento en el estudio. En cada grupo identificaron al macho dominante, lo esterilizaron y le pusieron un collar con un transmisor de radio. Finalmente, soltaron los grupos. Entre junio de 1986 y julio de 1988 probaron de localizar a cada macho esterilizado de 3 a 4 veces por año, desde un helicóptero. Los investigadores registraron el número de adultos y de potros en cada grupo que tenía un macho esterilizado (grupos tratados) y en cada grupo que no tenía un macho esterilizado (grupos no tratados).
A pesar de que los investigadores no pudieron registrar la tasa de nacimientos en los grupos de caballos, el porcentaje de potros en cada grupo es una buena aproximación.

Referencias: Eagle, T. C., Asa, C., and Garrott, R. et al. (1993), "Efficacy of Dominant Male Sterilization To Reduce Reproduction in Feral Horses," Wildlife Society Bulletin, 21(2), 116-121.

Encontraréis los datos de este estudio en el fichero **wildhorses.xls**.
Nombre de las variables:

1. **Adults**: número total de adultos en cada grupo
2. **Sterile_Males**: número de machos esterilizados encontrados en cada grupo
3. **Foals**: número de potros en cada grupo
4. **Year**: año
5. **Location**: F zona de manadas salvajes de Flanigan; B zona de manadas salvajes de Beaty Butte
6. **Date**: fecha de la observación
7. **Treatmen**t: 1 si está esterilizado (grupo tratado); 0 grupo no tratado.

a) Calculad la media, la varianza y los cuartiles para las variables *Adults* y *Foals* según las categorías de la variable *Treatment.* Haced también diagramas de caja. ¿Qué conclusión podéis sacar?
b) Para cada tratamiento, comparad las variables *Adults* y *Foals* según las categorías de la variable *Location.* Utilizad las mismas medidas numéricas y los mismos gráficos que en el apartado a). ¿Cómo son las distribuciones de caballos adultos y de potros en las dos zonas? ¿Se parecen (en cuanto a número, simetría, dispersión) o son muy diferentes?
c) Para cada tratamiento, comparad los percentiles 70 80 y 90 para la variable *Foals.*

Práctica 4. Transformaciones.

1. Los datos siguientes expresan el número de días transcurridos hasta la primera avería en cierto tipo de electrodomésticos:

534	875	435	654	432	984	321	765	453
765	564	982	873	567	871	658	564	399

a) Calculad la media, desviación típica, mediana y rango intercuartílico de las observaciones.
b) Encontrad la transformación lineal de la variable que representa el tiempo de duración en semanas.
c) Obtened la media, desviación típica, mediana y rango intercuartílico de los datos transformados. ¿Qué relación guardan con los valores obtenidos usando los datos originales?

2. Utilizad los datos del **ejercicio 2** de la **práctica 2**.
a) Hallad la media y desviación típica de estos datos.
b) Tipificad las observaciones.
c) Comprobad que la media de los datos tipificados es cero y que su desviación es uno.

3. Utilizad los datos del **ejercicio 3** de la **práctica 2.**
a) ¿Qué tipo de transformaciones pueden utilizarse para hacer más simétrica la distribución de la variable *gasto*?
b) Transformad los datos mediante las transformaciones propuestas en el apartado a)
c) ¿Cuál da lugar a una distribución más simétrica?
d) ¿Qué ocurre con los datos atípicos al aplicar la transformación del apartado c)?

4. La siguiente tabla muestra la esperanza de vida masculina en diferentes países del mundo.

País	*Esph*	*País*	*Esph*	*País*	*Esph*
Mozambique	43	Rep. Centro-afric.	45	Camerún	54
Sierra Leona	41	Pakistán	59	Rep. Kirguiz	62
Etiopía	49	Ghana	54	Georgia	69
Nepal	54	China	68	Uzbekistán	66
Tanzania	49	Tajikistán	67	Papúa Nueva-G	55
Uganda	43	Guinea	44	Perú	63
Bután	48	Mauritania	46	Guatemala	62
Burundi	46	Sri-Lanka	70	Congo	49
Malawi	44	Zimbawe	58	Marruecos	62
Bangladesh	55	Honduras	64	Rep. Dominicana	65
Chad	46	Lesotho	58	Ecuador	65
Guinea-Bisau	38	Rep. Egipto	60	Jordania	68
Madagascar	50	Indonesia	59	Rumania	67
Laos PDR	50	Myanmar	58	El Salvador	64
Ruanda	45	Somalia	47	Turkmenistán	63
Níger	44	Sudán	51	Moldavia	65
Burkina-Faso	47	Yemen	52	Lituania	66
India	61	Zambia	46	Bulgaria	68
Kenia	57	Costa de Marfil	53	Colombia	66
Mali	47	Bolivia	58	Jamaica	71
Nigeria	50	Azerbaiyán	67	Paraguay	65
Nicaragua	65	Filipinas	63	Namibia	58
Togo	53	Armenia	67	Kazajstán	64
Benín	49	Senegal	48	Túnez	67
Ucrania	66	Lituania	64	Rep. Irán	65
Argelia	67	Rep. Eslovaca	67	Panamá	71
Tailandia	67	Costa Rica	74	Rep. Checa	69
Polonia	66	Turquía	65	Fed. Rusa	64
…		…		…	

País	*Esph*	*País*	*Esph*	*País*	*Esph*
Chile	69	Argentina	68	Italia	74
Albania	70	Omán	68	Holanda	74
Mongolia	62	Eslovenia	69	Canadá	75
Rep. Siria	65	Puerto Rico	71	Bélgica	72
Sudáfrica	60	Corea	65	Finlandia	72
Estonia	65	Grecia	75	Emiratos Árabes	70
Brasil	64	Portugal	70	Francia	73
Botswana	66	Arabia Saudí	68	Austria	73
Malasia	69	Irlanda	73	Alemania	73
Venezuela	67	Nueva Zelanda	73	Estados Unidos	73
Bielorrusia	67	Israel	75	Noruega	74
Hungría	65	España	73	Dinamarca	72
Uruguay	69	Hong Kong	75	Suecia	75
México	67	Singapur	72	Japón	76
Trinidad-Tobago	69	Australia	74	Suiza	75
Gabón	52	Reino Unido	73		

a) Obtened el diagrama de caja de los datos.

b) ¿Qué tipo de transformaciones pueden utilizarse para disminuir la asimetría que presentan las observaciones?

c) Seleccionad de entre las anteriores la más adecuada para este conjunto de datos.

Práctica 5. Descripción conjunta de varias variables.

1. En un estudio sobre parásitos se consideró el número de garrapatas en el cuerpo de ratones. En un grupo de 44 ratones se obtuvieron los siguientes resultados:

Número de garrapatas:

0	1	1	1	0	2	3	4	0	5	0
0	0	0	1	0	0	0	2	3	2	1
1	2	0	2	0	4	1	1	0	0	2
1	0	0	1	0	0	1	1	0	0	6

Sexo:

M	H	H	H	H	M	M	M	M	H	H
H	H	H	M	M	H	M	H	H	M	M
M	H	H	H	M	M	H	H	H	M	H
H	M	M	H	H	M	M	H	M	M	H

a) ¿De qué tipo son las variables número *de garrapatas* y *sexo*?

b) Para la variable número *de garrapatas* calculad la media aritmética, la mediana, los cuartiles, la moda, el rango, la varianza, la desviación típica, el coeficiente de asimetría y el coeficiente de curtosis. ¿Qué información podéis sacar? ¿Para la variable *sexo* pueden calcularse también todas estas medidas?

c) Construid la tabla de frecuencias para la variable número *de garrapatas,* de manera que los puntos medios de los intervalos de clase coincidan con los valores que toma la variable. ¿Cuántos ratones tienen exactamente 3 garrapatas? ¿Qué porcentaje de ratones tienen más de cuatro garrapatas? ¿Qué porcentaje de ratones tienen al menos una garrapata? Haced la representación gráfica adecuada para este tipo de variable.

d) Haced un diagrama de caja para la variable número *de garrapatas.* ¿Hay valores anómalos?

e) Construid la tabla de frecuencias para la variable *sexo* y haced la representación gráfica correspondiente para este tipo de variable.

f) Obtened la tabla de contingencias para la variable *sexo* en las filas y la variable número *de garrapatas* en las columnas. ¿Cuántos ratones macho tienen 4 garrapatas? ¿Qué porcentaje representan respecto del total de machos? ¿y respecto del total de ratones? ¿Cuántos ratones hembra tienen más de 5 garrapatas? ¿Qué porcentaje representan respecta del total de hembras? ¿Qué porcentaje representan respecto del total de ratones?

g) Comprobad que las distribuciones marginales de las variables número *de garrapatas* y *sexo* coinciden con las tablas que habéis obtenido en los apartados c) y e).

2. Utilizad el fichero **europa.sf3**. Trabajaremos con las variables *renta, energía, vehículos y hospital*.

a) Haced un diagrama de dispersión matricial de estas. ¿Qué idea nos da este gráfico de la relación entre estas variables?

b) Calculad el coeficiente de asimetría y el coeficiente de curtosis. ¿Qué medida de dispersión usaríais para comparar estas variables? ¿Cuál presenta menos dispersión?

c) Calculad la matriz de covarianzas. Fijaos en las covarianzas de los pares (*renta, energía*) y (*renta, vehículos*). ¿Diríais que hay alguna que sea muy elevada?

d) Calculad la matriz de correlaciones. Fijaos en las correlaciones de los pares (*renta,energía*) y (*renta,vehículos*). ¿Diríais que son muy diferentes estas correlaciones? ¿Por qué a partir de la matriz de covarianzas no quedan tan claras estas relaciones entre las variables?

e) ¿Hay alguna variable que tenga correlación negativa? ¿Qué significado tiene el signo del coeficiente de correlación?

f) Haced un diagrama de dispersión de la variable *renta* sobre la variable *energía*, etiquetad los individuos mediante la variable *país*. ¿Quien son los individuos extremos, es decir, el más rico y el más pobre? ¿Qué relación parece que tengan las variables *renta* y *energía?*

g) Haced un diagrama de dispersión pero ahora de las parejas *hospital-renta* y *vehículos-renta.*

3. Usad el fichero **bebes.sf3**. Este fichero contiene el estudio realizado sobre 18 bebés de madres fumadoras y no fumadoras. De cada bebé se midió el *peso* (en Kg.) la *altura* (en cm.), si la madre *fuma* o no, el *grupo sanguíneo* y el número *de hermanos*.

a) Haced un estudio descriptivo (medidas numéricas y gráficos) de las variables *peso y altura* para el grupo de madres fumadoras y para el grupo de madres no fumadoras. ¿Qué conclusiones podéis sacar?
b) Recodificad la variable *peso* de la forma: si el peso pertenece al intervalo [0,2.75) consideraremos que el bebé es *pequeño*, si el peso pertenece al intervalo [2.75, 3.75), consideraremos que es *normal,* y si el peso pertenece al intervalo [3.75, 5), consideraremos que el bebé es *grande*. Llamaremos *codi_peso* a esta nueva variable.
c) Haced una tabla de contingencias con la variable *fuma* en las filas y la variable *codi_peso* en las columnas. Calculad el estadístico Chi-cuadrado y las medidas de asociación que creáis oportunas para decidir si hay alguna relación entre el peso de los bebés y el hecho que la madre sea fumadora.
d) Haced una tabla de contingencias con la variable *fuma* en las filas y la variable *grupo sanguíneo* en las columnas Decidid si hay alguna relación entre estas dos variables.

4. Usad el fichero **contami.sf3.** Con el fin de estudiar la relación entre el grado de contaminación ambiental y la climatología se han recogido datos durante 200 días y se han clasificado según el grado de *contaminación* (1=alta, 2=media, 3=baja) y según la *nubosidad* (1=intensa, 2=débil, 3=inexistente).

a) Construid la tabla de contingencias con la variable *contaminación* en files y la variable *nubosidad* en las columnas, de manera que se muestren los porcentajes de los diferentes grados de contaminación según el tipo de nubosidad.
b) Estudiad si hay dependencia del grado de contaminación en función de la nubosidad.
c) Recodificad la variable *nubosidad* en una de nueva que sólo tenga dos categorías (1=intensa, 2=débil/inexistente). Analizad si varía el resultado.

5. Las clasificaciones de 7 ciclistas en dos pruebas de pista han sido:

Ciclista	Prueba 1	Prueba 2
Álvarez	4	7
Díaz	2	5
Fernández	7	2
Gómez	6	6
Jiménez	1	4
López	5	1
Martínez	3	3

Usad Excel para analizar el grado de semejanza entre estas dos clasificaciones.

6. La siguiente tabla muestra el resultado del estudio realizado en una ciudad sobre el tipo de vehículo y el sexo de su propietario:

	Hombre	Mujer
Utilitario	537	155
Familiar	85	11
Gran berlina	12	3
Deportivo	14	6
Monovolumen	6	1

Mediante Excel analizad si hay asociación o independencia entre el sexo y el tipo de vehículo. En caso de que haya asociación, determinad el grado de la misma.

Práctica 6. Relaciones entre variables.

1. Se han medido dos variables continuas sobre 15 países diferentes. La variable *lifeexpf* es la esperanza de vida de las mujeres y *birthrate* es el número de nacimientos cada 1000 habitantes por cada un de los países (datos de 1992).

Country	lifeexpf	birthrate
Somalia	55	46
Tanzania	55	50
Zambia	59	48
Zaire	56	45
Algeria	68	31
Namibia	63	45
Burkina Faso	53	50
Cuba	79	18
Ecuador	72	28
North Korea	72	24
Mongolia	68	34
Thailand	71	20
Turkey	72	28
France	82	13
Netherlands	81	13

Se quiere estudiar si existe algún tipo de relación entre el número de nacimientos y la esperanza de vida.

a) Realizad un gráfico de dispersión con el número de nacimientos como eje X y usad la variable *country* para identificar a los individuos. Mirando el gráfico, ¿cómo es la esperanza de vida de las mujeres para los países que tienen un alto número de nacimientos?

b) ¿Qué tipo de relación parece que haya entre estas variables?

c) ¿Os parece adecuado modelar este comportamiento mediante una recta?

d) Calculad la recta de regresión de la esperanza de vida sobre el número de nacimientos. Si llamamos **x** al número de nacimientos e **y** a la esperanza de vida, ¿cuál seria la recta de regresión?

e) Para un incremento de una unidad en el número de nacimientos, ¿cuántas unidades esperaríamos que aumentase o disminuyese la esperanza de vida?

f) Si en 1992 el número de nacimientos en España era de 11 por cada 1000 habitantes, ¿qué esperanza de vida esperaríamos encontrar para las mujeres españolas en 1992?

g) Si sabemos que la esperanza de vida de las mujeres españolas en 1992 era exactamente de 82, ¿qué error estemos cometiendo al utilizar la predicción?

h) Teniendo en cuenta que el número de nacimientos toma valores en el intervalo comprendido entre 13 y 50, ¿sería adecuado realizar predicciones de la esperanza de vida de las mujeres de países con 70 nacimientos cada 1000 habitantes?

i) ¿Qué podéis decir del ajuste de este modelo lineal?

j) ¿Qué porcentaje de la variabilidad de la esperanza de vida queda explicado por las diferencias entre el número de nacimientos?

k) Haced un gráfico de los residuos sobre el número de nacimientos. ¿Para que países se obtiene una peor predicción de la esperanza de vida?

2. Se quiere estudiar si la edad en que un niño empieza a hablar predice su inteligencia posterior. En un estudio a 21 niños se registraron la *edad* (en meses) en que empiezan a hablar y la *puntación* obtenida en una prueba de aptitud hecha más adelante. Los datos son los siguientes:

Niño	Edad	Puntuación
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

a) Haced un gráfico de dispersión de la *puntuación* respecto de la edad. ¿Creéis que la relación es lineal? La asociación entre las dos variables ¿es positiva o negativa? ¿Cómo lo interpretáis?

b) Calculad la recta de regresión de la *puntuació*n respecto de la edad. ¿Cuál es la pendiente de la recta? Valorad el ajuste del modelo.

c) Haced una predicción de la *puntuación* de la prueba para un niño que ha empezado a hablar a los 15 meses. Calculad el error de predicción. ¿Qué indica el hecho que el residuo sea positivo?

d) Haced un diagrama de residuos. ¿Hay algún dato atípico? ¿Hay algún dato influyente?

e) Eliminad de los datos la observación 18 y calculad R^2. ¿Qué conclusiones podéis sacar del resultado?

f) Volved a calcular la recta de regresión dejando fuera la observación 19. ¿Diríais que la observación 19 es influyente? ¿Qué efecto tiene la exclusión del niño 19 sobre R^2? ¿Cómo lo explicaríais?

3. Se realiza un estudio del foto período (número de horas de luz al día) en aves acuáticas. Se quiere establecer una ecuación para predecir el tiempo de reproducción en función del foto período con el que se inició la reproducción. Los resultados obtenidos para un grupo de 11 patos del género *Aythya* son:

Horas de luz al día	*Tiempo* de reproducción
12.8	110
13.9	54
14.1	98
14.7	50
15.0	67
15.1	58
16.0	52
16.5	50
16.6	43
17.2	15
17.9	28

a) Haced un gráfico de dispersión de la variable *tiempo* en función de la variable *horas.* ¿Qué tipo de relación parece que haya entre estas variables?
b) Calculad la recta de regresión de la variable *tiempo* sobre la variable *horas*.
c) Valorad el ajuste del modelo lineal.
d) Pera un número de horas de luz de 15.5, ¿cuál es la predicción para el tiempo de reproducción?

4. Se realiza un experimento sobre 10 ratones para estudiar el crecimiento en función de diferentes dosis de una dieta suplementaria.

Individuo	*Suplemento*	*Crecimiento*
1	10	73
2	10	78
3	15	85
4	20	90
5	20	91
6	25	87
7	25	86
8	25	91
9	30	75
10	35	65

a) Realizad un diagrama de dispersión de la variable *crecimiento* en función de la variable *suplemento*. ¿Qué tipo de relación parece que haya entre las variables?
b) Comprobad que un modelo polinómico proporciona mejor ajuste que un modelo lineal.