Estadística Descriptiva

Estadística e Investigación Operativa, 1º EII

Antoni Torres Signes

Área de Estadística e Investigación Operativa
Escuela de Ingenierías Industriales

UNIVERSIDAD
DE MÁLAGA

- *Estadística descriptiva*
  Parte de la estadística que se ocupa de organizar, graficar, resumir y estudiar la información contenida en unos datos recogidos.
- *Población*
  Conjunto de elementos objeto del estudio estadístico.
- *Individuo*
  Elemento de la población sobre el que se estudia una o varias características.
- *Muestra*
  Cualquier subconjunto de la población.
- *Tamaño de la muestra*
  Número de elementos que forman parte de la muestra, n.
  También se considera el tamaño de la población, N, que puede ser desconocido.

- *Variable estadística*
  Cada una de las características que pueden estudiarse en la población. Según los valores que tome, clasificamos las variables en distintos tipos.
  - *Variable cualitativa*
    Los resultados posibles no son valores numéricos.
  - *Variable cuantitativa*
    Toma valores numéricos. Se diferencian en dos tipos.
    - *Discreta*
      Toma valores aislados, cuyo número puede ser finito o infinito numerable.
    - *Continua*
      Toma cualquier valor dentro de un intervalo. Por comodidad, sus valores se agrupan en intervalos, también llamados *clases*. Surgen así conceptos como *límite de clase*, *marca de clase*, *amplitud o tamaño de clase*.

2.1. Tabla de frecuencias

- Organización de los datos
  La información se ordena en base a *modalidades* o *clases*, es decir, valores o intervalos de la variable que agrupan las posibles respuestas.
  - Por elementos
    Cuando no hay una gran variedad de modalidades.
  - Por intervalos
    Cuando los datos presentan una importante variedad.
- Tabla de frecuencias
  Para una variable, relaciona cada modalidad con la cantidad de resultados que se han presentado de dicha modalidad.
  - *Frecuencia absoluta*
    Para cada modalidad, $i = 1, \dots k$, es el número de individuos que pertenece a dicha modalidad, n_i.
  - *Frecuencia relativa*
    Cociente entre frecuencia absoluta y tamaño de muestra, $f_i = n_i/n$.

2.1. Tabla de frecuencias

Ejemplo 1 (Variable cualitativa)

Se realiza una encuesta a 90 jóvenes (entre 18 y 34 años) que ha realizado alguna acción en el mercado de la vivienda durante los últimos doce meses. En concreto, se estudia qué tipo de búsqueda se ha realizado, ya sea en el mercado de arrendamiento, vivienda en propiedad, o en ambos.

Tipo de búsqueda	Frecuencia absoluta	Frecuencia relativa
Alquiler	44	0.4889
Compra	32	0.3555
Ambos	14	0.1556

- *Frecuencia absoluta acumulada*
  Para cada modalidad, $i = 1, \dots k$, es el número de individuos que pertenece a dicha modalidad o anteriores a ella, N_i.
- *Frecuencia relativa acumulada*
  Cociente entre frecuencia absoluta acumulada y tamaño de muestra, $F_i = N_i/n$.

2.1. Tabla de frecuencias

Ejemplo 2 (Variable discreta)

Se ha observado el número de nacimientos en un determinado año, en los municipios de menos de 500 habitantes de Andalucía, obteniendo la siguiente tabla de frecuencias.

Número de nacimientos	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
0	31	0.2844	31	0.2844
1	37	0.3394	68	0.6238
2	21	0.1927	89	0.8165
3	12	0.1101	101	0.9266
4	3	0.0275	104	0.9541
5	4	0.0367	108	0.9908
11	1	0.0092	109	1

2.1. Tabla de frecuencias

Ejemplo 3 (Variable continua)

Se ha observado la renta neta media declarada en un determinado año, en los municipios de menos de 300 habitantes de Andalucía, obteniendo la siguiente tabla de frecuencias.

Renta neta media declarada	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[2000, 4000)	1	0.0213	1	0.0213
[4000, 6000)	1	0.0213	2	0.0426
[6000, 8000)	9	0.1915	11	0.2341
[8000, 10000)	22	0.4681	33	0.7022
[10000, 12000)	10	0.2127	43	0.9149
[12000, 14000)	3	0.0638	46	0.9787
[14000, 16000)	1	0.0213	47	1

- La información se resume de forma que represente el comportamiento global de los datos, que suelen tener un volumen importante.
- *Medidas estadísticas*
  Valores que resumen las características de un conjunto de datos y facilitan la interpretación de su comportamiento. Se denominan *parámetros*, cuando se refieren a la población, y *estadísticos*, si se refieren a la muestra.
  - *Medidas de centralización*
    Representan características centrales de la distribución de los datos: media, mediana y moda.
  - *Medidas de posición*
    Indican qué parte de la distribución queda a cada lado de estas medidas: cuartiles, deciles, percentiles.
  - *Medidas de dispersión*
    Informan de la separación o concentración de los datos: varianza, desviación típica, rango o recorrido, coeficiente de variación.

3.1. Medidas de centralización

Media

- La media de una variable X, que se denota por $\bar{x}$, es el promedio de los datos con los que estamos trabajando. Representa el centro de gravedad de la distribución.
- Si tenemos los datos $x_1, x_2, \ldots, x_n$, la media viene dada por la expresión

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Si tenemos los datos en una tabla de frecuencias, con k modalidades,

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \sum_{i=1}^{k} x_i f_i$$

3.1. Medidas de centralización

Ejemplo 4

Se ha observado la longitud de grietas en micras producidas en el acero de unas calderas, obteniendo los siguientes datos:

24.9	12.7	21.2	30.2	25.8	18.5	10.3
14.0	27.1	45.0	24.2	8.9	32.4	11.8

La longitud media de las grietas es

$$\bar{x} = \frac{24.9 + 12.7 + \cdots + 11.8}{14} = 21.9286 \text{ µm.}$$

3.1. Medidas de centralización

Mediana

- Considerando los datos ordenados, la mediana, Me, es el valor que comprende al menos la mitad de la distribución inferior y al menos la mitad de la distribución superior
- Si el número de datos es impar, una vez ordenados los datos, la mediana es el valor central, que ocupa la posición $(n+1)/2$. Si es par, la mediana es la media de los dos datos centrales, que ocupan las posiciones $n/2$ y $(n/2+1)$.
- En tablas de frecuencias, consideramos $n/2$ y las frecuencias absolutas acumuladas. El primer valor cuya frecuencia absoluta acumulada supera el valor $n/2$, es la mediana. Si coincide la frecuencia absoluta acumulada con $n/2$, la mediana es la media entre el valor donde se produce la coincidencia y el siguiente valor de la variable.

3.1. Medidas de centralización

Ejemplo 5

En una calle se ha medido el ruido, en decibelios, que se produce en diferentes tiempos aleatorios a lo largo de una jornada, obteniendo los siguientes resultados:

94	110	74	65	60	90	83	87	75	114
69	94	124	91	90	102	77	125	108	65

Primero ordenamos los veinte datos,

60	65	65	69	74	75	77	83	87	90
90	91	94	94	102	108	110	114	124	125

La mediana de decibelios observados es el valor medio entre las posiciones $n/2 = 10$ y $(n/2 + 1) = 11$,

$$Me = 90 \text{ dB}.$$

3.1. Medidas de centralización

Moda

- La moda, Mo, es el valor o intervalo (marca de clase) más observado.
- Si la información la tenemos en forma de tabla de frecuencias, la moda es aquel valor o intervalo con mayor frecuencia.
- Puede haber dos o más modalidades con mayor frecuencia, con lo que puede haber dos o más modas.

3.1. Medidas de centralización

Ejemplo 6

En una muestra de mezclas de asfalto se midió la tensión de fractura en megapascales, obteniendo los siguientos datos (ya ordenados):

75	79	80	105	126	138	179	179	191
232	232	236	242	245	247	274	384	470

Para estos datos muestrales hay dos modas: 179 y 232 MPa, ya que estos valores se repiten dos veces cada uno.

3.2. Medidas de posición

Cuartiles

- Los cuartiles son valores que dividen la distribución ordenada de datos en cuatro partes.
- El primer cuartil, Q_1, comprende, al menos, el 25 % de los datos inferiores y el 75 % de los datos superiores. De igual modo, el segundo cuartil, Q_2, parte la distribución al 50 %, y el tercer cuartil, Q_3, al 75 y 25 %, respectivamente.
- Los cuartiles se obtienen con un procedimiento similar al de la mediana. Para el primer cuartil, buscamos el primer valor cuya frecuencia absoluta acumulada supere $n/4$. Si, en vez de superar, se iguala, obtenemos la media de dicho valor de la variable y el siguiente valor. Para el segundo y tercer cuartil hacemos lo mismo con $2n/4$ y $3n/4$, respectivamente.

3.2. Medidas de posición

Percentiles

- Análogamente a los cuartiles, los percentiles son valores que dividen la distribución ordenada de datos en cien partes.
- Para $r = 1, \ldots, 99$, cada percentil, P_r, comprende, al menos, el $r\,\%$ de los datos inferiores y el $(100 - r)\,\%$ de los datos superiores.
- Los percentiles se obtienen con un procedimiento similar al de los cuartiles. Para el percentil de orden r, buscamos el primer valor cuya frecuencia absoluta acumulada supere $rn/100$. Si, en vez de superar, se iguala, obtenemos la media de dicho valor de la variable y el siguiente valor.

3.2. Medidas de posición

Ejemplo 7

Los siguientes datos, ya ordenados, muestran el tiempo de fallo, en horas, de un material aislante eléctrico:

228	252	324	444	720	816	912
1296	1392	1488	2520	2856	3192	3528

Para estos datos muestrales se obtienen el primer cuartil y los percentiles de orden 10, 50, y 72:

$$Q_1 = 444;\ P_{10} = 252;\ P_{50} = \frac{912 + 1296}{2} = 1104;\ P_{72} = 2520.$$

3.3. Medidas de dispersión

Rango

Se llama *rango* o *recorrido* a la diferencia entre el valor más grande y el más pequeño que se han recogido.

Rango intercuartílico

Se llama rango o recorrido *intercuartílico* a la diferencia entre el tercer y primer cuartil, $RI = Q_3 - Q_1$.

3.3. Medidas de dispersión

Varianza

- Representa la dispersión de las observaciones respecto de la media. Generalmente, se utiliza la varianza muestral, S^2,

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

también en su fórmula alternativa

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

- Si trabajamos con tablas de frecuencias, con k modalidades,

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{k} x_i^2 n_i - n\bar{x}^2\right).$$

3.3. Medidas de dispersión

Desviación típica

- La *desviación típica* se define como la raíz cuadrada de la varianza. Generalmente, se utiliza la desviación típica muestral, $s = \sqrt{s^2}$.
- Representa la misma idea de la varianza, con la ventaja de en vez de expresarse en unidades al cuadrado, se expresa con las mismas unidades que la variable.

3.3. Medidas de dispersión

Coeficiente de variación

- Se define el *coeficiente de variación* como el cociente entre la desviación típica y la media,

$$CV = \frac{s}{|\bar{x}|}.$$

- No está referido a ninguna unidad de medida, con lo que permite comparar distribuciones distintas.
- Se puede utilizar en tanto por ciento,

$$CV = \frac{s}{|\bar{x}|} \times 100\,\%.$$

3.3. Medidas de dispersión

Ejemplo 8

Los siguientes datos reflejan el número de accidentes anuales ocurridos en una carretera durante 13 años. Obtener las medidas de dispersión vistas.

22	25	24	18	20	16	12
12	13	8	20	16	9	

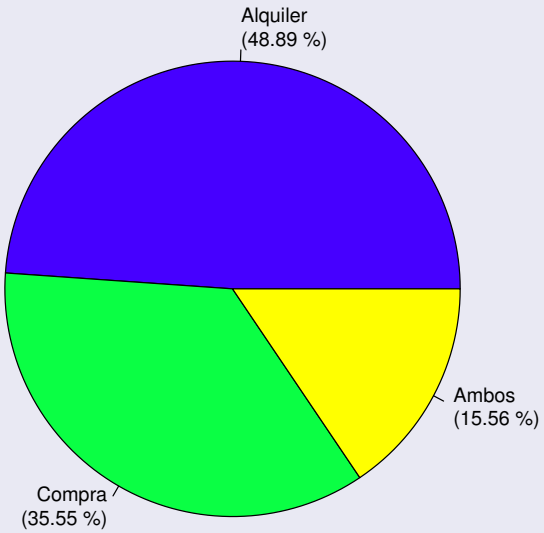
Se ha considerado también el coste, en miles de euros, de las compañías aseguradoras involucradas en los anteriores accidentes, obteniendo los siguientes resultados,

30.574	32.683	30.587	25.221	28.864	23.873	19.784
18.240	21.757	15.336	26.812	22.335	16.084	

¿Cuál de los dos conjuntos de datos tiene una media más representativa?

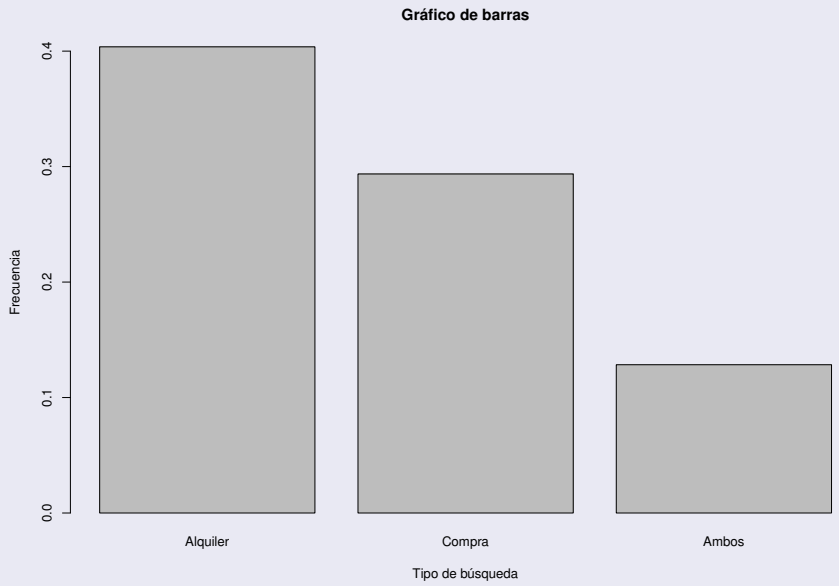
- Gráfico estadístico
  Representación visual de datos estadísticos que facilita la percepción de la información que desprenden los mismos.

Diagrama de sectores (variable cualitativa o cuantitativa discreta)



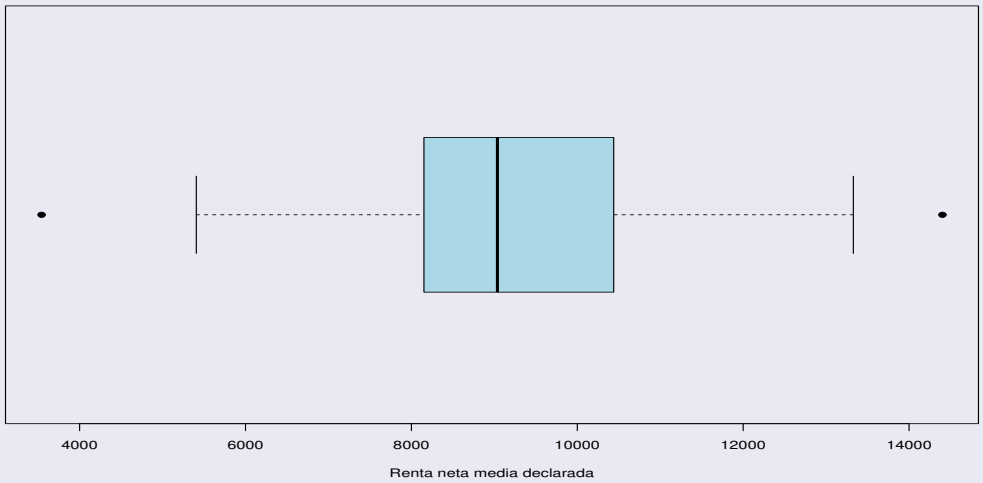
Se divide un círculo en sectores de área proporcional a las frecuencias de cada modalidad.

Diagrama de barras (variable cualitativa o cuantitativa discreta)



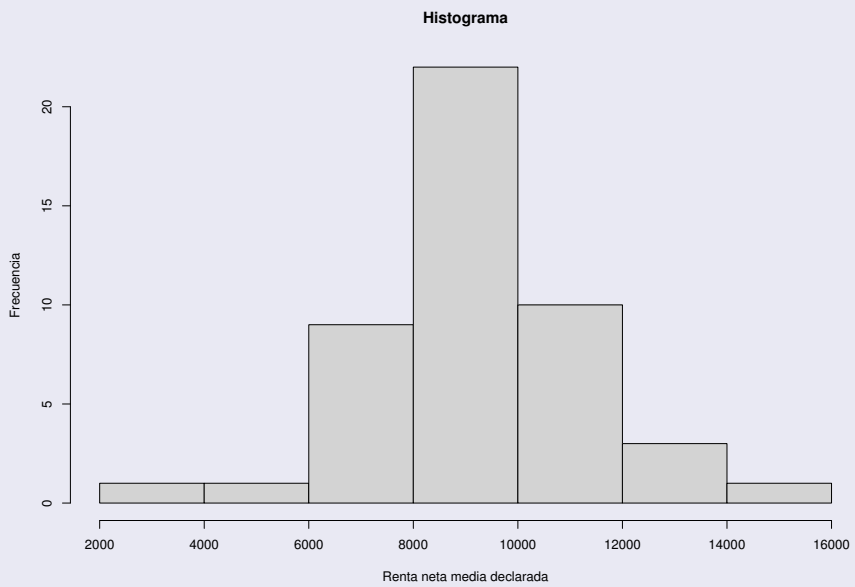
Representación gráfica sobre ejes de coordenadas a partir de barras rectangulares, con alturas que se corresponden a las frecuencias de cada modalidad.

Diagrama de caja y bigotes (variable cuantitativa)



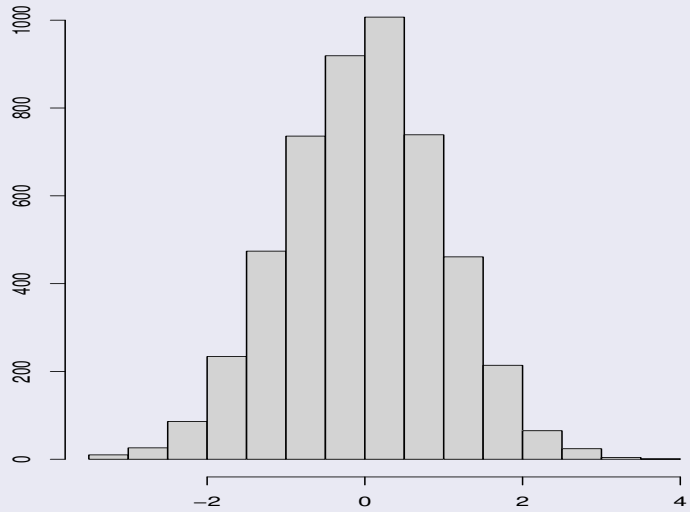
La caja representa el 50 % central de la distribución de los datos (RI). Incluye una línea vertical (Me) que parte la caja con el 25 % de las observaciones. El 25 % inferior y superior restantes de la distribución se representan a cada lado de la caja mediante los bigotes. Si hay observaciones *anómalas* o *atípicas*, es decir, aquellas que distan de la caja más de $(3/2)RI$, se suelen representar, como puntos aislados.

Histograma (variable cuantitativa continua)



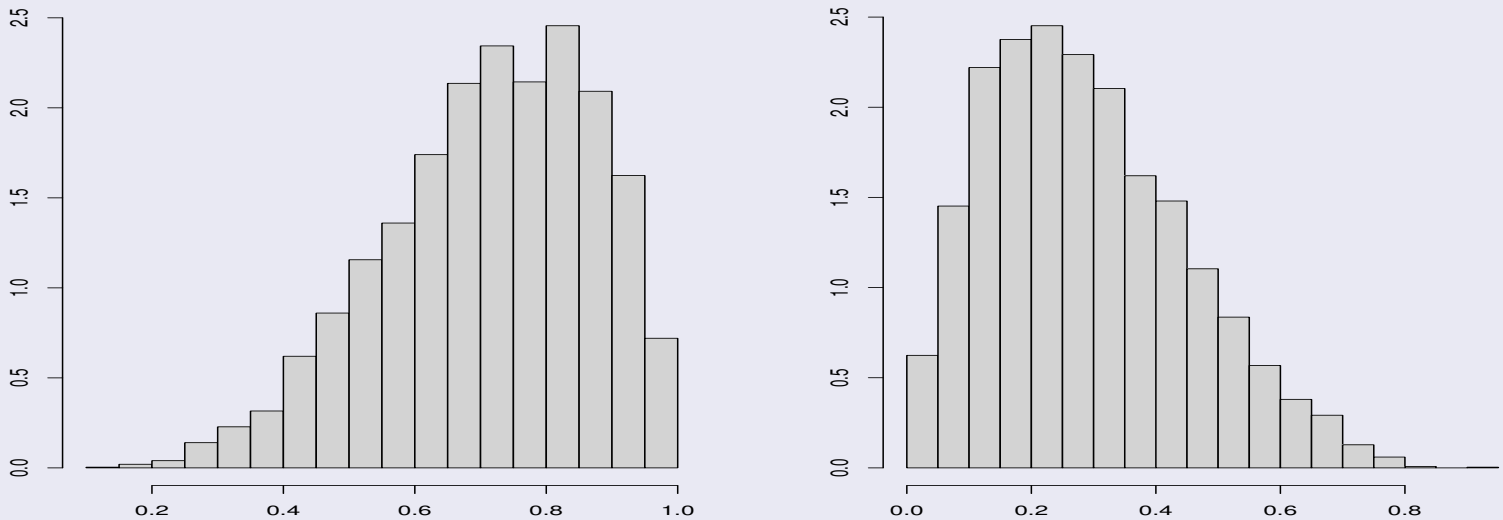
Representación gráfica sobre ejes de coordenadas a partir de rectángulos con alturas y bases que se corresponden respectivamente a frecuencias y amplitudes de cada intervalo o clase. Así, cada rectángulo tiene área proporcional a la frecuencia de la clase.

Distribución simétrica



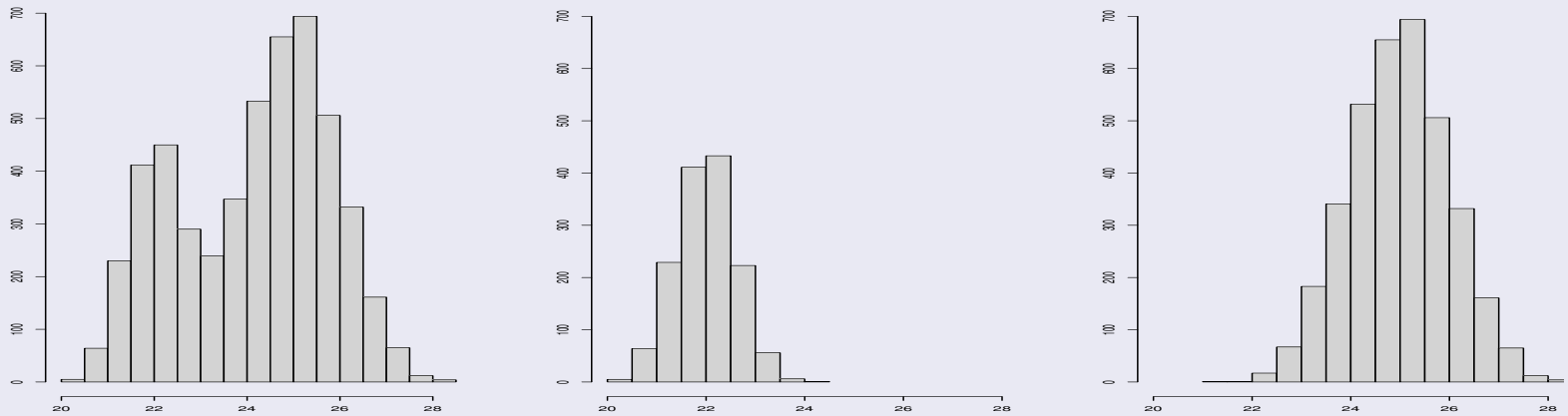
Estos dos histogramas representan la misma distribución de datos.
Tienen la misma forma, *simétrica*, aunque no indican lo mismo.
¿Dónde se situaría la media con respecto a la mediana?
¿Cómo sería el diagrama de caja?

Distribución asimétrica



Estos dos histogramas representan distribuciones asimétricas. El histograma de la izquierda muestra *asimetría a la izquierda*, mientras que el de la derecha presenta *asimetría a la derecha*.
¿Dónde se situaría la media con respecto a la mediana?
¿Cómo sería el diagrama de caja?

Distribución bimodal



El histograma de la izquierda representa el tiempo que tardan los autobuses urbanos en pasar por dos puntos del centro de una ciudad. Se aprecia *distribución bimodal*, es decir, dos modas (dos crestas). Esto puede ser debido a que se han mezclado dos poblaciones, como en este caso, donde no se ha diferenciado fin de semana de los días entre semana. Al separar respectivamente las dos poblaciones, se llega a los histogramas del centro y de la derecha.

- Hasta ahora hemos considerado un solo carácter de una población bajo estudio (unidimensional). Pero es habitual que interese más de un carácter. Si, para cada individuo, se estudian dos características, hablamos de *variables estadísticas bidimensionales* y, en general, *multidimensionales*.
- En esta sección estudiamos la relación entre dos variables estadísticas.
- En una variable estadística bidimensional (X, Y) observamos conjuntamente las dos características X e Y en los elementos de una población.
- Para cada elemento se obtiene un par de valores $(x_i,\ y_i)$, donde x_i es el valor para la variable X, e y_i es el valor para la variable Y, con $i = 1, \ldots, n$.

5.1. Organización de datos bidimensionales

- *Tablas de doble entrada*
  - Proporcionan información estadística conjunta para dos variables de cualquier tipo.
  - Las categorías, modalidades o clases, de cada variable se disponen respectivamente en las filas y columnas de la tabla.
  - Cada celda representa la frecuencia absoluta conjunta, $n_{i,j}$, o la frecuencia relativa conjunta, $f_{i,j}$, para $i = 1, \ldots, k$, $j = 1, \ldots, m$, modalidades respectivas de las variables X e Y.
  - Suele añadirse una fila y columna con los totales de cada fila y columna respectiva.
  - También se conocen como *tablas de contingencia*.

5.1. Organización de datos bidimensionales

Ejemplo 9

En los municipios de menos de 500 habitantes de Andalucía, se ha observado el número de Centros de Primaria y Consultorios en un determinado año, obteniendo los siguientes resultados:

C. Prim. \ Consult.	0	1	2	Total
0	1	74	8	83
1	1	23	2	26
Total	2	97	10	109

	Consultorios			
C. Prim.	0	1	2	Total
0	0.0092	0.6789	0.0734	0.7615
1	0.0092	0.2110	0.0183	0.2385
Total	0.0184	0.8899	0.0917	1

5.1. Organización de datos bidimensionales

- *Distribución marginal*
  Estudio del comportamiento de una variable con independencia de la otra. En una variable bidimensional, cada una de sus componentes, por separado, constituye una variable estadística unidimensional.
- *Distribución condicionada*
  La distribución condicionada de una variable dado un valor de la otra variable, por ejemplo, X dado $Y = y_j$, con $j = 1, \ldots, m$ modalidades, consiste en la distribución de la primera variable en el subconjunto donde la segunda variable toma el valor al que se condiciona.

5.2. Dependencia estadística

- Relación entre variables
- Variables independientes
  - Las variables X e Y se dice que son estadísticamente *independientes* si los valores de una de ellas no afecta a la distribución de la otra, es decir, si las distribuciones condicionadas son iguales para cualquier valor que se condicione.
  - También se observa independencia si la frecuencia relativa conjunta coincide con el producto de frecuencias relativas marginales.
- Estudiamos la *relación* entre las dos variables, es decir, el grado de dependencia entre ambas.
  Esta dependencia puede ser funcional, aleatoria, positiva, negativa, o sin dependencia.

5.2. Dependencia estadística

Ejemplo 10

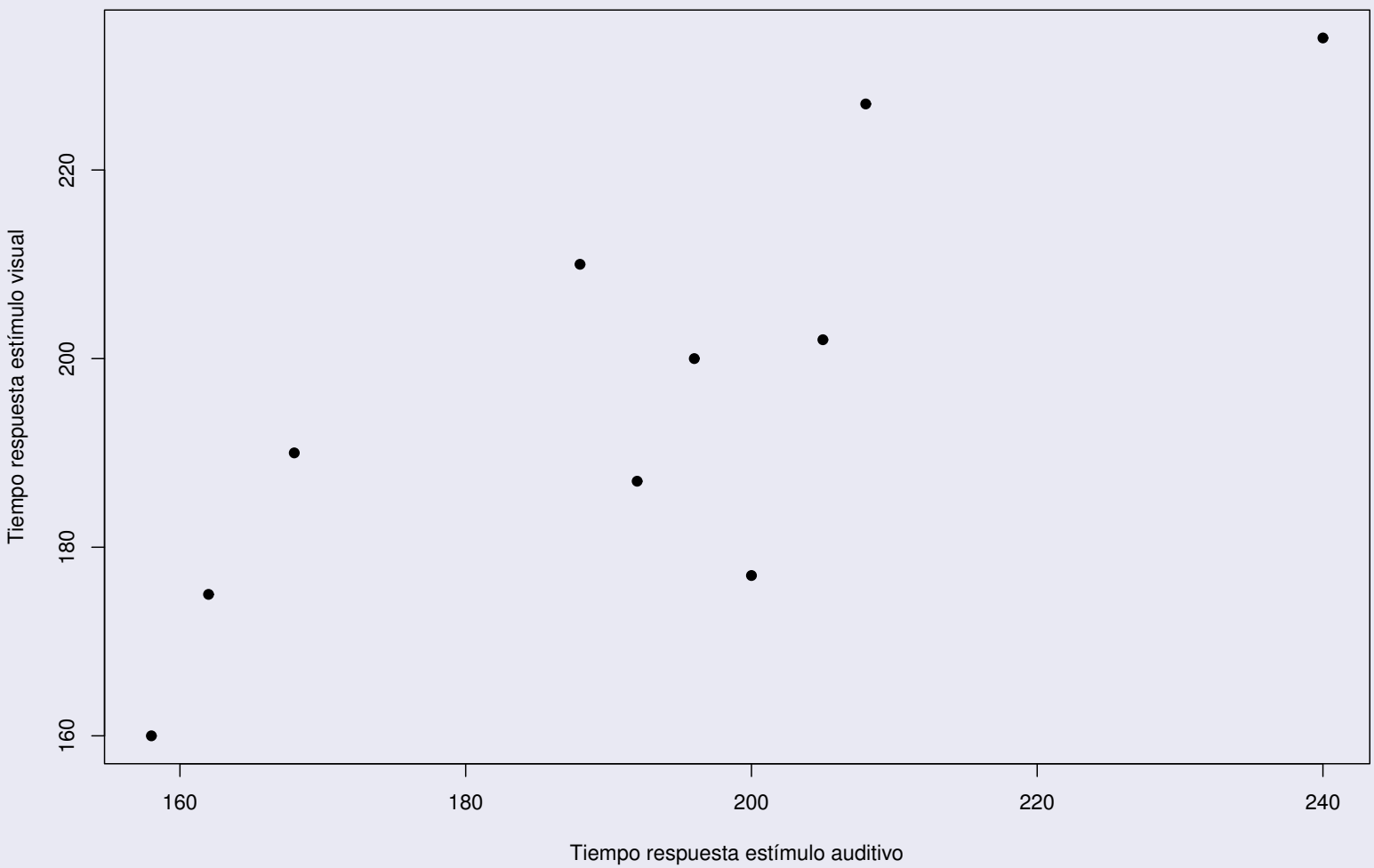
Se ha observado el tiempo, en milésimas de segundo, de reacción a estímulos visuales (V) y auditivos (A) en 10 pacientes, obteniendo los siguiente resultados:

V	160	227	210	190	177	187	202	234	175	200
A	158	208	188	168	200	192	205	240	162	196

- *Diagrama de dispersión*
  Dada una variable bidimensional (X, Y), el diagrama de dispersión o *nube de puntos* es la representación cartesiana de los puntos $(x_i, y_i),\ i = 1, \ldots, n$.
  El vector de medias, $(\bar{x},\ \bar{y})$, representa el centro de gravedad o centro de masas de la distribución conjunta.

5.2. Dependencia estadística

Diagrama de dispersión



6.1. Correlación y medidas bidimensionales

Covarianza

- Nos indica el tipo de relación entre variables y si dicha relación puede o no ser lineal. Generalmente se utiliza la covarianza muestral, $s_{X,Y}$,

$$s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}.$$

- Si trabajamos con tablas de doble entrada, con k y m modalidades respectivamente para X e Y,

$$s_{X,Y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{i,j} - n\bar{x}\bar{y}}{n-1}.$$

6.1. Correlación y medidas bidimensionales

Coeficiente de correlación lineal

- Se obtiene a partir de la covarianza y las desviaciones típicas marginales de cada variable:

$$r = \frac{s_{X,Y}}{s_X s_Y}, \qquad -1 \leq r \leq 1.$$

- Conocido también como *coeficiente de correlación de Pearson*, nos indica el tipo de relación entre variables, *directa* (si su signo es positivo) o *inversa* (signo negativo), y si la relación lineal entre las variables es *débil*, con r cercano a cero, o *fuerte*, con r cercano a los extremos, 1 y -1.

6.1. Correlación y medidas bidimensionales

Ejemplo (Continuación del Ejemplo 10)

Obtención de las medidas bidimensionales para el Ejemplo 10:

$$\begin{aligned}
\bar{x} &= 191.7; \\
\bar{y} &= 196.2; \\
s_X &= 24.6218; \\
s_Y &= 23.2561; \\
s_{X,Y} &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} \\
&= \frac{380320 - 10(191.7)(196.2)}{9} = 467.1778; \\
r_{X,Y} &= \frac{s_{X,Y}}{s_X s_Y} = \frac{467.1778}{(24.6218)(23.2561)} = 0.8159.
\end{aligned}$$

6.2. Recta de regresión

- La *regresión* es una técnica estadística que estudia la relación entre dos o más variables estadísticas. En la regresión lineal simple, una variable *respuesta* o *dependiente* Y, es explicada a partir de otra *regresora* o *independiente*, X.
- El coeficiente de correlación y el diagrama de dispersión nos informan de la relación lineal existente entre las variables bajo estudio.
- Cuando se tiene relación lineal entre las variables, se puede predecir una en función de la otra mediante una recta, $Y = a + bX + \mathcal{E}$.
- Para cada valor de la variable explicativa, X, la predicción de la recta produce un error, $e_i = y_i - (a + bx_i)$, con $i = 1, \ldots, n$.

6.2. Recta de regresión

- Para la estimación de los coeficientes de la recta, a y b, utilizamos el *método de mínimos cuadrados*, que minimiza la suma de cuadrados de los errores,

$$f(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

Derivamos respecto de cada parámetro y resolvemos el sistema

$$\frac{\partial f(a,b)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + bx_i)] = 0$$

$$\frac{\partial f(a,b)}{\partial b} = -2 \sum_{i=1}^{n} x_i [y_i - (a + bx_i)] = 0,$$

$$b = \frac{s_{X,Y}}{s_X^2}, \qquad a = \bar{y} - b\bar{x}.$$

A la pendiente de la recta, b, se conoce como *coeficiente de la regresión.*

6.2. Recta de regresión

- Así, la mejor predicción de Y mediante una función lineal de X es la recta

$$y = \bar{y} + \frac{s_{X,Y}}{s_X^2}(x - \bar{x}).$$

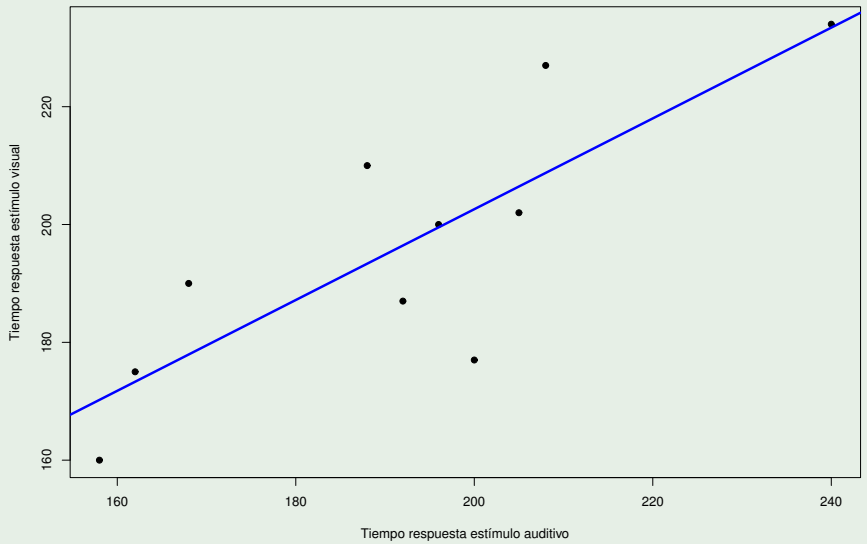
- Mediante la expresión obtenida, podemos predecir valores de la variable Y a partir de valores de la variable X.

6.2. Recta de regresión

Ejemplo (Continuación del Ejemplo 10)

Obtención de la recta de regresión para el Ejemplo 10:
Sustituyendo, con los valores obtenidos anteriormente, llegamos a la expresión de la recta,

$$y = 196.2 + \frac{467.1778}{24.6218^2}(x - 191.7) = 48.4714 + 0.7706x.$$



6.2. Recta de regresión

Coeficiente de determinación

- El *Coeficiente de determinación*, R^2, es un indicador de la calidad del modelo propuesto. Consiste en la proporción de variabilidad de la respuesta que explica el modelo.
- En la regresión lineal simple, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación lineal,

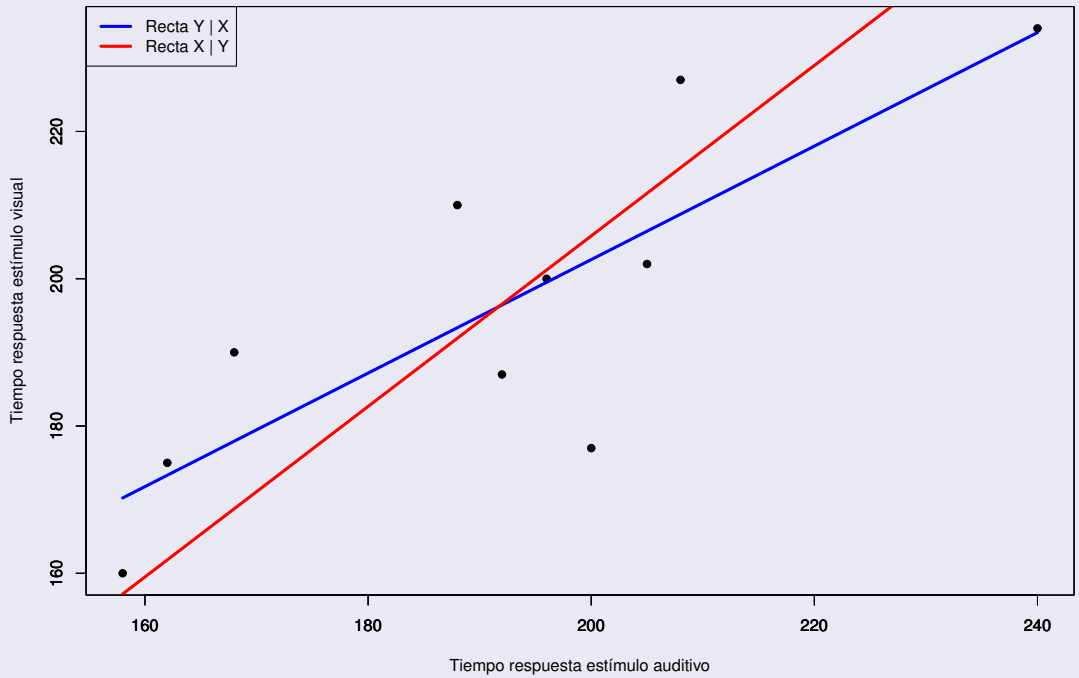
$$R^2 = r^2, \qquad 0 \leq R^2 \leq 1.$$

6.2. Recta de regresión

Recta de regresión X sobre Y

La recta obtenida anteriormente se conoce como recta de regresión Y sobre X, $Y \mid X$. También es posible, siguiendo el procedimiento visto, predecir X a partir de Y, $X \mid Y$,

$$x = \bar{x} + \frac{s_{X,Y}}{s_Y^2}(y - \bar{y}).$$



6.3. Transformaciones a la linealidad

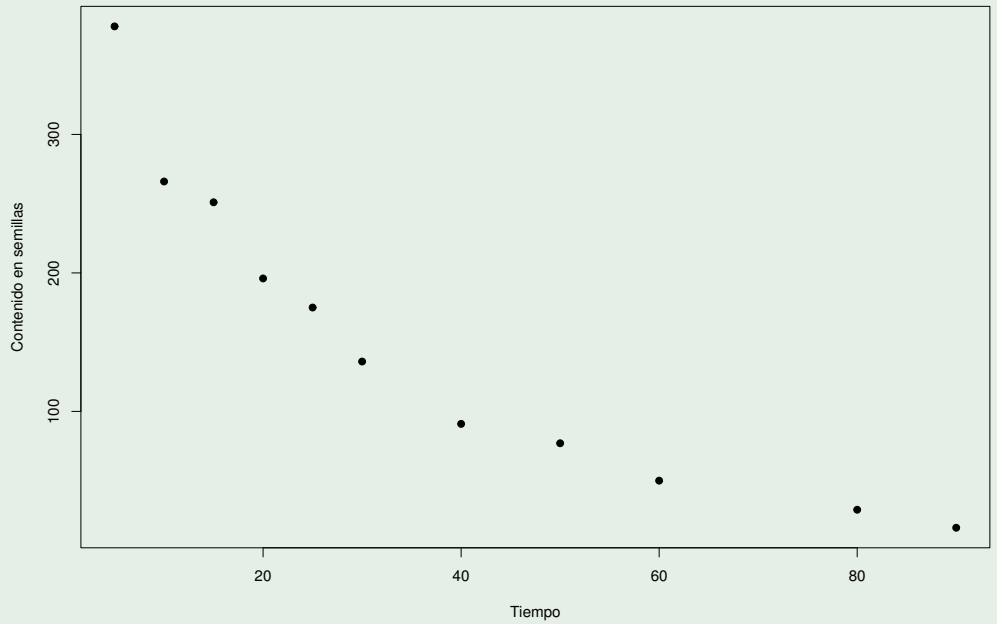
- El modelo de regresión lineal simple, $Y = a + bX + \mathcal{E}$, puede no ser apropiado debido a que la verdadera función de regresión sea no lineal.
- En ocasiones, una función no lineal puede expresarse como línea recta mediante algún tipo de transformación sobre la variable Y y/o la variable X.
- Estas transformaciones, aunque no sean lineales, llevan de manera directa al modelo lineal simple, cuyos parámetros se estiman con el método de mínimos cuadrados, visto anteriormente.

6.3. Transformaciones a la linealidad

Ejemplo 11

Un semillero estudia el comportamiento del contenido de un compuesto químico en semillas (Y), en nl/g, ante el paso del tiempo (X), en minutos, obteniendo los siguientes resultados:

Y	378	266	251	196	175	136	91	77	50	29	16
X	5	10	15	20	25	30	40	50	60	75	90



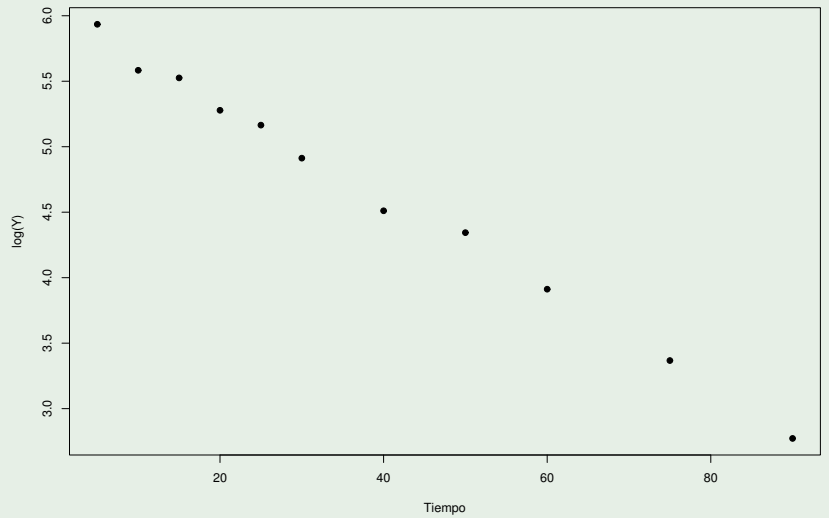
6.3. Transformaciones a la linealidad

Ejemplo (Continuación del Ejemplo 11)

- El gráfico anterior sugiere que una curva exponencial, $Y = ae^{bX}\mathcal{E}$, se ajustaría mejor a los datos.
- Tomando logaritmos en la anterior expresión, llegamos a un modelo lineal,

$$Y' = a' + bX + \mathcal{E}',$$

donde $Y' = \log(Y)$, $a' = \log(a)$, $\mathcal{E}' = \log(\mathcal{E})$.



6.3. Transformaciones a la linealidad

Ejemplo (Continuación del Ejemplo 11)

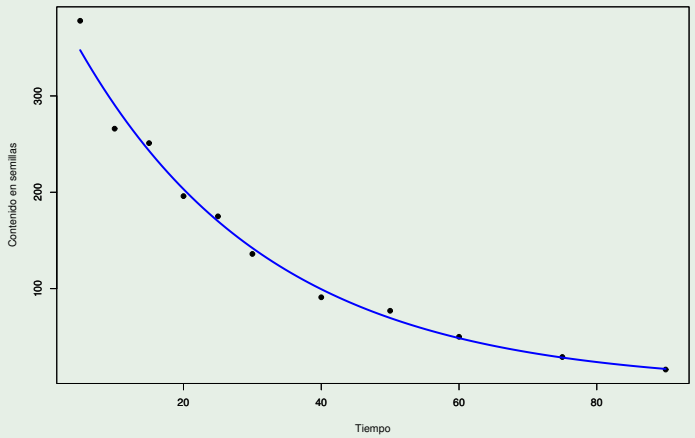
- Las estimaciones de los parámetros del modelo lineal son

$$a' = 6.0297, \qquad b = -0.0358.$$

- Deshacemos la transformación para llegar al modelo exponencial estimado para la regresión,

$$y = 415.5836\ e^{-0.0358x},$$

utilizado para la predicción de Y a partir de X.



6.4. Regresión polinomial

- Una función polinómica puede ser una buena aproximación para la regresión, en aquellos casos donde la relación lineal no parece adecuada.
- Consideremos un polinomio de segundo grado (aunque podría ser de mayor grado) como modelo de regresión,

$$Y = a + bX + cX^2 + \mathcal{E}.$$

- Análogamente a la regresión lineal, utilizamos el método de mínimos cuadrados para estimar los coeficientes del modelo, minimizando la siguiente expresión,

$$f(a, b, c) = \sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2)]^2.$$

6.4. Regresión polinomial

- Para ello, se obtiene el sistema

$$\frac{\partial f(a,b,c)}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i + cx_i^2)] = 0$$

$$\frac{\partial f(a,b,c)}{\partial b} = -2\sum_{i=1}^{n} x_i[y_i - (a + bx_i + cx_i^2)] = 0$$

$$\frac{\partial f(a,b,c)}{\partial c} = -2\sum_{i=1}^{n} x_i^2[y_i - (a + bx_i + cx_i^2)] = 0.$$

6.4. Regresión polinomial

- Operando, llegamos al siguiente sistema equivalente de ecuaciones, que se conocen como *ecuaciones normales de la regresión*,

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$
$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$
$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4.$$

- Resolviendo dicho sistema, llegamos a las estimaciones de los parámetros del modelo de predicción de Y a partir de X, mediante un polinomio de segundo grado.

Bibliografía

- Devore, J. L.: *Probabilidad y Estadística para Ingeniería y Ciencias.* Thomson, 2001
- Mendenhall, W.; Sincich, T.: *Probabilidad y Estadística para Ingeniería y Ciencias.* Prentice Hall, 1997
- Montgomery, D.; Runger, G. C.: *Probabilidad y Estaditica Aplicadas a la Ingeniería.* Mc Graw Hill, 2001
- Navidi, W.: *Estadística para ingenieros y científicos.* McGraw-Hill, 2006
- Ross, S. M.: *Introduction to Probability and Statistics for Engineers and Scientists.* Elsevier, 2009

Estadística Descriptiva

Estadística e Investigación Operativa, 1º EII

Antoni Torres Signes

Área de Estadística e Investigación Operativa
Escuela de Ingenierías Industriales

UNIVERSIDAD
DE MÁLAGA