

Tema 1. Estadística descriptiva

Estadística
Grado en Ingeniería Informática

Índice

1. Conceptos generales	1
2. Distribuciones de frecuencias	2
3. Representaciones gráficas	4
4. Medidas características: posición, dispersión, forma	8
4.1. Medidas de posición	8
4.2. Representación de medidas: el diagrama de caja	15
4.3. Tipificación de datos	17
5. Tablas de contingencia	19
6. Recta de regresión	21
6.1. Vector de medias. Covarianza y correlación	21
6.2. Método de Mínimos Cuadrados	23
6.3. Coeficiente de regresión. Coeficiente de determinación	25

La **Estadística descriptiva** abarca una serie de técnicas numéricas y gráficas para describir y analizar un conjunto de datos, a veces procedentes de una población mayor, pero sin extraer conclusiones (inferencias) sobre dicha población. En este tema se introducirán algunas técnicas descriptivas básicas, como la construcción de tablas de frecuencias, la elaboración de gráficas y las principales medidas descriptivas de centralización, dispersión y forma. La segunda parte del tema se dedicará a la introducción de algunas medidas descriptivas bidimensionales para variables categóricas (tablas de contingencia) y para variables continuas (recta de regresión)¹.

1. Conceptos generales


Frecuentemente, en un análisis estadístico el objetivo último es extraer conclusiones sobre un colectivo de interés denominado población. En ocasiones, el tamaño de la población (formada por individuos) puede hacer inabordable el estudio individualizado de las características de cada uno de ellos. Si se quisiera realizar un estudio sobre las horas diarias que pasan los jóvenes entre 15 y 20 años conectados a redes sociales, sería muy complicado y costoso realizar mediciones para cada uno de ellos. Para solucionar este problema, dichas mediciones se realizarán sobre una muestra.

Consideraremos los siguientes elementos:

- **Población:** colectivo de individuos sobre los que se quiere extraer alguna conclusión.
- **Individuo:** cada uno de los elementos de la población (unidad estadística).
- **Muestra:** subconjunto (representativo) de la población, que se selecciona con el objetivo de extraer información.

Las técnicas de **Estadística descriptiva** permiten describir y analizar un conjunto de datos, sin extraer conclusiones sobre la población a la que pertenecen. Se tendrá que recurrir a la **Inferencia Estadística**, que es la parte de la Estadística que trata las condiciones bajo las cuales la información derivada a partir de una muestra es válida para extraer conclusiones sobre la población de interés. Para aplicar cualquier técnica estadística será necesario analizar previamente el tipo de variable a la que se hace referencia.

- **Variable estadística:** cada una de las características consideradas con el propósito de describir a cada individuo de la muestra. Denotaremos las variables por X , Y , Z ,...
- **Tipos de variables:** distinguiremos dos tipos de variables. Las variables cualitativas o categóricas (aquellas que no se pueden expresar a través de una cantidad numérica) y las variables

¹Los contenidos de los temas se irán completando con las prácticas de ordenador. Además, en los apuntes se incluirán algunos ejemplos breves de código  (el que utilizaremos en las prácticas).

cuantitativas (se puede expresar a través de un número). A su vez, las primeras pueden clasificarse en nominales (no existe orden subyacente de las clases o categorías) u ordinales (sí existe orden subyacente de las clases o categorías); mientras que las cuantitativas pueden clasificarse en discretas y continuas, según el tipo de valores que tomen. En la Tabla 1 se incluyen algunos ejemplos.

Tipo	Clases	Ejemplo
Cualitativa	Nominal Ordinal	Sexo, raza, color de ojos,... Grado de contaminación, calificación,...
Cuantitativa	Discreta Continua	Nº de hermanos, nº de materias, ... Peso, altura, ...

Tabla 1. Tipos de variables estadísticas.

Ejemplo. En la web de Gapminder (<http://www.gapminder.org>) están disponibles una gran cantidad de datos de distintos países, relativos a factores sociales, económicos, medioambientales, etc. En concreto, una de las bases disponibles procede de un estudio realizado por la Organización de Naciones Unidas en el que se recoge información sobre el ratio de ordenadores personales por cada 100 habitantes. Se tiene datos para varios años, pero para la ilustración de las técnicas de este tema, nos quedaremos con los correspondientes al año 2006. Los países están clasificados en regiones geográficas: América, Este de Asia y Pacífico, Europa y Asia Central, Medio Este y Norte de África, Sur de Asia, África Subsahariana. Por tanto tendremos dos variables: X (ordenadores personales por cada 100 habitantes, cuantitativa continua) e Y (región geográfica, cualitativa nominal).

2. Distribuciones de frecuencias

Las tablas de frecuencias (véase Tabla 2) son una de las técnicas básicas para el resumen de información a partir de una muestra de datos. Su construcción es sencilla, pero en conjuntos de datos de un tamaño moderado o grande su cálculo puede resultar laborioso. También se pueden obtener utilizando cualquier paquete estadístico.

- **Tablas de frecuencias:** las tablas de frecuencias se utilizan para representar la información contenida en una muestra de tamaño n extraída de una población, (x_1, \dots, x_n) .
- **Clase:** cada uno de los valores que puede tomar una variable (cualitativa o cuantitativa discreta). Se denotan como: $c_i, i = 1, \dots, k$. El número de individuos de la muestra en cada clase (o modalidad) c_i se denota por n_i .

- **Frecuencia absoluta:** para cada clase c_i , la frecuencia absoluta es $n_i, i = 1, \dots, k$.
- **Frecuencia relativa:** para cada clase c_i , la frecuencia relativa es $f_i = n_i/n, i = 1, \dots, k$.
- **Frecuencia absoluta acumulada:** la frecuencia absoluta acumulada de una clase c_i es $N_i = \sum_{j=1}^i n_j = n_1 + \dots + n_i, i = 1, \dots, k$.
- **Frecuencia relativa acumulada:** la frecuencia relativa acumulada de una clase c_i es $F_i = \sum_{j=1}^i f_j = f_1 + \dots + f_i = \frac{N_i}{n}, i = 1, \dots, k$.

A partir de sus definiciones, se pueden demostrar algunas propiedades de las frecuencias absolutas y relativas que se calculan en las tablas de frecuencias. Así, se tiene que:

- Las frecuencias absolutas: $0 \leq n_i \leq n, i = 1, \dots, k$.
- Las frecuencias relativas: $0 \leq f_i \leq 1, i = 1, \dots, k$.
- Las frecuencias absolutas acumuladas: $N_k = \sum_{j=1}^k n_j = n_1 + \dots + n_k = n$.
- Las frecuencias relativas acumuladas: $F_k = \sum_{j=1}^k f_j = f_1 + \dots + f_k = 1$

Obsérvese que tiene sentido interpretar las frecuencias acumuladas si las categorías bajo estudio tienen un cierto orden.

Modalidad o clase	Frecuencia absoluta	Frecuencia relativa	Fr. abs. acumulada	Fr. rel. acumulada
c_1	n_1	f_1	N_1	F_1
c_2	n_2	f_2	N_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots
c_i	n_i	f_i	N_i	F_i
\vdots	\vdots	\vdots	\vdots	\vdots
c_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Tabla 2. Ejemplo de tabla de frecuencias.

En el caso de variables cualitativas o cuantitativas discretas con pocos valores, es posible determinar las clases de la variable. Sin embargo, en el caso de variables cuantitativas continuas (o cuantitativas discretas con muchos valores), se tendrán que construir clases artificiales de manera que se agrupen valores en intervalos. Estas nuevas clases se denominan intervalos de clase.

■ **Intervalos de clase:** para variables cuantitativas continuas, se agrupan los distintos valores obtenidos en la muestra en intervalos. Cada intervalo representará una *clase*. A partir de una muestra, los intervalos de clase se construyen de la siguiente forma:

- Denotamos por $e_1 < \dots < e_k$ los extremos de los k intervalos de clase. Cada intervalo será de la forma $[e_i, e_{i+1})$, a excepción de en ocasiones el último que podría ser cerrado en ambos lados, $[e_{k-1}, e_k]$ cuando el máximo de la muestra coincide con el extremo superior.
- Amplitud del intervalo: $a_i = e_{i+1} - e_i$.
- Marca de clase: $c_i = \frac{e_i + e_{i+1}}{2}$.
- Para seleccionar el número de intervalos, consideramos el entero más próximo a \sqrt{n} , donde n es el tamaño de la muestra observada. El número de intervalos suele estar entre 5 y 20. Para determinar la amplitud de los intervalos (en principio, todos de la misma amplitud), tenemos que ver antes cuál es el rango de variación de los datos (diferencia entre el máximo y el mínimo), y construir los intervalos de manera que cubran todo el rango. Habitualmente se escogen intervalos de igual longitud, pero no es una condición obligatoria.

Ejemplo. Sobre el conjunto de datos de ordenadores personales por cada 100 habitantes para los distintos países, podemos construir la tabla de frecuencias para la variable X (región geográfica). Así, en la Tabla 3, podemos ver que hay datos de 196 países. De ellos, el 26 % son países de Europa y Asia Central y el 4.1 % pertenecen al Sur de Asia. Se puede observar que las frecuencias relativas suman 1 (fila: Total). En este caso, no tiene sentido considerar frecuencias acumuladas, pues las categorías (en este caso las regiones) no tienen ningún orden subyacente.

Región	Fr. abs.	Fr. rel
América	40	0.204
Este de Asia y Pacífico	32	0.163
Europa y Asia Central	51	0.260
Medio Este y Norte de África	20	0.102
Sur de Asia	8	0.041
África Subsahariana	45	0.230
Total	196	1

Tabla 3. Tabla de frecuencias para la región geográfica de los países. Datos sobre número de ordenadores personales por cada 100 habitantes.

3. Representaciones gráficas

La clasificación de variables que se ha expuesto en la sección anterior, distinguiendo entre variables cualitativas y cuantitativas (discretas y continuas) es de crucial importancia a la hora de construir representaciones gráficas. De modo esquemático, se introducen las principales técnicas de representación

para variables cualitativas, variables cuantitativas discretas y cuantitativas continuas. En el caso de variables cuantitativas discretas, si tienen pocos valores, se puede hacer uso de las representaciones descritas para variables cualitativas (diagramas de barras y sectores). Si por el contrario toman muchos valores, entonces se pueden utilizar las representaciones para variables cuantitativas continuas.

Variables cualitativas. Para la representación de variables cualitativas se suelen utilizar el diagrama de barras o el diagrama de sectores. Para construir un diagrama de barras, en uno de los ejes se representan las categorías o clases de la variable que se quiere representar y se dibujan barras de altura/longitud proporcional a la frecuencia de cada clase (absoluta o relativa). En el diagrama de sectores también se representan las distintas clases y su frecuencia, de manera que el círculo se reparte de forma proporcional a la frecuencia de cada clase. Un ejemplo de diagrama de barras (horizontal) para las regiones geográficas de los países se puede ver en la Figura 1, y uno en vertical en la Figura 2 (izquierda).

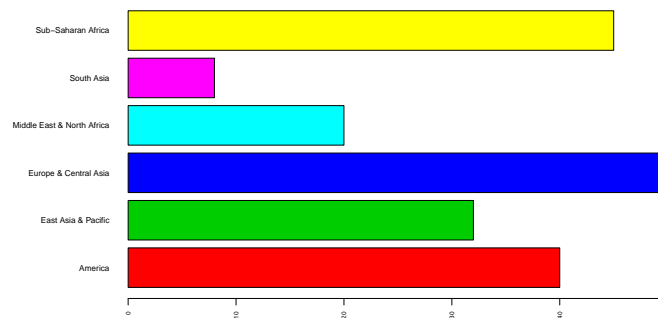


Figura 1. Diagrama de barras para las regiones geográficas.

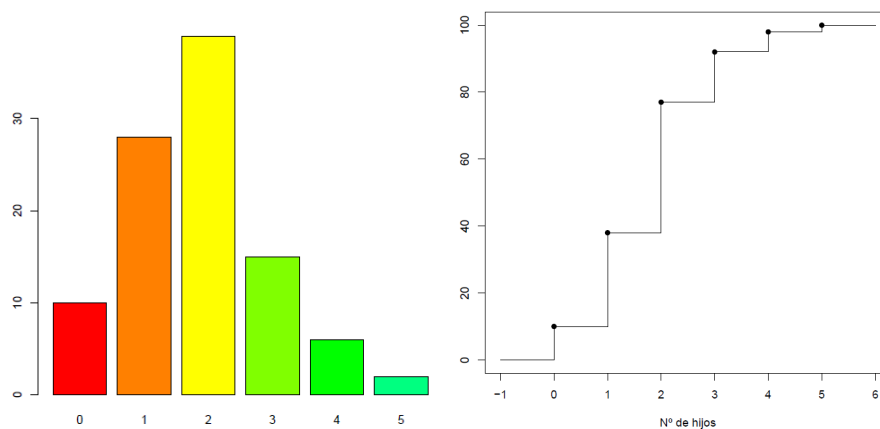


Figura 2. Diagrama de barras y diagrama acumulativo de frecuencias para el número de hijos de una familia.

Variables cuantitativas discretas. Además del diagrama de barras descrito para las variables cualitativas, que también se puede utilizar para variables cuantitativas discretas, para la representación de este tipo de variables se tiene el diagrama acumulativo de frecuencias. El diagrama acumulativo de frecuencias se construye representando, para cada clase de la variable c_i , los puntos (c_i, N_i) (o bien (c_i, F_i)) y uniéndolos con segmentos horizontales y verticales, de forma que se obtiene una función escalonada. Si se utilizan las frecuencias relativas acumuladas, el valor máximo del diagrama acumulativo se alcanza en el 1, mientras que si se construye con las frecuencias absolutas acumuladas, el máximo será el número de datos de la muestra. Consideremos, por ejemplo, una variable que cuente el número de hijos de una familia. Esta tomará valores discretos, y en una cantidad finita. En la Figura 2 se muestran el diagrama de barras y el diagrama acumulativo de frecuencias absolutas para este ejemplo.

Variables cuantitativas continuas. En el caso de variables cuantitativas continuas, podemos construir el polígono (acumulativo) de frecuencias, de igual modo que el diagrama acumulativo de frecuencias explicado para variables cuantitativas discretas, pero considerando las marcas de clase de cada intervalo e_i en la representación. Sin embargo, es más usual el histograma.

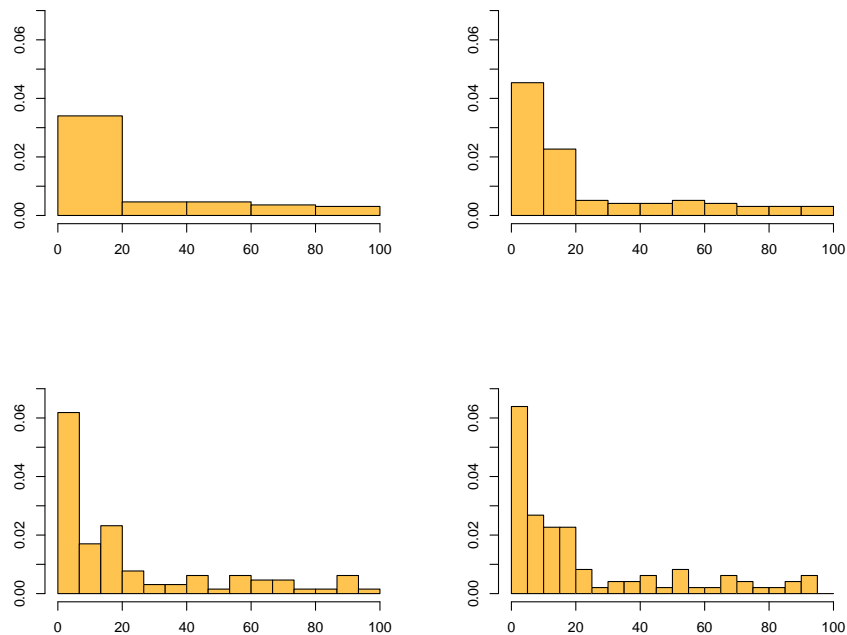


Figura 3. Histograma del porcentaje de ordenadores por cada 100 habitantes para diferentes intervalos.

El histograma equivale en cierto modo al diagrama de barras, pero en el caso continuo, de forma que las barras aparecen contiguas. En el eje horizontal se representan los intervalos de clase de la

variable, y sobre ellos se levantan barras de altura $h_i = n_i/a_i$ (o bien $h_i = f_i/a_i$), donde n_i es la frecuencia absoluta de cada intervalo (f_i es la frecuencia relativa) y a_i es la amplitud del mismo. Si el histograma se construye con frecuencias relativas, la suma de las áreas de las barras es igual a 1. El histograma da una idea clara de la *distribución* de los datos, pero es muy sensible a la elección de los intervalos de clase (véase Figura 3). Además, a partir del histograma, se pueden hacer algunos cálculos sobre porcentajes de datos. Por ejemplo, ¿qué porcentaje de países tienen entre 60 y 70 ordenadores personales por cada 100 habitantes? ¿Y entre 70 y 75? Para responder a esta pregunta habría que calcular el área del histograma considerando como base este intervalo.

Como se puede observar, la distribución del porcentaje de ordenadores por cada 100 habitantes es muy asimétrica por lo que para la correcta aplicación de otras técnicas estadísticas será de interés realizar transformaciones. En este caso, cabe señalar que la transformación logarítmica corrige la asimetría positiva (asimetría hacia la derecha). El histograma de la variable transformada puede verse en la Figura 4.

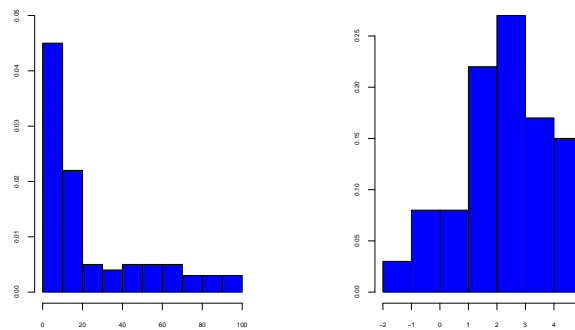



Figura 4. Histogramas del porcentaje de ordenadores por cada 100 habitantes y de su transformación logarítmica.

Códigos para el ejemplo. Las gráficas anteriores se han obtenido en  con los siguientes comandos. Se utiliza x para denotar la variable de interés:

```
tabla<-table(x) #frecuencias absolutas
prop.table(tabla) #frecuencias relativas
barplot(tabla,cex.names=0.75,col=2:7,las=2,horiz=TRUE) #diagrama de barras
pie(tabla,col=2:7) #diagrama de sectores
par(mfrow=c(1,2)) #dividimos la pantalla gráfica en 2 columnas
hist(whole_data,freq=F,col=4) #histogramas
hist(log(whole_data),freq=F,col=4)
```


4. Medidas características: posición, dispersión y forma

Denotando por X la variable estadística de interés (cuantitativa, en general, excepto en el caso de la moda) y por x_i la observación en el individuo i , se introducirán en esta sección algunas de las principales medidas características para describir la información contenida en una muestra x_1, \dots, x_n de tamaño n . Dichas medidas se utilizan para resumir la información atendiendo a tres aspectos principales:

alrededor de qué valores se encuentran los datos,

cuánto se dispersan,

si se distribuyen de manera similar a una *campana de Gauss*, que será el modelo que se tome como referencia y que se estudiará en temas posteriores.

Por ello, se distinguirán tres tipos de medidas: de posición, de dispersión y de forma.

4.1. Medidas de posición

Las medidas de posición o localización nos indican el valor o valores alrededor de los cuales se sitúan los datos observados. Distinguiremos medidas de localización de tendencia central (media, mediana y moda) y de tendencia no central (cuartiles, deciles y percentiles).

4.1.1. Medidas de posición de tendencia central

Como medidas de posición de tendencia central se introducirán la media aritmética o media muestral, la mediana y la moda. Estas medidas nos proporcionan valores alrededor de los cuales se distribuyen los datos observados en la muestra.

Media aritmética. Se define como:

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

La media aritmética (media muestral) presenta las siguientes propiedades, que son fáciles de deducir a partir de la definición:

- Toma valores entre el mínimo y el máximo:

$$\min\{x_1, \dots, x_n\} \leq \bar{x} \leq \max\{x_1, \dots, x_n\}.$$

- La media aritmética es lineal. Es decir, si consideramos los datos $y_i = ax_i + b$, la media de los nuevos datos se obtendrá como $\bar{y} = a\bar{x} + b$.
- La media de las desviaciones con respecto a la media es cero:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

- La media de los cuadrados de las desviaciones con respecto a una constante es mínima para la media:

$$\bar{x} = \arg \min_a \frac{1}{n} \sum_{i=1}^n (x_i - a)^2.$$

El valor de la media no tiene porqué pertenecer al conjunto de posibles valores de la variable. Por ejemplo, puede resultar que el número medio de hermanos de una muestra no sea un número entero.

Uno de los problemas que presenta la media es que no es una medida robusta, es decir, su valor se ve influenciado por datos anormalmente altos o bajos. Los datos que difieren numéricamente de las demás observaciones se denominan valores atípicos. Algunas modificaciones para corregir la falta de robustez son la media truncada y media recortada. En la media truncada, un porcentaje de los datos atípicos se elimina del cálculo y para obtener una media recortada, estos valores atípicos se substituyen por el punto de corte, es decir, el dato inmediatamente inferior a los que se eliminan, para datos altos, y el inmediatamente superior para los datos bajos.

Otra modificación es la media ponderada en la cual se asigna distintos pesos a las observaciones. En la media aritmética cada observación tiene una contribución de peso $1/n$ al valor de \bar{x} . En la media ponderada, cada observación tendrá una ponderación ω_i , de tal modo que $\sum_{i=1}^n \omega_i = 1$.

En el caso de que se disponga de datos agrupados en una tabla de frecuencias, la media aritmética se calcula como:

$$\bar{x} = \sum_{i=1}^k c_i f_i = \frac{\sum_{i=1}^k c_i n_i}{n},$$

donde c_i es la marca de clase, f_i (n_i) las frecuencias relativas (absolutas) y k denota el número de intervalos de clase de los que se dispone. Las propiedades anteriormente descritas también se aplican a este caso.

Mediana. Si suponemos que los datos de la muestra están ordenados de menor a mayor, la mediana es el valor hasta el cual se encuentran el 50 % de los casos. Por tanto, la mediana dejará la mitad de las observaciones por debajo de su valor y la otra mitad por encima. Así, si la muestra consta de un número impar de datos (n impar), la mediana será el dato central. Si el tamaño de la muestra n es par, entonces se tomará como mediana la media de los dos datos centrales.

En el caso de tener la variable continua representada en una tabla de frecuencias, podemos definir el intervalo mediano, que será aquel cuya frecuencia relativa acumulada en el extremo inferior es menor que $1/2$ y en el extremo superior mayor que $1/2$.

La mediana, a diferencia de la media, es una medida robusta ya que su valor se ve poco afectado por la presencia de datos atípicos. Si de una muestra se obtienen la media y la mediana y sus valores difieren sustancialmente, esto será indicativo de la presencia de datos atípicos. Además, la media aritmética de las diferencias absolutas con respecto a una constante es mínima para la mediana. Es decir, si denotamos la mediana por Me tenemos:

$$Me = \arg \min_a \frac{1}{n} \sum_{i=1}^n |x_i - a|.$$

Moda. Para variables discretas o cualitativas, la moda es el valor o valores que más se repiten. Esto implica que la moda no tiene porqué ser única. Para variables cuantitativas continuas, el intervalo modal es aquel con mayor frecuencia. La moda se denotará por Mo . Si los datos se encuentran agrupados, se puede obtener el intervalo modal como aquel que tiene una mayor frecuencia.

4.1.2. Medidas de posición de tendencia no central

Como medidas de posición de tendencia no central, introduciremos los cuartiles, deciles y percentiles.

- **Cuartiles.** Los cuartiles Q_1 , Q_2 y Q_3 dividen la muestra en cuatro partes iguales, de manera que por debajo de Q_1 tenemos el 25 % de los datos, entre Q_1 y Q_2 se encuentra otro 25 % y por encima de Q_3 otro 25 %. La idea de dividir la muestra en partes iguales se puede generalizar a la construcción de los deciles (d_1, \dots, d_9 , dividen la muestra en 10 partes iguales) y los percentiles (p_1, \dots, p_{99} , dividen la muestra en 100 partes iguales).

En general, se define el **cuantil** de orden p ($0 < p < 1$) como el valor que deja por debajo una proporción al menos p de observaciones. El cuantil p se denotará por q_p .

4.1.3. Medidas de dispersión absoluta

Las medidas de posición o localización indican en torno a qué valores se sitúan los datos, pero para obtener una descripción más precisa de los mismos, es necesario conocer cuál es la dispersión que presentan. Las medidas de dispersión absoluta dependen de las unidades en las que se miden las observaciones, siendo las más conocidas la varianza muestral y la desviación típica muestral, que no es más que la raíz cuadrada de la varianza muestral.

- **Varianza (s^2) y desviación típica (s) muestrales.** Dada una muestra x_1, \dots, x_n , si consideramos la media muestral \bar{x} como medida de posición de tendencia central, se podría pensar en

medir la dispersión a través de las diferencias de los valores a la media muestral: $(x_i - \bar{x})$, para todo $i = 1, \dots, n$. Una forma de contabilizar todas estas diferencias sería a través de la suma: $\sum_{i=1}^n (x_i - \bar{x})$. Por las propiedades de la media muestral, vimos que la media de las diferencias con respecto a la media muestral es nula, así que esta expresión siempre resultará cero. En este caso, las diferencias positivas y negativas de los datos a la media muestral se compensan, por lo que para hacerlas positivas podríamos pensar en medir estas diferencias al cuadrado: $(x_i - \bar{x})^2$. De este modo se obtiene la varianza cuya fórmula es:

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

La varianza muestral está medida en las unidades de los datos al cuadrado, por lo que no se puede comparar directamente con las medidas de posición, por ejemplo, con la media. Para obtener una medida en las unidades de los datos, se considera la desviación típica muestral:

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

La varianza muestral tiene las siguientes propiedades, fáciles de deducir a partir de la definición:

- Toma valores no negativos, puesto que se trata de un promedio de valores no negativos (diferencias al cuadrado).
- La varianza muestral no es lineal. Si consideramos los datos $y_i = ax_i + b$, la varianza muestral de los nuevos datos será $s_y^2 = a^2 s_x^2$. Es decir, la varianza no se ve afectada por traslaciones (sumar o restar una constante), pero sí por los cambios de escala al multiplicar los valores de la muestra por un factor.
- Una expresión alternativa para el cálculo de la varianza muestral es:

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

Aunque la varianza muestral es la medida más intuitiva, en temas posteriores se introducirá una nueva medida de dispersión, denominada cuasi-varianza:

$$s_c^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{ns^2}{n-1}.$$

La diferencia entre varianza muestral y cuasi-varianza radica en el denominador. En la varianza, se hace un promedio, dividiendo por el número de datos mientras que en la cuasi-varianza, se divide por

el número de datos de los que obtenemos información, conociendo la media muestral.

Consideremos el siguiente ejemplo: supongamos que tenemos una muestra de tamaño $n = 4$, $\{2, x, 6, 8\}$ cuya media es $\bar{x} = 5.25$. Con esta información es fácil deducir que $x = 5$. En general, si se conoce el valor de la media y $(n - 1)$ valores de la muestra, podemos determinar el que falta. Esta corrección es importante en muestras pequeñas o de tamaño moderado. Al igual que para la varianza, también se puede definir la cuasi-desviación típica, s_c .

Otras medidas de dispersión absoluta (es decir, que también dependen de las unidades de los datos) son el rango muestral (R) y el rango intercuartílico (RIC) que se definen de la siguiente manera:

$$R = \max\{x_i\} - \min\{x_i\}$$

$$RIC = Q_3 - Q_1.$$


Para el cálculo del rango se utilizan sólo dos observaciones, la más grande y la más pequeña, por lo que se ve afectado por la presencia de datos atípicos.

4.1.4. Medidas de dispersión relativa

Las medidas de dispersión absoluta dependen de las unidades de los datos, por lo que no son adecuadas para comparar diferentes variables. Una de las medidas de dispersión relativa (no depende de las unidades de los datos) más usual es el coeficiente de variación:

$$CV = \frac{s}{\bar{x}}, \quad (\bar{x} > 0)$$

El coeficiente de variación permite comparar variables aunque estas estén registradas en distintas unidades de medida. También es de utilidad para comparar variables que, aunque de la misma magnitud, están en escalas distintas. Por ejemplo, para comparar las longitudes del diámetro del tímpano (normalmente, entre 8 y 10 milímetros) y de la columna vertebral (en centímetros), podríamos transformar todas las observaciones a la misma escala pero seguramente la dispersión (medida en desviación típica) que encontraríamos en las longitudes del diámetro del tímpano sería prácticamente nula.

Ejemplo. Para los datos de número de ordenadores personales por cada 100 habitantes podemos calcular en  las medidas características que hemos presentado. Denotamos por x la variable de interés.

```
> summary(x)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.30  3.49  12.35  23.13  37.25  94.58
```

El comando `summary` devuelve los valores mínimo, máximo, media y cuartiles. En este caso, el valor mínimo es de 0.3 ordenadores personales por cada 100 habitantes y el máximo de 94.58. En promedio, habría 23 ordenadores personales (aproximadamente) por cada 100 habitantes. Además, un 25 % de los países tendrían menos de 3.49 ordenadores por cada 100 habitantes, y un 25 % de los países tendrían más de 37.25 ordenadores por cada 100 habitantes (así se interpretan Q_1 y Q_3). El valor de la mediana es de $Q_2 = 12.35$, lo cual quiere decir que la mitad de los países de la muestra disponen de menos de 12.35 ordenadores por cada 100 habitantes. El valor de la mediana es inferior al de la media, como cabía esperar dada la asimetría positiva de los datos (recuerda la gráfica del histograma).

El resumen que obtenemos con `summary` nos da directamente medidas de localización (de tendencia central y no central) y nos permite calcular dos medidas de dispersión (rango y rango intercuartílico), pero no tenemos varianza y/o desviación típica. Podemos obtener medidas de dispersión como sigue (además de tener los rangos de otro modo):

```
> diff(range(x));IQR(x)
```

```
[1] 94.28
```

```
[1] 33.76
```

Con el comando `range` obtenemos los valores mínimo y máximo y a través de su diferencia (con la función `diff`) calculamos el rango. La función `IQR` (interquartile range) nos devuelve el rango intercuartílico. El cálculo de varianza y desviación típica puede hacerse utilizando dos comandos muy sencillos que nos dan la cuasi-varianza y la cuasi-desviación típica:

```
> var(x);sd(x)
```

```
[1] 702.9521
```

```
[1] 26.51324
```

La diferencia entre varianza y cuasi-varianza radica en el denominador que se utiliza. Así, podríamos hacer la siguiente transformación:

```
> n<-length(x)
```

```
> var(x)*(n-1)/n;sd(x)*sqrt((n-1)/n)
```

```
[1] 695.9225
```

```
[1] 26.38034
```

El comando `length(x)` devuelve el número de elementos de la variable (en este caso son 100, ya que de los 196 países había 96 que no tenían información) y `sqrt` permite calcular la raíz cuadrada. Puede observarse que la diferencia no es excesivamente grande (en un caso dividimos por 99 y en otro caso por 100, en el cálculo de la varianza). Cabe destacar que las unidades de la varianza (y de la

cuasivarianza) son las de la variable de interés al cuadrado (que en este caso no tiene sentido físico) y que no son comparables con las propias mediciones de la variable. En nuestro ejemplo, la desviación típica muestral es del 26.38 %. Utilizaremos este dato más adelante.

4.1.5. Medidas de forma

Consideraremos dos medidas que proporcionan una idea de la forma de la distribución los datos. Su cálculo no es tan sencillo como el de las medidas de posición y dispersión estudiadas y lo que nos interesa es su interpretación.

- **Coefficiente de asimetría.** El coeficiente de asimetría de Fisher toma valor 0 cuando la distribución de los datos es simétrica con respecto a la media. Valores positivos de este coeficiente indicarán la presencia de asimetría positiva (más datos con valores superiores a la media), mientras que valores negativos son indicativos de una asimetría negativa (más datos con valores inferiores a la media). Dichas situaciones aparecen reflejadas en la Figura 5. El coeficiente de asimetría se calcula como:

$$\gamma_F = \frac{1}{s^3} \frac{(x_1 - \bar{x})^3 + \dots + (x_n - \bar{x})^3}{n} = \frac{1}{s^3} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3.$$

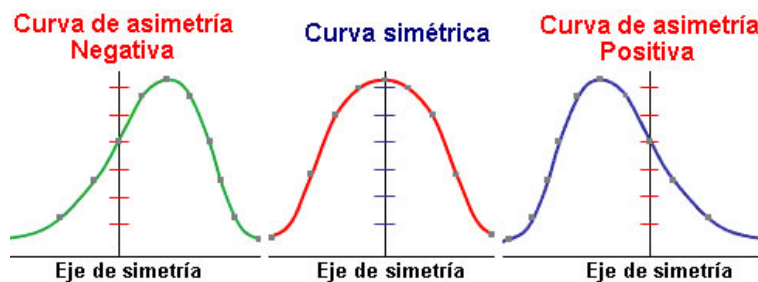


Figura 5. Interpretación del coeficiente de asimetría.

Además, para cuantificar la asimetría de una muestra de datos, también podemos utilizar los cuartiles. Si la distribución es simétrica, la distancia entre Q_3 y Q_2 (que contiene un 25 % de la muestra) y entre Q_2 y Q_1 (otro 25 %), debería ser la misma (es decir, $Q_3 - Q_2 = Q_2 - Q_1$). Así, si $Q_3 - Q_2 > Q_2 - Q_1$, es indicativo de asimetría positiva. Por otro lado, si $Q_3 - Q_2 < Q_2 - Q_1$, tendríamos indicios de asimetría negativa. Para que el resultado no dependa de la dimensión de los datos, podemos utilizar el siguiente índice de asimetría que toma valores en $[-1, 1]$, basado en los cuartiles:

$$\gamma_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}.$$

- **Coeficiente de curtosis.** El coeficiente de curtosis mide el grado de apuntamiento de la distribución. Su fórmula es:

$$\gamma_C = \frac{1}{s^4} \frac{(x_1 - \bar{x})^4 + \dots + (x_n - \bar{x})^4}{n} = \frac{1}{s^4} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

Si $\gamma_C > 3$, se dice que la distribución de frecuencias es leptocúrtica. Si $\gamma_C < 3$, la distribución de frecuencias es platicúrtica. Dichas situaciones aparecen reflejadas en la Figura 6. También se puede modificar la expresión anterior y considerar $\gamma_C^* = \gamma_C - 3$, ya que 3 es el valor del coeficiente cuando los datos vienen de una distribución Normal (que es la de referencia). De este modo, tendremos distribuciones leptocúrticas si $\gamma_C^* > 0$ y platicúrticas si $\gamma_C^* < 0$.

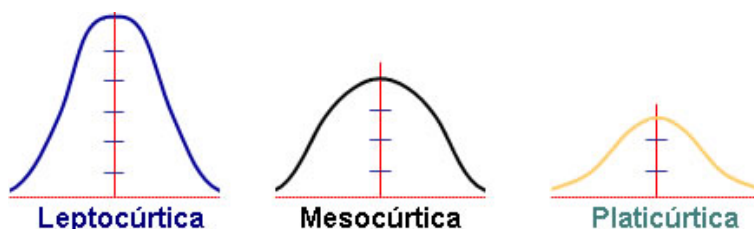


Figura 6. Interpretación del coeficiente de asimetría.

4.2. Representación de medidas: el diagrama de caja

Las representaciones gráficas que se han descrito en la sección anterior utilizan los datos observados para su construcción o la información que se obtiene en las tablas de frecuencias. A partir de las medidas características que se han descrito, se puede construir una nueva representación, que se conoce como diagrama de caja.

El diagrama de caja se construye a partir de las siguientes medidas que aparecen reflejadas en la Figura 7:

- El primer y el tercer cuartil, Q_1 y Q_3 , que delimitan la caja central. La longitud de la caja viene dada por el *RIC*, que es una medida de dispersión absoluta.
- Los límites inferior y superior se calculan como:

$$LI = \max\{\min\{x_i\}, Q_1 - 1.5(Q_3 - Q_1)\},$$

$$LS = \min\{\max\{x_i\}, Q_3 + 1.5(Q_3 - Q_1)\}.$$

En el cálculo de los límites inferior y superior se utiliza el $RIC = Q_3 - Q_1$.

- La mediana (Q_2) se representa con una línea horizontal en la caja central.

El diagrama de caja se utiliza para determinar los valores atípicos de la muestra, que son datos que difieren numéricamente de los demás. Formalmente, los datos atípicos son aquellos datos que quedan fuera del intervalo (LI, LS) . Si en lugar de considerar los límites inferior y superior construimos el intervalo (LI_e, LS_e) donde $LI_e = Q_1 - 3RIC$ y $LS_e = Q_3 + 3RIC$, los datos que caen fuera de este intervalo se denominan extremos. Algunos paquetes estadísticos hacen la distinción entre atípicos y extremos, representándolos de distintas formas en las salidas gráficas.

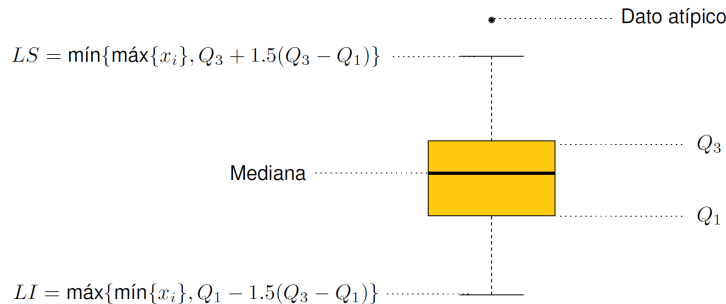


Figura 7. Diagramas de caja o boxplot.

En la Figura 8 se muestran los diagramas de caja para el número de ordenadores por cada 100 habitantes y para su logaritmo. Se puede observar la presencia de datos atípicos altos, representados con puntos. Sin embargo, un problema del diagrama de caja es que no permiten observar la presencia de multimodalidad. Para tener una información más completa sobre el comportamiento de la muestra de datos se deben utilizar conjuntamente el diagrama de cajas y el histograma.

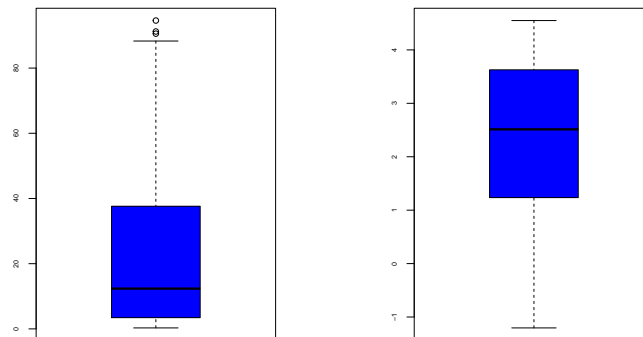


Figura 8. Diagramas de caja para el número de ordenadores personales por cada 100 habitantes (izquierda) y su logaritmo (derecha).

Ejemplo. Los diagramas de caja se han obtenido con los siguientes comandos de .

```
> par(mfrow=c(1,2))
> boxplot(x,col=4)
> boxplot(log(x),col=4)
```

En el ejemplo que nos ocupa, nos puede interesar analizar un poco más en profundidad los datos. Como comentamos, aunque los análisis mostrados hasta ahora se realizaron para los datos de 2006, existen datos de años anteriores. Podríamos calcular la media para cada año y así ver la evolución del número medio de ordenadores personales por cada 100 habitantes. Lo mismo se puede hacer con la mediana. Estos datos se muestran en la Figura 9, en negro para la media y el rojo para la mediana. Puede verse que los valores de la media son siempre superiores a los de la mediana, indicando que el comportamiento asimétrico que se observaba para los datos de 2006 aparece también en los otros años. De hecho, esta última afirmación se ve corroborada en la Figura 10 donde se representan los cuartiles para cada año (así como el máximo y el mínimo, en negro). Puede verse que la distancia entre el tercer cuartile y la mediana (puntos verdes y azules) se va haciendo cada vez mayor, lo que indica que la asimetría va aumentando a medida que pasan los años.

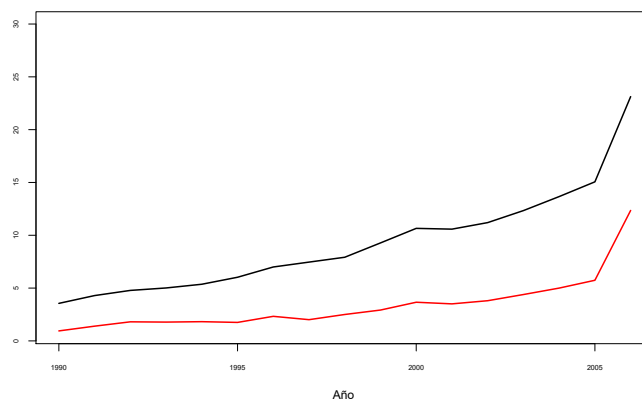


Figura 9. Evolución de la media y la mediana para le número de ordenadores personales por cada 100 habitantes. Línea negra: media. Línea roja: mediana

4.3. Tipificación de datos

El coeficiente de variación, como ya hemos visto, se utiliza para comparar variables. Si lo que queremos es comparar individuos de distintos grupos, debemos utilizar la **tipificación de datos**. A partir de una muestra x_1, \dots, x_n con media \bar{x} y varianza s^2 , los datos tipificados se construyen como:

$$z_i = \frac{x_i - \bar{x}}{s}$$

de manera que la muestra resultante z_1, \dots, z_n tendrá media 0 y varianza 1. La tipificación de datos permite comparar distintos grupos, así como la posición relativa de las observaciones en cada uno.

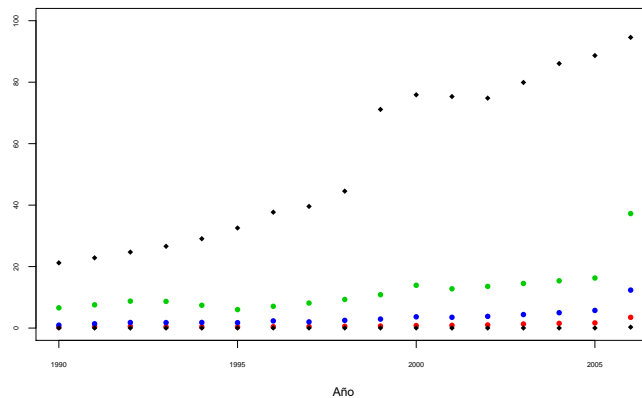


Figura 10. Evolución de los valores mínimo, máximo y cuartiles.

Ejemplo. Para los datos del ejemplo, podríamos considerar las distintas regiones geográficas. En cada una de ellas, tenemos las siguientes medias y desviaciones típicas: Consideremos los datos para

Región	Media	Desv. típica
América	23.23	30.68
Este de Asia y Pacífico	25.04	26.29
Europa y Asia Central	38.97	29.08
Medio Este y Norte de África	10.70	7.86
Sur de Asia	5.47	8.22
África Subsahariana	5.67	6.81

Tabla 4. Medias y desviaciones típicas del número de ordenadores personales por cada 100 habitantes para el año 2006.

España y para los Emiratos Árabes, que eran de 36.68 y 30.06 ordenadores personales por cada 100 habitantes, respectivamente. Observamos que el valor para los Emiratos es inferior que para España, pero si los ponemos en relación con la zona a la que pertenecen (Europa y Asia Central, para España, y Medio Este y Norte de África, para los Emiratos) vemos que el valor para España está por debajo de la media de su región, mientras que para los Emiratos está por encima. Teniendo en cuenta además la dispersión (desviación típica) de los datos de ambas regiones, tenemos los valores tipificados que son de -0.072 y 2.46 , respectivamente. El valor de España está cerca del cero (el valor original está cerca de la media) y es negativo (valor original menor que la media). Si hiciéramos una comparativa sobre las posiciones que ocupan España y los Emiratos con respecto a esta variable en su zona de referencia, los Emiratos estarían por delante de España en número de ordenadores personales por cada 100 habitantes.

5. Tablas de contingencia

En muchos casos se miden simultáneamente en cada individuo de una población dos o más variables con la finalidad de establecer la existencia (o inexistencia) de relaciones entre ellas. Por ejemplo, ¿existe relación entre el PageRank de una página web y el número de visitas que recibe? ¿existe relación entre el sexo y la nota que se saca en un determinado examen? ¿existe relación entre el género y el equipo de fútbol preferido?

Las situaciones que pueden aparecer en este tipo de problemas son muy variadas ya que dependen de la naturaleza de las variables que hayamos medido. Aquí nos centraremos en aquellas situaciones en que se han recogido simultáneamente dos variables que forman grupos dentro de la población. Dejaremos para más adelante, el análisis de la relación entre dos variables continuas.

A modo de ejemplo consideraremos los datos que aparecen resumidos en la Tabla 5, recogidos por Fisher en 1940. La variable X es el color de ojos mientras que Y es el color del pelo para un grupo de escolares escoceses.

$X \backslash Y$	rubio	pelirrojo	castaño	oscuro	negro
claros	688	116	584	188	4
azules	326	38	241	110	3
castaños	343	84	909	412	26
oscuros	98	48	403	681	85

Tabla 5. Tabla de contingencia del color de ojos y el color del pelo de escolares escoceses (Fisher, 1940)

La información recogida en el ejemplo anterior se puede entonces representar en una **tabla de contingencia**, que en general, tendría el aspecto que se muestra en la Tabla 6, donde n_{ij} es la frecuencia absoluta de la clase o categoría (c_i, d_j) , $i = 1, \dots, k$ y $j = 1, \dots, l$, es decir el número de veces que

$X \backslash Y$	d_1	\dots	d_j	\dots	d_l
c_1	n_{11}	\dots	n_{1j}	\dots	n_{1l}
\vdots	\vdots		\vdots		\vdots
c_i	n_{i1}	\dots	n_{ij}	\dots	n_{il}
\vdots	\vdots		\vdots		\vdots
c_k	n_{k1}	\dots	n_{kj}	\dots	n_{kl}

Tabla 6. Esquema de una tabla de contingencia con frecuencias absolutas.

$X \backslash Y$	d_1	\dots	d_j	\dots	d_l
c_1	f_{11}	\dots	f_{1j}	\dots	f_{1l}
\vdots	\vdots		\vdots		\vdots
c_i	f_{i1}	\dots	f_{ij}	\dots	f_{il}
\vdots	\vdots		\vdots		\vdots
c_k	f_{k1}	\dots	f_{kj}	\dots	f_{kl}

Tabla 7. Esquema de tabla de contingencia con frecuencias relativas.

se observó dicho dato. El tamaño de la muestra se calculará sumando todas las frecuencias absolutas

$$n = \sum_{i=1}^k \sum_{j=1}^l n_{ij} = \sum_{i,j} n_{ij}.$$

La frecuencia relativa de la clase conjunta (c_i, d_j) se define como

$$f_{ij} = \frac{n_{ij}}{n}, \quad \text{con } i = 1, \dots, k, j = 1, \dots, l,$$

que también se pueden representar en una tabla como se muestra en la Tabla 7. En el ejemplo del color de pelo y ojos de los escolares escoceses obtendríamos la Tabla 8.

$X \backslash Y$	rubio	pelirrojo	castaño	oscuro	negro
claros	0.130	0.022	0.110	0.035	0.001
azules	0.045	0.007	0.045	0.021	0.001
castaños	0.065	0.016	0.172	0.078	0.005
oscuros	0.019	0.009	0.076	0.129	0.016

Tabla 8. Tabla de frecuencias relativas para los datos del color de ojos y el color del pelo de escolares escoceses.

Las distribuciones de frecuencias de las variables X e Y , obtenidas sin tener en cuenta la existencia de una segunda variable, se denominan **distribuciones marginales**. Por ejemplo, considerando el esquema de la Tabla 6, para la variable X la frecuencia absoluta de la clase c_i , que denotaremos mediante $n_{i\bullet}$, se calculará sumando todos los datos de la fila i

$$n_{i\bullet} = n_{i1} + n_{i2} + \dots + n_{il}.$$

Análogamente, su frecuencia relativa, se calculará mediante $f_{i\bullet} = \frac{n_{i\bullet}}{n}$. Usualmente las distribuciones marginales aparecerán en los márgenes de la tabla de contingencia tal como se muestra en la Tabla 9.

$X \backslash Y$	y_1	\dots	y_j	\dots	y_l	
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1l}	$n_{1\bullet}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_i	n_{i1}	\dots	n_{ij}	\dots	n_{il}	$n_{i\bullet}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_k	n_{k1}	\dots	n_{kj}	\dots	n_{kl}	$n_{k\bullet}$
	$n_{\bullet 1}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet l}$	n

Tabla 9. Tabla de contingencia con distribuciones marginales.

6. Recta de regresión

Existen muchas situaciones que requieren el análisis combinado de dos ó más variables cuantitativas continuas, debido a las posibles relaciones entre ellas. Una forma de representar la relación entre ellas es a través de la **recta de regresión**. En esta sección introduciremos las medidas características conjuntas más usuales en este contexto (vector de medias y matriz de varianzas-covarianzas) y veremos cómo se construye una recta de regresión.²

6.1. Vector de medias. Covarianza y correlación

Supongamos que tenemos una variable bidimensional (X, Y) y que disponemos de las observaciones en una muestra de tamaño n , $\{(x_i, y_i)\}_{i=1}^n$. Se denomina **vector de medias** al vector cuyas componentes son las medias muestrales de las variables: (\bar{x}, \bar{y}) .

Para representar la dispersión podemos considerar los valores de las varianzas de cada variable por separado, es decir, s_x^2 y s_y^2 , pero quedaría sin resumir la variabilidad conjunta de ambas. Por eso debemos introducir la **covarianza** muestral. La covarianza entre dos variables X e Y , que es una medida que indica la variabilidad conjunta de X e Y , se calcula como:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}.$$

A partir de las varianzas y la covarianza se obtiene la **matriz de varianzas-covarianzas**:

$$S = \begin{pmatrix} s_x^2 & S_{xy} \\ S_{xy} & s_y^2 \end{pmatrix}$$

²La recta de regresión es parte del análisis descriptivo correspondiente al modelo de regresión, que estudiaremos más adelante.

Covarianza y correlación

El signo de la covarianza proporciona información sobre el tipo de relación que existe entre las variables. De este modo:

- a) Si la relación entre las variables es directa, entonces $S_{xy} > 0$.
- b) Si la relación entre las variables es inversa, entonces $S_{xy} < 0$.
- c) Si no existe relación lineal entre las variables, entonces $S_{xy} = 0$.

Las parejas de datos (x_i, y_i) con $i = 1, \dots, n$, de las dos variables (X, Y) , se pueden representar a partir de una **nube de puntos** o **diagrama de dispersión**. Esta representación gráfica se construye representando sobre un plano los valores de los puntos observados. En la Figura 11 podemos ver dos ejemplos de relaciones entre variables. La covarianza de los datos de la izquierda es positiva, mientras que la covarianza de los datos de la derecha es negativa. Así, diremos que la relación entre X e Y es directa cuando valores altos de X se corresponden con valores altos de Y . La relación se dice que es inversa si valores altos de X se corresponden con valores bajos de Y , o viceversa.

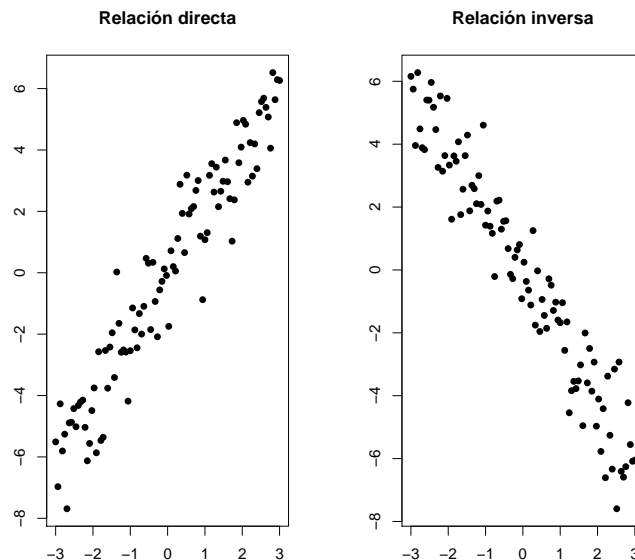


Figura 11. Ejemplo de diagramas de dispersión. Relaciones directa e inversa.

Al igual que ocurría con la varianza muestral, la covarianza está afectada por las unidades de medida de las variables, por lo que definiremos una medida característica para explicar la relación lineal entre variables que sea adimensional: el **coeficiente de correlación lineal**. A partir de una muestra de datos

$\{(x_i, y_i)\}_{i=1}^n$, el coeficiente de correlación lineal se calcula como:

$$r_{xy} = \frac{S_{xy}}{s_x s_y},$$

donde S_{xy} es la covarianza muestral y s_x, s_y son las respectivas desviaciones típicas muestrales.

El coeficiente de correlación lineal no tiene dimensiones y toma valores en $[-1, 1]$. Valores cercanos a 1 indicarían una relación lineal directa, mientras que valores cercanos a -1 darían una relación lineal inversa. Si $r_{xy} = 0$ diríamos que las variables están incorreladas.

Cuando existe una relación lineal entre dos variables, podemos tratar de buscar una formulación matemática que describa una en función de otra. La regresión lineal simple consiste en aproximar los valores de una variable a partir de los de otra utilizando una relación de tipo lineal. La recta de regresión de Y sobre X tendrá la siguiente expresión:

$$Y = a + bX + \varepsilon,$$

donde a representa la ordenada en el origen o intercepto, b es la pendiente (indica la razón de cambio en Y cuando X varía en una unidad) y ε el error desconocido (que tiene media 0). La variable X se denomina variable explicativa o independiente, mientras que la variable Y será la variable respuesta, o variable dependiente.

6.2. Método de Mínimos Cuadrados

En la práctica, a partir de los datos $\{(x_i, y_i)\}_{i=1}^n$ podremos estimar los valores de a y b . El objetivo será obtener los valores a y b que nos proporcionen los residuos más pequeños. Los residuos son las diferencias entre los valores observados de la variable respuesta y_i y los valores que proporciona el ajuste $\hat{y}_i = \hat{a} + \hat{b}x_i$ y vienen dados por:

$$e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i, \quad i = 1, \dots, n.$$

En la Figura 12, los segmentos verticales son los residuos, que representan la diferencia entre el valor observado y el valor que daría la recta ajustada.

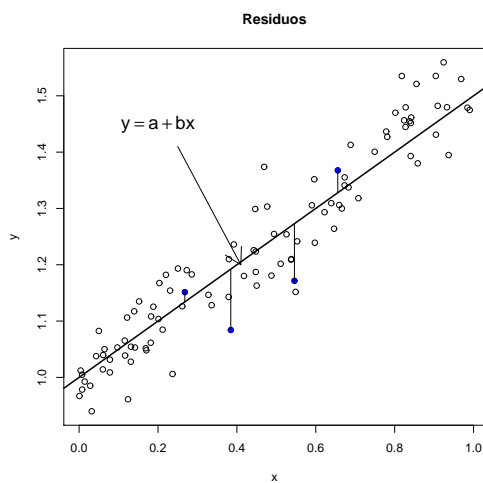


Figura 12. Residuos a minimizar en el método de mínimos cuadrados. Los segmentos verticales representan los residuos e_i .

El **Método de Mínimos Cuadrados** consiste en minimizar la suma de los cuadrados de los residuos, por lo que se buscan los valores a y b que minimizan:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

A partir del Método de Mínimos Cuadrados, se obtienen los valores para \hat{a} y \hat{b} :

$$\hat{b} = \frac{S_{xy}}{s_x^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x},$$

donde \bar{y} y \bar{x} denotan las medias muestrales de y_1, \dots, y_n y x_1, \dots, x_n , respectivamente; s_x^2 es la varianza muestral de X :

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

y S_{xy} es la **covarianza** muestral entre X e Y . En la Figura 12, representamos la recta ajustada, con a y b obtenidos por el método de Mínimos Cuadrados. Se puede comprobar que la recta de regresión ajustada por Mínimos Cuadrados pasa por el vector de medias (\bar{x}, \bar{y}) .

La recta de regresión de Y sobre X se puede utilizar para predecir valores de Y conocidos los valores de X , pero no al revés. Además, las predicciones con la recta de Y sobre X sólo son razonables cuando el valor de X para el que queremos hacer la predicción se encuentra entre el mínimo y el máximo de los valores observados para la variable.

Si quisiéramos hacer predicciones sobre el valor de X dado un valor de Y , tendríamos que utilizar la recta de regresión:

$$X = c + dY + \varepsilon_2, \quad \text{con} \quad \hat{d} = \frac{S_{xy}}{s_y^2}, \quad \hat{c} = \bar{x} - \hat{d}\bar{y}.$$

6.3. Coeficiente de regresión. Coeficiente de determinación

Coeficiente de regresión.

Se denomina **coeficiente de regresión** a la pendiente (parámetro b) de la recta de regresión de Y sobre X . Este coeficiente proporciona información sobre el comportamiento de la variable respuesta Y en función de la variable explicativa X y tiene el mismo signo que la covarianza.

- a) Si $b > 0$, al aumentar los valores de X también aumentan los valores de Y .
- b) Si $b < 0$, al aumentar los valores de X , los valores de Y disminuyen.

Coeficiente de determinación

Una medida para determinar cómo de bueno es el ajuste del modelo es el **coeficiente de determinación** (R^2) que mide la proporción de variabilidad de Y que explica X a través de la recta de regresión.

El coeficiente de determinación es el cuadrado del coeficiente de correlación lineal, y toma valores entre 0 y 1. Si R^2 toma valores próximos a 1, esto será indicativo de un buen ajuste. El coeficiente de determinación del modelo de regresión lineal simple viene dado por:

$$R^2 = r_{xy}^2 = \frac{S_{xy}^2}{s_x^2 s_y^2}.$$

El coeficiente de determinación, y por tanto, la variabilidad explicada por la recta de regresión de Y sobre X y la de X sobre Y es el mismo.

Anexo. Resumen

Medidas características			
Muestra x_1, \dots, x_n			
Tipo		Medida	Fórmula/definición
Posición	Central	Media	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
		Mediana	es el punto hasta el cual se encuentran el 50 % de los casos en la muestra
		Moda	es el valor que más se repite
	No central	Cuartiles	los cuartiles Q_1 , Q_2 y Q_3 dividen la muestra en cuatro partes iguales
		Deciles	ídem pero en 10 partes
		Percentiles	ídem pero en 100 partes
Dispersión	Absoluta	Varianza	$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
		Desviación típica	$s = \sqrt{s^2}$
		Cuasi-varianza	$s_c^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
		Cuasi-desviación típica	$s_c = \sqrt{s_c^2}$
		Rango	$R = \max\{x_i\} - \min\{x_i\}$
		Rango intercuartílico	$RIC = Q_3 - Q_1$
	Relativa	Coef. de variación	$CV = \frac{s}{\bar{x}}, \quad \bar{x} > 0$
		Rango relativo	$RR = R/\bar{x}, \quad \bar{x} > 0$
Forma	Asimetría	Coef. γ_F	$\frac{1}{s^3} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$
	Curtosis	Coef. γ_C	$\frac{1}{s^4} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$

Tabla 10. Tabla resumen de medidas características.

Muestra $(x_1, y_1), \dots, (x_n, y_n)$	
Vector de medias	(\bar{x}, \bar{y})
Covarianza	$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$
Matriz de varianzas-covarianzas	$S = \begin{pmatrix} s_x^2 & S_{xy} \\ S_{xy} & s_y^2 \end{pmatrix}$
Coeficiente de correlación lineal	$r_{xy} = \frac{S_{xy}}{s_x s_y}$
Recta de regresión $Y = a + bX + \varepsilon$	$\hat{b} = \frac{S_{xy}}{s_x^2}$ $\hat{a} = \bar{y} - \hat{b}\bar{x}$

Tabla 11. Tabla resumen de medidas características bidimensionales y ajuste de regresión utilizando el método de mínimos cuadrados.