Capítulo 1

Estadística descriptiva

1. Introducción. El propósito de la estadística
2. Tipos de datos
3. Descripción de datos mediante tablas
4. Descripción de datos mediante gráficos
5. Medidas características de una variable
6. Transformaciones lineales y su efecto en las medidas características
7. Transformaciones no lineales que mejoran la simetría
8. Relación entre dos variables. La recta de regresión

[0] Apuntes realizados por Ismael Sánchez para la asignatura de Estadística. Universidad Carlos III de Madrid.

1.1. Introducción. El propósito de la estadística.

¿Qué es la estadística?¿Para qué sirve la estadística en ingeniería? Aunque ambas preguntas pueden tener una respuesta muy compleja, se puede decir resumidamente que la estadística es **la ciencia que nos permite analizar un conjunto de datos observados con el fin de conocer las características del fenómeno real que los ha generado**. De esta forma, si conseguimos conocer el fenómeno que genera los datos podremos anticipar cómo serán los siguientes datos que se observen, o podremos aprender a provocarlos o evitarlos. Por ejemplo, supongamos que un fabricante de cierto componente electrónico desea saber si el procedimiento de fabricación A es más recomendable que el procedimiento B. Para ello, y en función de sus limitaciones presupuestarias, fabrica unos cuantos componentes con el procedimiento A y otros con el B. Mediante técnicas estadísticas podrá conocer con suficiente confianza si el método A es mejor o peor que el B, o por el contario, los datos no muestran evidencia suficiente para decidir al respecto. El fabricante podrá utilizar este resultado para decidir modificaciones en el procedimiento de fabricación de futuras piezas.

Es muy importante darse cuenta de que el fabricante no tiene un interés especial en conocer las diferencias entre los componentes **ya fabricados**, sino en tomar una decisión que afecte a los que **se fabricarán** en el futuro. Supongamos, por ejemplo, que se han fabricado 100 componentes: 50 con el procedimiento de producción A y los otros 50 con el procedimiento de producción B. Después de analizar estos 100 componentes se ha visto que 10 componentes del procedimiento A son defectuosos, mientras que con el procedimiento B sólo son defectuosos 8. ¿Qué método deberíamos implantar para producir a gran escala?. El fabricante no duda de que en esos 100 componentes, el método B fue mejor, pero ¿qué garantías ofrecen esos datos de que esa ventaja se mantendrá en el futuro?¿Podemos hacer el enunciado **general** de que el método de producción B es mejor que el A para fabricar esos componentes? Mediante el uso de la estadística podremos concluir si a la vista de los datos es razonable decantarse por el procedimiento B o por el contrario la ventaja observada en esos 100 componentes no es suficiente como para extrapolarla con un mínimo de seguridad. Un elemento a destacar en esta última frase es la idea de que no es seguro que en el futuro se observe el mismo resultado que en la muestra de datos. Para que la estadística sea útil debe proporcionar una idea del riesgo que cometemos al generalizar lo observado en los datos. Dedicaremos una parte importante de este texto a proporcionar medidas del riesgo o confianza de las decisiones basadas en datos.

Puede decirse entonces que la estadística es una **herramienta de aprendizaje a partir de la observación**, pues nos ayuda a extraer **conclusiones generalizables** a partir de un conjunto de datos observados. Este proceso por el que a partir de los hechos, ascendemos lógicamente hasta la ley o principio que los contiene se denomina **inducción** o **inferencia**. Es decir, a partir de lo particular construimos lo general. La inducción es el proceso contrario a la **deducción**. En la deducción, a partir de una ley general podemos explicar cualquier resultado particular. Por ejemplo, partiendo de la ley general que dice que todo cuerpo abandonado a cierta altura sin obstáculo tiende a caer por acción de la gravedad de la Tierra, podemos deducir que si una manzana madura se desprende del árbol, no quedará flotando en el aire sino que caerá. El tener una ley general que permita, mediante la deducción, describir un fenómeno es muy útil, pues permite predecir y explicar cualquier resultado particular. Pero, ¿cómo llegamos a dicha ley general? Precisamente en el ejemplo de la caída de la manzana por acción de la gravedad, Newton llegó a la ley general usando el camino inverso: la inducción.

Los fenómenos reales que interesan en ingeniería son con frecuencia demasiado complejos para ser previstos utilizando sólo los principios físicos que los rigen en condiciones ideales. Es frecuente

encontrar discrepancias entre lo que se observa y lo que predice un modelo teórico. La razón es que en el modelo teórico sólo es posible tener en cuenta un número limitado de factores, mientras que en la realidad influyen muchos más: rozamientos, heterogeneidad de materiales, impurezas, interacciones, etc. Como consecuencia, y debido al efecto de dichos factores que no contempla el modelo teórico, siempre existirá una incertidumbre sobre el resultado final, una posible discrepancia entre el valor observado y el valor previsto. En otras ocasiones no se dispone de ningún marco teórico que explique la realidad, por lo que debemos construir dicho modelo teórico exclusivamente a partir de la observación.

La principal contribución de la estadística a la resolución de problemas en ingeniería es su utilización como *herramienta de aprendizaje*. El aprendizaje que se obtiene mediante la estadística forma parte del denominado **método científico**. El método científico de aprendizaje consiste en la aplicación de forma iterativa de un proceso de **inducción-deducción** que puede resumirse en los siguiente pasos (ver figura 1.1):

1. Una hipótesis inicial o teoría permite construir un modelo que explique la realidad. Este modelo lleva, mediante un proceso de **deducción,** a ciertas consecuencias que deberían observarse si el modelo utilizado fuese realmente válido.

2. La observación de los datos reales pueden entonces revelar una discrepancia observada respecto a lo que predice el modelo. A esta discrepancia le llamaremos **error experimental**.

3. Dicho error experimental observado conduce, mediante un proceso de **inducción**, a la **modificación del modelo**.

4. Este proceso se **itera** hasta que se tiene un modelo capaz de explicar los datos observados de forma suficientemente precisa.

Este ciclo de deducción-inducción es la base de la aplicación de la estadística. La deducción se realiza a partir de los modelos que pueda ofrecer la ciencia o los que haya podido elaborar el investigador de forma empírica usando cualquier metodología estadística. Los datos observados serán tanto más informativos cuanto más cuidadosa haya sido su recogida, lo que precisa de un diseño del experimento adecuado. Finalmente, la interpretación del error observado ha de hacerse en el marco de la variabilidad que todo experimento tiene, y por tanto ha de ser valorado estadísticamente.

La mayoría de los ingenieros ignora que la estadística es una herramienta de aprendizaje muy eficaz, debido principalmente a su **visión determinista de la realidad**. Esta visión determinista está producida por su instrucción, basada en leyes físicas o matemáticas, tales como las ecuaciones de Maxwell en electromagnetismo, las leyes de la Termodinámica, o las leyes de la Mecánica de Newton. Todo este conjunto de leyes científicas son muy útiles e indispensables en la actividad de cualquier científico, pero están basadas en entornos y materiales ideales. Asímismo, ignoran la interacción que puede suceder cuando se combinan varios factores en la realidad. Cuando un ingeniero prevee un resultado a partir de una formulación teórica está empleando un razonamiento deductivo. Cualquier desviación en el valor observado respecto al valor previsto será en general o bien ignorado si el error es pequeño o bien puede provocar el fracaso del proyecto. Por esta razón, **el uso del conocimiento científico de forma puramente deductiva es ineficaz**. Mediante la utilización de la estadística, el ingeniero aprovechará el error de experimentación para convertirlo, mediante inducción, en un mayor conocimiento.

Ejercicios 1 .....

1. Resume el contenido de esta sección en una sola frase

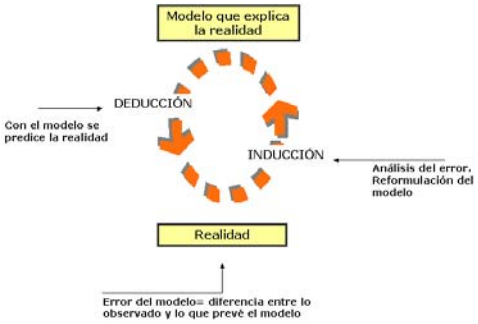


Figura 1.1: El método científico de aprendizaje. Ciclo inducción-deducción

2. Indica cuáles de los siguientes razonamientos son inductivos y cuáles deductivos

   a) Después de lanzar varias veces un dado, un jugador piensa que el dado está trucado

   b) El estudio de la serie de temperaturas sugiere que se está produciendo un cambio climático

   c) En el piso de enfrente no debe vivir nadie, pues nunca he visto las luces encendidas

   d) Como este procesador es más rápido, tardará menos en ejecutar mi programa

   e) Tras un experimento para decidir el material a utilizar para un sistema de sujección, se anota la tensión de rotura de 20 componentes del material A y otros 20 del material B. Tras ver los datos se decide que se usará el material A pues parece que será más resistente que el B.

   f) La energía cinética de un cuerpo de masa 10 Kg y velocidad 5 m/s es de $E = 1/2mv^2 = 25$ (julios)

3. Una persona ha lanzado una moneda 20 veces y ha sacado cara 15 veces. ¿Podemos concluir que esa persona tiene una forma de lanzar que favorezca que salga cara?

. . . . . .

1.2. Tipos de datos

En estadística, la materia prima son los datos y el producto final son un conjunto de conclusiones sobre el fenómeno que ha originado esos datos. Un dato es un valor concreto que toma una variable en un momento dado en un individuo. Los **individuos** serán los objetos de los que obtenemos los datos. Los individuos pueden ser desde personas a simples objetos. Denominaremos **variable** a la característica de interés, y que puede tomar un valor diferente en cada individuo. Si la característica de interés tomase siempre el mismo valor, ya no sería una variable, sino una **constante**. Una variable podría tomar valores diferentes incluso en un mismo individuo, si las condiciones cambiasen cada vez que repetimos la medición de la variable.

Un **dato** es el valor observado de una variable en un individuo en una medición. Se suele decir también que **un dato es la realización de una variable**. Por ejemplo, una variable puede ser la temperatura de una habitación, y un dato será la temperatura de una habitación en un instante dado. La estatura es una variable, y mi estatura es un dato. A efectos de lo que nos interesa en estadística, los datos se pueden clasificar según varios criterios.

Clasificación de los datos según su naturaleza

Según su naturaleza, los datos se pueden dividir en:

- **Datos cuantitativos:** cuyos valores son números que representan cantidades. A su vez los dividiremos en
  - Continuos: si pueden tomar cualquier valor dentro de un intervalo real. Por ejemplo, longitudes, pesos, tiempos, voltajes, etc
  - Discretos: si sólo pueden tomar valores enteros. Por ejemplo, número de cilindros de un vehículo, número de miembros de una familia, clientes que llegan por unidad de tiempo a un puesto de servicio, etc.
- **Datos cualitativos:** son aquellos cuyos valores son códigos que representan atributos. Por ejemplo: color de ojos, marca preferida de un producto, tipo de procesador de un ordenador. A los datos cualitativos también se les denomina categóricos, pues clasifican a los individuos según las diferentes categorías en que se divide la variable cualitativa.

Clasificación de los datos según su representatividad

Según su representatividad, los datos pueden formar toda la **población** o ser sólo una **muestra**.

- **Población:** es el conjunto de todos los datos posibles de una variable. Por ejemplo, en un aula de 83 alumnos, la población de estaturas son las 83 estaturas.
- **Muestra:** es un subconjunto de la población. Por ejemplo, en el aula de 83 alumnos, una muestra de estaturas sería la estatura de sólo 20 de ellos.

No se debe confundir la población con el conjunto de valores diferentes que puede tomar la variable. En el caso de las estaturas de los 83 alumnos, los valores posibles de altura son infinitos, al ser la estatura una variable continua. En el caso de la variable 'número de hermanos del alumno' posiblemente la variable tome valores entre 0 y 5, pero la población serán los 83 datos formados por el número de hermanos de los 83 alumnos, existiendo datos repetidos. **Tampoco debe confundirse la población con los individuos** sobre los que se mide la variable. En el ejemplo de la variable estastura de los 83 estudiantes, la población son las 83 estaturas y no los 83 estudiantes.

Para poder manejar mejor el concepto de población y muestra en estadística, será util definir previamente el concepto de experimento. En estadística, el término **experimento** tiene un significado muy amplio:

Definición 1 *Experimento es cualquier procedimiento de obtención de un dato, dadas una condiciones de experimentación. Si manteniendo las condiciones de experimentación constantes volvemos a obtener otro dato, estaremos repitiendo ese experimento.*

Un experimento puede ser tanto observar un dato registrado, como producirlo en un laboratorio, o bien obtenerlo mediante una encuesta. Si se cambian las condiciones de experimentación, el experimento será distinto. Por tanto, para definir un experimento no sólo hay que especificar la variable que vamos a medir o evaluar, sino en qué condiciones de experimentación lo vamos a hacer. De esta forma, si los datos se obtienen midiendo elementos sin reposición, el experimento será diferente a si la medición se hace con reposición. Una vez definido experimento, podemos definir población:

Definición 2 *Población es el conjunto de datos que se obtienen al repetir un experimento concreto todas las veces posibles.*

Cuando pensamos en poblaciones como colección de las mediciones de todos los individuos con cierta característica (por ejemplo, las 83 estaturas de los 83 estudiantes de un aula), estamos refiriéndonos a experimentos obtenidos sin reposición. La población será finita y tendrá 83 elementos. Cuando las mediciones se obtiene con reposición, como cuando se lanza repetidamente un dado, el experimento ya es distinto; las poblaciones serán de tamaño infinito y sólo existirán conceptualmente.

En estadística, habitualmente analizamos sólo una muestra cuyo tamaño depende de los recursos que se posean para la investigación. Sin embargo, el interés estará en aportar conocimientos no sobre la muestra sino sobre toda la población, es decir, **inferir cómo será la población a partir de la información limitada de la muestra**. Es muy posible que los elementos que constituyen la muestra hayan sido seleccionados al azar y sólo tendrán interés como representantes de la población que es objeto de estudio. Es importante entonces haber definido bien la población. En ocasiones, una investigación termina en fracaso si no se sabe a ciencia cierta a qué colectivo (población) representan los datos analizados. Dada una muestra de datos, no siempre es fácil deducir cuál es la población de la que proceden. Por ejemplo, si analizamos el gasto en teléfono móvil de 30 alumnos de primer curso del grupo de tarde de la titulación de Ingeniería Informática de la Universidad Carlos III de Madrid, en el campus de Leganés ¿a qué colectivo podemos extender dichas conclusiones? ¿A todos los estudiantes universitarios? ¿A los de la Carlos III? ¿A los jóvenes españoles? ¿A los jóvenes de clase media?¿A los estudiantes de turno de tarde? ¿A todos los estudiantes del campus de Leganés?. La respuesta no es sencilla. Es mucho más efectivo pensar primero en qué población estamos interesados y pensar después en algún procedimiento que nos permita obtener una muestra representativa.

Ejercicios 2 .....

1. A continuación, se presentan unas variables y un conjunto de datos de cada una. Decir qué conjunto de datos son una muestra y cuáles toda la población. En el caso de muestras, decir cómo se obtendría toda la población.

   a) Variable: consumo de energía de un hogar español el pasado domingo. Datos: consumo de todos los hogares de municipios de más de 100.000 habitantes.

   b) Variable: tiempo de ejecución de un programa en un ordenador PIV a 3.06 GHz. Datos: tiempos de ejecutar el programa 150 veces.

   c) Variable: longitud de una pieza metálica producida por la máquina M1. Datos: las longitudes de todas las piezas producidas hasta el momento actual.

   d) Variable: habitantes de un municipio español. Datos: censo de población.

e) Variable: tiempo de atención a un cliente que acude a un puesto de servicio. Datos: tiempo de atención a todos los clientes que acudieron ayer.

2. Un analista desea estudiar el tiempo de acceso de un conjunto de ordenadores a una red local. Para hacer el análisis necesita tomar datos del tiempo que se ha tardado en acceder a dicha red. Los datos los toma accediendo a la red desde el ordenador de su mesa entre las 21:00 y las 21:15 horas de un viernes por la tarde, tomando un total de 25 datos (tiempos de acceso). A continuación, utilizando esos datos elabora un informe sobre dicha red. ¿Qué comentarios te sugiere esa forma de analizar la variable?

.....

Clasificación de los datos según el procedimiento de obtención

Otra forma de clasificar los datos es según el procedimiento de obtención de los mismos. Los datos para el análisis pueden recogerse de muchas maneras, pero fundamentalmente se pueden establecer dos tipos de datos según su forma de obtención: datos observacionales y datos experimentales. En un problema de análisis de datos concreto habrá datos de ambos tipos.

- **Datos observacionales**: Son los datos que se recogen de forma **pasiva**. Simplemente observamos un proceso o a un conjunto de individuos y medimos el valor de la variable. Por ejemplo, podemos estar interesados en el tiempo que tarda en ejecutarse un programa. Tras ejecutar 50 veces dicho programa en un ordenador medimos los respectivos 50 tiempos de ejecución. Esos 50 datos son observacionales. Las principales características de los datos observacionales son:

  1. El investigador no ejerce ningún control sobre dicha variable.
  2. Los valores que toman la variable son conocidos por el analista después de haber seleccionado a los individuos a analizar, es información 'a posteriori'. El analista no puede determinar el valor de la variable. A lo sumo puede hacer una **selección** de aquellos datos que correspondan a determinados individuos con vistas a especializar su estudio.

- **Datos experimentales**: Son aquellos cuyo valor los fija el investigador. Por ejemplo, el analista puede estar interesado en el tiempo de ejecución de un programa en el ordenador A y en el ordenador B. En cada ordenador ejecuta el programa 50 veces. El tiempo de ejecución es un dato observacional, pero el ordenador en que se realiza, A o B, es un dato experimental, pues lo decide el analista.

A la variable cuyo valor es el objetivo del experimento se le suele denominar **variable respuesta**, pues su valor puede interpretarse como la respuesta a la influencia de cierto conjunto de variables. A estas otras variables cuya influencia puede determinar el valor de la variable respuesta se les suele denominar **factores**. Los valores que finalmente tome la variable respuesta serán entonces datos observacionales. **El investigador manipula deliberadamente algunos factores con el fin de poder establecer una relación de causalidad entre los factores y la variable respuesta**. El analista fija el valor que tomarán dichos factores, y por tanto esos valores serán **datos experimentales**.

Ejemplo 1 *Supongamos que un analista quiere saber la influencia de la temperatura de una CPU en su velocidad. Para ello ejecuta una serie de programas y contabiliza su tiempo de ejecución.*

Repite dicha tarea 30 veces, 10 en cada una de las temperaturas T_1, T_2, y T_3. Al final de estos experimentos, el analista tendrá una tabla de datos similar a la siguiente:

Experimento	*Temperatura*	*Tiempo (msg)*
1	T_1	*34*
2	T_1	*23*
⋮	⋮	⋮
11	T_2	*42*
12	T_2	*39*
⋮	⋮	⋮
30	T_3	*54*

*Se tienen así dos variables cuantitativas continuas: temperatura y tiempo. La temperatura es un factor, y sus valores son **datos experimentales**, pues los valores de la temperatura los decide el analista. El tiempo es la variable respuesta, y sus valores son **datos observacionales** que ha tenido que cronometrar el analista de alguna forma. Puede verse en la tabla anterior que aun manteniendo la temperatura constante, hay variabilidad en los tiempos de ejecución. Esa variabilidad indica que habrá más factores aparte de la temperatura que determinen el tiempo de ejecución.*

La principal ventaja de utilizar datos experimentales es que es más fácil establecer relaciones de causalidad entre los factores y la variable respuesta, pues el analista puede observar la evolución de los resultados a medida que va manipulando los factores. La capacidad de poder elegir los valores de las variables que interesen hace que se necesiten menos datos para sacar conclusiones que si se usasen datos observacionales. Hay por tanto una clara ventaja económica en este tipo de datos. La **necesidad de economizar** a la hora de recoger datos es muy importante en ingeniería, pues los costes de manipular procesos industriales o realizar ensayos pueden ser muy elevados.

Otra ventaja de los datos experimenales es que permiten **provocar situaciones** de interés que difícilmente puedan observarse en la realidad o que se precise de demasiado tiempo de observación. A veces, un experimento con factores será la única forma de obtener información, como sucede en el diseño de nuevos productos o modificaciones en el diseño de productos existentes. Otras veces, como ocurre en economía o sociología, la única fuente de información serán datos observacionales.

Por ejemplo, si medimos el tiempo que tarda en ejecutarse 50 veces un programa en un ordenador y lo comparamos con otras 50 ejecuciones del mismo programa en otro ordenador, tendremos que:

- El tiempo de ejecución es la **variable respuesta**, y sus valores serán datos cuantitativos y continuos.
- Los 100 valores de tiempo de ejecución (50 por cada ordenador) serán **datos observacionales**. Además serán una **muestra** de los infinitos valores que se podrían haber obtenido si se hubiese ejecutado ese programa indefinidamente, lo que en la práctica sería imposible.
- El tipo de ordenador es un **factor** que puede afectar a la **variable respuesta** y que tiene dos valores: ordenador A y ordenador B. La variable 'tipo de ordenador' tendrá 50 valores A y otros 50 valores B. Son datos elegidos por el analista y por tanto son **datos experimentales**, además de **cualitativos**.
- Habremos realizado dos experimentos. Un experimento será cronometrar el tiempo de ejecición de un programa en el ordenador A, y el otro experimento será cronometrar el tiempo de ejecución en el ordenador B. Cada experimento se ha repetido 50 veces.

El resultado del análisis estadístico de los datos con las técnicas que se verán en esta asignatura permitirá extraer conclusiones sobre la **población**, es decir, sobre las posibles ejecuciones del programa que puedan realizarse en el futuro. Dichas conclusiones pueden calificarse como un ejercicio de **inducción** o **inferencia** sobre la población general a partir de los resultados particulares de la muestra.

A todo el proceso de obtención de datos entre los que suele haber datos experimentales se le suele denominar coloquialmente experimento, lo que puede provocar confusión con la definición de experimento expuesta anteriormente, pues estaríamos diciendo que un experimento está compuesto por experimentos. En general, el contexto en que se utilicen evitarán la confusión. Cuando exista ambigüedad, se denominará **conjunto de experimentos** al conjunto de todos ellos y **experimento elemental** a cada uno experimentos propiamente dichos.

Ejercicios 3 .....

1. Haz un esquema del contenido de esta sección mediante un diagrama de árbol
2. ¿Puede un dato ser cualitativo y experimental al mismo tiempo? Propón dos ejemplos
3. ¿Puede un dato ser experimental y observacional?
4. ¿Puede un conjunto de datos ser muestrales y cualitativos?
5. Supongamos que tenemos presupuesto para tomar 200 datos y con ellos queremos saber si acudir a una academia de cálculo aumenta las probabilidades de aprobar. Diseña un experimento (conjunto de experimentos elementales) para hacer dicha comparación y clasifica cada tipo de dato según los criterios de esta sección. No olvides especificar a qué población extenderás los resultados.

.....

1.3. Descripción de datos mediante tablas

1.3.1. Tablas de frecuencias univariantes

La primera tarea, antes de entrar en análisis estadísticos más complejos, es tener una **idea general** de cómo son esos datos: distribución, valores medios, medidas de dispersión, detección de valores excesivamente altos o bajos; así como alguna representación gráfica que nos de una idea del conjunto. El objetivo es **resumir** para poder ver cuáles son los patrones más importantes que guíen posteriores análisis. Una primera forma de resumir la información es mediante una tabla que nos diga qué valores diferentes hemos observado en nuestra variable de interés y cuántos datos hay de cada valor. Esta tabla recibe el nombre de **tabla de distribución de frecuencias o tabla de frecuencias**. Una tabla de frecuencias para una variable es una tabla en la que se contabiliza para cada valor distinto observado la frecuencia observada (absoluta o relativa) de individuos que toman dicho valor.

Ejemplo 2 *La tabla que se muestra a continuación corresponde al número de cilindros de una muestra de 155 automóviles (fichero* ***cardata.sf****, contenido dentro de la base de datos de* Statgraphics*). La variable de interés es el número de cilindros, que es cuantitativa discreta. La tabla que se muestra es la que proporciona el programa* Statgraphics.

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	3	1	0,0065	1	0,0065
2	4	104	0,6710	105	0,6774
3	5	3	0,0194	108	0,6968
4	6	30	0,1935	138	0,8903
5	8	17	0,1097	155	1,0000

Tabla de frecuencias del número de cilindros

La tabla puede contener tanto la frecuencia absoluta como la frecuencia relativa, que se calcula simplemente dividiendo la frecuencia absoluta entre el número total de observaciones. De la tabla anterior es fácil obtener conclusiones que sería difícil obtener de la inspección cruda de los 155 datos. Puede verse que la mayoría de los coches tienen 4 cilindros. Puede verse también que los coches con un número impar de cilindros son realmente minoritarios. Por otra parte, el número de coches con seis u ocho cilindros son casi el 30 % de los automóviles, lo que forma una cantidad nada despreciable.

Ejemplo 3 *La tabla siguiente corresponde al mes de nacimiento de un conjunto de 95 estudiantes de primer curso de ingeniería industrial (fichero* ***alumnosindustriales.sf3****). En la tabla puede verse que los meses más frecuentes son, curiosamente, enero y junio.*

Class	Value	Frequency	Relative Frequency
1	Enero	15	0,1579
2	Febrero	5	0,0526
3	Marzo	10	0,1053
4	Abril	9	0,0947
5	Mayo	10	0,1053
6	Junio	13	0,1368
7	Julio	9	0,0947
8	Agosto	7	0,0737
9	Septiembre	6	0,0632
10	Octubre	1	0,0105
11	Noviembre	3	0,0316
12	Diciembre	7	0,0737

Mes de nacimiento de 95 estudiantes de primer curso de Ingeniería Industrial

La tabla anterior, en la que aparecen todos los valores diferentes y su frecuencia, pierde su utilidad si la variable es cuantitativa con muchos valores diferentes. En ese caso la tabla sería muy extensa y las frecuencias de cada valor serían siempre muy pequeñas. La tabla sería poco informativa y perdería su valor como instrumento para resumir los datos. Este problema surge tanto con variables continuas como con variables discretas que tengan muchos valores diferentes. En estos casos, se puede construir una tabla agrupando el rango de valores en intervalos, también llamados *clases*. En general, el rango de valores se divide en intervalos de la misma longitud.

Ejemplo 4 *A continuación se muestra la tabla de la variable precio (en dólares) de los 155 coches utilizados en los ejemplos anteriores (fichero cardata.sf)*

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000

Tabla de frecuencias del precio de los coches

En esta tabla es muy fácil sacar conclusiones. Puede verse que alrededor del 80 % de los coches tienen un precio entre 2000 y 6000 dólares, y que coches más baratos es casi imposible encontrar. El número de clases se ha de decidir en función del tamaño muestral. Demasiadas clases restaría capacidad de síntesis a la tabla, y demasiado pocas ocultaría detalles de los datos que pueden ser importantes. En general, se suele tomar un número de clases alrededor del valor $\sqrt{n}$, donde n es el número de datos.

Ejercicios 4 .....

1. ¿Cuál es la utilidad principal de hacer una tabla de frecuencias?
2. ¿Cómo harías una tabla de frecuencias para la variable: tiempo acceso a un servidor, de la que se tienen 3000 observaciones comprendida en el rango de valores [0,135] segundos?
3. ¿Cómo harías una tabla de frecuencias para la variable: modelo de vehículo de los conductores de Valladolid, cuando se dispone de 100.000 observaciones y un total de 150 modelos diferentes?
4. El fichero **TiempoaccesoWeb.sf3** tiene dos variables que miden el tiempo (en segundos) que tardó un ordenador en conectarse a internet desde que se pulsó el icono correspondiente hasta que se abrió la página de inicio. Las dos variables son **Ordenador_casa** y **Ordenador_uni** que contienen los tiempos de acceso utilizando un ordenador en un domicilio particular y uno de la Carlos III. Cada variable tiene 55 observaciones tomadas de forma consecutiva. Se pide

   a) Construye una tabla de frecuencias de cada variable

   b) ¿Qué porcentaje de las veces se accede desde casa a la Web en menos de 6 segundos?

   c) ¿Y desde la universidad?

.....

1.3.2. Tablas de frecuencias bivariantes. Distribución marginal y condicionada

Si se tiene más de una variable de cada individuo, puede realizarse igualmente la tabla de **frecuencias conjunta**. Por ejemplo, con dos variables se tendrá una tabla de doble entrada

donde cada dimensión corresponderá a una variable y cada celda de la tabla tendrá el número de individuos que tengan los correspondientes valores según la fila y columna en que se encuentre. Dicho valor recibe el nombre de **frecuencia conjunta** (absoluta o relativa)

Ejemplo 5 *La tabla siguiente muestra la tabla de* ***distribución de frecuencias conjunta*** *para las variables: número de cilindros y año de fabricación de los 155 coches del fichero* ***cardata.sf*** *anterior.*

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Tabla de distribución conjunta del número de cilindros (filas) y año de fabricación (columnas)

En esta tabla puede verse, por ejemplo, que hay 25 coches que se han fabricado en el año 81 y que además tienen 4 cilindros. En los márgenes de la tabla aparecen también los totales for filas y por columnas, que son precisamente las frecuencias de cada variable por separado. A esta distribución de frecuencias univariante que aparece en las tablas multivariantes se les denomina **distribuciones marginales**, por estar situados en los márgenes. Las distribuciones marginales son por tanto las distribuciones univariantes. Cada una de las filas o columnas por separado nos mostrará la distribución de frecuencias de una variable cuando la otra variable toma un valor determinado. A esta distribución de frecuencias, que es también univariante, se le denomina **distribución condicionada**. Por ejemplo, los valores de la primera columna de la tabla es la distribución del número de cilindros condicionada a que los coches son del año 78.

Las tablas de frecuencias multivariantes también pueden contener frecuencias relativas. En este caso es importante distinguir si el interés está en la frecuencia relativa conjunta o condicionada. La tabla siguiente muestra la tabla de distribución de frecuencias relativa conjunta, donde en cada celda se encuentra el porcentaje que representan los elementos de dicha celda respecto al total (155 coches).

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
	0,00%	0,00%	0,65%	0,00%	0,00%	0,65%
4	17	12	25	22	28	104
	10,97%	7,74%	16,13%	14,19%	18,06%	67,10%
5	1	1	1	0	0	3
	0,65%	0,65%	0,65%	0,00%	0,00%	1,94%
6	12	6	2	7	3	30
	7,74%	3,87%	1,29%	4,52%	1,94%	19,35%
8	6	10	0	1	0	17
	3,87%	6,45%	0,00%	0,65%	0,00%	10,97%
Column Total	36	29	29	30	31	155
	23,23%	18,71%	18,71%	19,35%	20,00%	100,00%

Tabla de distribución de frecuencias relativa bivariante.

Por ejemplo, los 25 coches que se fabricaron en el 80 y además tienen 4 cilindros suponen el $100\times(25/155)\,\% = 16{,}13\,\%$ del total de los coches. Si lo que se desea es la distribución de frecuencias relativas condicionada a la variable cilindros, los porcentajes se obtendrán dividiendo las frecuencias absolutas entre el total de la fila. Se tendrá la siguiente tabla:

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
	0,00%	0,00%	100,00%	0,00%	0,00%	100%
4	17	12	25	22	28	104
	16,35%	11,54%	24,04%	21,15%	26,92%	100%
5	1	1	1	0	0	3
	33,33%	33,33%	33,33%	0,00%	0,00%	100%
6	12	6	2	7	3	30
	40,00%	20,00%	6,67%	23,33%	10,00%	100%
8	6	10	0	1	0	17
	35,29%	58,82%	0,00%	5,88%	0,00%	100%
Column Total	36	29	29	30	31	155

Tabla de distribución de frecuencias relativas condicionada al número de cilindros

En esta tabla puede verse, por ejemplo, que los 25 coches fabricados en el año 80 y que tienen 4 cilindros son el $100\times(25/104)\,\% = 24{,}04\,\%$ de los coches de 4 cilindros.

Ejercicios 5 .....

1. Un proceso productivo tiene dos líneas de producción: línea A y línea B que funcionan de forma totalmente independiente (distintas máquinas, distintos operarios, etc). Un analista toma nota al final de cada línea del número de defectos que tienen los artículos que van produciendo ambas líneas. Primero toma nota de 50 artículos de la línea A y después va a la línea B y toma nota de 50 artículos de dicha línea. La tabla siguiente muestra un ejemplo

de cómo son los datos:

	Número de defectos	
Artículo	Línea A	Línea B
1	0	1
2	1	1
3	2	0
⋮	⋮	⋮
50	1	0

Cuando tiene 50 artículos inspeccionados por cada línea construye una tabla de frecuencias bivariante, resultando la siguiente tabla:

Número de defectos en los artículos de la Línea B

Número de defectos en los artículos de la Línea A	0	1	2	3	Row Total
0	10	4	1	2	17
	20,00%	8,00%	2,00%	4,00%	34,00%
1	6	4	2	1	13
	12,00%	8,00%	4,00%	2,00%	26,00%
2	4	2	1	2	9
	8,00%	4,00%	2,00%	4,00%	18,00%
3	4	3	4	0	11
	8,00%	6,00%	8,00%	0,00%	22,00%
Column Total	24	13	8	5	50
	48,00%	26,00%	16,00%	10,00%	100,00%

¿Qué opinión te merece la realización de esta tabla?

2. En un aula con 25 chicos y 14 chicas se pregunta quién fuma, resultando la siguiente tabla

	Fuma	No fuma	total
Chico	12	13	25
Chica	8	6	14
total	20	19	39

   a) ¿Qué proporción de estudiantes fuma?

   b) ¿Qué propoción de chicas no fuma?

   c) ¿Qué proporción de estudiantes son chicos y fumadores?¿Qué son el resto de los alumnos?

   d) ¿Cuál es la distribución marginal de frecuencias relativas del sexo de los alumnos?

   e) ¿Cuál es la distribución marginal de frecuencias absolutas del sexo de los alumnos?

   f) ¿Cuál es la distribución de frecuencias relativas de la variable sexo condicionada a que son alumnos fumadores?

. . . . . .

1.4. Descripción de datos mediante gráficos

Sin duda, la forma más rápida y eficaz de resumir la información de un conjunto de datos es mediante un gráfico. **Al realizar cualquier análisis estadístico, primero hay que visualizar los datos mediante gráficos, y posteriormente realizar una descripción basada**

en valores numéricos. Si en esta sección hemos presentado las tablas de frecuencias antes de las representaciones gráficas es sólo por motivos pedagógicos. Hay muchísmoas formas de representar gráficamente la información contenida en un conjunto de datos. Aquí mostraremos sólo las representaciones gráficas más habituales.

1.4.1. Diagrama de barras

Un diagrama de barras es la representación gráfica de una tabla de frecuencias donde los datos están sin agrupar. Consiste simplemente en dibujar un rectángulo por cada valor de la variable, de altura proporcional a la frecuencia. Es útil para variables cualitativas o cuantitativas discretas, pero con pocos valores diferentes. La Figura **??** muestra el gráfico de barras del número de cilindros de los 155 coches. Este gráfico dibuja los valores de la tabla que se mostró más arriba.

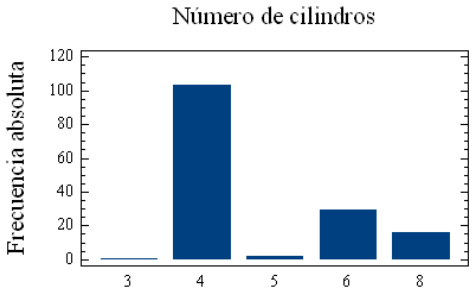
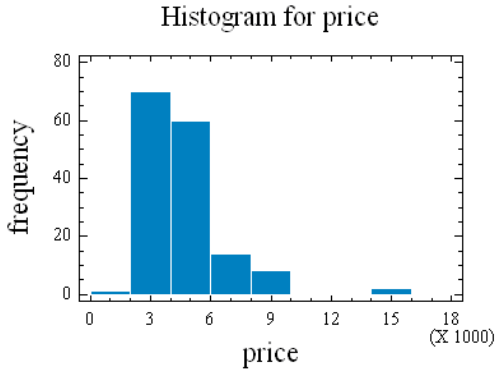


Diagrama de barras del número de cilindros de 155 coches

1.4.2. Histograma y polígono de frecuencias

El histograma es una de las herramientas gráficas más útiles para resumir un conjunto de datos de una variable. Un histograma es la representación gráfica de una tabla de frecuencias donde los datos han sido agrupados por intervalos. Por ejemplo, el histograma correspondiente a la tabla del precio (*price*) de los automóviles del fichero *cardata* mostrada anteriormente es



Precios de 155 coches (cardata.asf)

donde cada rectángulo corresponde con una clase, y la altura es proporcional a la frecuencia relativa de dicha clase.

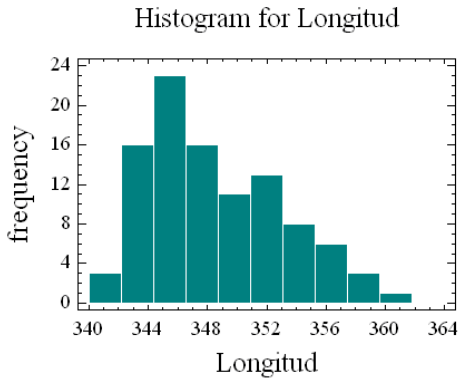
En un histograma hay que fijarse en las tendencias generales de los datos, como son:

- **Concentraciones**: por concentraciones nos referimos a aquellos rectángulos en los que hay mayor proporción, aldededor de los cuales se agrupan otros rectángulos de frecuencia decreciente. En el caso del histograma de precios existe una concentación en torno al valor 3000 $. Si existe más de una concentración será indicio de que los datos son heterogéneos y que podrían proceder de más de una población diferente.
- **Huecos**: que sería indicio aún mayor de que los datos proceden de poblaciones diferentes.
- **Valores atípicos**, por ser demasiado altos o bajos. Un dato será atípico si se separa mucho del patrón general que siguen los datos.
- **Asimetrías**: que indican hacia dónde tienden a desplazarse los datos cuando nos alejamos de las zonas de concentración. Cuando la asimetría es tal que la cola de la distribución de los datos apunta hacia la derecha, hacia $+\infty$, diremos que hay **asimetría positiva**. Este es el caso del historama de precios. Cuando la cola de la distribución apunta hacia $-\infty$ la asimetría se denomina **negativa**.

Es aconsejable hacer varios histogramas variando ligeramente el número de clases, de esta forma estaremos más seguros de que las características que observamos (asimetrías, concentraciones) no se deben a una agrupación casual de los datos.

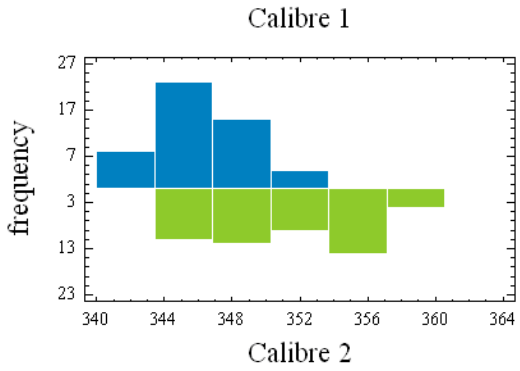
En el caso de nuestro histograma de precios podemos decir que se trata de una distribución con una sola concentración en torno a los 3000 $. Hay una asimetría positiva encontrándose un par de observaciones demasiado altas en comparación con el resto de los datos. Estos datos alejados no deben calificarse de atípicos pues al haber una asimetría hacia valores altos es razonable que haya algún valor más distante en dicha dirección. **Esos datos alejados no están fuera del patrón general de la distribución.**

Ejemplo 6 *El gráfico siguiente muestra el histograma de las longitudes de 100 clavos del mismo tipo (de una caja de clavos de un mismo tipo comprada en una ferretería). Los datos se encuentran en el fichero* ***longitudclavos.sf3****). Los clavos fueron medidos por dos personas que usaron instrumentos de medida diferentes (cada persona usaba su propio calibre).*



Histograma de las longitudes de los clavos medidas por cada calibre

Este histograma muestra que hay dos concentraciones, que sugieren que los clavos pueden proceder de dos poblaciones distintas. Sería interesante comprobar si las dos concentraciones coinciden con los clavos medidos por cada uno de los calibres. La figura siguiente, en la que se muestra el histograma de las mediciones realizadas con cada calibre, muestra que en efecto es el calibre utilizado el que produce esas dos concentraciones y que es muy probable que los clavos sean realmente homogéneos. El calibre 2 parece que proporciona mediciones con más dispersión y de valores en promedio superiores al calibre 1. Podría ***inferirse*** *que el Calibre 2 es de peor calidad que el Calibre 1.*



Mediciones de 50 clavos similares usando el calibre 1 y otros 50 usando el calibre 2

Ejemplo 7 *La Figura 1.2 muestra el histograma de los valores de velocidad de viento (m/s) registrados en un parque eólico durante varios meses (fichero* ***parqueeolico1.sf3****). Cada dato es la velocidad media registrada durante una hora, y se tienen 14000 datos. En el histograma (a) de esta figura se muestra un histograma con sólo 12 clases, donde puede apreciarse una concentración alrededor der los 7 m/s, y una asimetría positiva. Puede apreciarse una segunda concentración en torno a los 3 m/s. Para ver en más detalle esta posible segunda concentración, y aprovechando que se tienen muchos datos de esta variable, podemos ver el histograma (b) que está realizado con* $\sqrt{14000} \approx 118$ *clases. En (b) puede verse realmente que hay una primera concentración en torno a los 2 m/s y una segunda en torno a los 7 m/s. Vemos también que la segunda concentración no es muy acusada, siendo muy frecuente registrar valores más altos que 7 m/s.*

Ejemplo 8 *El fichero* ***defragmenta.sf3*** *contiene datos del tiempo que un ordenador tarda en escribir un fichero de 300 Mb en su disco duro. Se hacen dos experimentos; uno en el que el disco duro está desfragmentado, y otro en el que el disco duro tiene una fragmentación del 40 %. Cada*

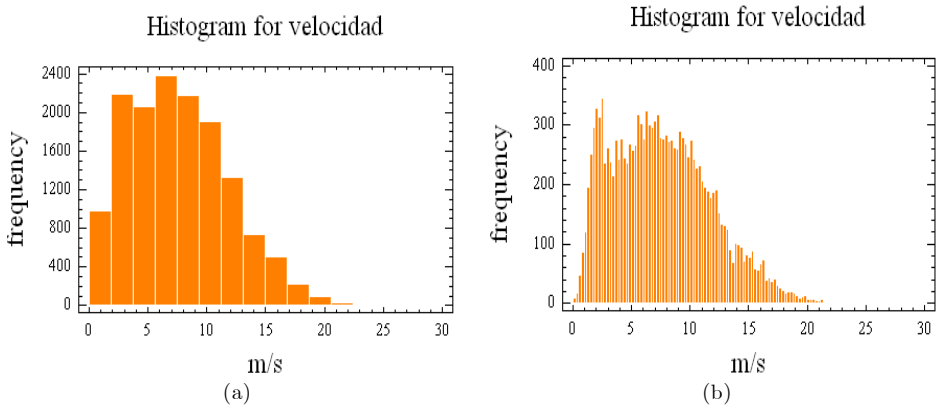
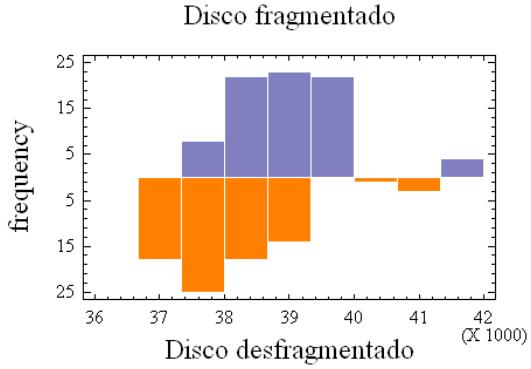


Figura 1.2: Distribución de las velocidades del viento registradas en un parque eólico

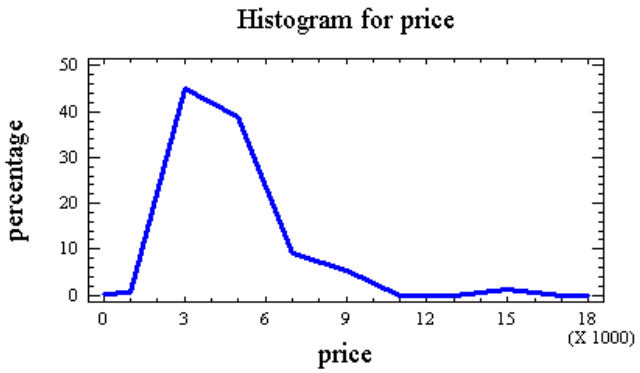
experimento se repite 79 veces. La figura siguiente muestra los dos histogramas.



Tiempo que se tarda en escribir un fichero de 300 Mb en un disco duro fragmentado o desfragmentado.

Puede verse que el tiempo tiende a ser mayor cuando el disco está fragmentado, aunque las diferencias no parecen muy grandes. Parece que hay otros factores importantes que afectan a la velocidad de escritura aparte del estado de fragmentación del disco duro. Ambas distribuciones tienen una zona de mayor concentración y no son muy asimétricas. En ambas distibuciones puede observarse un reducido grupo de datos alejados del resto, y que pueden considerarse atípicos. Esos atípicos aparecen tanto con el disco desfragmentado como con el disco sin desfragmentar. Por tanto no parecen que sean debidos a causas que no se vayan a volver a repetir.

El **polígono de frecuencias** consiste en una línea poligonal que resulta de unir los puntos centrales de la parte superior de los histogramas. La siguiente figura muestra el polígono de frecuencias del histograma de precios de los coches



Polígono de frecuencias del precio de 155 vehículos

Dependiendo de cada conjunto de datos, el polígono de frecuencias nos puede ayudar mejor que el histograma a hacernos una idea de cómo son los datos, sobre todo si tenemos un tamaño muestral grande.

1.4.3. Diagrama de tartas

Este diagrama se utiliza cuando hay pocos valores diferentes de la variable. Consta de un círculo dividido en porciones. Las diferentes porciones en las que se divide el círculo son proporcionales a la frecuencia de cada valor. A continuación se muestra el diagrama de tarta para los datos del número de cilindros de una muestra de 155 automóviles.

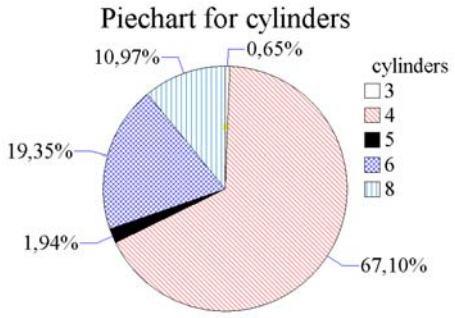
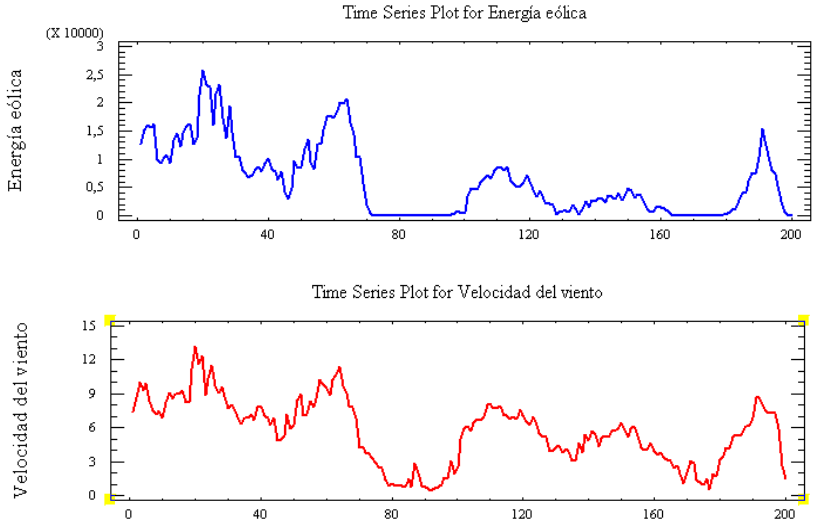


Diagrama de tarta del número de cilindros de 155 coches

1.4.4. Series temporales

Es el dibujo de la evolución temporal de una variable. El eje de abcisas es el tiempo y el eje de ordenadas es el valor observado en cada instante. El gráfico temporal permite visualizar la tendencia de una variable. A continuación se muestra el gráfico de la evolución de la energía eólica de los datos del Ejemplo 7 durante 200 horas, así como la evolución de las velocidades registradas en esas horas. En estos dos gráficos puede verse cómo tanto la energía eólica como la velocidad del ciento siguen evoluciones paralelas. Puede verse también cómo por debajo de cierto valor de velocidad, la generación de energía ha sido nula. Esto ha ocurrido en el periodos comprendido

entre las horas 70 a 100 y también entre las horas 160 y 180.



1.4.5. Gráficos de dispersión

Este gráfico, también conocido como gráfico bivariante o gráfico XY representa la información de dos variables para un conjunto de individuos. Para cada individuo se tienen dos datos: la información de la variable x y la de la variable y. En unos ejes de ordenadas se representa cada punto colocando en el eje x el valor de la primera variable y en el y el de la segunda. Se tendrán tantos puntos como tamaño de la muestra.

Ejemplo 9 *El siguiente gráfico muestra la información de una muestra de 40 vehículos. Para cada vehículo se tiene la potencia del motor (eje x) y la velocidad máxima (eje y). El gráfico de estos 40 puntos, uno por vehículo, revela que a mayor potencia mayor velocidad máxima, siendo esta relación lineal (la nube de puntos se extiene a lo largo de una línea recta imaginaria)*

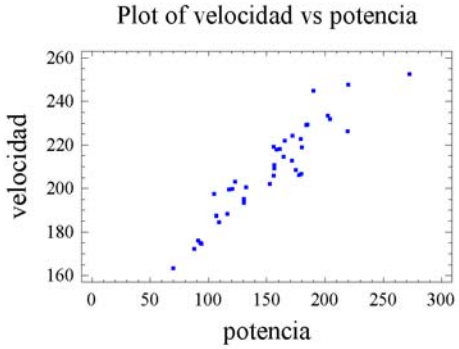
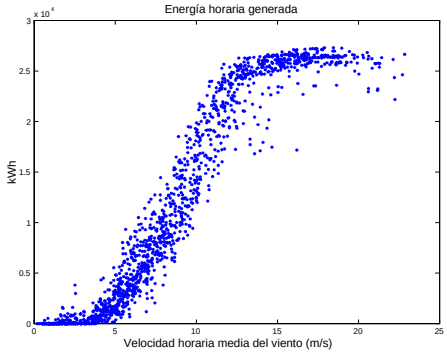


Gráfico de dispersión para la velocidad y la potencia de un conjunto de coches

Ejemplo 10 *La siguiente figura muestra el gráfico de dispersión de la energía generada en un parque eólico y la velocidad del viento (fichero* ***parqueeolico1.sf3****). Cada punto tiene dos valores:*

la energía producida en una hora y la velocidad registrada del viento en dicha hora. El eje X representa la velocidad (m/s) media registrada en una hora, mientras que el eje Y representa la energía (kwh) generada en ese tiempo. Puede verse en la figura que hay una clara relación entre la velocidad del viento y la energía, pero en este caso dicha relación es no lineal. A velocidades bajas, la producción energética es casi nula. A partir de cierta velocidad mínima, la energía aumenta de forma creciente. A partir de cierta velocidad máxima, la energía generada en una hora apenas varía y se mantiene próxima a cierto nivel máximo, que es la capacidad máxima de la instalacción.



Potencia horaria media generada en un parque eólico en función de la velocidad del viento

1.5. Medidas características de un conjunto de datos

En esta sección continuamos con nuestro propósito de buscar formas de **resumir la información de un conjunto de datos**, con el objetivo de poder ver sus características más relevantes. Ahora bucamos medidas que mediante un solo número se resuma alguna característica importante de los datos, a las que llamaremos medidas características. Resumir toda la información de nuestra muestra en un solo número puede ser muy arriesgado, por lo que el uso de estas medidas características **ha de hacerse siempre acompañando a herramientas gráficas**, como las expuestas en la sección anterior. Dividiremos las medidas características en medidas de centralización y medidas de dispersión.

1.5.1. Medidas de centralización

Las medidas de centralización que se exponen en esta subsección tienen como objetivo indicar el valor alrededor del cual tienden a situarse los datos. Estas medidas nos darán una idea de la magnitud de la variable. Alrededor de este valor indicado por la medida de centralización, los datos pueden estar dispersos de multiples formas. Sobre medidas características relacionadas con la dispersión tratará la siguiente subsección. Existen muchas medidas de centralización. Veremos las más importantes.

Media aritmética

En general, usaremos letras mayúsculas para denotar a las variables, y letras minúsculas para hacer referencia a valores observados genéricos de la variable (es decir, para los datos). Sea X la

variable de interés y sean $x_1, x_2, ..., x_n$ las n observaciones que se poseen de dicha variable. La media aritmética, o simplemente media, es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{1.1}$$

En el caso en que la variable sea discreta y existan valores repetidos, la media puede calcularse sumando cada valor observado por su frecuencia relativa de aparición. Si denotamos por $x_1, x_2, ..., x_J$ a cada uno de los valores distintos de la variable X y $n_1, n_2, ..., n_J$ el número de veces que hemos observado cada uno de estos valores tendremos que la frecuencia relativa de aparición de cada uno es $f_r(x_j) = n_j/n$, $j = 1, ..., J$, con $n = \sum n_j$. Entonces la media puede calcularse como

$$\bar{x} = \sum_{j=1}^{J} x_j f_r(x_j). \tag{1.2}$$

Ejemplo 11 *Utilizando los datos del número de cilindros de los 55 coches del fichero cardata.sf, mostrados en el Ejemplo 2 podemos calcular el número medio de cilindros de un coche. Como hay muchos coches con el mismo número de cilindros, lo más cómodo es utilizar la expresión (1.2) a los datos de la tabla incluida en el Ejemplo 2. Se tiene entonces que*

$$\begin{aligned} \bar{x} &= 3 \times 0{,}0065 + 4 \times 0{,}6710 + 5 \times 0{,}0194 + 6 \times 0{,}1935 + 8 \times 0{,}1097 \\ &= 4{,}84 \ \textit{cilindros} \end{aligned}$$

Ejercicios 6

1. Sea el siguiente conjunto de datos de la variable x, $x = \{2, 2, 2, 4, 4, 4, 5, 5\}$. Calcula la media aritmética usando tanto la expesión (1.1) como (1.2) y comprueba que se obtiene el mismo resultado.

2. Se tienen n artículos manufacturados, de los cuales d son defectuosos y $n - d$ son aceptables. Asignamos a cada artículo una variable x que toma valor 1 si el artículo es aceptable y 0 si es defectuoso. Demuestra que $\bar{x}$ es la proporción de artículos aceptables.

3. Un ordenador recibe datos de una variable de forma secuencial. Cada vez que recibe un dato, calcula la media aritmética de todos los datos recibidos hasta ese momento. Esta forma de operar puede ser poco eficiente cuando el número de observaciones sea ya muy elevado, pues tendrá que almacenar todos los datos y luego sumarlos. Con el fin de simplificar el cálculo se propone calcular la media haciendo el mínimo número de operaciones posible y con unos requerimientos de memoria mínimos. Llamemos $\bar{x}_n$ a la media de n observaciones y $\bar{x}_{n-1}$ a la media con sólo $n-1$ observaciones. Demuestra que $\bar{x}$ se puede calcular de forma recursiva como

$$\bar{x}_n = \bar{x}_{n-1} + \frac{1}{n}\left(x_n - \bar{x}_{n-1}\right), \tag{1.3}$$

y por tanto no hace falta almacenar todos los datos, sino sólo $\bar{x}_{n-1}$ y n. El número de operaciones es también mucho menor que un cálculo que no sea recursivo.

La media puede interpretarse como el centro de gravedad de los datos. Si imaginamos que un histograma fuese un objeto con masa, la media artitmética estará localizada en aquel punto del eje X que dejase al histograma en equilibrio. La Figura 1.3 ilustra esta idea de media como centro de gravedad de una distribución.

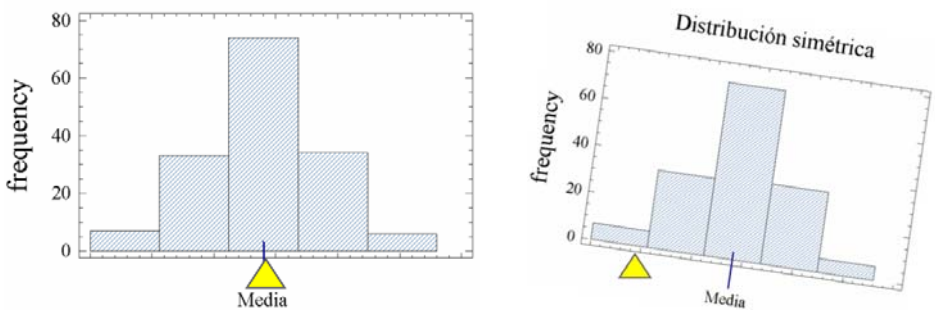


Figura 1.3: Localización de la media aritmética como centro de gravedad de la distribución

Esta interpretación sirve para entender mejor el comportamiento de la media. Por ejemplo, si la media es el centro de gravedad de los datos, es fácil entender que en un conjunto de datos que tengan un histograma simétrico con los datos concentrados alrededor de un eje de simetría, la media se encuentre justo en el centro de dicho histograma, bajo el intervalo de mayor frecuencia (ver Figura 1.4). Por el contrario, si la distribución es asimétrica, el centro de gravedad se verá desplazado respecto a la clase más frecuente en la dirección donde esté la cola de dicha distribución, tal y como ilustra la Figura 1.4.

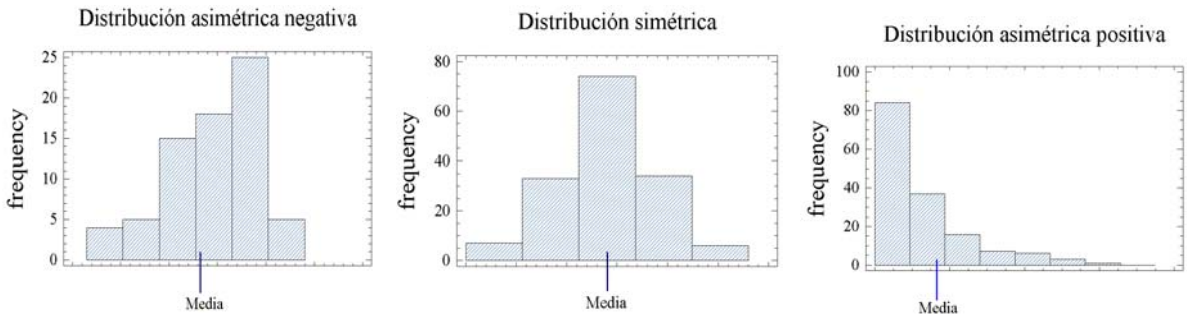


Figura 1.4: Localización de la media aritmética según la asimetría de la distribución

Un problema que presenta la media, y que está relacionada con la propiedad anterior, es su **sensibilidad a observaciones que tengan un valor elevado** (en valor absoluto), pues en ese caso la media se desplazará enormemente hacia dicho valor, perdiéndose su capacidad de servir como medida resumen del orden de magnitud de los datos. **Esta sensibilidad ante variables atípicas la tienen todas las medidas características basadas en sumas, pues un término muy grande en relación a los demás dominará la suma total**. En presencia de datos atípcos, por tanto, la media aritmética es una medida de centralización poco interesante. Debido a este problema, es necesario el uso de otras medidas de centralización que no utilicen sumas.

Mediana

La mediana es una medida de centralización que no está basada en sumas sino en el orden de las variables. La mediana de un conjunto de datos es el valor x_m tal que hay el mismo número de valores por encima y por debajo de él. Para calcular la mediana se ordenarán los datos de menor a mayor. Si el tamaño de la muestra es par, la mediana será el dato que ocupe la posición central. Si el número de datos es impar, la mediana será la media de los dos valores que están en en el centro. Veamos un ejemplo.

Ejemplo 12 *Sea X la variable de la que se observan los valores, ordenados de menor a mayor,*

$$\{3,\ 5,\ 6,\ 9,\ 12,\ 24,\ 27,\ 31,\ 33,\ 34\}. \tag{1.4}$$

Se tienen 10 observaciones. Al ser un número par de valores la mediana será:

$$x_m = \frac{12+24}{2} = 18,$$

mientras que la media es

$$\bar{x} = \frac{3+5+6+\cdots+34}{10} = 18{,}4$$

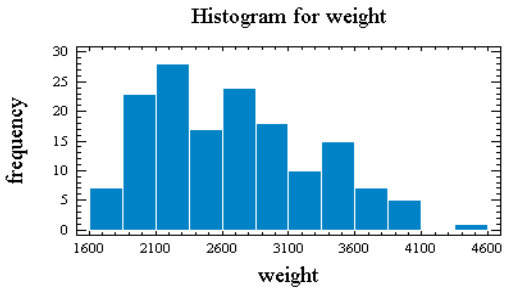
La mediana, al contrario que la media, no es sensible a que alguna observación tenga un valor atípico. Por ejemplo, si el valor más alto fuese 340 en lugar de 34 se tendría que la mediana seguiría manteniendo su valor, sin embargo la media aumentaría a

$$\bar{x} = \frac{3+5+6+\cdots+340}{10} = 49.$$

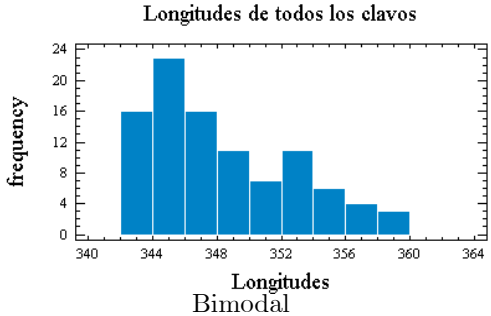
En estadística, a la propiedad de **ser insensible ante observaciones atípicas se le denomina robustez**. Se dice entonces que la mediana es una medida **robusta**. De la misma manera que la mediana es menos sensible que la media a valores atípicos, también será menos sensible que la media a asimetrías. Por esa razón, en distribuciones asimétricas, la media está más desplazada hacia la cola de la distribución que la mediana.

Moda

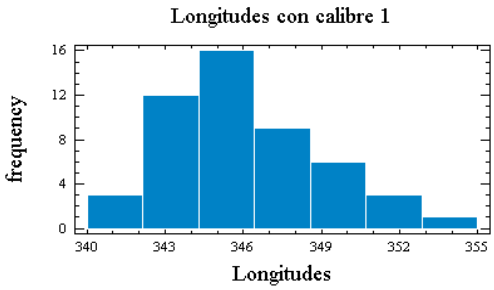
La moda es el valor más frecuente. Por ejemplo, con la variable número de cilindros de los vehículos la moda son 4 cilindros. En el caso de variables continuas o discretas pero con muchos valores diferentes, la moda es de escasa utilidad, pues es posible que no exista ningún valor repetido. En estos casos se toma como valor para la moda, o intervalo modal, **el intervalo de mayor frecuencia** (o también el del punto medio de dicho intervalo. Por ejemplo, en el caso de la variable *precio del vehículo* mostrada anteriormente, el intervalo modal es el de los 2000 a 4000 dólares, cuyo punto medio es 3000. También es frecuente usar el término moda a cualquiera de los máximos relativos de un histograma. Si el histograma presenta una sola concentración diremos que se trata de una distribución unimodal, como en el caso de los precios de los vehículos. Si el histograma presenta dos concentraciones, como sucede en el histograma de las longitudes de los clavos, se dice que la distribución es bimodal, con una moda en torno a los 346 mmm y otra en torno a 352 mm. la Figura 1.5 muestra varios histogramas con diferente número de intervalos modales.



Trimodal



Bimodal



Unimodal

Figura 1.5: Ejemplo de distribuciones de datos continuos con uno o más intervalos modales.

1.5.2. Medidas de dispersión

En esta sección se introducen medidas que resumen la dispersión de los datos alrededor de las medidas características

Varianza

Sea X la variable de interés, de la que se tienen las observaciones $x_1, x_2, ..., x_n$. La varianza s_x^2 es

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

que mide la dispersión alrededor del centro de gravedad $\bar{x}$ mediante el promedio de las desviaciones cuadráticas. El motivo de usar desviaciones cuadráticas está en que lo que nos interesa es la dispersión de cada dato alrededor de la media, independientemente de que sea por exceso o por

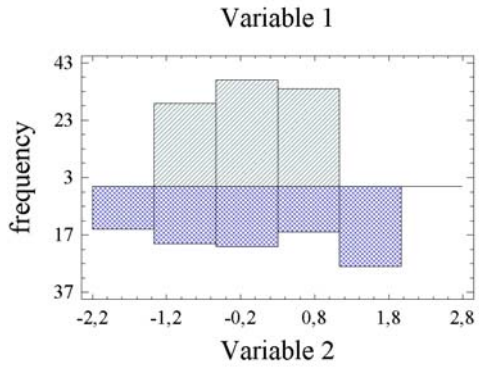
defecto. El signo de $x_i - \bar{x}$ no es relevante. Otra opción sería trabajar con desviaciones absolutas. A esta medida de dispersión se le suele denominad MAD (mean absolute deviations) y es

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

El trabajar con valores absolutos es, en general, más complicado que con funciones cuadráticas, por lo que el uso de la varianza está más extendido. Las unidades de la varianza es la de la variable x al cuadrado. Si queremos una medida de dispersión con las mismas unidades que x y que no esté basada en valores absolutos, usaremos la **desviación típica**, s_x, que no es más que la raíz cuadrada de la varianza, es decir,

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

La figura siguente muestra los histogramas de dos conjuntos de datos con misma media (cero) pero distinta varianza. La variable 1 tiene varianza 0.6, mientras que la variable 2 tiene varianza 1.2 y por eso está más dispersa. Al estar más dispersa, los rectángulos del histograma, que expresan la frecuencia relativa, serán más bajos.



Ejemplo 13 *En el Ejemplo 6 se presentaron los datos del fichero longitudclavos.sf, en el que se presentaban las mediciones de 100 clavos del mismo tipo; 50 medidos con el Calibre 1 y los otros 50 con el Calibre 2. En el Ejemplo 6 se mostraron los histogramas de ambos calibres, en los que se apreciaba que el Calibre 2 tenía una mayor dispersión en sus mediciones, lo que le hace menos preciso. La variabilidad en las mediciones de un calibre es un claro indicador de su calidad. Esa variabilidad se puede medir con la varianza. La varianza de los datos de ambos calibres es:*

$$\begin{aligned} \textit{Varianza Calibre 1=7.25 mm}^2 \\ \textit{Varianza Calibre 2=21.47 mm}^2 \end{aligned}$$

Como no hay ninguna razón a priori para que las longitudes de los 50 clavos medidos con el Calibre 2 tengan más varianza que los medidos con el Calibre 1, concluiremos que esa varianza es debido al error de medición que comete ese calibre. Se concluye así que el Calibre 1 parece de mejor calidad que el Calibre 2.

Una medida alternativa de la varianza es la llamada **cuasivarianza** $\hat{s}^2$ que es igual que la varianza solo que se divide por $n-1$ en lugar de por n. Es decir, la cuasivarianza tiene la siguiente expresión

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

Analogamente, la **cuasidesviación típica** $\hat{s}$ es

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

La razón matemática por la que se usa con mucha frecuencia la cuasivarianza en lugar de la varianza se verá en los próximos temas. La razón intuitiva es que los n términos del numerador $(x_i - \bar{x})^2$ no son independientes, pues si conocemos la media y $n-1$ datos, seremos capaces de deducir el dato n-ésimo. Por tanto, en el numerador sólo hay $n-1$ términos informativos.

El uso de la cuasivarianza es tan popular que muchos programas estadísticos (entre ellos el *Statgraphics*) calculan la cuasivarianza y la cuasidesviación típica en lugar de la varianza y la desviación típica respectivamente, y además lo denominan varianza y desviación típica.

La varianza y la desviación típica dependen de las unidades de la variable X. Para comparar dispersiones de variables que tengan unidades distintas podemos emplear el llamado **Coeficiente de Variación**, que se define como

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

y es adimensional (y sólo definida si $\bar{x} \neq 0$). En general se multiplica por 100 y así expresa un porcentaje. Por ejemplo, si queremos comparar la dispersión de las variables potencia y velocidad del conjunto de coches que se representó gráficamente más arriba tenemos:

$$\begin{aligned} \text{CV de potencia} &= 28.4\,\% \\ \text{CV de velocidad} &= 10.2\,\% \end{aligned}$$

por lo que concluimos que los datos de potencia están más dispersos alrededor de su media que los de la velocidad.

Meda

La varianza, así como la desviación típica, presenta el mismo problema que la media $\bar{x}$ en cuanto a sensibilidad ante observaciones anómalas. La sensibilidad le viene por tres razones fundamentales:

1. Depende de la media muestral, que es sensible a valores atípicos
2. Contiene desviaciones al cuadrado, que hace que una desviación $(x_i - \bar{x})$ muy grande se vea muy amplificada
3. Es, a su vez, un promedio, por lo que se verá dominado por los términos más grandes

Por esta razón se construye la Meda, que es una medida de dispersión robusta. la Meda es la mediana de las desviaciones respecto a la mediana en valor absoluto. Sea x_m la mediana de un conjunto de datos. Entonces

$$\text{Meda=Mediana}\left\{|x_1 - x_m|, ..., |x_n - x_m|\right\}.$$

Los pasos para calcular la Meda son entonces los siguientes

1. Calculamos la mediana de los datos, y la denominaremos x_m
2. Calculamos, para cada dato x_i, $i = 1, 2, ..., n$, la diferencia respecto a esta mediana en valor absoluto, es decir, $|x_1 - x_m|, ..., |x_n - x_m|$.
3. Calculamos la mediana de estas desviaciones absolutas. Ese valor es entonces la Meda

Como puede verse, la Meda no tiene ninguno de los tres elementos anteriores que hacían a la varianza sensible a atípicos.

Rango

También llamado recorrido. El rango es la amplitud de los datos, es decir

$$\text{Rango} = \text{máx}(x_1, x_2, ..., x_n) - \text{mín}(x_1, x_2, ..., x_n).$$

Por ejemplo, en el conjunto de datos (3,4,5,6,7,8) el rango es 8-3=5, y en el conjunto de datos (-9,2,16,-2,5,7,6,-12,23) el rango es 23-(-12)=35.

Cuartiles

Los cuartiles son los valores que dividen un conjunto de datos ordenados de menor a mayor en cuatro partes con igual (aproximadamente) número de datos. Los cuantiles son tres: Q_1, Q_2 y Q_3; también llamados primer, segundo y tercer cuartil respectivamente. El más popular es Q_2, que divide la muestra en dos partes de igual número de datos, y que es la **mediana**. Q_1 será el valor tal que el 25 % de los datos (ordenados de menor a mayor) son menores o iguales que él, y el 75 % de los datos son mayores que él. El cuantil Q_3 es el valor tal que el 75 % de los datos son menores o iguales que él y el 25 % de los datos es mayor que él. A la diferencia $Q_3 - Q_1$ se le denomina **rango intercuartílico**, y contiene al 50 % de los datos que están más cerca (en posición) de la mediana. En muchas ocasiones, especialmene con pocas observaciones, la elección de Q_1 y Q_3 es ambigua y **existen métodos distintos de cálculo que pueden llevar a valores diferentes**. En cualquier caso, la idea básica de todos los métodos es la misma y los resultados muy parecidos. Vamos a ver un método de cálculo y lo aplicaremos al conjunto de datos $x = \{1, 3, 5, 7, 9, 10\}$.

El método más sencillo para calcular los cuartiles consiste en calcular primero Q_2, y después calcular Q_1 y Q_3 como las medianas de la mitad de los datos que queda a cada lado de Q_2 (excluyendo dicho valor). Si aplicamos este criterio a los datos de arriba tenemos que $Q_2 = (5 + 7)/2 = 6$. Este valor produce una división de los datos en dos grupos. A la izquierda de Q_2 quedan {1,3,5}, siendo la mediana de este grupo de datos el número 3. Por tanto $Q_1 = 3$. Análogamente, a la derecha de Q_2 quedan los datos {7,9,10}, siendo la mediana el número 9.Por tanto, $Q_3 = 9$.

Veamos ahora el caso del cálculo de cuartiles con un número impar de datos, y utilizando el mismo método que en el párrafo anterior. Supongamos el conjunto de datos ordenados $x = \{1, 1, 3, 3, 5, 9, 11, 14, 15\}$. La mediana será el dato en la posición central, es decir, $Q_2 = 5$. A la izquierda de Q_2 quedan los valores {1,1,3,3}, siendo su mediana (1+3)/2=2, y por tanto $Q_1 = 2$. A la derecha de Q_2 quedan los datos {9,11,14,15}, y al ser su mediana (11+14)/2=12.5, tenemos que $Q_3 = 12{,}5$.

Otro método muy popular para calcular cuartiles, que no necesariamente conduce al mismo resultado que el anterior, es el siguiente: comenzamos localizando el segundo cuartil, que es el dato que está en la posición m

$$m = \text{Posición de } Q_2 = \frac{1}{2}(n + 1).$$

A esta posición le llamaremos **profundidad de la mediana**. Si la profundidad es un número entero, esa será la posición en la que se encuentre la mediana. Si no es un número entero, entonces acabará en .5. Entonces la mediana será la media de los números que se encuentren en las posiciones más cercanas, anterior y posterior, a dicho valor. Por ejemplo, si la profundidad es 5.5, la mediana será la media de los valores situados en posición 5ª y 6ª, lo que coincide de nuevo con la definición tradicional. En nuestro ejemplo con $x = \{1, 3, 5, 7, 9, 10\}$ se tiene que

$$m = \frac{1}{2}(6 + 1) = 3{,}5$$

y la mediana será Q_2 =(5+7)/2=6. A partir de esta definición de profundidad de la mediana definiremos la **profundidad del cuartil** que será q siguiente

$$q = \frac{1}{2}([m] + 1),$$

donde, como se definió antes $[m]$ es la parte entera de la profundidad de la mediana. El primer cuartil, Q_1 será el dato en posición q (o la media de los datos adyaccentes si q no es entero), mientras que el tercer cuartil, Q_3, será el dato en posición $(n + 1) - q$ (o la media de los adyaccentes si q no es entero). En nuestro ejemplo con $x = \{1, 3, 5, 7, 9, 10\}$ se tiene que $[m] = 3$ y por tanto

$$q = \frac{1}{2}(3 + 1) = 2,$$

Los cuartiles serán entonces

$$\begin{aligned} Q_1 &= 3, \\ Q_3 &= 9. \end{aligned}$$

A la distancia entre Q_1 y Q_3 se le llama **rango intercuartílico** (RI):

$$\mathrm{RI} = Q_3 - Q_1.$$

El rango intercuartílico es el rango de valores entre los que se encuentra aproximadamente el 50 % de los datos en posiciones centrales.

Percentil

Una medida similar al cuartil es el denominado **percentil p**. El **percentil p** de un conjunto de datos es el valor tal que el **p** por ciento de los datos es menor o igual que él, y el (100-p) % de los datos es superior a él. Esta definición puede ser ambigua, pues puede que en nuestro conjunto de datos no haya ningún valor que deje exactamente una proporción **p** por debajo de él. En esos casos se podrían proponer diversas formas de interpolar entre los valores más próximos al percentil exacto, o simplemente tomar la obervación más próxima al percentil. La forma más habitual de calcular percentiles es la siguiente:

1. Se ordenan los N datos de menor a mayor
2. El dato observado en posición i-ésima será el percentil

$$\text{percentil dato i-ésimo} = 100\left(\frac{i - 0{,}5}{N}\right) \tag{1.5}$$

3. Cualquier otro percentil se calcula interpolando linealmente entre los dos percentiles que correspondan con datos observados.

Por ejemplo, para el conjunto de datos $x = \{1, 3, 5, 7, 9, 10\}$ tendremos que el número 5, que es el dato en posición 3, es el percentil

$$100\left(\frac{3-0{,}5}{6}\right) = 41{,}67.$$

De la expresión (1.5) obtenemos que el percentil p estará en la posición

$$\begin{aligned} p &= 100\left(\frac{i-0{,}5}{N}\right) \Rightarrow \frac{pN}{100} = i - 0{,}5 \\ &\Rightarrow i = \frac{pN}{100} + 0{,}5. \end{aligned}$$

Por ejemplo, con los datos anteriores, el percentil 25 (Q_1) estará en la posición

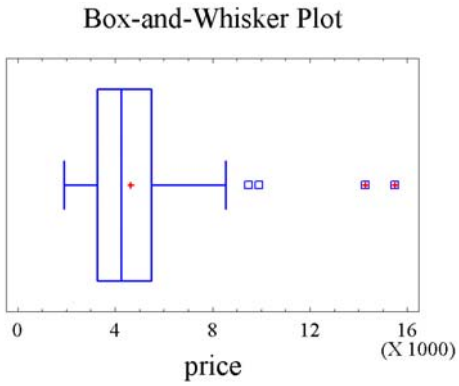
$$i = \frac{25 \times 6}{100} + 0{,}5 = 2$$

que corresponde con el valor 3, tal y como se calculó más arriba.

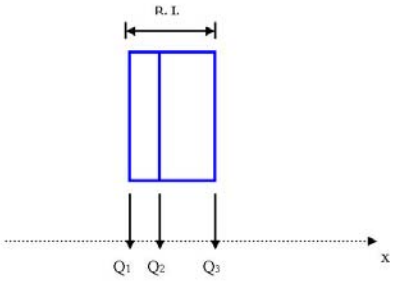
En la práctica, medidas características como los percentiles y cuartiles los calcularemos con ayuda de una aplicación estadística.

1.5.3. Diagrama de caja o box-plot

Al conjunto de los tres cuartiles junto con el mínimo y el máximo de los datos se le suele denominar el **resumen de las 5 cifras** (mínimo, Q_1, Q_2, Q_3, máximo). El diagrama de la caja, o box-plot, es precisamente la representación gráfica de resumen de las 5 cifras. Un ejemplo del diagrama de caja usando los datos del precio de los automóviles puede verse en la siguiente figura

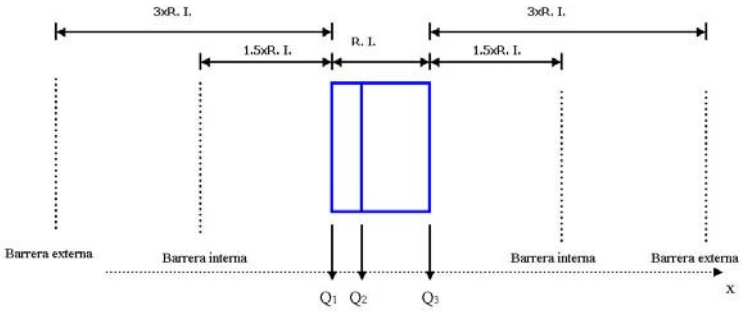


La construción de este diagrama se hace de la siguiente manera. Sobre la recta real se coloca una caja donde el extremo izquierdo corresponde con Q_1 y el derecho con Q_3. En el interior de la caja se coloca otro segmento que corresponde con Q_2, tal y como muestra la siguiente figura



Caja de un diagrama de caja

Esta caja representa la localización del 50 % de los datos en posición central. El ancho de la base de esta caja es el **rango intercuartílico o RI**. Además de la caja, se marcan unas areas que servirán para distinguir aquellos valores que se alejan demasiado de la caja, es decir, que se alejan mucho de este 50 % de datos que están en posición central. Se marcan 2 zonas. La primera se denomina barrera interior y se coloca a una distancia de 1.5×RI a cada lado de la caja. La segunda barera se denomina barrera externa y se coloca a $3\times RI$ de los extremos de la caja, tal y como se muestra en la siguiente figura.



Caja y barreras externa e interna

Una vez establecidas dichas barreras se marcan aquellas observaciones que caen más allá de dichas barreras, utilizando símbolos diferentes en cada zona. A los datos que caen en dichas zonas se les etiquetará como **atípicos**. A los atípicos que están más allá de la barrera externa se les añade el calificativo de **atípicos extremos**. A continuación se dibujan unos segmentos que salen de ambos lados de la caja hasta la observación más alejada, pero que se encuentre dentro de las barreras. Si no hubiese ningún dato atípico, los segmentos llegarían hasta el valor mínimo y el máximo. En el box-plot construido anteriormente con los datos de los precios puede apreciarse que hay 4 datos atípicos, dos de ellos extremos.

En un box-plot hay que fijarse principalmente en el rango intercuartílico, el rango descontando atípicos, la asimetría y los atípicos. El box-plot de los precios muestra asimetría positiva, pues

la distancia entre cuartiles aumenta de derecha a izquierda. Al haber asimetría positiva, **este carácter de atípicos debe interpretarse con cautela**, pues si bien es cierto que se encuentran en las zonas más allá de las barreras, su presencia es compatible con la asimetría de la distribución. El que una observación venga marcada y etiquetada como atípica en un box-plot es sólo una convención. Es el analista el que debe decidir si un dato es realmente atípico. Es frecuente añadir la información de la posición de la media muestral dentro de la caja. En el gráfico de los precios, la media muestral está indicada con el signo +.

Ejemplo 14 *El siguiente box-plot corresponde a los datos de las mediciones de los clavos realizadas por dos personas que usaban calibres distintos y de las que se mostró el histograma anteriormente (longitudclavos.sf). Ahora mostramos el box-plot con las longitudes realizadas por cada persona. Puede verse que ambas muestras son diferentes. El calibre 2 parece peor que el calibre 1 pues sus mediciones tienen más dispersión. Los valores medios y las medianas son también muy diferentes, lo cual es también indicio de que ambos calibres tienen muy distinta calidad (los clavos proceden de la misma caja y son seleccionados por cada persona al azar).*

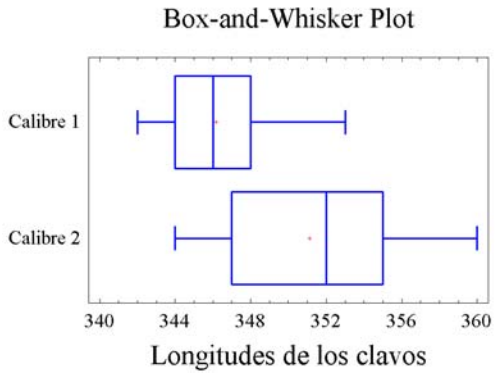


Diagrama de caja de las longitudes de los clavos

Ejercicios 7 .....

¿Qué opinión te merece el siguiente box-plot?

(1.6)

.....

1.5.4. Otras medidas de forma

Medidas de asimetría

Definimos coeficiente de asimetría como

$$CA = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}.$$

Este coeficiente también recibe el nombre de coeficiente de asimetría de Fisher. Este coeficiente toma los siguientes valores:

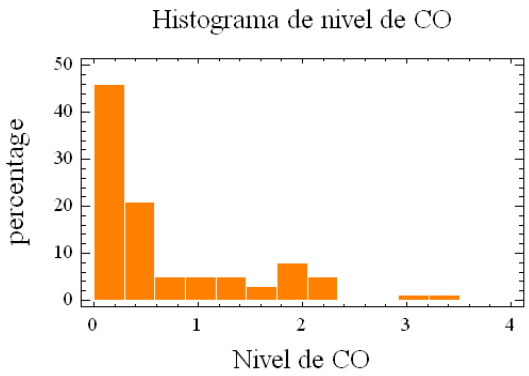
$$\begin{aligned} CA &= 0; \text{ si la distribución es perfectamente simétrica} \\ CA &> 0; \text{ si hay asimetría positiva} \\ CA &< 0; \text{ si hay asimetría negativa} \end{aligned}$$

Un coeficiente de asimeytría mayor que 1 en valor absoluto puede considerarse alto. Otra medida, menos popular, de medir la asimetría es mediante la comparación de la media con la mediana, mediante el siguiente coeficiente llamado coeficiente de sesgo (CS):

$$CS = \frac{3\,(\bar{x} - x_m)}{s},$$

que tiene la misma interpretación que CA.

Ejemplo 15 *El histograma siguiente muestra la distribución de frecuencias relativas de la concentración de monóxido de carbono (CO) de los gases emitidos por 100 vehículos. Esta concentración de CO se midió en una estación ITV a 100 turismos consecutivos (fichero COitv.sf3). Puede apreciarse en el histograma que la distribución tiene una asimetría positiva muy acusada.*

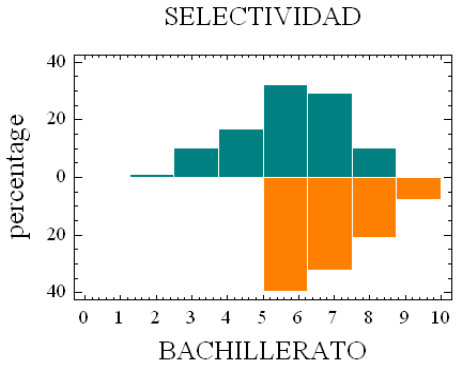


Concentración de monóxido de carbono (CO) en los gases de escape de 100 vehículos en una ITV.

El coeficiente de asimetría, CA, es $CA = 1{,}46$ *que es un valor muy elevado.*

Ejemplo 16 *El fichero califselectividad.sf3 contiene las notas del expediente de bachillerato de 106 alumnos de un instituto de secundaria de Madrid que se presentaron a las pruebas de acceso a la universidad. El fichero contiene también las notas que esos mismos alumnos obtuvieron en*

selectividad. La figura siguiente muestra los histogramas de ambos grupos de datos así como un conjunto de medidas características calculadas con el Statgraphics..



Summary Statistics

	SELECTIVIDAD	BACHILLERATO
Count	106	106
Average	5,70726	6,82453
Variance	2,14378	1,28587
Standard deviation	1,46416	1,13396
Skewness	-0,264352	0,707996

Puede verse en los histogramas que la distribución de notas de selectividad tiene una ligera asimetría negativa, siendo el coeficiente de asimetría CA=-0.2643 (según el Statgraphics), que es un valor pequeño. Sin embargo, las notas de bachillerato son asimétricas positivas, con unj coeficiente CA=0.708, que no es pequeño. Viendo más despacio los histogramas podemos encontar la explicación a esta diferente asimetría en ambas calificaciones. Las notas de selectividad tiene tanto valores aprobados como suspensos. Sin embargo, las notas de bachillerato sólo son de alumnos aprobados. Recuérdese que estos datos son de los alumnos que se presentaron a la prueba de selectividad, por lo que serán alumnos que han superado ya el bachillerato. Esto supone un truncamiento de los valores posible del expediente de bachillerato. Ese truncamiento hace que la distribución de notas sea asimétrica positiva.

Medidas de apuntamiento

El apuntamiento o curtosis mide lo puntiaguda que es la distribución. El coeficiente que mide este apuntamiento es

$$CAp = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4}.$$

Este coeficiente estará próximo a 3 si el histograma tiene una forma parecida a una campana (mesocúrtica), mayor que 3 si es más puntiaguda que una campana (leptocúrtica), o menor que 3 si es menos puntiaguda que una campana (platicúrtica). Es frecuente también expresar el coeficiente de curtosis como $CAp - 3$, con lo que valores positivos indicarán un fuerte apuntamiento y valores negativos indican una distribución más plana.

1.6. Transformaciones lineales y su efecto en las medidas características

Sea X la variable de interés de la que tenemos un conjunto de n datos, $x_1, ..., x_n$. Queremos saber las consecuencias sobre las medidas características de hacer una transformación lineal de X, es decir, queremos saber cómo serán las medidas características de $Y = a + bX$.

Efecto en la media

Vamos a demostrar que la media sufre la misma transformación lineal que las observaciones, es decir $\bar{y} = a + b\bar{x}$.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (a + bx_i)}{n} = \frac{\sum_{i=1}^{n} a}{n} + \frac{b\sum_{i=1}^{n} x_i}{n} = a + b\bar{x}. \tag{1.7}$$

Efecto en la mediana

Como la mediana es el dato en la posición central, y la transformación lineal no cambia la posición relativa de las observaciones, se tiene que

$$y_m = a + bx_m.$$

y el mismo razonamiento es válido para los cuartiles. Por tanto

$$\begin{aligned} Q_1(y) &= a + bQ_1(x) \\ Q_3(y) &= a + bQ_3(x) \end{aligned}$$

Efecto en la moda

Al ser una transformación lineal se tiene que el dato más repetido lo seguirá siendo, solo que dicho valor será el dato transformado. Por tanto

$$\text{Moda}(y) = a + b \times \text{Moda}(x).$$

Efecto en la varianza

Usando un desarrollo similar a (1.7) se tiene que

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^{n} (a + bx_i - a - b\bar{x})^2}{n} = \frac{\sum_{i=1}^{n} b^2 (x_i - \bar{x})^2}{n} = bs_x^2$$

y por tanto, para la desviación típica se tiene que

$$s_y = |b| s_x.$$

Es importante ver que la constante a no afecta a estas medidas de dispersión. Este resultado es intuitivo, pues a sólo es una traslación de los datos, por lo que la dispersión de los datos alrededor de la media no cambia.

Efecto en el rango y rango intercuartílico

Al ser ambos rangos diferencia de observaciones se tiene que de nuevo a no afecta. Es fácil deducir que, al no afectar la combinación lineal a la posición relativa de los datos,

$$\begin{aligned} \text{Rango}(y) &= \text{máx}(y_i) - \text{mín}(y_i) = \text{máx}(a + bx_i) - \text{mín}(a + bx_i) \\ &= a + b\,\text{máx}(x_i) - a - b\,\text{mín}(x_i) = b\text{Rango}(x). \\ \text{RI}(y) &= Q_3(y) - Q_1(y) = a + bQ_3(x) - a - bQ_1(x) = b\text{RI}(x) \end{aligned}$$

Ejemplo 17 *Para la producción de cierta aleacción metálica es muy importante tener controlada la temperatura del horno donde se realiza la aleacción. En el horno hay instalados varios termopares de wolframio (los termopares de wolframio son unos termómetros que permiten medir la temperatura por encima de los 1000º C). En un instante dado, las mediciones de temperatura realizadas por los termopares* $(x_i, i = 1, ..., n)$ *tienen las siguientes medidas características:*

$\bar{x} = 1651^oC$	$Mediana = 1652,3^oC$
$s_x^2 = 298,7^oC^2$	$Q_1 = 1638^oC$
$s_x = 17,28^oC$	$Q_3 = 1663^oC$
$Rango = 87,07^oC$	

Si se sabe que la temperatura real es un 5 % superior a la que miden los termopares, indicar cuáles serían las medidas características de las temperaturas reales en grados Kelvin (ºK=º C+273).

SOLUCIÓN:

Llamemos T_m *a la temperatura que mide el termopar, y* T_r *a la temperatura real en* ^oK. *Entonces*

$$T_r = 1{,}05T_m + 273$$

lo que implica una transformación lineal de la temperatura medida. Si llamamos y_i *a las temperaturas reales en grados Kelvin de cada termopar, se tiene que:*

$$\begin{aligned}
\bar{y} &= a + b\bar{x} = 273 + 1{,}05\bar{x} = 2006,5^oK \\
s_y^2 &= b^2 s_x^2 = 1{,}05^2 s_x^2 = 329^oK^2; \\
s_y &= |b| s_x = 1{,}05 s_x = 18,1^oK;
\end{aligned}$$

La posición relativa de los datos no cambia en una transformación lineal, por tanto

$$\begin{aligned}
Rango &= y_{\text{máx}} - y_{\text{mín}} = (1{,}05x_{\text{máx}} + 273) - (1{,}05x_{\text{mín}} + 273) = 1{,}05(x_{\text{máx}} - x_{\text{mín}}) = 91{,}4^oK \\
Mediana &= 1{,}05x_{mediana} + 273 = 2007,9^oK; Q_1 = 1992,9^oK; Q_3 = 2019{,}1^oK.
\end{aligned}$$

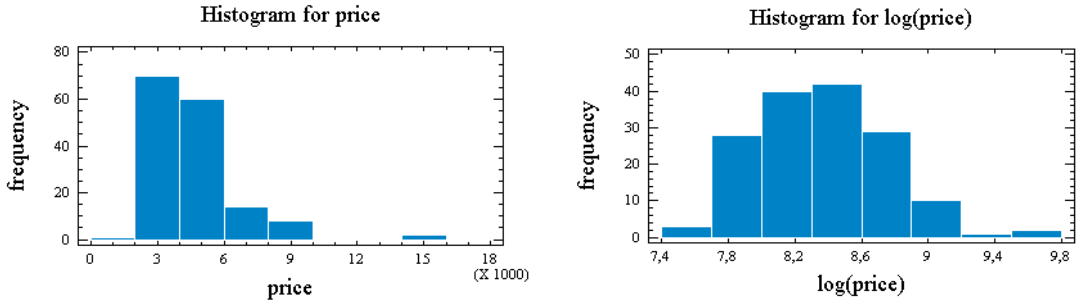
1.7. Transformaciones no lineales que mejoran la simetría

En los próximos capítulos aprenderemos un conjunto de técnicas estadísticas muy útiles para sacar conclusiones de un conjunto de datos. Muchas de estas técnicas sólo son aplicables cuando los datos son unimodales y simétricos. Aunque esta restricción pueda parecer muy fuerte, este tipo de distribuciones es bastante frecuente en datos reales. También es muy frecuente que cuando no se cumple ese patrón lo sea sólo porque los datos son asimétricos. La solución en esos caso es aplicar una transfornmación no lineal que produzca simetría.

Transformación de datos con asimetría positiva

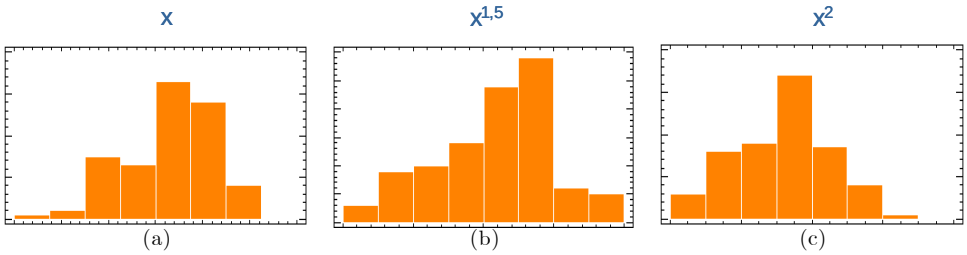
Si un conjunto de datos presenta asimetría positiva se puede intentar transformarlo en otro que sea más simétrico mediante transformaciones del tipo $\sqrt{x}$, $\log(x)$, ó x^c con $c < 1$, por ejemplo $c = -1$. Este tipo de transformaciones hacen que los números más grandes se reduzcan mucho, mientras que los más pequeños se reducirán poco. Si por ejemplo tuviésemos los datos x :{10, 20, 200, 5000} y calculamos sus logaritmos (neperianos) tendríamos $\log(x)$ ={ 2.3, 3.0, 5.3, 8.6}, que están más próximos entre sí que los originales. Vemos así cómo el número 5000 pasa a ser 8.6 tras la transformación, lo que implica una reducción importante, mientras que el número 10 pasa a 2.3, que es una reducción mucho menor.

Con este tipo de transformaciones aplicadas a datos con asimetrías positivas, lo que se consigue es comprimir la cola de la derecha de la distribución, obteniéndose una distribución más simétrica. Cuanto mauyor sea la asimetría positiva, más 'fuerte' debe ser la capacidad transformadora. Esta mayor capacidad transformadora se consigue con valores menores de c, donde la transformación logarítmica $\log(x)$ puede interpretarse como el límite cuando $c \to 0$. Por tanto, la transformación logarítmica es más fuerte que transformaciones con $c > 0$, pero menos que transformaciones con $c < 0$. Las transformaciones más frecuentes son $x^{1/2}$, $\log(x)$ y x^{-1}. El siguiente histograma muestra un ejemplo de distribución asimétrica positiva que tras una transformación logarítmica se hace más simétrica. La variable es el precio de los vehículos del fichero *cardata*.



Transformaciones de datos con asimetría negativa

Si un conjunto de datos presenta asimetría negativa, su transformación en una distribución simétrica se hará mediante las transformaciones x^c, con $c > 1$. En estos casos, lo que se hace es expandir la parte de la derecha del histograma que compense la cola de la izquierda, y así conseguir una mayor simetría. La figura siguiente muestra el histograma de un conjunto de datos de una variable X con asimetría negativa (figura (a)). Para mejorar su simetría se hace primeramente la transformación $x^{1,5}$ que parece ser insuficiente (figura (b)). Finalmente, una transformación más fuerte x^2 consigue una simetría suficiente (figura (c)).



Hay que tener cuidado con la presencia de valores que hagan las operaciones inviables, como tomar logaritmos de números menores de 1. En esos caso se suma una cantidad a todos los datos de forma que se pueda tomar logaritmos. También ha de tenerse cuidado con la presencia de datos positivos y negativos cuando se eleve a una potencia par, pues al perderse el signo negativo los datos transformados no guardarán ninguna relación con los originales. De nuevo, en esos caso, se suma una misma cantidad a todos los datos para que todos sean positivos antes de elevar al cuadrado.

1.8. Relación entre dos variables. La recta de regresión

En la sección anterior se presentaron un conjunto de medidas que resumen alguna característica de un conjunto de datos de una variable. Estas medidas características son un resumen numérico de las propiedades que se hayan visualizado utilizando gráficos tales como el histograma, el diagrama de barras, etc.

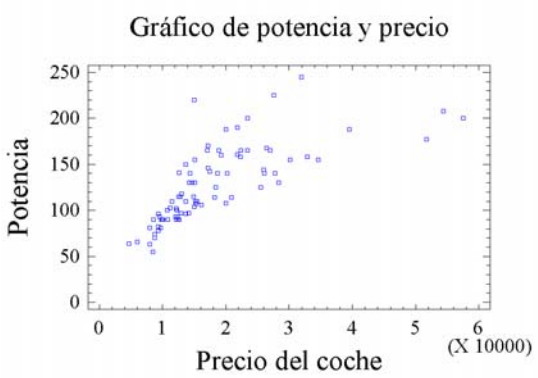
En esta sección presentaremos medidas que resuman la asociación entre dos variables cuantitativas. Estas medidas serán resúmenes numéricos de las relaciones que se hayan detectado usando los gráficos de dispersión presentados anteriormente. Nos centraremos sólamente en relaciones lineales entre dos variables. En esos casos, el gráfico de dispersión mostrará una nube de puntos alrededor de cierte linea recta imaginaria. Para resumir el grado de relación lineal entre dos variables se usan las siguientes medidas: (1) coeficiente de covarianza, (2) coeficiente de correlación (3) recta de regresión.

1.8.1. Coeficiente de covarianza

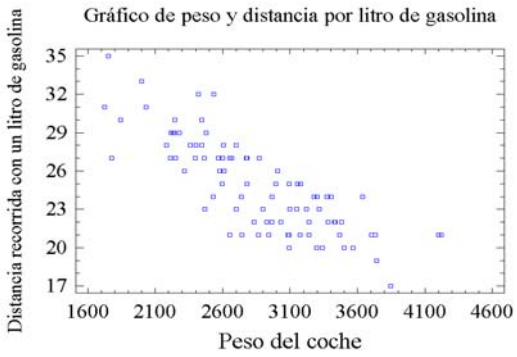
Supongamos que para un conjunto de n individuos se tiene información de dos variables x e y. Entonces la covarianza o coeficiente de covarianza se define como

$$\text{cov}(x, y) \equiv s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{n}.$$

Este coeficiente de covarianza tomará valores positivos si hay una relación lineal positiva entre ambas variables, es decir, si al aumentar una de ellas también lo hace la otra. El siguiente gráfico de dispersión ilustra esta covarianza positiva. El gráfico (a) muestra el precio del coche y su potencia para un conjunto de 90 coches europeos, donde la covarianza es 302764. Por el contario, si la relación lineal es negativa, el coeficiente de covarianza será negativo, como sucede con los datos de peso del coche y la distancia que recorrerá con un litro de combustible, que posee una covarianza de -1600



(a) Relación entre precio y potencia de 90 coches



(b) Relación entre peso y consumo de 90 coches

Si entre ambas variables no hay relación, la covarianza será próxima a cero. En estos casos, el diagrama de dispersión muestra una nube de puntos sin ningún patrón de relación. El signo del coeficiente de covarianza nos indica claramente el signo de la relación lineal que exista entre las variables. Sin embargo, al depender el coeficiente de covarianza de las unidades de x y de y , el valor concreto no es fácil de interpretar. Por ejemplo, en los dos gráficos anteriores, no sabríamos decir, apartir de los valores de covarianzas, que relación lineal es más fuerte. Para el gráfico (a)

la covarianza es 302764 dólares×CV, mientras que para el gráfico (b) la covarianza es de -1600 kilogramos×millas, que no es en absoluto comparable con dólares×CV.

Sería conveniente por tanto encontrar alguna medida característica que resuma la relación lineal de forma adimensional. Esta medida es el coeficiente de correlación, que se muestra a continuación.

1.8.2. Coeficiente de correlación

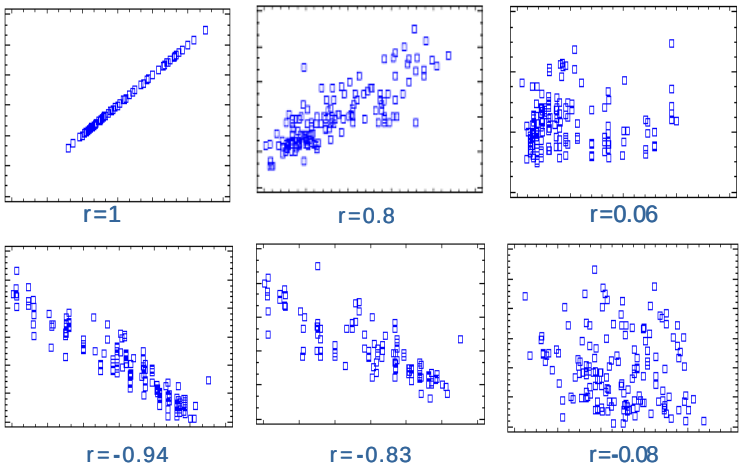
La información que suministra el coeficiente de correlación es la misma que la covarianza. Es un coeficiente que mide el grado de relación lineal entre dos variables tomadas en los mismos individuos, pero usando un valor adimensional. Se define como

$$r \equiv r_{xy} \equiv r(x,y) = \frac{\text{cov}(x,y)}{s_x s_y}.$$

Puede demostarse que estará siempre entre -1 y 1. Su interpretación es

$$\begin{aligned} r &= 0; \text{ no hay relación lineal} \\ r &> 0; \text{ relación lineal positiva} \\ r &< 0; \text{ relación lineal negativa} \end{aligned}$$

Si $r = 1$ tendremos una relación lineal positiva perfecta, en el sentido de que los datos estarán perfectamente alineados según una recta de pendiente positiva. Análogamente, si $r = -1$ tendremos una relación lineal negativa perfecta. En el caso de las variables precio y potencia mostradas anteriormente la correlación es 0.73 que es positiva y alta. En el caso de las variables peso y distancia recorrida, la correlación es -0.82 que es negativa y muy alta. Cuanto más próxima esté la nube de puntos a una línea recta más próximo estará el coeficiente de correlación a la unidad (en valor absoluto). Por el contrario, cuanto más dispersa esté la nube de puntos, la correlación estará más próxima a cero. A continuación se muestra una serie de figuras donde se representan conjuntos de datos de diferente coeficiente de correlación.

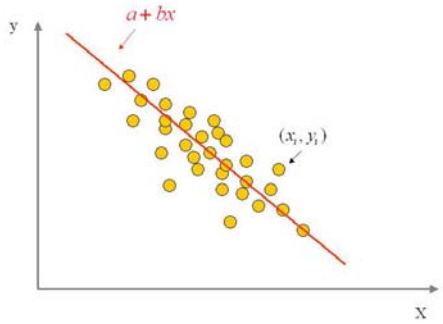


1.8.3. La recta de regresión

Definición de la recta de regresión

Nuestro interés en el cálculo de la correlación es medir la proximidad de la nube de puntos a una línea recta imaginaria. Lo que vamos a hacer ahora es obtener la ecuación de esa linea recta que sirva de resumen de la relación entre ambas variables. Es importante recalcar que esta línea recta es sólo una aproximación de la relación entre ambas variables. Cuando más próximo a ± 1 esté el coeficiente de correlación, mayor será la capacidad de aproximación de dicha recta. A este procedimiento de buscar una recta que aproxime el comportamiento de una nube de puntos le llamaremos **ajuste de una recta**.

Nuestro objetivo es encontrar la recta $a+bx$ que mejor resuma esa tendencia lineal que muestra la nube de puntos, como se ilustra en esta figura.



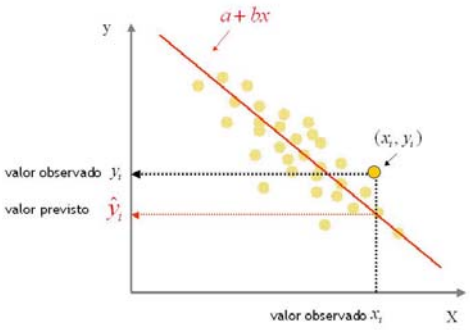
Conjunto de datos y recta que resume su tendencia lineal

Si la nube de puntos formase una línea recta perfecta (lo que ocurre sólo si la correlación entre ambas es ± 1) tendríamos que $y = a + bx$, y los valores de a y b los podríamos calcular usando sólo un par de puntos. En cualquier otro caso, si la correlación es diferente a ± 1, es imposible encontrar una línea recta que pase por todos los puntos.

Dada una nube de puntos formada por un conjunto de datos de dos variables, existen muchos criterios para ajustar una recta, lo que llevaría a calcular rectas diferentes según el criterio que empleemos. Nuestro interés es en la recta que nos ayude a predecir el valor de y a partir de la observación de x. A la variable que queremos saber su valor le denominamos **variable respuesta**, y se le suele reservar la letra y. A la variable que vamos a usar para predecir el valor de la variable respuesta le denominaremos **variable explicativa**, y se le suele reservar la letra x. Al valor que resulta de aplicar la recta $a + bx$ para un valor de $x = x_i$ dado, le llamaremos predicción, y lo denotaremos por el símbolo $\hat{y}$, o también $\hat{y}(x_i)$, o $\hat{y}_i$. La recta que predice el valor de y cuando se conoce el de x puede expresar entonces como

$$\hat{y} = a + bx.$$

La siguiente figura muestra la diferencia entre un valor observado y_i y un valor previsto $\hat{y}(x_i)$ por la recta $\hat{y} = a + bx$.



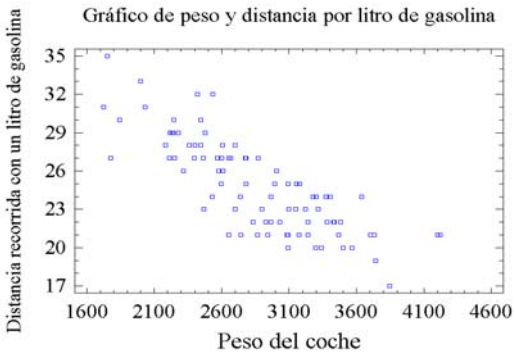
La teoría estadística nos dice que la recta que mejor hace esas predicciones de y es la que viene dada por los siguientes valores

$$b = \frac{\text{cov}(x,y)}{s_x^2}, \tag{1.8}$$

$$a = \bar{y} - b\bar{x}. \tag{1.9}$$

A la recta $\hat{y} = a + bx$ que utiliza los valores (1.9) y (1.8) le denominaremos **recta de regresión**.

Ejemplo 18 *El fichero cardata.sf tiene datos de 155 vehículos. Entre estos datos tenemos las variables mpg=distancia recorrida con un galón de combustible, y la variable weight=peso del vehículo. El gráfico de dispersión basado en estos 155 vehículos es*



(b) Datos de peso y consumo de coches

donde puede verse que hay una fuerte tendencia lineal negativa, con un coeficiente de correlación de -0.82. Queremos calcular la recta de regresión que me ayude a dar un valor aproximado de la distancia que recorrera un vehículo conocido su peso. Del análisis de los datos tenemos que

cov(mpg,weight)=–3688.24
var(mpg)=54.42
var(weight)=363630
media mpg=28.79
media weight=2672.2

Calcularemos la recta de regresión que nos ayude a predecir la distancia recorrida (mpg) en función del peso (weight). Por tanto nuestras variables son y =mpg y x =weight. La recta de regresión es entonces

$$\begin{aligned} b &= \frac{cov(x,y)}{s_x^2} = \frac{-3688{,}24}{363630} = -0{,}01014 \\ a &= \bar{y} - b\bar{x} = 28{,}79 - (-0{,}01014) \times 2672{,}2 = 55{,}89 \end{aligned}$$

Supongamos ahora que tenemos un vehículo que pesa 2600 unidades y que no conozcamos su consumo. Usando como aproximación la anterior recta de regresión podemos predecir que la distancia que recorra con un galón de combustible será

$$\textit{Distancia prevista}=\hat{y}(x = 2600) = a + b \times 2600 = 55{,}89 + (-0{,}01014) \times 2600 = 29{,}52 \textit{ millas.}$$

Algunas propiedades de la recta de regresión

Antes se ha mencionado que la recta de regresión es la recta que "mejor"predice y dado un valor de x. Sin embargo, no se ha dicho en qué sentido esa predicción es mejor. Vamos a aclarar este punto. Si aplicamos la recta de regresión a un valor de x del que hayamos observado el valor de y, veremos que, en general, el valor que predice la recta de regresión $\hat{y}$ no será igual que el observado (¿siempre?). Tendremos una discrepancia. Llamaremos **error de predicción**, o **residuo**, a esa discrepancia, y lo denotaremos con la letra e. Es decir,

$$\text{residuo}=e_i = y_i - \hat{y}_i.$$

Si hemos construido la recta de regresión con n pares de puntos (x_i, y_i), $i = 1, 2, ..., n$, tendremos n residuos $e_1, ..., e_n$. Puede demostrarse que la recta de regresión definida con (1.8) y (1.9) es la recta que tiene los residuos más pequeños. Más concretamente, la recta de regresión es la recta tal que los residuos que se obtienen tienen suma de cuadrados $\sum_{i=1}^{n} e_i^2$ mínima. Por esa razón a esta recta también se le denomina **recta de mínimos cuadrados**, o **minimo cuadrática**.

Otra propiedad que cumple la recta de regresión es que pasa por el punto $(\bar{x}, \bar{y})$. Es inmediato comprobar este resultado usando el valor de a en (1.9). Para $x = \bar{x}$ se tiene que

$$\hat{y}(x = \bar{x}) = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

Por último, hay que mencionar que la recta de regresión $\hat{y} = a + bx$ está diseñada para predecir y a partir de x y no debemos usarla para predecir un valor de x dado y. Es decir, si observamos un valor de y y predecimos el valor de x haciendo $\hat{x} = (y - a)/b$ ya no estamos usando la mejor recta que prediga x. No estamos haciendo el mejor uso de nuestros datos. Debemos en ese caso calcular una nueva recta intercambiando los papeles de x e y. Volviendo al Ejemplo18, si nuestro interés es encontar la predicción del peso de un vehículo si conociésemos la distancia que recorre con un galón de combustible, lo mejor es volver a construir una nueva regresión específica para esa predicción. Ahora la variable explicativa es x =mpg y la variable respuesta es y =weight, y tendremos que

$$\begin{aligned} b^* &= \frac{\text{cov}(x,y)}{s_x^2} = \frac{-3688{,}24}{54{,}42} = -67{,}774 \\ a^* &= \bar{y} - b\bar{x} = 2672{,}2 - (-67{,}774) \times 28{,}79 = 4623{,}4. \end{aligned}$$

Entonces, si sabemos que un vehículo ha recorrido 10 millas con un galón de combustible, el peso que podemos predecir para ese vehículo será de

$$\text{Peso previsto}=\hat{y}(x=10)=a^*+b^*\times 10=4623{,}4+(-67{,}774)\times 10=3945{,}7.$$