UNIDAD DIDÁCTICA 2

ESTADÍSTICA DESCRIPTIVA

Objetivo: El objetivo de esta Unidad Didáctica es introducir los conceptos más elementales de la Estadística, así como familiarizar al alumno con algunas técnicas, sencillas pero poderosas, de Estadística Descriptiva.

Contenido

1. Estadística descriptiva unidimensional

Como ya se mencionó en la UD 1, la Ciencia Estadística tiene un doble objetivo:

- La generación y recopilación de datos que contengan información relevante sobre un determinado problema
- El análisis de dichos datos con el fin de extraer de ellos dicha información.

El primer paso en el análisis de un conjunto de datos debe ser siempre un tratamiento descriptivo sencillo de los mismos. Dicho tratamiento busca poner de manifiesto las características y regularidades existentes en los datos y sintetizarlas en un número reducido de parámetros o mediante representaciones gráficas adecuadas.

En este primer nivel del análisis, puramente descriptivo, no se pretende todavía extrapolar conclusiones de los datos a la población de la que éstos han sido extraídos, lo que constituirá el objeto de las técnicas de Inferencia Estadística.

1.1. Conceptos básicos

1.1.1. Poblaciones

En la terminología estadística se denomina **población** al **conjunto de todos individuos o entes que constituyen el objeto de un determinado estudio y sobre los que se desea obtener ciertas conclusiones**.

Ejemplo 1: en un estudio sobre la intención de voto de los ciudadanos españoles la población la constituirá el conjunto de los aproximadamente 30 millones de españoles con derecho a voto.

Ejemplo 2: en un estudio sobre el desarrollo de la tristeza de los cítricos en la Comunidad Valenciana la población estará formada por la totalidad de árboles de cítricos existentes en esta Comunidad.

Ejemplo 3: al realizar en una industria el control de calidad en recepción de una partida de piezas, la población estará constituida por la totalidad de las piezas que componen la partida.

Los ejemplos anteriores tratan en todos los casos de poblaciones con una existencia física real, constituidas por un número finito, aunque posiblemente muy elevado, de individuos.

Aunque pueda parecer sorprendente no es ésta la situación más frecuente en la práctica, sino que en general las poblaciones a estudiar son de carácter abstracto, fruto del necesario proceso de conceptualización que debe preceder al estudio científico de cualquier problema real.

Ejemplo 4: un ejemplo trivial sacado de los juegos de azar sirve para ilustrar la idea anterior. Se desea estudiar si un dado es correcto o está trucado. ¿Qué quiere decir la afirmación de que el dado es correcto? En la práctica, que si se tira un número muy elevado de veces los seis resultados posibles saldrán aproximadamente con la misma frecuencia. Al abordar este problema nos referiremos a la población abstracta constituida por infinitos lanzamientos del dado en cuestión, población sobre la que deseamos

estudiar si la frecuencia relativa con la que se presentan los seis resultados posibles son idénticas.

Ejemplo 5: en una investigación sobre el rendimiento de una nueva variedad de trigo, la población sobre la que interesa obtener resultados podrían constituirla todas las parcelas plantadas con dicha variedad que puedan existir en el futuro.

Ejemplo 6: en un estudio sobre la eficiencia de diversos algoritmos de encaminamiento de mensajes entre nudos en una red de procesadores, la población a investigar la constituirían todos los mensajes que puedan llegar a generarse en la red.

Como se desprende de los ejemplos anteriores los "**individuos**" que forman una población pueden corresponder a entes de naturaleza muy diversa (personas, árboles, piezas, lanzamientos de dados, parcelas, mensajes, etc...). En los casos de los tres primeros ejemplos dichos individuos tienen una existencia real, previa a la realización del estudio. En casos como los de los ejemplos 4 5, y 6 los individuos que constituyen la población pueden irse generando mediante la realización de un determinado proceso (lanzar un dado, plantar una parcela con la variedad, emitir un mensaje desde un nudo,...). A estos procesos, que en sucesivas realizaciones pueden ir generando los diferentes individuos de la población les denominamos **experimentos aleatorios**.

1.1.2. Variables aleatorias

Concepto

¡En toda población real existe VARIABILIDAD!.

La vida útil de varios componentes electrónicos idénticos no es la misma; el número de asignaturas en las que se matricula un alumno también varía de uno a otro; el *troughput* de un sistema cambia de un instante a otro; el número que sale al lanzar el dado varía de unas tiradas a otras; unos mensajes tienen retardos más elevados que otros.

A c**ualquier característica que puede constatarse en cada individuo de una población** se le denomina **característica aleatoria**.

Así el partido a que piensan votar los individuos (Ejemplo 1) la ausencia o presencia de tristeza en los árboles (Ejemplo 2), el rendimiento obtenido en las parcelas (Ejemplo 5) o el retardo de un mensaje (Ejemplo 6) son características aleatorias.

Muchas características aleatorias se expresan numéricamente, por ejemplo, el número de puntos obtenidos al lanzar un dado, el tiempo hasta el fallo de un tipo de monitor o el retardo de un mensaje. A este tipo de características aleatorias se las denomina **variables aleatorias**.

Cuando una característica aleatoria es de tipo cualitativo, como por ejemplo el partido político a votar, destino correcto o no de un mensaje, nada impide codificar numéricamente sus diferentes alternativas y tratarla como una variable aleatoria.

> **NOTA:** hay que tener cuidado en estos casos porque operaciones perfectamente legítimas con características intrínsecamente numéricas (como, por ejemplo, sumar y promediar los rendimientos de diferentes sistemas informáticos) carecerían de sentido.

Variables discretas y variables continuas

Cuando el conjunto de los valores que podría tomar una determinada variable aleatoria es discreto (es decir, finito o infinito numerable) se dice que dicha **variables es de tipo discreto** (a veces a las variables de este tipo se les denomina también **atributos**), por oposición a aquellos casos en que dicho conjunto es un infinito continuo en los que la **variable** se denomina **continua**.

Ejemplos de variables discretas serían el número de puntos al lanzar un dado, el número de bajadas de un archivo de Internet, el número de errores en un programa de ordenador y también cualquier variable que se origine al codificar las diferentes alternativas de una característica cualitativa (sexo, partido votado, tipo de tarjeta gráfica, etc...).

Ejemplos de variables continuas serían todas las características que se miden sobre una escala de naturaleza básicamente continua (estaturas, pesos, rendimientos, tiempos, resistencias, etc...)

Variables k-dimensionales

Cuando sobre cada individuo de la población se estudian K características diferentes (todas ellas expresables numéricamente) se tiene una **variable aleatoria K-dimensional**. Por ejemplo si en la población constituida por los estudiantes de la UPV se estudia el sexo, la edad, la estatura y el peso, estaremos ante una variable aleatoria de dimensión 4.

En estos casos es frecuente utilizar los valores de aquellas componentes cuya naturaleza intrínseca es cualitativa (por ejemplo el sexo) para dividir la población inicial en subpoblaciones (en nuestro caso: chicos y chicas) entre las cuales interesa estudiar las diferencias en las pautas de variabilidad existentes en las otras componentes de la variable aleatoria, por ejemplo para estudiar cómo difieren las pautas de variabilidad del peso o la estatura entre chicos y chicas en la UPV.

Ejercicio 1: el peso y el modelo de un *Tablet PC* ¿constituyen una variable aleatoria bidimensional? ¿Y el número de líneas de código y el número de errores en los programas preparados por una empresa de software? ¿Y el contenido de leucocitos en la sangre de individuos alcohólicos y no alcohólicos? ¿Y las estaturas del marido y de la mujer en los matrimonios jóvenes de un país?

NOTA: es importante darse cuenta de la diferencia entre una variable aleatoria K-dimensional, en la que las K variables se miden sobre los individuos de una única población, y un conjunto de K variables aleatorias unidimensionales, definidas sobre K poblaciones distintas.

Muestras. Datos estadísticos

En general no resulta posible estudiar la totalidad de los individuos de una población para obtener información sobre ésta. Incluso cuando esta posibilidad existe técnicamente,

como es el caso al tratar de poblaciones reales finitas, dicho procedimiento suele ser impracticable por consideraciones económicas.

En consecuencia para obtener información sobre una población hay que limitarse a analizar sólo un subconjunto de individuos de la misma. A éste subconjunto se le denomina **muestra**.

La forma de seleccionar los individuos que han de constituir la muestra tiene, como es lógico, una importancia capital para garantizar que ésta permita obtener conclusiones que puedan extrapolarse validamente a la población de la que la muestra procede. No hay que olvidar nunca que el objeto final del estudio es siempre la población y que la muestra es sólo un medio para obtener información sobre ésta.

Con el fin de permitir inferir conclusiones válidas sobre una población la muestra debe ser "representativa" de ésta. En teoría la única forma de garantizar la representatividad de una muestra es seleccionando al azar los individuos que la van a componer, de forma que todos los individuos de la población tengan "a priori" una probabilidad idéntica de pertenecer a la muestra. Aunque ésta forma de proceder rara vez sea aplicable de forma estricta en la práctica, siempre hay que extremar las precauciones para que la forma real de obtener la muestra sea lo más parecida posible a la ideal.

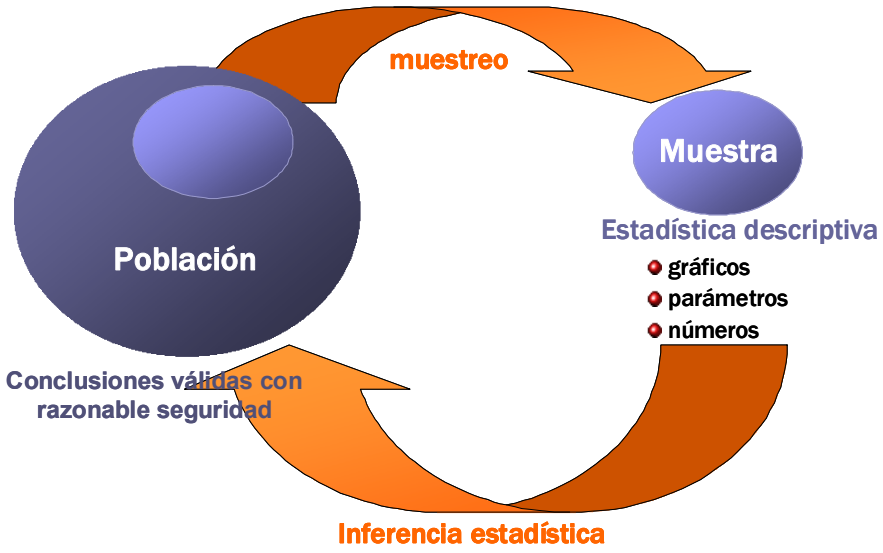
En realidad en muchos casos un conocimiento previo sobre la población es indispensable para decidir si una muestra puede considerarse o no representativa de la misma.

Ejercicio 2: se desea estudiar la relación que existe entre la estatura y el peso en la juventud española. El conjunto de los alumnos matriculados en Estadística en 1º del Grado en Ingeniería Informática de Valencia ¿puede considerarse una muestra representativa de la población a efectos del estudio en cuestión?

Dicho conjunto ¿puede considerarse una muestra representativa para estudiar las tendencias políticas en la juventud española? ¿Y para estudiar el nivel cultural? ¿Y para estudiar la característica aleatoria color de los ojos?

Cuando la población estudiada es real (Ejemplos 1, 2 y 3 del apartado anterior) la muestra se forma, como hemos señalado, seleccionando de la forma más aleatoria posible un conjunto de individuos de la misma. Cuando se muestrea una población abstracta, del tipo de las mencionadas en los Ejemplos 4 y 5, la forma de "extraer" una muestra no es más que realizar un cierto número de veces el experimento aleatorio que genera los individuos de la población. Por ejemplo, lanzar varias veces el dado, lanzar varias ejecuciones de una rutina que implementa un algoritmo en estudio o generar un conjunto de mensajes en la red de multiprocesadores.

Los valores observados para la variable aleatoria en los individuos que forman la muestra constituyen los **datos estadísticos**. El tratamiento de dichos datos con el fin de poner de manifiesto sus características más relevantes y sintetizarlas en unos pocos parámetros o mediante representaciones gráficas adecuadas, es el objeto de la **Estadística Descriptiva**. El análisis de los mismos con el fin de obtener conclusiones que, con un margen de confianza conocido, sean extrapolables a la población de la procede la muestra constituye el objeto de la **Inferencia Estadística**.



1.2. Tablas de frecuencias

El conjunto de valores observados, relacionados en el orden en el que han sido obtenidos, constituye el material inicial a partir del cual debe llevarse a cabo el análisis estadístico descriptivo.

Si el número de datos no es muy reducido, su interpretación se facilita presentándolos agrupados en una tabla.

Variables aleatorias cualitativas

Cuando la variable estudiada es cualitativa, o cuantitativa con un número reducido de valores posibles, los datos pueden sintetizarse en una **tabla** como la adjunta, en la que, en este caso, se pretende describir el funcionamiento simultáneo de un tipo de robots multiprocesador:

Nº de procesadores funcionando (X_i)	**Frecuencia absoluta** Nº de robots **(n_i)**	**Frecuencia relativa** % robots **$f_i = n_i / N$**
0	10	6,25%
1	35	21,88%
2	60	37,50%
3	55	34,37%
Total	160	100%

En esta tabla, para cada valor X_i constatado en la muestra se refleja la **frecuencia absoluta n_i** o número de veces que dicho valor ha sido observado en la muestra. Dado que las frecuencias absolutas dependen del número total N de observaciones, suele ser conveniente reflejar también en la tabla las **frecuencias relativas f_i** que no son más que los cocientes n_i/N.

Variables aleatorias continuas

Cuando la variable estudiada es de tipo continuo, y dado que el número de datos de la muestra es obviamente finito, nada impediría en principio emplear un procedimiento de tabulación similar al expuesto para el caso discreto. Sin embargo como será difícil encontrar valores repetidos de las X_i (de hecho si la variable se midiera con suficiente precisión la probabilidad de encontrar valores repetidos sería nula) la tabla resultante sería excesivamente prolija y casi tan difícil de interpretar como los datos iniciales. Por ello se acostumbra a proceder a un agrupamiento de los datos, dividiendo el campo de variación en un conjunto de K intervalos de igual longitud y anotando los límites y el valor central de cada intervalo, así como el número de observaciones constatadas en el mismo.

No es posible determinar "a priori" la amplitud óptima que deben tener los intervalos y, en consecuencia, el número de éstos. Un número excesivo de intervalos plantea el problema de conducir a una tabla muy prolija y difícil de interpretar, pero si el agrupamiento es excesivo se pierde una parte importante de la información contenida en los datos. En general valores entre 5 y 15 intervalos (dependiendo en parte del tamaño N de la muestra) suelen ser razonables, no estando en general justificado un nivel mayor de desagregación.

La siguiente tabla recoge, a título de ejemplo, el resultado de la tabulación en 11 intervalos de los valores del ratio entre los dos diámetros en 815 hojas de tabaco.

Limite del intervalo	Centro del intervalo X_i	Número de observaciones n_i
1,55 - 1,65	1,60	3
1,65 - 1,75	1,70	12
1,75 - 1,85	1,80	40
1,85 - 1,95	1,90	97
1,95 - 2,05	2,00	157
2,05 - 2,15	2,10	204
2,15 - 2,25	2,20	183
2,25 - 2,35	2,30	75
2,35 - 2,45	2,40	31
2,45 - 2,55	2,50	9
2,55 - 2,65	2,60	4

Con vistas a aumentar la información para un número determinado de intervalos se recurre a veces a establecer éstos con tamaños desiguales, más amplios en las zonas con pocos datos y más estrechos en las de mayor frecuencia de observaciones. Esta práctica, sin embargo, no es en general aconsejable, puesto que la información contenida en la tabla resulta más difícil de captar en un simple examen de la misma. En cambio puede resultar conveniente dejar dos intervalos abiertos en ambos extremos de la tabla, con el fin de recoger los pocos valores extremos observados.

En el establecimiento de intervalos conviene definir con precisión los límites de éstos y el tratamiento a dar a los valores que caigan exactamente sobre los mismos.

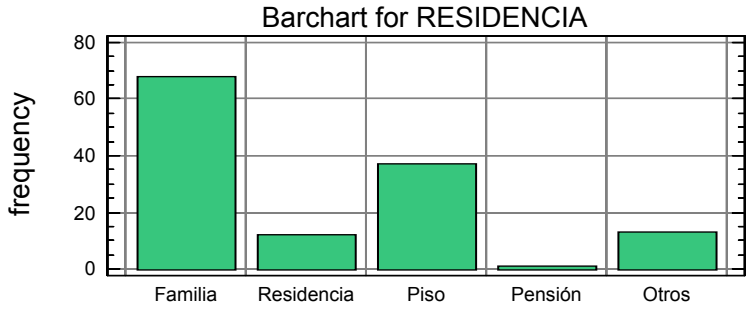
Señalemos por último que aunque una variable estudiada sea de tipo discreto también puede ser aconsejable agrupar los valores para su tabulación en el caso de que el campo de variabilidad de los datos sea muy amplio.

Ejercicio 3: Discuta el alumno la afirmación de que si la variable es continua y el sistema de medida es suficientemente preciso la probabilidad de encontrar dos valores repetidos es nula. ¿Por qué se pierde mucha información en la tabulación si el número de intervalos considerado es muy pequeño?

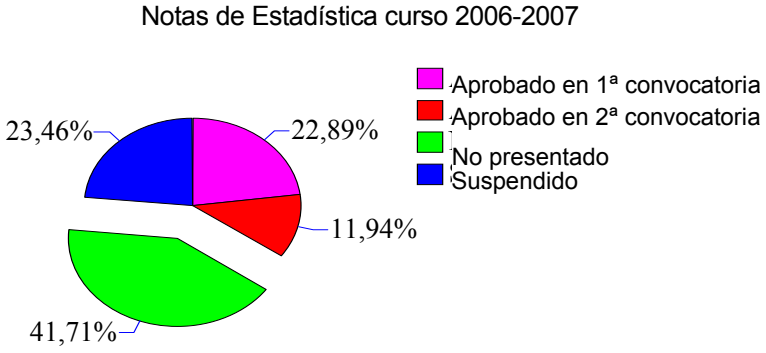
1.3. Diagrama de barras y tarta

En el análisis descriptivo de variables de naturaleza **cualitativa**, es muy habitual representar las frecuencias con las que se ha presentado en la muestra las diferentes alternativas construyendo un **diagrama de barras**, en el que a cada una de dichas alternativas se le hace corresponder una barra cuya altura es proporcional a la frecuencia (absoluta o relativa) con la que la misma ha aparecido.

A continuación se muestra un diagrama de barras relativo al lugar de residencia durante el curso de los alumnos (archivo **curs8990.sf3**).



Alternativamente es posible construir un **diagrama de tarta**, repartiendo la superficie total de un círculo en sectores cuyas áreas sena proporcionales a las frecuencias (absolutas o relativas) observadas en la muestra para cada una de las alternativas posibles de la variable cualitativa estudiada, tal como se aprecia en la siguiente figura.

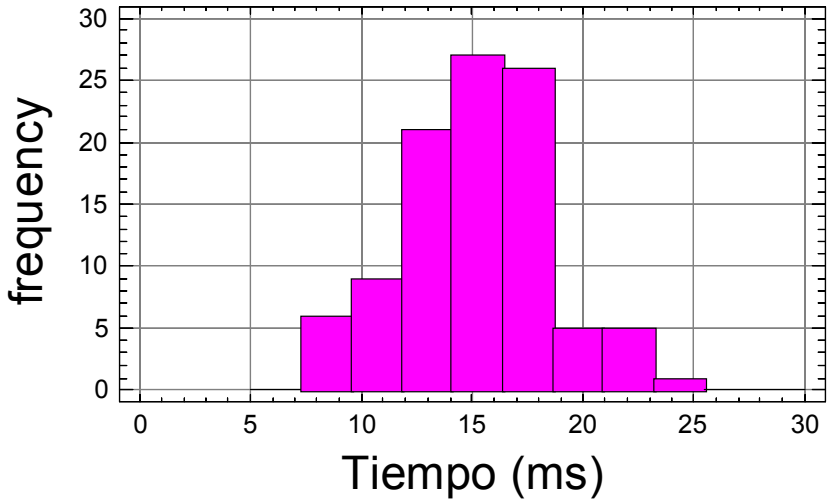


1.4. Histograma

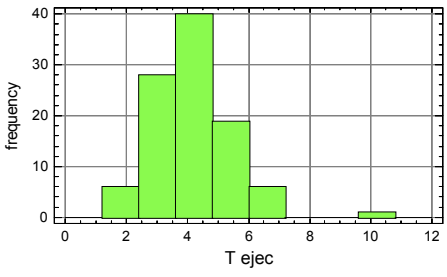
Un **histograma** no es más que una determinada representación gráfica de un conjunto de valores observados de una variable **cuantitativa (o discreta, pero con un número elevado de valores diferentes)**.

En el eje horizontal de las abscisas se representan los valores tomados por la variables en cuestión, agrupados en tramos de la forma habitual si la variables es continua. Sobre cada tramo se levanta una barra de altura proporcional a la frecuencia (es indiferente que sea absoluta o relativa) de valores observados en el tramo considerado.

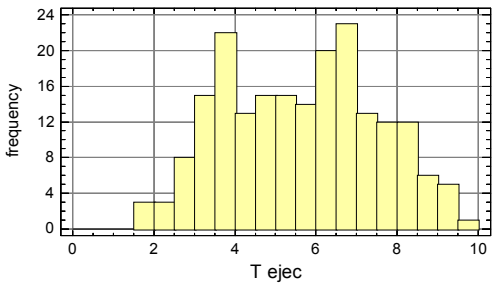
En la siguiente figura se recoge un histograma que representa los tiempos de ejecución (ms) de una muestra de 100 programas.



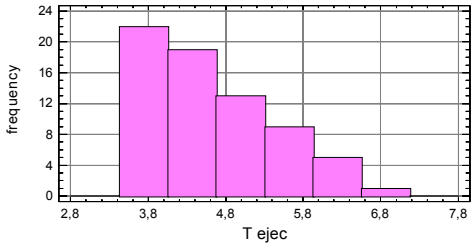
Los histogramas de frecuencias constituyen una poderosa herramienta para el análisis descriptivo de datos, pues permiten muchas veces poner claramente de manifiesto problemas como los que se muestran a continuación:



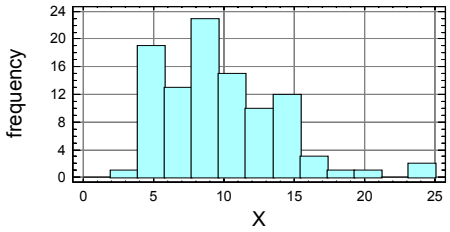
existencia de datos anómalos



mezclas de poblaciones distintas



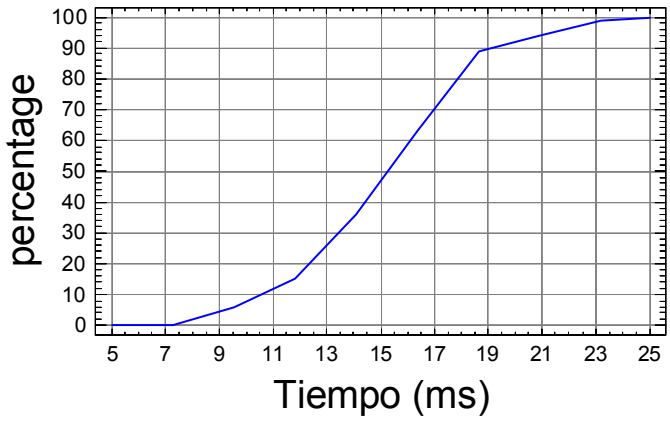
datos artificialmente modificados



no normalidad de los datos

Un mínimo de 40 ó 50 datos es aconsejable para construir un histograma. El número adecuado de tramos depende del tamaño de la muestra. Una regla empírica que conduce a valores razonables es utilizar como número de tramos un entero cercano a la raíz cuadrada del número de datos. En cualquier caso no es frecuente, ni presenta en general ventaja alguna, trazar histogramas con más de 15 ó 20 tramos.

Otro tipo de gráfico que resulta interesante, sobretodo por su relación con conceptos que se verán en la Unidad Didáctica 4, es el **diagrama de frecuencias acumuladas** o **polígono de frecuencias**. En este caso las abscisas levantadas sobre el límite superior de cada intervalo corresponde a la frecuencia acumulada, es decir, a la suma a la suma de las frecuencias consideradas en todos los intervalos anteriores al considerado (incluyendo las de éste). La gráfica, tal como se observa en la figura siguiente, tiene forma de una línea quebrada no decreciente. En general se opera con frecuencias relativas y la altura final es, por tanto, igual a 1 (o a 100 si las frecuencias relativas se expresan en porcentajes).



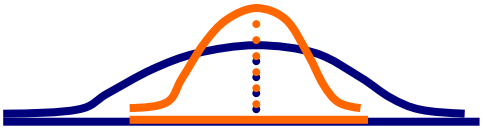
El diagrama de frecuencias acumuladas permite responder directamente a preguntas como ¿qué porcentaje de los programas tienen un tiempo de ejecución menor o igual que 17 ms?

1.5. Parámetros de posición

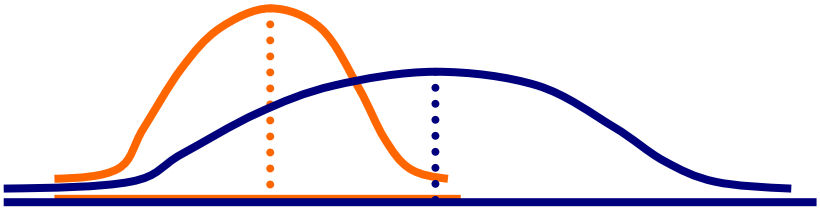
Las tablas y gráficas que acabamos de estudiar contienen la totalidad, o al menos una gran parte, de la información existente en la muestra. Uno de los primeros problemas que

se plantea en Estadística es el de sintetizar esta información, reduciéndola a un número limitado de parámetros más fáciles de manejar y comparar entre sí.

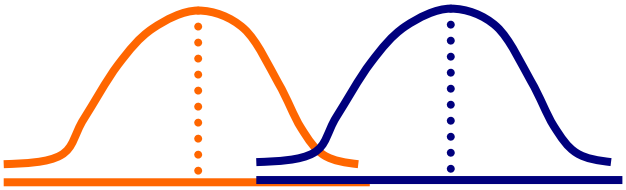
Fundamentalmente la pauta de variabilidad constatada en un conjunto de observaciones relativas a una variable cuantitativa unidimensional puede caracterizarse por dos tipos de parámetros que definan respectivamente la **posición** y la **dispersión** de las observaciones. En las siguientes figuras, en la que hemos sustituido por comodidad los histogramas de frecuencias por curvas continuas, se ve claramente el sentido de ambos términos.



Idéntica posición y distinta dispersión



Distinta dispersión y distinta posición



Idéntica dispersión y distinta posición

Ejercicio 4: Dibujar los histogramas de frecuencias (o unas curvas continuas que los aproximen) para la variable ESTATURA en los jóvenes españoles, diferenciando el relativo a los chicos del de las chicas. Dibujar también ambos histogramas hipotéticos para la variable COEFICIENTE INTELECTUAL.

En el presente apartado nos ocuparemos de los parámetros más utilizados para caracterizar la posición de un conjunto de datos, dejando para el siguiente el estudio de los parámetros de dispersión.

1.5.1. Media

El parámetro de posición más utilizado en la práctica es la **media aritmética** de los datos. Se denota como $\overline{X}$ y su cálculo se realiza mediante la fórmula bien conocida:

Donde N es el número de individuos de la muestra, o sea el número de datos.

$$\overline{X} = \frac{\sum x_i}{N}$$

La media sintetiza la información existente en la totalidad de los datos en un número que da una idea clara sobre la posición de los mismos.

Ejercicio 5: Con el objeto de determinar la calidad de cierto componente electrónico, se ha tomado una muestra de 11 componentes, midiéndose sus tiempos de funcionamiento sin averías (meses). Los resultados en horas son 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52. Obtener el tiempo medio sin averías.

En algunos casos particulares la media puede resultar una medida de posición algo engañosa. Este es el caso en concreto con datos muy asimétricos, en los que unos pocos valores extremos (en general por la cola derecha del histograma) pueden influir excesivamente sobre el valor de la media.

Ejercicio 6: Al preguntar un viajero a un botones de un hotel que propina le daban normalmente, éste respondió que la media de aquel día había sido 10 €. En efecto de los 10 viajeros de aquel día 9 le habían dado 1 € y uno 100 €. La media no era evidentemente en este caso una medida adecuada de la posición de los datos. ¿Cuál considera el alumno que era una medida adecuada de la posición de los datos mencionados?

1.5.2. Mediana

En caso de datos muy asimétricos o con algunos valores extremos puede ser aconsejable usar la mediana como una medida de posición alternativa en vez de la media.

La mediana (M_e) puede definirse intuitivamente como el valor central de los observados. Más precisamente, si se ordenan las N observaciones de menor a mayor la mediana se define como el valor:

- Que ocupa la posición (N+1)/2, si N es impar
- Media entre los valores que ocupan las posiciones N/2 y (N/2)+1, si N es par

Ejercicio 7: ¿cuál sería la mediana de los datos recogidos en el ejemplo mencionado de las propinas al botones del hotel?

Ejercicio 8: calcular la mediana de los datos del Ejercicio 5 (Tiempo de funcionamiento sin averías)

Ejercicio 9: Calcular las medianas de las variables EDAD, ESTATURA, PESO y TIEMPO con los datos de la encuesta y compararlos con las medias respectivas. Constatar la sensible diferencia entre ambos parámetros para la variable TIEMPO, y comprobar mediante un histograma que la distribución de esta variable es muy asimétrica.

1.5.3. Cuartiles

El primer cuartil de un conjunto de datos se puede definir de forma aproximada como el valor C1 tal que la cuarta parte de los datos son inferiores a él y tres cuartas partes de los datos son superiores al mismo. De forma más precisa C1 es el primer cuartil si el número de datos $\leq$ C1 es mayor que N/4 y el número de datos $\geq$ C1 es mayor que 3N/4.

De forma simétrica se define el tercer cuartil C3 como el valor tal que el número de datos $\leq$ C3 es mayor que 3N/4 y el número de valores $\geq$ C3 es mayor que N/4.

Entre los dos cuartiles C1 y C3 se encuentra el 50% central de los datos observados.

Ejercicio 10: Calcular el primer y tercer cuartil de los datos del ejemplo sobre propinas en el hotel.

Ejercicio 11: Calcular el primer y tercer cuartil de los datos del ejemplo sobre Tiempo de funcionamiento sin averías.

Ejercicio 12: Calcular los dos cuartiles de las variables ESTATURA y PESO con los datos de la encuesta. Repetir el cálculo por separado para los chicos y las chicas y comentar los resultados obtenidos.

1.6. Parámetros de dispersión

Como hemos señalado toda población real se caracteriza por la presencia de variabilidad en los valores de las variables que puedan observarse en la misma. Para describir un conjunto de datos estadísticos, y tener en consecuencia una idea sobre la pauta de variabilidad existente en la población de la que procede la muestra, no es suficiente por tanto con disponer de una medida de la posición de dichos datos, sino que es preciso también cuantificar de alguna forma el grado de dispersión existente en los mismos.

Ejercicio 13: ¿Para una persona que no sabe nadar es suficiente saber que la profundidad media de un lago es 1,40 m para lanzarse al baño en el mismo? Por cierto, ¿cuál sería la población y cuál la variable aleatoria en este caso? ¿Aclararía mucho la decisión el conocer además la profundidad mediana del lago?

Intuitivamente la idea de dispersión de un conjunto de datos es bastante clara. El conjunto de datos 3, 3, 3, 3, y 3 tiene una dispersión nula. Los datos 1, 3, 5, 7 y 9 tienen dispersión, pero menos que los datos 1, 5, 10, 15 y 20. ¿Cómo puede precisarse esta idea intuitiva mediante un índice que cuantifique la mayor o menor dispersión de unos datos? Diferentes parámetros pueden utilizarse al respecto.

1.6.1. Recorrido

La medida de dispersión más sencilla para un conjunto de observaciones es el recorrido, que no es más que la diferencia entre el mayor y el menor de los datos. Aunque útil en muestras pequeñas (el recorrido se utiliza frecuentemente en el control de procesos industriales, donde es habitual tomar periódicamente muestras de tamaño 5), el recorrido presenta el inconveniente de que ignora gran parte de la información existente en la muestra, además de depender del tamaño de la muestra (de una misma población, muestras más grandes tendrán en general recorridos más altos que los de muestras pequeñas).

$$R = x_{Max} - x_{Min}$$

Ejercicio 14: Calcular el recorrido de los datos del Tiempo de funcionamiento sin averías.

Ejercicio 15: Suponiendo que, por error, en los datos del Tiempo de funcionamiento sin averías se hubiera anotado el 8º dato como 150 en vez de 62 ¿cuál sería ahora el recorrido?

1.6.2. Varianza. Desviación típica

Dado que la media es en la mayor parte de los casos un buen parámetro de posición, parece lógico tomar como medida de dispersión algún parámetro relacionado con la magnitud de las desviaciones de los datos observados respecto a su media.

El valor medio de estas desviaciones será siempre cero (al anularse las desviaciones positivas con las negativas) por lo que no puede utilizarse como medida de dispersión.

La medida de dispersión más utilizada en Estadística es la denominada **varianza** o, alternativamente, su raíz cuadrada a la que se denomina **desviación típica**.

La varianza no es más que el promedio de los cuadrados de las desviaciones de los datos respecto a su media. Consideraciones teóricas que no son del caso en este momento hacen que en el cálculo de dicho promedio la suma de los cuadrados de las desviaciones se divida por N-1 en vez de por N

$$\text{Varianza} = S^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}$$

En general se prefiere utilizar como medida descriptiva de la dispersión la desviación típica, pues resulta más fácil de interpretar al venir expresada en las mismas unidades que los datos primitivos. Sin embargo las propiedades estadísticas son más sencillas con las varianzas. La desviación típica viene medida en las mismas unidades que los datos primitivos.

$$S = \sqrt{S^2}$$

Ejercicio 16: Calcular la varianza y desviación típica de los datos del ejemplo sobre Tiempo de funcionamiento sin averías.

En algunos casos interesa disponer de un indicador de dispersión que sea adimensional. Un ejemplo lo tendríamos si pretendiésemos comparar la precisión de dos sistemas de medida de ciertas características que den las determinaciones en escalas diferentes. En estas situaciones puede usarse el **coeficiente de variación** que no es más que el cociente entre la desviación típica y la media.

$$CV = \frac{S}{\overline{X}}$$

1.6.3. Intervalo intercuartílico

Por último en aquellos casos en que la media no es un indicador adecuado de posición (como sucede en distribuciones muy asimétricas), tampoco resultará la desviación típica (basada en las desviaciones respecto a la media) un parámetro adecuado de dispersión. En estos casos se utiliza a veces con dicho fin el **intervalo intercuartílico** que no es más que la diferencia entre el tercer y el primer cuartil.

$$\mathbf{II = C_3 - C_1}$$

El intervalo intercuartílico es un **indicador robusto** de dispersión, de la misma forma que la mediana es un indicador robusto de posición, puesto que ambos parámetros resultan poco influidos por la existencia de algún valor anormal (por ejemplo, debido a un error en la introducción de datos) entre las observaciones.

Ejercicio 17: Obtener el intervalo intercuartílico de los datos de los ejercicios 15 y 16 y comentar el resultado.

Ejercicio 18: En los datos de ESTATURA de las chicas modificar un dato poniéndolo en metros en vez de en centímetros. Calcular la media, desviación típica, mediana e intervalo intecuartílico de los nuevos datos de ESTATURA de las chicas y compararlos con los valores que se obtienen tras corregir el dato erróneo. ¿Qué se observa?

1.7. Parámetros de asimetría y curtosis

Como ya se ha comentado las variables aleatorias continuas presentan frecuentemente una pauta de variabilidad que se caracteriza por el hecho de que los datos tienden a acumularse alrededor de un valor central, decreciendo su frecuencia de forma aproximadamente simétrica a medida que se alejan por ambos lados de dicho valor. Ello conduce a histogramas que tienen forma de curva en campana (la famosa campana de Gauss, denominada así en honor del célebre astrónomo que estableció, junto con Laplace, la distribución Normal al estudiar la variabilidad en los errores de sus observaciones)

Para estudiar este tipo de pauta de variabilidad se ha establecido un modelo matemático, la **distribución Normal**, de extraordinaria importancia en toda la Inferencia Estadística. Toda distribución Normal viene completamente caracterizada por su media y su desviación típica, es decir por sus parámetros de posición y de dispersión.

Sin embargo un problema frecuente al estudiar datos reales es, precisamente, analizar hasta qué punto la distribución Normal resulta un modelo adecuado, puesto que pautas de variabilidad que se alejen sensiblemente de la Normal pueden exigir el recurso a tratamientos estadísticos especiales o ser el síntoma de anomalías en los datos.

Con este fin se utilizan los coeficientes de asimetría y de curtosis, que se estudian a continuación.

1.7.1. Coeficiente de asimetría

Si unos datos son simétricos lo son respecto a su media, y la suma de los cubos de las desviaciones de los datos respecto a dicha media $\Sigma(x_i-x)^3$ será nula. Por el contrario dicha suma será positiva si los datos presentan una cola alargada hacia la derecha y negativa si la presentan hacia la izquierda.

Se define el coeficiente de asimetría CA como el promedio (dividiendo por N-1 en vez de por N) de los cubos de las desviaciones respecto a la media, dividido por el cubo de la desviación típica. La división por s^3 tiene por objeto obtener un coeficiente adimensional, o sea que no dependa de la escala en que vengan los datos.

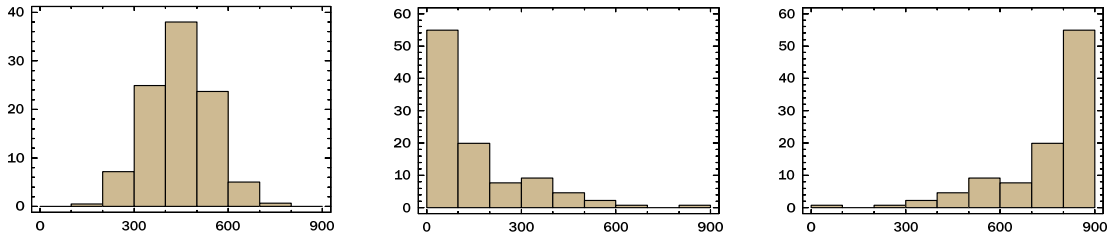
$$CA = \frac{\sum (X_i - \bar{X})^3/(N-1)}{s^3}$$

En contextos inferenciales, cuando el objetivo es analizar hasta qué punto es verosímil que la muestra observada proceda de una población en la que la variable sigue una distribución normal, se utiliza el **coeficiente de asimetría estandarizado (CAE)**. En

muestras que proceden de poblaciones normales, el CAE está comprendido (en el 95% de los casos) entre -2 y 2.

NOTA: el CAE no es más que el CA dividido por una estimación de la desviación típica con la que puede fluctuar en las muestras este coeficiente debido al azar del muestreo.

En la siguiente figura se reflejan los histogramas posibles (simplificando su representación usando curvas continuas) de unos datos simétricos (CA=0), de otros con asimetría positiva (CA > 0) y de otros con asimetría negativa (CA < 0).



Datos simétricos | Asimetría positiva | Asimetría negativa

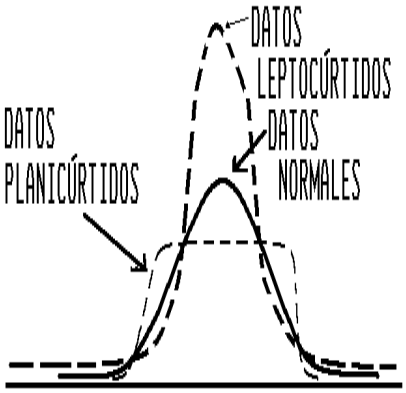
1.7.2. Coeficiente de curtosis

Un conjunto de datos se dice que es **leptocúrtico** si presenta valores muy alejados de la media con mayor frecuencia de la que cabría esperar para unos datos normales que tuvieran la misma desviación típica. Obviamente, para compensar estos valores extremos un histograma de datos leptocúrticos es más apuntado en las cercanías de la media que lo que lo sería el de unos datos normales con la misma desviación típica. Frecuentemente valores elevados de la **curtosis** de un conjunto de datos suele ser síntoma de que entre los mismos se incluyen observaciones anómalas (por ejemplo errores de trascripción o algún individuo perteneciente a una población distinta de la estudiada).

En el otro sentido unos datos se denominan **planicúrticos** si valores alejados de la media aparecen con una frecuencia menor que la que cabría esperar si los datos siguieran una distribución normal con la misma desviación típica. Para compensar este hecho, el histograma de unos datos planicúrticos aparece más plano en el entorno de la media que lo que lo sería el de unos datos normales con idéntica varianza.

Así como la leptocurtosis estaba en general asociada a la presencia de datos anómalos, una planicurtosis excesiva puede revelar que los datos han sido artificialmente censurados para eliminar los valores considerados extremos.

La siguiente figura refleja los histogramas (sustituidos por curvas continuas) de tres distribuciones de datos con idénticas medias y desviaciones típicas, pero que difieren en su curtosis.



El grado de curtosis de un conjunto de datos se mide mediante el **coeficiente de curtosis CC**, que es el cociente entre el promedio (dividiendo por N-1 en vez de por N) de las cuartas potencias de las desviaciones respecto a la media y la desviación típica elevada a 4. En datos que siguen exactamente una distribución normal el coeficiente de curtosis resulta igual a 3, por ello en general el CC se define restando 3 al mencionado cociente:

$$CC = \frac{\sum (X_i - \bar{X})^4 / (N-1)}{s^4} - 3$$

Por tanto un conjunto de datos será leptocúrtico si su CC es mayor que 0 y planicúrtico si su CC es negativo. Obviamente cuanto mas difiere de 3 el coeficiente CC, más acusada es la característica de curtosis correspondiente.

Al igual que sucedía para el coeficiente de asimetría, en contextos inferenciales, cuando el objetivo es analizar hasta qué punto es verosímil que la muestra observada proceda de una población en la que la variable sigue una distribución normal, se utiliza el **coeficiente de curtosis estandarizado (CCE)**. En muestras que proceden de poblaciones normales, el CCE está comprendido (en el 95% de los casos) entre -2 y 2.

NOTA: el CCE no es más que el CC dividido por una estimación de la desviación típica con la que puede fluctuar en las muestras este coeficiente debido al azar del muestreo.

En la siguiente figura se reflejan los histogramas posibles (simplificando su representación usando curvas continuas) de unos datos simétricos (CA=0), de otros con asimetría positiva (CA > 0) y de otros con asimetría negativa (CA < 0).

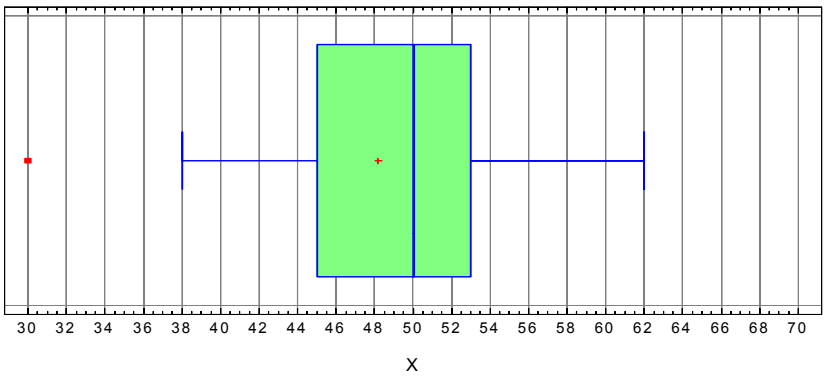
Ejercicio 19: Calcular los coeficientes de asimetría y curtosis de la ESTATURA en chicos y chicas y comparar los resultados obtenidos. Obtener también dichos coeficientes para la variable TIEMPO.

1.8. Diagrama Box-Whisker

Un diagrama Box-Whisker (traducido literalmente "Caja-Bigote") es una representación gráfica sencilla de un conjunto de datos. Presenta, frente a un histograma, la ventaja de no exigir un número elevado de datos para su construcción, además de resultar más sencillo su manejo cuando el objetivo es comparar distintos conjuntos de datos.

La figura adjunta refleja un diagrama Box-Whisker para los valores del Tiempo de funcionamiento sin averías.

La "caja" comprende el 50% de los valores centrales de los datos, extendiéndose entre el primer cuartil y tercer cuartil (45 y 53 en la figura). La línea central corresponde a la mediana (50 en la figura). Los "bigotes" se extienden desde el menor (38) al mayor (62) de los valores observados y considerados "normales". Aquellos valores extremos que difieren del cuartil más próximo en más de 1,5 veces el intervalo intercuartílico, se grafican como puntos aislados (como sucede en la figura con el valor 30) por considerar que pueden corresponder a datos anómalos ("outliers" en la terminología estadística).



Los diagramas Box-Whisker resultan una herramienta extremadamente práctica para la comparación de las pautas de variabilidad existentes en distintos conjuntos de datos.

Ejercicio 20: Comparar la distribución de la ESTATURA entre chicos y chicas mediante los diagramas Box-Whisker correspondientes.

2. ESTADÍSTICA DESCRIPTIVA BIIDIMENSIONAL

2.1. Tablas de frecuencias cruzada

Cuando se está considerando una variable aleatoria bidimensional, un primer análisis de la relación existente entre las dos características en estudio puede llevarse a cabo a partir de la construcción de una tabla de frecuencias cruzada, que recoja la frecuencia con que se ha observado cada combinación de valores posibles de ambas variables. En el caso de que una o ambas de las variables sea de tipo continuo será preciso proceder a un agrupamiento en intervalos de sus valores.

A este tipo de tablas de frecuencias, que son especialmente útiles cuando las dos variables son de naturaleza cualitativa, se les denomina en Estadística **tablas de contingencia**.

La tabla siguiente está obtenida a partir de las respuestas dadas por los alumnos a la encuesta, y cruza las variables **SEXO** y **REPITE.**

REPITE SEXO	SI	NO	Row Total
CHICOS	5 10.9	41 89.1	46 64.8
CHICAS	1 4.0	24 96.0	25 35.2
COLUMN TOTAL	6 8.5	65 91.5	

Como hemos señalado cada casilla recoge el número de individuos que tienen los valores correspondientes para las dos variables (SEXO y REPITE). A la derecha de la tabla se recogen las frecuencias totales, tanto absolutas como relativas (estas últimas expresadas en %), para los dos valores de SEXO. A estas frecuencias se les denomina **frecuencias marginales**. En la parte inferior de la tabla se recogen las frecuencias marginales para la variable REPITE.

Con el fin de estudiar si la proporción de repetidores es similar en los dos sexos conviene calcular la frecuencia relativa de cada casilla respecto al total de la fila correspondiente. Estas frecuencias relativas, que se recogen en la tabla en % en la parte inferior de cada casilla, se denominan **frecuencias relativas condicionales** de REPITE en función de los valores de SEXO.

Es importante que las frecuencias relativas se calculen adecuadamente respecto al total de la fila o de la columna correspondiente, según sea relevante para los objetivos perseguidos en un determinado estudio.

Ejercicio 21: Calcular a partir de los datos de la tabla anterior las frecuencias relativas condicionales de SEXO frente a REPITE. ¿Cuál de los dos conjuntos de frecuencias condicionales consideras que puede prestarse a una interpretación más interesante?

Cuando sobre cada individuo de la población se observan **dos** características aleatorias de naturaleza cuantitativa, se tiene una variable aleatoria bidimensional cuantitativa.

Ejemplo 1: en la población constituida por los estudiantes universitarios españoles se observa la ESTATURA (cms) y el PESO (kgs) de cada estudiante. Una muestra de esta variable bidimensional puede estar constituida por los 130 pares de valores constatados en los 130 alumnos que respondieron a la encuesta

Ejemplo 2: para el control del consumo de energía en calefacción en una factoría durante los meses de invierno se anota diariamente el CONSUMO (termias) y la TEMPERATURA diaria (ºC a las 12). Una muestra de esta variable bidimensional puede estar constituida por los 57 pares de valores de CONSUMO y TEMPERATURA constatados en 57 días laborables del invierno de 1985.

En el Apartado anterior se expuso cómo podía describirse, mediante una Tabla de Contingencia, la relación entre las dos componentes de una variable bidimensional en el caso de que ambas fueran de tipo **cualitativa.**

Cuando las dos variables sean de tipo **cuantitativo**, y especialmente cuando se trate de variables **continuas** (como sucede en los dos ejemplos anteriores) es posible utilizar técnicas más adecuadas para describir y analizar la relación existente entre ambas.

NOTA: utilizaremos frecuentemente la expresión "dos variables aleatorias" por ser más cómoda que "dos componentes de la variable aleatoria bidimensional", aunque ésta última es más correcta

Por supuesto es posible, en primer lugar, construir una **tabla de frecuencias cruzada** entre las dos variables, aunque será necesario previamente agruparlas en intervalos.

La siguiente tabla (construida mediante el Statgraphics previa una recodificación de las variables) refleja las frecuencias observadas para cada combinación de tramos de ESTATURA y PESO.

ESTATURA PESO	145 155	155 165	165 175	175 185	185 195	Row Total
40 55	9 75.0	17 44.7	0 .0	0 .0	0 .0	26 20.0
55 70	3 25.0	18 47.4	31 53.4	5 29.4	0 .0	57 43.8
70 85	0 .0	3 7.9	24 41.4	12 70.6	3 60.0	42 32.3
85 99	0 .0	0 .0	3 5.2	0 .0	2 40.0	5 3.8
Column Total	12 9.2	38 29.2	58 44.6	17 13.1	5 3.8	130 100

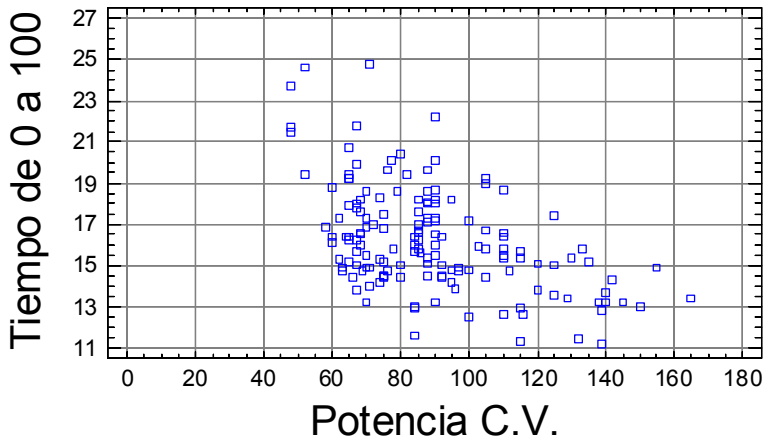
En el margen derecho se recogen las frecuencias (absolutas y relativas, estas últimas expresadas como %) de los 4 tramos considerados para PESO. Estas frecuencias, que están obtenidas sumando para todos los valores posibles de ESTATURA se denominan **marginales**.

Dentro de cada columna se recogen las frecuencias observadas para los diferentes tramos de PESO en los individuos cuya ESTATURA se halla en el tramo considerado. Las frecuencias relativas están calculadas respecto a la frecuencia total de la columna considerada y se denominan **frecuencias relativas condicionales**. Así de los individuos cuya ESTATURA está en el tramo 145-155 el 75% pesan entre 40 y 55 Kg. y el 25% entre 55 y 70 Kg., mientras que de los que miden entre 175 y 185 cm. el 29.4% pesan entre 55 y 70 Kg. y el 70.6% pesan entre 70 y 85 Kg.

2.2. Diagramas de dispersión

Una forma sencilla de describir gráficamente las relaciones constatadas entre dos variables consiste en representar cada observación por un punto en un plano cuya abscisa sea el valor de la primera variable y cuya ordenada sea el de la segunda. A este tipo de gráfico se le denomina diagrama de dispersión.

La siguiente figura refleja el diagrama de dispersión de la variable ACELERACIÓN (medida como tiempo de 0 a 100 km/h) frente a POTENCIA (C.V.) de una muestra de coches de diferentes marcas y modelos.



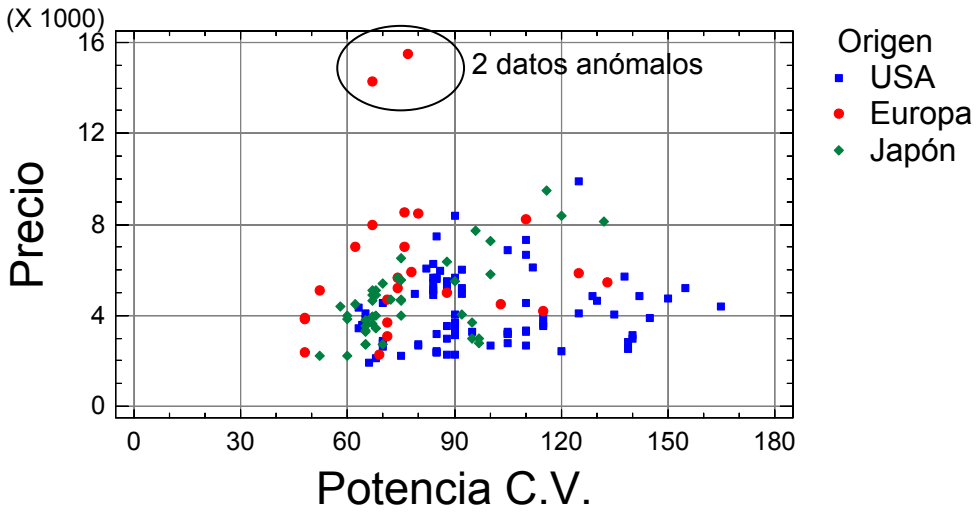
El diagrama pone claramente de manifiesto una relación negativa (inversa) entre las dos variables estudiadas, que se refleja en una nube de puntos cuyo eje principal tiene un sentido decreciente, como consecuencia del hecho de que, en términos generales, los coches más potentes tienen una mayor aceleración (menor tiempo de 0 a 100) que los menos potentes.

Ejercicio 22: Para estudiar un ejemplo en el que el Diagrama de Dispersión pone claramente en evidencia una relación negativa entre dos variables obtener el diagrama para las variables TEMPER y CONSUMO del fichero GAS.

En general cuanto más estrechamente se agrupen los puntos del diagrama de dispersión alrededor de una recta más fuerte es el grado de relación lineal existente entre las dos variables consideradas.

Otra característica de este tipo de gráficos es que nos permite representar dos variables en función del valor que tome una tercera de tipo cualitativo, así como

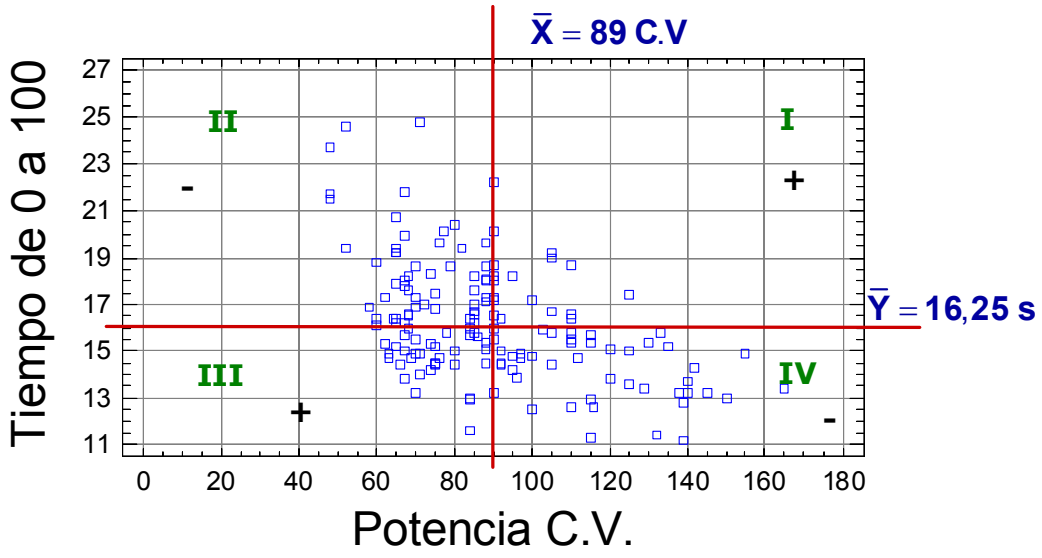
detectar datos anómalos. Por ejemplo, el diagrama de dispersión que se muestra a continuación relaciona la variable PRECIO (miles de $) frente a POTENCIA (C.V.), pero los puntos se han codificado según la procedencia de los mismos (ORIGEN):



2.3. Covarianza y Coeficiente de Correlación

Con el fin de cuantificar en un índice numérico el grado de relación lineal existente entre dos variables se utilizan en Estadística dos parámetros: la covarianza y el coeficiente de correlación.

Para dar una idea intuitiva del concepto de covarianza vamos a razonar sobre el siguiente diagrama de dispersión, correspondiente a las variables POTENCIA (C.V) y ACELERACIÓN (tiempo en segundos de 0 a 100), en el que hemos trazado una línea horizontal a la altura del valor medio de la segunda variable (16,25 s) y una línea vertical situada sobre el valor medio de la primera variable (89 C.V.)



En este caso, en el que existe claramente una fuerte relación negativa, la mayor parte de los puntos caen en los cuadrantes 2 y 4. Por el contrario cuando la relación existente sea positiva la mayoría de los puntos caerán en los cuadrantes 1 y 3.

Si consideramos el signo que para cada punto x_i,y_i del diagrama tiene el producto $(x_i\text{-media}_X)(y_i\text{- media}_Y)$ vemos que éste resulta positivo en los cuadrantes 1 y 3 y negativo en los cuadrantes 2 y 4. Por lo tanto el producto anterior será **en promedio** positivo si existe una relación creciente (o sea positiva) entre las dos variables (es decir si la Y tiende a crecer cuando lo hace la X) y negativo si la relación existente es decreciente (o sea negativa).

Por definición la **covarianza** entre dos variables no es más que el promedio de los productos de las desviaciones de ambas variables respecto a sus medias respectivas.

NOTA: por consideraciones que no son del caso, y de forma similar a como se procedió al definir la varianza, el promedio se calcula dividiendo por N-1 en vez de por N

$$\mathbf{COV_{(X,Y)} = S_{X,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}}$$

La covarianza presenta el inconveniente de que depende de las dimensiones en que se expresan las variables. Así la covarianza entre POTENCIA y ACELERACIÓN será 1000 veces más pequeña si la segunda variable se mide en milisegundos que si se mide en segundos.

Para obviar este problema se utiliza universalmente en Estadística, como medida del grado de relación lineal existente entre dos variables, el **coeficiente de correlación lineal** que no es más que la covarianza dividida por el producto de las desviaciones típicas de las dos variables.

$$\mathbf{r_{X,Y} = \frac{COV_{X,Y}}{S_X S_Y}}$$

Se demuestra fácilmente que el coeficiente de correlación entre dos variables se mantiene inalterable si cualquiera de ellas sufre una transformación lineal.

El coeficiente de correlación lineal (**r**) tiene una serie de propiedades que lo hacen especialmente adecuado para medir el grado de correlación (lineal) entre dos variables:

- **r** está siempre comprendido entre -1 y +1.
- Los valores extremos, -1 y +1, sólo se alcanzan si existe una relación lineal exacta entre la Y y la X, o sea, si los puntos del diagrama de dispersión están exactamente alineados en una recta. (el valor +1 se obtiene si la recta es creciente y el – 1 si es decreciente)

- Cuando no existe relación alguna entre las dos variables, **r** = 0. En la práctica será "cercano" a cero debido al azar del muestreo.
- Cuanto más estrecho es el grado de relación lineal existente entre dos variables más cercano a 1 es el valor de **r** (o a -1 si la relación es decreciente). Por el contrario un valor de **r** nulo o cercano a 0 indicará una relación lineal inexistente o muy débil.

Ejercicio 23: Calcular los coeficientes de correlación entre ESTATURA y PESO, entre EDAD y ESTATURA y entre TEMPER y CONSUMO y contrastarlos con el aspecto de los diagramas de dispersión correspondientes. ¿Hasta qué punto las diferencias de peso entre los alumnos están asociadas a las diferencias de estatura entre ellos?

Es importante resaltar que tanto la covarianza como el coeficiente de correlación miden sólo el grado de relación lineal existente entre dos variables. Dos variables pueden tener una relación estrecha y sin embargo resultar **r** cercano a cero por ser dicha relación no lineal.

Ejercicio 24: Introducir, usando el Statgraphics dos variables: una X de valores -3,-2,-1, 0,1,2,3 y otra Y de valores 9,4,1,0,1,4,9. Dibujar el Diagrama de Dispersión y hallar el coeficiente de correlación entre ambas. ¿Están relacionadas las variables? ¿Lo están linealmente?

Ejercicios resueltos

Apartado 2.A.1 del Capítulo 2 del libro de R. Romero y L.R. Zúnica "Métodos Estadísticos en Ingeniería" SPUPV 637

Ver boletín correspondiente en PoliformaT (EST GII: Recursos / 04 | Ejercicios)

Para saber más

- Descartes: Terminología estadística y representaciones gráficas. Instituto de Tecnologías Educativas (ITE) del Ministerio de Educación: http://recursostic.educacion.es/descartes/web/materiales_didacticos/Estadistica_descriptiva_1/estadistica_indice.htm#obje

- Descartes: **Variables estadísticas, Tablas de frecuencias y Gráficos**. Instituto de Tecnologías Educativas (ITE) del Ministerio de Educación: http://recursostic.educacion.es/descartes/web/materiales_didacticos/iniciacion_estadististica_fjgarcia/FGG990_UD.htm

- **Adivina qué valores tendrán los parámetros a partir de un histograma. Rice Virtual Lab in Statistics**: http://www.ruf.rice.edu/~lane/stat_sim/descriptive/index.html

- **Correlation. VESTAC.** Java Applets for Visualization of Statistical Concepts (**Basics-Correlation**): http://www.kuleuven.ac.be/ucs/java/gent/Ap7.html

- Adivina cuál es el coeficiente de correlación de las nubes de puntos que genera aleatoriamente este applet de Java : http://goo.gl/8pog

- Análisis de regresión lineal con Excel : http://goo.gl/VW8S

- Instrucciones para algunas calculadoras Casio : http://goo.gl/XXPN

- A propósito de las relaciones causa-efecto, y como curiosidad, un poco de latin: http://goo.gl/YkNw

- En el episodio 1 de la 2.a temporada de Numb3rs se habla de diagramas de dispersion : http://goo.gl/MUkC

Fuentes

Métodos Estadísticos en Ingeniería (Romero Villafranca, Rafael) | Material docente previo de V. Giner (DEIOAC) | Material docente previo de R. Alcover (DEIOAC) | Material docente previo de A. Calduch (DEIOAC) | Material docente previo de J. R. Navarro (DEIOAC) | Material docente previo de Carmen Capilla (DEIOAC) |Material docente previo de E. Vázquez (DEIOAC)