CAPÍTULO 2
ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

2.1 INTRODUCCIÓN

Como ya se comentó en el Capítulo 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.

El objetivo de este capítulo es familiarizar al alumno con algunas técnicas, sencillas pero poderosas, de Estadística Descriptiva para el estudio de variables unidimensionales

Tras exponerse cómo puede sintetizarse la información contenida en un conjunto de datos mediante una tabulación adecuada o mediante un histograma, se estudian los diferentes tipos de parámetros (posición, dispersión, asimetría y curtosis) que pueden utilizarse para caracterizar la pauta de variabilidad constatada en las observaciones. También se introduce un tipo de representación gráfica, los diagramas Box-Whisker, de gran utilidad práctica.

2.2 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.

a) Cuando la variable estudiada es de tipo 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 la gravedad de un ataque de mosca del mediterráneo a partir del número de picadas constatado en 200 naranjas

Picadas (X_i)	Numero de naranjas (ni)	Frecuencia relativa fi = ni / N
0	48	0.24
1	106	0.53
2	32	0.16
3	14	0.07

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.

b) 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 normalmente 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 la longitud y la anchura 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.

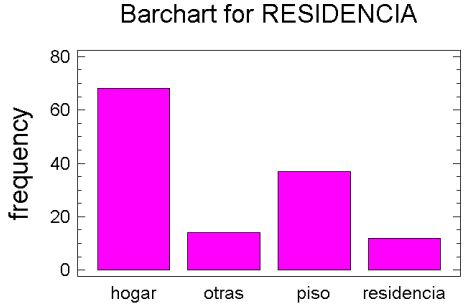
Autoevaluación: reflexionar sobre la afirmación de que si la variable es continua y el sistema de medida fuera suficientemente preciso la probabilidad de encontrar dos valores repetidos es prácticamente nula.
¿Por qué se pierde mucha información en la tabulación si el número de intervalos considerado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

Autoevaluación: Tabular los valores constatados para las variables DIGITO, POLITICA y PROBLEMA en la encuesta realizada. Estudiar en particular la frecuencia con la que aparecen los distintos dígitos. A la vista de los resultados ¿parece admisible la hipótesis de que cuando se enuncian dígitos supuestamente "al azar" los pares aparecen con la misma frecuencia que los impares?

2.3 DIAGRAMAS DE BARRAS Y DE TARTA

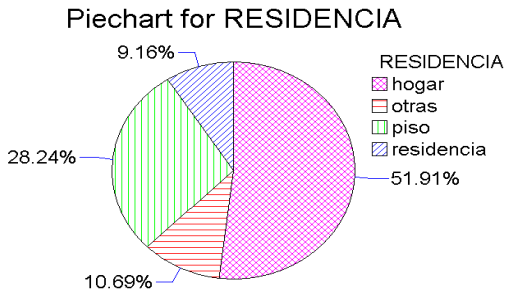
En el análisis descriptivo de variables de naturaleza cualitativa, es muy habitual representar las frecuencias con las que se han presentado en la muestra las diferentes alternativas construyendo un **diagrama de barras**, en el que a cada una de dichas alternativas se le hace corresponder una barra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

En la siguiente figura se recoge 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 sean proporcionales a las frecuencias

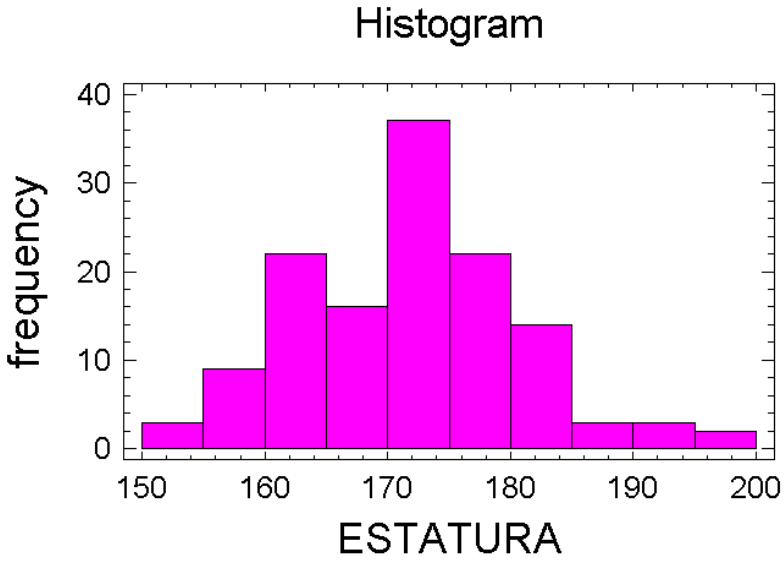
observadas en la muestra para cada una de las alternativas posibles de la variable cualitativa estudiada, tal como se aprecia en la siguiente figura



2.4 HISTOGRAMAS

Un histograma es una representación gráfica de un conjunto de valores observados de una variable **cuantitativa** continua (o discreta pero con un número elevado de valores diferentes). En el eje horizontal de las abscisas se representan los valores tomados por la variable en cuestión, agrupados en tramos de la forma habitual. 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 mismo.

La siguiente figura recoge el histograma correspondiente a las estaturas de los alumnos que respondieron a la encuesta (archivo curs8990.sf3)



Los histogramas de frecuencias constituyen una poderosa herramienta para el análisis descriptivo de datos, pues permiten muchas veces poner claramente de manifiesto problemas como

- existencia de datos anómalos
- mezclas de poblaciones distintas
- datos artificialmente modificados
- No normalidad de los datos
- etcétera...

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.

Autoevaluación: Una determinada dimensión generada en el mecanizado de unas piezas debe diferir como máximo en 5 unidades del valor nominal. Los datos reflejados en un gráfico de control referidos a 100 piezas y medidos en diferencias respecto al nominal, son los siguientes:

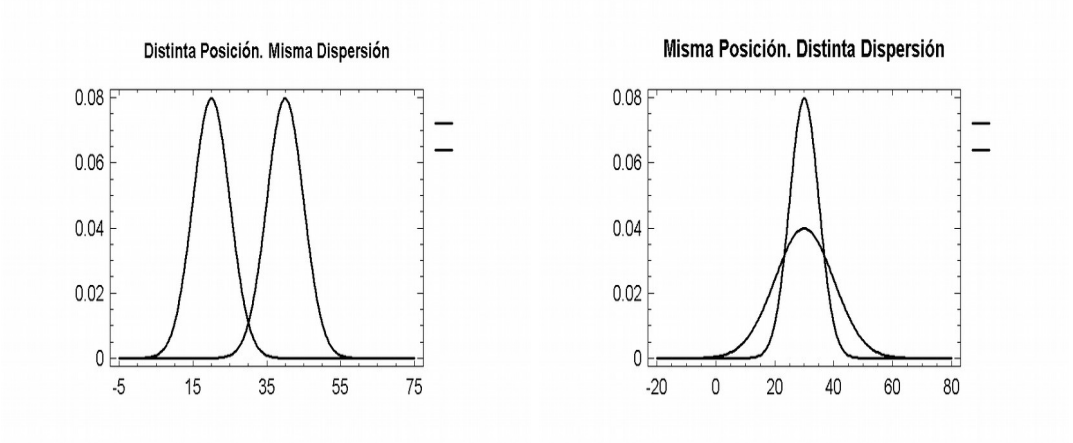
-5 -3 0 3 5 0 4 4 0 0 5 3 2 5 -2 5 2 4 -2 5 0 2 0 3 0 5 5 4 7 -1 4 4 -2 0 5 4 3 2 -2
1 2 0 3 1 0 -2 3 2 -1 -3 0 2 2 0 0 3 2 3 0 0 4 0 0 -4 0 0 0 0 5 3 0 -3 0 0 -2 5 -2 0 1
0 -3 5 1 2 4 5 3 -3 5 -1 3 0 3 4 4 -4 0 0 0

Obtener un histograma de los datos anteriores y discutir las conclusiones que se deducen del mismo. (Ver respuesta en el Anejo al final del Tema)

2.5 PARAMETROS 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 la siguiente figura, en la que hemos sustituido por comodidad los histogramas de frecuencias por curvas continuas, se ve claramente el sentido de ambos términos.



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.

2.5.1 Media

El parámetro de posición más utilizado en la práctica es la media aritmética de los datos. Su cálculo se realiza mediante la fórmula bien conocida:

$$\text{media muestral} \quad \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

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

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.

Autoevaluación: Calcular y comparar la media de ESTATURA y de PESO de chicos y chicas a partir de los resultados de la encuesta realizada en clase

La media tiene una serie de propiedades que la hacen especialmente idónea como medida de posición, y que pueden deducirse de forma inmediata a partir de su definición. Por ejemplo si una variable Y es una transformada lineal de otra variable X (Y=a+bX), la media de Y resulta ser la misma transformada lineal de la media de X ($\bar{y} = a + b\bar{x}$).

También si una variable Z es la suma de las dos componentes X e Y de una variable bidimensional, la media de Z resulta igual a la suma de las medias de X e Y (Z=X+Y $\Rightarrow$ $\bar{z} = \bar{x} + \bar{y}$). Esta propiedad se generaliza de forma inmediata a la suma de K variables (expresado con más propiedad: a la suma de las K componentes de una variable K-dimensional)

Sin embargo, 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.

2.5.2 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 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 $\frac{N+1}{2}$ si N es impar
- media entre los valores que ocupan las posiciones $\frac{N}{2}$ y $\frac{N}{2}+1$ si N es par

Autoevaluación: En una empresa de 500 operarios se considera la variable salario mensual de cada empleado; ¿qué serían en este ejemplo la media y la mediana de los datos?

Autoevaluación: 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.

Autoevaluación: La LD50 de un insecticida (ver ejemplo en la tercera Autoevaluación del apartado 2.2.2.3) ¿a qué parámetro de la distribución de la variable considerada corresponde? ¿Qué otro

parámetro podría haberse definido en vez de la LD50? ¿Por qué consideras que se eligió en su día el primer parámetro? (Ver respuesta en el Anejo al final del Tema)

2.5.3 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 simétrica se define el tercer cuartil C3 como el valor tal que tres cuartas partes de los datos son inferiores a él y una cuarta parte de los datos son superiores al mismo.

Una forma sencilla de calcular los cuartiles C1 y C3 es hallando las medianas de la mitad inferior y de la mitad superior de los datos.

De acuerdo con la definición dada, entre los dos cuartiles C1 y C3 se encuentra el 50% central de los datos observados.

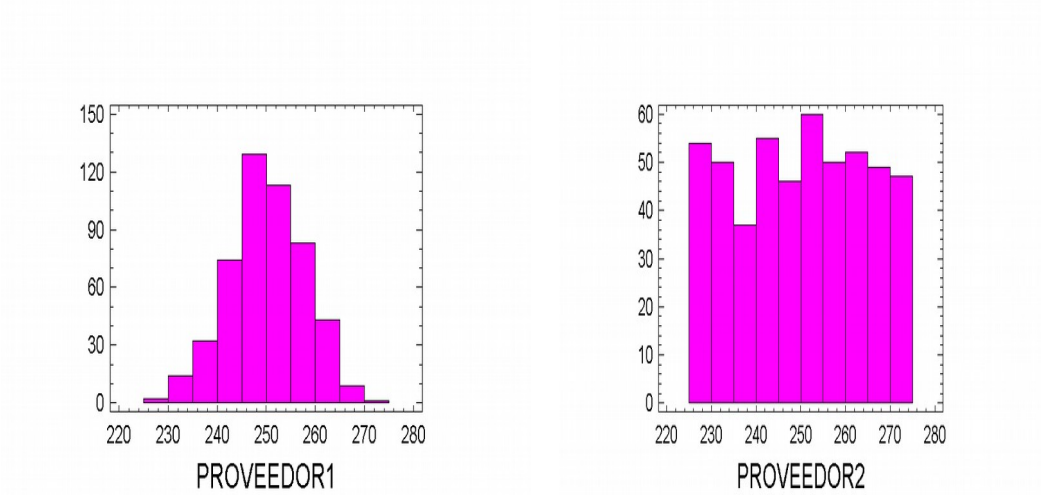
Autoevaluación: Calcular el primer y tercer cuartil de los datos del ejemplo sobre sobre mecanizado de cierta pieza.
Calcular los dos cuartiles de las variables ESTATURA y PESO con los datos de la encuesta. Repetir el cálculo por separado para los chicos y las chicas y analizar los resultados obtenidos

2.6 PARÁMETROS DE DISPERSIÓN

Como hemos señalado toda población real se caracteriza por la presencia de variabilidad en los valores de la misma en los individuos de la población. 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.

Autoevaluación: Para una persona que no sabe nadar ¿es suficiente saber que la profundidad media de un lago es 1,40 mts 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?

Autoevaluación: Como otro ejemplo de la importancia del concepto de dispersión supongamos que una empresa automovilística ha determinado mediante estudios ergonómicos que la dureza óptima de los asientos es 250 newtons. A sus posibles proveedores de asientos les exige que la dureza de los asientos que le vendan no difieran en más de un 10% de dicho valor, o sea que esté comprendida entre 225 y 275 newtons. La dureza de los asientos ofrecidos por dos posibles proveedores presenta una pauta de variabilidad sintetizada por los siguientes histogramas:



En ambos casos los proveedores cumplen las especificaciones de la empresa poseyendo las dos variables consideradas la misma media deseada de 250 nwts. ¿Puede considerarse que la elección entre ambos proveedores es por tanto irrelevante? ¿En qué difieren las pautas de variabilidad de las durezas entre ambos proveedores? ¿Cuál resulta preferible? ¿Por qué?

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

2.6.1 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 más pequeñas)

2.6.2 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.

Varianza: $s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$ Desviacion Tipica: $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$

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.

Se demuestra fácilmente que si $Y = a + bX \Rightarrow s^2(Y) = b^2 s^2(X)$

Sin embargo, a diferencia de lo que sucedía con la media, la varianza de la suma de variables aleatorias no es en general igual a la suma de sus varianzas. Como veremos más adelante en este texto (apartado 5.7.1) sí que se cumple, sin embargo, que cuando en una población dos variables aleatorias X_1 y X_2 son independientes (concepto que se precisará más adelante en el capítulo 4) la varianza de su suma es igual la suma de sus varianzas.

Muy frecuentemente las variables aleatorias reales siguen pautas de variabilidad que se caracterizan por histogramas que se asemejan a campanas aproximadamente simétricas. La Estadística ha establecido un modelo matemático de este tipo de variables aleatorias, la denominada Distribución Normal o de Gauss (Esta distribución se estudia más adelante en el Capítulo 6). En datos que siguen una distribución normal se cumplen aproximadamente las siguientes propiedades:

- Dos tercios de los datos difieren de la media menos de S
- El 95% de los datos difiere de la media menos de 2S
- La práctica totalidad de los datos (en teoría más del 99.7%) difieren de la media menos de 3S

Autoevaluación: Utilizando el Statgraphics comprobar si se cumplen aproximadamente los valores anteriores con los datos de las variables PESO, analizando sólo los datos de los chicos.

La desviación típica viene medida en las mismas unidades que los datos primitivos. 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, expresándose generalmente en porcentaje

$$\text{Coeficiente de Variacion: } CV = \frac{s}{\bar{x}} 100$$

2.6.3 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.

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.

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

2.7 PARÁMETROS DE ASIMETRÍA Y DE 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.

2.7.1 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 $\sum_{i=1}^{N}(x_i - \overline{x})^3$ será nula, al compensarse valores positivos y negativos. Por el contrario dicha suma será positiva si los datos presentan una cola alargada hacia la derecha (pues los datos muy alejados de la media por la derecha contribuyen a la suma anterior con valores positivos muy grandes, al estar elevados al cubo) y será negativa (por un motivo análogo) si la cola alargada se presenta hacia la izquierda.

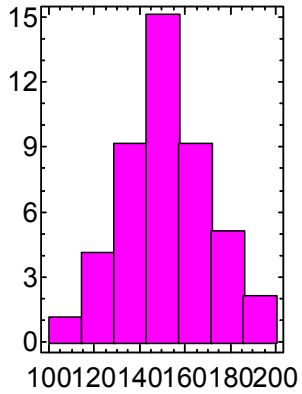
Se define el coeficiente de asimetría CA como el promedio (dividiendo por N-1 en vez de por N) de los cubos de las desviaciones respecto a la media, dividido por el cubo de

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.

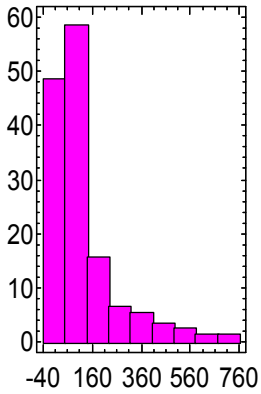
$$\text{Coeficiente de Asimetria: } CA=\frac{\sum (x_i-\bar{x})^3 \big/ (N-1)}{s^3}$$

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**, que nos es más que 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 muestras que proceden de poblaciones normales, este coeficiente de asimetría estandarizado está comprendido (en el 95% de los casos) entre -2 y +2.

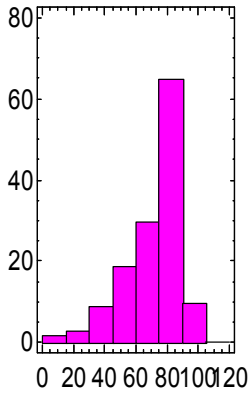
En la siguiente figura se reflejan los histogramas posibles de unos datos simétricos (CA=0), de otros con asimetría positiva (CA mayor que 0) y de otros con asimetría negativa (CA menor que 0).



Datos Simétricos



Datos con Asimetría Positiva



Datos con Asimetría Negativa

2.7.2 Coeficiente de Curtosis

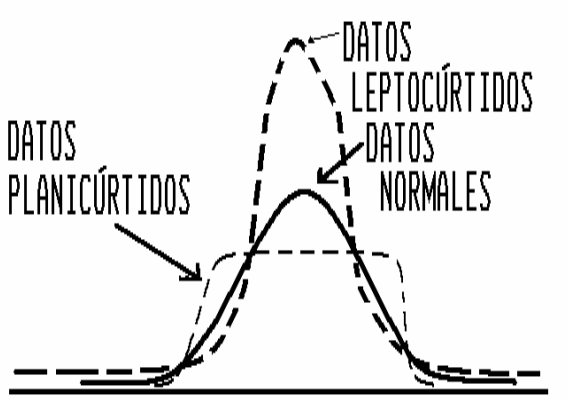
Un conjunto de datos se dice que es **leptocúrtico** si presenta valores muy alejados de la media con mayor frecuencia de la que cabría esperar para unos datos normales que tuvieran la misma desviación típica. Obviamente, para compensar estos valores extremos, un histograma de datos leptocúrticos es 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. También la mezcla en una misma muestra de datos procedentes de dos poblaciones con diferentes medias, produce histogramas planicúrticos.

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 se basa en 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 este cociente resulta igual a 3. Por ello en general el coeficiente de curtosis 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 mayor sea en valor absoluto el coeficiente CC, más acusada es la característica de curtosis correspondiente.

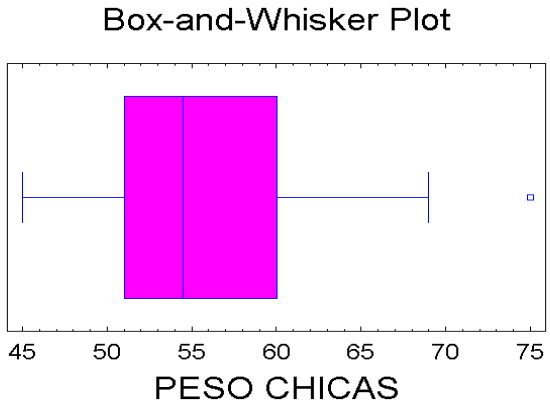
Al igual que sucedía para el coeficiente de asimetría, en contextos inferenciales, cuando el objetivo es analizar hasta qué punto es verosímil que la muestra observada proceda de una población en la que la variable sigue una distribución normal, se utiliza el **coeficiente de curtosis estandarizado**,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

Autoevaluación: 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

2.8 DIAGRAMAS BOX-WHISKER

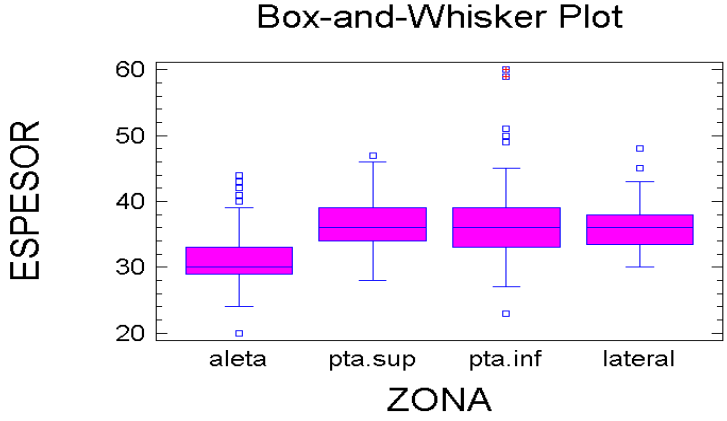
Un diagrama Box-Whisker (traducido literalmente "Caja-Bigote") es una representación gráfica sencilla de un conjunto de datos. Presenta, frente a un histograma, la ventaja de no exigir un número elevado de datos para su construcción, además de resultar más sencillo su manejo cuando el objetivo es comparar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

La figura adjunta refleja un diagrama Box-Whisker para los valores de la variable PESO en las chicas (archivo curs8990.sf3)



La "caja" comprende el 50% de los valores centrales de los datos, extendiéndose entre el primer cuartil y tercer cuartil (51 y 60 en la figura). La línea central corresponde a la mediana (54.5 en la figura). Los "bigotes" se extienden desde el menor (45) al mayor (69) 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 75) por considerar que pueden corresponder a datos anómalos ("outliers" en la terminología estadística). En este ejemplo concreto el outlier corresponde realmente al peso de un chico que se equivocó al poner su código de sexo al rellenar la encuesta

Los diagramas Box-Whisker resultan una herramienta extremadamente práctica para la comparación de las pautas de variabilidad existentes en distintos conjuntos de datos. Como ejemplo, en la siguiente figura se representan en un diagrama Box-Whisker múltiple (posicionado verticalmente) los espesores de imprimación medidos en la planta de pintura de una factoría automovilística, diferenciados según la zona del coche en la que se miden



Se aprecia claramente que la pauta de variabilidad del espesor es claramente diferente en la aleta delantera, con valores sensiblemente más bajos que en el resto del vehículo. También se aprecia la existencia de algunos valores anómalamente altos en la zona inferior de la puerta (debidos probablemente a goteo de la pintura)

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

2.A AUTOEVALUACIONES RESUELTAS Y EJERCICIOS

2.A.1 Respuesta a algunas Autoevaluaciones

Autoevaluación: El contenido en zumo y el calibre de las naranjas de un huerto ¿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 creatinina 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?

En primer lugar hay que ver con claridad cuál es la población implicada en cada caso. Los individuos de esta población son los entes sobre los que se miden o constatan la variable, o variables, consideradas. Por lo tanto se trata de ver si las dos variables planteadas en cada caso se miden o no sobre el mismo individuo. Sólo en este caso se tratará de una variable bidimensional y tendrá sentido plantearse, por ejemplo, si las dos variables están o no relacionadas (cuestiones del tipo: ¿en los individuos en los que la primera variable tiene valores altos, la segunda variable suele también tener valores altos?)

En el primer ejemplo la población la constituyen las naranjas y sobre cada una de ellas se pueden medir las dos variables $\Rightarrow$ variable bidimensional

En el segundo ejemplo la población la constituyen los programas preparados por la empresa y sobre cada uno de ellos se puede contar el número de líneas de código y el número de errores $\Rightarrow$ variable bidimensional

En el tercer ejemplo, sin embargo, no hay una población en la que sobre cada individuo se puedan medir las dos variables indicadas. Hay realmente dos poblaciones, la de individuos alcohólicos y la de individuos no alcohólicos, y en los individuos de cada una de ellas se puede medir una variable (que se define igual en ambos casos). Se trata, por tanto, de dos variables unidimensionales.

También podría plantearse este caso como la existencia de una única población (la formada por todos los individuos) en la que se "miden" dos variables: el contenido de creatinina en sangre (variable cuantitativa), y una variable cualitativa con dos alternativas que indica si el individuo es o no alcohólico y que divide la población inicial en dos subpoblaciones, entre las cuales puede interesar comparar la pauta de variabilidad de la primera variable.

En el cuarto ejemplo la población la constituyen los matrimonios jóvenes (habría que precisar bien esta definición) y sobre cada uno de ellos se puede constatar la edad del marido y la de la mujer $\Rightarrow$ variable bidimensional

Autoevaluación: En el estudio de insecticidas se define la LD50 (Dosis Letal 50) de un producto como aquella dosis mínima que administrada a ratas provoca la muerte al 50% de las mismas. Al estudiar la LD50 de un determinado producto: ¿Cuál es la población implicada, y cual la variable aleatoria considerada?

¿Donde radica la variabilidad en este contexto? En el hecho de que hay ratas que resisten dosis mayores y otras que resisten dosis menores. Por tanto la población implicada sería la población de las ratas, y la variable aleatoria en cuestión sería la dosis mínima que administrada a la rata considerada provoca su muerte.

(Observar, por cierto, que es difícil medir experimentalmente esta variable. En general se dará a una rata una dosis D; si la rata muere es que el valor X de la variable considerada en la rata en cuestión es menor que D, mientras que si sobrevive dicho valor X será mayor que D. Se trata éste de una situación que exige tratamientos estadísticos avanzados especiales, dado que los datos están "censurado", en el sentido de que en vez de conocerse, como sucede habitualmente, los valores de la variable estudiada en los individuos de la muestra, lo único que se conocen son ciertas cotas de estos valores)

Autoevaluación: En una factoría interesa cuantificar, con el fin de controlar el consumo de energía (utilizada en su mayor parte en la climatización de las naves), la relación existente entre el consumo diario de electricidad y la temperatura media del día correspondiente. ¿Cuál es en el contexto anterior la población implicada y la variable aleatoria considerada?

La población implicada estaría constituida por los días (posiblemente habrá que concretarla refiriéndose sólo a días laborables del periodo invernal en el que funciona la calefacción en la fábrica) sobre los que se mide la variable bidimensional TEMPERATURA (por ejemplo, temperatura a las 12 del mediodía), y CONSUMO de energía dicho día.

Autoevaluación: Una determinada dimensión generada en el mecanizado de unas piezas debe diferir como máximo en 5 unidades del valor nominal. Los datos reflejados en un gráfico de control referidos a 100 piezas y medidos en diferencias respecto al nominal, son los siguientes:

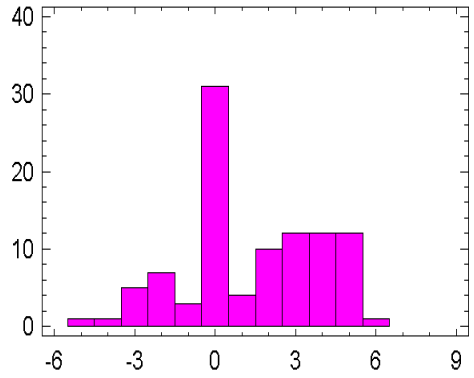
-5 -3 0 3 5 0 4 4 0 0 5 3 2 5 -2 5 2 4 -2 5 0 2 0 3 0 5 5 4 7 -1 4 4 -2 0 5 4 3 2 -2
1 2 0 3 1 0 -2 3 2 -1 -3 0 2 2 0 0 3 2 3 0 0 4 0 0 -4 0 0 0 0 5 3 0 -3 0 0 -2 5 -2 0 1
0 -3 5 1 2 4 5 3 -3 5 -1 3 0 3 4 4 -4 0 0 0

Obtener un histograma de los datos anteriores y discutir las conclusiones que se deducen del mismo.

Vamos a analizar con detalle este ejemplo, correspondiente a un caso real fruto de la experiencia personal de los autores en sus relaciones con la industria.

Una visión superficial de los datos no parece resaltar ningún hecho especial. Se constata que dos de los cien valores están fuera del intervalo de especificación [-5 +5], lo que haría pensar en un proceso que produce aproximadamente un 2% de piezas defectuosas. Por otra parte se aprecia un gran número de valores en el nominal 0, y la media de los cien datos es 1.4, que indicaría sólo un ligero descentrado respecto a dicho valor nominal óptimo. ¡La realidad, sin embargo, es muy diferente! y para ponerlo de manifiesto es suficiente una herramienta estadística tan sencilla como la construcción de un histograma.

La siguiente figura recoge el histograma de los datos (obtenido de forma que cada uno de los 13 valores registrados al redondear los datos corresponda a un tramo)



Aunque los 4 primeros tramos (de izquierda a derecha) parecen iniciar la forma esperada para una campana, en los 3 siguientes aparece ya una clara anomalía, con valores bajos para -1 y +1 y un valor excepcionalmente alto para el 0. La baja frecuencia de los valores -1 y +1 se explicó al ver in situ el útil con el que se medía la dimensión considerada. Se trataba de un útil analógico en la que la medida final la indicaba una aguja sobre una escala que tenía en el centro el valor nominal teórico buscado (el 0), hacia la derecha las desviaciones positivas (+1, +2, ...) y hacia la izquierda las negativas (-1, -2, ...). En esta situación es muy probable que los operarios cuando la aguja se posicionaba en la zona central cercana al cero, tuvieran tendencia a apuntar este valor nominal sin detenerse a precisar si realmente estaba la aguja más cerca del -1 o del +1.

Esta anomalía que acabamos de detectar es simplemente una curiosidad (es frecuente encontrarla en este tipo de situaciones), pero tiene poca relevancia práctica, pues realmente no es muy grave que, por ejemplo, un valor que es +0.8 se anote como 0 en vez de como +1. Por otra parte, este fenómeno del “robo” de algunos valores por parte del 0 a los -1 y +1, en modo alguno basta para justificar la elevadísima frecuencia observada para este valor nominal.

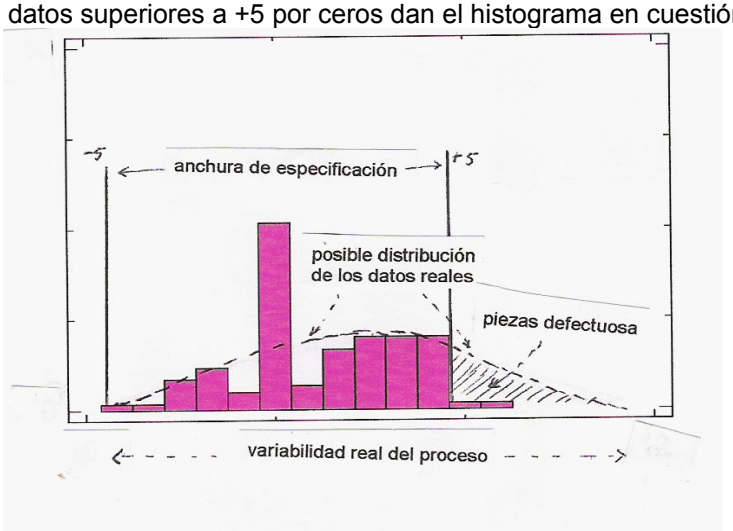
El problema realmente importante aparece cuando se analiza la parte final derecha del histograma. Se aprecia con gran claridad que a partir del límite superior de especificación (+5) los datos prácticamente desaparecen, dando lugar a un histograma “en acantilado”. ¡Habitualmente los procesos bajo control generan datos con histogramas aproximadamente campaniformes, y es imposible que un proceso real funcione tal como indica el histograma, con una elevada frecuencia de valores +4 y +5 y prácticamente sin valores más elevados!

La explicación más verosímil a la vista del histograma es que los operarios falsean sistemáticamente los datos cuando se salen de especificación, tendiendo simplemente a apuntarlos al valor nominal (o quizás en algún caso en el +5) lo que justificaría de paso la elevada frecuencia obtenida para el 0.

La causa de esta práctica muy probablemente radicaba en algún “jefe”, de un tipo desgraciadamente frecuente, que culpabiliza a los operarios de los problemas de un proceso (que los operarios no han diseñado) y en frases del tipo “menos Estadística y más trabajar” o “que sea la última vez que veo un dato fuera de especificación”.

Pero analicemos las consecuencias. Desde luego el “jefe” ha conseguido su objetivo (ya no ve datos fuera de especificación), pero su gestión ¿ha mejorado el proceso o lo ha dejado igual? Ni una cosa ni otra, su gestión ha empeorado el proceso, y no sólo, y es lo más importante, por crear un ambiente de trabajo orientado a “buscar culpables” que fomenta el falseo de datos y otras prácticas muy perniciosas, sino también por una cuestión técnica que vemos a continuación.

En la siguiente figura hemos superpuesto sobre el histograma de los datos falseados, la curva que corresponde muy probablemente a la distribución real de los datos del proceso (que tras la sustitución de los datos superiores a +5 por ceros dan el histograma en cuestión).



Lo primero que se aprecia es que el proceso debe estar produciendo realmente del orden de un 30% de piezas defectuosas, en vez del 2% que se deducía de los datos. (No deja de ser notable, que una herramienta tan sencilla como un histograma detecte que, aunque en los datos sólo hay un 2% de piezas malas, en realidad probablemente hay 15 veces más)

Pero además la curva dibujada pone claramente de manifiesto que el proceso no es capaz de producir todas las piezas dentro del intervalo [-5 +5], porque su variabilidad natural es superior a esta anchura de especificación. Esta variabilidad natural no depende de lo laboriosos o cuidadosos que sean los operarios, sino más bien de cuestiones, como la precisión de las máquinas de mecanizado o la calidad de las materias primas, sobre las que la decisión corresponde precisamente al "jefe". Por lo tanto, hagan lo que hagan los operarios el proceso producirá piezas fuera de especificación.

Pero el proceso, aun manteniendo su variabilidad, produciría menos piezas fuera de especificación si se centrara adecuadamente en su valor nominal (centrar un proceso, es decir modificar su media, es generalmente mucho más fácil que reducir su dispersión). El proceso está realmente muy descentrado, con un valor medio que probablemente será del orden de 4. ¡Pero como los datos están falseados este problema no se detecta, pues aparentemente la media es 1.4, bastante cercana al nominal!

En definitiva, el histograma ha puesto e evidencia las muy perniciosas consecuencias que pueden derivarse de ciertos estilos de dirección, desgraciadamente frecuentes en la industria española.

Autoevaluación: La LD50 de un insecticida (ver respuesta a la segunda Autoevaluación en este Anejo) ¿a qué parámetro de la distribución de la variable considerada corresponde? ¿Qué otro parámetro podría haberse definido en vez de la LD50? ¿Por qué consideras que se eligió en su día el primer parámetro?

La LD50 sería la mediana de la variable considerada.

Podría, en principio, haberse elegido otra medida de posición, como por ejemplo la media que es la utilizada más habitualmente. Sin embargo las variables del tipo "duración hasta que ..." o "resistencia hasta el fallo ..." suelen tener distribuciones asimétricas positivas, por lo que la mediana es un indicador de posición más aconsejable.

Adicionalmente, el carácter "censurado" de la variable estudiada (ver en este mismo Anejo la respuesta a la segunda Autoevaluación) hace más sencillo, utilizando procedimientos estadísticos elementales, calcular la mediana (pues basta con ver que la mitad de las ratas

resisten más de un determinado valor) que la media (que en principio necesitaría conocer la resistencia individual de cada rata)

2.A.2 Ejercicios adicionales

1. *Los siguientes datos recogen los volúmenes de facturación (en millones de pesetas anuales) de una muestra de 11 empresas de un determinado sector*

25 110 42 10 8 180 70 14 56 17 30

1. *Dibujar un diagrama Box-Whisker de los datos*
2. *A la vista del diagrama ¿qué signos cabe esperar que tengan el coeficiente de asimetría y el coeficiente de curtosis de los datos?*
3. *¿Qué indicadores serían adecuados para describir la posición y la dispersión de estos datos?*

2. *En una muestra de dos datos la media muestral* 0 *es igual a 4 y la varianza muestral* s^2 *es igual a 8. Calcular el recorrido y los coeficientes de asimetría y de curtosis de la muestra.*

3. *Para controlar un proceso se toman cada 4 horas 5 piezas, midiéndose en cada una de ellas la variable de interés. En la tabla siguiente se recogen los resultados obtenidos en las últimas 8 muestras.*

31.2	*28.8*	*21.6*	*23.4*	*32.4*	*28.9*	*23.3*	*27.6*
18.6	*36.7*	*21.5*	*33.9*	*21.3*	*29.6*	*18.3*	*21.1*
24.7	*20.5*	*19.5*	*21.3*	*20.4*	*20.8*	*18.6*	*32.1*
29.5	*34.7*	*21.7*	*17.1*	*20.1*	*20.8*	*23.1*	*29.8*
31.6	*21.8*	*26.6*	*29.1*	*27.7*	*29.5*	*33.7*	*32.5*

a) *Calcular la media, desviación típica y coeficientes de asimetría y curtosis de los datos. ¿Son los datos aproximadamente simétricos?*
b) *Constatar si, como postula la teoría, aproximadamente dos tercios de los datos difieren de la media menos de un desviación típica. ¿Qué conclusión se obtiene?*
c) *Construir un histograma de los datos (10 tramos entre 17 y 37) y deducir a partir del mismo la causa el resultado anormal obtenido en el apartado anterior.*