Apuntes
508 Estadística

Ingeniería
técnica en
Diseño Industrial

Curso 2002/03

Irene Epifanio

LÉEME

El material de este curso lo componen:

-  Apuntes de teoría y problemas
-  Hojas de problemas con soluciones
-  Hojas de sesiones de prácticas con el ordenador
-  Cuestionarios de autoevaluación
-  Otro tipo de material extra: formulario, tablas, información complementaria
-  PERO SOBRE TODO: acudir y atender a las clases, y a las tutorías cuando se necesiten

- Los apuntes tienen como objetivo facilitar el seguimiento de las clases, sirviendo de apoyo a éstas, puesto que recogen los contenidos básicos del curso. Se trata de que la atención se centre en las explicaciones de clase y que el hehco de tomar notas no la dificulte. Se ha partido de cero, y se ha intentado presentar los conceptos de forma simple, siempre ayudándose de ejemplos-problemas, evitando las demostraciones que pueden encontrarse en la bibliografía o en tutorías. Se trata de que las herramientas estadísticas se entiendan, se sepan utilizar e interpretar, ya que su uso va a ser completamente aplicado. Estos apuntes van ligados a las clases, no se entienden como algo separado de ellas, pues **las clases** son, en realidad, **esenciales**.

- Los problemas aparecen por un lado en los apuntes de teoría marcados con  y serán resueltos en clase. Por otro lado, aparecen en las hojas de problemas. Estas hojas están compuestas por una serie de problemas, todos ellos con su solución para que así podáis comprobar si se han realizado correctamente: recordad que vale más resolver uno por vosotros mismos que copiar mil. Además, al comienzo de las hojas de problemas de cada tema aparece generalmente un ejemplo resuelto de nivel algo superior (no asustarse). Algunos de los problemas de estas hojas se resolverán en clase, mientras que el resto se dejarán para que los trabajéis en casa. De nuevo, podéis recurrir a la bibliografía o a tutorías si queréis afianzar más algún tema.

- En las prácticas, se trabajarán conceptos vistos en clase de teoría-problemas con la ayuda del ordenador (el papel del ordenador es fundamental, como se verá) y otros conceptos nuevos. Todas las prácticas están constituidas por una primera parte de explicaciones de comandos y conceptos y otra parte con los problemas a resolver (con datos diferentes para cada persona). Es **muy recomendable** asistir a las prácticas, habiéndose leído con anterioridad las hojas de prácticas y habiendo repasado también los conceptos que se vayan a tratar en esa sesión.
- Las cuestiones de autoevaluación pretenden fomentar el estudio diario, ya que la continuidad es fundamental Asimismo, las cuestiones de autoevaluación os permitirán comprobar vuestros progresos y detectar posibles errores, en cuyo caso ya sabéis que contáis con las tutorías.
- En mi página ( hhtp://www3.uji.es/∼epifanio, curs 2002-03) podréis encontrar el material disponible en cada momento, así como el de años anteriores.
- Si habéis llegado a leer hasta aquí (thanks!), habréis comprobado que la palabra más repetida ha sido: TUTORÍAS. En cuanto tengáis cualquier duda, no *dudéis* en acudir a solucionarla, no dejéis que "se enrede la madeja". Mi deseo es ayudaros y las tutorías están para eso.
- Confío en que me digáis aquello que encontréis mejorable (cualquier sugerencia será bienvenida!), para ello también os pasaré un cuestionario sobre vuestra opinión en diversos aspectos.
- De verdad espero que todo esto os sirva de ayuda y sobre todo que no nos tengamos que ver las caras en septiembre ... Ahora abrochémonos el cinturón y ¡al ataquerrrrr! ☺

Castellón, febrero 2003

Irene Epifanio

Lista de Símbolos

-  : ¡Ojo!, observación
-  Definición (a leer)
-  ¡Alto! esto es muy importante
-  Ejemplo-problema hecho en clase (a escribir se ha dicho)
-  Información complementaria
-  Sesión de prácticas con el ordenador
-  se verá más adelante
-  Ya se ha visto (a recordarlo)
-  Dificultad
-  Nota
-  Curvas peligrosas donde no debes estrellarte: ir con cuidado

Tema 1

Descripción de una muestra

1.1. Introducción

A continuación, se presentan varios ejemplos del tipo de problemas que seremos capaces de resolver al final del curso (espero!).

Ejemplo 1.1. Estamos interesados en comparar 3 tipos de hojalata A, B y C, según la corrosión que se produce en botes llenados con cierto líquido de pH determinado al cabo de 3 meses. La corrosión la mediremos por el contenido de Sn (estaño) en el líquido. Para ello, llenaremos botes de los 3 tipos y mediremos su contenido en Sn al cabo de 3 meses. Supongamos que obtenemos:

A	13	8	10	11	8	
B	1	0	3	0		
C	8	5	10	3	7	3

¿Existe diferencia significativa entre los 3 tipos de hojalata? ¿Hay un tipo mucho mejor que los otros? ¿Pueden clasificarse los tipos de hojalata en diversos grupos homogéneos en el sentido que dentro de cada grupo no difieran significativamente?

El problema podría complicarse si, por ejemplo, sabemos que la temperatura de almacenamiento (5^o, 15^o y 25^o) también afecta a la corrosión, entonces deberíamos plantear un diseño experimental adecuado al problema.

Ejemplo 1.2. El departamento de calidad de cierta empresa quiere conocer la proporción de piezas que no tienen una longitud dentro de las tolerancias, son defectuosas cuando el proceso se encuentra bajo control. Tras recoger una muestra, la proporción de piezas defectuosas se estima en un 2 %. Después de un tiempo, se sospecha que ha habido una descorrección, con lo cual la proporción de piezas defectuosas habría aumentado. Se toma una muestra de 100 piezas y resulta que hay 5 defectuosas, ¿podemos concluir que ha aumentado la proporción o se debe únicamente al azar?

Ejemplo 1.3. Se pretende diseñar un ratón ergonómico para niños de 7 a 9 años. Hemos de conocer la forma de su mano derecha por lo que hemos de tomar distintos datos antropométricos de un conjunto de niños. Supongamos que estamos interesados en la longitud de su dedo índice. Realizamos un estudio piloto con 30 niños, de los que obtenemos una media de 6 cm y

una desviación típica de 0.4 cm [1]. Si deseamos poder afirmar con un 95 % de confianza que la media es imprecisa como mucho en 0.1 cm, ¿cuántos datos deberíamos tomar? Una vez tomados, calcula el intervalo de confianza al 95 % para la media y entre qué valores se encontrarán el 90 % de los niños.

Veamos ahora de qué se encarga la Estadística. La ciencia Estadística tiene un doble objetivo:

- La generación y recopilación de datos que contengan información relevante sobre un determinado problema (Muestreo)
- El análisis de dichos datos con el fin de extraer de ellos dicha información. El primer paso en el análisis de los datos consistirá en describirlos a través de ciertas medidas y gráficas, lo cual nos facilitará su comprensión (Estadística descriptiva). Sin embrago, buscamos ir más allá y poder sacar conclusiones basadas en dichos datos. Para ello, podremos recurrir a plantear un modelo matemático (teoría de la probabilidad) que nos permitirá después extraer las conclusiones que nos interesan (Inferencia estadística).

Por tanto, un modelo estadístico constará de varias partes: a) Muestreo (tema 5), b) Estadística descriptiva (temas 1 y 2), c) Confección de un modelo matemático (teoría probabilidad)(temas 3, 4 y 5), d) Inferencia estadística (tema 6). Como puede verse, cada $PARTE$ del programa se corresponde con estas partes. También, en la asignatura 'Diseño conceptual' en el primer semestre ya os aparecieron los cuestionarios-encuestas para analizar (obtención de información conocida a nivel personal) que implica las partes a) y b) en primer término. Además, el año próximo, en el primer semestre en 'Ergonomía', utilizaréis diversas partes (fundamentalmente c) y d)), mientras que en el segundo semestre en 'Metodologías del Diseño Industrial' la parte c).

En resumen, la **Estadística**: estudia los métodos científicos para recoger (hacer un muestreo), organizar, resumir y analizar datos (estadística descriptiva), así como para obtener conclusiones válidas (inferencia estadística) y tomar decisiones razonables basadas en tal análisis.

Definamos ahora algunos conceptos básicos:


Población: conjunto de todos los individuos que son objeto de estudio y sobre los que queremos obtener ciertas conclusiones. **Ejemplos:**

- Todos los niños entre 7 y 9 años (ejemplo 1.3)
- Todos los botes fabricados y por fabricarse de los tres tipos de hojalata (ejemplo 1.1)

Como puede verse, a veces las poblaciones existen físicamente y son finitas aunque muy grandes, en cambio otras veces la población es de carácter abstracto. En general, en lugar de hacer un estudio de todos los elementos que componen la población (hacer un **censo**), se escoge un conjunto más reducido.


Muestra: es un subconjunto, una parte de la población que seleccionamos para un estudio.

Es deseable que la muestra extraída "se parezca" a la población, es decir, "que sea como la población pero en tamaño reducido". El objetivo es que la muestra sea representativa de

[1] en este tema estudiaremos la media y la desviación típica

la población. Notemos que si la muestra es mala, las conclusiones extraídas no serán válidas, podrían ser erróneas. En el tema 5, se estudiará el muestreo con más detalle ().

Ejemplo: si para obtener medidas para el ejemplo 1.3 acudiéramos a un entrenamiento de baloncesto de niños entre 10 a 11 años, ¿obtendríamos una muestra representativa de la población o sesgada?

Tamaño muestral: es el número de observaciones de la muestra, N.

Variable aleatoria: es una característica aleatoria que podemos expresar numéricamente, es la característica que estamos midiendo en cada individuo. Una característica aleatoria será una característica que tomará un valor para cada individuo.

Las variables aleatorias las denotaremos con letras mayúsculas: X, Y, ...

Las variables aleatorias pueden clasificarse en:

- Cualitativas o categóricas: expresan una cualidad
- Cuantitativas: tienen propiamente carácter numérico

Variable cualitativas. Las variables cualitativas a su vez se subdividen en: ordinales o no ordinales, según si las categorías pueden o no disponerse bajo un orden con sentido.

Ejemplos de variables cualitativas no ordinales:

- Sexo de una persona: 1 = Mujer, 2 = Hombre
- Adicción al tabaco: 1 = Fuma, 2 = No fuma
- Tipo de defectos de un frigorífico defectuoso: 1 = Termostato, 2 = Compresor, 3 = Motor, 4 = Cableado, 5 = Revestimiento, 6 = Otros
- **Ejemplo 1.4:** los alumnos de 1^o ETDI quieren irse de viaje de fin de curso para celebrar que han aprobado y para sacarse unos euros deciden vender gorras. Quieren conocer el color preferido por los compradores potenciales, por tanto, les interesa la variable aleatoria: "Color de la gorra preferido por los miembros de la UJI", con posibles valores: 1 = Negro, 2 = Blanco, 3 = Rojo, 4 = Otros.

Ejemplos de variables cualitativas ordinales:

- Interés sobre una determinada materia: 1 = Bajo, 2 = Medio, 3 = Alto
- Cualquiera de las de la encuesta de evaluación: 1 = Muy desfavorable, 2 = Desfavorable, 3 = Indiferente, 4 = Favorable, 5 = Muy favorable

Las **variables cuantitativas** también se dividen en dos:

- Discretas: toman valores discretos, es decir, en un conjunto numerable (podemos contar los posibles valores que pueden adoptar). Existen "espacios"entre los posibles valores que puede adoptar la variable.

- Continuas: como indica su nombre, toman valores en un conjunto no numerable. "Los valores que adoptan estas variables, pueden estar tan cercanos como se quiera".

Ejemplos de variables discretas:

1. Número de piezas defectuosas en un lote de 100 piezas
2. Número de caras obtenidas al lanzar una moneda 20 veces
3. Número de $5's$ al lanzar un dado 60 veces

En los tres casos anteriores los valores que pueden adoptar son finitos: en a) de 0 a 100, en b) de 0 a 20, en c) de 0 a 60. Sin embargo, podría no ser así, podría adoptar valores discretos no limitados:

1. Número de manchas de más de $1mm^2$ en una lámina
2. Número de defectos en 2m de cable
3. Número de veces al mes que va al cine un estudiante de ETDI

Ejemplos de variables continuas:

1. **Ejemplo 1.1:** contenido de Sn en el líquido de botes A
2. **Ejemplo 1.3:** longitud de la mano de niños de 7 a 9 años
3. Peso de ciertas piezas
4. Tiempo de vida (duración) de ciertos motores
5. Dureza de cierto material
6. Resistencia de cierto producto
7. Notas de estudiantes de ETDI
8. Euros gastados con el móvil en un mes por un estudiante de la UJI

Observación: *La distinción entre variables continuas y discretas no es rígida. Las variables continuas anteriores corresponden a medidas físicas que siempre pueden ser redondeadas, por ejemplo, la longitud podemos medirla hasta el milímetro más cercano o el peso hasta el gramo más cercano. Aunque estrictamente hablando, la escala de dichas medidas sea discreta, las consideraremos continuas como una aproximación a la verdadera escala de medida.*

Resumiendo, las variables aleatorias pueden ser:

1. Categóricas o cualitativas
   a) No ordinales
   b) Ordinales
2. Cuantitativas
   a) Discretas
   b) Continuas

1.2. Distribución de frecuencias

Ejemplo 1.4. Hemos realizado una encuesta a los alumnos de 1^o ETDI para conocer su color preferido de gorra, obteniendo un total de 150 valores: 1, 1, 1, 4, 2, 1, 1, 1, 3, 1, 4, 4, 2, 1, 1, 4, 4, 1, 4, etc. Así presentados los valores, resulta difícil interpretarlos. Para obtener una visión general podríamos emplear una **tabla de frecuencias**.

Color	Frecuencia absoluta	Frecuencia relativa
1 = Negro	78	78/150 = 0.52 = 52 %
2 = Blanco	10	10/150 = 0.07 = 7 %
3 = Rojo	20	20/150 = 0.13 = 13 %
4 = Otros	42	42/150 = 0.28 = 28 %
Total	150	1 100 %

Si la variable estudiada toma muchos valores distintos, en lugar de construir una macro tabla difícil de interpretar, los valores se agrupan. [Nota: en la página 14 de Ras podéis encontrar una forma de agruparlos].

Ejemplo 1.5. Tabla de frecuencias de las notas de Estadística de ETDI en junio de 2001.

Intervalo (límites de clase)	(Marca de clase)	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumuluada
[0, 2.5)		9			
[2.5, 5)		21			
[5, 7.5)		63			
[7.5, 10]		46			

Frecuencia absoluta: número de observaciones en el intervalo

Frecuencia relativa: número de observaciones en el intervalo / tamaño muestral; suma 1; indica el porcentaje de observaciones en el intervalo

Frecuencia acumulada: suma de las frecuencias de los intervalos anteriores, incluyendo el actual. Indica el número de observaciones por debajo del extremo superior de la clase. Obviamente, el último valor es el tamaño muestral.

Frecuencia relativa acumulada: frecuencia acumulada / tamaño muestral. Indica el porcentaje muestral por debajo del extremo superior de la clase. El último valor será 1 (100 %).

Ejemplo 1.5.:

- ¿Cuántos sacaron menos de 5?
- ¿Cuántos aprobaron?
- ¿Qué porcentaje sacó ≥ 2.5?
- De los extremos de los intervalos, ¿cuál es la menor nota por debajo de la cual se encuentra aproximadamente el 66 %?

Normalmente, las clases son de igual anchura, pero podrían no serlo:

Intervalo	Frec. abs.	Frec. rel.	Frec. acum.	Frec. rel. acum.
[0, 5)	30			
[5, 7)	44			
[7, 8.5)	39			
[8.5,10]	26			

En la práctica 1 se trabajará más este apartado con el ordenador.

1.3. Métodos gráficos

Los gráficos nos permiten también ilustrar la distribución de los datos.

Histograma: pueden ser de frecuencias absolutas, relativas, acumuladas o relativas acumuladas, según que represente la altura de la barra.

Ejemplo 1.5.:

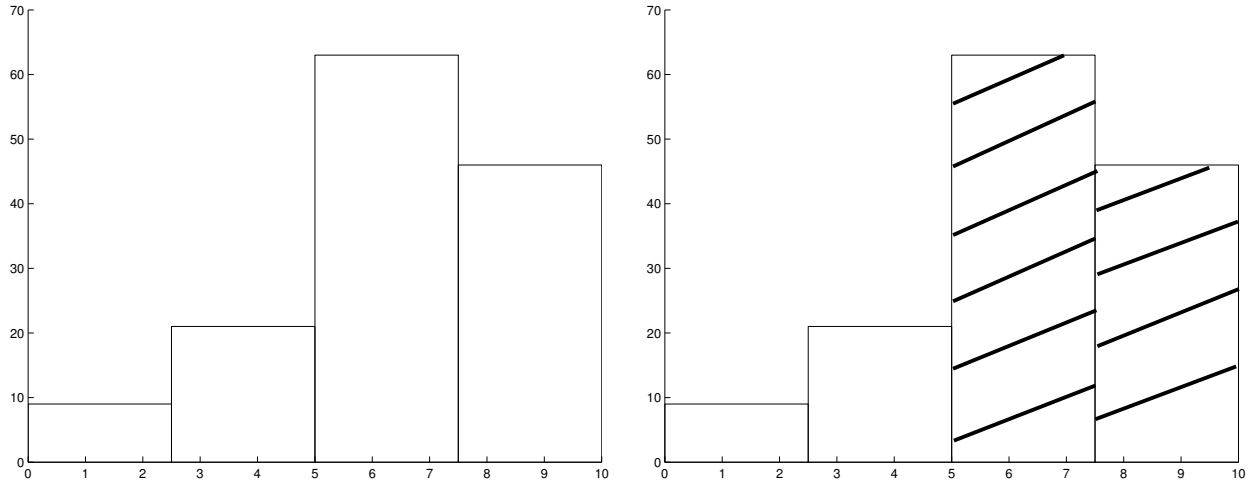


Figura 1.1: Histograma de frecuencias absolutas del ejemplo 1.5.

Los histogramas nos muestran como se distribuyen (como se reparten) los datos, las cimas de las barras indican la forma de la distribución. Además, el área de cada barra es proporcional a

la correspondiente frecuencia. **Ejemplo:** el área rayada del ejemplo anterior es el 78.4 % del área total de todas las barras, por tanto, el 78.4 % de las notas están en las correspondientes clases, o sea, el 78.4 % de las notas están entre 5 (inclusive) y 10.

Hay muchos más métodos gráficos: diagramas de barra, de sectores, polígonos de frecuencias, diagrama de cajas (*boxplot*), Pareto,... Algunos de ellos verán en los problemas del tema 1 (fotocopias del libro de Ras) y en la práctica 1 .

1.4. Medidas descriptivas

Además de las gráficas, otra forma de resumir los datos es mediante parámetros numéricos, que podemos dividir en:

- Medidas de posición o centrales: dan cuenta de la posición de las observaciones
- Medidas de dispersión: indican la dispersión (variabilidad) de los datos
- Medidas de forma: miden la forma de distribuirse los datos

Medidas de posición: media, mediana, moda y percentil.

Media: si tenemos una muestra $\{x_1, x_2, ..., x_N\}$,

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{1.1}$$

[Nota: podría ser que no todas las observaciones tuvieran la misma importancia, en ese caso hablaríamos de **media ponderada**, $\sum_{i=1}^{N} w_i \cdot x_i / \sum_{i=1}^{N} w_i$, utilizada por ejemplo para obtener la nota final de la asignatura: $0.8 \times$ Nota examen $+ 0.2 \times$ Nota prácticas].

Ejemplo 1.6: Nota media de 5 prácticas: $\{10, 8, 9, 7, 9\}$ (Calculadora: $\boxed{\overline{x}}$)

Ejemplo 1.7: Nota media de 5 prácticas: $\{10, 8, 9, 0, 9\}$

Como acabamos de ver, si tenemos valores extremos (*outliers*, ver página 13 del libro de Ras), muy alejados de la mayoría, la media puede distorsionarse, en esos casos es mejor usar: la mediana. En general, la media es mejor puesto que en su cálculo implicamos todos los datos y no sólo el orden como en la mediana.

Mediana: es el valor central de la muestra, en el sentido que la mitad de los valores son menores que la mediana y la mitad son más grandes. Para calcularla: ordenados los datos de menor a mayor $\{x_1, x_2, ..., x_N\}$, la mediana es:

$$\begin{cases} x_{\frac{N+1}{2}} & \text{si N es impar,} \\ \frac{x_{\frac{N}{2}}+x_{\frac{N}{2}+1}}{2} & \text{si N es par.} \end{cases} \tag{1.2}$$

✍ **Ejemplo 1.6 y 1.7:**

Moda: es el valor muestral con la frecuencia más alta, el que más se repite (la moda no tiene porqué ser un único valor).

✍ **Ejemplo 1.6 y 1.7:**

Cuartil: la mediana dividía los datos en dos partes iguales. Cuando dividimos el conjunto ordenado de datos en cuatro partes iguales, los puntos de división se llaman cuartiles. Por tanto, el primer cuartil es un valor tal que la cuarta parte de los valores de la muestra son más pequeños que él y las tres cuartas partes restantes son más grandes. El segundo cuartil es la mediana y el tercer cuartil es tal que las tres cuartas partes de valores son más pequeños que él y la cuarta parte más grandes.

Percentil: Si en lugar de dividir la muestra en 2 ó 4 partes, se divide en 100 partes iguales, los puntos de división se llaman percentiles. Por tanto, el k-ésimo percentil, P_k, es un valor tal que al menos el $k\,\%$ de las observaciones están en el valor o por debajo de él, y al menos el $(100-k)\,\%$ están en el valor o por encima de él. ✍ **Ejemplo:** ¿A qué percentil corresponde el primer, segundo y tercer cuartil?

¿Cómo calcular el k-ésimo percentil?

1. Ordenar las N observaciones de menor a mayor
2. Calcular $\frac{N\cdot k}{100}$
   a) Si $\frac{N\cdot k}{100}$ no es un entero: considerar el entero inmediato posterior y determinar el valor ordenado correspondiente
   b) Si $\frac{N\cdot k}{100}$ es un entero, digamos j: calcular la media de las observaciones ordenadas j-ésima y $(j+1)$-ésima

✍ **Ejemplo 1.5.:**

- ¿Qué significa $P_{95} =$?

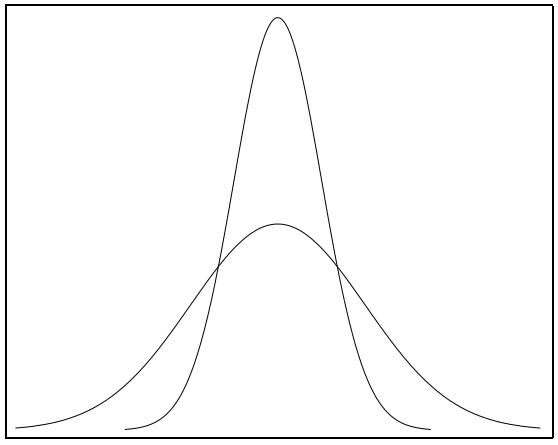
- ¿Qué significa $P_{97} =$?

✍ **Ejemplo 1.6.:** Calcula P_{25} y P_{75}

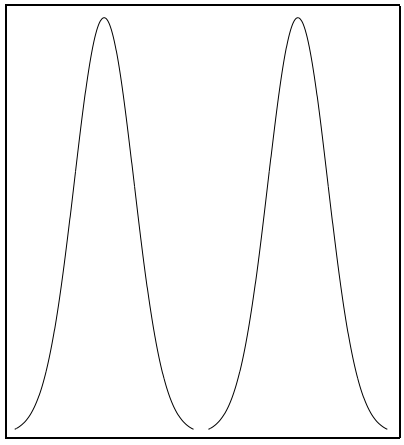
✍ **Ejemplo 1.7.:** Calcula P_{25} y P_{75}

Una medida de posición no es suficiente para describir los datos, porque no informa acerca de la variabilidad de los datos. **Ejemplo 1.8.:** La nota media de prácticas es 5.2 tanto para $\{0, 2, 5, 9, 10\}$ como para $\{4, 5, 5, 6, 6\}$, sin embargo, claramente su dispersión es distinta.

Si representamos los histogramas mediante curvas continuas, apreciaremos la distinción entre posición y dispersión.



misma posición y diferente dispersión

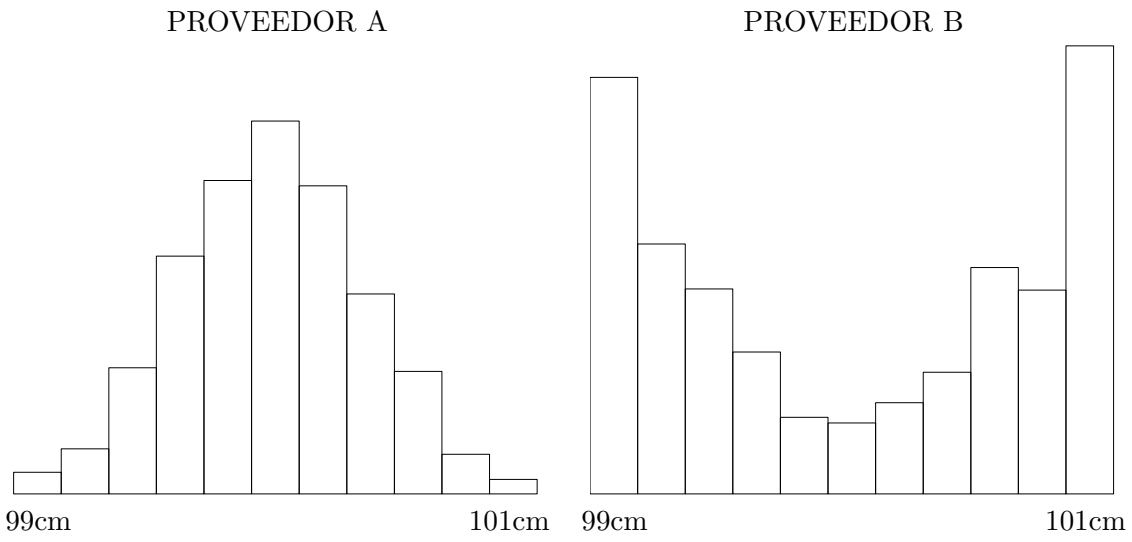


distinta posición y misma dispersión

Algunos ejemplos de la importancia del concepto de dispersión:

✍ **Ejemplo 1.9.:** ¿Para una persona que no sabe nadar es suficiente saber que la profundidad media del lago es 1.4m para lanzarse al mismo? ¿Y si conociésemos la mediana?

✍ **Ejemplo 1.10.:** Una empresa desea comprar varillas de 1m. de longitud, aunque está dispuesta a aceptar aquellas varillas que no difieran en más de 1cm del metro de longitud. La longitud de las varillas ofrecidas por dos proveedores presentan una pauta de variabilidad sintetizada por los siguientes histogramas:



Ambos proveedores cumplen las especificaciones y tienen media 1m, pero ¿cuál escogerías?


Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica o estándar, coeficiente de variación.


Rango o recorrido: Diferencia entre el mayor y menor valor de la muestra.


Ejemplo 1.6 y 1.7:


Rango intercuartílico: Diferencia entre el tercer y primer cuartil.


Ejemplo 1.6 y 1.7:

Al igual que ocurría con la mediana (en los parámetros de posición), el rango intercuartílico es un indicador robusto, puesto que le afectan poco los valores extremos de la muestra. De todas maneras, ambas medidas (rango y rango intercuartílico) son bastante pobres, porque ignoran gran parte de la información muestral (se basan en el orden). Por ello, la medida de dispersión más usada es la varianza:


Varianza:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N-1} \tag{1.3}$$

Fórmula alternativa:

$$s^2 = \frac{\sum_{i=1}^{N} {x_i}^2 - N \cdot \overline{x}^2}{N-1} = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_N}^2 - N \cdot \overline{x}^2}{N-1} \tag{1.4}$$

Calculadora: $\boxed{\sum x^2}$, $\boxed{\overline{x}}$ o bien $\boxed{\sigma_{N-1}}$, $\boxed{x^2}$

Observación: comprobación de la fórmula alternativa

$$\left(\sum_{i=1}^{N} (x_i - \overline{x})^2 = \sum_{i=1}^{N} (x_i^2 - 2 \cdot \overline{x} \cdot x_i + \overline{x}^2) = \sum_{i=1}^{N} x_i^2 - 2 \cdot \overline{x} \cdot \sum_{i=1}^{N} x_i + N \cdot \overline{x}^2 = \sum_{i=1}^{N} x_i^2 - 2 \cdot N \cdot \overline{x}^2 + N \cdot \overline{x}^2 = \sum_{i=1}^{N} {x_i}^2 - N \cdot \overline{x}^2 \right)$$

Por la fórmula 1.3 puede apreciarse que a mayor varianza, mayor dispersión, pues calculamos desviaciones de la media al cuadrado. Por esto último (cuadrados), **la varianza siempre será mayor o igual que cero**. [RECORDAD: NUNCA NEGATIVA, SIEMPRE POSITIVA ...].

Ejemplo 1.6 y 1.7:

¿Por qué dividir por N - 1, en lugar de por N? Por razones técnicas que ya se comentarán más adelante (tema 5); una justificación intuitiva sería considerar el caso en que N=1 (un único valor muestral). Si N es grande no habrá apenas diferencia.

Ejemplo 1.3: si sólo observáramos 1 niño (N=1) y nos diera como medida 7 cm, ¿cuál sería s^2? ¿Y si dividiéramos por N?

La varianza es muy apropiada por ciertas propiedades (si dos variables son independientes, la varianza de la suma es la suma de las varianzas), pero tiene un problema: cambia las unidades de los datos, ya que hacemos un cuadrado. Para resolverlo se usa la raíz cuadrada de la varianza:

Desviación típica o estándar:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}} = \sqrt{s^2} \tag{1.5}$$

Calculadora: $\boxed{\sigma_{N-1}}$

Ejemplo 1.6 y 1.7:

Coeficiente de variación: en el problema 8 del tema 1, se verá ▶▶|.

Medidas de forma: recordemos que miden la forma en como se distribuyen los datos, para ello se compara con una forma predeterminada, de referencia: la distribución Normal (▶▶| tema 5), que tiene forma de campana (la campana de Gauss):

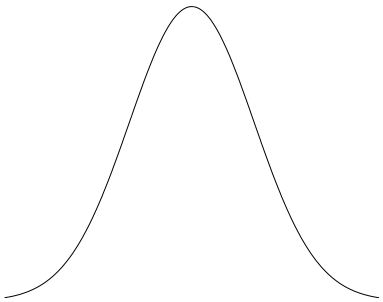


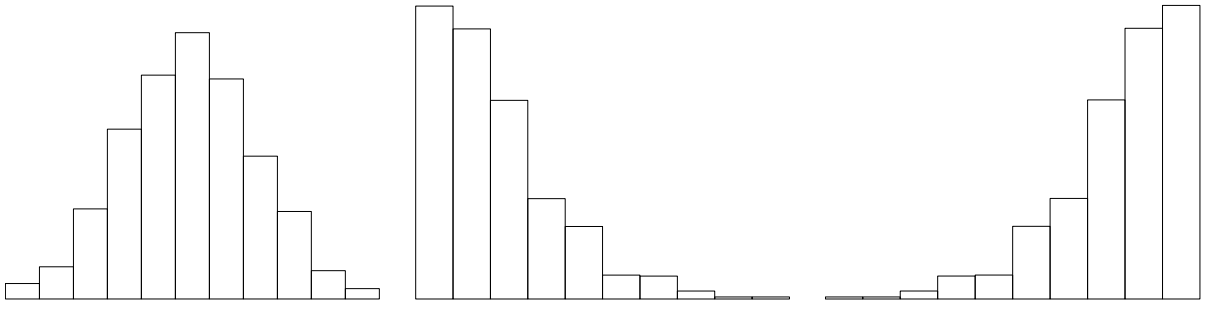
Figura 1.2: Datos acumulados en torno a la media, simétrica respecto de la media, forma de campana

Medidas de forma: coeficiente de asimetría, coeficiente de curtosis o apuntamiento.

Coeficiente de asimetría:

$$CA = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3/N - 1}{s^3} \tag{1.6}$$

- $CA = 0$: datos simétricos
- $CA > 0$: datos con asimetría positiva o a derechas
- $CA < 0$: datos con asimetría negativa o a izquierda



$CA = 0$ $CA > 0$ $CA < 0$

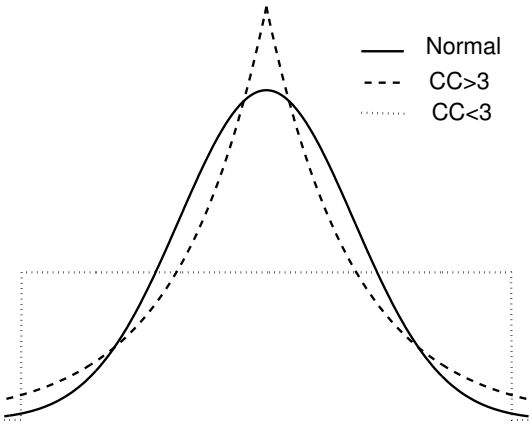
Coeficiente de curtosis: indica lo apuntada que es la distribución

$$CC = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4/N - 1}{s^4} \tag{1.7}$$

- $CC = 3$: los datos tienen forma de campana de Gauss [Nota: a veces se resta 3 al coeficiente para centrarlo en el cero]

- $CC > 3$: la distribución es más apuntada, picuda que la Normal

- $CC < 3$: es menos apuntada, más achatada que la Normal



En la práctica 1 se trabajará más este apartado.

1.5. Descripción de la población

Hasta ahora hemos examinado diversas formas de describir una muestra. Aunque la descripción de un conjunto de datos es a veces de interés por sí misma, normalmente lo que se pretende es generalizar y extender los resultados más allá de la limitación de la muestra. La población es realmente el foco de interés.

Como ya vimos (apartado 1.1.), el proceso de sacar conclusiones sobre una población basándonos en las observaciones de una muestra de dicha población, es la Inferencia Estadística.

Puesto que las observaciones se realizan únicamente en la muestra, las características de la población nunca se conocerán exactamente. Para poder inferir ("deducir, concluir, tomar decisiones") de una muestra a la población, necesitaremos un lenguaje (paralelo al muestral) para describir la población.

Variables categóricas: podemos describir la población simplemente indicando la proporción de la población en cada categoría.

Ejemplo 1.4.:

Toda la población, todos los miembros de la UJI

Color	Frecuencia relativa (proporción)
1 = Negro	0.57
2 = Blanco	0.14
3 = Rojo	0.09
4 = Otros	0.2

La muestra de alumnos de 1^o ETDI

Color	Frecuencia relativa (proporción)
1 = Negro	0.52
2 = Blanco	0.07
3 = Rojo	0.13
4 = Otros	0.28

La proporción muestral de una categoría es una *estimación* de la correspondiente proporción poblacional (en general desconocida). Puesto que no tienen porqué ser iguales (aunque sí que querríamos que fuesen cuanto más iguales mejor), las denotaremos con letras diferentes:

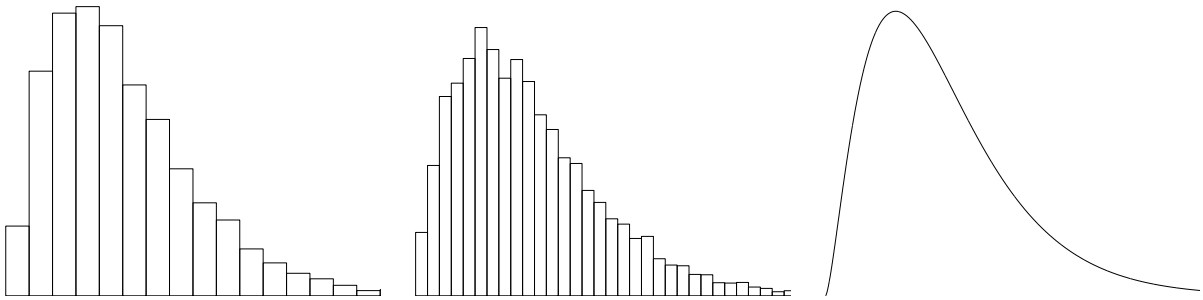
- p = proporción de la población
- $\hat{p}$ = proporción de la muestra

Variables cuantitativas: para variables cuantitativas, la media, varianza, desviación típica, etc. son descripciones de la población. Estas cantidades se calculan con los datos muestrales y constituyen una *estimación* de las correspondientes cantidades para la población. La media de la población la denotaremos mediante la letra μ, la varianza y desviación típica de la población con σ^2 y σ respectivamente. Recordemos que la media muestral era $\overline{x}$, la varianza muestral, s^2 y la desviación típica, s. Notemos que $\overline{x}$ es una *estimación* de μ (desconocida) y s es una estimación de σ (desconocida).

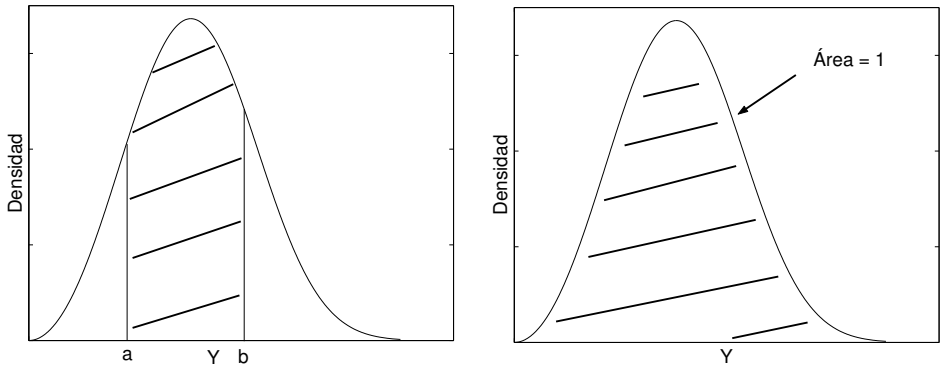
[♪ Nota: las cantidades poblacionales las denotamos con letras griegas que se corresponden con las respectivas letras latinas, para las cantidades muestrales].

Ejemplo 1.3.: con la muestra de 30 niños obtenemos $\overline{x} = 6$ y $s = 0.4$. La media de la población (todos los niños entre 7 y 9 años) la llamamos μ y no la conocemos. La desviación típica de la población (todos los niños entre 7 y 9 años) la llamamos σ y no la conocemos.

El histograma también es una buena herramienta que nos informa sobre la distribución de frecuencias de la población. Si, además, la variable es continua, podemos emplear una curva suave para describirla. Esta curva puede verse como una idealización del histograma con clases muy estrechas. Esta curva que representa la distribución de frecuencias, es la **curva de densidad**.

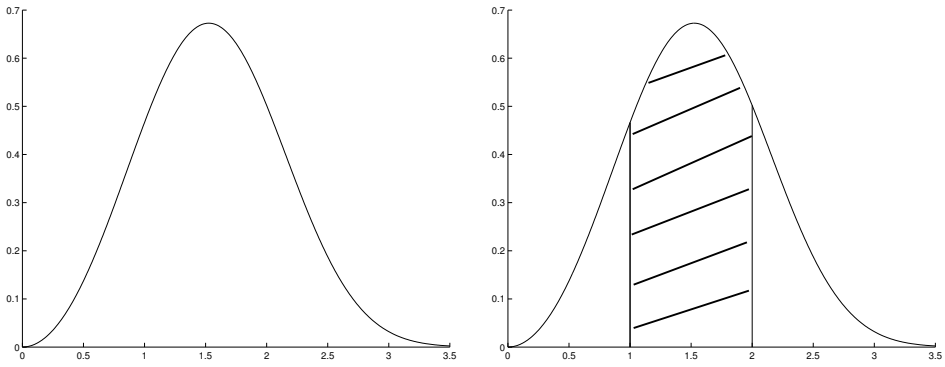


Interpretación de la densidad: el área bajo la curva de densidad entre los valores a y b equivale a la proporción de valores de la variable Y entre a y b.



Debido a la forma en que la curva es interpretada, el área bajo la curva entera debe ser igual a 1.

Ejemplo 1.11.: Supongamos que nos interesa la variable X = tiempo (en miles de horas) de vida de cierta clase de termostatos y que se distribuye según la siguiente curva de densidad:



Él área rayada es igual a 0.61, lo cual indica que el 61 % de los valores de la variable están entre 1 y 2.

Para calcular las áreas bajo las curvas de densidad, necesitaríamos integrar, pero ... usaremos tablas (formulario).

Observación: *¿Cuál sería la frecuencia relativa de un valor concreto, por ejemplo 6 cm, de la variable del ejemplo 1.3? La respuesta es cero (el área es cero). Aunque parezca extraño que la frecuencia relativa de una longitud igual a 6 cm sea cero, pensemos un poco. Si estamos midiendo hasta el milímetro más cercano, entonces, en realidad estamos preguntando la frecuencia relativa entre 5.95 cm y 6.05 cm, que no es cero. Pensemos en la longitud como una variable continua idealizada. Es similar al hecho de que una línea de 1 m, está compuesta de puntos, cada uno de ellos de longitud cero.*

En resumen, una medida numérica calculada a partir de los datos es un estadístico. La correspondiente medida numérica que describe la población es un parámetro. En la siguiente tabla se recogen las más importantes:

Medida	Valor muestral (estadístico)	Valor poblacional (parámetro)
Proporción	$\hat{p}$	p
Media	$\overline{x}$	μ
Desviación típica	s	σ

Problemas del tema 1

1. Al medir la altura de un grupo de 10 estudiantes, se encontró un valor medio de 172.6 cm y una desviación estándar de 11.4 cm. Dos estudiantes (con alturas respectivas 169.5 cm y 187.5 cm) se extraen del grupo. ¿Cuál es la altura media de los 8 estudiantes que quedan? (*Sol: 171.125 cm*).

2. La siguiente tabla muestra la cantidad de tierra de las regiones de un cierto país, junto con el porcentaje de tierra cultivada en cada región.

REGIÓN	CANTIDAD DE TIERRA (M/ha)	% Cultivado
Norte	421	46.7
Sur	350	21.0
Este	259	8.7
Oeste	80	18.8

   Calcula el porcentaje de tierra cultivada en la totalidad del país. (*Sol: 27.97 %*).

3. *Outliers*: ejemplo 1 de Ras (página 12) y práctica 2 de ordenador.

4. *Boxplot* (diagrama de cajas): ejemplo 4 de Ras (página 19) y práctica 1 de ordenador.

5. Diagramas Pareto: ejemplo 5 de Ras (páginas 20 y 21) (en la práctica 3 de ordenador se verá también).

6. Un termómetro inglés mide la temperatura en 10 puntos distintos de un gran horno. La temperatura la proporciona en grados Fahrenheit.

   Datos: 475 500 460 460 470 475 465 510 450 480

   ¿Cuáles son la temperatura media, mediana, la varianza y la desviación típica en grados Fahrenheit y en grados Celsius? ¿Cuál es la relación que hay entre ellas? (práctica 1 del ordenador). [♪ Nota: Grados Celsius = (5/9)· Grados Fahrenheit - 160/9].

   (*Sol: grados Fahrenheit: 474.5, 472.5, 341.3889, 18.47671; grados Celsius: 245.83, 244.72, 105.3669, 10.2648; relación: 245.83 = (5/9) · 474.5 - 160/9, 244.72 = (5/9) · 472.5 -160/9, 105.3669 =* $(5/9)^2$ *· 341.3889, 10.2648 =* $|5/9|$ *· 18.47671*).

7. En la tabla siguiente se muestra el punto de fusión de 48 filamentos metálicos.

320	331	322	323	324	325	326	320	335	324	314	328
325	320	318	314	312	319	318	317	313	308	318	310
322	329	327	305	313	324	320	316	314	328	329	308
329	330	317	321	323	322	316	319	327	310	311	324

En base a esos datos se puede construir una tabla de frecuencias como la siguiente:

Límites del intervalo	F. absoluta	F. relativa	F. acumulada	F. relativa acumulada
(300, 305]	1	2.083333	1	2.083333
(305, 310]	4	8.333333	5	10.416667
(310, 315]	7	14.583333	12	25
(315, 320]	13	27.083333	25	52.083333
(320, 325]	12	25	37	77.083333
(325, 330]	9	18.75	46	95.833333
(330, 335]	2	4.166667	48	100

a) ¿Cuántos productos tienen punto de fusión menor o igual que 315?
b) ¿Cuántos productos tienen punto de fusión entre 316 y 325?
c) ¿Qué porcentaje de productos tienen punto de fusión menor o igual que 320?
d) ¿Qué porcentaje de productos tienen punto de fusión mayor que 325?
e) ¿Cuál es el punto de fusión más pequeño (de entre los extremos de los intervalos) por debajo del cuál se encuentra aproximadamente el 95 % de los productos?

(*Sol: 12, 25, 52.08 %, 22.92 %, 330*).

8. **Coeficiente de variación** (práctica 2 del ordenador): supongamos que hemos medido la longitud de unas vigas y que al calcular su variación hemos obtenido una desviación típica de 1 cm. Después medimos el diámetro de unas arandelas y obtenemos también una desviación típica de 1 cm. Es obvio que ambas dispersiones no significan lo mismo. Para dar cuenta de este efecto se puede definir el **coeficiente de variación**, que es el cociente entre la desviación típica y la media ($s/\overline{x}$). De esta manera, es posible comparar el grado de dispersión relativa de dos distribuciones, es muy útil cuando se busca comparar la variabilidad de dos o más conjuntos de datos que difieren de manera considerable en la magnitud de las observaciones. Notemos que cuanto mayor sea este coeficiente mayor será la dispersión (en términos relativos). [♪ Nota: a veces se multiplica por 100, para expresarlo en porcentaje].

   Un fabricante de tubos de televisión, produce tubos de dos tipos, A y B, que tienen vidas medias de $\overline{x}_A = 1495$ horas y $\overline{x}_B = 1875$ horas y desviaciones típicas: $s_A = 280$ horas y $s_B = 310$ horas. ¿Qué clase de tubo tiene mayor dispersión (en términos absolutos)?, y ¿cuál mayor dispersión en términos relativos?(*Sol: B (310); A (0.187)*)

9. Calculad la media y desviación típica de los datos del ejercicio 6 después de haberles restado la media y haberlos dividido por su desviación típica ($z_i = \frac{x_i - \overline{x}}{s}$). Esta nueva variable es una **variable tipificada o estandarizada**. (*Sol:* $\overline{z} = 0$, $s_Z = 1$).