Capítulo 1

Introducción

1. La necesidad de la estadística en la toma de decisiones.
2. Dos enfoques que no desarrollaremos en el curso.
   - Teoría de la decisión estadística: Presa de decisions en contextos d'incertesa. La informació continguda en una base de dades.
   - La toma de decisiones desde la perspectiva de las ciencias de la computación.
3. Tipos de datos.
4. Clasificación de los métodos de análisis de los datos en función del tipo de datos y los objectivos planteados.

1.1 La necesidad de la estadística en la toma de decisiones

Consideremos las siguientes preguntas:

- ¿Cuánto tiempo tarda una furgoneta en ir desde Barcelona a Tarragona?
- ¿Cuánto cuesta ese viaje por cada una de las posibles rutas?
- Una empresas de reparto de pizzas a domicilio encarga una campaña de publicidad basada en el reparto de folletos por los buzones. ¿Se incrementan sus ventas los días posteriores a la campaña? Si es así, ¿cuánto tiempo persiste el efecto de la campaña? ¿Cuándo deberá repetirla?
- Uno de los repartidores de folletos decide tirar los suyos a un contenedor de basura. ¿Será capaz el gerente de la compañía de buzoneo de detectar esta irregularidad?
- Una cadena de comida rápida tiene 16 establecimientos repartidos por Barcelona. ¿Hay una diferencia significativa en los beneficios obtenidos por unos y otros establecimientos? Si es así, ¿qué factores pueden explicar las diferencias observadas? ¿Pueden controlarse algunos de esos factores?
- Una empresa fabrica material eléctrico (enchufes, interruptores). ¿Qué porcentaje de las piezas fabricadas son defectuosas? ¿Se mantiene este porcentaje constante a lo largo de los meses?
- Si se introduce un cambio en la cadena de montaje, ¿cómo afecta éste a la proporción de piezas defectuosas? ¿Produce mejoras que compensen la inversirón que supone el cambio?
- ¿Cómo puede esa empresa segurarse de que la cadena de montaje está *bajo control*?
- Una inmobiliaria acaba de terminar un edificio que ocupa toda una manzana de las del ensanche barcelonés. Hay 250 nuevas viviendas listas para entregar a sus compradores. ¿Cómo puede la constructora asegurarse de que no hay defectos visibles en esas viviendas? Por

ejemplo, ¿habrá muchas bañeras que hayan sido dañadas durante las labores realizadas en los pisos con posterioridad a su instalación?

Estos ejemplos ponen de manifiesto que las decisiones en el mundo de la empresa se toman a menudo en circunstancias que involucran incertidumbre.

Las decisiones se deben tomar, a pesar de ello. Conviene tener en cuenta esa incertidumbre y tratar de reducirla a la hora de decidir.

Las técnicas estadísticas que estudiaremos en el curso nos serán útiles para administrar la incertidumbre inherente a los procesos de decisión.

Estas técnicas complementarán el *buen sentido* del decisor.

Cuanta mayor sea la información de que dispone el decisor, más fácil será su tarea.

En el mundo actual la información se almacena principalmente en forma numérica en archivos informáticos.

Cuanto mayor sea nuestra capacidad de extraer información de estos archivos, mejor será nuestra posición como decisores.

La estadística es una disciplina científica que está dedicada a la extracción de información de conjuntos de datos. Por eso tiene un papel fundamental en el proceso de toma de decisiones.

Posibles definiciones de Estadística:

- Conjunto de métodos científicos que permiten recolectar, analizar e interpretar información numérica con el objetivo de establecer y/o corroborar hipótesis sobre los individuos a que hacen referencia los datos.
- *La estadística no es una colección de recetas, sino un método intelectual para extraer información de los datos.* (D. Moore)
- Metodología de análisis de los datos que permite separar patrones de comportamiento sistemático, de los hechos observados por puro azar.

Fases de un estudio estadístico orientado a la toma de decisiones:

1. Planteamiento de la pregunta relevante y delimitación de la población para la que ha de ser válida la respuesta que le daremos al final del estudio.
2. Producción de los datos: muestreo o diseño de experimentos.
3. Análisis de los datos: organización de los datos, resumen y extracción de la información relevante que contienen los datos (estadística descriptiva).
4. Extracción de conclusiones relevantes. Inferencia estadística.

EJEMPLO. Una compañia de transportes por carretera.

Una compañía de transporte de mercancías por carretera con camiones de gran tonelaje quiere ofrecer un nuevo servicio: transporte de pequeñas mercancías, que por sí solas no justifican el viaje de un camión, pero que pueden acomadarse en camiones que hagan el recorrido deseado y que no estén totalmente llenos.

La empresa quiere saber cuánto ha de cobrar por cada tonelada adicional que se incluya en un camión parcialmente cargado cuya ruta y carga principal ya está acordada. Se ha determinado que el único coste adicional que genera la inclusión de esta mercancía extra es el consumo de combustible, medido en litros por cada 100 kilómetros recorridos.

La pregunta que se hace la empresa es ésta: ¿en cuánto aumenta el consumo de combustible si se incluye en un camión una tonelada de mercancía adicional?

Información disponible: La empresa tiene un archivo infomático con datos históricos de los viajes reañlizados por sus camiones. En total hay 5.428 viajes grabados. Entre otras cosas, están registrados el peso (toneladas) de la carga transportada y el consumo (l/100km) de combustible.

¿Qué hacer? ¿Analizaremos todos los datos o sólo una parte de ellos (una muestra)? ¿Qué análisis de los datos nos permitirá responder a la pregunta anterior?

Seleccionamos una muestra aleatoria de 40 viajes. (¿Por qué 40? ¿Son suficientes?)
Los datos están en la tabla 1.1.

Peso (TM)	Consumo (litros/100km)	Peso (TM)	Consumo (litros/100km)
27.2160	44.3757	24.9480	47.0382
36.2880	58.7978	32.6592	55.9979
34.0200	52.2647	28.5768	46.1159
21.7728	32.6654	35.8344	60.3054
37.1952	61.8924	32.6592	53.4525
26.3088	47.9982	27.2160	46.1159
33.5664	52.2647	36.2880	54.6956
24.0408	39.8629	27.6696	42.7620
36.2880	67.1974	30.8448	57.3637
34.4736	52.2647	34.0200	53.4525
28.5768	47.0382	29.4840	47.9982
32.6592	51.1285	36.7416	58.7978
29.0304	44.3757	35.3808	53.4525
28.1232	47.9982	37.6488	61.8924
35.8344	57.3637	27.6696	48.9981
28.5768	47.0382	28.1232	47.9982
34.9272	51.1285	34.4736	52.2647
23.1336	41.2616	33.5664	55.9979
35.3808	54.6956	22.6800	38.5559
35.8344	54.6956	24.9480	50.0406

Tabla 1.1: Ejemplo de la compañía de transporte por carretera. Datos de consumo y toneladas transsprotadas por 40 camiones

El análisis de los datos y **la extracción de conclusiones** se verán a lo largo del curso.

1.2 Dos enfoques que no desarrollaremos en el curso

1.2.1 Teoría de la decisión estadística

Toma de decisiones en contextos de incertidumbre

El valor de la información contenida en una base de datos

REFERENCIA: Peña (1995) Vol. 1. Capítulo 5.7

1.2.2 La toma de decisiones desde la perspectiva de las ciencias de la computación

Hace unos años se hablaba de informática de gestión.

Recientemente se ha empezado a hablar también de informática de decisión.

Los científicos y expertos que trabajan en las ciencias de la computación se han planteado hace pocos años cómo extraer información relevante de las bases de datos informatizadas, el tamaño medio de las cuales crece a gran velocidad.

EJEMPLO. Grandes bases de datos

- Las empresas de telefonía disponen de información pormenorizada de todas las llamadas enviadas o recibidas por ellas: momento de la llamada, duración, origen, destino, coste real, coste facturado, etc. Pueden saber qué tipos de servicios usa con mayor frecuancia cada cliente y dirigir así de forma más eficiente sus campañas de márketing. Pueden saber cómo afecta la entrada en el mercado de nuevas compañías.

- Hipermercados. Con las tarjetas de cliente los hipermercados consiguen identificar a cada comprador y así pueden estudiar la evolución de su cesta de la compra: frecuencia, lista detallada de todas las compras, productos que sienpre compra juntos, productos que ha dejado de comprar, etc. Las campañas de márketing y la disposición de los productos en la tienda se benefician de esta información.
- Compañías de seguros. Información variada de todos los asegurados: datos demográficos, servicios contratados, grado de utilización de los servicios, etc. Esto es útil para intentar captar nuevos clientes o rechazar otros, ampliar los servicios prestados a los clientes, etc.
- Bancos: datos de créditos pagados y no pagados, por ejemplo.

Pero ... las bases de datos son inmensas: millones y millones de datos, megas de memoria que entran cada día (e incluso cada hora) a formar parte de la base de datos.

Esto conlleva que muchas de las técnicas estadísticas habituales no sean operativas, porque para llevarlas a cabo es necesaria la presencia de un operador: con tales cantidades de datos harían falta meses de trabajo de estos operadores para obtener información que se precisa para dentro de una semana.

En el campo de las ciencias de la computación se han desarrollado técnicas específicas y otras se han tomado prestadas de la estadística o la inteligencia artificial, para analizar estas grandes bases de datos de forma eficiente. Muchas de estos nuevos procesos automatizan de forma eficiente algoritmos y procedimientos ya conocidos.

Estas técnicas se han agrupado bajo el nombre de **data mining**, término que podría traducirse por **minería de datos** o *explotación de datos*.

El lema de la minería de datos es convertir los datos en conocimiento.

EJEMPLO. Algunas aportaciones curiosas del *data mining*.

- Las personas que compran un coche Renault resultan ser mejores pagadores de créditos que el resto.
- Las personas que tienen ojos claros sufren accidentes de tráfico nocturnos que el resto.

1.3 Tipos de datos

Variable estadística es cada rasgo o característica que puede ser observada en los elementos de una población. Generalmente no toman el mismo valor en todos ellos.

EJEMPLO. Algunas variables estadísticas.

Estado civil, nivel de estudios, fecha (expresada en años), altura de una persona (expresada en centímetros).

Las variables estadísticas se clasifican según distintos criterios.

1. **Según cuántos valores puede tomar la variable.**
   Datos discretos: la variable toma sólo un número finito de valores (ej.: estado civil, nivel de estudios) o como mucho, una cantidad numerable

(ej.: goles marcados en un partido de futbol).
Datos continuos: Pueden tomar como valor cualquier número real que esté en un cierto intervalo (Ej: estatura, edad).

2. **Según la calidad de los datos.**
Hay operaciones que se pueden hacer con unas variables y no se pueden hacer con otras: ordenar, sumar, dividir.
No todas las variables son de la misma calidad. Las escalas de medida pueden ser de distinto tipo: nominal (*sexo*), ordinal (*nivel de estudios*), de intervalo (*fecha*), de razón (*estatura*).
Una clasificación relacionada es la que divide las variables en cualitativas o categóricas (incluye las variables nominales) y cuantitativas (incluye las ordinales, de intervalo y de razón).

Los valores de una variable estadística, sea del tipo que sea, pueden ser observados con distintos esquemas. Ello da lugar a distintas estructuras en los datos recogidos. Distinguiremos aquí tres de estos esquemas:

Datos de sección cruzada. Se observan una o más variables estadísticas en varios elementos de la población en el mismo instante.
Ejemplo: A principios del año 2000, se pregunta a los empleados de una empresa por los siguientes datos: *Número de hijos*, *Ingresos brutos anuales*, *¿dispone de seguro médico privado?*.

Datos de serie temporal. Se observa el valor de una variable en el mismo *individuo* periodicamente durante un lapso de tiempo prolongado. En ocasiones ese individuo es único: la población sólo tiene un elemento.
Ejemplo: Para el *individuo* Catalunya, se observa el día 15 de cada mes el número de personas registradas como desempleados en las oficinas del INEM de Catalunya. Esto se repite durante varios años.

Datos de Panel. Se observa el valor de una variable en varios *individuos* periodicamente durante un lapso de tiempo prolongado. Es una combinación de datos de sección cruzada y datos de series temporales.
Ejemplo: En la población de empresas del sector industrial, se seleccionan 100 de ellas y se anota su volumen de ventas cada mes durante unos años.

En los estudios de sección cruzada las observaciones de las variables estadística se registran usualmente en un fichero informático.

Habitualmente se mide el valor de p variables en n individuos de un población.

Los datos observados se ordenan para formar una matriz de datos.

	Variable 1	Variable 2	$\ldots$	Variable p
Observación 1	x_{11}	x_{12}	$\ldots$	x_{1p}
Observación 2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Observación n	x_{n1}	x_{n2}	$\ldots$	x_{np}

En el caso de las series temporales, la estructura de los ficheros de datos es como sigue. Se mide el valor de una variable en T instantes consecutivos (T fechas). Los datos observados se uelen almacenar

Fecha	Variable
1	x_1
2	x_2
$\vdots$	$\vdots$
T	x_T

1.4 Clasificación de los métodos de análisis de los datos en función del tipo de datos y los objectivos planteados

Tipo de datos	Objetivos	
Sección cruzada	Describir	Modelizar
Una variable discreta	(Tema 2)	(Tema 3)
Una variable continua	(Tema 2)	(Tema 3)
Dna variables discretas	Tablas de contingencia	Pruebas de independencia
Dos variable continuas	Regresión simple	Modelo de regresión simple (una variable explica la otra)
Una variable discreta y una continua	(Tema 2)	Análisis de la varianza (la variable discreta explica la otra; puede haber más de una discreta)
Varias variables discretas	Análisis de Correspondencias, análisis de conglomerados, escalamiento multidimensional	
Varias variables continuas	Componentes principales, análisis de conglomerados, escalamiento multidimensional	
Una variable continua y varias variables continuas que la explican	Modelo de regresión múltiple	Modelo de regresión múltiple
Una variable discreta y varias variables continuas que la explican	Análisis discriminante	Análisis discriminante, regresión logística
Una variable continua y varias variables discretas y continuas que la explican	Análisis de covarianzas	Análisis de covarianzas
Varias variables continuas y varias variables continuas que las explican	Correlación canónica	Correlación canónica
Varias variables continuas y varias variables discretas que las explican	Análisis de la varianza multivariante	Análisis de la varianza multivariante
Series temporales		
Una variable continua	(Tema 5)	(Tema 5)
Una variable continua y varias continuas que la explican		*Econometría*
Esquemas más complicados son poco habituales		