

Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS
MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 1. Introducción

Objetivos

- Definir e identificar el papel de la estadística como ciencia en el análisis de datos
- Diferenciar entre estadística descriptiva e inferencial
- Entender la importancia del muestreo en estadística y su problemática
- Encontrar la forma más adecuada de tabular y representar datos
- Ser capaces de interpretar distintos tipos de gráficos

Contenido

- ¿Qué es la estadística?
- Utilidad de la Estadística
- Taxonomía de la estadística
- Conceptos previos
- Tipos de variables
- Tipos de estudios estadísticos (una taxonomía)
- Razonamiento Estadístico
- Próxima sesión

¿Qué es la estadística?

La estadística es la **ciencia** y, posiblemente, también el **arte** de aprender de los datos. Como disciplina, se ocupa de al menos:

- la recopilación,
- el análisis
- la interpretación
- Extrapolación
- La predicción
- la presentación eficaces de los resultados
- la comunicación

La estadística está en el centro del tipo de razonamiento cuantitativo necesario para realizar importantes avances en las ciencias, como la medicina y la genética, y para tomar decisiones importantes en los negocios y la política pública.

La materia prima de la estadística son la incertidumbre y la variación.

¿Qué es la estadística?

La estadística es un campo altamente **interdisciplinar**; la investigación en estadística **encuentra aplicación en prácticamente todos los campos científicos**



las cuestiones de investigación en los distintos campos científicos motivan el desarrollo de nuevos métodos y teoría estadística.

Utilidad de la Estadística

Entender, medir, evaluar y predecir. Apoyar la planeación y la toma de decisiones consciente.

Se ha convertido en una herramienta esencial en muchas disciplinas:

- Medicina: Evidencia de la investigación clínica, factores de riesgo de enfermedades, comparación de tratamientos, etc. Epidemiología.
- Psicología: La Psicometría es la rama de la Psicología Experimental que se encarga de la medición y cuantificación de los procesos psicológicos y las capacidades cognitivas.
- Finanzas: Riesgos financieros, detección de fraudes, previsiones financieras, pronóstico, etc.
- Marketing: campañas de venta en Marketing, Marketing dirigido, fidelidad, etc.
- Medio ambiente: Previsiones meteorológicas, cambio climático, etc.
- Ingeniería: Medición de errores del sistema, diseño de experimentos, control de calidad, fiabilidad, etc.
- Ciencias Sociales: Evaluación de políticas, programas y proyectos. Desempleo, demografía, elecciones, Educación, etc.
- Deportes: Calificar a los jugadores, rendimiento de equipos, encontrar puntos débiles del adversario, etc.
- Investigación. Da soporte bibliométrico a las propuestas y proyectos, entre otros aspectos.
- Entretenimiento. Contenidos interactivos. Contenidos de streaming (TV y radio). Contenidos bajo demanda (música y cine)
- IA
- ...

Utilidad de la Estadística

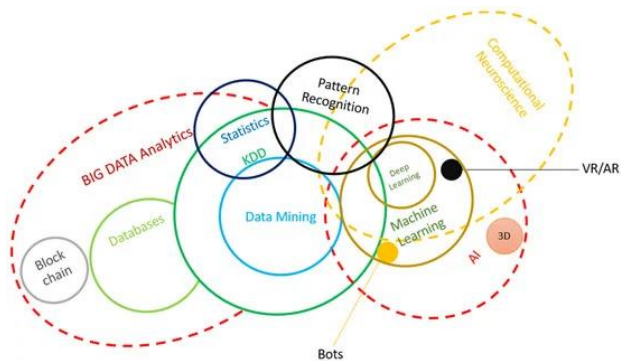
...la estadística, como campo científico interdisciplinario, desempeña un papel sustancial tanto para la comprensión teórica y práctica de la IA y para su futuro desarrollo. La estadística podría incluso considerarse un elemento central de la IA. Con sus conocimientos especializados de la evaluación de datos, empezando por la formulación precisa de la pregunta de investigación y pasando por la fase de diseño del estudio hasta el análisis y la interpretación de los resultados, la estadística **es un socio natural de otras disciplinas en la enseñanza, la investigación y la práctica...**

Friedrich, S., Antes, G., Behr, S. et al. **Is there a role for statistics in artificial intelligence?**. Adv Data Anal Classif (2021).

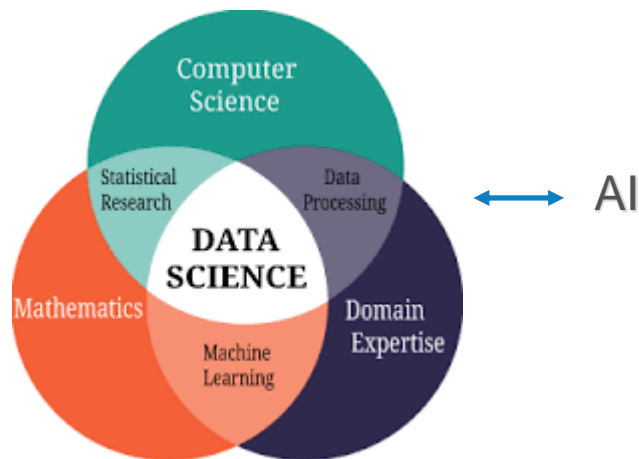
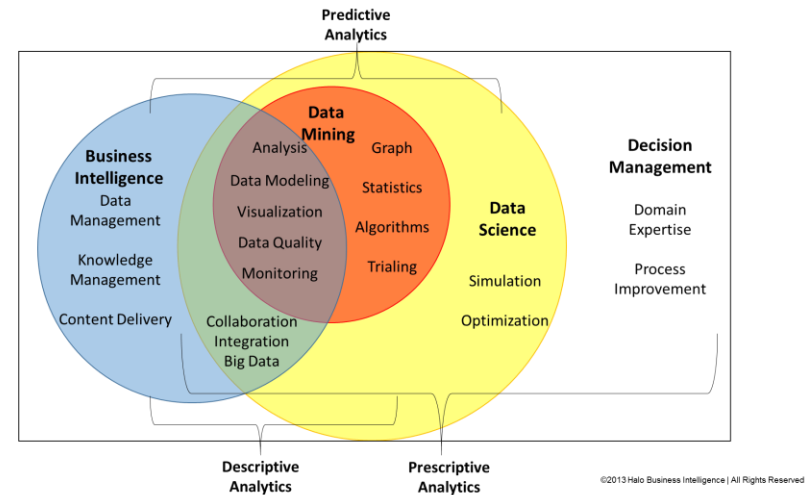
<https://doi.org/10.1007/s11634-021-00455-6>

<https://link.springer.com/content/pdf/10.1007/s11634-021-00455-6.pdf>

Utilidad de la Estadística

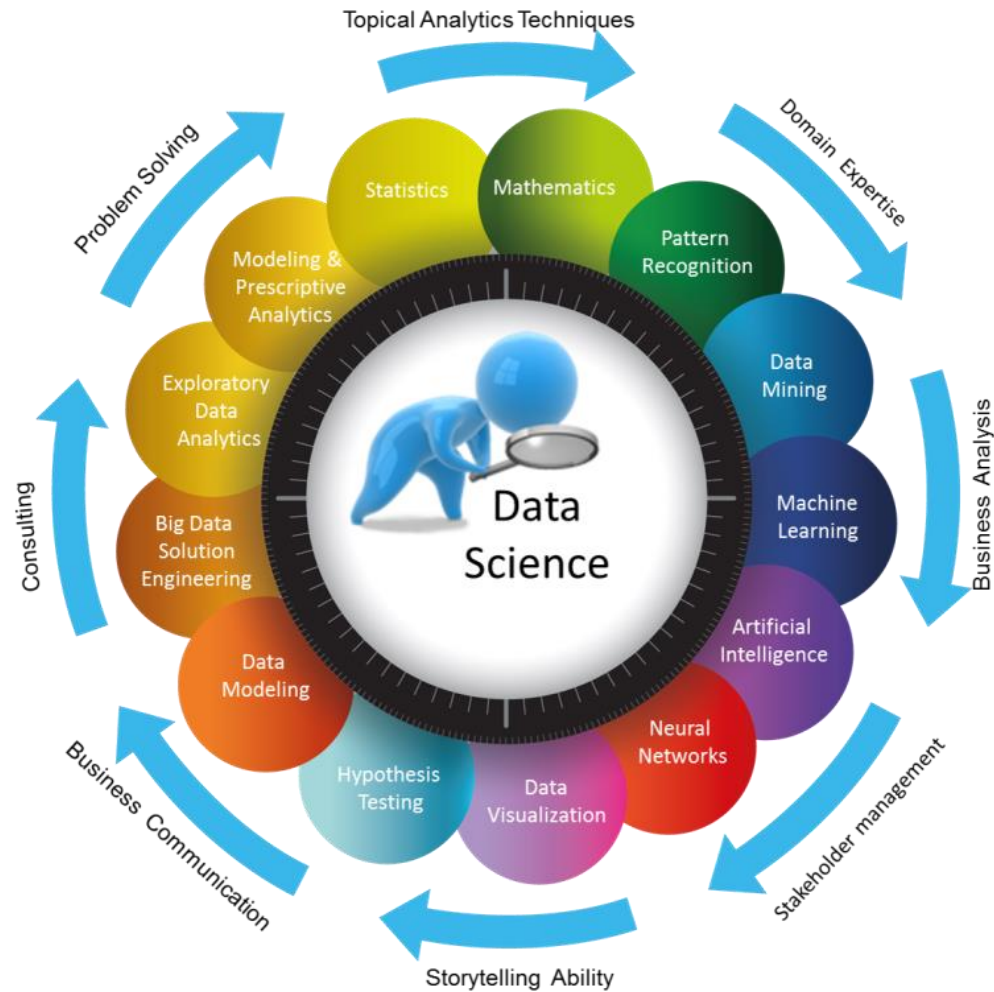


<https://www.mdpi.com/2504-2289/6/1/18/htm>



<https://library.lclark.edu/digital/hpc>

Utilidad de la Estadística



<https://sameerdhanrajani.wordpress.com/2015/01/12/data-science-the-new-monetization-model-for-analytics-industry/>

Taxonomía

ESTADÍSTICA DESCRIPTIVA



Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos. Podemos incluir la estadística multivariada (encuentra relaciones entre variables, define segmentos, clasifica, perfila, ...). Modelos no supervisados.

ESTADÍSTICA INFERENCIAL



A partir del cálculo de probabilidades y datos muestrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.
Incluye:
Muestreo, Pronóstico de series temporales. Modelos supervisados.

VISUALIZACIÓN

Sobre datos-variables aleatorias X de interés

Conceptos previos

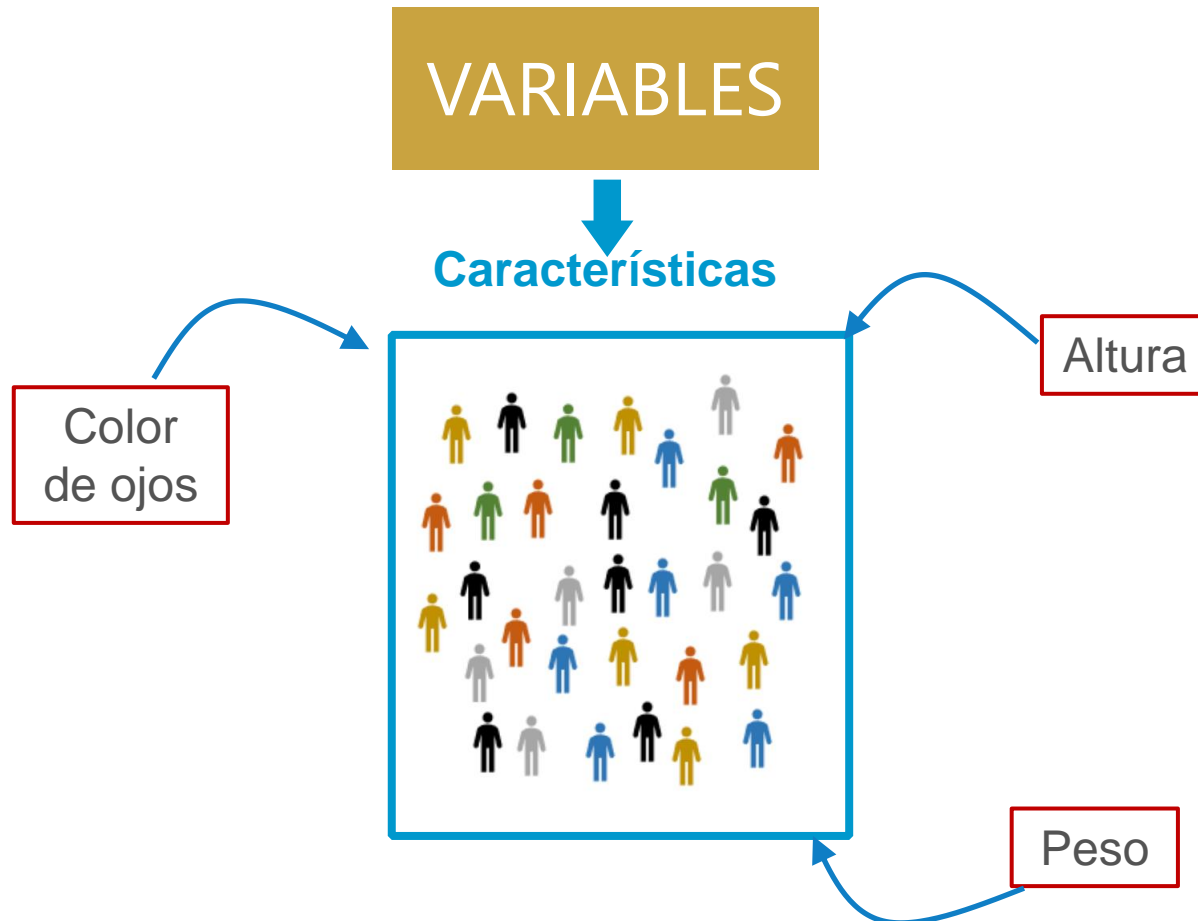
- Variable: Características y atributos, preferiblemente medibles de una población.
- Dato: No es solo un Número. Número sobre algo. Información en contexto. Es una historia.
- Individuo: Es un elemento del conjunto “objeto de estudio” sobre lo que se recolecta la información.
- Población: conjunto finito o infinito de individuos que son objeto de estudio.
- Muestra: Cualquier subconjunto de la población. Puede ser aleatoria o no aleatoria. Ideal que sea aleatoria y representativa (recoja la dispersión de la población de origen).

El Censo es un tipo de muestra que tiene en cuenta toda la población.

Taxonomía



Conceptos previos



Conceptos previos

Ej. Los datos `Pennstate2.csv` corresponden a una **encuesta realizada a 205 estudiantes de una clase de estadística para estudiantes de ciencias sociales y del comportamiento**. La encuesta se realizó en el semestre de primavera de 2000. Estas son las primeras filas de los datos (salida de R):

```
> head(PennState2)
```

	Sex	EarPrces	Tattoo	CDs	Height	HtChoice	Looks	Friends
1	Female	2	No	25	70	70	4	Opposite
2	Male	0	No	47	77	77	6	NoDiff
3	Male	2	No	50	71	71	5	Opposite
4	Female	2	No	50	65	65	4	Opposite
5	Male	4	Yes	50	74	72	5	NoDiff
6	Female	9	Yes	70	61	65	7	Opposite

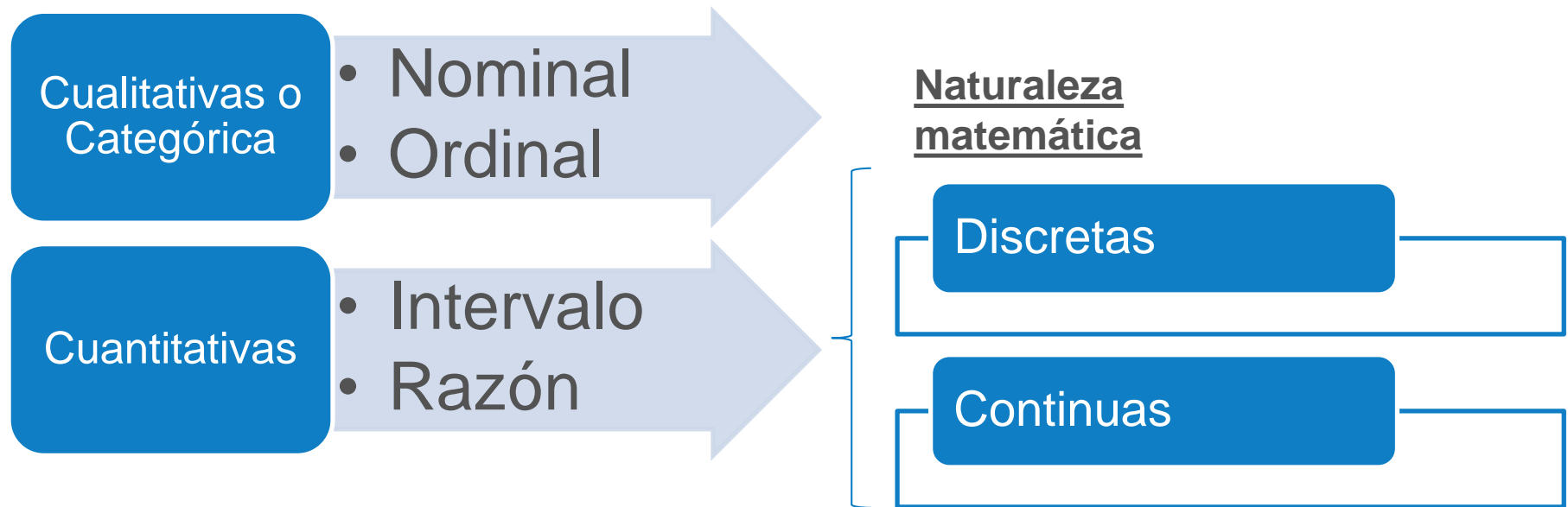
Dato

Variable

Unidad
estadística,
individuo o
Registro

Tipos de variables

El tipo de variable es importante: afecta a lo que podemos hacer con ella, al tipo de análisis que podemos hacer. Los métodos estadísticos y econométricos que se usan dependen del tipo de variable



Tipos de variables

Nominal:

Los valores son “categorías” mutuamente excluyentes.

Las categorías son valores diferentes por una cualidad, no por una cantidad.

Ningún “valor” se puede decir que sea mayor o menor que otro.

Ejemplos:

País; región, tipo de producto, marca, ...

Ordinal:

Sus valores son “categorías” mutuamente excluyentes.

Existe Orden. Este tipo de variables nos permite establecer relaciones de igualdad/desigualdad y a su vez, podemos identificar si una categoría es mayor o menor que otra.

Ejemplos:

Clase social (baja, media, alta), nivel de riesgo,...

Tipos de variables

Intervalo o razón:

Son variables numéricas cuyos valores representan magnitudes y la distancia entre los números de su escala es igual. Con este tipo de variables podemos realizar comparaciones de igualdad/desigualdad, establecer un orden dentro de sus valores y medir la distancia existente entre cada valor de la escala. Las variables de intervalo carecen de un cero absoluto, por lo que operaciones como la multiplicación y la división no son realizables.

Transformación $Y=ax+b$

Ejemplos:

Temperatura, escala medición en notas (0-5; 1-8) ...

Razón:

Las variables de razón poseen las mismas características de las variables de intervalo, con la diferencia que cuentan con un cero absoluto; es decir, el valor cero (0) representa la ausencia total de medida, por lo que se puede realizar cualquier operación *Aritmética* (Suma, Resta, Multiplicación y División) y *Lógica* (Comparación y ordenamiento). Este tipo de variables permiten el nivel más alto de medición.

Transformación $Y=ax$

Ejemplos:

Rentabilidad, retorno, ingreso, precio,...

Tipos de variables

Cuantitativas. Por naturaleza matemática.

Discreta:

El número de valores posibles entre dos valores dados es finito. Los valores pueden ser números enteros.

Ejemplos:

Son el resultado de contar (personas en el hogar, tamaño de los municipios, número de acciones,...)

Continua:

El número de valores posibles entre dos valores es infinito.

Ejemplos:

Rentabilidad, retorno, ingreso, precio,...

Tipos de variables

Hay ocho columnas de datos:

C1 Sex: Masculino o Femenino

C2 EarPrces: Perforaciones en las orejas Total de perforaciones en las dos orejas

C3 Tatto: El estudiante tiene un tatuaje (Sí o No)

C4 CDs: Estimación del estudiante sobre el número de CDs de música que posee

C5 Height: Estatura auto declarada en pulgadas

C6 HtChoice: Altura que el estudiante elegiría tener si pudiera elegir cualquier altura

C7 Looks: En una escala de 0-9, importancia de la apariencia frente a la personalidad. (0) nada importante (9) Muy importante

C8 Friends: Sexo con el que es más fácil hacer amigos (opuesto, igual, indiferente)

```
> head(PennState2)
```

	Sex	EarPrces	Tattoo	CDs	Height	HtChoice	Looks	Friends
1	Female	2	No	25	70	70	4	Opposite
2	Male	0	No	47	77	77	6	NoDiff
3	Male	2	No	50	71	71	5	Opposite
4	Female	2	No	50	65	65	4	Opposite
5	Male	4	Yes	50	74	72	5	NoDiff
6	Female	9	Yes	70	61	65	7	Opposite

Dicotómica

**Cuantitativa
Discreta**

**Cuantitativa
Continua**

???

**Cualitativa
Nominal**

Tipos de variables

Otras clasificaciones

- Según comportamiento:
Lineales Vs No Lineales
- Según Carácter temporal:
Series de Tiempo Vs Corte Transversal Vs Paneles de Datos
- Según “causalidad”:
Exógena Vs Endógena
- Por dependencia:
Dependiente(respuesta-explicada) o independiente (explicativa-predictora), variables omitidas.
- Otras
Variables dicotómicas (binarias, dummy)

Tipos de estudios estadísticos (una taxonomía).

Ejemplo. Encuestas.
Estadísticas de un
fenómeno-covid-19



Observacionales

Son aquellos en los recogemos datos observando por lo que no intervenimos ni alteramos a los individuos de ningún modo.

Experimentales

Aplicamos tratamientos y luego observamos sus efectos sobre sus sujetos, que aquí pasan a llamarse unidades experimentales.



Ejemplo. Ensayos clínicos.
RCT

Razonamiento Estadístico

Pensamiento crítico tiene en cuenta las siguientes preguntas:

- ¿Cuál es el objetivo del estudio? (comprensión del problema en su contexto)
- ¿Cuál es la unidad de análisis?
- ¿Cuáles son las fuentes de los datos?
- ¿Qué tipos de variables se tienen?
- ¿Con qué tipo de muestreo han sido obtenidos los datos? (error de muestreo)
- ¿Existen variables (intermedias) que influyan en los resultados y se hayan omitido?
- ¿Las tablas y gráficas resumen adecuadamente los datos?
- ¿Qué tipos de errores y sesgo (“manipulación” voluntaria o involuntaria) que podemos encontrar?
- ¿Las conclusiones se extraen directa y naturalmente de los datos?

Instalación de R

Para instalar estos programas se puede usar lo siguiente:

Instalar R para Windows:

<https://cran.r-project.org/bin/windows/base/>



Instalar R para MacOS:

<https://cran.r-project.org/bin/macosx/>

Instalar Rstudio para Windows o MacOS:

Descarga e instala la última versión que corresponda.

<https://www.rstudio.com/products/rstudio/download/>

OS	Download	Size	SHA-256
Windows 10/11	 RStudio-2022.07.2-576.exe	190.49 MB	b38bf925
macOS 10.15+	 RStudio-2022.07.2-576.dmg	224.49 MB	35028d02

□ Tema 1: Introducción a la estadística.

- Distribución de frecuencias.
- Tabulación de variables
- Gráficas básicas
- Elegir gráficos adecuados
- Aplicación de las TIC
- Retos de la estadística en el Big Data



www.unir.net