

Introducción

Estadística Computacional - 2025

Ricardo Ñanculef, Carolina Saavedra
jnancu@inf.utfsm.cl, carolina.saavedra@usm.cl

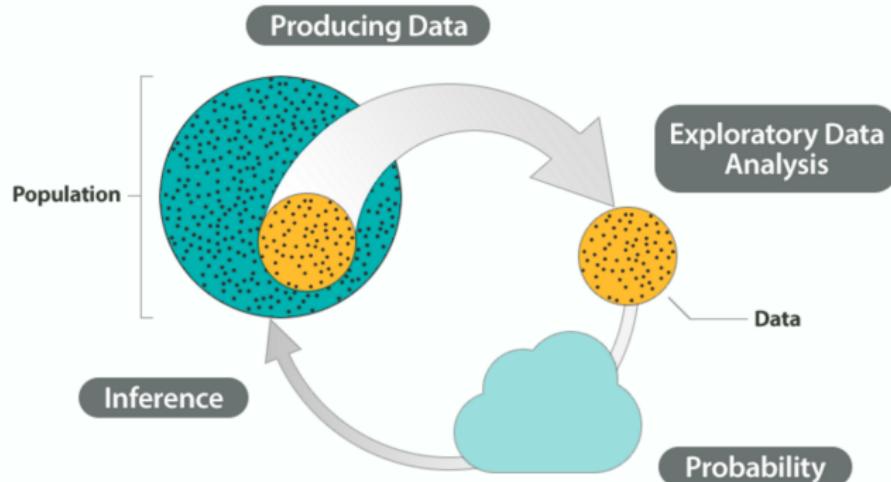
Departamento de Informática UTFSM



Departamento de Informática
Universidad Técnica Federico Santa María

Propósito y Partes de la Estadística

- ▶ Objetivo: Obtener conclusiones acerca de un fenómeno aleatorio mediante la recolección y análisis de **datos** experimentales.
- ▶ Partes: Teoría de Probabilidades, Inferencia, Análisis Exploratorio de Datos.



Propósito y Partes de la Estadística

Objetivo de la Teoría de Probabilidades

Obtener conclusiones acerca de un fenómeno aleatorio a partir de un **modelo** para el fenómeno.

Ejemplo

1. Los procesos de una empresa dependen de la correcta operación de un cierto equipo. La empresa está interesada en anticipar una falla en estos equipos.
2. Si el tiempo que dura el equipo se denota por X , un modelo posible para X es $X \sim \text{Exp}(\theta = 2)$.
3. A partir del modelo anterior, ¿Cuál es la probabilidad de que un cierto equipo dure más de $T = 2$ años antes de tener que enviarlo a mantenimiento?
4. La respuesta se traduce en calcular $p = P(X > 2) = e^{-1} \approx 0.368$.

Propósito y Partes de la Estadística

Objetivo de la Inferencia Estadística

Obtener conclusiones acerca de un fenómeno aleatorio a partir de **datos** experimentales.

Ejemplo

1. Los procesos de una empresa dependen de la correcta operación de un cierto tipo de equipo. La empresa está interesada en anticipar una falla en estos equipos.
2. Las últimas 10 veces que una de las unidades ha fallado, el tiempo de vida del equipo (en años) eran de $\{1.2, 1.5, 2.8, 2.5, 1.5, 3.5, 2.2, 1.6, 1.2, 2.4\}$.
3. A partir de los datos disponibles, ¿Cuál es la probabilidad de que un cierto equipo dure más de $T = 2$ años antes de tener que enviarlo a mantenimiento?

Propósito y Partes de la Estadística

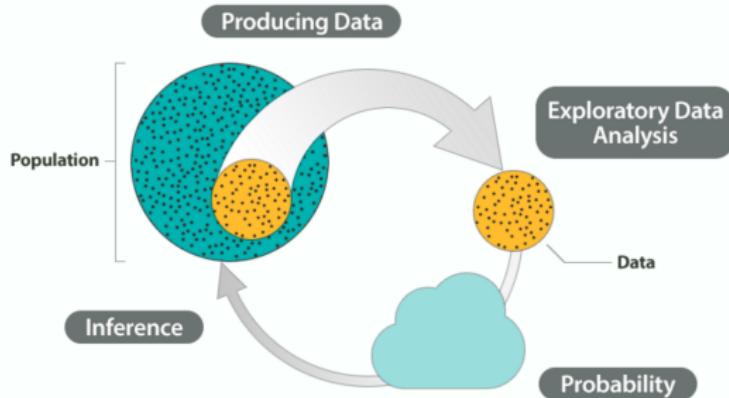
- La Inferencia Estadística y la Teoría de Probabilidades se pueden “reconciliar” haciendo que nuestros modelos del fenómeno dependan de los datos.

Ejemplo

1. A partir de los datos $\{1.2, 1.5, 2.8, 2.5, 1.5, 3.5, 2.2, 1.6, 1.2, 2.4\}$, podemos concluir nuestros equipos han durado en promedio $\bar{X} = 2.04$ años.
2. Si el tiempo que dura el equipo se denota por X , un modelo razonable para X es entonces $X \sim \text{Exp}(\theta = \bar{X} = 2.04)$.
3. La respuesta se traduce en calcular $p = P(X > 2) = e^{-2/2.04} \approx 0.375$.

¿Porqué es Necesario un Modelo?

- ▶ Si los datos son $\{1.2, 1.5, 2.8, 2.5, 1.5, 3.5, 2.2, 1.6, 1.2, 2.4\}$, no podríamos estimar $p = P(X > 2)$ simplemente como $\hat{p} = 5/10 = 0.5$?



Poblaciones y Muestras

Conceptos claves

- ▶ **Población:** conjunto de todos los individuos o elementos que nos interesa caracterizar.
- ▶ **Variable:** característica que nos interesa medir en un individuo de la población.
- ▶ **Muestra:** subconjunto finito de la población que nos permite estudiar la variable de interés.
- ▶ **Datos:** conjunto de mediciones de la variable de interés en la muestra.

Poblaciones y Muestras

Ejemplo

- ▶ Fenómeno Aleatorio: Sobrepeso Infantil en Chile.
- ▶ Pregunta: ¿Cuántos niños chilenos tiene sobrepeso ($IMC > 25$)?
- ▶ **Población:** conjunto de **todos los niños chilenos al 2018**.
- ▶ **Variable:** IMC del niño, x .
- ▶ **Muestra:** subconjunto finito de niños que efectivamente se pesaron para concluir algo acerca de la población.
- ▶ **Datos:** IMC de los niños seleccionadas para estudio: x_1, x_2, \dots, x_n .

¿Porqué es Necesario un Modelo?

- ▶ Si estudiando 10 niños estudiados, 2 son obesos, ¿podemos concluir que *más de un 15 % de todos los niños chilenos* son obesos?
- ▶ Si estudiando 100 niños estudiados, 20 son obesos, ¿podemos concluir que *más de un 15 % de todos los niños chilenos* son obesos? minutos?

Problema Fundamental de la Estadística

- ▶ ¿Podemos obtener conclusiones sobre toda la población a partir de una muestra finita?
- ▶ ¿Qué tan probable es que nuestras conclusiones sean equivocadas?

Poblaciones y Muestras

Observaciones

- ▶ En muchos casos, los datos disponibles no han sido obtenidos seleccionando explícitamente elementos de una población pre-definida.
- ▶ **Población Conceptual:** En estos casos, el dilema estadístico fundamental sigue estando presente. Para abordarlo, es útil pensar en los datos como si hubiesen sido generados de una población conceptual, hipotética, o virtual.
- ▶ **Ejemplo:** en el caso de la duración de los equipos, la población es conceptual, y corresponde a todos los equipos de características similares a los actuales que la empresa podría emplear para sus procesos.
- ▶ **Poblaciones como Distribuciones (Desconocidas):** Para fines matemáticos, y sobre todo en el caso de poblaciones hipotéticas, podemos pensar en la población como la distribución de probabilidad (real y desconocida) desde de la cuál se toma un número finito de observaciones.

Poblaciones, Variables y Muestras

- ▶ Si estudiando 10 niños estudiados, 2 son obesos, ¿podemos concluir que *más de un 15 % de todos los niños chilenos* son obesos?
- ▶ Si estudiando 100 niños estudiados, 20 son obesos, ¿podemos concluir que *más de un 15 % de todos los niños chilenos* son obesos? minutos?

Tipos de Muestreo

Diremos que la muestra que se obtiene de la Población es **aleatoria** si cada elemento de la Población tiene la misma oportunidad de ser muestreado. Otros muestreos siguen criterios que dificultan el análisis y la validez de un razonamiento inductivo (e.g. muestreo por conveniencia o por juicio).

- ▶ Por ahora asumiremos que la muestra es de tipo **aleatorio** y que esta ya ha sido recogida dando origen a un **dataset** que debe ser explorado/analizado en busca de conclusiones.
- ▶ En el caso de poblaciones conceptuales, el supuesto anterior se puede formalizar usando la definición matemática de independencia.

Outline

ANÁLISIS EXPLORATORIO DE DATOS

Análisis Exploratorio de Datos

Objetivo

Generar representaciones gráficas y numéricas que describan y resuman los datos disponibles para **formular hipótesis** acerca de la población, detectar posibles patrones o relaciones entre las variables estudiadas, además de identificar errores o eventos anómalos en las mediciones.

- ▶ Visualización: Presentación Gráfica de Datos.
- ▶ Agregación: Medidas de Tendencia y Dispersión.

- ▶ En la definición anterior, formular/detectar/identificar podrían cambiarse por “exponer” .

Análisis Exploratorio de Datos

Visualización

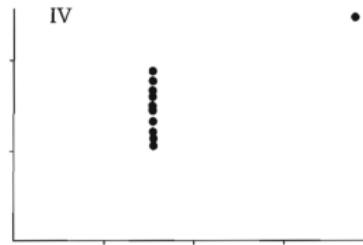
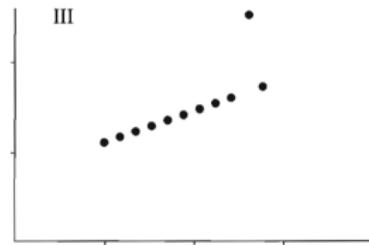
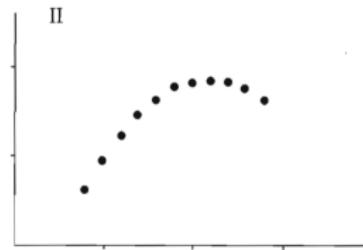
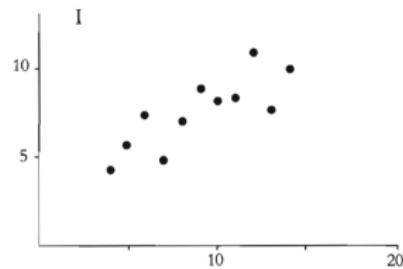
Uso de recursos gráficos (áreas, colores, líneas, sombras, etc) para resumir y presentar datos.



Análisis Exploratorio de Datos

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Análisis Exploratorio de Datos

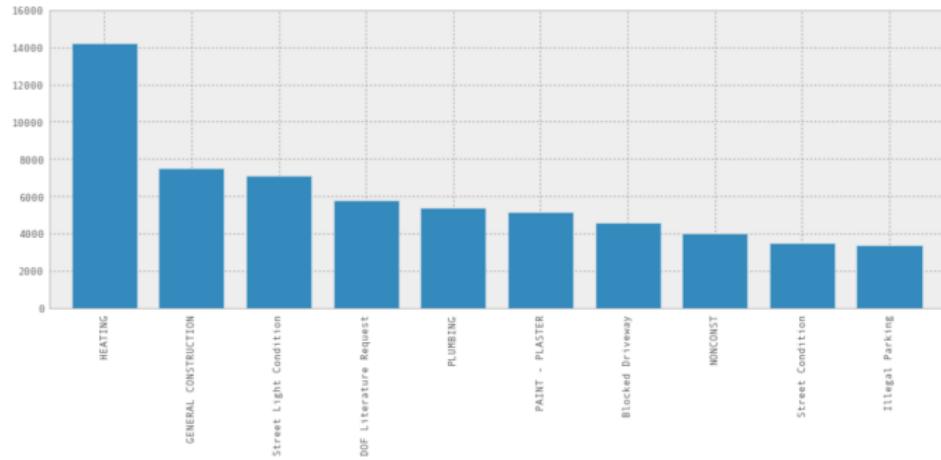


Análisis Exploratorio de Datos

	Agency Name	Complaint Type	Descriptor	Location Type
1	Department of Transportation	Street Condition	Pothole	
2	New York City Police Department	Noise - Street/Sidewalk	Loud Music/Party	Street/Sidewalk
3	New York City Police Department	Noise - Commercial	Banging/Pounding	Club/Bar/Restaurant
4	New York City Police Department	Noise - Commercial	Banging/Pounding	Club/Bar/Restaurant
5	New York City Police Department	Noise - Commercial	Banging/Pounding	Club/Bar/Restaurant
6	New York City Police Department	Noise - Street/Sidewalk	Loud Music/Party	Street/Sidewalk
7	New York City Police Department	Blocked Driveway	No Access	Street/Sidewalk
8	New York City Police Department	Noise - Commercial	Loud Music/Party	Store/Commercial
9	Department of Transportation	Street Condition	Pothole	
10	New York City Police Department	Noise - Commercial	Loud Music/Party	Store/Commercial
11	New York City Police Department	Illegal Parking	Blocked Hydrant	Street/Sidewalk
12	New York City Police Department	Blocked Driveway	No Access	Street/Sidewalk
13	New York City Police Department	Noise - Commercial	Loud Music/Party	Store/Commercial
14	New York City Police Department	Noise - Commercial	Loud Music/Party	Club/Bar/Restaurant
15	New York City Police Department	Noise - Commercial	Loud Music/Party	Club/Bar/Restaurant
16	Department of Health and Mental Hygiene	Rodent	Rat Sighting	1-2 Family Dwelling
17	New York City Police Department	Noise - Commercial	Banging/Pounding	Club/Bar/Restaurant
18	New York City Police Department	Noise - Commercial	Banging/Pounding	Club/Bar/Restaurant

Análisis Exploratorio de Datos

```
In [12]: complaint_counts[:10].plot(kind='bar')  
Out[12]: <matplotlib.axes.AxesSubplot at 0x7ba2290>
```



Análisis Exploratorio de Datos

Medidas de Tendencia y Dispersión

Medidas de agregación numéricas para resumir y presentar datos.

Table 1. Summary statistics for the mobile-monitoring sampling data.

Parameter	Observations (<i>n</i>)	Mean ± SD	Median	5th percentile	95th percentile
UFP concentration (particles/cm ³)	8,225	44,000 ± 24,800	39,800	15,900	87,500
PM _{2.5} concentration (µg/m ³) ^a	8,354	36 ± 30	29	10	129
PAH concentration (ng/m ³)	7,453	76 ± 55	55	8	212
Traffic count per minute					
WB	9,598	13.7 ± 2.3	13.2	10.2	17.7
BQE	9,553	36.7 ± 6.5	38.4	24.5	44.9
Wind speed (m/sec)	7,913	1.3 ± 1.0	0.9	0.4	3.6
Temperature (°C)	9,441	26.3 ± 3.5	26.7	19.8	30.7
RH (%)	9,441	45.8 ± 11.3	45.4	27.3	66.4

^aMeasured using DustTrak, which has a known but consistent bias by a factor of 2.5–3 relative to gravimetric measurements (Chang et al. 2001).

Tipos de Datos I

Taxonomía de Stevens (1946)

De acuerdo a los valores que una variable puede tomar, podemos distinguir distintos tipos de variable:

- ▶ **Cualitativo**: valores no son operables aritméticamente.
- ▶ **Cuantitativo ó Numérico**: valores son operables aritméticamente.

Las variables de tipo cualitativo pueden ser:

- ▶ **Categóricas**: valores sólo soportan la operación $=$ y \neq . Ejemplo:
 $\{banana, manzana, pera\}$.
- ▶ **Ordinales**: valores soportan además las operaciones $<$ y $>$. Ejemplo:
 $\{pesimo, malo, medio, bueno, excelente\}$.

Tipos de Datos II

Taxonomía de Stevens (1946)

Las variables de tipo cuantitativo pueden ser:

- ▶ **Intervalares:** Cuando el valor 0 es arbitrario y no indica “ausencia de variable”. Operaciones posibles: $=, \neq, <, >, +, -$. Asigna números para posición en escala. Ejemplo: latitud/longitud, fechas, talla de zapatos.
- ▶ **De Razón:** Cuando existe un 0 absoluto. Operaciones posibles: $=, \neq, <, >, +, -, \cdot, /$. Los cuocientes entre valores tienen sentido. Ejemplo: peso, altura, temperatura en K.

Tipos de Datos II

Taxonomía de Stevens (1946)

Las variables de tipo cuantitativo pueden ser:

- ▶ **Intervalares:** Cuando el valor 0 es arbitrario y no indica “ausencia de variable”. Operaciones posibles: $=, \neq, <, >, +, -$. Asigna números para posición en escala. Ejemplo: latitud/longitud, fechas, talla de zapatos.
- ▶ **De Razón:** Cuando existe un 0 absoluto. Operaciones posibles: $=, \neq, <, >, +, -, \cdot, /$. Los cuocientes entre valores tienen sentido. Ejemplo: peso, altura, temperatura en K.

Datos Discretos versus Continuos

Además, las variables de tipo cuantitativo pueden ser:

- ▶ **Discretas:** toman un conjunto finito o infinito enumerable de valores (\mathbb{N}, \mathbb{Z})
- ▶ **Continuas:** toman valores en un subconjunto completo de la recta real.

Tipos de Datos III

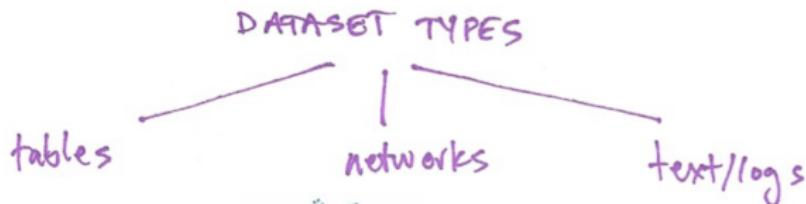
Dimensionalidad

En ocasiones nos interesará medir más de una cantidad sobre cada individuo de la Población. En este caso x será un vector en vez de un escalar. Llamaremos **dimensionalidad** al Número de cantidades que componen la medición.

- ▶ **Univariadas:** 1 cantidad.
- ▶ **Bivariadas:** 2 cantidades.
- ▶ **Multivariadas:** muchas cantidades.

Tipos de Datos IV

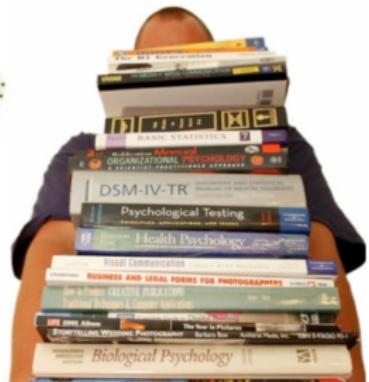
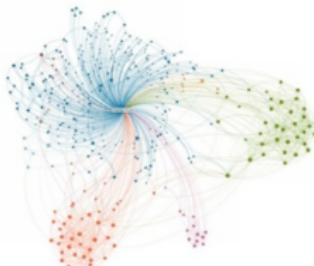
Nuevas Taxonomías



Google Docs

FriendFeed Audience

Site Name	Category	Components	Unique Visits	Country	Re Page	Page Views	Google
friendfeed.com	Online Cor.	1600000	1500000	0.1	2000000		
twinkl.org	Online Cor.	400000	300000	0	740000		
freeebookson.com	Online Cor.	43000	180000	0	130000		
christianmag.com	Online Cor.	29000	29000	0	74000		
christianmag.com	Home & G.	29000	29000	0	93000		
budget.com	Home & G.	24000	71000	0	340000	TRUE	
web-strategists.com	Online Cor.	32000	34000	0	86000		
twinkl.com	Home & H.	20000	34000	0	570000		



Outline

VISUALIZACIÓN DE DATOS

Análisis Exploratorio de Datos

Visualización de Datos - Métodos Pictóricos

Generar representaciones gráficas que describan los datos disponibles y que permitan obtener **hipótesis preliminares** acerca del **Fenómeno estudiado**.

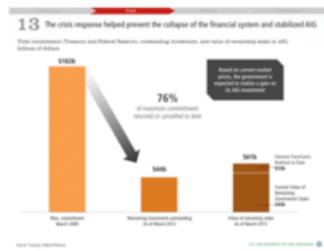
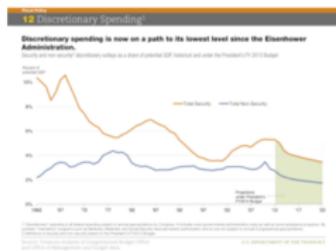
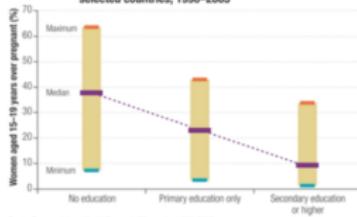


Figure 2 Adolescence pregnancy rates by educational level, selected countries, 1990–2005



Sencillo de visualizar y entender.

Análisis Exploratorio de Datos

Frecuencias Absolutas y Relativas

- ▶ Sea x una variable discreta de interés y sea n el tamaño de la muestra.
- ▶ Sean c_1, c_2, \dots, c_K los distintos valores posibles.
- ▶ Las **frecuencias absolutas** correspondientes al valor c_i se denotan n_i y se definen como el número de ocurrencias del valor c_i en la muestra.
- ▶ Las **frecuencias relativas** (o densidades) correspondientes al valor c_i se denotan f_i se definen como $f_i = n_i/n$.

Análisis Exploratorio de Datos

Frecuencias Absolutas y Relativas

- ▶ Sea x una variable discreta de interés y sea n el tamaño de la muestra.
- ▶ Sean c_1, c_2, \dots, c_K los distintos valores posibles.
- ▶ Las **frecuencias absolutas** correspondientes al valor c_i se denotan n_i y se definen como el número de ocurrencias del valor c_i en la muestra.
- ▶ Las **frecuencias relativas** (o densidades) correspondientes al valor c_i se denotan f_i se definen como $f_i = n_i/n$.
- ▶ También se puede definir las **frecuencias acumuladas** hasta el valor i como

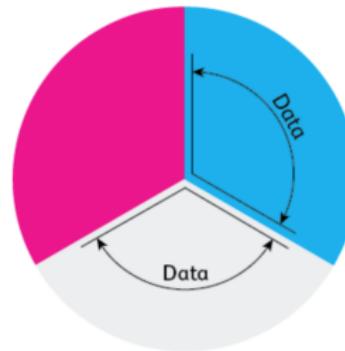
$$N_i = \sum_{j=1}^i n_j$$

$$F_i = \sum_{j=1}^i f_j$$

Análisis Exploratorio de Datos

Diagrama de Torta (*Pie chart*)

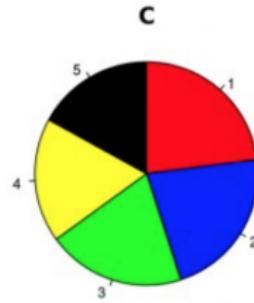
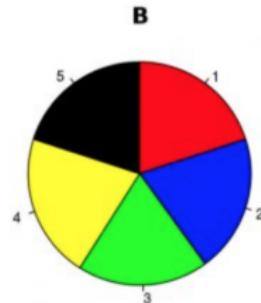
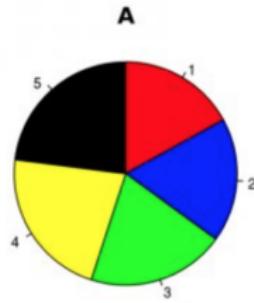
Representación gráfica a forma de disco con diferentes sectores correspondientes a cada valor posible de la variable. El área de cada sector es proporcional a la frecuencia del valor correspondiente.



Análisis Exploratorio de Datos

Diagramas de Torta (*Pie chart*)

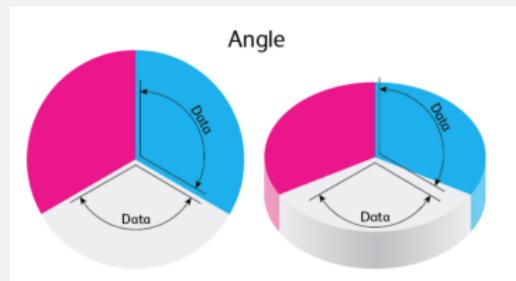
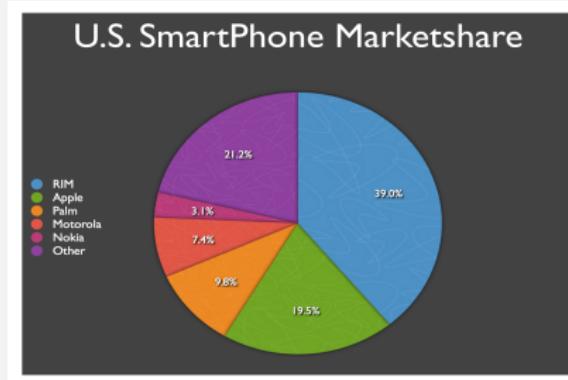
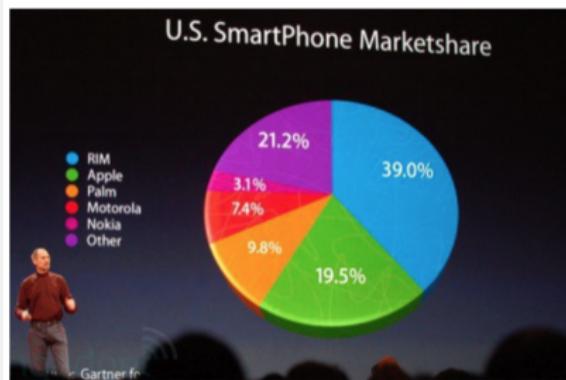
- ▶ Inferir valores desde ángulos o áreas es complejo para muchas personas.
- ▶ Comparar ángulos o áreas entre tortas no es fácil.



- ▶ Más apropiado cuando la variable puede tomar pocos valores.
- ▶ Ampliamente considerado un **mal tipo de gráfico** que es mejor evitar.

Análisis Exploratorio de Datos

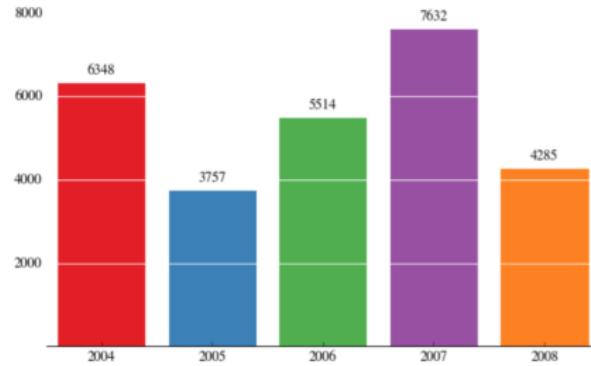
Distorsiones en Diagramas de Torta



Análisis Exploratorio de Datos

Diagramas de Barras (*bar chart*)

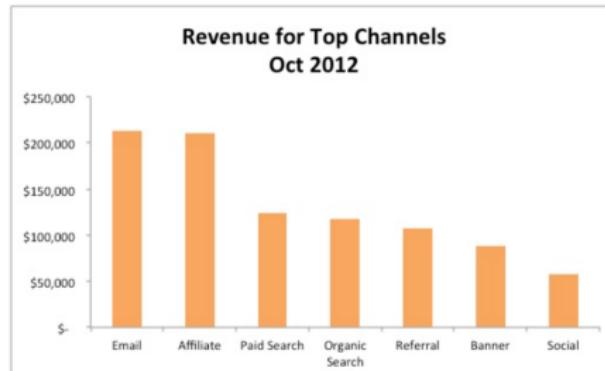
Representación gráfica como secuencia de barras de igual ancho, donde la longitud de cada barra es proporcional a la frecuencia del correspondiente valor en la muestra. Áreas son también proporcionales.



Análisis Exploratorio de Datos

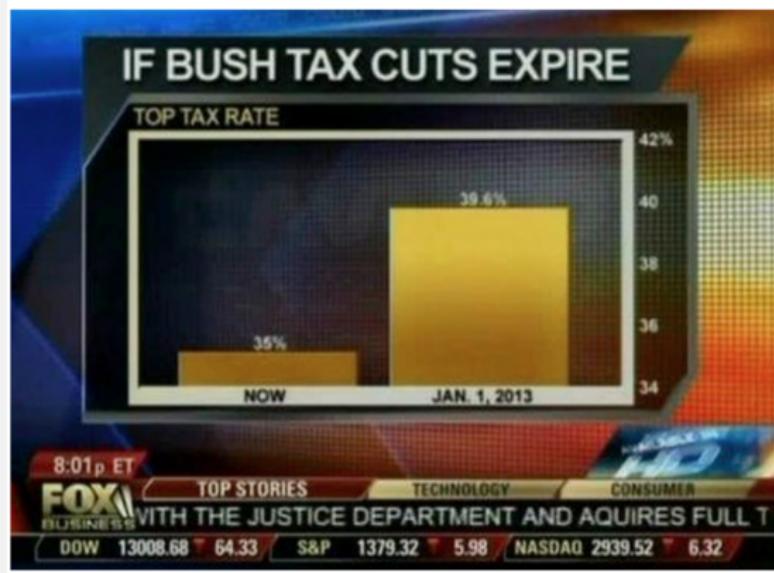
Orden en Diagramas de Barras

- ▶ Si hay un orden “natural” en los valores posibles, las barras se suelen ordenar con ese criterio.
- ▶ *Criterio de Pareto:* las barras se ordenan en forma decreciente de acuerdo a la frecuencia de cada valor de x .

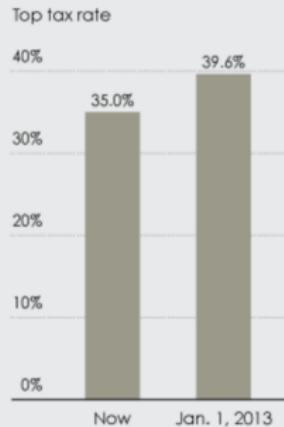


Análisis Exploratorio de Datos

Distorsiones en Diagramas de Barras (*bar chart*)



If Bush tax cuts expire...



Análisis Exploratorio de Datos

Tabligráma o Diagrama Tallo-Hoja (*steam and leaf*)

En una tabla se disponen dos columnas. Se seleccionan 1 o más dígitos significativos que se anotarán en la primera columna (tallo). Los dígitos menos significativos se escriben en la segunda columna (hoja).

15	455677888888
16	000000122333345556677799
17	001233344456788
18	05

Análisis Exploratorio de Datos

Tabligráma o Diagrama Tallo-Hoja (*steam and leaf*)

En una tabla se disponen dos columnas. Se seleccionan 1 o más dígitos significativos que se anotarán en la primera columna (tallo). Los dígitos menos significativos se escriben en la segunda columna (hoja).

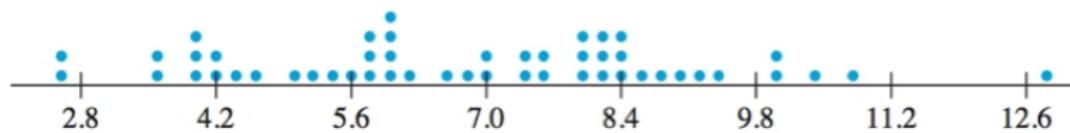
15	455677888888
16	000000122333345556677799
17	001233344456788
18	05

- ▶ Útil para identificar valores típicos, dispersión, presencia valores fuera gráfica.
- ▶ Libertad en estructura a través de simbología.

Análisis Exploratorio de Datos

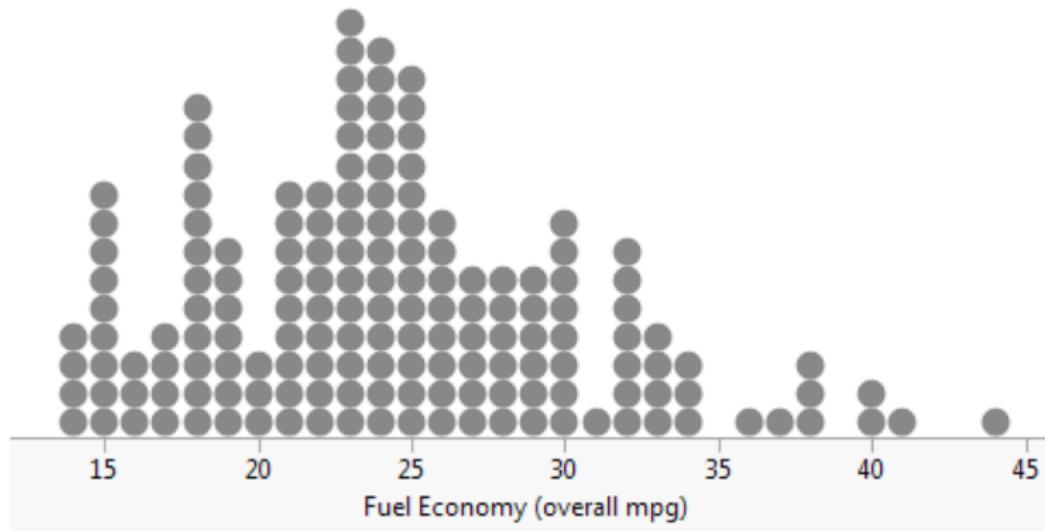
Diagramas de Puntos (*Dot Plots*)

- ▶ Una de las mejores representaciones para variables numéricas (*Según Cleveland*).
- ▶ Útil para dataset pequeños o pocos valores distintos.
- ▶ Cada valor posible se representa usando un punto. Si el valor ocurre más de una vez se superponen los puntos verticalmente.



Análisis Exploratorio de Datos

Diagramas de Puntos (*Dot Plots*)



Análisis Exploratorio de Datos

- ▶ ¿Qué sucede con variables numéricas con muchísimos valores diferentes?

Análisis Exploratorio de Datos

- ▶ ¿Qué sucede con variables numéricas con muchísimos valores diferentes?

Histogramas

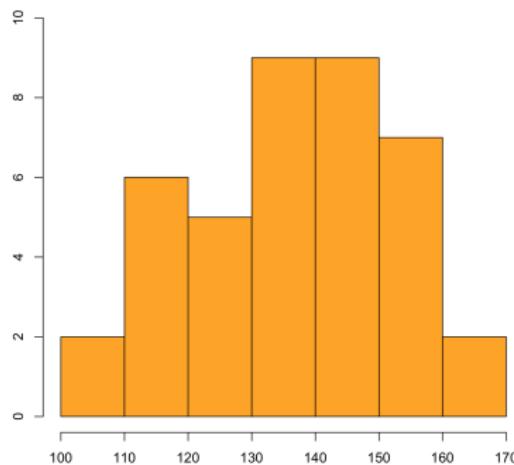
Representación simplificada/resumida de la distribución de frecuencias. Similar al *bar chart* pero:

- ▶ Los valores posibles de x se agrupan en K clases o subconjuntos de valores.
- ▶ Se registra la frecuencia de cada clase: Número de observaciones que caen dentro del conjunto.

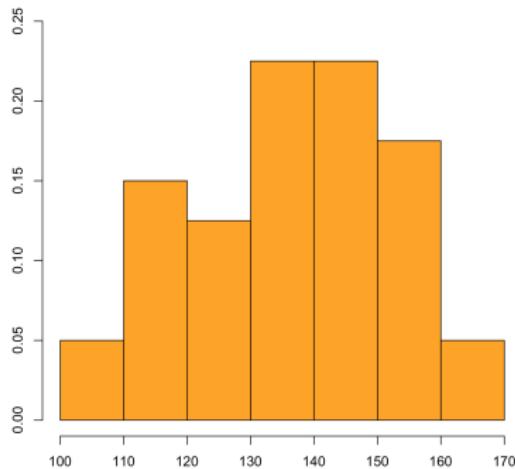
Análisis Exploratorio de Datos

Histogramas

► Frecuencias absolutas



► Frecuencias relativas



Análisis Exploratorio de Datos

Construcción de Histogramas

Dado un valor de K , sea

- ▶ $x_{\max} = \max(x_1, x_2, \dots, x_n)$
- ▶ $x_{\min} = \min(x_1, x_2, \dots, x_n)$.

Calculamos

- ▶ Rango: $R = x_{\max} - x_{\min}$.
- ▶ Amplitud: $A = R/K$.
- ▶ La primera clase de valores C_1 se define como el intervalo $[x_{\min}, x_{\min} + A]$.
- ▶ La clase i , con $i > 1$, se define como el intervalo $(x_{\min} + (i - 1) \cdot A, x_{\min} + i \cdot A]$.

Análisis Exploratorio de Datos

Reglas para determinar K :

- ▶ Sturges: $K = \lceil \log_2(n) \rceil + 1$.
- ▶ Raíz cuadrada: $K = \lceil \sqrt{n} \rceil$.
- ▶ Freedman-Diaconis: $A = 2 \cdot \text{IQR} \cdot n^{-1/3}$.

¿Qué significa un valor $K = 1$ ó $K = n$? ¿Son útiles?

Análisis Exploratorio de Datos

Construcción de Histogramas (Ejemplo)

Construya un histograma para los siguientes datos (use $K = 7$):

10	7	8							
11	1	2	3	7	9				
12	0	3	3	4	6	8			
13	1	2	2	4	5	6	7	8	
14	0	1	2	3	3	5	7	8	8
15	0	2	3	3	8	8			
16	0	0	1	3					

Análisis Exploratorio de Datos

Construcción de Histogramas (Ejemplo)

Construya un histograma para los siguientes datos (use $K = 7$):

10	7	8					
11	1	2	3	7	9		
12	0	3	3	4	6	8	
13	1	2	2	4	5	6	7
14	0	1	2	3	3	5	7
15	0	2	3	3	8	8	
16	0	0	1	3			

- $R = 56$, $A = 8$, Clases [107, 115], (115, 123], (123, 131], (131, 139], (139, 147], (147, 155], (155, 163].

Análisis Exploratorio de Datos

Construcción de Histogramas (Ejemplo)

clase	marca de clase	n_i	f_i	F_i
[107, 115]				
(115, 123]				
(123, 131]				
(131, 139]				
(139, 147]				
(147, 155]				
(155, 163]				

- Marca de clase: valor representativo de la clase. Usualmente el punto medio del intervalo.

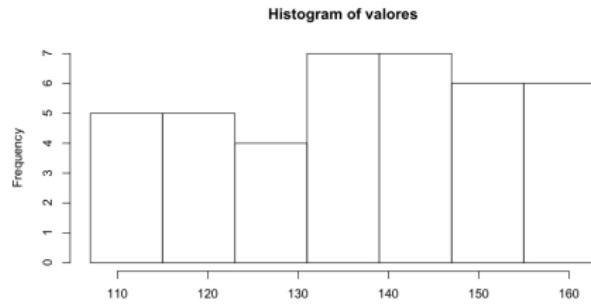
Análisis Exploratorio de Datos

Construcción de Histogramas (Ejemplo)

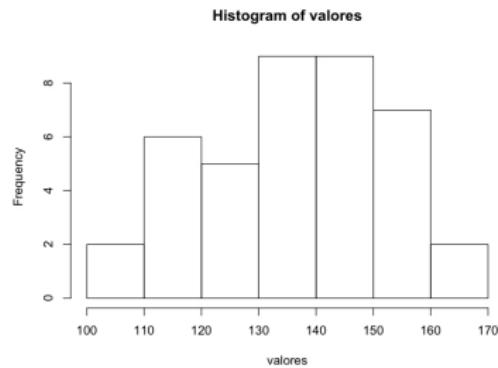
clase	marca de clase	n_i	f_i	F_i
[107, 115]	111	5	0.125	0.125
(115, 123]	119	5	0.125	0.25
(123, 131]	127	4	0.1	0.35
(131, 139]	135	7	0.175	0.525
(139, 147]	143	7	0.175	0.7
(147, 155]	151	6	0.15	0.85
(155, 163]	159	6	0.15	1.0

- Marca de clase: valor representativo de la clase. Usualmente el punto medio del intervalo.

Análisis Exploratorio de Datos



(a) A mano



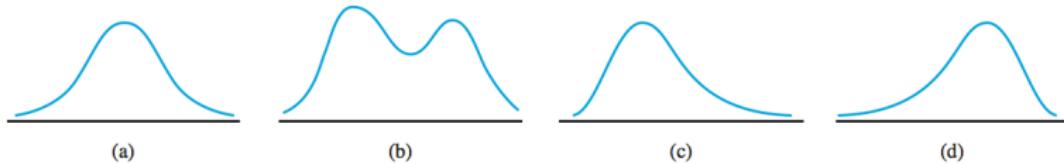
(b) software R

Variantes: intervalos de la forma $[a, b)$ y clases no equi-espaciadas.

Análisis Exploratorio de Datos

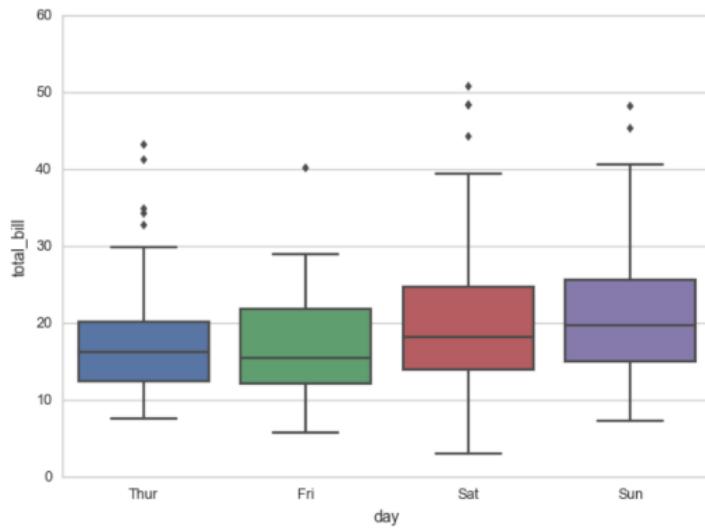
Características de un Histograma

- ▶ Moda: Clase de valores que se observa con mayor frecuencia. *unimodal, bimodal, multimodal.*
- ▶ Moda Local: Clase de valores que se observa con mayor frecuencia en torno a sus vecinos (picos del histograma).
- ▶ Sesgo o Asimetría: “inclinación” del histograma con respecto a su moda. La definiremos matemáticamente luego.



Boxplots

- ▶ En muchos problemas es necesario **comparar** el comportamiento de muestras para determinar si existen diferencias significativas entre ellas.
- ▶ Un **Boxplot** o **Diagrama de Caja** es un modo conveniente de presentar gráficamente la tendencia y dispersión de un conjunto de datos.



Outline

MEDIDAS DE TENDENCIA Y DISPERSIÓN

Medidas de Tendencia

Definición

Valor numérico que representa el conjunto de observaciones muestrales.

- ▶ **Moda:** valor o clase de valores con mayor frecuencia (puede no existir).
- ▶ **Media Muestral:** promedio aritmético de los valores en la muestra:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i .$$

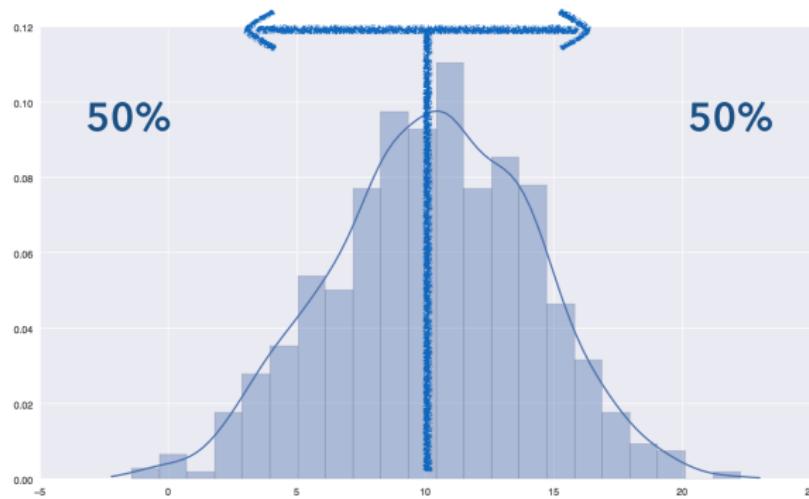
- ▶ **Mediana Muestral:** estadística de orden, que divide la muestra en dos grandes mitades, $(-\infty, \tilde{x}]$ y $[\tilde{x}, \infty]$, cada una con frecuencia (relativa) de 0.5.

$$|\{x_i : x_i \leq \tilde{x}\}| = |\{x_i : x_i \geq \tilde{x}\}|$$

Medidas de Tendencia

Mediana en un Histograma

Conceptualmente, se trata de un valor que divide el histograma empírico de frecuencias en dos partes iguales: las áreas hasta antes de la mediana y después de la mediana son iguales.



Medidas de Tendencia

Cálculo de la Mediana

Si tenemos todo el conjunto de datos, podemos calcular la mediana como sigue:

- ▶ Sean $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ las observaciones ordenadas en forma creciente.
- ▶ Si n es impar, la mediana es simplemente

$$\tilde{x} = x_{((n+1)/2)} .$$

- ▶ Si n es par, la mediana se suele elegir como sigue:

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2} .$$

Medidas de Tendencia

Ejemplo

Calcule la media y la mediana correspondiente a las siguientes observaciones, provenientes de un estudio acerca del ingreso mensual de ex estudiantes sansanos durante el primer año de egreso (en miles de pesos):

{1200, 600, 1000, 700, 800, 1300, 900}.

- ▶ Observaciones ordenadas: {600, 700, 800, 900, 1000, 1200, 1300}.
- ▶ Como $n = 7$ es impar, la mediana es simplemente

$$\tilde{x} = x_{((n+1)/2)} = x_{(4)} = 900.$$

Medidas de Tendencia

Ejemplo

Calcule la media y la mediana correspondiente a las siguientes observaciones, provenientes de un estudio acerca del ingreso mensual de ex estudiantes sansanos durante el primer año de egreso (en miles de pesos):

{1200, 600, 1000, 700, 800, 1300, 900, 1400}.

- ▶ Observaciones ordenadas: {600, 700, 800, 900, 1000, 1200, 1300, 1400}.
- ▶ Como $n = 8$ es par, la mediana es simplemente

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2} = 0.5 \cdot x_{(4)} + 0.5 \cdot x_{(5)} = 0.5(900 + 1000) = 950$$

Media versus Mediana

- ▶ ¿Qué medida es preferible: la media o la mediana?

Ejemplo

Calcule la media y la mediana correspondiente a las siguientes observaciones, provenientes de un estudio acerca del ingreso mensual de ex estudiantes sansanos durante el primer año de egreso (en miles de pesos):
 $\{1200, 600, 1000, 700, 800, 1300, 900, 100000\}$.

- ▶ 100000 es claramente una observación extrema, que no se ajusta a la tendencia del resto de las observaciones.

Media versus Mediana

Ejemplo

Calcule la media y la mediana correspondiente a las siguientes observaciones, provenientes de un estudio acerca del ingreso mensual de ex estudiantes sansanos durante el primer año de egreso (en miles de pesos):

{1200, 600, 1000, 700, 800, 1300, 900, 100000}.

- ▶ Observaciones ordenadas: {600, 700, 800, 900, 1000, 1200, 1300, 100000}.
- ▶ Como $n = 8$ es par, la mediana es

$$\tilde{x} = \frac{900 + 1000}{2} = 950.$$

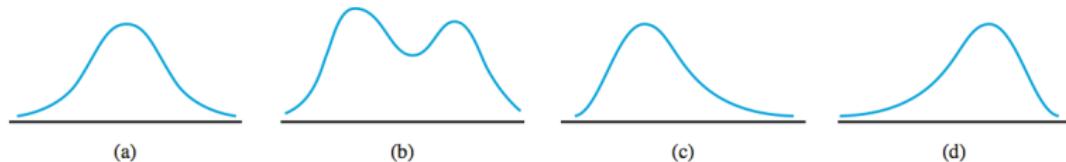
- ▶ La media en cambio es

$$\bar{x} = \frac{600 + 700 + 800 + 900 + 1000 + 1200 + 1300 + 100000}{8} = 13312.5$$

Media versus Mediana

- ▶ ¿Qué medida es preferible: la media o la mediana?
- ▶ **Media:** Es muy sensible a valores extremos (outliers). En presencia de estos valores no refleja la tendencia de la gran mayoría de las observaciones.
- ▶ **Mediana:** Es super resistente a a valores extremos (outliers). No toma en cuenta la magnitud de las observaciones, excepto aquella que divide la muestra en dos proporciones idénticas.
- ▶ Hay medidas de tendencia que intentan ser un punto medio entre estos extremos: ver **medias trucadas** o **recortada** en texto guía.

Media versus Mediana



- ▶ ¿Qué se puede decir de posición de la media y la mediana cuando el histograma idealizado tiene sesgo negativo o positivo?
- ▶ **Sesgo:** $\bar{x} - \tilde{x}$ (Asimetría, *Skew*).

Media Muestral \bar{x} versus Media Poblacional μ

- **Media Poblacional:** Si la Población es finita podríamos medir todos los valores de x en la Población: x_1, x_2, \dots, x_m y luego calcular el promedio

$$\mu = \frac{\sum_{i=1}^m x_i}{m} .$$

- **Media Muestral:** Se calcula en la muestra.
- Obviamente \bar{x} y μ , en general, difieren. Por eso el apellido “**muestral**”.
 - Lo mismo sucede con otras medidas de tendencia.

Dos muestras con media, mediana y moda muestrales iguales, ¿Son iguales?

Medidas de Dispersión

Definición

Valor numérico que mide la dispersión, variabilidad o grado concentración de los datos, usualmente en torno a un valor de **tendencia**.

► Valores Categóricos:

- **Índice de Variación:** $1 - f_M$ donde f_M es la frecuencia (relativa) de la moda.
- **Entropía:**

$$\mathbb{H}(f) = - \sum_{i=1}^n f_i \log f_i$$

Medidas de Dispersión

Definición

Valor numérico que mide la dispersión, variabilidad o grado concentración de los datos, usualmente en torno a un valor de tendencia.

► Valores Numéricos:

- **Rango:** $x_{(n)} - x_{(1)} = \max(S) - \min(S)$, donde S es el dataset.
- **Varianza Muestral:** La denotaremos s^2 o bien $\hat{\sigma}^2$ y se calcula como:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 .$$

- **Desviación Estándar Muestral:** Se denota s o bien $\hat{\sigma}$, calculado como:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} .$$

Medidas de Dispersión

Varianza Poblacional versus Varianza Muestral: $n - 1$ versus n

Si la Población es finita podríamos medir todos los valores de x en la Población: x_1, x_2, \dots, x_m . Luego, podríamos calcular el promedio de estos valores

$$\mu = \frac{\sum_{i=1}^m x_i}{m} .$$

y el promedio de las diferencias en torno a la media,

$$\sigma^2 = \frac{\sum_{i=1}^m (x_i - \mu)^2}{m} .$$

- s^2 se puede considerar una aproximación de σ^2 calculada en la muestra. Como reemplazamos μ por \bar{x} el valor calculado será en general menor que la verdadera varianza. Para compensar este efecto se divide por $n - 1$ y no por n ,

Medidas de Dispersión

Rango Inter-Cuartílico IQR

¿Cómo medir dispersión en torno a la mediana? Queremos:

- ▶ **Medida basada en frecuencia:** Si la mediana se define en términos del histograma de frecuencias empíricas, su medida de dispersión debiera responder al mismo criterio.
- ▶ **Resistencia a outliers:** Si la mediana busca ser resistente a observaciones extremas, su medida de Dispersión debiera ser también resistente a outliers.

Medidas de Dispersión

Rango Inter-Cuartílico IQR

¿Cómo medir dispersión en torno a la mediana? Queremos:

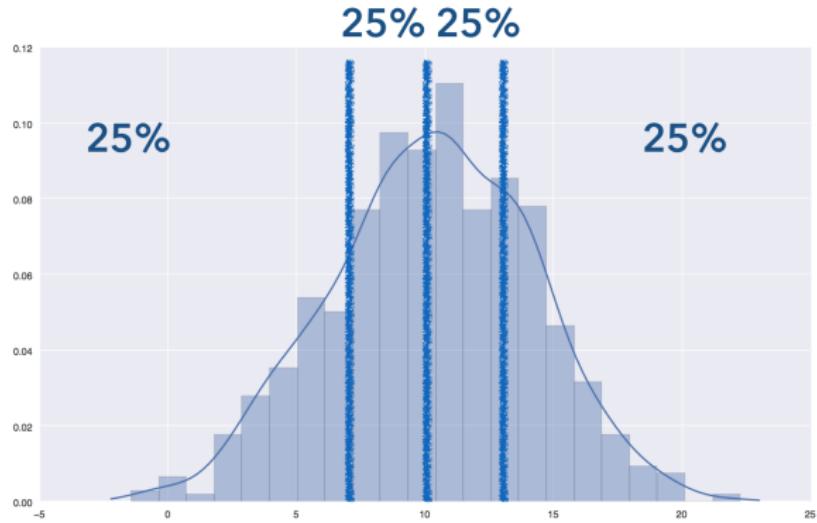
- ▶ **Medida basada en frecuencia:** Si la mediana se define en términos del histograma de frecuencias empíricas, su medida de dispersión debiera responder al mismo criterio.
- ▶ **Resistencia a outliers:** Si la mediana busca ser resistente a observaciones extremas, su medida de Dispersión debiera ser también resistente a outliers.

Existen varias otras medidas para calcular la dispersión de los datos: coeficiente de variación, desviación absoluta promedio, *median absolute deviation*.

Cuartiles

Definición

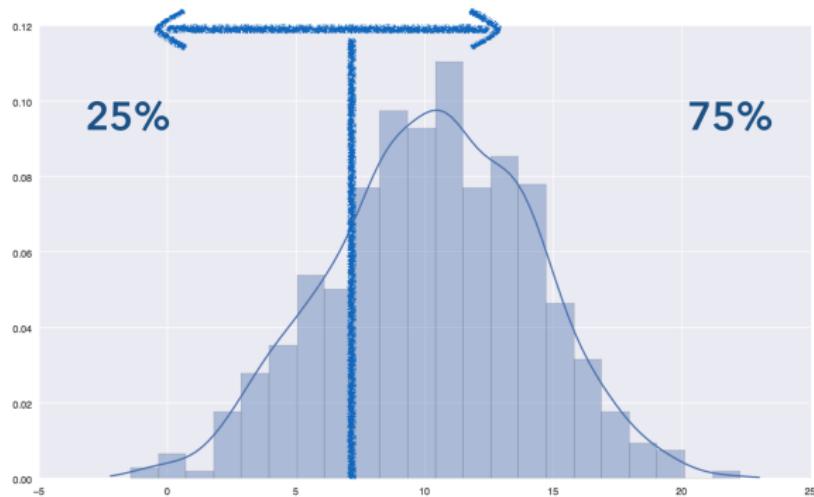
Llamaremos Cuartiles 1, 2 y 3, a los valores que dividen el histograma de frecuencias en 4 partes de igual proporción. Usaremos los símbolos Q_1 , Q_2 , Q_3 para denotar estos valores. Notar que Q_2 coincide con \tilde{x} .



Cuartiles

Definición

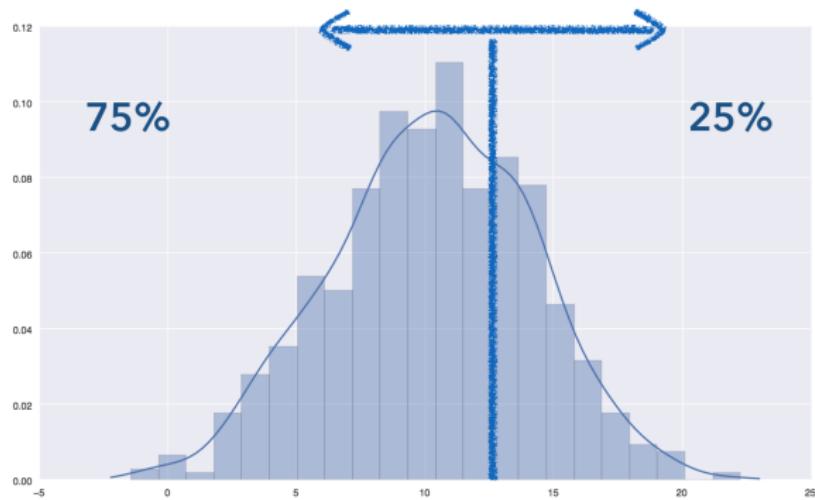
Llamaremos Cuartiles 1, 2 y 3, a los valores que dividen el histograma de frecuencias en 4 partes de igual proporción. Usaremos los símbolos Q_1 , Q_2 , Q_3 para denotar estos valores. Notar que Q_2 coincide con \tilde{x} !



Cuartiles

Definición

Llamaremos Cuartiles 1, 2 y 3, a los valores que dividen el histograma de frecuencias en 4 partes de igual proporción. Usaremos los símbolos Q_1 , Q_2 , Q_3 para denotar estos valores. Notar que Q_2 coincide con \tilde{x} !



Cuartiles

Cálculo de los Cuartiles (Libro Guía y Software R)

Supongamos que disponemos de todas las observaciones $S = \{x_1, x_2, \dots, x_n\}$.

- ▶ Calculamos \tilde{x} , la mediana del conjunto.
- ▶ Sea $S_1 = \{x_{(1)}, x_{(2)}, \dots, x_{(s)}\}$, el conjunto de observaciones menores o iguales que la mediana, es decir $S_1 = \{x_i \in S : x_i \leq \tilde{x}\}$. (si \tilde{x} no está en el conjunto de observaciones no se toma en cuenta. Esto ocurre si n es par).
- ▶ Sea $S_2 = \{x_{(s)}, x_{(s+1)}, \dots, x_{(n)}\}$, el conjunto de observaciones mayores o iguales que la mediana, es decir $S_2 = \{x_i \in S : x_i \geq \tilde{x}\}$. (si \tilde{x} no está en el conjunto de observaciones no se toma en cuenta. Esto ocurre si n es par).
- ▶ Q_1 se define como la mediana de S_1 .
- ▶ Q_3 se define como la mediana de S_2 .
- ▶ Q_2 es la mediana de todo el conjunto, i.e. \tilde{x} .

Cuartiles

Ejemplo 1

Consideremos las observaciones 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37. Determine todos los cuartiles de la muestra. ($n = 12$)

Cuartiles

Ejemplo 1

Consideremos las observaciones 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37. Determine todos los cuartiles de la muestra. ($n = 12$)

- ▶ Mediana $\tilde{x} = 15$
- ▶ $S_1 = \{2, 3, 5, 7, 11, 13\}$ (15 NO está en el conjunto de observaciones).
- ▶ $S_2 = \{17, 19, 23, 29, 31, 37\}$ (15 NO está en el conjunto de observaciones).
- ▶ Q_1 es la mediana de S_1 ($n = 6$), i.e. $Q_1 = (5 + 7)/2 = 6$.
- ▶ Q_3 es la mediana de S_2 ($n = 6$), i.e. $Q_3 = (23 + 29)/2 = 26$.
- ▶ Q_2 es la mediana de todo el conjunto, i.e. $Q_2 = 15$.

Cuartiles

Ejemplo 2

Consideremos las observaciones 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31. Determine todos los cuartiles de la muestra. ($n = 11$)

Cuartiles

Ejemplo 2

Consideremos las observaciones 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31. Determine todos los cuartiles de la muestra. ($n = 11$)

- ▶ Mediana $\tilde{x} = 13$
- ▶ $S_1 = \{2, 3, 5, 7, 11, 13\}$ (13 sí está en el conjunto de observaciones).
- ▶ $S_2 = \{13, 17, 19, 23, 29, 31\}$ (13 sí está en el conjunto de observaciones).
- ▶ Q_1 es la mediana de S_1 ($n = 6$), i.e. $Q_1 = (5 + 7)/2 = 6$.
- ▶ Q_3 es la mediana de S_2 ($n = 6$), i.e. $Q_3 = (19 + 23)/2 = 21$.
- ▶ Q_2 es la mediana de todo el conjunto, i.e. $Q_2 = 13$.

Cuartiles

Ejemplo 3

Consideremos las observaciones 2, 3, 5, 7, 11, 13, 17, 19, 23, 29. Determine todos los cuartiles de la muestra. ($n = 10$)

- ▶ Mediana $\tilde{x} = 12$
- ▶ $S_1 = \{2, 3, 5, 7, 11\}$ (12 NO está en el conjunto de observaciones).
- ▶ $S_2 = \{13, 17, 19, 23, 29\}$ (12 NO está en el conjunto de observaciones).
- ▶ Q_1 es la mediana de S_1 ($n = 5$), i.e. $Q_1 = 5$.
- ▶ Q_3 es la mediana de S_2 ($n = 5$), i.e. $Q_3 = 19$.
- ▶ Q_2 es la mediana de todo el conjunto, i.e. $Q_2 = 12$.

Medidas de Dispersión

Rango Inter-Cuartílico IQR

Ahora sí, podemos definir una medida de dispersión para la mediana. Llamaremos Rango Inter-Cuartílico (IQR) a la diferencia entre Q_3 y Q_1

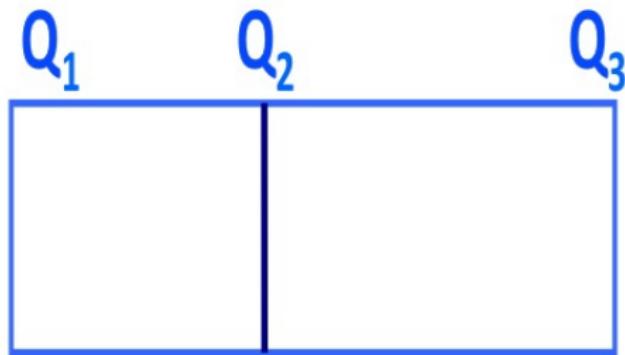
$$\text{IQR} = Q_3 - Q_1 .$$

► **Observaciones Importantes:** Por definición, tenemos

1. Un 50 % de las observaciones están entre Q_3 y Q_1 .
2. Un 25 % de las observaciones están entre Q_2 y Q_1 .
3. Un 25 % de las observaciones están entre Q_2 y Q_3 .
4. Un 50 % de las observaciones están fuera del intervalo $[Q_1, Q_3]$.

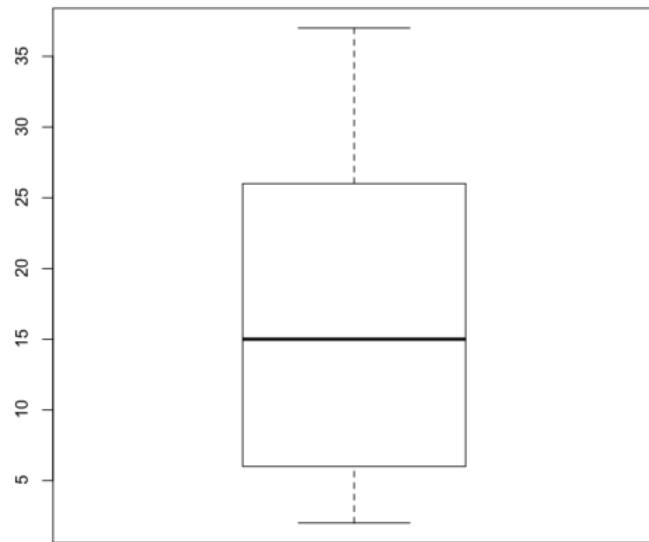
Boxplots

- Un **Boxplot** o **Diagrama de Caja** es un modo conveniente de mostrar la tendencia/Dispersión de los datos, usando las estadísticas de orden Q_1, Q_2, Q_3 :



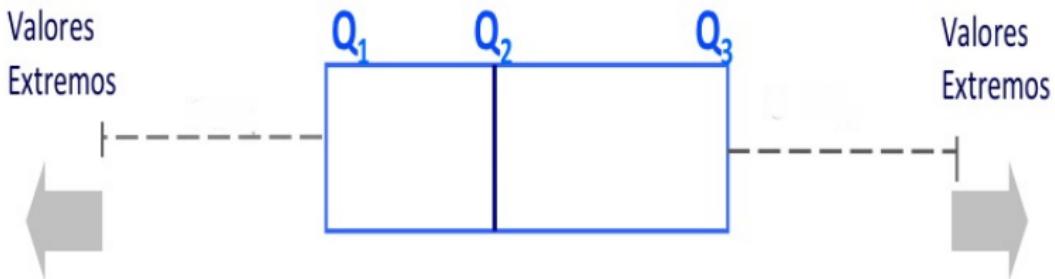
Boxplots

- Ejemplo, para el conjuntos de datos: 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37,



Boxplots

- ▶ Los **Bigotes** del Boxplot representan el rango de las observaciones que no son atípicas o extremas (outliers). ¿Cómo detectar observaciones atípicas?
- ▶ **Criterio:** Usar una tolerancia de $1.5 \cdot \text{IQR}$ en torno a la caja principal.



Boxplots

Calculando los Bigotes

El rango de tolerancia es entonces el intervalo $[L_{\inf}, L_{\sup}]$ donde

$$L_{\inf} = Q_1 - 1.5 \cdot \text{IQR}$$

$$L_{\sup} = Q_3 + 1.5 \cdot \text{IQR} .$$

- ▶ **Bigote Mayor (B^+):** Es la observación más grande que está dentro del rango de tolerancia, i.e. $B^+ = \arg \max_i \{x_i : x_i \leq L_{\sup}\}$
- ▶ **Bigote Menor (B^-):** Es la observación más pequeña que está dentro del rango de tolerancia, i.e. $B^- = \arg \min_i \{x_i : x_i \geq L_{\inf}\}$

Boxplots

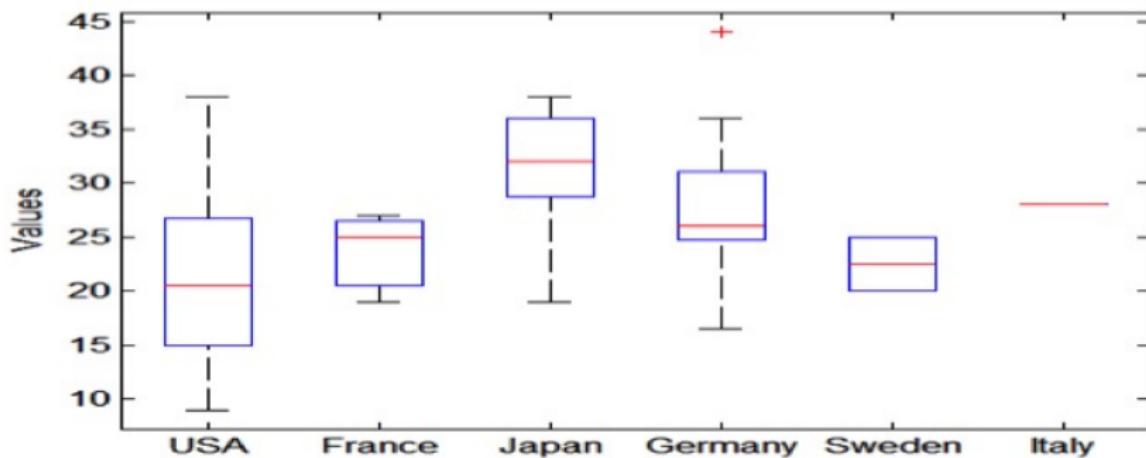
Ejercicio

En un estudio sobre la resistencias de ciertas botellas de vidrio, se recogió una muestra de tamaño 20 obteniendo las siguientes estadísticas: $Q_1 = 196.0$, $Q_2 = 202.2$, $Q_3 = 216.8$. Si las tres observaciones menores eran 125.8; 188.1 y 193.7, mientras que las tres observaciones mayores eran 221.3; 230.5 y 250.2.

- ▶ Determine si hay outliers.
- ▶ Construya un boxplot para los datos.

Boxplots

Comparación de muestras



- En general, la diferencia entre las muestras es **más significativa** mientras **menos se sobreponen** las respectivas cajas.