¿QUÉ ES LA ESTADÍSTICA?

Es el arte de realizar inferencias y sacar conclusiones a partir de datos imperfectos.

ÁREAS DE LA ESTADÍSTICA

- **Diseño**: Planeamiento y desarrollo de investigaciones
- **Descripción**: Resumen y exploración de datos.
- **Inferencia**: Hacer predicciones o generalizaciones acerca de características de una población basadas en la información de una muestra de la población.

ALGUNAS DEFINICIONES

POBLACIÓN $\Rightarrow$ total de sujetos o unidades de análisis de interés en el estudio

MUESTRA $\Rightarrow$ cualquier subconjunto de la población en la cual se recolectarán los datos

PARÁMETRO $\Rightarrow$ es una medida resumen calculada sobre la población

ESTADÍSTICO $\Rightarrow$ es una medida resumen calculada sobre la muestra

EJEMPLO de BASE DE DATOS

Caso	Sexo	Provincia	Edad	PAS	
1	F	1	35	110	
2	M	2	28	120	⇐ **REGISTRO**
3	M	3	59	136	
	⇑			⇑	
	VARIABLE			**OBSERVACIÓN**	

TIPOS DE DATOS

- **CATEGÓRICOS O CUALITATIVOS**
  Se registra la presencia de un atributo.
- **NUMÉRICOS O CUANTITATIVOS**
  Resultan de contar o de registrar una magnitud
- **OTROS** (rangos, censurados, scores, etc.)

DATOS NUMÉRICOS

Clasificación

- DISCRETOS
  Número finito de valores posibles
- CONTINUOS
  Pueden tomar infinitos valores en un rango.

¿Cómo clasificar la variable edad?

¿POR QUÉ ES IMPORTANTE IDENTIFICAR EL TIPO DE DATOS?

Porque

- el tipo de datos DETERMINA el método de análisis apropiado y
- cada método de análisis estadístico es específico para un cierto tipo de datos.

ESTADÍSTICA DESCRIPTIVA

La *estadística descriptiva* o *análisis exploratorio de datos* ofrece modos de presentar y evaluar las características principales de los datos a través de tablas, gráficos y medidas resúmenes.

GRÁFICOS
DATOS NUMÉRICOS

Objetivo $\Rightarrow$ Poner de manifiesto características sobresalientes de los datos

- Gráfico de tallo-hojas
- Histograma
- Gráfico de caja (box-plot)

GRÁFICO DE TALLO - HOJAS

Tasas de neumonía cada 1000 habitantes. Año 2000.

Tasa	Tasa
0.00	3.86
1.28	3.98
1.60	4.01
1.67	4.22
2.19	4.38
2.87	4.84
3.01	4.92
3.16	5.50
3.20	7.36
3.21	8.07
3.33	9.29
3.37	10.83

0	**0**
1	**2 6 6**
2	**1 8**
3	**0 1 2 2 3 3 8 9**
4	**0 2 3 8 9**
5	**5**
6	
7	**3**
8	**0**
9	**2**
10	**8**

GRÁFICO DE TALLO - HOJAS

¿Qué información nos brinda este gráfico?

- El *rango* de las observaciones.
- La *forma* de la distribución. Simetría.
- Cuántos picos o modas tiene la distribución.
- Si hay valores que se apartan del conjunto.

Características

- No hay pérdida de información
- Apropiado para muestras pequeñas

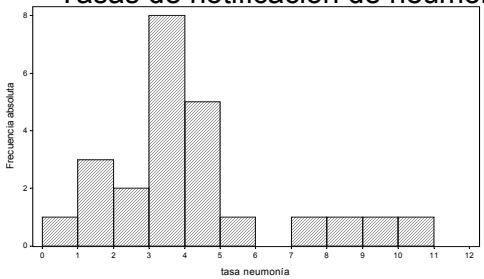
TABLA DE FRECUENCIAS

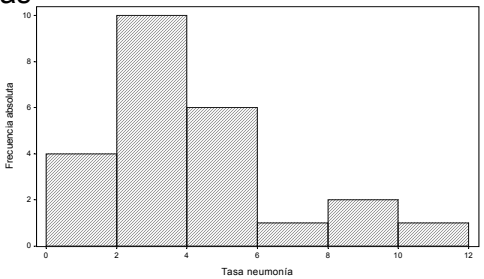
Tasas de notificación de neumonías.

Intervalo	Frecuencia (f_i)	Frecuencia relativa porcentual (fr_i)	Frecuencia acumulada (fa_i)	Frecuencia relativa acumulada (fra_i)
[0, 1)	1	4.2	1	4.2
[1, 2)	3	12.5	4	16.7
[2, 3)	2	8.3	6	25.0
[3, 4)	8	33.3	14	58.3
[4, 5)	5	20.8	19	79.2
[5, 6)	1	4.2	20	83.3
[6, 7)	0	0.0	20	83.3
[7, 8)	1	4.2	21	87.5
[8, 9)	1	4.2	22	91.7
[9, 10)	1	4.2	23	95.8
[10, 11)	1	4.2	24	100.0

HISTOGRAMA

Tasas de notificación de neumonías



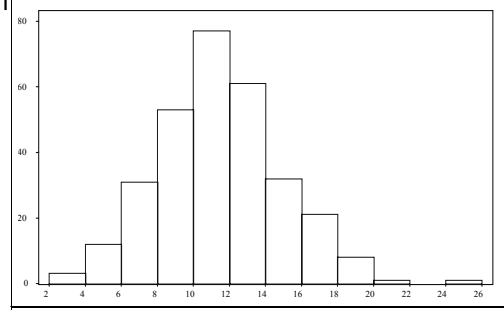


Intervalo	Frecuencia (f_i)	Frecuencia relativa porcentual (fr_i)	Frecuencia acumulada (fa_i)	Frecuencia relativa acumulada (fra_i)
[0, 1)	1	4.2	1	4.2
[1, 2)	3	12.5	4	16.7
[2, 3)	2	8.3	6	25.0
[3, 4)	8	33.3	14	58.3
[4, 5)	5	20.8	19	79.2
[5, 6)	1	4.2	20	83.3
[6, 7)	0	0.0	20	83.3
[7, 8)	1	4.2	21	87.5
[8, 9)	1	4.2	22	91.7
[9, 10)	1	4.2	23	95.8
[10, 11)	1	4.2	24	100.0

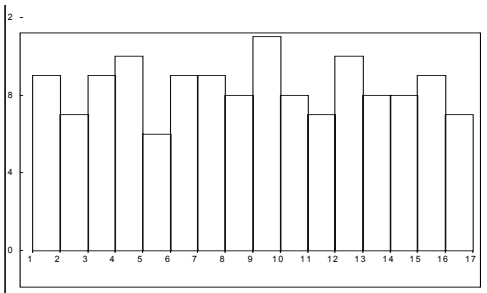
HISTOGRAMA

- ¿Cuántas clases?
- Proporciones representadas en áreas
- ¿Los intervalos deben ser de la misma longitud?
- Area total bajo el histograma 100%
- ¿Qué se ve?
  - Forma de la distribución.
  - Si hay agrupamientos.
  - Si hay datos atípicos.

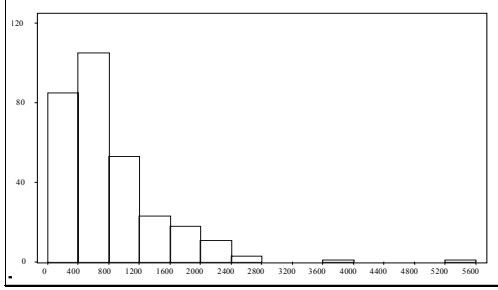
DISTRIBUCIÓN ACAMPANADA



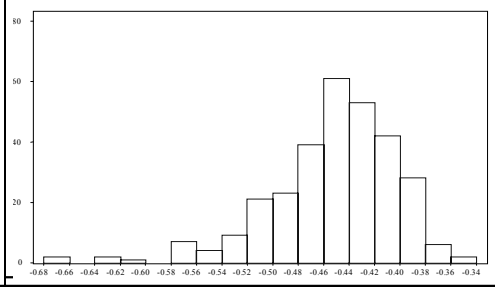
DISTRIBUCIÓN UNIFORME



ASIMETRIA DERECHA



ASIMETRÍA IZQUIERDA



HISTOGRAMAS

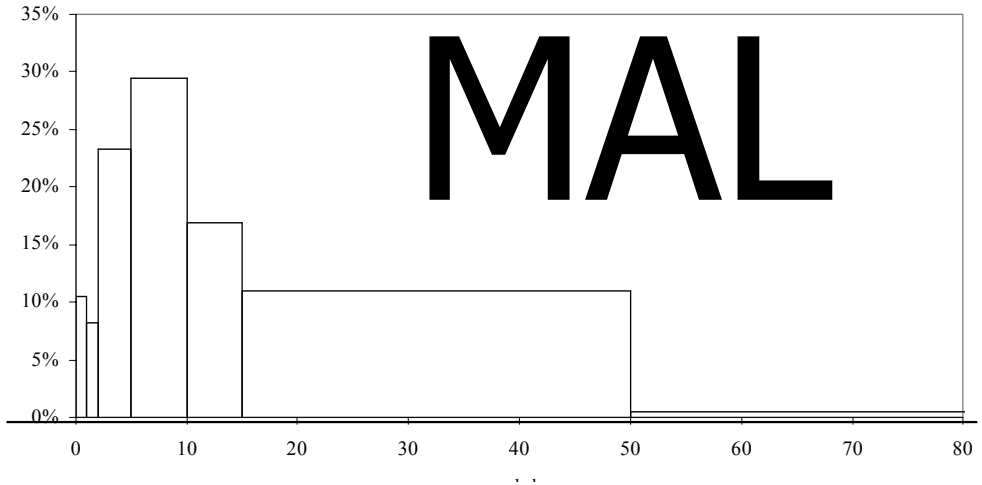
Intervalos de clase de diferente longitud

Notificaciones de casos de rubéola. Argentina, año 2000.

Categoría (años)	Frecuencia (f_i)	Frecuencia relativa (f_r)
[0, 1)	497	10.5%
[1, 2)	387	8.2%
[2, 5)	1100	23.3%
[5, 10)	1389	29.4%
[10, 15)	798	16.9%
[15, 50)	521	11.0%
≥ 50	28	0.6%
Total	4720	100.00%

HISTOGRAMAS
Intervalos de clase de diferente longitud

Notificaciones de casos de rubéola. Argentina, año 2000.



HISTOGRAMAS
Intervalos de clase de diferente longitud

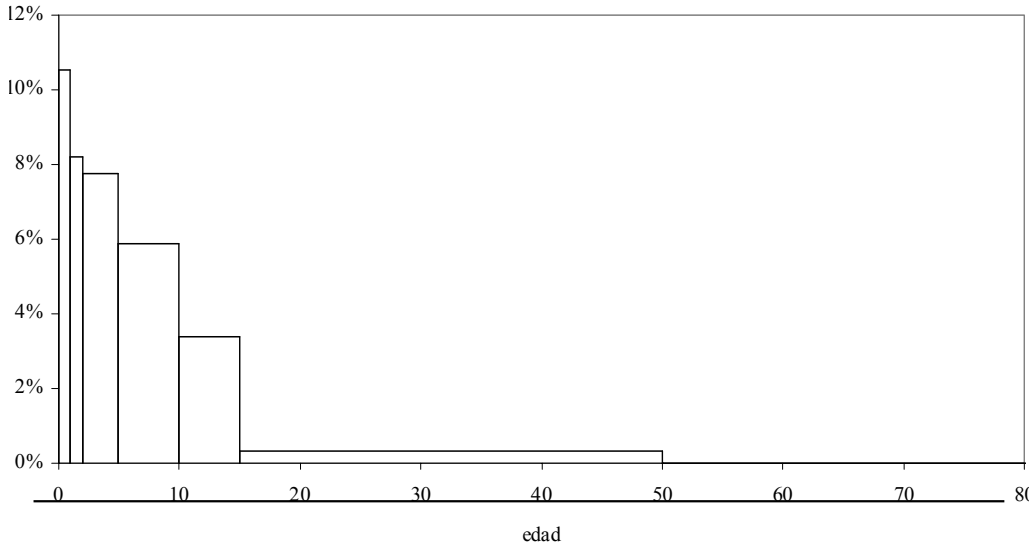
Notificaciones de casos de rubéola. Argentina, año 2000.

$$\text{Escala densidad} \Rightarrow \quad \text{altura de la barra} = \frac{\text{frecuencia en el intervalo}}{\text{longitud del intervalo}}$$

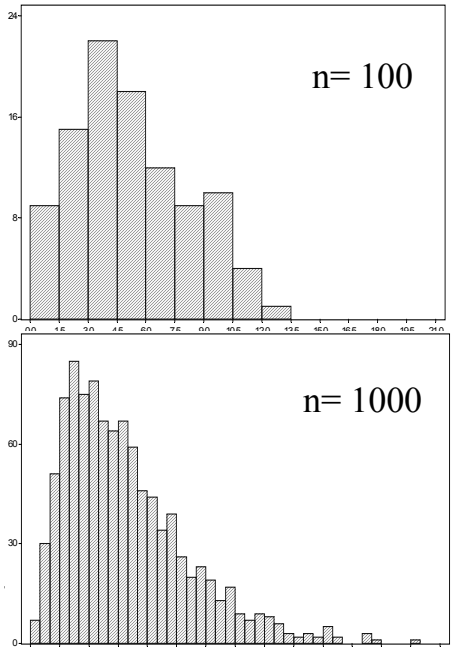
Categoría (años)	Frecuencia (f_i)	Frecuencia relativa (f_r)	Escala densidad
[0, 1)	497	10.5%	10.53%
[1, 2)	387	8.2%	8.20%
[2, 5)	1100	23.3%	7.77%
[5, 10)	1389	29.4%	5.89%
[10, 15)	798	16.9%	3.38%
[15, 50)	521	11.0%	0.32%
≥ 50	28	0.6%	0.01%
Total	4720	100.00%	--

HISTOGRAMA USANDO ESCALA DENSIDA

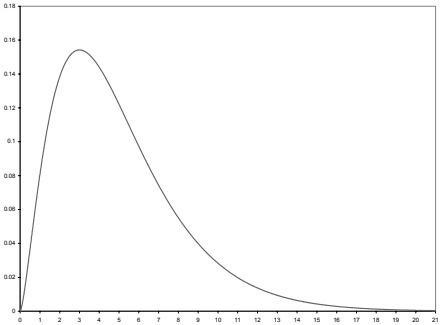
Notificaciones casos de rubéola. Año 2000.



DISTRIBUCIÓN MUESTRAL



DISTRIBUCIÓN POBLACIONAL

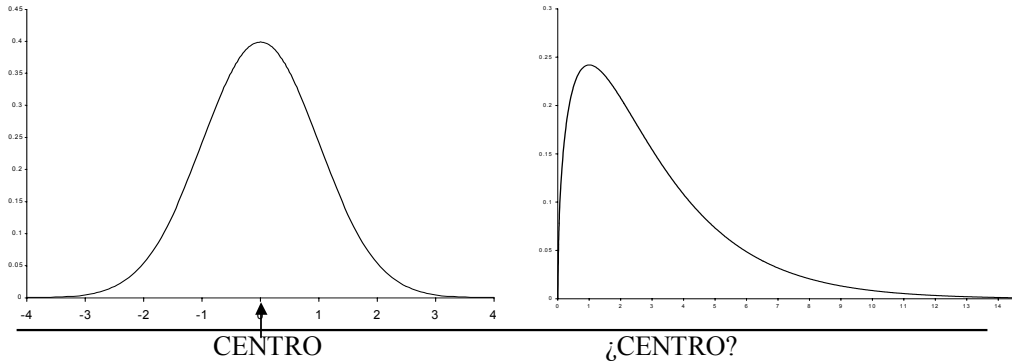


MEDIDAS RESÚMENES
Datos numéricos

- Resumir $\Rightarrow$ pasar de una visión detallada a una generalización simple e informativa tratando de preservar las características esenciales.
- *Medidas de posición o localización* $\Rightarrow$ describen un valor alrededor del cual se encuentran las observaciones.
- *Medidas de dispersión o escala* $\Rightarrow$ pretenden expresar cuan variable es un conjunto de datos.

MEDIDAS DE POSICIÓN O CENTRALIDAD

Pretenden dar una idea de dónde se encuentra el *centro* de una distribución.



EL PROMEDIO O MEDIA ARITMÉTICA

Media muestral. Observaciones $X_1, X_2, ..., X_n$

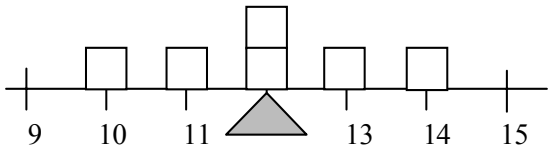
$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Media poblacional.

$$\mu = E(X)$$

PROPIEDADES DE LA MEDIA

- Se usa para datos numéricos.
- Representa el *centro de gravedad* de los datos.



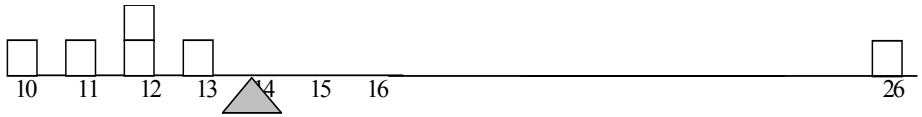
- La suma de las distancias a la media es cero.
- Es muy sensible a la presencia de datos atípicos (OUTLIERS).

PROPIEDADES DE LA MEDIA

- Es muy sensible a la presencia de datos atípicos (OUTLIERS).

  Modificamos 1 dato en el ejemplo anterior

  $X_2 = 14 \rightarrow X_2 = 26$



- Es una buena medida del centro de la distribución cuando ésta es simétrica.

MEDIANA MUESTRAL

- La *mediana* es el dato que ocupa la posición central en la muestra ordenada.
- Para calcularla
  - Ordenamos los datos de menor a mayor
  - La mediana es el dato que ocupa la posición (n+1)/2 en la muestra ordenada

Ejemplos:

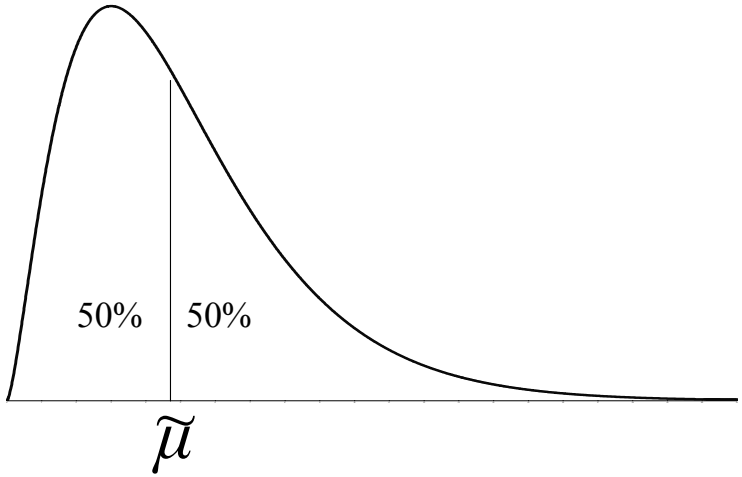
n impar

$X_1 = 10$ $X_2 = 14$ $X_3 = 12$ $X_4 = 18$ $X_5 = 11$

n par

$X_1 = 10$ $X_2 = 14$ $X_3 = 12$ $X_4 = 18$ $X_5 = 11$ $X_6 = 23$

MEDIANA POBLACIONAL

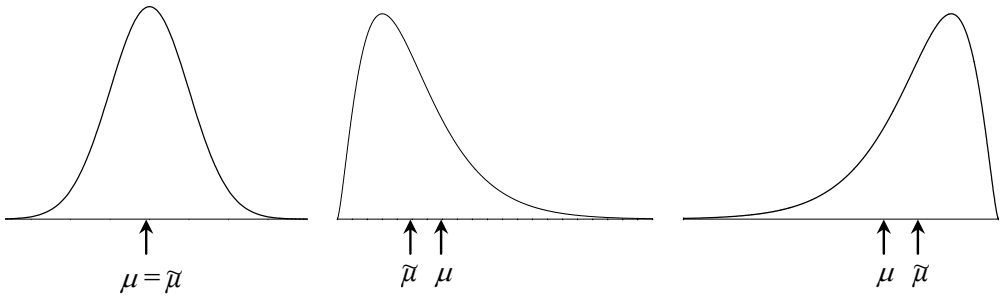


PROPIEDADES DE LA MEDIANA

- Apropiada para *datos numéricos y ordinales*
- Distribución simétrica $\Rightarrow \quad \overline{X} = \tilde{X}$
- Asimetría derecha $\Rightarrow \quad \overline{X} > \tilde{X}$
- Asimetría izquierda $\Rightarrow \quad \overline{X} < \tilde{X}$

1) 12, 13, 14, 15, 16
2) 12, 13, 14, 15, 20
3) 2, 13, 14, 15, 16

PROPIEDADES DE LA MEDIANA



PROPIEDADES DE LA MEDIANA

- Medida de posición *robusta*

  10 11 12 12 13 14 media = 12 mediana = 12

  10 11 12 12 13 26 media = 14 mediana = 12

- Insensible a la distancia de los datos al centro

10	11	12	13	14
10	11	12	13	100
0	11	12	12	12
10	11	12	100	100

COMPARACIÓN DE MEDIA Y MEDIANA

	VENTAJAS	DESVENTAJAS
MEDIA	Usa toda la información que proveen los datos. Es de manejo algebraico simple.	Muy sensible a la presencia datos outliers.
MEDIANA	Representa el centro de la distribución (en un sentido claramente definido). Robusta a la presencia de outliers. Útil para datos ordinales.	Usa muy poca información los datos.

MEDIA α-podada

- Compromiso entre la media y la mediana
- Forma de cálculo:
  - Ordenamos los n datos de menor a mayor
  - Excluimos los $n.\alpha$ datos más pequeños y los $n.\alpha$ datos más grandes.
  - Calculamos el promedio de los datos restantes

Ejemplo

$X_1 = 85$ $X_2 = 98$ $X_3 = 99$ $X_4 = 95$ $X_5 = 98$

MEDIA α-podada

- *¿Cómo elegimos α?*
- *¿Cuándo usar esta medida?*
- *¿Qué hacer cuando el número de datos que debe excluirse no es entero?*
- *¿Cuál de las tres medidas de posición preferir?*

  Datos normales usar la MEDIA, es más eficiente.

PERCENTILES

- Otro modo de resumir $\Rightarrow$ PERCENTILES
- El *percentil p%* de un conjunto de datos es el dato que deja una fracción p% de las observaciones debajo de él y (1 – p)% encima de él.

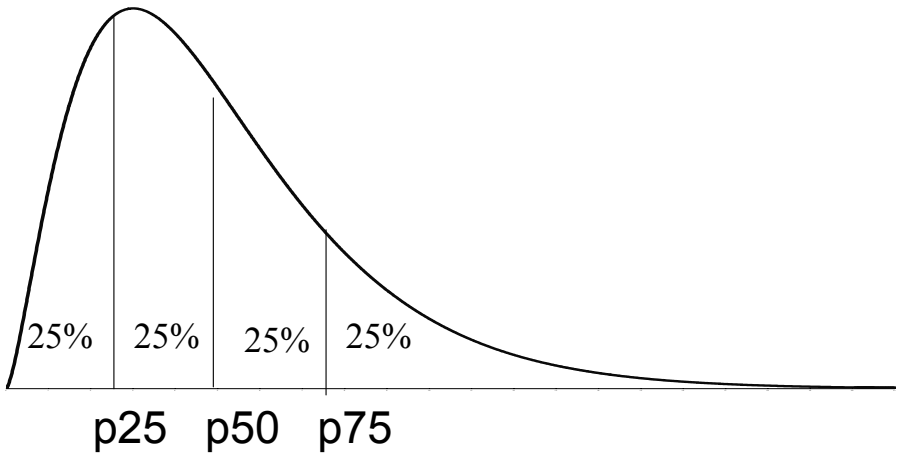
Ejemplo: Niñas recién nacidas (a término)

percentil 10% peso = 2450 g

percentil 90% peso = 3370 g

CUARTILES

- Particionan la distribución en 4 porciones



¿Cómo se calculan los cuartiles de una muestra de n observaciones?

- Ordenar los datos de menor a mayor.
- El cuartil inferior es el dato que ocupa la posición $(n+1)/4$ en la muestra ordenada.
- El cuartil superior es el dato que ocupa la posición $3(n+1)/4$ en la muestra ordenada.
- Si la posición resulta ser un número decimal, promediamos los datos que se encuentran a izquierda y derecha de la posición obtenida.

CUARTILES

osición	1	2	3	4	5	6	7	8	9	10	11	12	1
Datos	104	112	134	146	155	168	170	195	246	302	338	412	6

- Cuartil inferior C_I
  - posición (n+1)/4 = 3.5
  - C_I = (134+146)/2 = 140
- Cuartil inferior C_S
  - posición 3(n+1)/4 = 10.5
  - C_S = (302+338)/2 = 320

MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Muestra A:	55	55	55	55	55	55	55
Muestra B:	47	51	53	55	57	59	63
Muestra C:	39	47	53	55	57	63	71

- Las *medidas de dispersión o variabilidad* describe cuan cercanos se encuentran los datos entre ellos, o cuán cerca se encuentran de alguna medida de posición.

RANGO MUESTRAL

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Muestra A:	55	55	55	55	55	55	55	Rango = 55 – 55 = 0
Muestra B:	47	51	53	55	57	59	63	Rango = 63 – 47 = 16
Muestra C:	39	47	53	55	57	63	71	Rango = 71 – 39 = 32

Características y propiedades

- Es muy simple de obtener.
- Es extremadamente sensible a outliers.
- Ignora la mayoría de los datos.
- En general aumenta cuando aumenta n.

DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL

La *varianza* de una muestra de observaciones $X_1, X_2, ..., X_n$, es

$$s^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

La *desviación estándar muestral:*

$$s = \sqrt{s^2}$$

Varianza y desviación estándar de una población con N unidades

$$\sigma^2 = \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

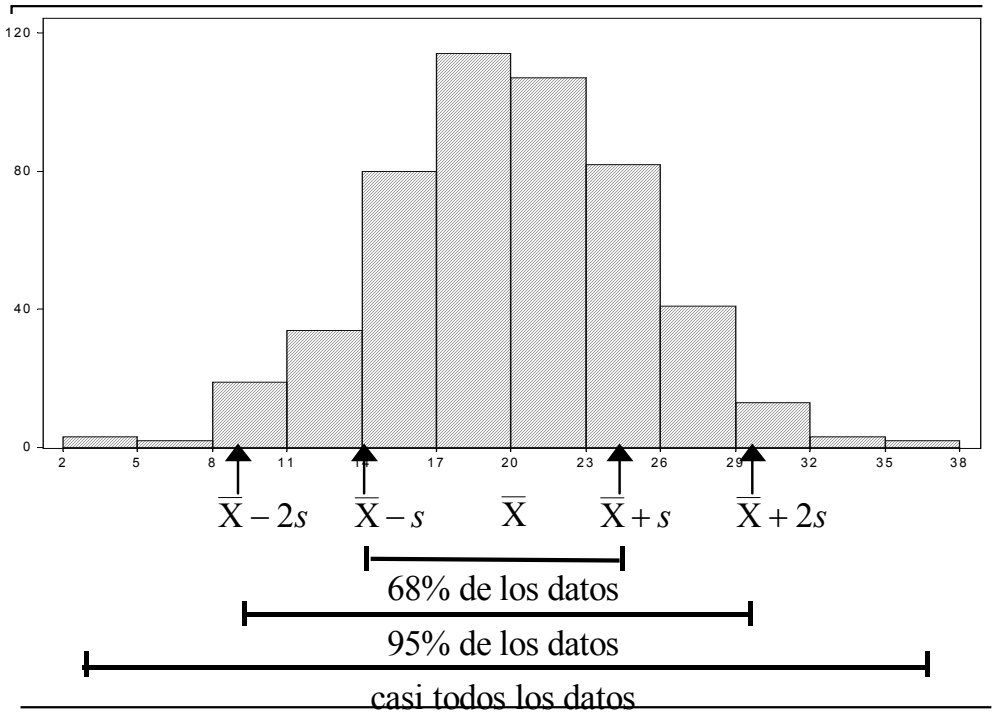
¿Por qué usamos (n - 1) en el denominador de S?

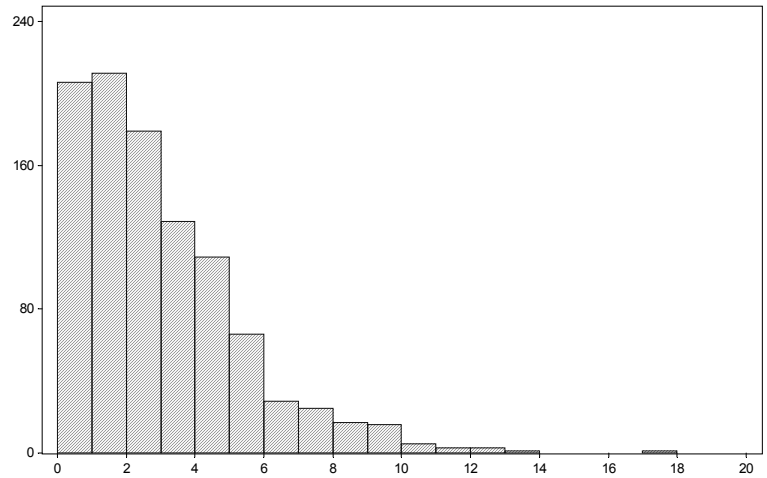
VARIANZA Y DESVÍO ESTÁNDAR

- *Interpretación de s - Regla empírica*

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces:

- Alrededor 68% de los datos caen a menos de 1 DS de la media.
- Alrededor 95% de los datos caen a menos de 2 DS de la media.
- Prácticamente todos los datos caen a menos de 3 DS.





¿Validez de la regla empírica?

Media = 3 s = 2.45

PROPIEDADES DE LA DESVIACIÓN ESTÁNDAR

- $s = 0$ solamente cuando todos los datos son iguales, de otro modo $s > 0$.
- s es una medida de dispersión *muy sensible* a la presencia de datos outliers.
- s mide dispersión alrededor de la media, usarla como medida resumen acompañando a la media.

MAD (Median absolute deviations)

- Definimos la *MAD* de una muestra $X_1, X_2, ..., X_n$ como

$$MAD = mediana\ (|\ X_i - \tilde{X}\ |)$$

- Es ROBUSTA
  - se basa en la mediana como medida de posición
  - toma mediana de las desviaciones

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana.
- Calculamos la distancia de cada dato a la mediana.
- Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
- Buscamos la mediana de las distancias sin signo.

MAD

- Ejemplo:

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

- mediana = 170
- Distancias a la mediana

  –66, –58, –36, –24, –15, –2, 0, 25, 76, 132, 168, 242, 508
- Distancias sin signo (absolutas) ordenadas

  0, 2, 15, 24, 25, 36, 58, 66, 76, 132, 168, 242, 508

MAD = 58

PROPIEDADES MAD

- Robusta
- Si la distribución es acampanada y simétrica la MAD y el desvío estándar *s* se relacionan del siguiente modo:

$$s \cong 1.48 \text{ MAD}$$

- Si la distribución es muy asimétrica *s* >> MAD

DISTANCIA INTERCUARTIL

- El *rango intercuartil o distancia intercuartil* de un conjunto de datos es la distancia entre los dos cuartiles

$$D_I = C_S - C_I$$

Interpretación

- Indica el rango donde se encuentra el 50% central de las observaciones.

Distancia intercuartil. Propiedades

- Si todos los datos son iguales $D_I = 0$.
- D_I puedes ser igual a cero aún cuando no todos los datos sean iguales.

  Ejemplo 5 12 12 12 12 12 20 $n = 7$ $C_I = 12$ $C_S = 12$ $D_I = 0$
- Es una medida robusta de dispersión.
- Cuando la distribución es simétrica y acampanada

$$D_I \cong \frac{4}{3} s$$

- Para distribuciones muy asimétricas *s* > D_I

DISTANCIA INTERCUARTIL

- Ejemplo

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

$$D_I = C_S - C_I = 320 - 140 = 80$$

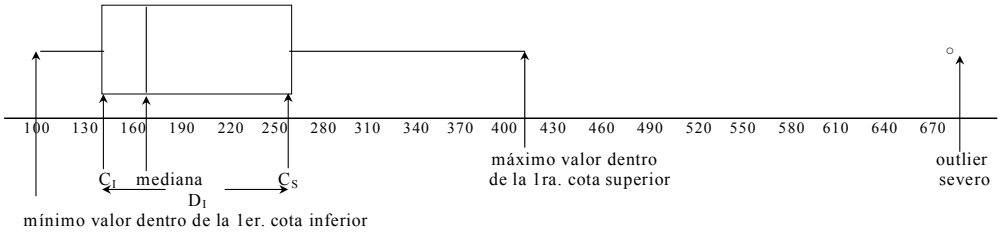
GRÁFICO DE CAJA (Box-plot)

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	67

- C_I = 140 = 170 C_S = 320 D_I = 320 – 140 = 80

  2ª cota inferior = $C_I - 3\ D_I = 140 - 3 \cdot 80 = -100$
  1ª cota inferior = $C_I - 1.5\ D_I = 140 - 1.5 \cdot 80 = 20$
  1ª cota superior = $C_S + 1.5\ D_I = 320 + 1.5 \cdot 80 = 440$
  2ª cota superior = $C_S + 3\ D_I = 320 + 3 \cdot 80 = 580$

GRÁFICO DE CAJA (Box-plot)



BOX - PLOT

- *¿Qué se observa?*
  - Un dato outlier.
  - Distribución con asimetría derecha.
- *Un box-plot muestra*
  - Una medida de posición robusta (mediana)
  - Una medida de posición robusta (D_I)
  - Simetría o no de la distribución
  - Criterio para detectar datos outliers
  - 5 números resúmenes