Elementos de la Estadística Básica

Juancho, Andrés y Fito

23 de Junio del 2020

Qué es? Cómo nace?

1. La Estadística a es la ciencia que se encarga de recoger, organizar e interpretar los datos.

Qué es? Cómo nace?

1. La Estadística a es la ciencia que se encarga de recoger, organizar e interpretar los datos.
2. Es esencial para interpretar los datos que se obtienen de la investigación científica.

Qué es? Cómo nace?

1. La Estadística a es la ciencia que se encarga de recoger, organizar e interpretar los datos.
2. Es esencial para interpretar los datos que se obtienen de la investigación científica.
3. La Estadística (del latín, Status o ciencia del estado) se ocupaba sobre todo de la descripción de los datos fundamentalmente sociológicos: datos demográficos y económicos(censos de población, producciones agrícolass, riquezas, etc.), principalmente por razones fiscales.

Qué es? Cómo nace?

1. La Estadística a es la ciencia que se encarga de recoger, organizar e interpretar los datos.
2. Es esencial para interpretar los datos que se obtienen de la investigación científica.
3. La Estadística (del latín, Status o ciencia del estado) se ocupaba sobre todo de la descripción de los datos fundamentalmente sociológicos: datos demográficos y económicos(censos de población, producciones agrícolass, riquezas, etc.), principalmente por razones fiscales.
4. Posteriormente (s. XVIII) su uso se extiende a problemas físicos (principalmente de Astronomía).

Para qué sirve?

1. Análisis de muestras (inferencias)

Para qué sirve?

1. Análisis de muestras (inferencias)
2. Descripción de datos (resumir)

Para qué sirve?

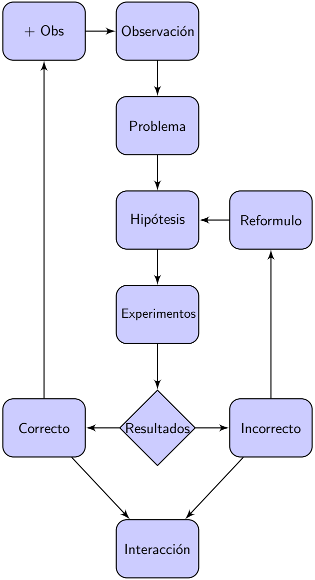
1. Análisis de muestras (inferencias)
2. Descripción de datos (resumir)
3. Contraste de hipótesis (validar)

Para qué sirve?

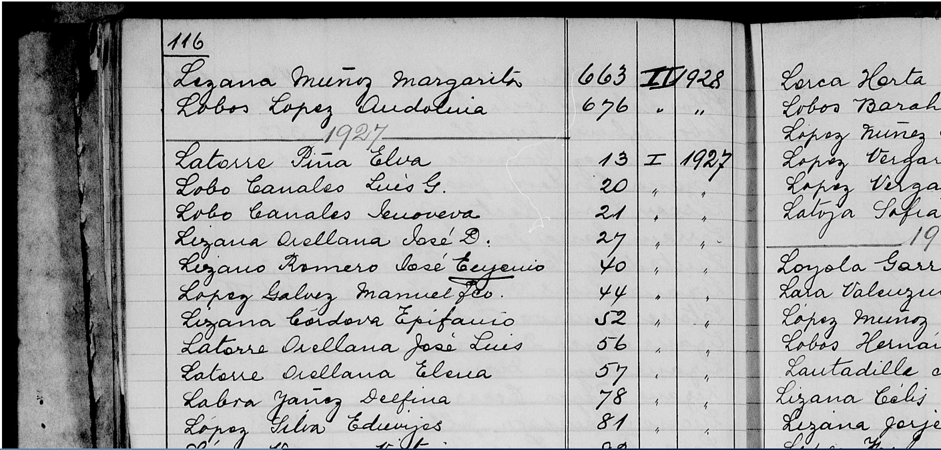
1. Análisis de muestras (inferencias)
2. Descripción de datos (resumir)
3. Contraste de hipótesis (validar)
4. Medición de relaciones entre variables estadísticas

Para qué sirve?

1. Análisis de muestras (inferencias)
2. Descripción de datos (resumir)
3. Contraste de hipótesis (validar)
4. Medición de relaciones entre variables estadísticas
5. Predicción


+ Obs
Observación
Problema
Hipótesis
Reformulo
Experimentos
Correcto
Resultados
Incorrecto
Interacción

Registro de nacimientos 1885-1930, Chile, Valle del Cachapoal.



116

Lizana Muñoz Margarita	663	II	1928
Lobos Lopez Audolina	676	"	"
1927			
Latorre Piña Elva	13	I	1927
Lobo Canales Luis G.	20	"	"
Lobo Canales Genoveva	21	"	"
Lizana Orellana José D.	27	"	"
Lizano Romero José Eugenio	40	"	"
Lopez Galvez Manuel Fco.	44	"	"
Lizana Córdova Epifanio	52	"	"
Latorre Orellana José Luis	56	"	"
Latorre Orellana Elena	57	"	"
Labra Yañez Delfina	78	"	"
López Silva Eduvijes	81	"	"

Lorca Horta
Lobos
López Muñoz
López
López
Latoja Sofia
Loyola
Lara
López Muñoz
Lobos
Lizana Celis
Lizana

Lenguaje utilizado

1. Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. (finita o infinita). Pro ejemplo: Las estrellas de la vía lactea, los habitantes de un país etc.

Lenguaje utilizado

1. Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. (finita o infinita). Pro ejemplo: Las estrellas de la vía lactea, los habitantes de un país etc.
2. A un subconjunto (o parte) de una población se le denomina **muestra**. Por ejemplo, algunas estrellas de la vía lactea.

Lenguaje utilizado

1. Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. (finita o infinita). Pro ejemplo: Las estrellas de la vía lactea, los habitantes de un país etc.
2. A un subconjunto (o parte) de una población se le denomina **muestra**. Por ejemplo, algunas estrellas de la vía lactea.
3. A la cantidad de elementos de una muestra se le llama **tamaño** de una muestra.

Lenguaje utilizado

1. Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. (finita o infinita). Pro ejemplo: Las estrellas de la vía lactea, los habitantes de un país etc.
2. A un subconjunto (o parte) de una población se le denomina **muestra**. Por ejemplo, algunas estrellas de la vía lactea.
3. A la cantidad de elementos de una muestra se le llama **tamaño** de una muestra.
4. El caso particular, en que una muestra incluye a *todos* los elementos posibles se denomina **censo**.

Lenguaje utilizado

1. Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. (finita o infinita). Pro ejemplo: Las estrellas de la vía lactea, los habitantes de un país etc.
2. A un subconjunto (o parte) de una población se le denomina **muestra**. Por ejemplo, algunas estrellas de la vía lactea.
3. A la cantidad de elementos de una muestra se le llama **tamaño** de una muestra.
4. El caso particular, en que una muestra incluye a *todos* los elementos posibles se denomina **censo**.
5. *En particular, nos interesan variables cuantitativas en esta primera parte*.

Representación Gráfica

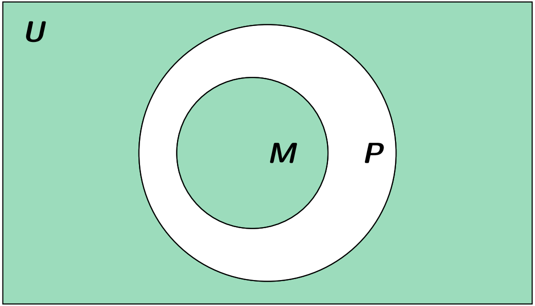


Figure: *

Conceptos o lenguaje mencionado. La **U** representa un universo de objetos de estudio. La **P** es una población dentro de ese universo. La **M** es una muestra.

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:
   - Discretos: Puedes numerar esos valores (listarlos). Por ejemplo: número de casas en un barrio (1,2,3,. . .)

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:
   - Discretos: Puedes numerar esos valores (listarlos). Por ejemplo: número de casas en un barrio (1,2,3,. . .)
   - Contínuos: No puedes listarlos. Por ejemplo: Temperatura del agua: 37.9876 C°.

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:
   - Discretos: Puedes numerar esos valores (listarlos). Por ejemplo: número de casas en un barrio (1,2,3,...)
   - Contínuos: No puedes listarlos. Por ejemplo: Temperatura del agua: 37.9876 C°.
2. **Dimensión**: Es el número de variables estadísticas que representan nuestro objeto de estudio. Por ejemplo para calcular la velocidad de un auto necesitas saber la distancia recorrida y el tiempo que demoró. Esta variable llamada *velocidad* tiene 2 dimensiones.

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:
   - Discretos: Puedes numerar esos valores (listarlos). Por ejemplo: número de casas en un barrio (1,2,3,. . .)
   - Contínuos: No puedes listarlos. Por ejemplo: Temperatura del agua: 37.9876 C°.
2. **Dimensión**: Es el número de variables estadísticas que representan nuestro objeto de estudio. Por ejemplo para calcular la velocidad de un auto necesitas saber la distancia recorrida y el tiempo que demoró. Esta variable llamada *velocidad* tiene 2 dimensiones.
   - Comenzaremos con el estudio de variables discretas!

Lenguaje Utilizado

1. **Variable Estadística**: es el símbolo o carácter que representa nuestro objeto de estudio. Y puede tener diferentes tipos de *valores*:
   - Discretos: Puedes numerar esos valores (listarlos). Por ejemplo: número de casas en un barrio (1,2,3,...)
   - Contínuos: No puedes listarlos. Por ejemplo: Temperatura del agua: 37.9876 C°.
2. **Dimensión**: Es el número de variables estadísticas que representan nuestro objeto de estudio. Por ejemplo para calcular la velocidad de un auto necesitas saber la distancia recorrida y el tiempo que demoró. Esta variable llamada *velocidad* tiene 2 dimensiones.
   - Comenzaremos con el estudio de variables discretas!
   - Para ellos adicionaré un término más llamado **rango de una variable discreta**, que es la diferencia entre el máximo y el mínimo valor observado.

Estudio de Variables Discretas

Supongamos que estudiamos el crecimiento una población pequeña y nos interesan los nacimientos. Entonces, pasamos por aquella población preguntando el número de hijos obteniendo la siguiente lista: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$. Entonces, relacionemos los conceptos vistos!.

- La variable estadística es: **número de nacimientos**.

Estudio de Variables Discretas

Supongamos que estudiamos el crecimiento una población pequeña y nos interesan los nacimientos. Entonces, pasamos por aquella población preguntando el número de hijos obteniendo la siguiente lista: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$. Entonces, relacionemos los conceptos vistos!.

- La variable estadística es: **número de nacimientos**.
- La **muestra** que tomamos tiene **tamaño**: 20.

Estudio de Variables Discretas

Supongamos que estudiamos el crecimiento una población pequeña y nos interesan los nacimientos. Entonces, pasamos por aquella población preguntando el número de hijos obteniendo la siguiente lista: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$. Entonces, relacionemos los conceptos vistos!.

- La variable estadística es: **número de nacimientos**.
- La **muestra** que tomamos tiene **tamaño**: 20.
- La variable es **discreta**, pues toma solo los valores $k = 1, 2, 3, 4, 5$. También podemos decir esto matemáticamente a través de la siguiente nomenclatura $x_1 = 1, x_2 = 2, \ldots, x_5 = 5$. De forma general, nuestras variables x_i con $i = 1, \ldots, 5$ toma valores discretos entre 1 y 5.

Estudio de Variables Discretas

Supongamos que estudiamos el crecimiento una población pequeña y nos interesan los nacimientos. Entonces, pasamos por aquella población preguntando el número de hijos obteniendo la siguiente lista: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$. Entonces, relacionemos los conceptos vistos!.

- La variable estadística es: **número de nacimientos**.
- La **muestra** que tomamos tiene **tamaño**: 20.
- La variable es **discreta**, pues toma solo los valores $k = 1, 2, 3, 4, 5$. También podemos decir esto matemáticamente a través de la siguiente nomenclatura $x_1 = 1, x_2 = 2, \dots, x_5 = 5$. De forma general, nuestras variables x_i con $i = 1, \dots, 5$ toma valores discretos entre 1 y 5.
- Ahora explicaré como se construye una tabla de frecuencias de variable discreta, utilizando este ejemplo!.

Estudio de Variables Discretas

- **Frecuencia Absoluta**: Número de veces que aparece repetido el valor en cuestión. Por ejemplo para nuestro conjunto: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$; el valor 1 aparece repetido 6 veces. Por lo tanto la **frecuencia absoluta** de la variable **1** es igual a **6**. Denotaremos por n_i este concepto.

Estudio de Variables Discretas

- **Frecuencia Absoluta**: Número de veces que aparece repetido el valor en cuestión. Por ejemplo para nuestro conjunto: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$; el valor 1 aparece repetido 6 veces. Por lo tanto la **frecuencia absoluta** de la variable **1** es igual a **6**. Denotaremos por n_i este concepto.
- **Frecuencia Relativa**: Es la división entre la frecuencia absoluta y el tamaño de la muestra. Para el mismo caso anterior tendríamos: $f_1 = \frac{1}{6} = 0.30$ que se lee *"la frecuencia relativa de 1 es igual a 1 dividido en 6."*

Estudio de Variables Discretas

- **Frecuencia Absoluta**: Número de veces que aparece repetido el valor en cuestión. Por ejemplo para nuestro conjunto: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$; el valor 1 aparece repetido 6 veces. Por lo tanto la **frecuencia absoluta** de la variable **1** es igual a **6**. Denotaremos por n_i este concepto.
- **Frecuencia Relativa**: Es la división entre la frecuencia absoluta y el tamaño de la muestra. Para el mismo caso anterior tendríamos: $f_1 = \frac{1}{6} = 0.30$ que se lee *"la frecuencia relativa de 1 es igual a 1 dividido en 6."*
- **Frecuencia Acumulada**: Suma de las frecuencias absolutas de los valores inferiores o igual a x_i , o número de medidas por debajo, o igual, que x_i. Denotaremos este concepto por N_i .

Tabla de frecuencias en variable discreta

-

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

Tabla de frecuencias en variable discreta

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

- **Frecuencia Relativa Acumulada**: División entre la frecuencia acumulada y el total de observaciones. Denotado por F_i.

Tabla de frecuencias en variable discreta

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

- **Frecuencia Relativa Acumulada**: División entre la frecuencia acumulada y el total de observaciones. Denotado por F_i.
- Interpretemos! **multiplicamos por 100**.

Tabla de frecuencias en variable discreta

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

- **Frecuencia Relativa Acumulada**: División entre la frecuencia acumulada y el total de observaciones. Denotado por F_i.
- Interpretemos! **multiplicamos por 100**.
- f_i: porcentaje en relación al total.

Tabla de frecuencias en variable discreta

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00
- **Frecuencia Relativa Acumulada**: División entre la frecuencia acumulada y el total de observaciones. Denotado por F_i.
- Interpretemos! **multiplicamos por 100**.
- f_i: porcentaje en relación al total.
- F_i: porcentaje por debajo de la variable en cuestión.

Tabla de frecuencias en variable discreta

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00
- **Frecuencia Relativa Acumulada**: División entre la frecuencia acumulada y el total de observaciones. Denotado por F_i.
- Interpretemos! **multiplicamos por 100**.
- f_i: porcentaje en relación al total.
- F_i: porcentaje por debajo de la variable en cuestión.
- Tarea (revisar libro)

Cómo caracterizo la distribución de mis datos?

- **Medidas de centralización**. (valor promedio, entorno de que valor se distribuyen)

Cómo caracterizo la distribución de mis datos?

- **Medidas de centralización**. (valor promedio, entorno de que valor se distribuyen)
- Medidas de dispersión. (variabilidad en relación al promedio)

Cómo caracterizo la distribución de mis datos?

- **Medidas de centralización**. (valor promedio, entorno de que valor se distribuyen)
- Medidas de dispersión. (variabilidad en relación al promedio)
- Momentos. (caso que generaliza los anteriores).

Cómo caracterizo la distribución de mis datos?

- **Medidas de centralización**. (valor promedio, entorno de que valor se distribuyen)
- Medidas de dispersión. (variabilidad en relación al promedio)
- Momentos. (caso que generaliza los anteriores).
- Asimetría y curtosis. (Grado de simetria en la distribución).

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

- $\overline{\mathbf{x}}$ representa el $\mathbf{x}$ promedio.

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

- $\overline{\mathbf{x}}$ representa el **x** promedio.
- N el número total de muestras o tamaño de la muestra.

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

- $\overline{\mathbf{x}}$ representa el **x** promedio.
- N el número total de muestras o tamaño de la muestra.
- Ejemplo: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

- $\bar{\mathbf{x}}$ representa el **x** promedio.
- N el número total de muestras o tamaño de la muestra.
- Ejemplo: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$
- $N = 20$

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

- $\bar{\mathbf{x}}$ representa el **x** promedio.
- N el número total de muestras o tamaño de la muestra.
- Ejemplo: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$
- $N = 20$
- $\displaystyle\sum_{k=1}^{N} x_k = 2 + 1 + 1 + \cdots + 3 + 2 + 1 = 45$

Promedio

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad (1)$$

- $\overline{\mathbf{x}}$ representa el **x** promedio.
- N el número total de muestras o tamaño de la muestra.
- Ejemplo: $C = \{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$
- $N = 20$
- $\sum_{k=1}^{N} x_k = 2 + 1 + 1 + \cdots + 3 + 2 + 1 = 45$
- $\overline{\mathbf{x}} = 2.25$

Mediana

Una medida de centralización importante es la mediana M_e . Se define ésta como una medida central tal que, con los datos ordenados de menor a mayor, el 50 % de los datos son inferiores a su valor y el 50% de los datos tienen valores superiores. Es decir, la mediana divide en dos partes iguales la distribución de frecuencias. Ejemplo:

$$C_o = \{1, 1, 1, 1, 1, 1, 2, 2, 2, \mathbf{2}, \mathbf{2}, 2, 2, 3, 3, 3, 3, 4, 4, 5\} \tag{2}$$

- Como $N = 20$ es par la primera mitad tiene valor superior 2, y la segunda mitad valor inferior 2. Entonces debo calcular: $M_e = \frac{2+2}{2} = 2$. La mediana es igual a dos.

Mediana

Una medida de centralización importante es la mediana M_e . Se define ésta como una medida central tal que, con los datos ordenados de menor a mayor, el 50 % de los datos son inferiores a su valor y el 50% de los datos tienen valores superiores. Es decir, la mediana divide en dos partes iguales la distribución de frecuencias. Ejemplo:

$$C_o = \{1, 1, 1, 1, 1, 1, 2, 2, 2, \mathbf{2}, \mathbf{2}, 2, 2, 3, 3, 3, 3, 4, 4, 5\} \tag{2}$$

- Como $N = 20$ es par la primera mitad tiene valor superior 2, y la segunda mitad valor inferior 2. Entonces debo calcular: $M_e = \frac{2+2}{2} = 2$. La mediana es igual a dos.
- Si fuera impar, por ejemplo:

$$D_o = \{1, 1, 1, 2, 2, \mathbf{3}, 3, 3, 3, 4, 5\} \tag{3}$$

La mediana es igual a 3.

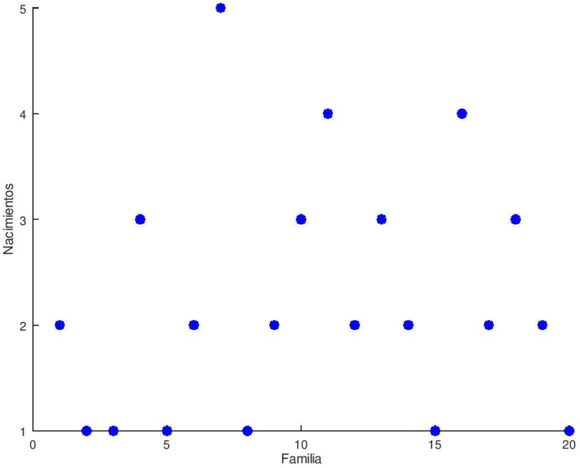
Moda

Aquél valor que tiene frecuencia máxima. El que más se repite!. En nuestro ejemplo sería:

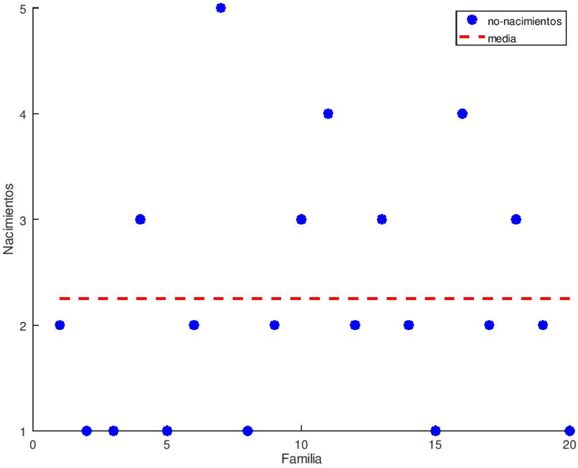
$$C_o = \{1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5\} \tag{4}$$

el valor 2.

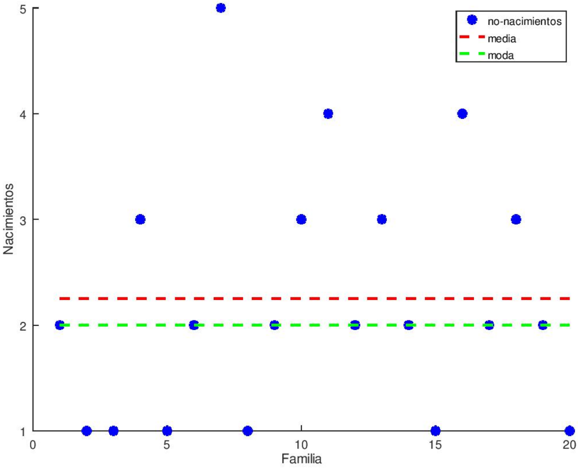
Utilizando puntos



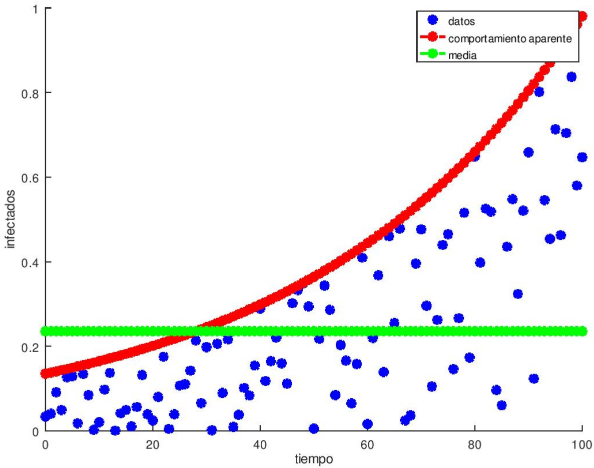
Utilizando puntos



Utilizando puntos



Centralizar no necesariamente caracteriza

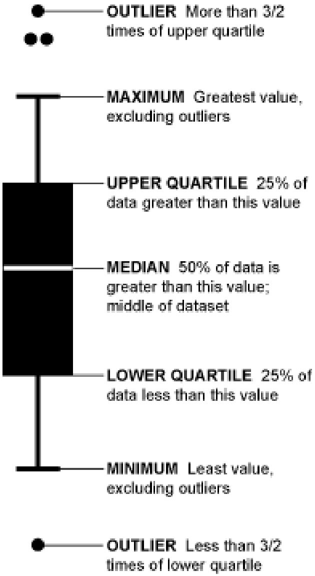


Conclusiones

- Usos e poquito de historia.
- Lenguaje técnico.
- Permite Caracterizar Datos.
- Inferir.
- Organizar.
- Frecuencias.
- Centralización.

Cuartiles, deciles y percentiles

- Divide en 3 partes los datos.
- El primer cuartil será la medida tal que el 25% de los datos son menores a este valor.
- El segundo cuartil es equivalente a la mediana.
- El tercer cuartil será la medida tal que el 75% de los datos son inferiores a este valor.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

Número de hijos en una muestra de 20 familias.

x_i	N_i
1	6
2	13
3	16
4	19
5	20

$1-1-1-1-1--1-2-2-2-2--2-2-2-3-3--3-3-4-4-5$

$$N/4 = 20/4 = 5 \rightarrow Q_{\frac{1}{4}} = 1$$
$$N/2 = 20/2 = 10 \rightarrow Q_{\frac{1}{2}} = M_e = 2$$
$$3N/4 = 20/4 = 15 \rightarrow Q_{\frac{3}{4}} = 3$$

Ejemplo 1: Python

```
import numpy as np
import statistics
import matplotlib.pyplot  as plt
# variable          # tiempo
x = [1,2,3,4,5]     t = [0,1,2,3,4]
# tamano x
L = len(x)
# calculo de la media             # truco para ver la medi
meanx = statistics.mean(x)        mean_vec = meanx * np.or
```