Statistica 1: Prof Maurizio Pratelli (1)

E-Mail: Maurizio.Pratelli@unipi.it

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

- Software: R-project
- Libro: S. Ross Probabilità e Statistica per l'ingegneria e le Scienze. Apogeo.
- Appunti: Statistica significa Raccogliere Dati. Un campione serve ad avere informazioni sulla popolazione.

Es: Elezioni politiche. Campione → Sondaggio sulle intenzioni di voto.

I campioni diversi danno risultati diversi

Esempio: 1200 componenti elettroniche.

Quante sono difettose?

Si prende un campione ad esempio di 30. Di questi 30 sono difettose

$\frac{2}{30} = 0.07$ → 7% sono difettose. (stima)

Quanto è affidabile la stima?!

La statistica si fa sempre su un campione e non sull'intera popolazione.

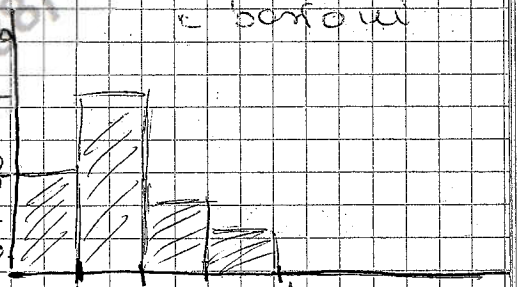
Statistica → teorica Descrittiva: Senza un modello probabilistico.
→ Probabilità.
→ Inferenza statistica: Interpretare i dati con un modello probabilistico (Capitolo 2 del libro).

Statistica descrittiva: Si hanno dei dati che si vogliono descrivere.

Es: Popolazione di 30 anni → celibi
→ sposati
→ separati
→ divorziati

87 dati — 29 celibi → frequenza assoluta
— 31 Sposati
— 17 Separati
— 10 Divorziati

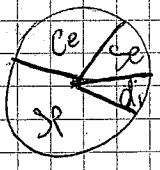
Istogramma con barre



$\frac{29}{87}$ ⇒ frequenza relativa

la somma delle frequenze relative deve essere 1.

Istogramma a torta



MASTER COPY
Tel. 050 9312126
Cell. 328 9337745

- Diagrammi "Stem and leaf"

Esempio:

6.28
8.48
8.12
4.21
4.32
6.18
6.51

6	. 28 . 48 . 51
4	. 21 . 37
8	. 12 . 49

Esempio:

6.28
6.18
6.36
6.21
6.24
6.17

6.1	. 07 . 08
6.2	. 08 . 04 . 08
6.3	. 06

Dati Numerici: $X = (x_1, \dots x_n)$ $n = \text{length}(x)$ (lunghezza)

- Media (Empirica) → $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ $\text{mean}(x) = \frac{\text{SUM}(x)}{\text{length}(x)}$
- Mediana → 17 | 41 | 29 | 21 | 32 | 41 | 13
- Moda (Se esiste) 13 | 17 | 21 | 29 | 32 | 41 | 41

mediana

I Dati si mettono sempre in ordine crescente.

(4)

Se i numeri sono dispari la mediana si trova al posto $n+1$. Se sono pari

ES: 13 | 17 | 21 | 28 | 32 | 41 | 41 | 44

30,5 → la media tra i due.

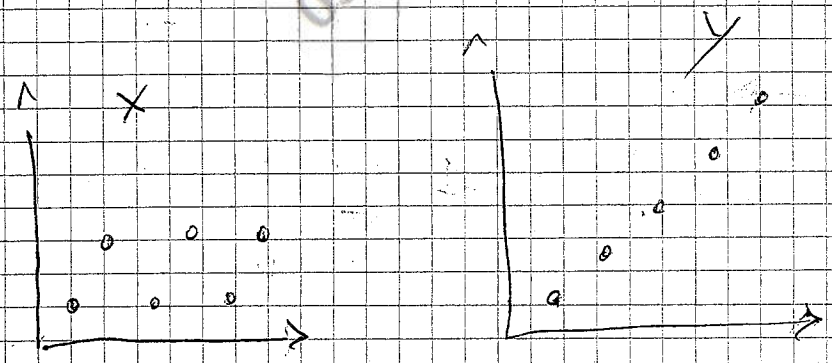
Moda: Il termine più frequente se entrambi.

Varianza campionaria: $Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ Misura della dispersione dei dati

Deviazione standard (scarto quadratico medio): $Sd_x = \sqrt{Var(x)} = \frac{\sqrt{(x_i - \bar{x})^2}}{\sqrt{n-1}}$

$Var = 0$ Se $x_1 = x_2 = \dots = x_n$

Plot (x)



$\bar{x} = \bar{y}$ $Sd(x) \ll Sd(y)$

Linguaggio R

Var(x) =

```
> x
> xbar = sum(x) / length(x)
```

= <-

stessa cosa

oppure

```
xbar <- sum(x) / length
```

```
> Var(x) = sum((x - xbar)^2) / (length(x) - 1)
```

comando predefinito → nel programmare funzione solo mean(x)

Definizione:

```
Sol(x) = sqrt(Var(x))
```

codice

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

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \bar{x}^2 - \sum_{i=1}^{n} 2x_i \cdot \bar{x} = \sum x_i^2 + n\bar{x}^2 - 2\sum x_i \cdot \bar{x}$$

$$\sum_{i=1}^{n} x_i \cdot \bar{x} = \bar{x} \sum_{i=1}^{n} x_i = n\bar{x}^2$$

Disuguaglianza di Chebychev (empirica)

$d > 0$ $\quad \#\{i : |x_i - \bar{x}| > d\} \leq \dfrac{(n+1)\,Var(x)}{d^2}$

↳ var dimensionale → n° decimali