STATISTICA I

raccogliere dati, analizzare e trarre previsioni

26/09/18

- POPOLAZIONE
- CAMPIONE

↓ due campioni diversi danno risultati diversi

ES: 1200 componenti elettroniche prodotte difettose?
prendo un campione di 30
trovo 2/30 difettosi

$\frac{2}{30} \pm 0.07$ = 7% difettosi

STIMA ↑

Quando è affidabile questa stima?

Statistica si divide in:

STATISTICA

- DESCRITTIVA: descrivere oggetti senza interpretazione probabilistica
- PROBABILITÀ
- INFERENZA STATISTICA: interpretare i dati con un modello probabilistico

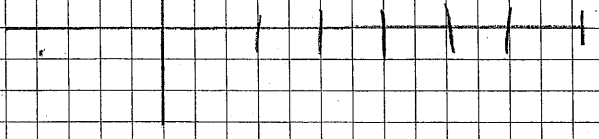
- DATI: rappresentare, descrivere...

ES: 40 auto → celibi, sposati, separati, divorziati

87 dati

		FREQUENZA ASSOLUTA
28	celibi	
31	coniugati	
17	separati	
10	divorziati	

ISTOGRAMMA a BASTONI



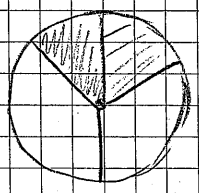


diagramma a torta

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

29/87 FREQUENZA RELATIVA

Diagrammi "Stem and Leaf" Ramo e foglia

↓ mettere in evidenza la prima cifra

es:

6,28
8,48
8,12
7,21
7,37
6,68
6,51

⇒

6 | .28, .68, .51
7 | .21, .37
8 | .12, .48

6.28
6.19
6.36
6.21
6.24
6.17

⇒

6.1 | .07, .03
6.2 | .01, .04, .08
6.3 | .06

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

• Dati numerici

$x = (x_1, \dots x_n)$

n = length (x) $(\in \mathbb{R})$

media (empirica) → $\bar{x} = \frac{x_1 + \dots x_n}{n}$ = mean (x) = sum (x) / length (x)

mediana

moda (se $\exists$) = le carattere più frequente

17 | 41 | 29 | 21 | 29 | 13 | 41 | 44

13 17 29 (32) 41 41 44

↑ mediana

$2n+1$ al posto $n+1$

↓ disporre

se pari ($2n$) ⇒ media tra i due centrali

13 17 21 29 | 32 41 41 44

↓

30,5

Varianza campionaria $\mathrm{Var}(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ $\left(\frac{1}{n}\right)$

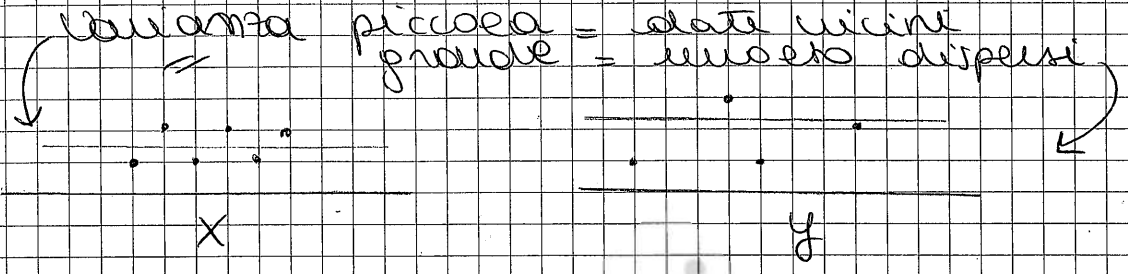
MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Deviazione Standard (scarto quadratico medio) $\sigma_X = \sqrt{Var(X)} = plot(X)$

↓ misura della dispersione dei dati

CASO LIMITE: $Var(X)=0 \iff x_1=x_2=\dots=x_n$

variabile piccola = dati vicini
variabile grande = molto dispersi



$sd(X) \ll sd(Y)$

in R

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

calcolo predefinito

$sd(x) = sqrt(var(x))$

$X = (x_1 \dots x_n)$

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

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

dim dell' OSS.

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

$= \sum_{i=1}^{n} x_i\bar{X} = \bar{X}\cdot\sum_{i=1}^{n} x_i = n\bar{X}^2$ ← diventa

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Disuguaglianza di Chebicheff (empirica)

$d > 0$ — quanto che si scostano dalla media $> d$

$$\#\{i : |x_i - \bar{x}| > d\} \leq \frac{(n-1)\,\mathrm{var}(x)}{d^2}$$

cardinalità.

es: $A = \{1, 2, 3\}$

$\#A = 3$

DIM

$$d^2 \cdot \#\{i : |x_i - \bar{x}| > d\} \leq \mathrm{var}(x)(n-1) \quad \left[= \sum_{i=1}^{n} (x_i - \bar{x})^2\right]$$

da dim.

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \geq \sum_{i : |x_i - \bar{x}| \geq d} (x_i - \bar{x})^2 + \sum_{i : |x_i - \bar{x}| > d} (x_i - \bar{x})^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \geq \sum_{i : |x_i - \bar{x}| \geq d} (x_i - \bar{x})^2 \geq d^2 \cdot \#\{i : |x_i - \bar{x}| > d\}$$

coppia di dati (x, y) $\quad length(x) = length(y)$

$(x_1, y_1) \; \ldots \; (x_n, y_n)$

covarianza (empirica) $\quad \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = cov(x,y)$

coefficiente di correlazione (empirico)

$$\rho(x,y) = \frac{cov(x,y)}{sd(x)\, sd(y)}$$

dim: $\quad sd(x) = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$

$$\rho(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} \leftarrow \langle a, b \rangle$$

$\|a\| \nearrow \qquad \nwarrow \|b\|$

$|\rho(x,y)| \leq 1$ diz. di Shwarz

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Sfruttando $a, b \in \mathbb{R}^n$

$\langle a, b \rangle \leq \|a\| \cdot \|b\|$

- prendo $a = (x - \bar{x}) = (x_1 - \bar{x} \dots)$

  $b = (y - \bar{y})$

$r > 0$ → "vanno nello stesso senso" se aumento uno, aumento l'altro

$r < 0$ "vanno in senso opposto"

$r = 0$ sono "slegate"

con limite

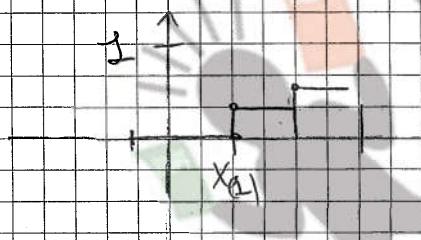
$r = 1$

- $y_i = a x_i + b \quad a > 0$

  $a < 0 \quad r = -1$

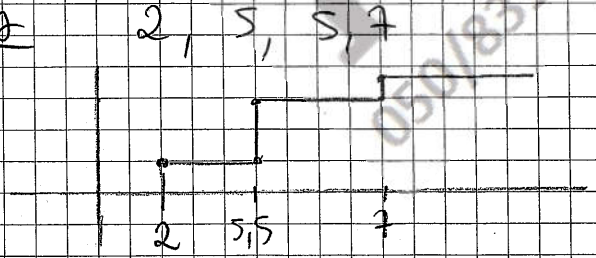
$X (x_1, x_2 \dots x_n)$

$x_{(1)}, x_{(2)}, \dots x_{(n)}$ ordinati crescendo



c.d.f. (cumulative distribution function) ovvero → funzione di ripartizione empirica

es



in R → `x = c(2, 3, 67)` c lo definisce

`> x = rep(50, 2, 6)` = 50 numeri e cosa fa ogni e se

- `rep` → ripetere
- `rnorm` → numeri distribuiti

`> x = rnorm(x, type "l")` un particolare type

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745