28/2/18 ALTEZZA PESO SESSO SCUOLA PROVENIENZA

Maurizio Pratelli

M = 1 classico = 1 Pisa = 1
F = 2 scientifico = 2 Toscana = 2
altro = 3 Altro = 3

R (R-project) (programma gratuito di base da scaricare)

R-studio

→ variante preferita

Rcmdr Probabilità e statistica per l'ingegneria e per le Scienze Applicate

Franchetta Statistica multivariata con R Pisa University Press

ARGOMENTI CORSO

1) Statistica descrittiva (raccogliere dati, descriverli ma senza interpretazione di un modello → senza probabilità)

2) Probabilità e variabili aleatorie

3) Inferenza statistica (→ statistica con la probabilità)

è in cuore della materia

Popolazione e campione

↓ ↓

certi dati parte della popolazione di cui si esaminano dei dati

popolazione: sondi elettorali in Sardegna

popolazione = elettori in Sardegna

campione ← "exit pool" "sondaggi elettorali"

parte anche piccola della popolazione che ci serve per rilevare info

Supponiamo di avere una indagine su caratteri "qualitativi" (sesso, scuola e provincia sono caratteri qualitativi)

frequenza → assoluta

↓

relativa

↓

si esprime in % solitamente informa di più di quella assoluta

es. 197 studenti

64 ragazze

frequenza assoluta

$\frac{64}{197}$ = frequenza relativa delle ragazze

il partito ABC ha ottenuto 1 228 796 voti

↳ freq. assoluta

≈ 17,76 % voti → freq. relativa

Moda = carattere più frequente

147 numeri tutti diversi
a priori

⇒ dovremmo avere un carattere quantitativo $X = (x_1, \dots, x_n)$

→ con questi numeri tutti diversi ha senso parlare di MEDIA

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

media empirica (o campionaria) del carattere

$X = (x_1, \dots, x_n)$ n - numeri

carattere quantitativo

Varianza campionaria (campione)

$$\hookrightarrow = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ quando abbiamo i campioni conviene usare questa varianza

Varianza empirica

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

$$Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ → scarto quadratico medio

$$\sigma(X) = \sqrt{Var(X)}$$ → deviazione standard

A cosa serve la varianza?

È una misura della dispersione

$Var(x) = 0 \iff$ tutti i dati sono uguali

$Var(x)$ "piccola" $\iff$ dati simili

$Var(x)$ "grande" $\iff$ dati diversi

Disuguaglianza $(d > 0)$ → dati numerici

$$\#\{x_i : |x_i - \bar{x}| > d\}$$

(cardinalità) quanti dati si scostano dalla media per più di d

disuguaglianza di Čebyšëv

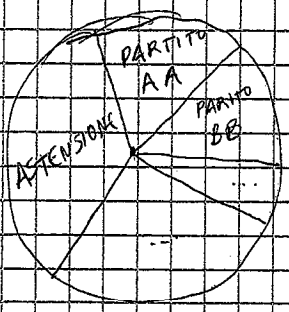
$$\#\{x_i : |x_i - \bar{x}| > d\} \leq \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{d^2} \geq 1$$

è un numero maggiore di d

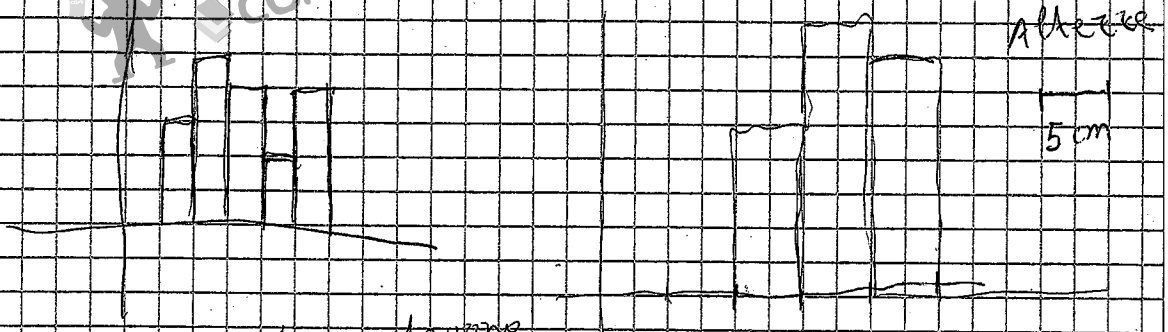
$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)} \cdot \frac{(n-1)}{d^2} = Var(x) \cdot \frac{(n-1)}{d^2}$$

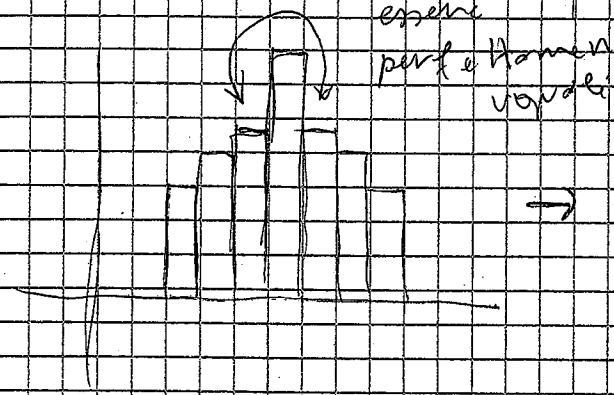
più la varianza è piccola più sono pochi i valori che si scostano dalla media di d

Istogrammi a torta



ISTOGRAMMA A BASTONI (A CANNE)

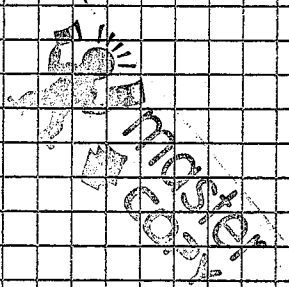




→ istogramma normale

↓

ha un andamento simmetrico rispetto al centro



right skewed (sbilanciato verso destra)

[left skewed]

bimodale