Imp statistica => utilizzare i dati raccolti per trarne info

ARTE DI APPRENDERE DATI

Statistica = elaborare i dati e ricavare informazioni su un insieme di oggetti→POPOLAZIONE
Si considera un sottogruppo della popolazione, detto CAMPIONE, poiché considerare tutta la popolazione è pressoché impossibile data la complessità.
Per far si che il campione sia rappresentativo della popolazione, si sceglie a caso.
La statistica nasce a Firenze, e deriva da Stato poiché veniva fatta da esso.

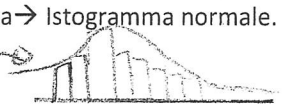
Statistica descrittiva→ si occupa dei metodi di esposizione e sintesi dei dati → raccogliere e descrivere i dati
2 Ipotesi:
-raccogliamo dati numerici
-il campione considerato non è troppo numeroso cioè i dati si suddividono in un numero di valori non troppo elevato

master COPY

Sia v uno dei valori
f = frequenza assoluta = quanti dati sono uguali a v.
$\frac{f}{n}$ = ~~frequenza assoluta~~ RECAPITO, n = ampiezza del campione (numero dei dati raccolti)

Grafico a barre→ ascisse, valori ottenuti. Ordinate, frequenze.
Se i dati raccolti sono elevati, il grafico a barre diventa complicato e dunque ricorro alle classi. Tipicamente il numero adeguato di classi è 10. Il grafico a barre con la suddivisione in classi diventa un istogramma.

1 picco {
- Per molti dati numerici l'istogramma assume una forma a campana→ Istogramma normale.
- Right skewed→ istogramma presenta una coda più lunga a destra
- Left skewed→ istogramma presenta una coda più lunga a sinistra



epiché = BIMODALE

Statistiche:
Grandezze calcolabili dai dati $(x_1, \dots, x_n) = x$ che ci danno informazioni sui dati

1. Media campionaria → media aritmetica dei dati

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

2. Varianza campionaria→ ci dice come i dati sono concentrati/dispersi attorno alla media

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$$

La varianza è sempre ≥ 0 poiché somma di scarti quadratici.
$(x_i - \bar{x})^2$ distanza di ogni valore dalla media.
Verrebbe da dividere per n, ma per motivi tecnici dividiamo per n-1

master COPY
master COPY COPISTERIA

3. Varianza empirica→ non adatta allo studio di un campione

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

4. Deviazione standard campionaria

$$\sigma = \sqrt{Var(x)}$$

Il significato della varianza è dato dalla disuguaglianza di Chebyshev→ ci da un limite superiore della frazione dei dati che si allontana dalla media

DISUGUAGLIANZA DI CHEBYSHEV

$$\text{(*)}\left\{ x_i : |x_i - \bar{x}| > d \right\} \leq \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{d^2}, \quad d > 0$$

(*) cardinalità

$\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{d^2}$ limite superiore

Vale per un campione $x = (x_1, \dots, x_m)$ con media $\bar{x}$ e $Var(x) > 0$

→ Il numero dei dati che si discosta dalla media più di un certo valore d non può essere superiore a quel rapporto

$$\rightarrow Var(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 \rightarrow \sum_{i=1}^{m}(x_i - \bar{x})^2 = Var(x)(m-1)$$

① $$\boxed{\# \left\{ x_i : |x_i - \bar{x}| > d \right\} \leq \frac{m-1}{d^2} Var(x)}$$ disuguaglianza in termini di VAR(x)

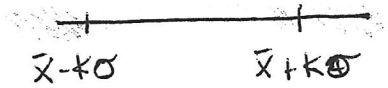
⇓

② ["almeno una frazione di $\left(1 - \frac{1}{k^2}\right)$ dei dati cade tra $\bar{x} - k\sigma, \bar{x} + k\sigma$ "]

Dim

① ⇒ ②

Scelgo $d = *\sigma$



$S_k := \left\{ x_i : |x_i - \bar{x}| \leq k\sigma \right\}$

$C_k = \left\{ x_i : |x_i - \bar{x}| > k\sigma \right\}$ m dati

$\# S_k + \# C_k = m \rightarrow \# C_k = m - S_k$

dis. di Cheby

$$\# C_k \leq \frac{m-1}{k^2\sigma^2} Var(x) \rightarrow \# C_k \leq \frac{m-1}{k^2}$$

$$m - \# S_k \leq \frac{m-1}{k^2} \rightarrow -\# S_k \leq \frac{m-1}{k^2} - m$$

$$\# S_k \geq m - \frac{m-1}{k^2} \rightarrow \frac{\# S_k}{m} \geq 1 - \frac{m-1}{mk^2} \Rightarrow 1 - \frac{1}{k^2}$$

$\frac{m-1}{m} \leq 1$

$$\frac{\# S_k}{m} \geq 1 - \frac{1}{k^2} \Rightarrow \text{②} \quad *$$

Esiste una regola empirica che rispetta la disuguaglianza di Chebyshev, ma la migliora e vale per dati che hanno una distribuzione a campana cioè un istogramma normale.

La regola empirica dice che:

68% $\bar{x} \pm \sigma$

95% $\bar{x} \pm 2\sigma$ → entro 3 deviazioni standard ci stanno tutti i dati

98% $\bar{x} \pm 3\sigma$

Coppie di dati numerici

x e y due tipi di grandezze

(x_i, y_i) → campione bivariato

Ci interessa studiare se esiste qualche relazione fra le due grandezze

Correlazione positiva→ se a valori elevati di x corrispondo valori elevati di y

$$(x_i - \bar{x}) > 0 \quad \rightarrow \quad (y_i - \bar{y}) > 0$$

Correlazione negativa→ se a valori elevati di x corrispondo valori bassi di y

$$(x_i - \bar{x}) > 0 \quad \rightarrow \quad (y_i - \bar{y}) < 0$$

Per misurare questa correlazione introduco la covarianza

Covarianza campionaria

$$\mathrm{cov(x,y)} = \frac{\sum_{i=i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Coefficiente di correlazione

$$\mathrm{r = cor(x,y)} = \frac{\mathrm{cov(x,y)}}{\sigma(x)\sigma(y)} \qquad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Proprietà:

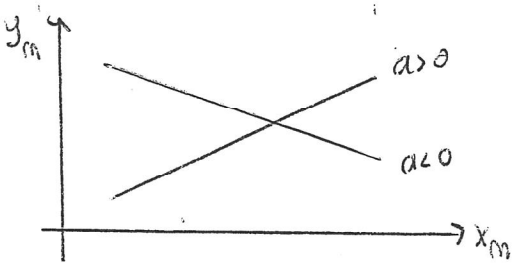
r>0 → correlazione positiva

r<0 → correlazione negativa

$-1 \leq r \leq 1$

$y_i = ax_i + b$ a>0 → r = 1 valore massimo

$y_i = ax_i + b$ a<0 → r = -1 valore minimo



Identità fondamentale

Utile per velocizzare il calcolo della var(x)

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

Dim

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

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

def $\bar{x}$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad \square$$

master COPY

Funzione di ripartizione empirica

Funzione costruita a partire da un campione numerico $y = (y_1, \dots, y_n)$.

$x_1, \dots, x_k$ i k valori distinti assunti dai dati del campione, e li ordiniamo in ordine crescente, chiamandoli:

$$x_{(1)}, \dots, x_{(k)} : x_{(1)} \le x_{(2)} \le \cdots \le x_{(k)}$$

$$F_n(x) = \sum_{i:\, x_i \le x}^{n} \frac{f_i}{n}$$

Sommo per tutti i valori ≤ x le frequenze relative.

Se $x \le x_{(1)}$ non ho alcuna frequenza da sommare→ funzione vale 0

In ogni punto $x_{(k)}$ la funzione fa un salto pari alla frequenza relativa di $x_{(k)}$

La funzione di ripartizione è una funzione a gradini :

0 $x \le x_{(1)}$

1 $x \ge x_{(k)}$

master COPY COPISTERIA

Esempio

Raccogliamo le altezze di un campione di 10 bambini did 3 anni

$y = (y_1, \dots, y_{10})$ n=10

100 98 102 96 95 98 105 90 95 98

y_1 y_2 y_3 ……y_{10}

Valori distinti: 100 98 102 96 95 105 90

x_1 x_2 x_3 …… x_7 k=7

Li riordiniamo:

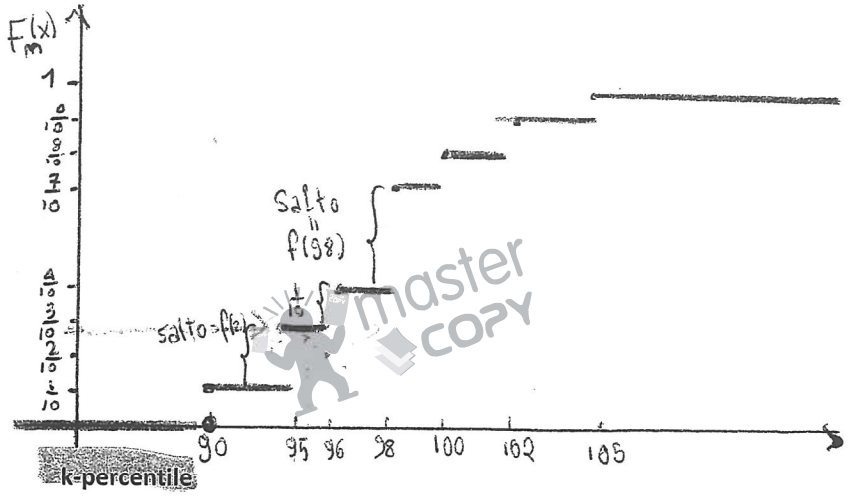
$x_{(1)} = 90$ $x_{(2)} = 95$ $x_{(3)} = 96$ $x_{(4)} = 98$ $x_{(5)} = 100$ $x_{(6)} = 102$ $x_{(7)} = 105$

master COPY

Per ciascuno di questi valori ci segniamo le frequenze

Valori	Frequenza assoluta	Frequenza relativa
90	1	1/10
95	2	2/10
96	1	1/10
98	3	3/10
100	1	1/10
102	1	1/10
105	1	1/10

Funzione di ripartizione empirica:



$F_m(x_{(1)}) = \frac{f_1}{n}$

$F_m(x_{(2)}) = \frac{f_1+f_2}{n}$

k-percentile

Def→ sia k un numero intero t.c $0 \leq k \leq 100$. Dato un campione di dati numerici se ne esiste uno che è:
- maggiore o uguale di almeno il k-percento dei dati
-minore o uguale di almeno il (100-k) percento dei dati
Allora k è il percentile k-esimo

Il k-esimo percentile è un numero q t.c. il $\frac{k}{100}$dei dati sta prima di q e il $(1-\frac{k}{100})$dei dati sta dopo q.

Il 25-esimo percentile si dice PRIMO QUARTILE
Il 50-esimo percentile si dice SECONDO QUARTILE
Il 75-esimo percentile si dice TERZO QUARTILE

Come determinare il percentile k-esimo di un campione di numerosità n?
Chiamo $\beta = \frac{k}{100}$
$n\beta$ dati devono essere $\leq q$
$n(1-\beta)$ dati devono essere $\geq q$

Se $n\beta$ non è un intero, il dato che soddisfa la richiesta di essere il k-esimo percentile, è quello che occupa la posizione data da $n\beta$ arrotondando all'intero successivo

Se $n\beta$ è intero, sia il dato $x_{n\beta}$ sia il suo successivo (in particolare tutti i numeri compresi fra essi) soddisfano le richieste e quindi la quantità cercata è la media di questi due valori.

Dal grafico della funzione di ripartizione empirica si può leggere direttamente i percentili.

Esempio
Dati numerici:
90 95 95 96 98 98 98 98 100 102 105
k=30→ cercare il 30-esimo percentile significa cercare il numero q tale che il 30% dei dati sta prima di q e il 70% dei dati sta dopo q.
n=10 30% dei dati → 3 dati 70% dei dati → 7 dati
$n\beta$ è un intero $\beta = \frac{30}{100} \rightarrow \frac{30}{100} * 10 = 3$ → scelgo la media tra 95 e 96→ q=95.5
Sia 95 che 96 soddisfacevano la richiesta→ 3 dati minori di loro e 7 dati maggiori di loro

master COPY COPISTERIA