Appunti di Statistica

Prof. di Gesu

Appunti del corso Prof. Maurizio Pratelli

- lezioni su teams - Registro (ma è meglio seguire)
- Usa e-learning
- mail: giacomo.digesu@unipi.it

(Testo: "Probabilità e statistica per l'ingegneria e le scienze" S. Ross)

Statistica → scienza dei dati (analisi e interpretazione)
↳ al giorno d'oggi ce n'è una quantità enorme.

In questo corso → Rudimenti di "statistica classica"

statistica descrittiva statistica inferenziale

STATISTICA DESCRITTIVA

Abbiamo un insieme molto grande di dati

Problema: come li descriviamo in modo sintetico, efficiente, suggestivo?

esempio neuroscienziati cercano di comprendere quali parti del cervello si occupano di elaborare l'informazione visuale data dai colori.

Esperimento: prendono una persona e danno uno stimolo visivo che consiste nel mostrare delle figure in movimento alternando figure a colori e figure in B/N, in istanti $t = 1, \dots, T$ fanno una tomografia a risonanza del cervello.

- Otteniamo una quantità enorme di dati.
- Calcoliamo "correlazione" tra l'input dato dallo stimolo visivo e la risposta $x(t)$ al variare del tempo.
- selezioniamo i voxels che sono fortemente correlati con lo stimolo
- generiamo un'immagine che evidenzia in rosso le parti del cervello selezionate.

Tutta l'analisi descrittiva è stata fatta usando il software R

STATISTICA INFERENZIALE

Partiamo sempre da un insieme di dati. Vogliamo utilizzare i dati per:
- fare delle previsioni
- prendere delle decisioni

} basandoci su un modello matematico della realtà che vogliamo descrivere.

I modelli matematici considerati sono probabilistici.

Teoria della probabilità ci fornisce il linguaggio fondamentale per poter fare della statistica inferenziale.

La teoria della probabilità è una teoria astratta che descrive gli "esperimenti aleatori".

Esempi di esperimenti aleatori:
- lancio dado / moneta
- osservazione del meteo nei giorni futuri
- osservazione velocità di propagazione di un virus nel prossimo anno.

Vedremo due teoremi fondamentali:
- legge dei grandi numeri
- teorema limite centrale

esempio (statistica inferenziale)
① Problema dell' "Exit Poll": la popolazione vota per una carica cattera da due candidati A e B.
Viene estratto un campione (a caso) dalla popolazione che ci dice la preferenza espressa nel voto.
Vogliamo fare una previsione sull'esito del voto basandoci sulle indicazioni del campione estratto.
(concetto di "stimatore", "intervalli di confidenza")
② Raccogliamo i dati ottenuti lanciando 100 volte un dado.
Dobbiamo decidere in base ai dati se considerare il dado come "regolare" o "taroccato". Come faccio a decidere?
("test statistico" o "test d'ipotesi")

Nel corso tratteremo:
- Elementi di statistica descrittiva
- Introduzione all'utilizzo del software R / fogli elettronici
- Elementi di teoria della probabilità
- Elementi di Statistica inferenziale

04/03

Software R → sviluppato a partire dagli anni '90 da Ross Ihaka e Robert Gentleman, basandosi sul precedente software S.
→ è un linguaggio di programmazione e un ambiente di sviluppo specifico per l'analisi statistica dei dati

↳ caratteristiche:
- è gratuito
- cross-platform
- open-source
- molto diffuso per l'analisi statistica dei dati (gran parte delle tecniche statistiche classiche e moderne sono già implementate in R)

> "prompt"
frase tra virgolette "..."
3x200 → scrivo 3*2
nello script # commento, se voglio che sia stampata " "

05/03

La media (aritmetica o empirica) dei numeri $x_1, \dots, x_n$ è $\overline{x} = \frac{x_1 + \dots + x_n}{n}$

Il simbolo di sommatoria Σ (sigma) $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Moda, media e mediana sono indicatori di centralità, ossia riassumono in un solo valore centrale tutti i dati (con criteri diversi).

09/03

I quartili

Dati $x_1, \dots, x_n$ ordinati ($x_1 \leq \dots \leq x_n$), se non lo sono vanno ordinati.

es: $n = 13$ 1, 5, 5, 7, 9, 12, 29, 43, 72, 84, 84, 86, 86

mediana ↳ detta anche secondo quartile

valore minimo → quartile zero 1

valore massimo → quarto quartile 86

Dividendo ulteriormente: 5, 5, (7), 9, 12 — 7 = primo quartile

43, 72, (84), 84, 86 — 84 = terzo quartile

I quartili 1°, 2°, 3° non sono sempre univocamente determinati (quando non lo sono prendo una approssimazione, ad esempio la media)

Significato dei quartili:
- quartile zero → non ci sono valori più piccoli
- primo quartile → il 25% dei valori sono più piccoli
- secondo quartile → il 50% dei valori sono più piccoli
- terzo quartile → il 75% dei valori sono più piccoli
- quarto quartile → tutti i valori sono più piccoli

Alcuni indicatori di dispersione (o indici di variabilità)

Dati $x_1, \dots, x_n$

- $\delta = \frac{1}{n} \cdot (|x_1 - \bar{x}| + \dots + |x_n - \bar{x}|)$ più è grande, più i dati differiscono dalla media. ⇒ più c'è variabilità

  ↳ $\delta = 0 \iff x_1 = \dots = x_n = \bar{x}$

  δ = SCARTO MEDIO ASSOLUTO

- $s = \sqrt{\frac{1}{n} \cdot [(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]}$ o $s = \sqrt{\frac{1}{n-1} [(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]}$ ↳ è più usata

  s = DEVIAZIONE STANDARD (campionaria)

- $s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2$

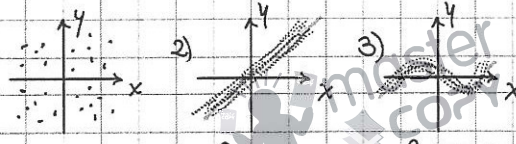
  s^2 = VARIANZA CAMPIONARIA

11/03

dati numerici
$x = (x_1, \dots, x_n)$
$y = (y_1, \dots, y_n)$

Il coefficiente di correlazione
Date delle coppie di dati $(x_1, y_1), \dots, (x_n, y_n)$
es: - Input, Risposte
- sondaggio e chiedere: peso e altezza // età e guadagno // h e €/anno ....
Il coefficiente di correlazione misura il "legame di natura lineare" tra i dati x e i dati y.

es:



in 2) e 3) c'è un legame, qual'è la differenza? Tipologia di legame

legame lineare → i dati sono vicini ad una retta
RETTA DI REGRESSIONE $y = ax + c$

legame periodico → i dati sono vicini ad una curva.

→ $cor(x,y)$ basso.

In generale il coefficiente di correlazione, $cor(x,y)$, tra i dati x e i dati y è un numero compreso tra -1 e 1 e si ha
- $cor(x,y) = 0$ → nessuna correlazione
- $cor(x,y) = 1$ → dati su una retta a pendenza positiva ($>x \Rightarrow >y$)
- $cor(x,y) = -1$ → dati su una retta a pendenza negativa ($>x \Rightarrow <y$)
- $cor(x,y)$ vicino a 1 → i dati si trovano vicino ad una retta a pendenza positiva (2)
- $cor(x,y)$ vicino a -1 → i dati si trovano vicino ad una retta a pendenza negativa

$$cor(x,y) = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2} \cdot \sqrt{\sum_{k=1}^{n} (y_k - \bar{y})^2}}$$

Possiamo riscrivere $cor(x,y)$ in termini della covarianza campionaria $cov(x,y)$ e di S_x, S_y (deviazione standard di x e y)

Covarianza campionaria
$$cov(x,y) = \frac{1}{n-1}\left[(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})\right]$$
$$= \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

osservazione:
$$S_x^2 = var(x) = cov(x,x)$$
$$S_y^2 = var(y) = cov(y,y)$$

$$cor(x,y) = \frac{cov(x,y)}{S_x \cdot S_y}$$

Per la disuguaglianza di Cauchy-Schwartz: $|cov(x,y)| \le \frac{S_x S_y}{\sqrt{var(x)}\sqrt{var(y)}}$

dim: 1) consideriamo il caso $\bar{x}, \bar{y} = 0$

Allora $cov(x,y) = \sum x_k y_k$ e $var(x) = \sum x_k^2$, $var(y) = \sum y_k^2$

$|\sum x_k y_k| \overset{?}{\le} \sqrt{\sum x_k^2}\sqrt{\sum y_k^2}$ → è la disuguaglianza di C.S. !! (vale sempre)

2) se $\bar{x} \neq 0$ e/o $\bar{y} \neq 0$

definisco $u = x - \bar{x}$ e $w = y - \bar{y}$ $(u_1 = x_1 - \bar{x}, \dots ; w_1 = y_1 - \bar{y}, \dots)$

allora $\bar{u} = 0$ e $\bar{w} = 0$

quindi mi ricaduco a 1) applicato a u e w.

Quindi abbiamo dimostrato che $-1 \le cor(x,y) \le 1$

Determinare m e c della retta di regressione $y = mx + c$

retta di regressione $m = m_*$, $c = c_*$ con m_*, c_* tali che

$$\min_{m,c} \sum_{k=1}^{n} (y_k - mx_k - c)^2 = \sum_{k=1}^{n} (y_k - m_* x_k - c_*)^2$$

Tra tutte le rette, quella di regressione è quella più vicina ai dati

Proposizione:

i) vale $m_* = \frac{cov(x,y)}{var(x)}$ $\left(= cor(x,y)\frac{S_y}{S_x}\right)$ e $c_* = \bar{y} - \frac{cov(x,y)}{var(x)}\bar{x}$

ii) vale $\underbrace{\sum_{k=1}^{n} (y_k - m_* x_k - c_*)^2}_{\text{la distanza dei punti dalla retta di regressione}} = \sum_{k=1}^{n} (y_k - \bar{y})^2 (1 - [cor(x,y)]^2)$

Conseguenze

1) da i) segue subito: $cor(x,y) \lesseqgtr 0 \iff m_* \lesseqgtr 0$

2) da ii) segue subito: $|cor(x,y)| = 1 \Rightarrow \sum |y_k - m_* x_k - c_*|^2 = 0$

ovvero i dati si trovano esattamente sulla retta di regressione

3) se $cor(x,y) = 0 \to m_* = 0$, $c_* = \bar{y}$ e quindi la migliore approssimazione a y_k è $m_* x_k + c_* = \bar{y}$ (indipendente da x_n!)

(sapere x_k non ci aiuta a indovinare y_k, il miglior "guess" è sempre $\bar{y}$)

4) da ii) segue che più $|cor(x,y)|$ è vicino a 1 e più i dati sono vicini alla retta di regressione

dim: (idea: si considera la funzione di due variabili $(m,c) \mapsto f(m,c) = \sum (y_k - mx_k - c)^2$ e si trova il minimo assoluto che è proprio (m_*, c_*))