Data Mining e Bioimmagini

Misure di posizione, acquisizione dato, Boxplot e R studio

Prof. Pierangelo Veltri – 27/09/2023- Autori: Maturo, Accetturo - Revisionatori: Accetturo

Misure di posizione

- Z-score: indica il numero di deviazioni standard di cui una data osservazione si allontana dalla media $Z = (X - \bar{x})/3$. *Serve ad indicare quanto si allontana un valore dal valore medio.*
- La **varianza** di un insieme di dati è una misura di variabilità corrispondente al quadrato della deviazione standard.
- **Varianza campionaria**: s^2 = quadrato della deviazione standard s
- **Varianza di una popolazione**: σ^2 quadrato della deviazione standard della popolazione σ
- **Calcolo outlier**: indicano valori significativi per distanza da valori netti;

Acquisizione del dato

Il prof ci tiene molto all'acquisizione del dato
La parte di analisi sui fenomeni è fortemente dipendente dall'acquisizione del dato.

#esempio un test su ragazzi in un'aula:

- si identifica il numero delle osservazioni "n"
- si calcola la somma dei valori $\sum x$
- si calcola la sommatoria dei quadrati originali $\sum x^2$
- si calcola il valore della deviazione standard rispetto ai dati campionari, cioè di quanto si allontanano.
  L'indicatore può servire rispetto ad uno studio di popolazione.
  Formula applicata $s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$

#esempio nascituri
Dati: 4000g, Peso medio nascituri= 3150. Dev standard s = 653
Calcolo: Z= 4000-3150/653 = 1,22
Risultato osservato 1.22 volte la dev standard lontani dalla media

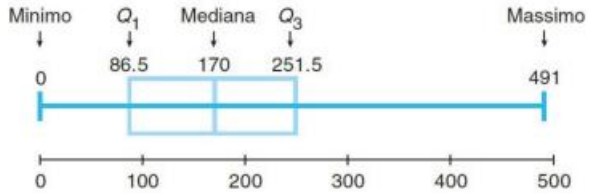
Percentili

I percentili sono misure di posizione, indicate con P_1, P_2,…, P_{99}, che dividono il dataset in 100 intervalli e sono utili per stabilire una soglia di riferimento.
#esempio
Riportiamo tutti i valori di TSH dei bambini a cuisi è fatto il test dell'ipotiroidismo e stabiliamo come valore soglia il 99esimo percentile.
Quindi tutti i bambini al di sotto della soglia avranno valori di TSH normali, mentre quelli superiori(che nel nostro caso corrispondono al 1%) verranno richiamati per ulteriori test.
Così facendo andremo a richiamare tutti i bambini al di sopra di tale soglia per ulteriori test.

BOXPLOT

Il boxPlot rappresenta i 5 numeri di cui è costituito un dataset e sono: il valore Minimo, Primo quartile Q1, Secondo quartile (o mediana) Q2, Terzo quartile Q3, valore Massimo.

Identifica un insieme di dati su un intervallo di valori ammissibili per la popolazione.
Ad esempio, si prende l'intervallo dei battiti al minuto riconosciuti, poi si prende la popolazione e si va a rappresentare il boxPlot.

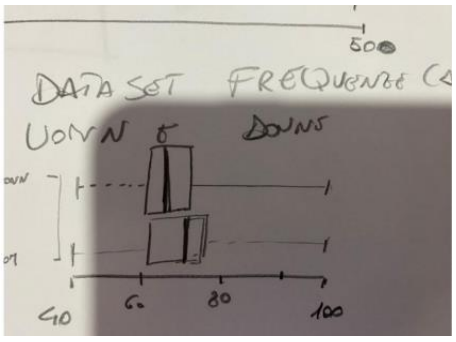


Ricordare che i dati che ne risultano devo essere sempre interpretati in base all'utilizzo che se ne fa

#esempio *rappresentazione dataset 40 fumatori*
Si inseriscono i livelli di cotinina di 40 fumatori in ordine crescente e inseguito andrebbero rappresentati i valori di frequenza o di utilizzo per la mappatura.

Tabella 2.9 Livelli di cotinina di 40 fumatori in *ordine crescente*.

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	265	266	277	284	289	290	313	477	491



#esempio *confronto tra due BoxPlot dei dataset che riportano la frequenza cardiaca di uomini e donne*

#Caso Studio
- Calcolo dei valori per la riabilitazione
L'arduino è un dispositivo programmabile che consente di fare delle acquisizioni di dati;
Sono stati selezionati 40 pazienti che hanno eseguito degli esercizi di riabilitazione per il recupero degli arti superiori.
Tramite questi esercizi viene acquisito il dato (quanto si riesce a tracciare in maniera continuativa e il tempo che si impiega per la lettera).
I dati sono scritti su un SD Card, per poi essere trasferiti nel PC in formato CSV e si aprono su Excel.

Su tale dataset è possibile fare analisi con i boxPlot e confrontare , ad esempio, il tempo medio che il paziente impiega e l'errore medio nel tracciamento della lettera.

Si rappresentano i risultati comparando nello stesso spazio di confronto i viola con i verdi che corrispondono alla popolazione utilizzata come campione di riferimento e alla popolazione utilizzata come campione patologico.

R STUDIO

- **Cosa è possibile fare?**
  Analizzare ed elaborare dati con caratteristiche e parametri delle risonanze magnetiche del brain *(volume corteccia periferica, brain matter, volume di fluido)* e altre informazioni numeriche, che sono il risultato di una fase di pre-processamento e misurazione che viene fatta sulla singola immagine.
  Molti dataset sono disponibili online con informazioni e descrizioni delle immagine.
  #Esempio nel campo medico si possono prendere dati di campioni dei pazienti patologici e di pazienti normali, per sviluppare dei box plot e metterli a confronto.

- **Come si presenta R studio?**
  Si presentano quattro finestre:
  - **Console:** dove si possono scrivere le istruzioni e dove viene scritto l'output numerico
  - **R-script**: dove si possono scrivere le istruzioni e salvarle per un successivo utilizzo
  - **environment-history**: dove si possono vedere le strutture dati utilizzate
  - **plots**: dove vengono scritti gli output grafici, l'help e package di R attivabili

- **Tips per l'utilizzo**
  - Le **istruzioni** possono essere scritte sulla console e poi eseguite battendo "invio".
  - Nella funzione **R-script** per eseguire le istruzioni su una o più righe cliccare il pulsante "run"
  - Le istruzioni sono automaticamente copiate sulla finestra console quando eseguite.
  - Sulla stessa riga si possono scrivere più istruzioni separate da " ; "
  - i **commenti** devono essere preceduti da #
  - l'output viene scritto sulla finestra **console** o sulla finestra **plot**

#Esempi
Traccia
Vengono misurate le pulsazioni cardiache al minuto di una classe dove il lancio di una moneta determina se devono correre per un minuto o no
Testa=correre, Croce = No

Valori misurati
pulse 1, pulse 2 Il numero di prime/seconde pulsazioni al minuto
ran indica il valore booleano della moneta (1: testa, 2:croce)
Smokes= se fumano, **sex=** genere, **height**= altezza, **weight**= peso
Activity= abitudine all'attività fisica (0:no – 1: poca – 2:moderata – 3: alta)

OBS	PULSE1	PULSE2	RAN	SMOKES	SEX	HEIGHT	WEIGHT	ACTIVITY
1	64	88	1	2	1	66.00	140	2
2	58	70	1	2	1	72.00	145	2
3	62	76	1	1	1	73.50	160	3

Come si caricano questi dati su R studio? 2 opzioni:

- Caricarli su foglio excel e in seguito caricarlo in R *(Più semplice)*
- Definire una matrice e inserire i dati direttamente dall'ambiente R

Come si estrapolano/leggono i dati ?

Si prende il nome della varibile con funzione lettura con il nome del file

```
pulse=read.table("C:/DATA/pulse.txt",header=T,row.name=1)
```

Come si accede alle variabili ?

- Procedere al nome della variabile, il nome del dataset seguito da $
  #esempio pulse$PULSE1
- Usando l'istruzione attach
  #esempio attach(pulse) per chiudere si utilizza detach(pulse)

#AltroEsempio
Indice di correlazione di Pearson (anche detto coefficiente di correlazione lineare) tra due variabili statistiche esprime relazione di linearità (o di vicinanza).
Più il valore sarà vicino a 1 e più ci sarà correlazione, mentre sarà vicino a 0 meno correlano.

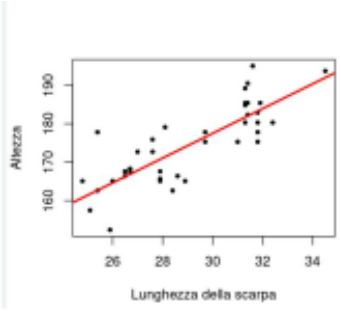
- Correlazione tra altezza e la lunghezza della scarpa

1. Si definisce la funzine di plot

```
plot(Lunghezza, Altezza, xlab="Lunghezza della scarpa (cm)",
     ylab="Altezza (cm)", col="red", pch=16)
```

pch: è una dimensione
xlab/ylab: indica cosa andrà scritto sul plotting

2. Si può avviare il test di correlazione che indica quanto le due variabili sono correlate
il valore di correlazione è 0.83



il dataset è interessante perché è vicino a 1 quindi dice che le due varibili sono correlate

Consiglio: è importante che il file si trovi all'interno della cartella sulla quale stiamo lavorando, altrimenti definire una variabile d'ambiente

3. Si arricchisce l'informazione, con altre informazioni cliniche (dati clinici) per vedere se l'informazione risultante è confermata.