STATISTICA: scienza dei dati, nel senso di analisi ed interpretazione.

La statistica moderna è legata al machine learning ovvero all'intelligenza artificiale.

STATISTICA CLASSICA

STATISTICA DESCRITTIVA

IDEA! più sono i dati e più precise sono le previsioni.

STATISTICA INFERENZIALE

- STATISTICA DESCRITTIVA (come riassumere i dati?)

La statistica descrittiva comprende generalmente un insieme di dati (generalmente molto grande)

=> l'obbiettivo è descrivere tale insieme in modo sintetico ed efficiente.

=> l'idea alla base di questa "riduzione" dell'insieme dei dati risiede nel concetto di CORRELAZIONE.

- STATISTICA INFERENZIALE

La statistica inferenziale ha ancora come punto di partenza un insieme di dati, ma questa volta vogliamo utilizzarli per fare delle previsioni e/o prendere delle decisioni.

=> il metodo della statistica inferenziale si distingue da altri metodi, anch'essi finalizzati a fare previsioni e/o prendere decisioni, per il semplice fatto che il metodo stesso è basato su un modello matematico che descrive la situazione reale: tale modello matematico è di tipo PROBABILISTICO.

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

TEORIA della PROBABILITÀ: è una teoria astratta che descrive gli esperimenti aleatori.

Gli elementi aleatori sono esempi molto concreti come per esempio:
- lancio di un dado/moneta
- osservazione del tempo a Pisa tra 1 settimana
- osservazione della velocità di propagazione di un virus nel prossimo anno.

Notazione matematica:

- simbolo di sommatoria $\sum$ ("sigma")

es. $\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \cdot \sum_{k=1}^{n} x_k$ → INDICE MUTO

Alcuni concetti fondamentali che poi andremo ad usare in R:

- MEDIA ⇒ $\frac{\text{somma dei dati}}{\text{numero dei dati}} = \bar{x}$ "media empirica".

- MEDIANA ⇒ dati i seguenti dati 3, 6, 2, 17, 9

⇓ li ordino

2, 3, [6], 9, 17

↳ è la MEDIANA

oppure quando si ha un numero pari di dati la cosa cambia leggermente:

dati i seguenti dati: 4, 3, 17, 2

⇓ li ordino

2, 3, 4, 17

- MODA ⇒ come suggerisce il nome stiamo parlando del valore che compare più spesso.

prendiamo i seguenti dati: 1, 7, 2, 9, 7, 7, 3, 2, 2

⇒ ho due mode: 7 e 2

oss. si possono avere anche più mode per lo stesso gruppo di dati.

⇒ media, mediana e moda sono chiamati INDICATORI DI CENTRALITÀ:

→ tutte e tre riassumono in un solo valore centrale tutti i dati; sono tre valori diversi, dunque si otterranno tre valori distinti ognuno ottenuto seguendo un criterio diverso.

⇒ si usano per sintetizzare tutto il campione: ciò potrebbe portare a risultati troppo approssimativi, dunque sarà necessario determinare degli intervalli di dispersività

QUARTILI

Partendo dai nostri dati ordinati $x_1, x_2, \dots, x_n$ ("ordinati" $\equiv x_1 \leq x_2 \leq \dots \leq x_n$) vediamo cosa sono i quartili a partire da un esempio:

Esempio.

Prendo una lista di 13 numeri:

1, 5, 5, 7, 9, 12, 29, 43, 72, 84, 84, 86, 86

=> in questo caso la mediana è 29 (valore centrale).

- la mediana è detta anche SECONDO QUARTILE ed è uno degli indicatori di centralità.
- il valore max ed il valore minimo sono detti rispettivamente QUARTO QUARTILE e QUARTILE ZERO.
- per trovare il PRIMO QUARTILE prendo i valori che trovo prima della mediana e seleziono tra questi il valore centrale
- per trovare il TERZO QUARTILE prendo i valori dopo la mediana e ne seleziono il valore centrale.

IDEA! "quartile" sta a suggerire che andiamo a suddividere i dati in "quattro parti".

=> mediana, I e III quartile possono non essere univocamente determinabili, e quando non lo sono si prendono delle approssimazioni (come per esempio la media aritmetica tra due valori)

Esempio

2, 8, 9, 9, 10, 11, 11, 11

- divido in quattro parti i miei dati.

max = 11

min = 2

III quartile = 11

mediana = 9,5

I quartile = 8,5

SIGNIFICATO QUARTILI

IDEA: avere un certo quartile mi dice che:

- QUARTILE ZERO: X valori più piccoli
- I QUARTILE: 25% dei valori sono più piccoli
- II QUARTILE: 50% dei valori sono più piccoli
- III QUARTILE: 75% dei valori sono più piccoli
- IV QUARTILE: tutti i valori sono più piccoli

INDICATORI di DISPERSIONE

=> ci dicono quanto/come i valori del nostro data-set variano attorno a quelli che abbiamo precedentemente chiamato "indicatori di centralità".

=> si chiamano INDICATORI di DISPERSIONE oppure INDICI di VARIABILITÀ

Dati $x_1, _, x_n$ voglio inizialmente vedere come questi si distribuiscono attorno alla media (quanto differiscono rispetto alla media):

IDEA: prendo ciascun dato x_i con $i=1,_,n$, e calcolo la distanza di ciascuno dalla media

$$\delta = \frac{1}{n}\left[(x_1-\bar{x})+(x_2-\bar{x})+_+(x_n-\bar{x})\right]$$ SCARTO MEDIO ASSOLUTO

NON È MOLTO USATO NELLA PRATICA

=> la media empirica è $\bar{x} = \frac{1}{n}(x_1+_+x_n)$

=> al posto dello scarto medio assoluto, che non è molto usato nella pratica, si usa la

DEVIAZIONE STANDARD:

$$s = \sqrt{\frac{1}{n-1}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+_+(x_n-\bar{x})^2\right]}$$ DEVIAZIONE STANDARD

=> rispetto allo scarto medio assoluto non si prendono le sole distanze, ma i quadrati delle distanze perché sono più precisi

=> lo scarto medio assoluto e la deviazione standard sono la stessa cosa: misurano la distanza tra la media e ciascun dato del mio data-set.

OSS Si usa la deviazione standard.

La varianza campionaria/scarto quadratico medio è il quadrato della deviazione Standard:

=> è un indicatore di dispersione (con una variabile).

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

VARIANZA CAMPIONARIA o SCARTO QUADRATICO MEDIO CAMPIONARIO.

COEFFICIENTE di CORRELAZIONE

=> misura in maniera quantitativa quanto una variabile influenza l'altra, (il coefficiente di correlazione, infatti, si misura tra una coppia di dati).

Date alcune coppie di dati come $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ le riscriviamo in maniera sintetica come

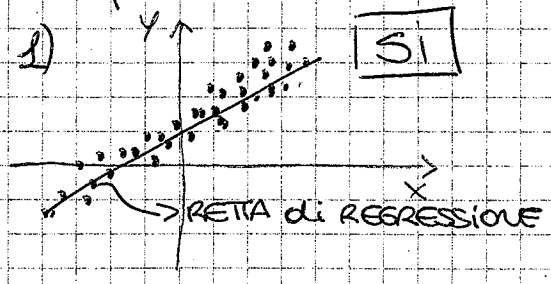
$X = (x_1, \dots, x_n)$

$Y = (y_1, \dots, y_n)$.

IDEA: le due variabili considerate sono generalmente legate per qualche ragione.

Il COEFFICIENTE di CORRELAZIONE misura il legame di natura lineare tra i dati X e i dati Y.

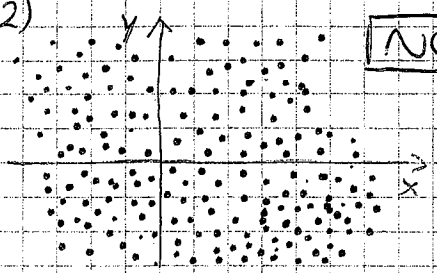
Esempio

1) 

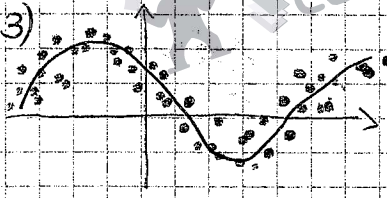
SI

=> dopo i dati si dispongono in varie posizioni interpretando le variabili X e Y come le coordinate di un punto nel piano.

=> in base alla disposizione dei dati possiamo capire se $\exists$ o meno un legame tra i dati di una coppia.

2) 

NO

3) 

SI

IDEA! se $\exists$ un legame tra i dati di una coppia allora conoscendo anche solo uno dei dati della coppia è possibile risalire all'altro dato.

Negli esempi precedenti vediamo che nel I esempio c'è un legame lineare, ovvero i dati sono allineati, all'incirca, attorno ad una retta, detta anche RETTA di REGRESSIONE.