Centro Stampa

N°1003

CLASSIFICAZIONE E INTERPRETAZIONE
TEORIA TEMI ESAME

DI BETTALE VALENTINA

APPLICATION DEVELOPMENT SVILUPPO DELLE APPLICAZIONI

COME POSSIAMO COSTRUIRE UN'APPLICAZIONE?

DAL PROBLEMA → SI ARRIVA AL SISTEMA
3 PASSI

PROBLEMA

1° **ANALISI** (CONCETTUALIZZAZIONE DEL PROBLEMA)

CAPIRE E STUDIARE BENE IL PROBLEMA X AVERE LE SPECIFICHE VOLUTE E X SAPERE QUALI SONO:

- I DATI A DISPOSIZIONE
- IL RUOLO E LA GESTIONE DELL'INCERTEZZA
- LE PROBLEMATICHE E CHE TIPO DI APPROCCIO BISOGNA UTILIZZARE

2° PROCESSO DI **COSTRUZIONE** DI UN MODELLO

GRAZIE A MATEMATICA, LOGICA, INFORMATICA..

E SCELTA DEL METODO CON CUI UTILIZZARE QUESTO MODELLO

MODELLO E METODO E IL TUTTO DEVE SUCCESSIVAMENTE ESSERE ADEGUATO AL SISTEMA SPECIFICO (TUNING)

ALLA FINE SI OTTIENE IL PROTOTIPO DEL SISTEMA

⚠ NEL CASO SI ACCORGA CHE MANCA QUALCOSA SI PUÒ TORNARE INDIETRO ALLA FASE DI ANALISI

3° PROCESSO DI **VERIFICA** E **VALIDAZIONE**

X IL PRODOTTO DEVE — SODDISFARE LE SPECIFICHE (VER.)
ESSERE FUNZIONALE X CHI LO UTILIZZA (VAL.)

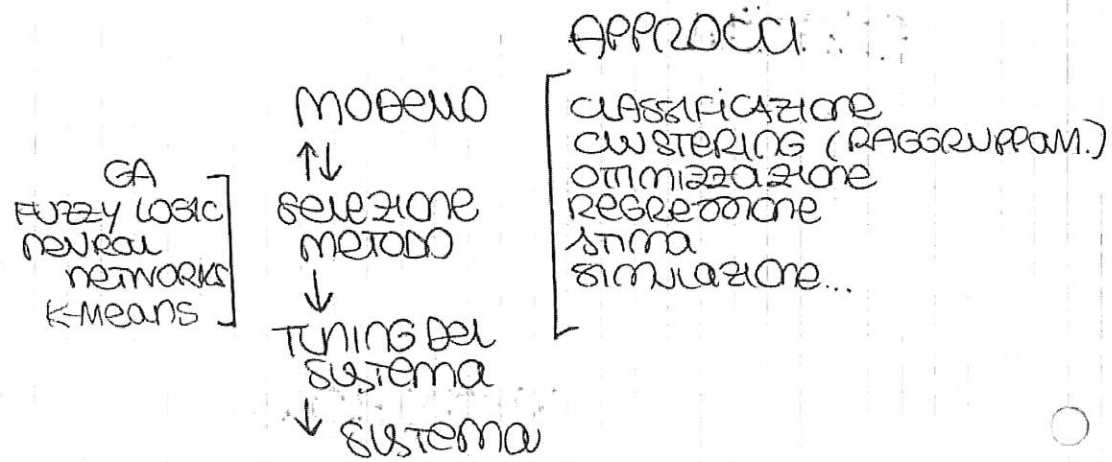
SISTEMA PRONTO

⚠ OGNI VOLTA CHE SI È RISOLTO UN PROBLEMA, CE NE SI CREA UN ALTRO (CERCHIO ∞)

NOI PARTIAMO DAL PUNTO ②, DAL CONSTRUCTION PROCESS

PROCESSO di COSTRUZIONE

INCLUDE:
- QUALE APPROCCIO USARE/SCEGLIERE X RISOLVERE IL PROBLEMA
- QUALE METODO E QUALE MODELLO SCEGLIERE
- IL TUNING (ADEGUAZIONE AL PROBLEMA)
- IL SISTEMA OTTENTO



1° PROBLEMA

COME SI POSSONO DESCRIVERE LE CARATTERISTICHE DI UN SET DI ELEMENTI?

COME **DESCRIVERE** I DATI?

UN ELEMENTO (O OGGETTO) È CARATTERIZZATO DA TANTI DATI

SI INCONTRA QUESTO PROBLEMA QUANDO SI PROVA AD ORGANIZZARE I DATI ACQUISITI DURANTE UN TRIAL

(= STUDIO CONTROLLATO E PROGETTATO DI ELEMENTI SELEZIONATI)

LA RISPOSTA AL PROBLEMA È LA **STATISTICA DESCRITTIVA**

METODI STATISTICI:

LE STATISTICHE POSSONO ESSERE DI 2 TIPI:

① **STATISTICA DESCRITTIVA**

USATA SEMPLICEMENTE X DESCRIVERE IL CAMPIONE CHE CI INTERESSA

È USATA - INNANZITUTTO X OTTENERE UN PRIMO SGUARDO SUI DATI
- POI X USARLI NEI TEST STATISTICI
- INFINE X INDICARE L'ERRORE ASSOCIATO AI RISULTATI E X GLI ALTRI GRAFICI

② STATISTICA INFERENZIALE

INFERENZA = DEDUZIONE GENERALIZZAZIONE

SI FANNO DELLE INFERENZE SU COME UN VALORE STATISTICO RAPPRESENTA TUTTA LA POPOLAZIONE

SERVE X VALIDARE LA RAPPRESENTATIVITÀ DEL CAMPIONE, ALTRIMENTI NON SI PUÒ ESTENDERE SUBITO UN RISULTATO DI UN TRIAL A TUTTA LA POPOLAZIONE
→ SELEZIONE TIPICA DEI TRIAL: CRITERI DI SELEZIONE X ANDARSENE

SONO TECNICHE CHE PERMETTONO DI USARE I CAMPIONI X FARE DELLE **GENERALIZZAZIONI** SULLA POPOLAZIONE DALLA QUALE SONO PRESI I CAMPIONI

⚠ LETTERE X INDICARE — POPOLAZIONE: μ, δ — CAMPIONI: $\bar{x}, s, a$

[LA POPOLAZIONE INCLUDE TUTTI GLI OGGETTI DI INTERESSE, MENTRE IL CAMPIONE È SOLO UNA PORZIONE DELLA POPOLAZIONE.

PARAMETRI → ASSOCIATI ALLE POPOLAZIONI. LETTERE GRECHE

STATISTICHE → ASSOCIATE AI CAMPIONI, LETTERE ROMANE

QUANDO SI USANO I METODI INFERENZIALI, È IMPORTANTE CHE IL CAMPIONE RAPPRESENTI ACCURATAMENTE LA POPOLAZIONE

∃ 5 MODI X ESTRARRE UN CAMPIONE

(TIPI DI ESTRAZIONE)

① **E. CASUALE**: (RANDOM)
CASUALITÀ TOTALE
OGNI ELEMENTO NELLA POPOLAZIONE HA LA STESSA PROBABILITÀ DI ESSERE ESTRATTO

È IL MODO PREFERITO DI ESTRAZIONE, MA È MOLTO DIFFICILE DA FARE

② **E. SISTEMATICA**: CAMPIONAMENTO SISTEMATICO OGNI n ELEMENTI
È + FACILE DEL RANDOM

EX: SEGNALI

③ E. conveniente (convenience) : non si cerca un campione specifico, ma chi c'è
ogni dato a disposizione è usato

Facile da fare, ma è forse la tecnica peggiore

Semplice da usare anche x l'organizzazione

④ E. cluster : divido la popolazione in gruppi e poi estraggo da essi degli elementi

La popolazione è solitamente divisa geograficamente

I gruppi sono chiamati cluster o block

→ si sceglie casualmente un gruppo e si usano TUTTI i suoi elementi

⑤ E. stratificato: la popolazione è divisa in gruppi chiamati strati, x qualche caratteristica, non geograficamente

ex: M e F

Un campione è preso da ogni strato usando ①, ② o ③

Parametri della statistica descrittiva

- Valor medio o media aritmetica ($\bar{x}$)

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

sommatoria di tutti gli elementi divisa x il numero di elementi stesso

- mediana

n dispari → $\frac{n+1}{2}$ ⇒ prendo il valore centrale

n pari → $\left(\frac{n}{2} + \left(\frac{n}{2}+1\right)\right)/2$ ⇒ prendo la media tra i 2 valori centrali

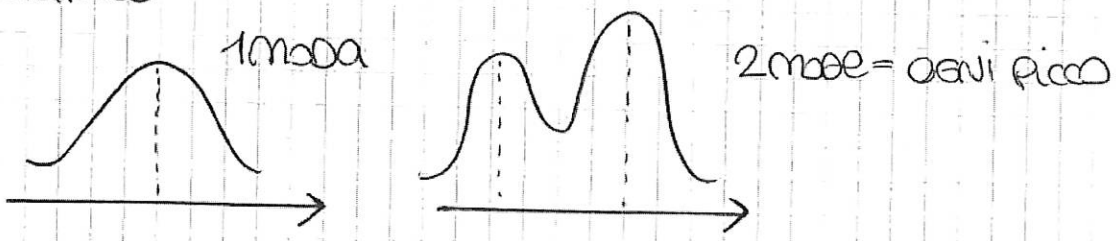
a sx e a dx della mediana devo avere lo stesso numero di elementi!

MODA

VALORE CHE SI RIPETE PIÙ FREQUENTEMENTE TRA TUTTE LE OSSERVAZIONI IN UN CAMPIONE

PUÒ NON ESSERE UNICA

VALORE CHE + FREQUENTEMENTE ASSUME LA VARIABILE



VARIANZA

DICE QUANTO È DISPERSA LA POPOLAZIONE:

$S\uparrow$: CAMPIONE DISOMOGENEO

$S\downarrow$: CAMPIONE OMOGENEO $\rightarrow \bar{x}$ SIGNIFICATIVO

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

DIPENDE DALL'UNITÀ DI MISURA DELLA VARIABILE!

DEVIAZIONE STANDARD

$s = \sqrt[2]{s^2}$

X TOGLIERE IL PROBLEMA DELLA DIPENDENZA E CONFRONTARE COME VARIANO 2 VALORI E LA LORO RIPRODUCIBILITÀ

COEFFICIENTE DI VARIAZIONE

$100\% \cdot \left(\frac{s}{\bar{x}}\right)$ VALORE ADIMENSIONALE!

QUESTO VALORE RIMANE SEMPRE LO STESSO INDIPENDENTEMENTE DA QUALI UNITÀ DI MISURA SI USANO

RANGE

DIFFERENZA TRA L'OSSERVAZIONE MASSIMA E LA MINIMA

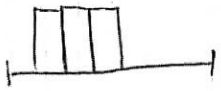
O

COPPIA FORMATA DAL VALORE MINIMO E DAL VALORE MASSIMO CHE TROVO NEL CAMPIONE

DÀ L'IDEA DI QUANTO È AMPIO IL CAMPO DI VALORI DELLA MIA VARIABILE

PERCENTILI

SI PUÓ DIVIDERE L'INTERO INTERVALLO DI VALORI IN CLASSI CON LO STESSO NUMERO DI ELEMENTI
IL PERCENTILE FA SÌ CHE LE BARRE SIANO TUTTE ALTE UGUALI, CIOÈ CONTENGANO LO STESSO NUMERO DI ELEMENTI.
⚠ NON TUTTI GLI INTERVALLI SARANNO UGUALI, ANZI SARÀ OGNI CHE CAMBIANO X AVERE IN OGNI BARRA LO STESSO NUMERO DI ELEMENTI



DISTRIBUZIONE DELLA FREQUENZA

È UNA VISUALIZZAZIONE ORDINATA DI CIASCUN VALORE E DELLA SUA FREQUENZA, CIOÈ QUANTE VOLTE APPARE NEL SET DI DATI

PUÓ ESSERE ESPRESSA COME UNA PERCENTUALE.

UTILE CON VARIABILI DISCRETE

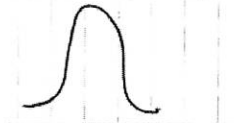
DATA UNA DISTRIBUZIONE GAUSSIANA SI DEFINISCONO ANCHE:

- **SKEWNESS**: misura della mancanza di simmetria
  misura se la distribuzione è simmetrica o no rispetto al valore medio

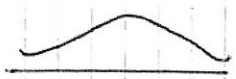
$$SK = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \bar{x})^3$$

$SK < 0$ a dx rispetto media
$SK > 0$ a sx rispetto media
$SK = 0$ gaussiana

- **KURTOSIS**: misura se la distribuzione è a picco o piana rispetto ad una normale



DISTRIBUZIONE + STRETTA
PICCO + ALTO
↑ K
K > 0 + alta



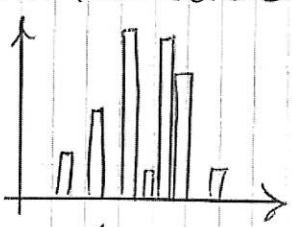
DISTRIBUZIONE + AMPIA
PICCO + BASSO
CURVA SPANCIATA
↓ K
K < 0 + piatta
RISPETTO A GAUSSIANA

$$k = \frac{1}{(n-1)s^4} \sum_{i=1}^{n} (x_i - \bar{x})^4$$

Ci sono inoltre 3 possibili rappresentazioni grafiche:

① DIAGRAMMA A **BARRE**

GLI ELEMENTI SONO SUDDIVISI IN GRUPPI E POI NE SI CALCOLANO LE FREQUENZE



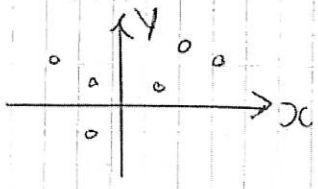
↓

IN EXCEL

- DATA ANALYSIS
- ISTOGRAMMA
- INPUT (COLONNA DI VALORI)
- BEAN (INTERVALLI 20-40-...160)
- OUTPUT (DOVE)
- LEGENDA

② DIAGRAMMA **CASUALE** (SCATTER)

CI SONO 2 VARIABILI, 1 SU ASCISSA E 1 SU ORDINATA E VENGONO INSERITI I VARI ELEMENTI COME PUNTI

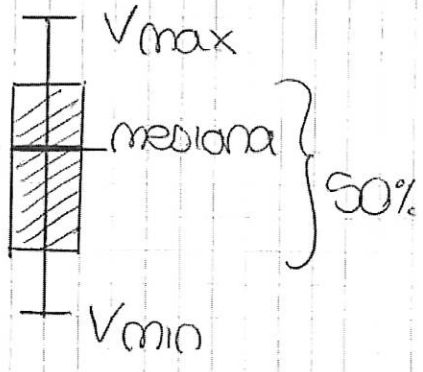


③ DIAGRAMMA A **BOX**

IL CAMPIONE VIENE DIVISO IN 4 PERCENTILI

È UN BOX CHE CONTIENE IL 50% DEI DATI

DICE SE LA POPOLAZIONE È EQUAMENTE DISTRIBUITA, GRAZIE ALLA POSIZIONE DELLA MEDIANA



VEDO LAB. 1

2° PROBLEMA che **relazione** c'è tra i parametri della popolazione e i corrispondenti parametri del campione?

OVVERO, DATO UN SET DI ELEMENTI PRESI DA UNA POPOLAZIONE, COME USARLO X STUDIARE QUESTA POPOLAZIONE?

- CON LA STATISTICA DESCRITTIVA SAPPIAMO COME È FATTO IL CAMPIONE
- POI SI PASSA ALLA STATISTICA INFERENZIALE X FARE DELLE **stime** E SAPERE COME SONO LEGATI TUTTI I PARAMETRI VISTI ↓

IL VALORE VERO DELLA POPOLAZIONE SI RAGGIUNGE SE AUMENTANO MOLTO I CAMPIONI

(QUESTO È UN PROBLEMA USUALE QUANDO SI CERCA DI ANALIZZARE I DATI ACQUISITI DURANTE UN TRIALS)

STATISTICA INFERENZIALE

SI BASA SUI CONCETTI DI:

- PROBABILITÀ [Pr(A)]: DEFINITA COME IL NUMERO DI EVENTI FAVOREVOLI DIVISO IL NUMERO TOTALE DEI POSSIBILI EVENTI
  CIOÈ SE SI VERIFICA UN EVENTO
  PUÒ ESSERE STIMATA CALCOLANDO LA FREQUENZA DEGLI EVENTI
- PROBABILITÀ CONDIZIONATA [Pr(A/B)]: PROBABILITÀ CHE SI VERIFICHI A, SE SI È GIÀ VERIFICATO B
- VARIABILI CASUALI: VARIABILI DELLE QUALI I POSSIBILI VALORI SONO I RISULTATI NUMERICI DI UN FENOMENO CASUALE
  SONO DI 2 TIPI < DISCRETE / CONTINUE

I PARAMETRI UTILIZZATI PER STUDIARE IL COMPORTAMENTO CONGIUNTO DI 2 VARIABILI SONO:

- COVARIANZA = MISURA DI COME I CAMBIAMENTI IN UNA VARIABILE CASUALE SONO ASSOCIATI CON EVENTI DI UNA SECONDA VARIABILE CASUALE
  CIOÈ È IL LIVELLO DI VARIAZIONE CONGIUNTO DI 2 VARIABILI ASSOCIATE LINEARMENTE
  $COV(x,y)$
- COEFFICIENTE DI CORRELAZIONE: $\rho = \frac{COV(x,y)}{\sigma_x \sigma_y}$

$-1 < \rho < \rho$

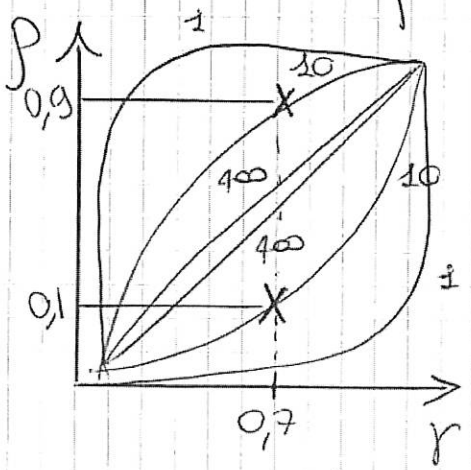
$\rho = 0$: 2 VARIABILI INDIPENDENTI / NON CORRELATE
$\rho = 1$: VARIABILI $\propto$
$\rho = -1$: VARIABILI INVERSAMENTE $\propto$

$\rho \to 0$: RELAZIONE TRA LE 2 VARIABILI MOLTO BASSA
$\rho \to \pm 1$: CORRELAZIONE ALTA! BASTA USARE SOLO + UNA VARIABILE DI 2

⚠ PERÒ AL VALORE DI QUESTA STIMA

Solitamente, x una stima veritiera bisogna avere
$\rho > 0,7 \quad \vee \quad \rho < -0,7$
se $-0,7 < \rho < 0,7$, DUBBI!! I punti sono molto asperi!
Grafica pp. 13

x analizzare ρ si usa il **DIAGRAMMA AD OCCHIO**



ci sono ≠ curve in base al numero di elementi analizzati

r è la stima della correlazione che si calcola sul campione su TOT elementi

Trovato r (ex r=0,7) lo interseco con le curve nel numero di elementi, poi parto i 2 valori a sx e vedo in che range sta ρ

$r=0,7 \rightarrow \rho \in [0,1 ; 0,9]$ x 10 elementi!
non va bene!
$r=0,9 \rightarrow \rho \in [0,6 - 0,8]$ x 100 elementi
intervallo ragionevole e significatività maggiore!
La numerosità del campione è quindi importante:
con maggiori elementi, la stima migliora!
"deduzioni" o "generalizzazioni"
Le inferenze vanno sempre documentate e ragionate, soprattutto quando si hanno pochi campioni! (ciò non vale x la statistica descrittiva)

3° PROBLEMA

BISOGNA COSTRUIRE DEI PARAMETRI / DELLE VARIABILI CHE SIANO IN GRADO DI **RAPPRESENTARE TUTTO** IL DATA SET

COME OTTENERE 1 ACCOPPIATO SET DI PARAMETRI CHE RAPPRESENTI IL SET DI DATI

X FARLO SI UTILIZZA LA **REGRESSIONE LINEARE**

= METODO DI APPRENDIMENTO STATISTICO UTILE ED AMPIAMENTE UTILIZZATO

LA REGRESSIONE LINEARE È UN SEMPLICE APPROCCIO LINEARE UTILIZZATO PER PREDIRE UNA RISPOSTA QUANTITATIVA Y SULLA BASE DI UNA SINGOLA VARIABILE X

SOLO 2 VARIABILI:

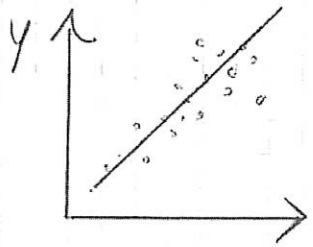
X = V. INDIPENDENTE

Y = V. DIPENDENTE

$Y = f(X)$

LA REGRESSIONE LINEARE ASSUME CHE ESISTE APPROSSIMATAMENTE UNA RELAZIONE LINEARE TRA X e Y E CHE PUÒ ESSERE RAPPRESENTATA DALLA RETTA DI REGRESSIONE:

$Y = \alpha + \beta X$ α: INTERCETTA ; β: PENDENZA



3 DIVERSI TIPI DI FITTING (LINEARE, QUADRATICO) = SONO DELLE EQUAZIONI CHE INTERPOLANO AL MEGLIO LA MIA DISPERSIONE DI PUNTI

(LA REGRESSIONE LINEARE È UN SOTTOGRUPPO DEI FITTING POSSIBILI)

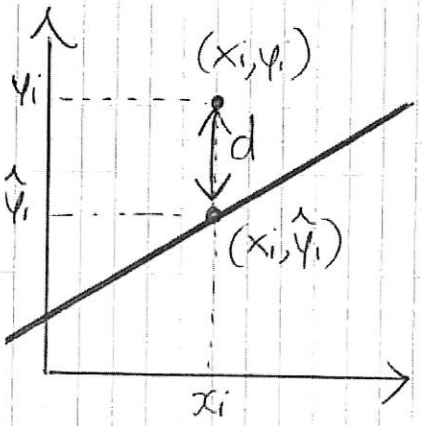
LA VARIANZA σ^2 (DISPERSIONE DEI PUNTI) È INDICE DELLA BONTÀ DELLA REGRESSIONE:

- I PUNTI DEVONO STARE LA MAGGIOR PARTE SULL'INTERPOLAZIONE E NON FUORI
- ALTRIMENTI NON È UN MODELLO OTTIMALE X I MIEI DATI (CIOÈ SE σ^2 È TROPPO ELEVATO)

X CALCOLARE I COEFFICIENTI α e β DELLA REGRESSIONE BISOGNA USARE I DATI ED IL **METODO dei MINIMI QUADRATI**.

I PUNTI $(x_1, y_1), \dots, (x_n, y_n)$ RAPPRESENTANO I DATI IN COPPIE DI n-OSSERVAZIONI, OGNUNA DELLE QUALI CONSISTE IN UNA MISURAZIONE DI X e Y.

L'OBBIETTIVO È STIMARE α e β IN MODO CHE IL MODELLO LINEARE DELLA RETTA SI ADATTI AI DATI DISPONIBILI: CIOÈ MUOVERE L'INTERCETTA α ED UNA PENDENZA β IN MODO CHE LA LINEA RISULTANTE SIA LA PIÙ VICINA POSSIBILE AI DATI.



PER OGNI PUNTO HO 2 COPPIE DI VALORI

(x_i, y_i) QUELLA REALE

$(x_i, \hat{y}_i)$ QUELLA CHE STA SULLA RETTA DELLA REGRESSIONE

$\hat{y}_i = f(x_i)$

MISURIAMO LA DISTANZA

$$d = y_i - \hat{y}_i$$

d RAPPRESENTA L'I-ESIMO **RESIDUO**, CIOÈ LA DIFFERENZA TRA IL VALORE REALE OSSERVATO E IL VALORE PREVISTO DAL MODELLO LINEARE SUPPOSTO.

SI VUOLE UNA RETTA CHE MINIMIZZA LA SOMMA (DEI QUADRATI) DI TUTTE QUESTE DISTANZE $S = \sum_{i=1}^{n} d_i^2$

OTTENENDO LE 2 FORMULE

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad ; \quad a = \bar{y} - b\bar{x}$$

$(\bar{x}, \bar{y})$ È UN PUNTO CHE STA SEMPRE SULLA RETTA

BISOGNA PERÒ VALUTARE SUCCESSIVAMENTE ANCHE LA **BONTÀ** DELLA REGRESSIONE EFFETTUATA, TRAMITE 2 VALORI:

- IL COMPONENTE DEI RESIDUI : $R_{es} = y_i - \hat{y}_i$
- IL COMPONENTE DELLA REGRESSIONE : $R_{eg} = \hat{y}_i - \bar{y}$

UNA BUONA REGRESSIONE SI HA SEMPRE SE IL VALORE ASSOLUTO DEL COMPONENTE DELLA REGRESSIONE È MAGGIORE DI QUELLA DEI RESIDUI (SOMMATORIA):

$$\|\hat{y}_i - \bar{y}\| > \sum_{i=1}^{N} \|y_i - \hat{y}_i\| \qquad \|Reg\| > \Sigma\|Res\|$$

(AMPIO) (ACCOLA)

DOPO 20, 4 ESEMPI CON FITTING

BUONO = VERDE → 1/1 → ABBIAMO BISOGNO DI $d \to 0$
MEDIO = GIALLO
MALE = ROSSO → 1/1 → SE $S = \Sigma_i d_i^2$ È GRANDE, BISOGNA TROVARE UN ALTRO METODO

AFFIDABILITÀ DELLA STIMA DEI COEFF. α E β:

LA VERA RELAZIONE TRA x E y NON È MAI CONOSCIUTA X LA POPOLAZIONE REALE.

CON LA REGRESSIONE LINEARE NON TOCCHERAI MAI TUTTI I DATI REALI, NON SI TROVERANNO MAI α E β PER UN CIO-POTRA AVVENIRE.

NUOVE DEFINIZIONI:

- DISTRIBUZIONE DI PROBABILITÀ (FUNZIONE DI PROBABILITÀ)

  x UNA VARIABILE DISCRETA (CASUALE)

  È UNA LISTA DI PROBABILITÀ ASSOCIATA CON OGNI VALORE POSSIBILE

- FUNZIONE DELLA DENSITÀ DI PROBABILITÀ

  x VARIABILI CONTINUE (CASUALE)

  È UNA FUNZIONE TALE CHE L'AREA SOTTO LA CURVA DI FUNZIONE DI DENSITÀ TRA 2 PUNTI QUALSIASI a E b DEVE UGUALE UNA PROBABILITÀ CHE x CADA TRA a E b. COSÌ, L'AREA TOTALE SOTTO LA CURVA DI FUNZIONE DI TUTTA x È = A 1.

TEST di IPOTESI

UN'IPOTESI STATISTICA È UN'ASSUNZIONE CIRCA UN PARAMETRO DI UNA POPOLAZIONE,

UN TEST DI IPOTESI È UNA PROCEDURA FORMALE USATO X ACCETTARE O RIGETTARE UN'IPOTESI STATISTICA!

CI SONO 2 TIPI DI IPOTESI STATISTICHE CHE VENGONO CONTRAPPOSTE:

H_0 : IPOTESI NULLA
SI VUOLE DIMOSTRARE
DICE CHE LE OSSERVAZIONI DEL CAMPIONE
RISULTANO PURAMENTE DAL CASO