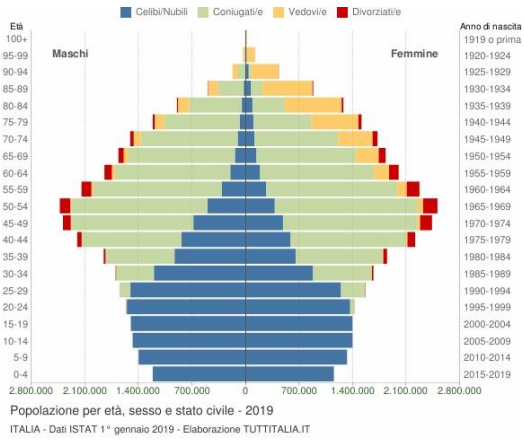
ANNO 2019-2020 PROF. LAGAZIO

STATISTICA DESCRITTIVA

Statistica: scienza che studia i metodi per elaborare i dati, metodi per raccogliere, organizzare e sintetizzare le info al fine di ottenere conoscenza.

ISTAT istituto per raccolta e info delle statistiche.

Diagramma: piramide dell'età della popolazione, tipica rappresentazione grafica per descrivere la struttura di una popolazione. Diviso a metà (maschi e femmine) ciascuna barra fa riferimento ad una classe di età

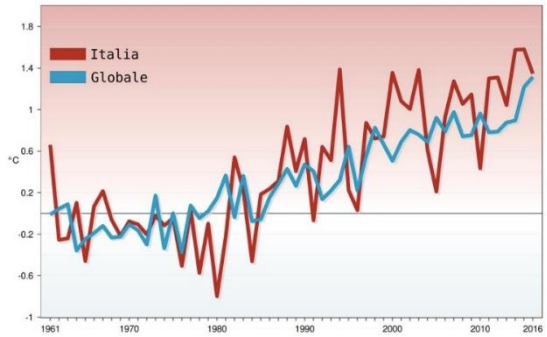


diversa.

Colpa forse della transizione demografica: passaggio da un'economia prevalentemente rurale a una prevalentemente industriale. In questo processo di industrializzazione cambia il comportamento demografico della popolazione (riduzione del numero di figli).

Piramide ora non è più, giustificata struttura della popolazione. Popolazione tende ad invecchiare e si tende a vivere più di 80 anni. Questo può avere ripercussioni circa la sostenibilità pensionistica.

Altro grafico: deviazioni medie della temperatura media di un anno rispetto a livello ritenuto normale



(anomalie)

Serie storica: insieme di informazioni ottenute nel tempo, che viene poi rappresentata su un diagramma (rappresentare un fenomeno nel tempo)

Fino a metà anni 80 anomalie oscillano intorno allo 0. Da 1985 ci sono anni più caldi e anni più freddi ma tutte le anomalie sono sopra lo 0 (caldi rispetto a valore di riferimento) le temperature progressivamente sono in aumento, tendenza a crescere. Cambiamento veloce come si vede dalla pendenza. La previsione è

basata su concetti e metodi statistici (tecnica di regressione lineare per previsioni) su questo grafico ci aspettiamo che continui ad aumentare nel futuro.

Abbiamo visto distribuzioni temporali e spaziali in comune hanno il fatto che studiano fenomeni di natura variabile (es epidemie). I fenomeni non variabili sono molto rari (principi di carattere deterministico = valori esatti in linea teorica) perché quando si osserva la realtà (es. Far cadere palla) vi sono delle piccole deviazioni. Non osserviamo mai lo stesso valore perché quando si effettua l'esperimento entrano in gioco tutta una serie di elementi che tendono ad influenzare il significato (per ottenere lo stesso tempo, nell'esempio della palla, dovremmo lasciar cadere la palla esattamente sempre alla stessa altezza). Ogni volta che si ripete l'esperimento si avrà un po' di variabilità. La maggior parte dei fenomeni che osserviamo empiricamente sono variabili e questo è tanto più vero se dal mondo della fisica ci si muove nell'ambito socioeconomico (in questo ambito vi sono gli esseri umani che sono estremamente variabili, per caratteristiche morfologiche (alti bassi) e nei comportamenti (es. Di acquisto, gusti e preferenze).

Quando le caratteristiche sono variabili c'è bisogno di statistica. (quando c'è variabilità)

ESEMPIO: dataset da indagine sui bilanci delle famiglie (Excel)

Ogni 2 anni Banchitalia seleziona un campione di famiglie e su queste rileva tutta una serie di caratteristiche (in particolare legate al bilancio familiare es. Fonti di reddito, ricchezza e come viene gestita)

CAMPIONE =8000 famiglie per capire le caratteristiche di tutti gli italiani, lo studio non viene fatto perché quelle 8000 famiglie siano interessanti, ma vengono prese in considerazione in quanto si presume siano rappresentative dell'intero universo delle famiglie italiane.

Quando si studia un fenomeno, in linea di principio lo si vuole investigare nella sua completezza, ma non sempre è possibile (prima si faceva censimento della popolazione, operazione mediante la quale viene censita, ovvero rilevata, tutta la popolazione sul territorio e rilevate tutta una serie di caratteristiche come età, tit. stud., statciv... Si può fare una volta ogni 10 anni anche perché per analizzare e pubblicare i risultati ci vogliono, ancora adesso, 3 4 anni. Operazione estremamente complessa, rilevare tutta la popolazione è oneroso in termini di tempo e denaro e non sempre è possibile). I censimenti non si fanno più in questo modo, il nuovo censimento (partito 2 anni fa) funziona in modo diverso, rilevando un sottoinsieme opportunatamente selezionato detto campione.

La statistica si può dividere in 2 grandi aree:

1. Statistica descrittiva = es. Piramide per età, andamento temperatura, diffusione epidemia, osservo un fenomeno e ne sintetizzo e descrivo le caratteristiche. La statistica descrittiva si usa sempre (sintetizzare informazioni, indifferente campione o popolazione, però poi è differente avere i dati completi o su campione; se si hanno i dati su tutta la popolazione la descrizione, ovvero gli strumenti della statistica descrittiva, mettono a disposizione tutti gli elementi per estrarre informazioni dai calcoli; se invece si ha un campione e vogliamo usare questo per capire le caratteristiche di un insieme più grande, dal quale questo campione è stato selezionato, occorre generalizzare i risultati).
2. Inferenza= passare dal particolare al generale, con l'operazione di inferenza si è soggetti ad errore dovuto fondamentalmente al campione, che potrebbe non essere una buona immagine della popolazione, si ipotizza che il campione sia rappresentativo della popolazione (i soggetti intervistati si spera abbiano un comportamento simile a quello di tutta la popolazione). Non si hanno garanzie su questo, i campioni rappresentativi non esistono. L'inferenza statistica è l'insieme di metodi e di tecniche che consentono di gestire l'errore dovuto al campionamento.

CONTINUO ESEMPIO 1 DA EXCEL: tipica rappresentazione di dataset

Caratteristiche generali:

- Colonne: per individuare le diverse caratteristiche, le caratteristiche si chiamano variabili (si hanno diversi valori in ciascuna colonna)
- Righe: si riferiscono ciascuna a una unità statistica diversa, i soggetti (nel es. le unità statistiche sono le famiglie) sui quali si rileva il fenomeno di interesse sono chiamati unità statistiche. Es. Famiglie, individui, entità inanimate…

Alcune delle variabili dell'esempio si riferiscono al capofamiglia (es. genere, Stat civ). Ogni colonna è una variabile, le variabili sono di natura diversa.

Dall'esempio: nomenclatura

n quest= nome famiglia

genere= questa variabile può assumere due valori diversi, i valori che la variabile può assumere si chiamano modalità. Il genere è una variabile che può assumere due modalità: maschio (1) e femmina (2)

Stat civ= variabile con 4 modalità celibe/nubile, coniugato, separato/divorziato, vedovo

Regione= variabile con 20 modalità

N comp= variabile che almeno idealmente assume valori all'interno dei numeri naturali (no limitata superiormente)

Età= misurata in anni compiuti

Y (reddito) = variabile che assume n valori

Le variabili vengono chiamate con X, Y, Z

X= MAIUSCOLO INDICA LA VARIABILE

x1, x2, x3, …, xi… xn = successione di tutti i valori osservati (valori che la variabile assume), rappresentazione di tutta la tabella di una colonna

ESEMPIO:

X=genere

x1=genere capofamiglia della famiglia 1

x2=genere capofamiglia della famiglia 2

xi= genere capofamiglia della famiglia i

n= indica il numero totale delle famiglie presenti nel collettivo studiato

Collettivo= insieme di unità statistiche (esaustivo o campione è irrilevante quando si fa statistica descrittiva)

Le variabili hanno natura diversa, genere 2 modalità mentre il tit studio può assumere 4 modalità, il reddito valori nel continuo, età, n comp, percettori di reddito sono conteggi. La natura delle variabili è molto importante perché a seconda del tipo di variabile che si prende in considerazione, cambia il modo di gestirla.

Una prima distinzione che si può fare è distinguere tra variabili:

- Qualitative

- Quantitative

Immaginare quale domanda fare per rilevare il valore della variabile:

-reddito si chiederà quanto guadagna il soggetto, altezza quanto sei alto (quanto, si vuole rilevare una quantità, caratteristiche di tipo quantitativo)

- titolo di studio si chiederà quale tit studio, genere quale genere (quale, si vuole rilevare una qualità, caratteristiche di tipo qualitativo)

Distinzione in base alle scale di misura:

Prendiamo in considerazione 2 variabili qualitative es. Stat civ e tit studio, entrambe 4 modalità ma differenza sostanziale: la variabile stato civile ha 4 modalità che non hanno ordinamento intrinseco (es anche genere);tit studio ha anch'esso 4 modalità, ma che possono essere ordinate (ha senso chiedersi quale di questi se un soggetto ha un tit di stud maggiore di un altro, confronto in termini di maggiore e minore.

Si hanno due scale di misure diverse

Per le variabili qualitative:

- Scala di misura nominale (o sconnessa) = relativa a Stat civ, genere, gruppo sanguigno… condizione professionale (questa variabile può avere diverse classificazioni più o meno articolate)
- Scala di misura ordinale = relativa a titstud per es. si può fare ordinamento e si possono fare confronti in termini di maggiore o minore.

In realtà, in alcuni casi il confine tra le variabili misurate su scala nominale e le variabili misurate su scala ordinale è molto labile. Es: voti, scale di misure diverse per esprimere esito di un esame trentesimi, centesimi, decimi, alfabetico. I voti, anche se da noi per esempio vengono trattati come quantità, non sono esattamente variabili quantitative e stanno in mezzo fra le variabili ordinabili e le variabili quantitative (diverse scale di misura). Tipicamente i voti vengono trattati come variabili quantitative (in Italia si calcola la media dei voti, che è proprio l'indice che viene utilizzato per trattare variabili quantitative)

Per le variabili quantitative:

- Scala di intervallo = molto rare, solo 1 importante per noi. Le variabili misurate su scala di intervallo sono quelle nelle quali lo 0 è fissato su base convenzionale. Due esempi di scala di intervallo: temperatura e tempo. La temperatura può essere misurata in °C, °F o °K lo 0 è diverso, la conseguenza è che non si possono prendere due valori e confrontarli fra rapporti. Esempio: oggi ci sono 10° ieri 5° non possiamo dire che oggi è il doppio più caldo di ieri, dire il doppio vuol dire fare il rapporto fra due valori. Se lo 0 è arbitrario il rapporto tra due valori non ha alcun senso (cambiando lo 0 cambia anche il rapporto) si possono confrontare intervalli di valore. Tempo viene misurato fissando l'origine (nostro calendario 2020anni fa anno 0) e misurando da quell'istante, esempio 20 febbraio non si dice mica essere il doppio del 10 febbraio (non ha nessun senso perché questo confronto dipende dall'origine, dallo 0),si possono fare confronti: se si contano i giorni si può fare per rapporto, due settimane è il doppio di una settimana, per dire ciò si prende un intervallo di tempo che intercorre fra due date, non la data, e lo si confronta con un altro intervallo di tempo compreso tra due date. Il tempo è una variabile molto importante in statistica, la serie storica (studio dell'evoluzione di fenomeni nel tempo) è una delle cose più importanti in statistica.
- Scale di rapporto = molto più frequenti, variabili dove lo 0 ha un suo significato, esempio reddito=0 significa proprio che quella famiglia, in quel particolare anno non ha avuto reddito, numero di percettori di reddito è pari a 0 all'interno della famiglia, significa che la famiglia non ha fonti di sostentamento (nessuno all'interno della famiglia percepisce reddito) .Se una variabile è misurata

su scala di rapporto i confronti possono essere fatti direttamente sul rapporto, una famiglia ha un Reddito doppio rispetto a un'altra, numero di componenti è doppio rispetto a un'altra (rapporto direttamente tra due valori) .

Ulteriore distinzione sulle variabili quantitative:

- Discrete = n comp, n percettori di reddito…, assumono valori nel discreto (0,1,2 lo 0 può essere incluso o meno). Una variabile è discreta quando assume un numero finito di valori o un'infinità numerabile. (es. n componenti è una variabile di conteggio, lo 0 non è incluso e teoricamente può assumere valori non limitati superiormente).
- Continue=età è una variabile misurata su scala continua, ogni secondo si è ogni secondo più vecchi (ma poi la rileviamo in anni compiuti). L'altezza ad es. con strumenti sofisticati potrebbe essere analizzata nel continuo, ma per le finalità pratiche e gli strumenti che abbiamo, l'altezza la misuriamo in centimetri. (almeno idealmente vi sono variabili che assumono valori nel continuo all'interno di un intervallo).

Questa distinzione è molto importante perché a seconda della natura della variabile, cambia il modo in cui la variabile viene analizzata.

LOGICHE DI ANALISI

La statistica descrittiva si preoccupa di descrivere ciò che si osserva e la descrizione di ciò che si osserva passa attraverso la sintesi. La quantità di informazioni che si raccoglie è estremamente articolata e complessa, poi si sintetizza. Difficile effettuare un'analisi, occorre sintetizzare con gli strumenti a disposizione. All'inizio gli strumenti vengono utilizzati in modo uni variato, ovvero analisi di una variabile per volta (analisi uni variata)

Quali sono gli strumenti per fare sintesi:

-costruzione di tabelle (comprimere l'informazione osservata in una tabella)

-grafici (analogo alla forma tabellare, sintesi dell'informazione)

-calcolo di opportuni indici (es. Media)

TABELLE: distribuzioni di frequenza

Prendiamo in considerazione la variabile più semplice ovvero il genere. Come sintetizzare l'informazione contenuta nella colonna della tabella? Si possono contare le famiglie con capofamiglia maschio e le famiglie con capofamiglia femmina. I valori calcolati vengono chiamate frequenze assolute.

Frequenza assoluta: valore che corrisponde al numero di volte che è stato osservato un certo valore della variabile. La frequenza assoluta è associata alla modalità, prendo la modalità maschio e conto quante volte la modalità si è presentata nel collettivo studiato. Questo può essere effettuato anche per tutte le altre variabili, es. Stat civ, regione...Per età e reddito in realtà questo processo non è particolarmente efficiente perché l'obiettivo della costruzione della distribuzione di frequenza è quello di sintetizzare l'informazione, non è facile trovare gli stessi identici valori, è quasi impossibile.

<u>Frequenze assolute</u>: ni (il genere per esempio ha 2 frequenze assolute maschio o femmina)

Caratteristica frequenze assolute: sommatoria che va da i=1 a c di ni è uguale a n.

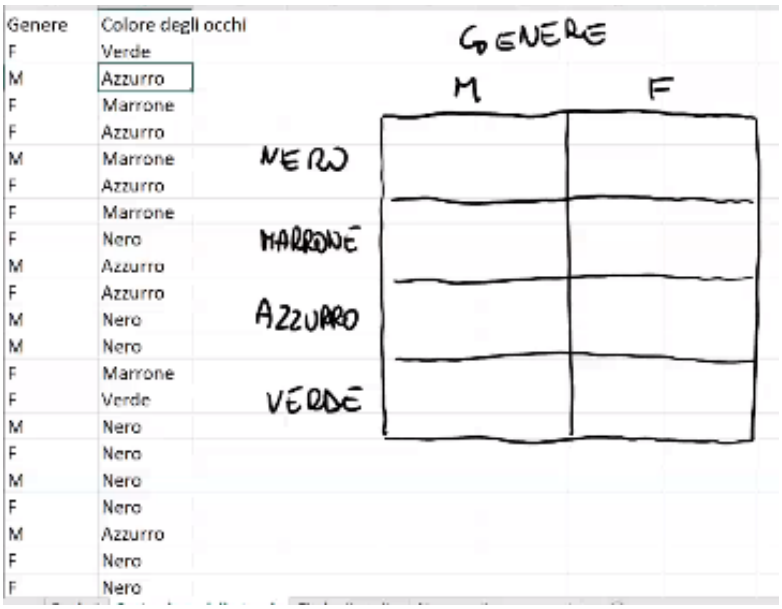
(c= numero di modalità della variabile (genere c=2); n= dimensione del collettivo)

<u>Frequenze relative</u>: fi = ni/n

STATISTICA MULTIVARIATA:

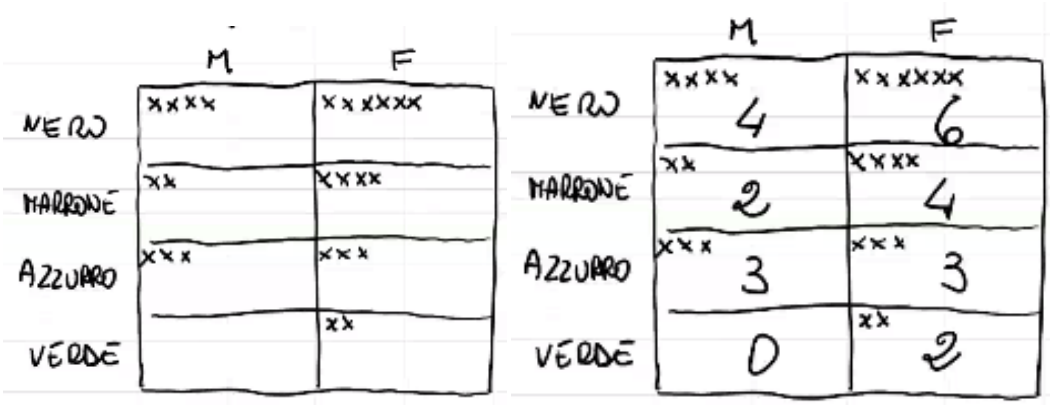
(aulaweb excel con esempi). Esempio genere e colore occhi (excel)

Primo strumento per fare sintesi è la distribuzione di frequenza (come per la statistica univariata, dove veniva presa singolarmente variabile per variabile) in questo caso è possibile costruire una distribuzione di frequenza che considera congiuntamente più variabili. Obiettivo finale di questo processo di sintesi è quello di capire se fra le variabili prese in considerazione vi è un legame (se vanno nella stessa direzione, esempio colore occhi e genere, femmine hanno colore tendenzialmente più scuro maschi o è vero il contrario o nè uno né l'altro). L'obiettivo è valutare se esiste una qualche forma di relazione (associazione). Si hanno 2 o più variabili prese in considerazione e valutate congiuntamente. Si osserva un insieme di unità statistiche, nell'es abbiamo due variabili. Si può costruire una tabella a doppia entrata (distribuzione doppia di frequenza).



Genere	Colore degli occhi
F	Verde
M	Azzurro
F	Marrone
F	Azzurro
M	Marrone
F	Azzurro
F	Marrone
F	Nero
M	Azzurro
F	Azzurro
M	Nero
M	Nero
F	Marrone
F	Verde
M	Nero
F	Nero
M	Nero
F	Nero
M	Azzurro
F	Nero
F	Nero

Si riportano poi le osservazioni e le si collocano nella tabella (questa operazione prende il nome di "Spoglio")



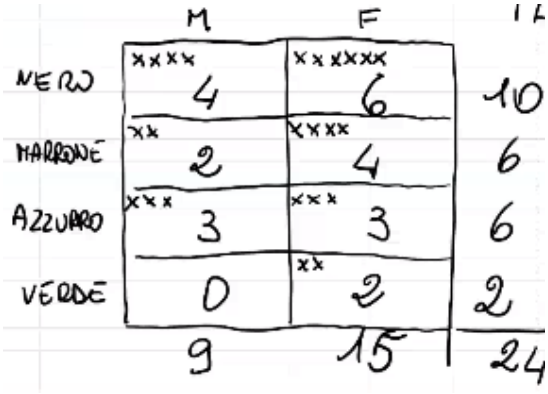
Viene definita come "Tavola o tabella di contingenza"

I valori all'interno della tavola indicano quante volte si sono presentate congiuntamente determinate caratteristiche (es 4 significa che di tutti i soggetti presi in considerazioni vi sono 4 soggetti che sono di genere maschile E hanno occhi neri, osserviamo congiuntamente le due variabili). Queste frequenze

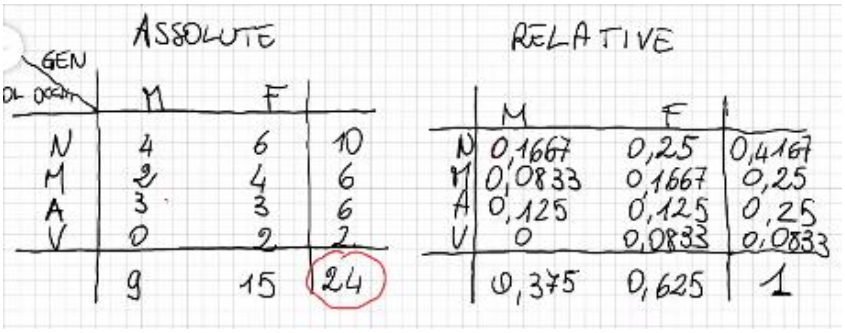
vengono chiamate FREQUENZE CONGIUNTE. Se si considerano tutte insieme le frequenze si ha la distribuzione di frequenza congiunta del genere e del colore degli occhi per il collettivo studiato:

Caratteristiche frequenze congiunte:

1. Se si sommano si ottiene la numerosità complessiva del collettivo
2. Marginalizzazione della tabella: se si sommano le frequenze congiunte per riga si va al margine. Nel momento in cui si marginalia si trascura uno dei due caratteri (vengono sommati soggetti maschili e femminili es. con colore nero, 10 soggetti hanno colore nero ma trascuriamo genere). Queste somme prendono il nome di frequenze Marginali (sono le frequenze che si otterrebbero in analisi uni variata) Si possono costruire due distribuzioni di frequenze marginali.



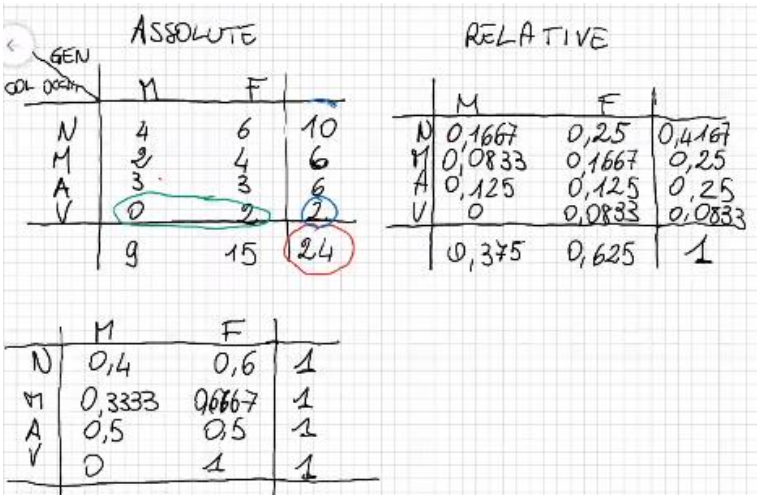
Si può costruire una tabella di contingenza delle frequenze relative (si ottengono dividendo ogni frequenza assoluta per la numerosità, nell'esempio 24)



Se sommiamo le frequenze congiunte otteniamo le frequenze marginali

Se si sommano tutte le frequenze congiunte si ottiene 1, se si sommano tutte le frequenze marginali si ottiene 1.

Se si osservano e analizzano le singole colonne o le singole righe della tabella. Si possono così calcolare le percentuali di riga (frequenza congiunta/frequenza marginale corrispondente) (la distribuzione marginale relativa si ottiene dividendo le frequenze assolute marginali per il totale e corrispondono alle distribuzioni delle frequenze relative semplici dei due caratteri)



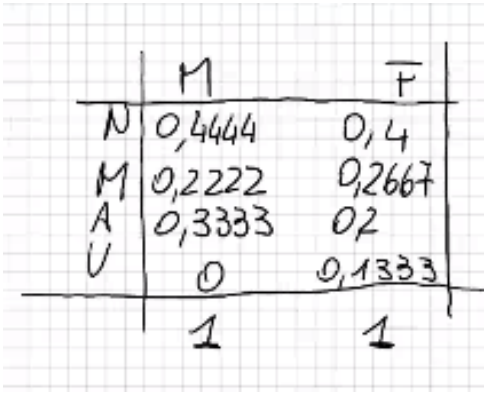
Restringere attenzione a una determinata caratteristica, questa operazione è detta CONDIZIONAMENTO (si considerano i soggetti con occhi neri e vedere quanti sono maschi e quante femmine)

Condizionamento: distribuzione di una variabile quando prediamo in considerazione un sottoinsieme delle unità statistiche che presentano una determinata variabile.

Si possono costruire tante distribuzioni condizionate quante sono le modalità della variabile (4+2 distribuzioni condizionate nel nostro esempio)

NB: le distribuzioni condizionate si possono calcolare solo in termini relativi, le frequenze assolute non sono confrontabili

Condizionamento per genere:



ESEMPIO

Mobilità del titolo di studio
fonte petretto pignataro Economia del capitale umano

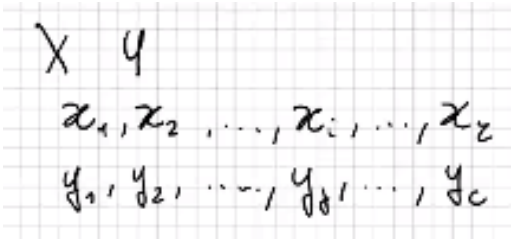
padre/figlio	Elem	Media inf	Media sup	Laurea	Totale
Elem	86	498	371	47	1002
Media inf	3	80	174	45	302
Media sup	2	6	92	48	148
Laurea	0	5	23	35	63
Totale	91	589	660	175	1515

ESEMPIO

Numero di componenti per genere

Num Comp	M	F	Totale
1	784	1407	2191
2	1616	929	2545
3	942	624	1566
4	816	536	1352
5	299	198	497
Totale	4457	3694	8151

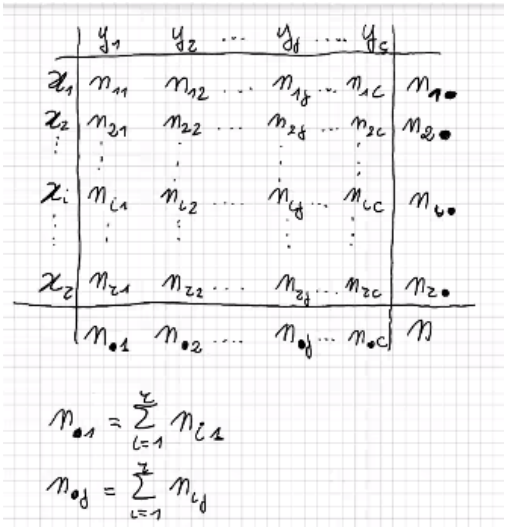
NB: QUESTE TABELLE POSSONO ESSERE COSTRUITE ANCHE SE LE VARIABILI SONO DI TIPO QUALITATIVO O SUDDIVISE IN CLASSI

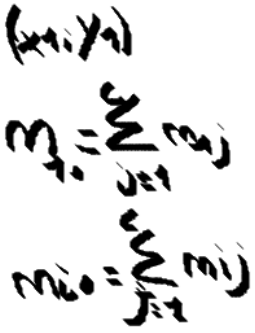


R = RIGHE (r sarà pari al numero di modalità della variabile)

C= COLONNE (c sarà pari al numero di modalità della variabile)

N =frequenze congiunte, 2 pedici si riferisce alla coppia di valori (es.x1y1 n11…)

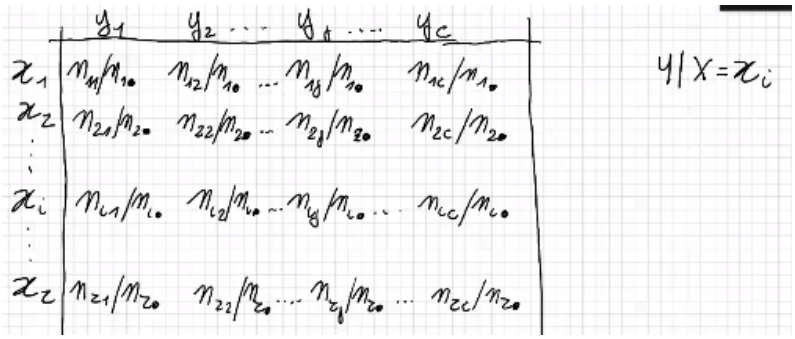




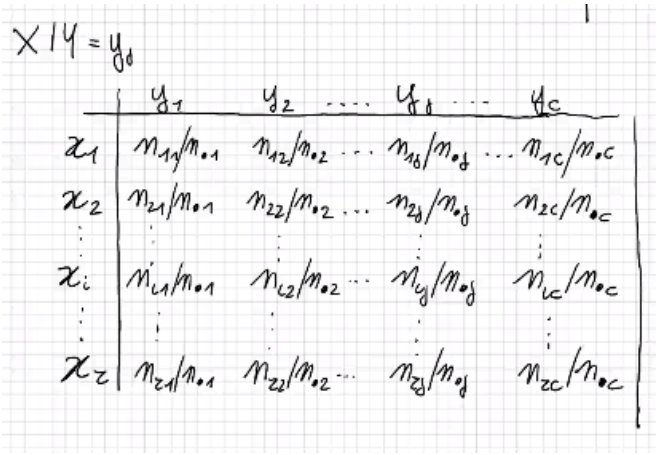
Frequenze marginali mantengono un pedice e omettono un altro (prima riga somma per j che va da 1 a c di n1j perciò n1.)

Per costruire le frequenze marginali: cndiionando pre riga (tutte le frequenze congiunte prim riga/ frequenza margianle)

Distribuzioni condizionate della variabile y dato un certo valore della variabile x (Y / X = xi , Y condizionato a X)



Distribuzione condizionata della variabile X dato un certo valore della variabile Y



Tavole di contingenza: per ciascuna coppia di modalità rileviamo quante volte questa coppia si è verificata, le frequenze sono chiamate frequenze congiunte (assolute o relative). Ai margini della tavola si possono costruire per somma (riga o colonna) le frequenze marginali (per riga della variabile X, per colonna della variabile Y) dicono come è distribuita una variabile quando non si tiene conto dell'altra. Distribuzione condizionata: si sceglie una modalità dell'altra variabile e si considerano solo le unità statistiche che presentano questa modalità. Le distribuzioni condizionate sono solitamente relative, abbiamo distribuzioni condizionati del carattere Y dato un certo valore del carattere X (e viceversa)

Num Comp	M	F	Totale
1	784	1407	2191
2	1616	929	2545
3	942	624	1566
4	816	536	1352
5	299	198	497
Totale	4457	3694	8151

distribuzione congiunta assoluta (riga e colonna totale riportano le distribuzioni marginali)

Un carattere di tipo quantitativo e uno di carattere qualitativo, in caso di caratteri quantitativi si possono calcolare gli indici di posizione (media), di variabilità (varianza).

Misure di posizioni: si può calcolare il numero medio di componenti marginale (non tenendo conto del genere) Numero di componenti = X

<u>MEDIA MARGINALE</u>

$$\bar{x} = \frac{1}{n} \sum x_i n_{i\bullet} = \text{MEDIA MARGINALE}$$

$$= \frac{1}{8151}(1 \times 2191 + 2 \times 2545 + \ldots + 5 \times 497) = 2{,}44$$

Si può ragionare anche in maniera condizionata, esempio prendiamo solo le famiglie con capofamiglia maschio (sottoinsieme di tutte le osservazioni) , calcola la media condizionata

MEDIA CONDIZIONATA

$$\bar{x}_{y=y_j} = \bar{x}_j = \sum_i x_i \left(\frac{n_{ij}}{n_{\bullet j}}\right) = \frac{1}{n_{\bullet j}} \sum x_i n_{ij}$$

FREQ COND

Nel nostro caso si possono calcolare due medie condizionate, possiamo calcolare tante medie condizionate quante sono le distribuzioni condizionate. (distribuzione condizionata n componenti con capofamiglia maschio e la distribuzione condizionata e n componenti con capofamiglia femmina)

$$\bar{x}_{y=M} = \frac{1}{4457}(1 \times 784 + 2 \times 1616 + \ldots + 5 \times 299) =$$

$$= 2{,}60$$

leggermente più alto rispetto a media marginale, sono di più le famiglie con capofamiglia maschio

$$\bar{x}_{y=F} = \frac{1}{3694}(1 \times 1407 + 2 \times 929 + \ldots + 5 \times 198) =$$

$$= 2{,}24$$

le famiglie con femmina tendono ad essere leggermente più piccole.

Proprietà medie condizionate:

la media delle medie condizionate è pari alla media marginale (questo è vero sempre)

PROPRIETÀ DELLE MEDIE CONDIZIONATE

$$\bar{x} = \frac{1}{n} \sum \bar{x}_j n_{\bullet j}$$

Se si applica la formula nell'esempio

$$\frac{1}{n} \sum \bar{x}_j n_{\bullet j} = \frac{1}{8151}(2{,}60 \times 4457 + 2{,}24 \times 3694) =$$

$$= 2{,}44$$

Si può calcolare la VARIANZA MARGINALE (O TOTALE) per le due variabili.

$$\sigma_x^2 = \frac{1}{n} \sum x_i^2 n_{i\bullet} - \bar{x}^2$$

VARIANZA MARGINALE (TOTALE)

Nell'esempio:

$$= \frac{1}{8151}\left(1^2 \times 2191 + 2^2 \times 2545 + \ldots + 5^2 \times 497\right) - 2{,}44^2$$
$$= 1{,}4813$$

Si può calcolare la VARIANZA CONDIZIONATA (esprime la variabilità intorno alla propria media delle unità della distribuzione condizionata

$$\sigma^2_{X|Y=y_j} = \frac{1}{n_{\cdot j}} \sum x_i^2 n_{ij} - \bar{x}^2_{Y=y_j}$$

Possiamo calcolare due varianze condizionate:

$$\sigma^2_{X|Y=M} = \frac{1}{4457}\left(1^2 \times 784 + 2^2 \times 1616 + \ldots + 5^2 \times 299\right) - 2{,}60^2 =$$
$$= 1{,}3599$$

$$\sigma^2_{X|Y=F} = \frac{1}{3694}\left(1^2 \times 1407 + 2^2 \times 929 + \ldots + 5^2 \times 198\right) - 2{,}24^2 =$$
$$= 1{,}5555$$

La varianza totale è maggiore o uguale (solo in caso molto particolare) alla media delle varianze condizionate (si perde quota di variabilità). Questo perché andiamo con la varianza condizionata si va a vedere la variabilità all'interno di ciascun gruppo senza tener conto che i gruppi possono essere diversi fa loro.

$$\text{VAR TOT} \geq \text{MEDIA VAR CONDIZIONATE}$$

NB: dati due caratteri X e Y, quantitativi, si può sintetizzare la distribuzione doppia mediante il punto di coordinate (x medio; y medio) chiamato punto medio o baricentro della distribuzione.

Cosa dicono in più le distribuzioni congiunte rispetto a quelle marginali? Il motivo per cui si ragiona congiuntamente su più caratteri, perché si cerca la relazione che lega le due variabili, valutare se siano dipendenti o indipendenti tra loro. (esempio colore occhi e colore capelli tendono ad essere associati)

RELAZIONE DI INTERDIPENDENZA = le variabili sono legate tra loro, si assume che i caratteri abbiano tutti lo stesso ruolo e che i legami tra essi siano bidirezionali (si influenzano reciprocamente)

Esempio: reddito famiglia e rilevazione dei diversi tipi di consumo (es. spesa per vacanze) si osserva che queste due variabili sono tra loro legate, in questo caso si parla di RELAZIONE DIPENDENZA DELLE DUE VARIABILI (come le modalità di un carattere "dipendono" da quelle di un altro carattere secondo un legame unidirezionale, relazione causa-effetto, alcune variabili influenzano altre, es. maggior disponibilità di reddito permette disponibilità di vacanze)

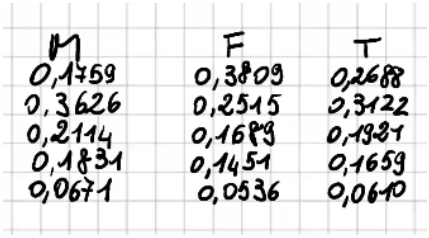
NB: un conto è che le variabili abbiamo comportamenti simili (associazione di tipo statistico), un conto è sovrapporre a questo legame una relazione di causa effetto (questo tanto più vero nelle scienze sociali). Il fatto che si sia una coincidenza/associazione di tipo statistico non significa che fra le due variabili ci sia una

relazione di causa effetto (quantità gelato venduto e numero di interventi dei bagnini, il legame tra le due variabili è dato da un andamento stagionale).

Esempio: nel caso di un esperimento che si misura l'efficacia sui soggetti, il legame statistico può avere interpretazione di tipo causale, perché vi è relazione coerente causa-effetto.

Esempio: se si prende il reddito e i consumi di tipo culturale (libri, mostre) di una famiglia si osserva che queste variabili sono legate tra loro, ma in parte la relazione non è dovuta a un rapporto diretto causa effetto. Potrebbe essere che il legame sia il manifestarsi degli effetti del grado di istruzione.

Num Comp	M	F	Totale
1	784	1407	2191
2	1616	929	2545
3	942	624	1566
4	816	536	1352
5	299	198	497
Totale	4457	3694	8151



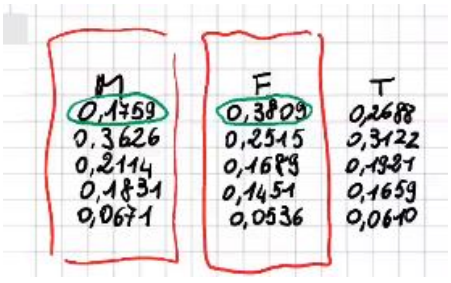
Per ragionare di questi legami bisogna vedere le distribuzioni condizionate (Distribuzione n componenti dato il genere)

C'è un legame tra le due variabili oppure no? Mentre per le famiglie con capofamiglia maschio la % di famiglie con componenti pari a 1 è meno del 20% per le femmine è quasi del 40%. La differenza osservata in questo modo è abbastanza importante e abbastanza ragionevole, la donna tende ad essere individuata come capofamiglia soprattutto quando lei è la famiglia (40% dei casi la donna è l'unico componente) per i maschi la percentuale è molto piccola. Tra le due variabili c'è una relazione, questa relazione si è percepita guardando le distribuzioni condizionate.

Situazione di indipendenza: quando la distribuzione condizionata per i maschi è uguale alla distribuzione condizionata per le femmine.

Diverse misure di associazione a seconda della natura delle variabili. Situazione generale, che va bene per tutte le variabili: dipendenza.

Quando le variabili tendono ad avere un comportamento congiunto, es. genere e n componenti, si è visto che le variabili sono legate tra loro, ciò viene individuato dalle distribuzioni condizionate. (Le famiglie con capofamiglia maschio tendono ad essere un po' più grandi delle famiglie con capofamiglia femmina)
Frequenza modale si vede dalla distribuzione condizionata.

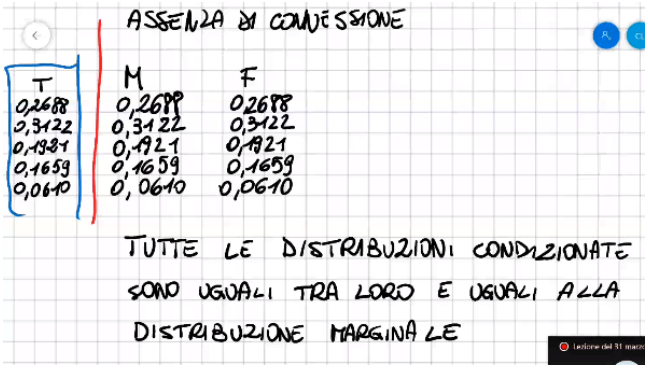


CONNESSIONE

Si può ragionare di dipendenza quando una delle due variabili fornisce informazioni sul valore dell'altra: se si sa che il capofamiglia è femmina ci si immagina una dimensione inferiore della famiglia, al contrario si tende a pensare a una dimensione familiare maggiore. Quando tra le variabili c'è un legame tale per cui conoscere il valore di una determina il valore dell'altra si dice ci sia ASSOCIAZIONE tra le due variabili→ le variabili sono CONNESSE
Situazione nella quale tra il genere e il numero di componenti non vi è relazione, è necessario che le distribuzioni condizionate e marginali siano uguali tra loro → ASSENZA DI CONNESSIONE

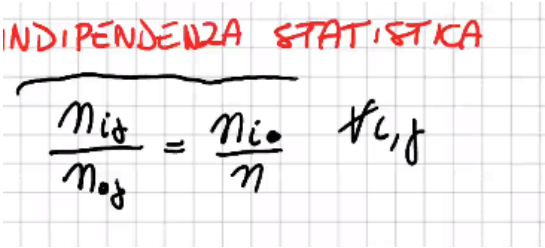
conoscere il valore di una delle due variabili non è informativo sul valore dell'altra. (conoscere genere non dice nulla sul numero di componenti)

La situazione di assenza di connessione è quella nella quale le distribuzioni condizionate sono tutte uguali tra loro e uguali alla distribuzione marginale



INDIPENDENZA STATISTICA: il numero di componenti è statisticamente indipendente dal genere

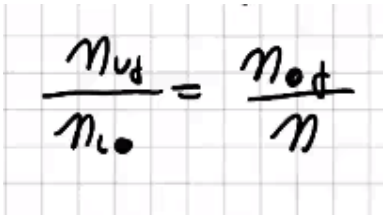
RELAZIONE DI INDIPENDENZA DI X DA Y: X è indipendente da Y se, qualunque si ala modalità con cui si manifesta il carattere Y, la distribuzione relativa condizionata di X non cambia.



X non dipende dal carattere Y (nell'esempio X numero di componenti, Y genere)

Le frequenze relative delle distribuzioni condizionate della X rispetto alla variabile Y devono essere tutte uguali fra loro e uguali alla distribuzione marginale relativa della X

RELAZIONE DI INDIPENDENZA STATISTICA DI Y DA X



SE X E' INDIPENDENTE DA Y E' VERO ANCHE IL CONTRARIO (scambio di numeratore e denominatore)

Se si procede per riga si costruiscono le distribuzioni condizionate del genere

	A	B	C	D
2				
3	Num Comp	M	F	Totale
4	1	784	1407	2191
5	2	1616	929	2545
6	3	942	624	1566
7	4	816	536	1352
8	5	299	198	497
9	Totale	4457	3694	8151

N COMP	M	F
1	0,3578	0,6422
2	0,6350	0,3650
3	0,6015	0,3985
4	0,6036	0,3964
5	0,6016	0,3984
TOT	0,5468	0,4532

Se la famiglia ha un numero di componenti superiore a 1, si tende a immaginare che il capofamiglia sia maschio. I caratteri sono interdipendenti, ovvero il legame è bidirezionale.

Misurare quanto è forte il legame tra le due variabili:

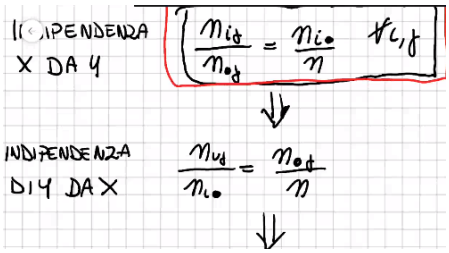
N COMP	M	F
1	0,3578	0,6422
2	0,6350	0,3650
3	0,6015	0,3985
4	0,6036	0,3964
5	0,6016	0,3984
TOT	0,5468	0,4532

M	F
0	1
1	0
1	0
1	0
1	0

quando si conosce il numero di componenti si può dire con certezza il genere del capofamiglia, apporto informativo è molto più elevato

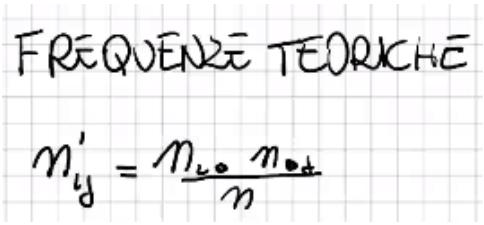
Misura di questa relazione passa attraverso una relazione tra le frequenze che deriva dalla relazione di indipendenza

CONDIZIONE DI INDIPENDENZA le congiunte sono uguali al prodotto delle marginali diviso la numerosità

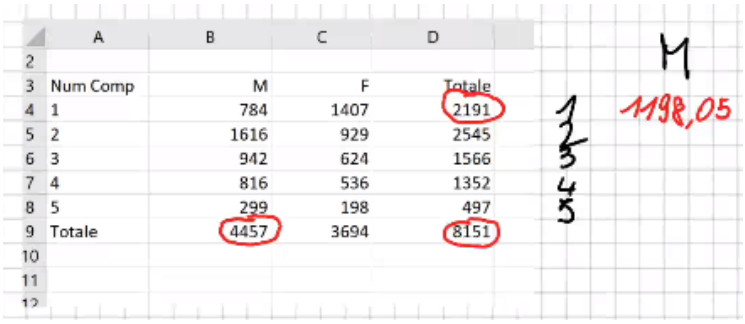


$$n_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

FREQUENZE TEORICHE: in caso di indipendenza, moltiplico la frequenza marginale di riga per la frequenza marginale di colonna e divido il prodotto per la numerosità

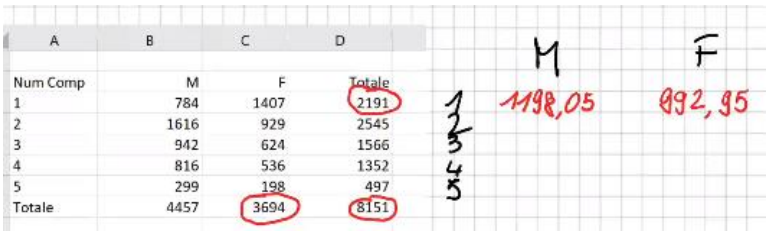


La frequenza teorica è qual valore che si deve avere in caso di indipendenza



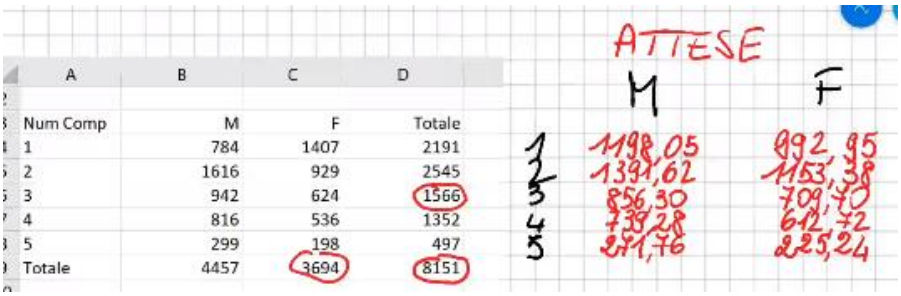
	A	B	C	D
2				
3	Num Comp	M	F	Totale
4	1	784	1407	2191
5	2	1616	929	2545
6	3	942	624	1566
7	4	816	536	1352
8	5	299	198	497
9	Totale	4457	3694	8151
10				
11				
12				

M
1198,05



A	B	C	D
Num Comp	M	F	Totale
1	784	1407	2191
2	1616	929	2545
3	942	624	1566
4	816	536	1352
5	299	198	497
Totale	4457	3694	8151

M F
1198,05 992,95



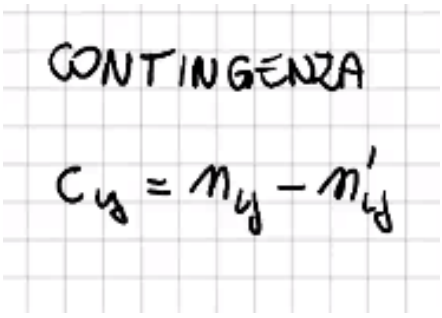
A	B	C	D
Num Comp	M	F	Totale
1	784	1407	2191
2	1616	929	2545
3	942	624	1566
4	816	536	1352
5	299	198	497
Totale	4457	3694	8151

ATTESE

	M	F
1	1198,05	992,95
2	1391,62	1153,38
3	856,30	709,70
4	739,28	612,72
5	271,76	225,24

Le attese non sono uguali a quelle osservate, capire quanto si è lontani dalla indipendenza.

Situazione di contingenza: misura la forza della relazione in termini di scostamento dal valore della frequenza teorica.

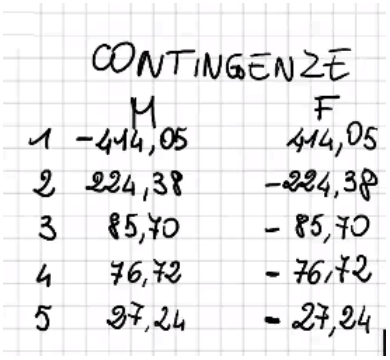


CONTINGENZA

$$C_{ij} = n_{ij} - n'_{ij}$$

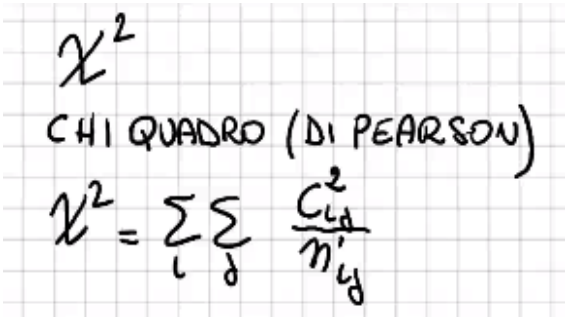
Se c'è indipendenza le frequenze osservate e attese devono essere tutte uguali a 0 , se non lo sono si è in una situazione di dipendenza

Si può calcolare la tavola delle contingenze: quanto le frequenze osservate sono diverse da quelle attese



Le contingenze sommano a 0 per riga e colonna (se abbiano due colonne cambio segno)

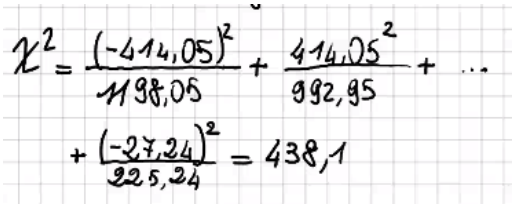
Grado di associazione delle variabili: sintesi delle contingenze: UNO DEGLI INDICI PIU' IMPORTANTI DELLA STATISTICA (<u>per caratteri qualitativi sconnessi</u>)



SOMMA DELLE CONTINGENZE AL QUADRATO DIVISO LE CORRISPONDENTI FREQUENZE TEORICHE

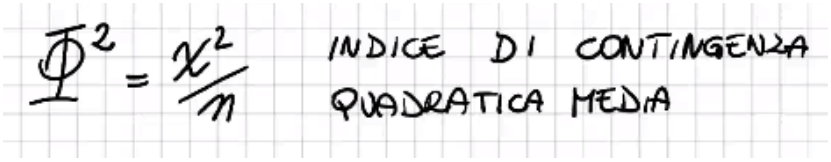
Questo indice è sempre maggiore di 0, è pari a 0 solo se i due caratteri sono perfettamente indipendenti (tutte le frequenze osservate sono uguali a quelle teoriche e le contingenze sono quindi pari a 0). L'indice assumerà valori tanto più grandi quanto più le frequenze osservate si differenziano da quelle teoriche.

Nell'esempio:



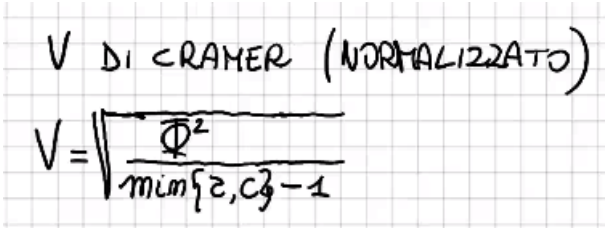
VALORE MASSIMO: dipende dalle dimensioni della tabella e dalla numerosità totale del collettivo studiato (non vi è un criterio univoco) per questo il Chi quadro viene normalizzato in due passaggi diversi:

Primo indice: (non dipende dalla numerosità totale)



(valore minimo pari a 0 il caso di indipendenza, valore massimo pari a 1, solo se il numero di righe o colonne è pari a 2, altrimenti l'indice è maggiore di 1) Per questo l'indice viene normalizzato:

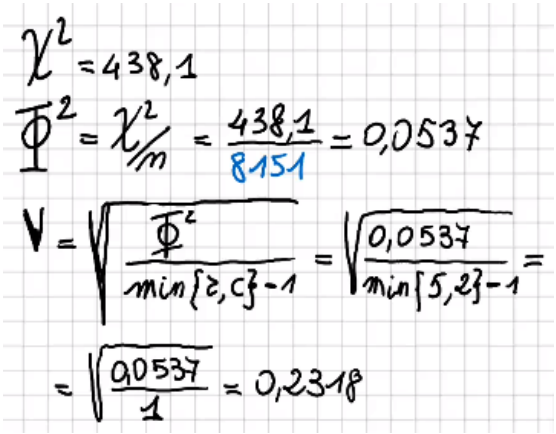
Secondo indice:



R e c sono numero di righe e numero di colonne

Al denominatore si ha il minore fra il numero di righe e il numero di colonne. Assume il valore minimo 0 quando si è in una situazione di indipendenza statistica e valore massimo pari a 1 quando si è in una situazione di massima interdipendenza o connessione tra le due variabili.

Dall'esempio:



Esercizio n. 2 Turno A del 18/04/2019

Una compagnia assicurativa sta confrontando il numero di polizze che quattro agenzie hanno stipulato nell'arco di un mese per le tre principali *Prestazioni Assicurative*. Si è potuta così costruire la seguente tabella a doppia entrata:

	Prestazioni assicurative			
Agenzia	Auto	Vita	Casa	Totale
1	22	28	26	76
2	20	42	16	78
3	53	31	5	89
Totale	95	101	47	243

(a) Calcolare la moda per le distribuzioni condizionate della variabile *Prestazioni assicurative*.

(b) Si misuri il grado di associazione tra le agenzie e le tipologie di *Prestazioni Assicurative* stipulate.

a) Determinare la moda, prestazione assicurativa più frequente basta guardare le frequenze assolute:
- Prima agenzia: prima riga, quale modalità più frequente (22,28,26) frequenza maggiore è 28 quindi la moda è vita
- Seconda agenzia la moda sarà fra 20,42,16 la frequenza più elevata è 42 e la modalità è vita
- Terza agenzia modalità più frequente auto

b) Misurare il grado di associazione→ indici di Chi quadro

$$\chi^2 = \sum_i \sum_j \frac{c_{ij}^2}{n_{ij}}$$

bisogna calcolare le frequenze attese e le contingenze

	Prestazioni assicurative			
Agenzia	Auto	Vita	Casa	Totale
1	22	28	26	76
2	20	42	16	78
3	53	31	5	89
Totale	95	101	47	243

ATTESE

A	V	C
29,71	31,59	14,70
30,49	32,42	15,09
34,79	36,99	17,21

$$\chi^2 = \sum_i \sum_j \frac{c_{ij}^2}{n_{ij}} = \frac{(22-29,71)^2}{29,71} + \frac{(28-31,59)^2}{31,59} +$$

non è necessario calcolare la tabella delle contingenze

$$\chi^2 = \sum_i \sum_j \frac{c_{ij}^2}{n_{ij}} = \frac{(22-29,71)^2}{29,71} + \frac{(28-31,59)^2}{31,59} + \ldots + \frac{(5-17,21)^2}{17,21} = 36,8$$

Per avere un indice interpretabile occorre normalizzare:

$$\Phi^2 = \frac{\chi^2}{n} = \frac{36,8}{243} = 0,1513$$

$$V = \sqrt{\frac{\Phi^2}{\min\{2, c-1\}}} = \sqrt{\frac{0,1513}{2}} = 0,275$$

CORRELAZIONE

Chi quadro misura la connessione confrontando le frequenze osservate e quelle teoriche (che si devono avere quando vi è indipendenza statistica)

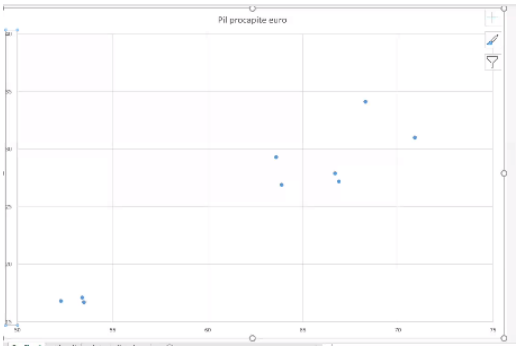
Relazione di interdipendenza fra <u>variabili di tipo quantitativo</u>

Regione	Tasso di attività	Pil procapite euro
Piemonte	66.9	27.2
Lombardia	68.3	34.1
Liguria	63.9	26.9
Toscana	66.7	27.9
Emilia Rom	70.9	31
Lazio	63.6	29.3
Campania	53.5	16.7
Puglia	53.4	17.1
Sicilia	52.3	16.8

Percentuale di popolazione attiva = popolazione in età lavorativa/ (numeratore tutti soggetti che lavorano o cercano lavoro) e popolazione totale

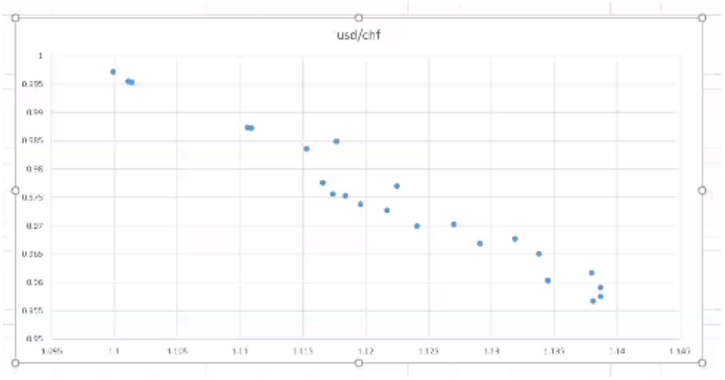
Primo passo: disegnare un diagramma: diagramma di dispersione: diagramma cartesiano si assegna una variabile a un asse e l'altra all'altro (irrilevante quale sull'asse delle X e quale sull'asse delle Y) i punti rappresentano i valori osservati (ogni punto rappresenta un'osservazione)

Modificando origine il diagramma cambia (la scala di misura è importante)



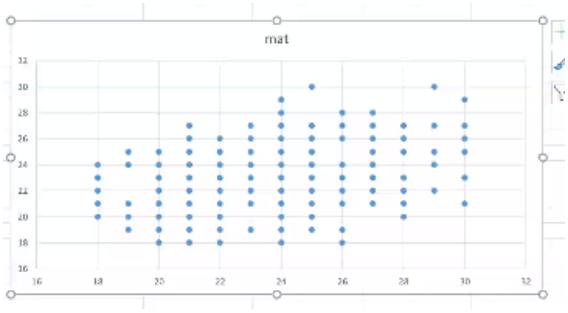
La distribuzione ha un particolare andamento, si nota che quando è alto il valore di una delle due variabili tendenzialmente corrisponde un valore alto anche dell'altra. Questa è una forma di associazione tra le due variabili, legame di tipo diretto (alto, alto; basso, basso)

ESEMPIO: Excel (foglio valute) tasso cambio euro/dollaro e dollaro/franco svizzero

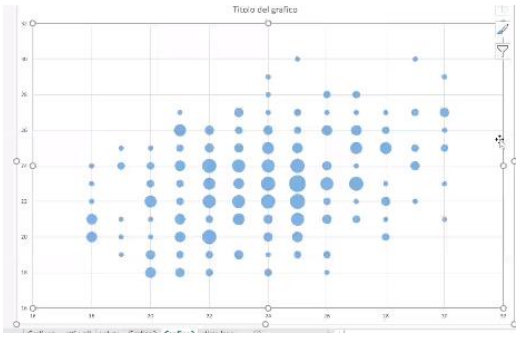


Quando uno dei due valori è alto tendenzialmente l'altro è basso (quando il dollaro si apprezza cresce il valore sull'asse delle ascisse e decresce quello sull'asse delle ordinate). Vi è una relazione tra le due variabili, ma di tipo inverso (quando una variabile è alta l'altra variabile è bassa e viceversa)

ESEMPIO: Excel voti statistica e matematica; distribuzione di frequenza (si hanno coppie di valori che non si sono mai presentate e coppie di valore che si presentano più volte.



Il diagramma di dispersione non tiene conto delle frequenze, per questo si utilizza un diagramma a bolle.

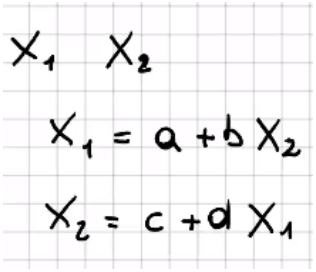


La dimensione dei punti che rappresentano l'osservazione non è costante, coppie che si presentano con più frequenza saranno pallini più grandi e viceversa osservazioni con poca frequenza saranno pallini più piccoli.

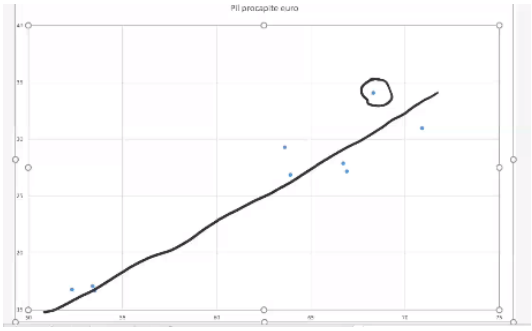
Si osserva una leggera tendenza a crescere delle due variabili (voto alto di una variabile tende ad essere relativamente più alto anche il voto dell'altra)

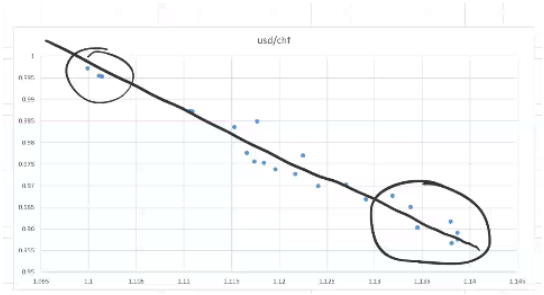
Questo tipo di associazione viene detta <u>CORRELAZIONE (relazione di interdipendenza</u>)

<u>CORRELAZIONE LINEARE</u>: particolare tipo di correlazione, quando una delle due variabili aumenta di una unità l'altra variabile aumenta di un valore tendenzialmente costante)

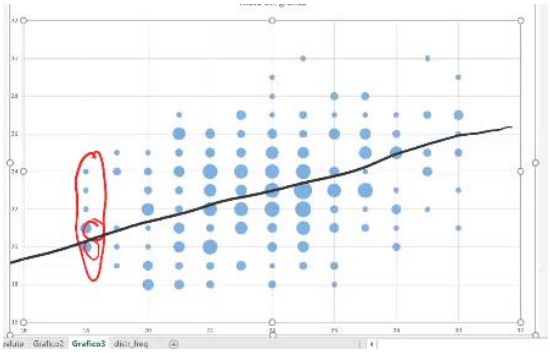


Se correlazione lineare perfetta: tutti i punti giacciono su una retta (mai nella realtà). Nella realtà i punti tendono a essere intorno a una retta

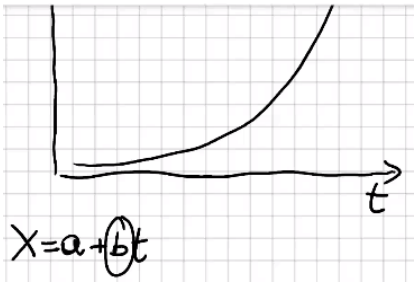




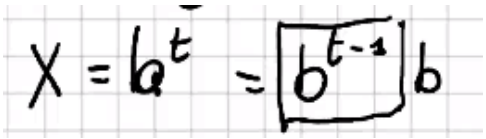
inclinata negativamente perché rappresenta la relazione tra le due variabili



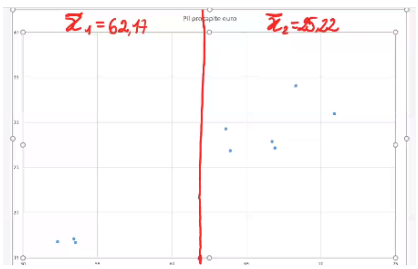
Nell'andamento lineare gli incrementi sono costanti (es. tempo sull'asse x e conteggio sull'asse delle Y, abbiamo un andamento lineare quando ogni giorno in più comporta un aumento di X pari a t) un incremento unitario di una variabile comporta un incremento costante dell'altra.



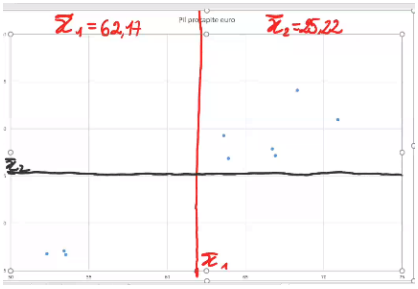
Un andamento esponenziale è del tipo b elevato alla t, il valore al tempo t sarà una certa percentuale del valore osservato il periodo precedente



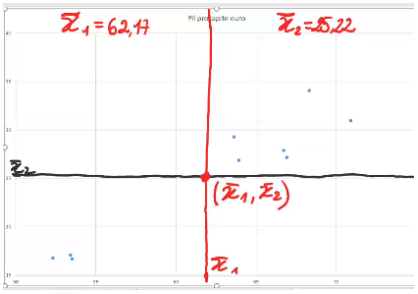
La relazione lineare può essere di tipo diretto o indiretto; può essere forte o debole. Strumento per misurare e distinguere in quale situazione ci si trova.



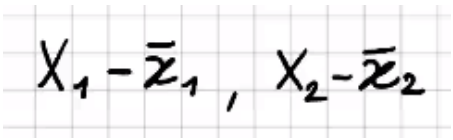
Nel descrivere la relazione che intercorre tra le due variabili abbiamo definito valori bassi o alti, alto o basso secondo un valore di riferimento che è la media. I valori medi di X e Y suddividono il piano in 4 quadranti. Stare a destra della media significa avere un valore alto (stare a sinistra significa avere un valore basso)



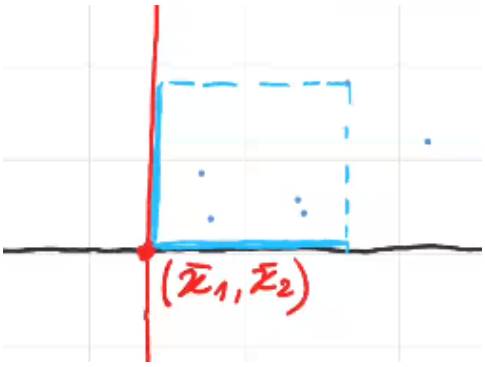
La linea orizzontale divide il grafico in due sezioni: sotto la media vi sono unti con valori del Pil bassi, sopra la media valori alti.



Si è così traslata l'origine degli assi in un punto di coordinate x medio, allora le coordinate saranno date dagli scarti dalla media per la prima e per la seconda variabile



Esempio:



Si può osservare:

- i punti che stanno nel primo quadrante sono punti per il quale si avranno scarti dalla media che saranno positivi per entrambe le variabili.

- nel secondo quadrante si avranno punti con scarto dalla media per la prima variabile negativo, per la seconda variabile lo scarto è positivo.

-terzo quadrante si avranno punti che avranno scarti negativi per entrambe le variabili.

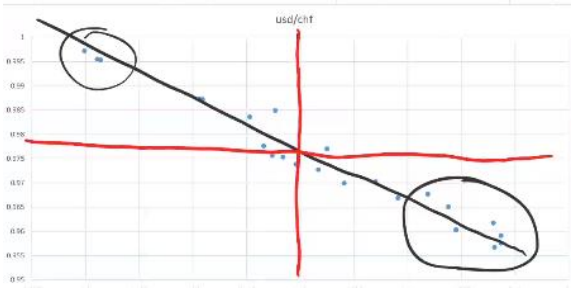
-quarto quadrante prima variabile i punti avranno scarto positivo e negativo per la seconda variabile.

(scostamenti concordi: scarti o entrambi positivi o entrambi negativi; scostamenti discordi: scarti postivi per una variabile e negativi per l'altra variabile)

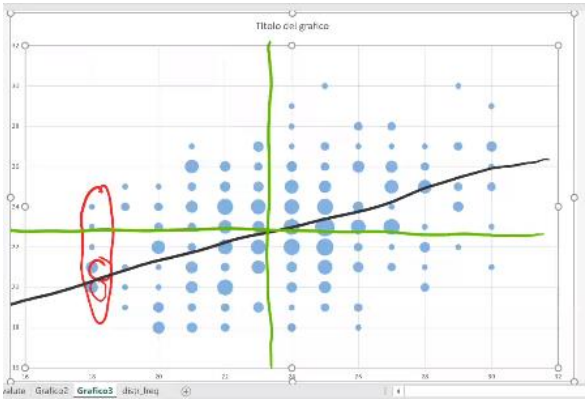
(i due caratteri presentano concordanza se la maggior parte degli scostamenti sono concordi, le variabili variano nella stessa direzione, a valori alti di una variabile corrispondono valori alti dell'altra variabile)

(i due caratteri presentano discordanza se la maggior parte degli scostamenti sono discordi, le variabili variano in direzioni opposte, a valori alti di una variabile corrispondono valori bassi dell'altra variabile e viceversa)

Se le variabili vanno nella stessa direzione la maggior parte dei valori osservati stanno nel primo e nel terzo quadrante (quando una variabile è alta è alta anche l'altra) quando si ha una relazione diretta i punti tendono a collocarsi nel primo e nel terzo quadrante.



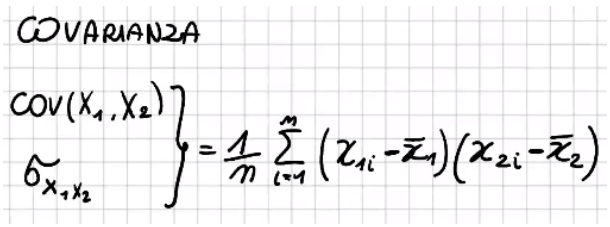
I punti in questa situazione stanno prevalentemente nel secondo e nel quarto quadrante, questo perché si è in una situazione in cui quando una delle due variabili è alta l'latra è bassa.



Qua non vi è una forte prevalenza dei punti a collocarsi in un determinato quadrante. Relazione debole

Prima misura di correlazione:

COVARIANZA: indice basato sugli scarti (media del prodotto degli scarti dal prodotto delle medie)



ESEMPIO 1: tutti i punti nel terzo e primo quadrante, gli scarti o sono entrambi positivi o entrambi negativi Entrambi positivi prodotto positivo, entrambi negativi prodotto positivo. Le situazioni nelle quali gli scarti hanno lo stesso segno sono prevalenti. Questa situazione viene detta concordanza

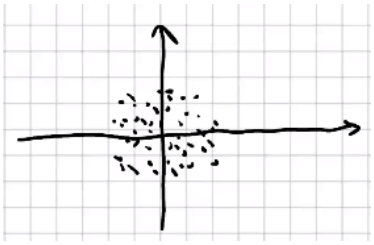
$COV(X_1, X_2) > 0$ CONCORDANZA

Se la covarianza è inferiore a 0 significa che i valori negativi sono prevalenti, la maggior parte degli scarti sono di valore discorde (esempio 2)

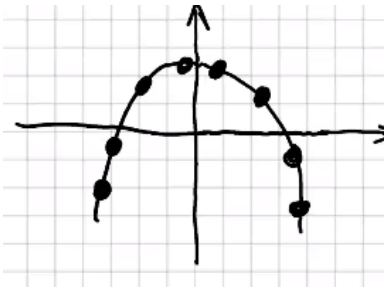
$COV(X_1, X_2) < 0$ DISCORDANZA

Se la covarianza è pari a zero è una situazione di assenza di relazione lineare (mai osservabile nella pratica)

$COV(X_1, X_2) = 0$ ASSENZA DI RELAZIONE LINEARE


addendi positivi e negativi tendono a bilanciarsi

NB: Il fatto che non esista relazione lineare non significa che non vi sia relazione tra le due variabili


relazione molto forte tra le due variabili, ma se si calcola la covarianza su questi valori questa risulta essere apri a 0. Questo perché la covarianza misura la linearità di una relazione, la relazione in questo caso è di tipo parabolico. Nel primo tratto le due variabili crescono congiuntamente, nel secondo tratto una variabile cresce e una decresce.

Covarianza = tendenza delle variabili a variare insieme, come variano insieme (stessa direzione, direzione opposta)

$$COV(X_1, X_2) = \frac{1}{n} \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

Può essere calcolata anche come:

$= \frac{1}{n} \sum x_{1i} x_{2i} - \bar{x}_1 \bar{x}_2$ media dei prodotti dei valori delle variabili e sottrarre il prodotto delle medie delle due variabili

Regione	Tasso di attività	Pil procapite eurc	$x_{1i} x_{2i}$
Piemonte	66.9	27.2	1819,68
Lombardia	68.3	34.1	2329,03
Liguria	63.9	26.9	1718,91
Toscana	66.7	27.9	
Emilia Rom	70.9	31	
Lazio	63.6	29.3	
Campania	53.5	16.7	
Puglia	53.4	17.1	
Sicilia	52.3	16.8	
			14475,16

$= \frac{1}{n} \sum x_{1i} x_{2i} - \bar{x}_1 \bar{x}_2$

$\bar{x}_1 = 62,17 \qquad \bar{x}_2 = 25,22$

$COV(X_1, X_2) = \frac{1}{9} 14475,16 - 62,17 \times 25,22 =$

$= 40,37$

Problema: come per la varianza la covarianza è espressa in una unità di misura che è il prodotto delle unità di misura delle variabili. Se si utilizzassero i valori espressi su scala ordinaria e non i valori percentuali, si otterrebbe un valore della covarianza differente. Si può individuare una regola: date due variabili e la loro covarianza, si applica una trasformazione lineare diversa a ciascuna delle due variabili

$X_1 \quad X_2 \quad COV(X_1, X_2)$

$Y_1 = a + bX_1 \quad Y_2 = c + dX_2$

quanto sarà la covarianza di Y?

$COV(Y_1, Y_2) = bd\ COV(X_1, X_2)$

La covarianza non è influenzata dalla traslazione, sarà in relazione alla covarianza tra le variabili X, ottenuta moltiplicando per due coefficienti, risente però del cambio di scala (nell'esempio potremmo moltiplicare per 1000 o dividere per 100)

Interpretazione difficile della covarianza, si può interpretare il segno (positivo le variabili sono concordi, segno negativo le variabili sono discordi)

<u>COEFFICIENTE DI CORRELAZIONE LINEARE</u>

Pari alla covarianza tra le due variabili divisa per il prodotto delle due deviazioni standard

NB: l'indice non viene normalizzato perché non è compreso tra 0 e 1 , è però sicuramente un indice relativo.

COEFFICIENTE DI CORRELAZIONE LINEARE

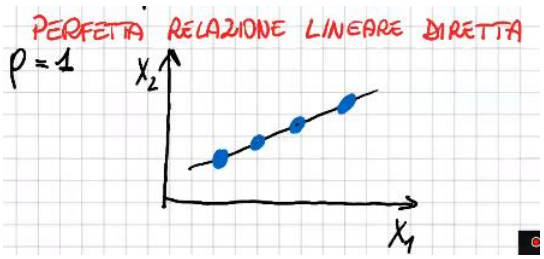
$\left.\begin{matrix} r_{X_1 X_2} \\ \rho_{X_1 X_2} \end{matrix}\right\} = \frac{COV(X_1, X_2)}{\sigma_{X_1}\ \sigma_{X_2}}$

Con questa operazione si elimina l'unità di misura (numero puro)

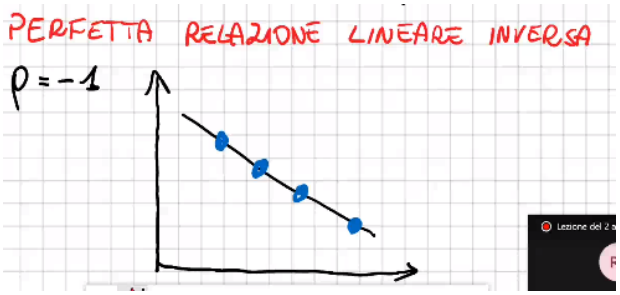
NB: questo perché $-\sigma_{x_1}\sigma_{x_2} \leq cov(X_1, X_2) \leq \sigma_{x_1}\sigma_{x_2}$

Caratteristica estremamente importante: è compreso tra due estremi che hanno un significato ben preciso.

$$-1 \leq \rho \leq 1$$



Tutti i valori osservati giacciono su una retta inclinata positivamente, relazione lineare perfetta tra le due variabili (i punti sono perfettamente allineati) (concordanza)

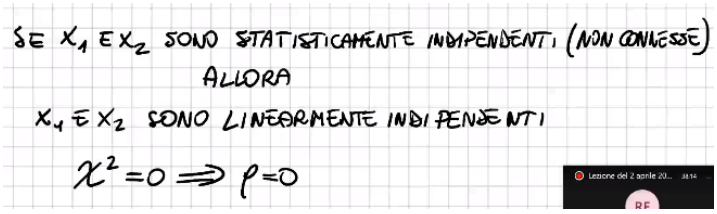


Situazione nella quale tutti i punti sono perfettamente allineati lungo una linea che è inclinata negativamente, relazione inversa (discordanza)

NB: la retta è una retta qualunque, la sua inclinazione può essere qualunque purché non orizzontale o verticale.

Entrambi sono casi teorici, nella pratica non è praticamente possibile avere casi di questo tipo

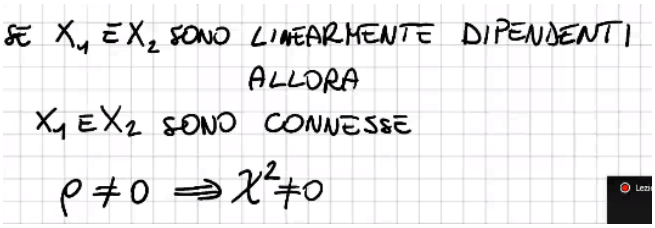
Indipendenza statistica prevede che le variabili non sono connesse, dire invece che sono linearmente indipendenti non significa che le variabili non siano connesse.



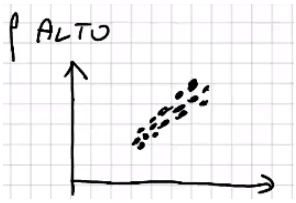
L'indipendenza statistica è una condizione sufficiente per l'indipendenza lineare

NB: può capitare che ro sia pari a 0 ma il Chi quadrato non sia pari a 0.

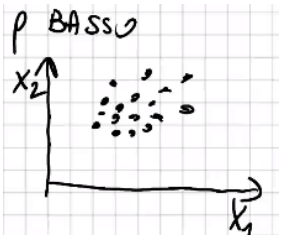
Al contrario:



Quando il coefficiente di correlazione è alto i punti tendono a disporsi molto vicini a una linea



Quando il coefficiente di correlazione è basso i punti tendono a disperdersi molto di più



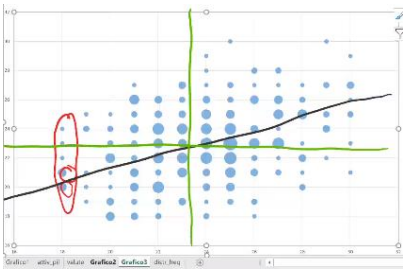
ESEMPIO PIL:

se si disegna una linea all'interno del diagramma si può notare che i punti sono piuttosto vicini

ESEMPIO VALUTE:

punti sono in molto vicini alla retta

ESEMPIO DIAGRAMMA A BOLLE

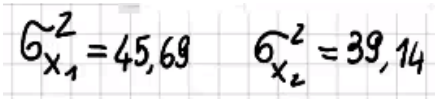


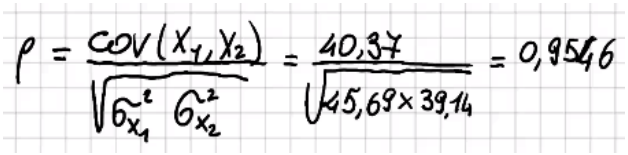
in questo caso vi è molta più dispersione

Tanto più il valore del coefficiente è prossimo a 1 o -1 tanto più la nuvola di punti ha una forma affusolata vicino alla retta (più facile disegnare retta), tanto più coefficiente vicino a 0 tanto più la nuvola di punti diviene indistinta (più difficile disegnare retta)

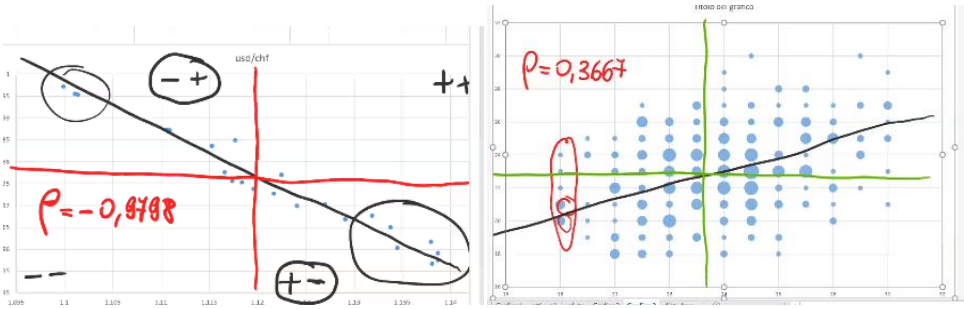
TORNANDO ALL'ES:

COV= 40,37 ma non dice se la relazione fra le variabili è forte o debole, si calcola allora il coefficiente di correlazione lineare. Prima occorre calcolare la deviazione standard per le due variabili (e quindi la varianza)





valore alto, prossimo a 1



Ne secondo caso la relazione è molto più debole

Calcolo della <u>covarianza nel caso di distribuzione di frequenza</u>

$$COV(X_1,X_2)=\frac{1}{n}\sum_i\sum_j x_{1i}\,x_{2j}\,n_{ij}-\bar{x}_1\bar{x}_2$$

Etichette di riga	18	19	20	21	22	23	24	25	26	27	28	29	30	Totale complessivo
18			1	4	1	1	1							11
19		1		1			2	1						6
20	4	3	1	1	5	2	2	1						19
21	3	2	4	4	2	2	4	2	5	1				29
22	2	2	7	2	6	5	7	3	3					37
23		3		5	6	4	6	3	2	3				32
24	2	1	3	3	7	7	5	5	3	1	1	1		40
25		2	4	6	8	10	7	4	2	2			1	46
26	1	2		3	2	6	3		4	1	2			24
27				2	1	7	1	5	4	1	2			23
28			2	1	3	1		5	2	1				15
29					1		3	2		2			1	9
30				1		1		2	1	3		1		9
Totale complessivo	12	16	26	33	42	46	42	33	26	15	5	2	2	300

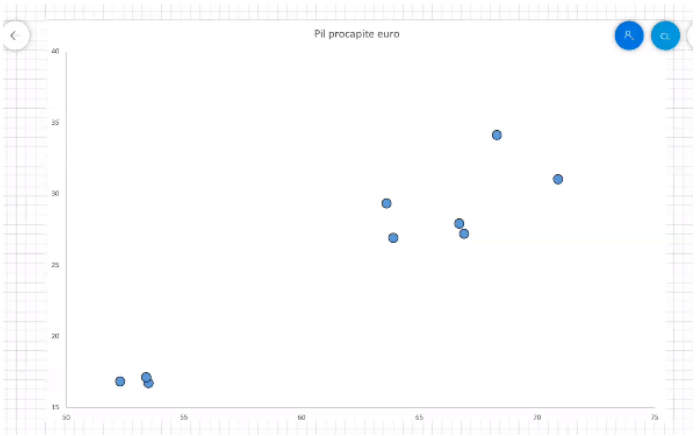
in questo caso (18*20*4) prodotto di valore di riga, valore di colona e frequenza

Quando si analizzano dati quantitativi solitamente di ha una distribuzione unitaria non una distribuzione di frequenza.

ANALISI DI REGRESSIONE, (regressione lineare)

Argomento connesso all'analisi di correlazione lineare, ma in realtà ha delle sue peculiarità che ne fanno un oggetto diverso. L'analisi di regressione è lo strumento più importante di tutta la statistica, più o meno tutti i metodi di carattere statistico possono essere riconducibili alla regressione.

Cosa significa fare regressione?



X = TASSO DI ATTIVITA'
Y = PIL PROCAPITE

Supponiamo che intercorra una relazione fra le due variabili di questo tipo: (relazione non perfetta)

$$Y=f(X)+\text{ERRORE}$$

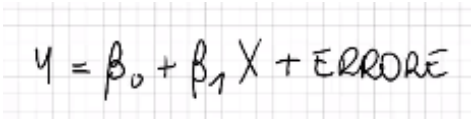
f(x) viene chiamata funzione sistematica

La variabile Y viene chiamata variabile di risposta, X variabile esplicativa

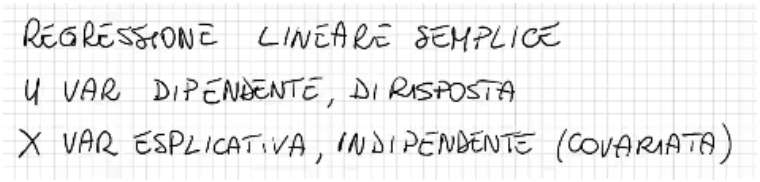
Una relazione di questo tipo dice che i valori che la Y assume dipendono dalla variabile X ovvero, quando diciamo che Y è in funzione di X diciamo che è stata considerata in funzione di diversi valori della variabile X

Osservando il diagramma vi sono diversi valori della Y (alcuni valori bassi, alcuni alti), la Y varia e si cerca di descrivere questa variabilità utilizzando un'altra variabile, la variabile X. Almeno in parte i valori di Y sono tra loro diversi perché sono stati osservati in corrispondenza di diversi valori della variabile X.

La f (x) viene generalmente esplicitata nel seguente modo:

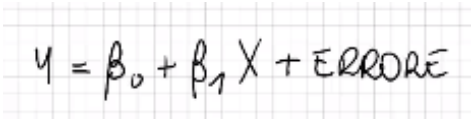


ANALISI DI REGRESSIONE LINEARE SEMPLICE (1 SOLA VARIABILE X E UNA SOLA VARAIBILE Y)

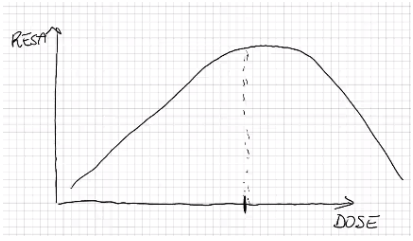


Occorre risolvere quindi un problema di regressione lineare, fare un'analisi di regressione lineare in prima battuta corrisponde a determinare beta 0 e beta 1 (intercetta e pendenza della retta)

DIFFERENZE FRA REGRESSIONE E CORRELAZIONE: modo in cui vengono trattate le variabili. Nella correlazione le variabili vengono considerate simmetricamente, la regressione è assimetrica, una variabile spiega e una variabile viene spiegata. Si cerca di interpretare la variabile di risposta in termini di variabile esplicativa (PIL pro capite, tasso di attività). Nel primo caso si parla di interdipendenza, nella regressione la relazione è di dipendenza (Y dipende da X) non necessariamente la dipendenza è di tipo causale.



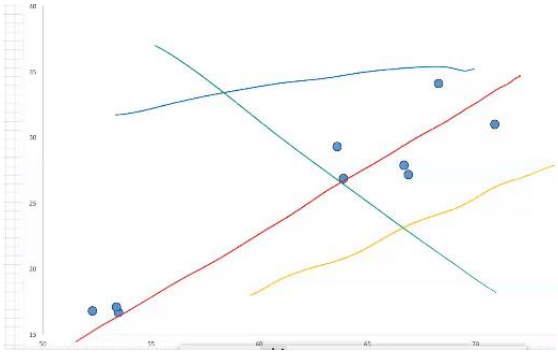
Si considera una funzione rappresentata da una retta per diversi motivi. E' una tecnica molto antica e estremamente semplice da utilizzare, con la regressione lineare si possono gestire relazioni più complesse, la retta è molto semplice da interpretare (due soli parametri beta 0 e beta 1 di interpretazione molto chiara: intercetta valore di Y quando X uguale a 0, ordinata del punto in cui la retta taglia asse Y e pendenza quanto varia Y all'incremento di una unità della variabile X), la retta è una buona approssimazione locale di qualunque funzione (esempio: X = dose fertilizzante, Y = resa di una pianta, ci si aspetta che in un primo tratto all'aumentare di X ci sia un incremento di Y, arrivati a un certo dosaggio presumibilmente un incremento di X non produrrà nessun effetto, poi a un certo punto un incremento di X avrà effetto negativo su Y, relazione sicuramente non lineare)



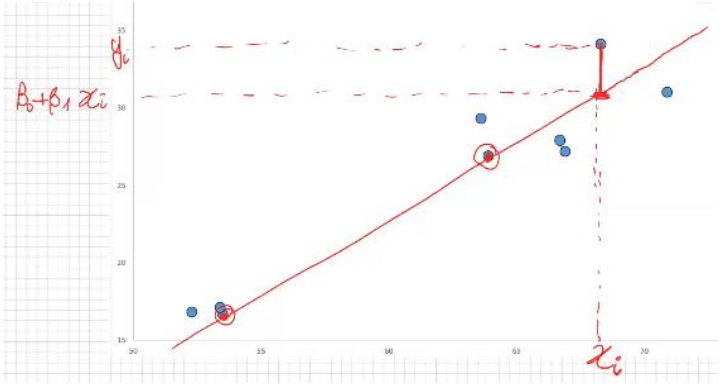
se si intercetta un determinato range lineare la retta rappresenta bene la relazione che intercorre tra le due variabili.

Non si conosce la vera forma funzionale tra le variabili, ma si utilizza la retta cosicché localmente questa ne dia una buona approssimazione.

Primo problema della regressione: quantificare i parametri (determinare valori di Beta 0 e Beta 1, ovvero individuare una retta)



Tutti noi disegneremmo una linea vicina a quella tracciata in rosso escludendo tutte le altre che possono essere tracciate, questo perché sceglieremo la retta che passerà più vicina alla nuvola di punti (distanza minima) la distanza punto retta viene calcolata in verticale, lungo la variabile Y.



Si calcola per ciascun punto la distanza punto retta

$$y_i - \beta_0 - \beta_1 x_i$$

Se il punto sta al di sopra della retta la differenza sarà positiva

Se il punto sta al di sotto della retta la differenza sarà negativa

Il segno è irrilevante, ciò che interessa è vedere se il punto è vicino a lontano, si ha una quantità per ciascun punto, occorre quindi sintetizzare:

$$S = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Questo problema si chiama <u>PROBLEMA DEI QUADRATI</u>

Elevando al quadrato si elimina il segno che è irrilevante, somma così si sintetizza. La retta migliore è quella che rende minima la somma. Si cercano i valori di Beta 0 e Beta 1 che rendono minima questa somma.

I valori dei coefficienti che minimizzano la somma (b1 e b0) soluzione al problema dei miniquadrati

$$S = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

MINIMI QUADRATI

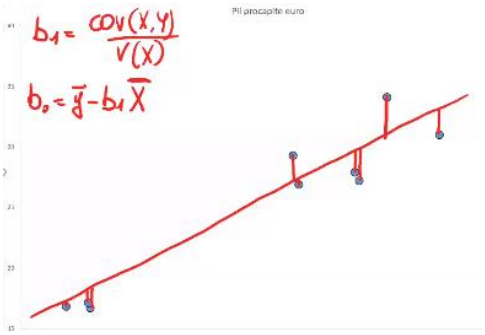
$$b_0 = \bar{y} - b_1 \bar{X}$$

$$b_1 = \frac{COV(Y,X)}{VAR(X)} = \frac{\sigma_{xy}}{\sigma_x^2}$$

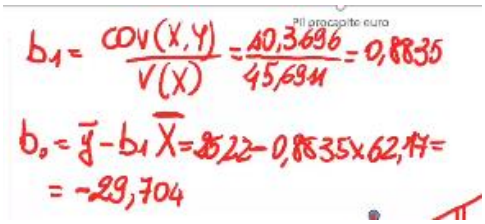
COEFFICIENTI DI REGRESSIONE

$$Y = b_0 + b_1 X$$

NB: sempre a parità della variabile X, le distanze vengono calcolate in relazione alla variabile Y



NB: b1 non è uguale al coefficiente di correlazione lineare (per quanto simile)



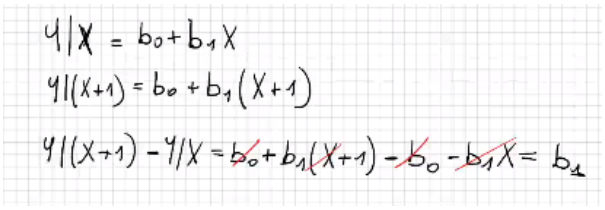
Come costruire la retta in excel:

- *Avere dati sul foglio excel*
- *Costruire il diagramma di dispersione (con attenzione a cosa si mette sugli assi)*
- *Andando poi su uno qualunque dei punti del diagramma (tasto dx) selezionare dal menu che appare "linea di tendenza", selezionare poi in fondo "visualizza equazione sul grafico" (che sarà pari all'equazione calcolata)*

Diversi scopi dell'analisi di regressione:

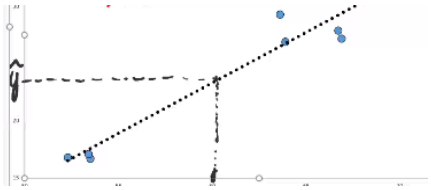
- -una prima è descrittiva, mediante questa semplice equazione si descrive la relazione fra le due variabili

- -secondo scopo è di carattere interpretativo, i coefficienti della retta hanno una loro interpretazione, l'intercetta solitamente è di scarso interesse mentre la pendenza dice di quanto varia la variabile di risposta (Y) al crescere unitario della variabile esplicativa (X) (nell'esempio se il tasso di attività aumenta di una unità tendenzialmente allora il pil pro capita aumenta di circa 883 euro qualunque sia il tasso di attività)



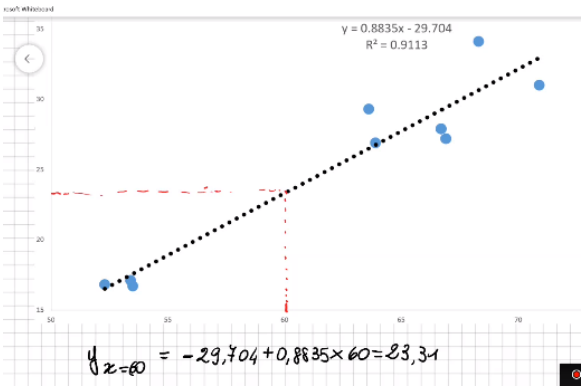
se b1 positivo Y aumenta, se negativo Y diminuisce

Altra finalità: può essere utilizzata per fare previsione. Si può fare in due contesti diversi: se viene fissato un valore della variabile X (esempio 60) quale è il valore della variabile Y? Si proiettano i valori sulla retta e si guarda il valore corrispondente di Y

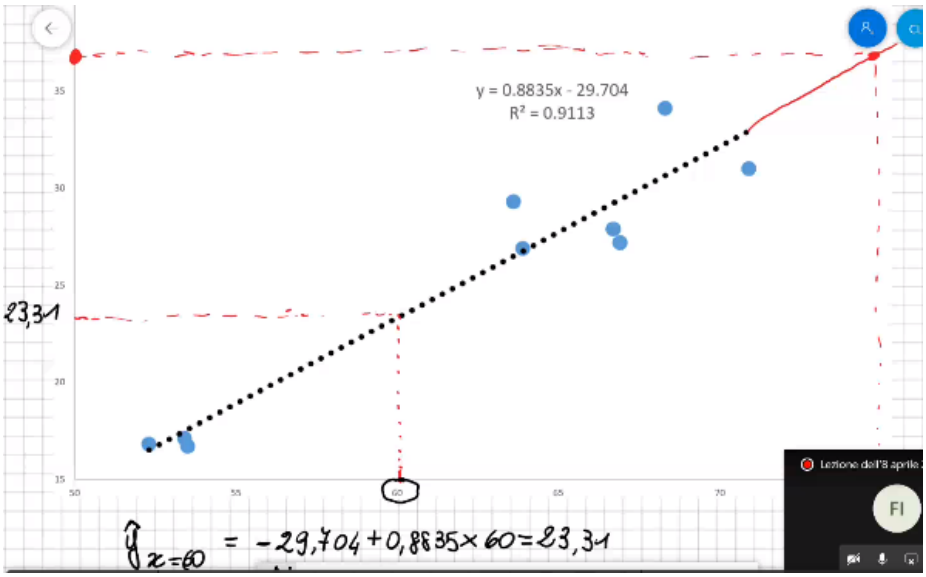


questa operazione viene chiamata operazione di interpolazione (questa previsione è soggetta ad errore perché ci si basa sulla retta anche se la relazione fra le variabili non è limitata alla componente sistematica vi è un errore).

Finalità vi è previsiva. Questa operazione di previsione talvolta va sotto il nome di <u>interpolazione.</u> Il valore della varaibile X per il quale stiamo cercando di fare la previsione è all'interno del range preso per la retta di regressione.



Supponiamo di voler fare previsione con un valore di X che è fuori dal range (es. 75). Si procede con la stessa logica



$$\hat{y}_{x=75} = -29,704 + 0,8835 \times 75 = 36,56$$

Questo tipo di previsione si chiama estrapolazione. Si differenzia dall'interpolazione perché il valore di X preso in considerazione è al di fuori del range.

Differenzia sostanziale dal punto di vista logico : con l'interpolazione alla componente sistematica si è soggetti ad errore e si può determinare la natura dell'errore stesso. Estrapolazione: quando si prende il valore di Y associato al valore di X anche in questo caso si è soggetti ad errori, ma quando si fa estrapolazione vi è una seconda fonte di errore legata al fatto che si prende la retta e la si prolunga al di fuori dei valori del range considerato; l'operazione di prolungmento è un'operazione estremamente delicata (retta è una buona approssimazione locale, per piccoli intervalli), quando ci si imita al range dei valori osservati della variabile X si possono fare molte considerazioni; al di fuori del range di valori osservati non si ha nessuna informazione (si spera che la retta sia valida anche al di fuori del range dei valori osservati) non si hanno elementi per sapere se la retta vada bene oppure no. Nelle operazioni di estrapolazione vi sono due fonti di errori: uno intrinseco al processo di previsione (la previsione si basa solo sulla componente sistematica tralasciando l'errore) e una secnda fonte di errore specifica del'estrapolazione, legata al fatto che noi utilizziamo il modello (componente sistematica =retta) anche in regioni sulle quali non siamo in grado di valutare se la retta vada bene oppure no (prolungamento retta fuori dal range di valori).

Se il prolungamento è breve può essere ragionevole pensare che la retta possa andare bene, se lo spostamento è di valori molto grandi la retta diventa molto azzardata.

CARATTERISTICHE DELLA RETTA DI REGRESSIONE (tutte riferite al fatto che la retta di regressione ha intercetta)

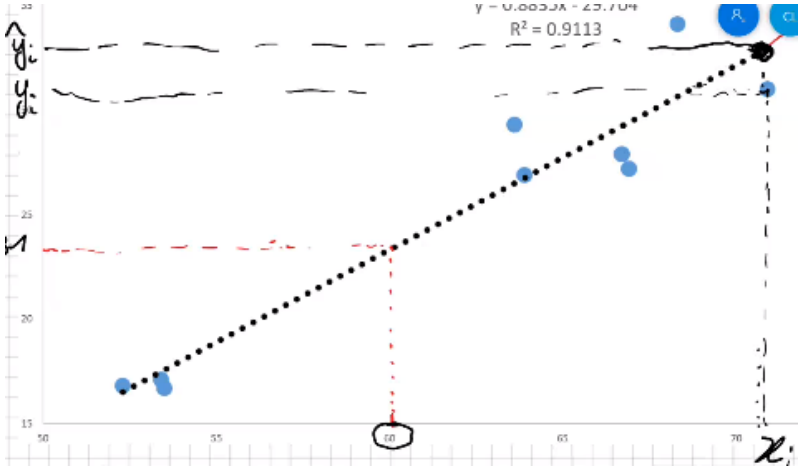
$\beta_0 = 0$ $\qquad Y = \beta_1 X + \text{ERRORE}$

Hp: in cui l'incertezza è pari a 0 (vincolo) Delle infinite rette che si possono considerare si limita l'attenzione a quelle che passano per l'origine → OPERAZIONE SCONSIGLIATA SEMPRE

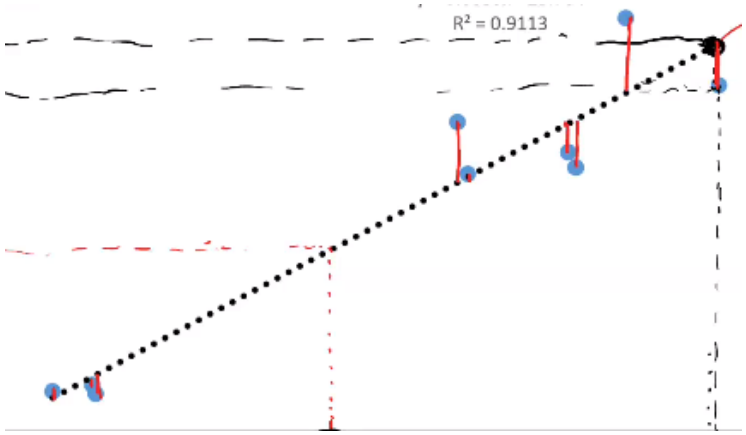
Quando si fa un modello di regressione si fa un modello completo, in cui l'intercetta non viene fissata ma viene lasciata nel modello. Conseguenze di lasciare l'intercetta:

1. La retta di regressione passa per il punto che ha come coordinate x medio e y medio
2. Dopo aver costruito la retta si determinano i residui.

$\hat{y}_i = b_0 + b_1 x_i$ VALORE TEORICO valori previsti dalla retta di regressione



$e_i = y_i - \hat{y}_i$ RESIDUO differenza tra valore osservato e valore teorico per ciascun punto osservato



Se il punto è sopra la retta il residuo è positivo

Se il punto è sotto la retta il residuo è negativo

Proprietà dei residui: se nel modello vi è l'intercetta allora la somma dei residui è uguale a 0

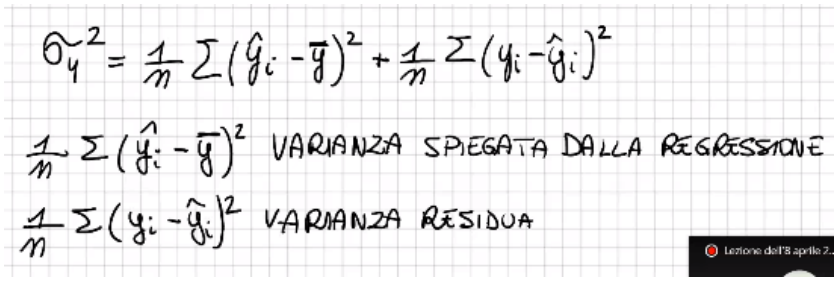
$\sum e_i = 0$ i residui pos e neg si bilanciano

3. La somma dei valori osservati è uguale alla somma dei valori teorici per la variabile y

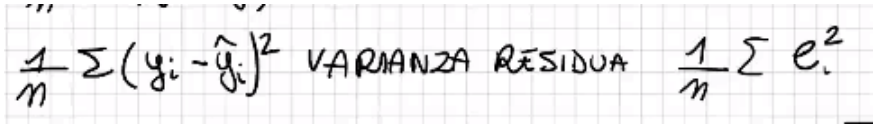
$$\sum y_i = \sum \hat{y}_i \qquad \bar{y} = \bar{\hat{y}}$$

4. Qualuque nuvola di punti con il metodo dei minimi quadrati consente sempre di individuare una retta

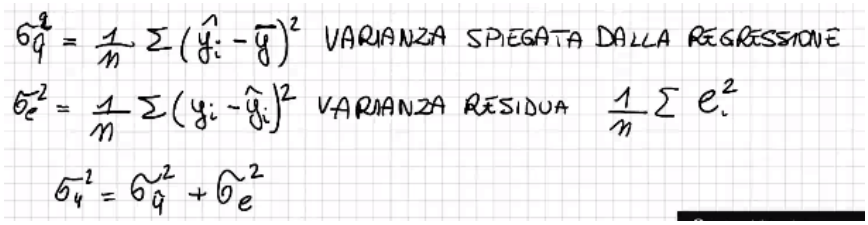
   Bontà di adattamento: capacità della retta di regressione di descrivere in modo adeguato ciò che si è osservato. Se la retta ha bassa bontà di adattamento la retta servirà poco, se la bontà è elevata significa che la retta è adeguata .Indice che misura questa caratteristica: per prima cosa introduciamo un'altra caratteristica (sempre con intercetta all'interno del modello) se si ha l'intercetta si può scomporre la varianza della variabile Y in due termini: il primo termine si chiama varianza spiegata dalla regressione, il secondo termine si chiama varianza dei residui (o varianza residua)



   Nel primo termine si prendono in considerazioni gli scarti dei residui dalla media
   Nel secondo si ha la somma dei residui al quadrato



   Allora

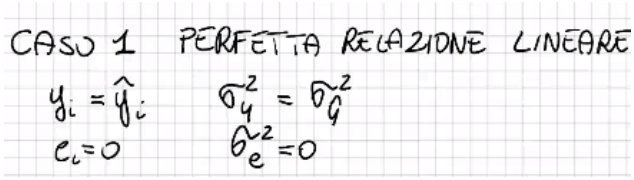


Questi termini sono importanti perché una delle motivazioni del modello di regressione è quello di studiare la variabilità della variabile Y in termini di variabile X. La varianza spiegata dalla regressione fornisce la quota di variabilità della variabile Y dovuta alla variabile X (si osserva Y in corrispondenza di diversi valori della X) la varianza residua fornisce la quota di variabilità della Y dovuta all'errore (variabilità dei residui, non dovuta alla regressione) .

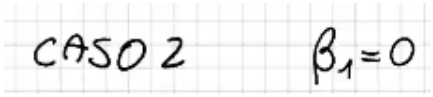
La varianza totale misura quanto è la variabilità totale della variabile Y. Quando si utilizza un modello di regressione si indivuduano due fonti di variabilità: un primo motivo per cui varia Y è perché varia X (variabilità sistematica perché ogni volta che X assume un certo valore Y assume un valore specifico) , alla struttra sistematica si deve aggiungere una parte di variabilità dovuta all'errore, una parte della variabilità è dovuta all'errore. Con la scomposizione della varianza della Y si è in grado di misurare l'importanza di queste componenti. La prima quota di variabilità spiega quanto Y è spiegato da X, la seconda quota misura la variabilià dei residui, e questo determina un'ulteriore variabilità non imputabile alla componente sistematica del modello ma imputabile all'errore.

Casi estremi:

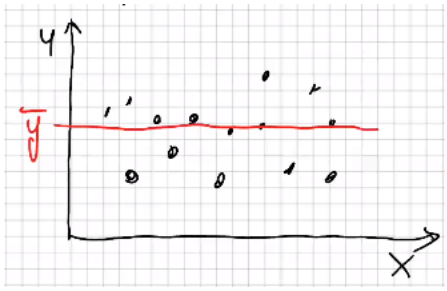
- La retta di regressione spiega tutto e non vi è errrore, perfetta relazione lineare fra le due variabili (punti perfettamente allineati sulla retta)



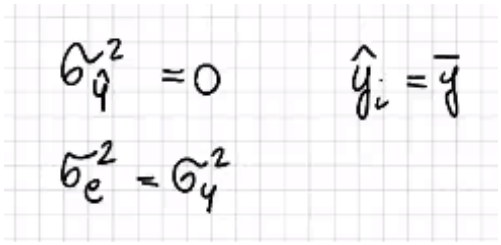
- Situazione opposta, la retta di regressione non spiega niente, tra la Y e la X non vi è alcuna relazione lineare



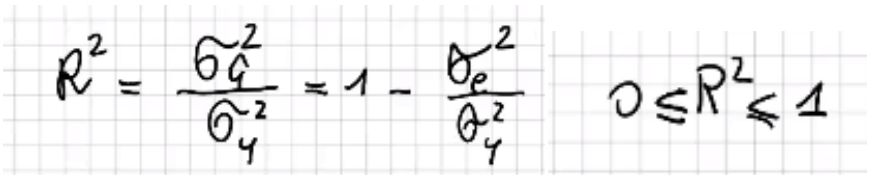
retta orizzontale, al variare di X Y non cambia

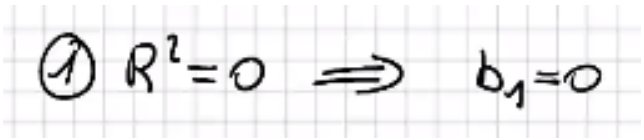


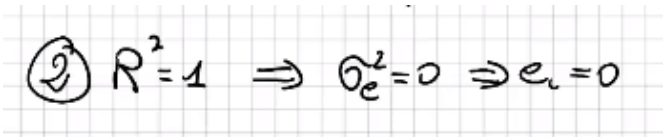
se la retta è orizzontale allora: tutta la variabilità della Y sta nei residui



L'indice di bontà di adattamento si chiama <u>coefficiente di determinazione lineare</u>:

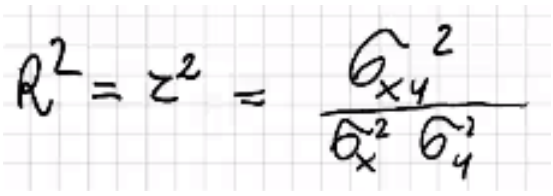






R dipende dal contesto in cui si analizza la retta di regressione, in ambito socio economico un valore pari a 0,5 è cosiderato più che accettabile perché l'errore è importante. Non esiste una regola per giudicare R quadro in mod assoluto.

Si può dimostrare che R quadro è pari al coefficiente di correlazioe lineare fra le due variabili (solo per la regresione lineare semplice)



molto spesso si utilizza il coeff di correlazione lienare al quadrato per calcolare R quadro

Esercizio n. 2 del 07/09/2015

Nel corso degli ultimi sette anni il fatturato (in milioni di euro) di una azienda ha seguito il seguente andamento:

Anno	2010	2011	2012	2013	2014	2015	2016
Fatturato	0.82	1.00	1.02	1.14	1.22	1.49	1.28

(a) Si calcoli il tasso medio di variazione del fatturato per il periodo considerato.

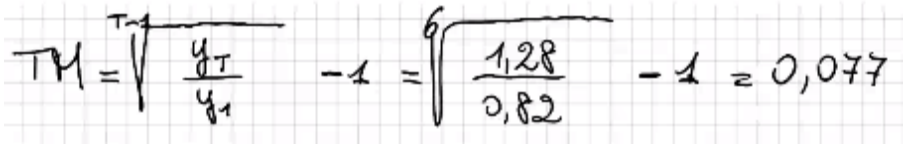
(b) Si calcolino i coefficienti della retta di regressione che descrive l'andamento del fatturato nel tempo.

(c) Si misuri la bontà di adattamento della retta di regressione individuata al punto precedente.

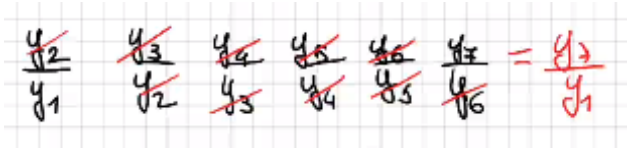
(d) Sulla base del tasso medio di variazione calcolato al punto (a) e del modello di regressione individuato al punto (b), si forniscano le corrispondenti previsioni di fatturato per il 2017.

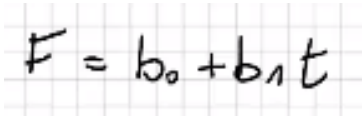
Si ha una serie storica , come si evolve nel tempo il fatturato

a) Modo iù veloce per calclarlo, senza calcolare tutti gli indici a base mobile:



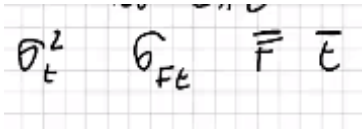
va bene perché :

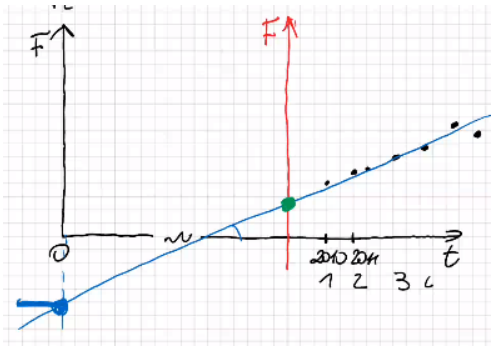


b) 

hp di relazione di tipo lineare, fatturato cresca linearmente nel tempo, da un anno all'altro il fatturato cresca di una quota costante.

Occorre perciò calcolare:





Utilizzare come tempo i valori 1,2,3,4 (traslare asse delle X)

non cambia la nuvola di punti, perciò non cambia la retta , ma cambia l'equazione della retta (la pendenza rimane inalterata) cambia l'intercetta.

$$\bar{t} = \frac{n+1}{2} = 4$$

$$\bar{F} = \frac{1}{n}\sum F_i = \frac{1}{7}(0{,}82 + 1 + \ldots + 1{,}28) = 1{,}14$$

$$\sigma_t^2 = \frac{1}{n}\sum t_i^2 - \bar{t}^2 = \frac{1}{7}(1^2 + 2^2 + \ldots + 7^2) - 4^2 = 4$$

$$\sigma_F^2 = \frac{1}{7}(0{,}82^2 + 1^2 + \ldots + 1{,}28^2) - 1{,}14^2 = 0{,}0407$$

$$\sigma_{Ft} = \frac{1}{n}\sum F_i t_i - \bar{F}\bar{t} = 0{,}3657$$

$$b_1 = \frac{\sigma_{Ft}}{\sigma_t^2} = \frac{0{,}3657}{4} = 0{,}09$$

$$b_0 = \bar{F} - b_1\bar{t} = 1{,}14 - 0{,}0914 \times 4 = 0{,}7729$$

La pendenza mi dice che mi posso aspettare che mediamente il fatturato è aumentato i 0,09 per ogni anno (circa 90mila euro l'anno).

c) $$R^2 = r^2 = \frac{\sigma_{Ft}^2}{\sigma_F^2 \sigma_t^2} = \frac{0{,}3657^2}{0{,}0407 \times 4} = 0{,}8216$$ valore molto alto

d) Previsione in due modi diversi

-con il tasso medio di variazione

$$\hat{F}_1 = y_{2016} \times (1 + TM) = 1{,}28 \times (1 + 0{,}077) = 1{,}3786$$

-con la retta di regressione

	1	2	3	4	5	6	7	8
Anno	2010	2011	2012	2013	2014	2015	2016	2017
Fatturato	0.82	1.00	1.02	1.14	1.22	1.49	1.28	

$$\hat{F}_2 = b_0 + b_1 \times 8 = 0{,}7729 + 0{,}0914 \times 8 = 1{,}5043$$

Previsioni diverse perché si sono utilizzati metodi diversi. Con il primo metodo si ipotizza che la crescita non sia di entità costante ma proporzionale alla crescita del periodo precedente (non ammontare costante ma 7% del valore dell'anno precedente)

ESERCIZI SULLA REGRESSIONE

PROVARSI IL TESTO DEI SEGUENTI ESERCIZI

16/2/2018 N.2
11/5/2018 N.3 c
18/4/2019 N.3 A B C D
7/6/2019 N.2

Esercizio n. 2 del 16/02/2018

Un ricercatore ha stimato la relazione tra voto di laurea (L) e voto all'esame di maturità (M) utilizzando un'analisi di regressione lineare. Sulla base delle 158 osservazioni a disposizione, il ricercatore ha ottenuto per l'intercetta della retta di regressione un valore pari a 43.51. Sapendo che $\sum L_i = 12768$, $\sum M_i = 12903$, $\sum L_i^2 = 1039930$ e $\sum M_i^2 = 1065605$ calcolare:

(a) il valore della pendenza della retta di regressione;

(b) il valore del coefficiente di determinazione lineare.

$\hat{L} = b_0 + b_1 M$

Si vuole spiegare il voto di laurea in funzione del voto di maturità

n = 158

b0=43,51

a)Calcolare b1

a) $b_1 = \frac{\sigma_{ML}}{\sigma_M^2}$ non possiamo utilizzare questa formula perché non si ha modo di determinare la covariazna fra le due variabili (COV richiede di conoscere i prodotti dei valori delle variabili, ma il testo non i fornisce)

Sappiamo che

$b_0 = \bar{L} - b_1 \bar{M}$ b0 è dato , le medie possono essere calcolate e perciò si ricava b1

$\bar{L} = \frac{1}{n} \sum L_i = \frac{12768}{158} = 80,81$

$\bar{M} = \frac{1}{n} \sum M_i = \frac{12903}{158} = 81,66$ si calcolano le medie, dopodichè si calcola b1

$b_1 = \frac{\bar{L} - b_0}{\bar{M}}$ $b_1 = \frac{80,81 - 43,51}{81,66} = 0,4612$

b)calcolare il coefficiente di determinazione lineare

relazione tra R quadro e la pendenza della retta di regressione

b) $R^2 = r_{ML}^2 = \frac{\sigma_{ML}^2}{\sigma_M^2 \sigma_L^2} = \left(\frac{\sigma_{ML}^2}{\sigma_M^4}\right) \frac{\sigma_M^2}{\sigma_L^2} = b_1^2 \frac{\sigma_M^2}{\sigma_L^2}$ (b_1^2)

. calcolare le varianze (media dei quadrati meno il quadrato della media)

$\sigma_L^2 = \frac{1}{n} \sum L_i^2 - \bar{L}^2 = \frac{1039930}{158} - 80,81^2 = 51,56$

$\sigma_M^2 = \frac{1}{n} \sum M_i^2 - \bar{M}^2 = \frac{1065605}{158} - 81,66^2 = 75,24$

$R^2 = b_1^2 \frac{\sigma_M^2}{\sigma_L^2} = 0,4612^2 \times \frac{75,24}{51,56} = 0,3103$

Esercizio n. 3 del 111/05/2018 Turno C

In tabella viene riportata la serie storica semestrale (dal primo semestre 2013 al secondo semestre 2016) delle importazioni via mare di merce containerizzata. I dati, espressi in Mld di Euro, si riferiscono all'intero traffico import Extra-UE che transita dai confini italiani.

Semestre	2013/I	2013/II	2014/I	2014/II	2015/I	2015/II	2016/I	2016/II
Importazioni	14.97	14.60	16.06	15.99	18.15	16.93	18.50	16.88

(a) Si calcolino i coefficienti della retta di regressione che descrive l'andamento delle importazioni nel tempo.

(b) Si misuri la bontà di adattamento della retta di regressione individuata al punto precedente.

(c) Si fornisca la previsione delle importazioni per il primo semestre del 2018.

$$\hat{I} = b_0 + b_1 t$$

a)Occorre trasformare la variabile tempo per facilitare il calcolo, purhè fatta coerentemente questa operazione ha come effetto di modificare l'interecetta senza modificare la pendenza.

t	1	2	3	4	5	6	7	8
Semestre	2013/I	2013/II	2014/I	2014/II	2015/I	2015/II	2016/I	2016/II
Importazioni	14.97	14.60	16.06	15.99	18.15	16.93	18.50	16.88

1 unità=1 semestre

a) $b_1 = \frac{\sigma_{It}}{\sigma_t^2}$ $\quad b_0 = \bar{I} - b_1 \bar{t}$

$\bar{I} = \frac{1}{n}\sum I_i = \frac{1}{8}(14,97+14,60+\ldots+16,88) = 16,51$

$\sigma_t^2 = \frac{1}{n}\sum t_i^2 - \bar{t} = \frac{1}{8}(1^2+2^2+\ldots+8^2) - 4,5^2 = 5,25$

$\sigma_I^2 = \frac{1}{n}\sum I_i^2 - \bar{I} = \frac{1}{8}(14,97^2+14,60^2+\ldots+16,88^2) - 16,51^2 = 1,68$

$\sigma_{It} = \frac{1}{n}\sum I_i t_i - \bar{I}\bar{t} = \frac{1}{8}(1\times 14,97+2\times 14,60+\ldots+8\times 16,88) - 4,5\times 16,51 =$

$= 2,35$

$b_1 = \frac{\sigma_{It}}{\sigma_t^2} = \frac{2,35}{5,25} = 0,4481$

$b_0 = \bar{I} - b_1 \bar{t} = 16,51 - 0,4481 \times 4,5 = 14,49$

b)bontà di adattamento quando si hanno tutte le osservazioni è meglio utilizzare il coeff di corr. lineare

$R^2 = r^2 = \frac{\sigma_{It}^2}{\sigma_I^2 \sigma_t^2} = \frac{2,35^2}{1,68\times 5,25} = 0,6267 = 62,67\%$

c)prevedere il primo semestre 2018

: Extra-UE che transita dai confini italiani.

1	2	3	4	5	6	7	8	9	10	11
)13/I	2013/II	2014/I	2014/II	2015/I	2015/II	2016/I	2016/II	2017/I	2017/II	2018/I
4.97	14.60	16.06	15.99	18.15	16.93	18.50	16.88			
1	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6

t =11 nel nostro caso

c) $\hat{Y}(t=11) = b_0 + b_1 \times 11 = 14,49 + 0,4821 \times 11 = 19,42$

ESERCIZI PROVA INTERMEDIA 2019

Esercizio n. 3 Turno A del 18/04/2019

Dai dati Istat per il 2015 a livello regionale si desume che il coefficiente di correlazione lineare tra numero di medici (M) e il numero totale di addetti nelle professioni sanitarie e infermieristiche (S) è pari a 0.9695. Sapendo che $\sum M_i = 233.102$, $\sum S_i = 330.602$, $\sum M_i^2 = 4428.999$ e $\sum S_i^2 = 8623.421$ calcolare:

(a) il valore dell'intercetta della retta di regressione di M su S;

(b) la varianza del numero di medici spiegata dal numero di addetti nelle professioni sanitarie.

$r_{ML} = 0,9695$

a) $\hat{M} = b_0 + b_1 S \qquad b_0 = \overline{M} - b_1 \overline{S}$

$b_1 = \frac{\sigma_{MS}}{\sigma_S^2}$

non si ha la covarianza, ma si ha il coefficiente di correlazione lineare e quindi si può:

$b_1 = \frac{\sigma_{MS}}{\sigma_S^2} = \frac{\sigma_{MS}}{\sigma_S \sigma_S} \frac{\sigma_M}{\sigma_M} = \left(\frac{\sigma_{MS}}{\sigma_M \sigma_S}\right) \frac{\sigma_M}{\sigma_S} = r_{MS} \frac{\sigma_M}{\sigma_S}$

calcolo le medie e poi le varianze

N non viene dato direttamente ma nel testo viene detto “a livello regionale” quindi visto che le regioni sono 20 n=20

$\overline{M} = \frac{1}{n} \sum M_i = \frac{233,102}{20} = 11,6551$

$\overline{S} = \frac{1}{n} \sum S_i = \frac{330,602}{20} = 16,5301$

$\sigma_M^2 = \frac{1}{n} \sum M_i^2 - \overline{M}^2 = \frac{4428,999}{20} - 11,6551^2 = 85,6086$

$\sigma_S^2 = 157,9268$

$b_1 = r_{MS} \times \sqrt{\frac{\sigma_M^2}{\sigma_S^2}} = 0,9695 \times \sqrt{\frac{85,6086}{157,9268}} = 0,7138$

$b_0 = 11,6551 - 0,7138 \times 16,5301 = -0,1441$

b)determinare la varianza spiegata dalla regressione

$$R^2 = \frac{\sigma_A^2}{\sigma_M^2} = 1 - \frac{\sigma_e^2}{\sigma_M^2} \qquad R^2 = \frac{\sigma_A^2}{\sigma_M^2} \Rightarrow \sigma_A^2 = R^2 \sigma_M^2 = 0,9695^2 \times 85,6086 = 80,4661$$

Esercizio n. 3 Turno B del 18/04/2019

Dai dati Istat per il 2015 a livello regionale si desume che l'intercetta della retta di regressione del numero di medici (M) sul numero totale di addetti nelle professioni sanitarie e infermieristiche (S) è pari a -0.1441. Sapendo inoltre che il coefficiente di correlazione lineare tra le due variabili è pari a 0.9695 e che $\sum_i M_i = 233.102$, $\sum_i S_i = 330.602$ e $\sum_i S_i^2 = 8623.421$ calcolare:

(a) il valore della pendenza della retta di regressione di M su S;

(b) la varianza residua della regressione.

$n = 20$ $\quad \bar{M} = 11,6551$

$\hat{M} = b_0 + b_1 S$ $\quad \bar{S} = 330,602$

$b_0 = -0,1441$ $\quad \sigma_S^2 = 157,9268$

$r = 0,9695$

a) b_1

$$b_0 = \bar{M} - b_1 \bar{S} \Rightarrow b_1 = \frac{\bar{M} - b_0}{\bar{S}} = \frac{11,6551 + 0,1441}{16,5301} = 0,7138$$

b) σ_e^2 $\quad R^2 = \frac{\sigma_A^2}{\sigma_M^2} = 1 - \frac{\sigma_e^2}{\sigma_M^2}$

$$\sigma_e^2 = (1 - R^2)\sigma_M^2$$

non possiamo calcolare la varianza totale, ma abbiamo il coefficiente di correlazione lineare, dobbiamo ricavare la varianza totate

$$R^2 = b_1^2 \frac{\sigma_S^2}{\sigma_M^2} \Rightarrow \sigma_M^2 = \frac{b_1^2}{R^2} \sigma_S^2 = \frac{0,7138^2}{0,9695^2} \times 157,9268 = 85,6080$$

$$\sigma_e^2 = (1 - R^2)\sigma_M^2 = (1 - 0,9695^2) \times 85,6080 = 5,1425$$

Esercizio n. 2 del 07/06/2019

Al fine di valutare gli effetti sulle vendite settimanali dovute alla dimensione dello spazi
la direzione marketing di una nuova catena di supermercati effettua un'analisi di regressione lineare del Fatturato (F) in funzione della Superficie espositiva (S). Di seguito vengono riportati, per ognuno dei 10 negozi presi in esame, i valori di Fatturato osservato e quelli previsti ($\hat{F}$) dal modello di regressione:

Supermercato	1	2	3	4	5	6	7	8	9	10
F	2353	1563	1817	1758	1943	2067	1807	2703	2693	2591
$\hat{F}$	2266	2186	2133	1882	1970	2055	1960	1968	2314	2561

(a) Determinare la varianza residua di regressione.

(b) Misurare la bontà di adattamento della retta di regressione.

(c) Sapendo che la varianza di S è pari a 13428, determinare il coefficiente angolare della retta di regressione.

(d) Sapendo inoltre che la media di S è pari a 358.7 si stimi il valore previsto del fatturato per un negozio con una superficie espositiva di 550 metri quadrati.

Si hanno già i valori teorici

$$\hat{F} = b_0 + b_1 S$$

a) $\sigma_e^2 = \frac{1}{n} \Sigma e_i^2$ $\qquad e_i = F_i - \hat{F}_i$

Più facile e veloce farlo con excel

	oss F	teor $\hat{F}$	residui e	residui^2 e^2	oss^2 F^2
	2353	2266	87	7569	5536609
	1563	2186	-623	388129	2442969
	1817	2133	-316	99856	3301489
	1758	1882	-124	15376	3090564
	1943	1970	-27	729	3775249
	2067	2055	12	144	4272489
	1807	1960	-153	23409	3265249
	2703	1968	735	540225	7306209
	2693	2314	379	143641	7252249
	2591	2561	30	900	6713281
somma	21295	21295	0	1219978	46956357
media	2129.5	2129.5	0	121997.8	4695636
					160865.5
					0.241616
					2.894523

$e_i = F_i - \hat{F}_i$

e_i^2

$\sigma_e^2 = \frac{1}{n} \Sigma e_i^2 = 121997{,}8$

b)bontà di adattamento, non possiamo sfruttare il coefficiente di correlazione lineare, dobbiamo ricorrere arra definizione di R quadro

	oss F	teor $\hat{F}$	residui e	residui^2 e^2	oss^2 F^2
	2353	2266	87	7569	5536609
	1563	2186	-623	388129	2442969
	1817	2133	-316	99856	3301489
	1758	1882	-124	15376	3090564
	1943	1970	-27	729	3775249
	2067	2055	12	144	4272489
	1807	1960	-153	23409	3265249
	2703	1968	735	540225	7306209
	2693	2314	379	143641	7252249
	2591	2561	30	900	6713281
somma	21295	21295	0	1219978	46956357
media	2129.5	2129.5	0	121997.8	4695636
					160865.5
					0.241616
					2.894523

$e_i = F_i - \hat{F}_i$

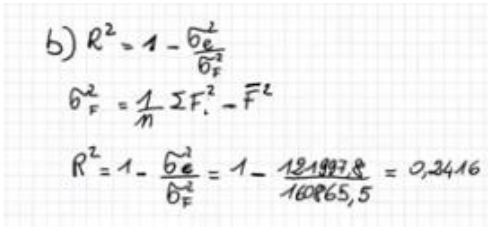
e_i^2

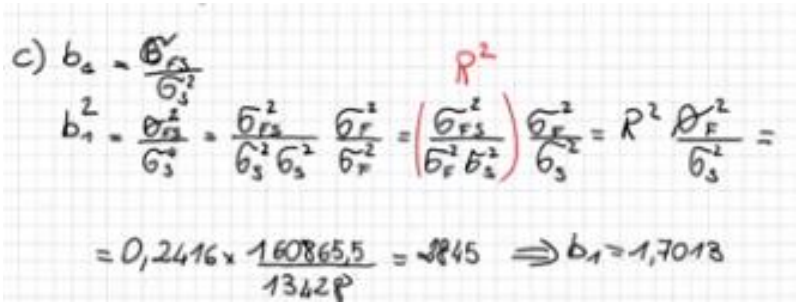
$\sigma_e^2 = \frac{1}{n} \Sigma e_i^2 = 121997{,}8$

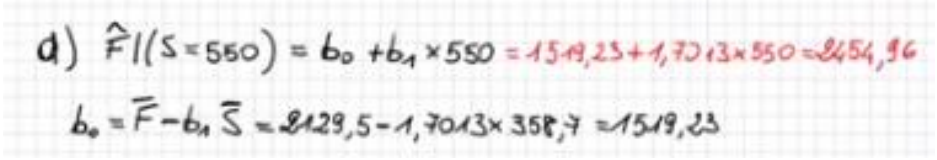
$\bar{F} = 2129{,}5$

$\sigma_F^2 = \frac{1}{n} \Sigma F_i^2 - \bar{F}^2 =$

$= 160865{,}5$







TEORIA DELLA PROBABILITA'

Dobbiamo necessariamente parlare di probabilità per quando parleremo di inferenza statistica (significa cercare di utilizzare quei metodi che ci consentono di generalizzare un risutato ossevato su un campione di unità statistiche a un insieme più grande). La probabilità entra in gioco nei meccanismi di selezione del campione (c.d. campioni probabilistici, le unità che entrano a far parte del campione vengono selezionate mediante procedure di casualizzazione).

DEFINIZIONI

Esperimento casuale/aleatori/prova → definizione molto generale, per esperimento si intende un insieme di procedure volte a produrre un risultato; l'esperimento casuale è un esperimento in cui non sono in grado di predire con certezza il risultato (prima dell'esperimento non si sa che cosa si osserverà, solo a posteriori si osserva l'esito). Il contrario di un esperimento casuale è un esperimento deterministico (ogni qualvolta che viene ripetuto produce esattamente lo stesso risultato. Esempi esperimenti casuali: lancio della moneta lancio del dado,esperimento per vaccino,somministrazione farmaco.

Si usa la parola "caso" perché anche se in linea teorica saremmo in grado di prevedere il risultato con certezza, è più facile utilizzare il calcolo delle probabilità. Esempio: il lancio della moneta si può descrivere nei termini della fisica classica (certa h rispetta al piano,forza sulla moneta,rotazione della moneta ecc..),ma l'esito finale è fortemente dipendente dalle condizioni iniziali e piccole variazioni hanno molta influenza sull'esito finale, per questo è più semplice utilizzare le regole della probabilità.

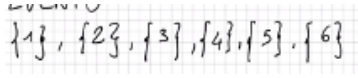
Per definizione un esperimento casuale può avere più esiti diversi:

Spazio campionario→ insieme degli esiti possibili di un esperimento indicato con S o Ω , di natura diversa a seconda dell'esperimento (lancio moneta S è fatto da testa e croce, nel lancio del dato gli elementi sono 6)

Eventi elementari →Elementi dello spazio campionario

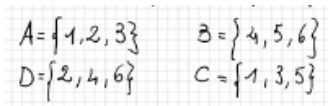
Evento → Ciascun risultato possibile dall'esperimento

Esempio: lancio dado



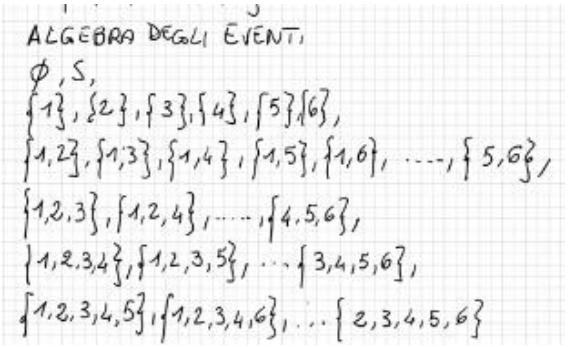
eventi elementari

Da questi si possono costruire altri eventi combinando gli eventi elementari



ALGEBRA DEGLI EVENTI → insieme di tutti i sottoinsiemi possibili dello spazio campionario

Contiene sempre l'insieme vuoto e lo spazio campionario a cui si aggiungono in primis gli eventi elementari, poi si possono formare tutti gli eventi formati da coppie di eventi elementari,poi tutti gli eventi formati da tre eventi elementari, poi formati da quattro,poi da cinque.

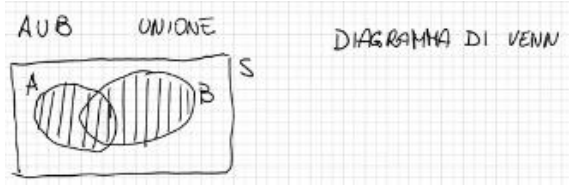


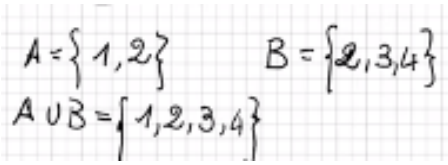
il numero degli elementi dell'algebra degli eventi solitamente è 2 elevato alla n

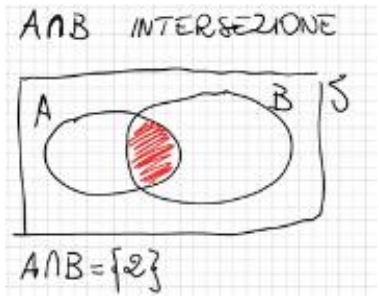
Si chiama così perché su questi eventi si possono definire una serie di operazioni:

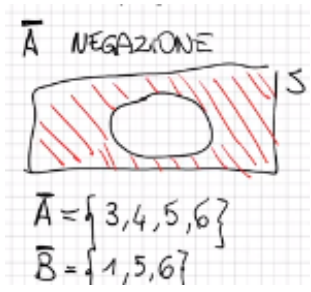
si prendono in considerazione due eventi A e B

NB: *rewind operazioni sugli insiemi*



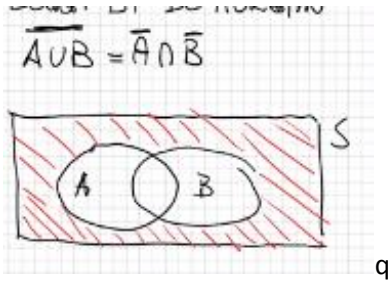




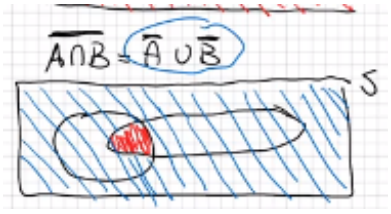


LEGGI DI DE MORGAN

lega le operazioni di unione,intersezione e negazione



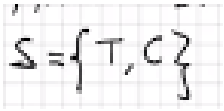
questa legge lega l'unione della negazione e l'intersezione delle negazioni



A intersecato b in rosso, la negazione è tutto il resto e può leggersi come tutto ciò che non sta dentro A e non sta dentro B (unione delle negazioni di A e B)

PROBABILITA' = sulla definizione si sono scontrati per decenni, vi sono molte definizioni nessuna esaustiva e soddisfacente.

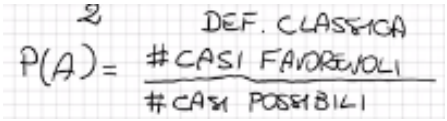
Esempio: lancio della moneta



La probabilità si applica agli eventi e si indica come P(A) nell'esempio

P(T) = 0,5 tutti diremmo subito che è 0,5 questo perché la prob è data dal rapporto tra i casi favorevoli e i casi possibili 1/2 = 1 è il n di casi favorevoli e 2 i casi possibili

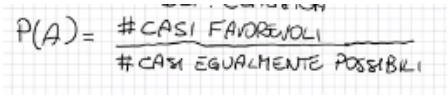
- ✓ <u>DEFINZIONE DI PROBABILITA' CLASSICA</u>



Es mazzo di carte da 40 quale prob che estraendo una carta a caso esca una figura (4 serie da 10 carte con 3 figura ciascuna) 12 figure su 40 carte la porb che esca figura è 12/40

Questa definizione va molto bene per i giochi di sorte basati su carte,estrazioni, può andare bene per il lancio della monet, ma in realtà presenta tutta una serie di problemi. Per esempio, quale è la probabilità che il genoa vinca la prossima partita di campionato? Non avrebbe senso applicare questa definizione, la prob che il genoa vinca sarà sempre 1/3 (3 risultati pareggio,vincita,perdita) è chiaro che la prob non può essere 1/3 qualunque sia la prossima partita di campionato. La definizione di probabilità classica ha 2 problemi molto grossi:

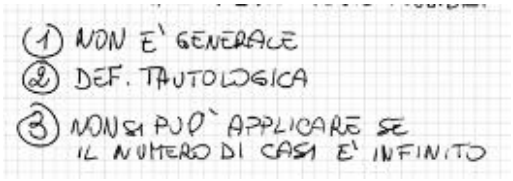
-il primo apperentemente si può correggere, ed è questo: non basta dire che i casi al denominatore sono i casi possibili, ci vuole condizione (quella per cui non si può applicare alle partite di calcio) i casi possibili devono essere egualmente possibili.(SE I CASI NON SONO EGUALMENTE POSSIBILI NON E' APPLICABILE)



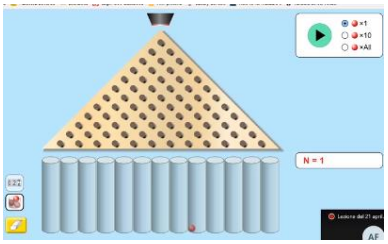
LA DEFINIZIONE NON E' GENERALE, PROBLEMA DI CARATTERE APPLICATIVO

-secondo problema collegato al primo, DI CARATTERE FILOSOFICO dire egualmente possibili significa dire egualmente probabili, ma se si usa il concetto di probabilità all'interno della definizione di probabilità non si sta, di fatto, dando nessuna definizione di probabilità (definizione circolare, definire qualcosa partendo dalla cosa stessa) vizio logico nella definizione

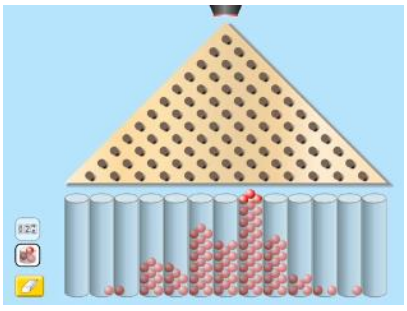
-terzo problema, sempre legato alla generalità della definizione, vi sono situazioni nelle quali i casi possibili sono in numero infinito, non si possono contare, non possiamo esprimerli in numero. La definizione NON E' APPLICABILE SE IL NUMERO DI CASI E' INFINITO



ALTRA DEFINIZIONE DI PROBABILITA'



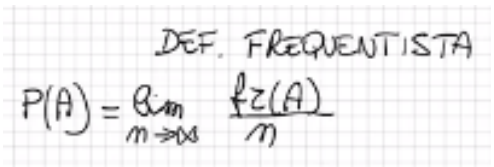
Quale è la prob che la pallina finisca nel quarto cilindro partendo da sinistra? Faccio cadere 100 palline per esempio



posso contare le palline cadute nel cilindro (9) in tutto le palline sono 100 e quindi posso approssimare la probabilità alla frequenza relativa

- ✓ <u>DEFINIZIONE FREQUENTISTA DI PROBABILITA'</u>

La probabilità è uguale al limite di n che tende ad infinito della frequenza realtiva per il numero di prove n (numero infiniteesimo di prove)



questa def ha meno problemi della def classica, è una definizione di tipo empirico, non c'è più il problema dei casi egualemnte possibili, è irrilevante perché è basata sul fatto che l'esperimento venga effettuato (a volte viene chiamata definizione di probabilità a posteriori). Questa

definizione è quella che si apllica generalmente in ambito statistico (logica proprio della statistica è la ripetizione del campionamento).

Questa definizione ha un unico limite: L'ESPERIMENTO DEVE ESSERE RIPETIBILE

(in ambito scientifico e socio economico l'esperiemento è ripetibile, ma esistono molte situazioni in cui non è così, la ripetività prevede che l'esperimento avvenga nelle stesse condizioni, usare questa definizione per una partita di calcio significa dire che la partita può essere ripetuta nelle stesse condizioni più volte e questo è ovviamente falso

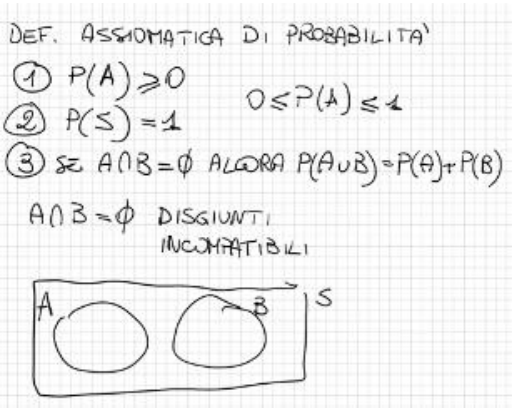
- ✓ <u>DEFINIZIONE SOGGETTIVA DI PROBABILITA' (DE FINETTI)</u>

La probabilità di un evento è la valutazione di un individuo razionale, sul verificarsi dell'evento e come tale può cambiare da individuo a individuo. Estramemente importante perché è alla base di tutto un filone della statistica. Può essere applicata a qualunque situazione (es partita di calcio ciascuno sulla base delle proprie informazione può esprimere una valutazione). Possiamo esprimere valutazioni diverse, ma devono comunque rispettare tutta una serie di regole. Il problema più grosso di questa definizione è quello che prende il nome di ELICITAZIONE occorre trovare un modo per conoscere le valutazioni di probabilità espresse da un individuo, tipicamente l'elicitazione è basata sul meccanismo della scommessa. Un altro limite riguarda la soggettività, introdurre valutazioni su questo comporta problemi non banali.

Come si è risolto quindi il problema della non univocabilità della definizione di probabilità

- ✓ <u>DEFINIZIONE ASSIOMATICA DI PROBABILITA'</u>

E' intrisincamente diverse da quelli precedenti, le prime danno una def e un modo per determinare la orbabilità. Questa definizione fornisce le regole a cui deve sottostare la probabilità. E' basata su 3 postulati



DEF. ASSIOMATICA DI PROBABILITA' (KOLMOGOROV, 1933)

-dato un evento A la probabilità è un numero non negativo

-la probabilità dello spazio campionario è pari a 1 (prob di un evento certo è 1)

I primi due assiomi dicono che la probabilità di un evento A è un numero compreso tra 0 e 1 (poi spesso definita in termini % per comodità)

-se 2 eventi non hanno nulla in comune (AintersecatoB=insieme vuoto),gli eventi sono disgiunti o incompatibili, se i due eventi sono separati allora la prob dell'unione è uguale alla somma delle probabilità degli eventi→ chiamato assioma di additività (la probabilità è una funzione additiva)

ESEMPIO: lancio di un dado

A= (1,2) B= (3)

Dalla definizione classica P(A)=2/6=1/3 P(B)=1/6