

ELEMENTI DI STATISTICA

PARTE 1

1. INTRODUZIONE	1
1.1 <i>La parola statistica</i>	1
1.2 <i>Cenni storici</i>	1
1.3 <i>Gli studiosi</i>	1
1.4 <i>La statistica moderna</i>	2
1.5 <i>Le fonti statistiche</i>	2
2. DEFINIZIONI	3
2.1 <i>Una definizione di statistica</i>	3
2.2 <i>I fenomeni collettivi</i>	3
2.3 <i>Unità statistiche, caratteri, modalità</i>	4
2.4 <i>Popolazione e campione</i>	6
2.5 <i>Le scale di misura</i>	6
3. FASI DELL'INDAGINE STATISTICA	7
3.1 <i>Piano di rilevazione</i>	7
3.2 <i>Rilevazione dei dati</i>	9
3.3 <i>Spoglio dei dati</i>	9
3.4 <i>Correzione degli errori</i>	9
3.5 <i>Elaborazione</i>	10
3.6 <i>Interpretazione</i>	10
4. SERIE STATISTICHE	10
4.1 <i>Definizione</i>	10
4.2 <i>Rappresentazione di una serie statistica</i>	10
4.3 <i>Tipi particolari di serie</i>	11
5. TABELLE STATISTICHE DI FREQUENZA	12
5.1 <i>Tabelle statistiche</i>	12
5.2 <i>Frequenze</i>	15
6. RAPPRESENTAZIONI GRAFICHE	17
ESERCIZI - 1	23
7.A. GLI INDICATORI STATISTICI	24
7.A.1 <i>Definizioni</i>	24
7.A.2 <i>Requisiti</i>	25
7.B. INDICI STATISTICI DI POSIZIONE O MEDIE LASCHE O MEDIE BASATE SULL'ORDINAMENTO DEI DATI	26
7.B.1 <i>Valori estremi</i>	26
7.B.2 <i>Valore centrale</i>	26
7.B.3 <i>Mediana</i>	26
7.B.4 <i>Proprietà della mediana</i>	29
7.B.5 <i>Quantili e quartili</i>	30
7.B.6 <i>Moda</i>	32
7.C. MEDIE ANALITICHE	33
7.C.1 <i>Media aritmetica</i>	33
7.C.2 <i>Proprietà della media aritmetica</i>	35
7.C.3 <i>Osservazioni sulla media aritmetica</i>	36

7.C.4 Caratteristiche della media aritmetica	36
ESERCIZI - 2	37
8.A. INDICI DI VARIABILITA'	37
8.B. INDICI DI VARIABILITA' BASATI SULL'ORDINAMENTO DEI DATI	39
8.B.1 Campo di variazione	39
8.B.2 Differenza interquartilica	39
8.C. MISURE DI DISPERSIONE	39
8.C.1 Devianza	40
8.C.3 Varianza	41
8.C.4 Scarto quadratico medio	41
8.C.5 Proprietà delle misure di dispersione	42
ESERCIZI - 3	42
9. CONCENTRAZIONE	42
9.1 Concetto e misura della concentrazione (nel caso di serie)	42
9.2 La curva di Lorenz	45
9.3 Indici di concentrazione	45
ESERCIZI - 4	46
11. SIMMETRIA E CURVA NORMALE	47
11.1 Simmetria e asimmetria	47
11.2 Box plot	48
11.3 La curva normale	48
ESERCIZI - 5	50
ESERCIZI RIEPILOGATIVI - A	51

PARTE 2

12.A ANALISI BIVARIATA	53
12.A.1 <i>Introduzione: serie e tabelle doppie</i>	53
12.A.2 <i>La connessione</i>	54
12.B. INDIPENDENZA E DIPENDENZA IN DISTRIBUZIONE (CONNESSIONE) PER CARATTERI QUALITATIVI	55
12.B.1 <i>Indipendenza in distribuzione</i>	55
12.B.2 <i>Connessione in distribuzione</i>	56
12.B.3 <i>Indici di connessione: L'indice Chi-quadrato di Pearson</i>	57
ESERCIZI - 6	59
12.C. REGRESSIONE PER DISTRIBUZIONI DOPPIE MISTE:	60
12.C.1 <i>Indipendenza in distribuzione</i>	60
12.C.2 <i>Dipendenza in media: il rapporto di correlazione η^2 di Pearson</i>	60
ESERCIZI - 7	63
12.D. CORRELAZIONE (=INTERDIPENDENZA) PER VARIABILI STATISTICHE DOPPIE (DISTRIBUZIONI PER CARATTERI QUANTITATIVI)	63
12.D.1 <i>Covarianza (misura assoluta di interdipendenza)</i>	63
12.D.2 <i>Coefficiente di correlazione di Bravais - Pearson (misura relativa di interdipendenza)</i>	64
ESERCIZI - 8	65
12.E. DIPENDENZA FRA VARIABILI QUANTITATIVE: REGRESSIONE LINEARE	66
12.E.1 <i>Indice di determinazione R^2</i>	67
ESERCIZI - 9	69
ESERCIZI RIEPILOGATIVI - B	70
ESERCIZI GENERALI	71

1. INTRODUZIONE

1.1 La parola statistica

L'etimologia della parola "Statistica" deriva dal vocabolo italiano "Stato" e fa riferimento, nella quasi totalità dei linguaggi europei, alla constatazione per cui le prime informazioni su fenomeni reali sono state raccolte ed organizzate ad opera degli *organismi statali* che ne erano anche i principali utilizzatori. La prima apparizione del vocabolo "statistica" inteso come raccolta di informazioni organizzate e gestite dallo Stato sembra essere quella dell'italiano Ghislini che, nel 1589, indica la Statistica come "descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato".

1.2 Cenni storici

La statistica è sorta in tempi antichissimi, fin dai primi insediamenti umani aventi una semplice organizzazione sociale. Si trovano documenti di rilevazioni di persone e di terreni nei nuraghi sardi e nei monumenti egiziani più antichi. Esistono notizie relative a rilevazioni statistiche fatte eseguire dall'imperatore cinese Yu, più di 4000 anni fa, allo scopo di ottenere notizie precise sulla situazione dell'agricoltura in ogni provincia, e quindi di poter ripartire equamente le imposte. Presso i Romani, Servio Tullio istituì la prima forma di censimento, chiamato allora "Census", che era effettuato ogni cinque anni e serviva a conoscere il numero dei cittadini, l'ammontare dei loro beni, l'andamento delle nascite e delle morti.

Possiamo però affermare che la statistica, come disciplina a sé stante e, come già detto, inizialmente finalizzata all'espressione di fenomeni riguardanti gli Stati, sia nata solo nel secolo XVII, in seguito alle grandi scoperte matematiche. Nel XVIII e nel XIX secolo, grazie all'introduzione di metodi matematici e al "calcolo delle probabilità" la statistica ha avuto notevoli impulsi e il suo campo delle applicazioni si è andato ampliando quando si è capito come utilizzare un insieme d'informazioni allo scopo di ricercare le cause del manifestarsi di taluni fenomeni.

Tale disciplina ha poi avuto una svolta decisiva quando ha aggiunto all'aspetto descrittivo anche l'aspetto investigativo, ottenuto mediante la messa in relazione dei fatti osservati con altri che ne possano essere causa o conseguenza.

Con questo nuovo approccio il campo dell'indagine statistica si è ampliato in modo eccezionale e i metodi statistici hanno trovato applicazione in moltissime discipline: economia, sociologia, fisica, biologia, genetica, psicologia, eccetera.

1.3 Gli studiosi

Fra i numerosi studiosi che hanno contribuito allo sviluppo della statistica possiamo ricordare il belga Adolfo Quetelet (1796-1874) che ha sostenuto il principio secondo cui: "le leggi che governano la Società sono fisse e immutabili, come quelle che governano i corpi celesti ed esistono fuori dal capriccio degli uomini". Numerose sono state, nel corso del tempo, le definizioni della statistica: c'era chi si limitava a considerarla solo un metodo, chi invece una scienza. Attualmente si possono dare varie definizioni di statistica e fra esse ci pare interessante quella proposta da B. Giardina il quale sostiene che la statistica in senso moderno sia <<l'applicazione dei metodi scientifici alla programmazione della raccolta dei dati, alla loro classificazione, elaborazione, analisi e presentazione e all'inferenza di conclusioni attendibili da essi>>.

Ma si possono ricordare anche nomi come quello di Cournot, indagatore dei fatti economici vissuto in Francia e facente grande uso anche del calcolo di probabilità e Grigor Mendel che applicò le osservazioni statistiche al campo della biologia.

Fu però Gustav Rumelin a dare una sistemazione della statistica. Egli individuò una distinzione fondamentale tra una statistica, che definì e che ancora oggi è definita, *metodologica*, che fa da

base e dà le linee direttrici di metodo per la raccolta dei dati, la loro elaborazione matematico probabilistica e l'interpretazione dei risultati, da una statistica *applicata* ai vari rami della conoscenza e della scienza, che assume denominazioni proprie a seconda dei campi di applicazione, come ad esempio la demografia, la biometria, l'antropometria e via dicendo. In particolare la statistica ha avuto grande applicazione nel campo degli studi biologici, ad opera di Ronald Fisher e di Karl Pearson, i cui studi sono stati presi come esempio anche nelle altre discipline statistiche per via della loro generalità e importanza metodologica.

In Italia i maggiori statisti sono stati Rodolfo Benini, Corrado Gini, Giorgio Mortara, Marcello Boldrini e Livio Livi.

1.4 La statistica moderna

Oggi la statistica è ampiamente utilizzata pressoché in tutti i campi di studio. Se ne fa sempre maggiore utilizzo nel campo delle ricerche di mercato, della pianificazione territoriale, in campo produttivo ed economico in generale.

Con la nascita dei grandi Stati europei, si attribuisce all'analisi statistica dei fenomeni collettivi un interesse pubblico che spinge progressivamente le nazioni occidentali a dotarsi di Istituti "centrali" di Statistica, deputati per legge alla raccolta, organizzazione e diffusione di dati sulla popolazione, sulle abitazioni, sulle risorse economiche e su tutti gli aspetti rilevanti della vita collettiva di una nazione, di una Comunità di stati (Unione Europea) o dell'intero pianeta (Nazioni Unite).

Oggi, gli organismi pubblici che istituzionalmente raccolgono e diffondono informazioni statistiche sono innumerevoli ed agiscono secondo una gerarchia di competenze che individua nell'Ente locale la sede prioritaria di raccolta del dato elementare, mentre la verifica, l'aggregazione e la pubblicazione sono di competenza dell'Ente centrale. In Italia tale ente è l'ISTAT (*Istituto Nazionale di Statistica*).

1.5 Le fonti statistiche

L'ISTAT è dipendente dalla Presidenza del Consiglio dei Ministri ed istituzionalmente accentra tutta la pubblicazione Statistica nazionale (www.istat.it), provvedendo a curare e diffondere il materiale raccolto in pubblicazioni a periodicità mensile, annuale, decennale o anche occasionale. Le rilevazioni contenute in tali pubblicazioni possono poi essere a carattere generale o relative a specifici campi di attività. Così, accanto all'Annuario Statistico Italiano e ai Bollettini mensili che sono a carattere generale, vi sono gli annuari dedicati alle nascite e ai decessi, pubblicazioni in campo sanitario e familiare. Accanto a queste vi sono poi le statistiche della Pubblica Amministrazione, le statistiche giudiziarie civili e penali, statistiche nel campo dei conti nazionali e statistiche sul tema del lavoro. Altre pubblicazioni esistono, poi, relativamente a tutte le attività produttive.

L'ISTAT non è, però, l'unica organizzazione che cura e gestisce l'ingente mole delle fonti statistiche nazionali. Ogni anno il Parlamento pubblica la *Relazione sulla situazione economica del Paese*, il Ministero del tesoro cura il *Bilancio di Previsione dello Stato*, la Banca d'Italia pubblica il *Bollettino statistico* e tante altre pubblicazioni sono effettuate dalle Camere di Commercio, Confindustria e altri enti come Enel, Rai e Ferrovie dello Stato.

A livello internazionale meritano un cenno le statistiche dell'ONU, del BIT (*Bureau international du travail*), quelli della OMS (*Organizzazione Mondiale della Sanità*), dell'UNESCO e della FAO. Infine va citato l'Istituto statistico delle comunità europee, l'EUROSTAT, che produce statistiche relative ai paesi aderenti all'unione.

2. DEFINIZIONI

2.1 Una definizione di statistica

La statistica può essere, quindi, definita come un insieme di metodi per lo studio dei *fenomeni collettivi* con lo scopo di metterne in luce le regolarità nascoste attraverso lo studio del modo con cui si sono manifestati i singoli casi, o dai risultati apparsi nei vari esperimenti.

Tutti gli studi statistici si basano sullo studio dei fenomeni e sulla individuazione dei fatti che determinano tali fenomeni al fine di determinare quali elementi possono essere considerati cause principali e quali sono le relazioni intercorrenti tra tali fatti. Ciò consente di individuare, ed isolare, le cause accidentali e di costruire dei modelli relativi al fenomeno preso in considerazione. Attraverso tali modelli è, poi, possibile effettuare delle previsioni in merito ai fenomeni stessi.

In genere si parla di una *statistica descrittiva*, cioè quel genere di analisi che si limita ad osservare le regolarità dei fenomeni, da una *statistica inferenziale*, che si propone di individuare i legami tra i fatti ed i fenomeni fino alla formulazione di un modello matematico che esprima l'andamento del fenomeno stesso.

2.2 I fenomeni collettivi

In generale, per fenomeno intendiamo tutto ciò che capita intorno a noi o che noi stessi provochiamo e che sia percepito e constatato. Possiamo distinguere fra fenomeni naturali (come il clima, il tramonto del sole, le onde del mare) e fenomeni riprodotti in laboratorio. Tutti i fenomeni che si presentano costantemente con le stesse caratteristiche sono chiamati *fenomeni tipici*. Tale è ad esempio la caduta di un corpo che, abbandonato ad una certa altezza, cade verticalmente verso il basso a causa della forza di gravità terrestre.

D'altra parte esistono dei fenomeni che si manifestano ogni volta con caratteristiche diverse e per i quali è difficile fare delle previsioni sul loro comportamento. Pensiamo, ad esempio, ai fenomeni meteorologici, che non sempre permettono di fare in anticipo delle previsioni sicure sulle condizioni del tempo nei giorni successivi, e che quindi possono essere definiti *fenomeni atipici*.

Se consideriamo, invece, i *fenomeni sociali* quali ad esempio le nascite, i matrimoni, le migrazioni, possiamo affermare che non è possibile stabilire delle leggi generali, se limitiamo il nostro studio ad un singolo caso, come invece avviene per i fenomeni tipici. Possiamo però affermare che se si effettuano delle osservazioni molto numerose su tali fenomeni, essi rivelano determinate caratteristiche uniformi, per cui si può concludere che, pur essendo singolarmente atipici, presentano, considerati collettivamente, una tipicità di comportamento che ci permette di studiare le leggi che li governano. I fenomeni di questo tipo sono chiamati *fenomeni collettivi*.

I fenomeni sociali sono innumerevoli e per tale motivo è sorta l'esigenza e l'utilità di studiare autonomamente gruppi di essi. Da cui le statistiche applicate, già individuate da Gustav Rumelin, quali la *Demografia*, che studia i fenomeni riguardanti le caratteristiche strutturali e dinamiche delle popolazioni umane; la *Statistica Sanitaria*, che studia quantitativamente i fenomeni legati al settore della sanità, relativamente sia alle strutture che al loro utilizzo e in relazione ai fenomeni patologici che colpiscono l'uomo; la *Biometria* che studia i fenomeni biologici degli organismi viventi, la *Statistica Giudiziaria*, la quale studia quantitativamente i fenomeni che ricadono nella sfera del diritto violato e che costituiscono oggetto di conoscenza delle Forze dell'ordine, dell'Autorità Giudiziaria, degli Istituti di prevenzione e Pena e di alcuni uffici amministrativi, o che derivano dall'attività della magistratura; la *Statistica Economica*, che studia i fatti cosiddetti economici, quali l'andamento dei prezzi, dei consumi, delle produzioni, eccetera.

2.3 Unità statistiche, caratteri, modalità * → vedi pag. seguente

I casi individuali oggetto dell'osservazione sono definiti *unità statistiche*. L'unità statistica è il più piccolo elemento sul quale si effettua un'osservazione. È, ad esempio, unità statistica ogni macchina prodotta da una fabbrica, ogni abitante di una data provincia, ogni alunno di una scuola. L'unità statistica può essere *semplice*, se corrisponde ad una singola persona o ad un oggetto (ad esempio, età della popolazione italiana, cilindrata delle automobili), *composta*, se è formata da un insieme di elementi (ad esempio, nuclei familiari, lotti di produzione).

Il risultato di un'operazione compiuta sulle unità statistiche (ad esempio, numero delle persone di sesso maschile nella popolazione italiana, prezzo medio di un certo bene) si definisce *dato statistico* o *carattere* ed è, in definitiva, un'informazione sul fenomeno che si vuole studiare. Il dato statistico può rappresentare o il numero delle unità statistiche che hanno in comune una caratteristica fissata (ad esempio, il numero degli alunni di una scuola, il numero di matrimoni in una certa regione) oppure un numero ricavato da più unità statistiche che serve a rappresentare un aspetto del fenomeno studiato (ad esempio, il reddito globale degli abitanti di una provincia, il loro reddito pro capite, il volume delle importazioni di un dato bene).

Si definisce *intensità* il numero che esprime l'ammontare, la misura o la grandezza di un carattere quantitativo di una unità statistica.

Modalità sono le diverse intensità o i diversi attributi che un carattere può assumere.

Le modalità secondo cui sono classificate le unità statistiche possono essere qualitative o quantitative. Le *modalità quantitative* sono espresse da numeri risultanti da misurazioni o da enumerazioni, come, ad esempio, la rilevazione dei redditi di una popolazione, dalle altezze dei militari di leva, del numero dei vani degli alloggi di un comune e così via. Le modalità quantitative possono essere continue o discrete; *continue*, se sono espresse da numeri reali e possono assumere tutti i valori di un intervallo (ad esempio, altezze, pesi), *discrete*, in caso contrario (ad esempio, numero dei vani delle abitazioni, numero degli addetti in un settore industriale).

Le *modalità qualitative* sono invece espresse da attributi, espressioni verbali, come, ad esempio, la rilevazione della popolazione italiana secondo lo stato civile, la rilevazione delle automobili prodotte nei vari mesi di un anno, la rilevazione della produzione di cereali nelle diverse regioni italiane. Le modalità qualitative si definiscono *ordinabili*, quando è possibile ordinarli secondo un criterio logico, o *sconnesse*, quando invece non è possibile ordinarle in alcun modo.

Si dice *frequenza* il numero che esprime quante volte una data modalità del carattere si presenta nella totalità delle unità rilevate.

***Unità statistiche, caratteri, modalità**

- UNITA' STATISTICA: è l'elemento di base della popolazione su cui si vuole effettuare la rilevazione



Sulla unità statistica vengono osservati uno o più caratteri



- CARATTERE: è la caratteristica oggetto di studio cioè l'aspetto che della unità statistica si vuole studiare.



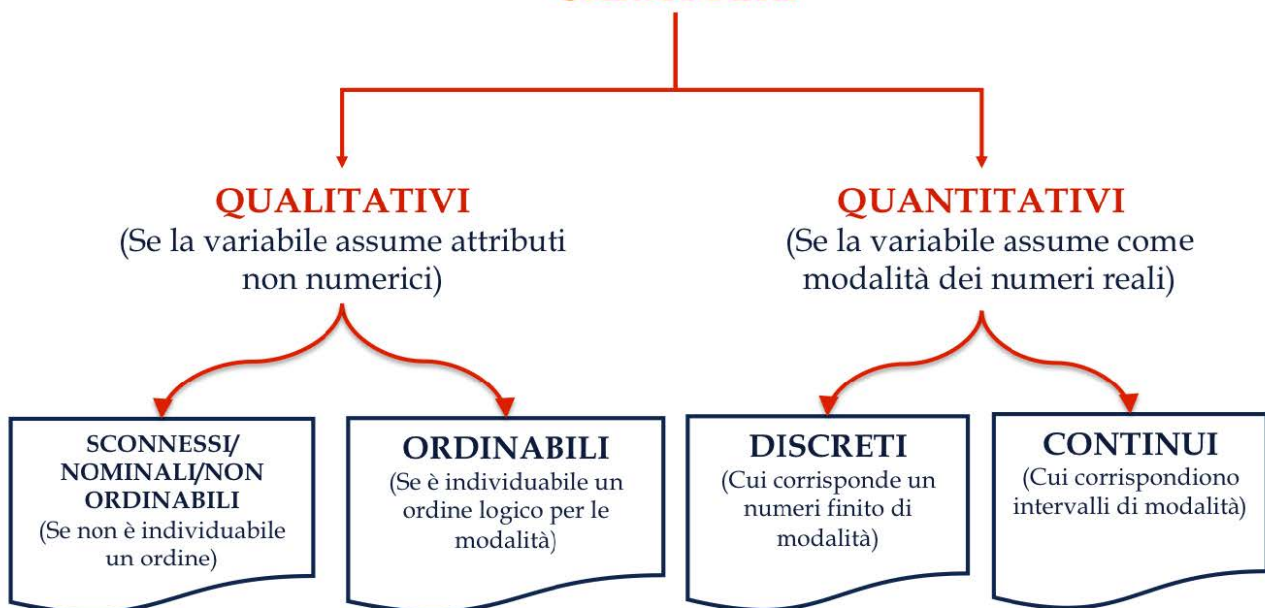
Ciascun carattere si esprime attraverso modalità



- MODALITA': è il numero (per i caratteri quantitativi) o l'attributo (per i caratteri qualitativi) che l'unità statistica manifesta.

* * * *

CARATTERI



La distinzione fra variabili discrete e continue è concettuale.

Talune variabili teoricamente continue possono essere considerate discrete a causa della limitatezza degli strumenti di misura che li rendono di fatto discreti.

Quelle variabili, invece, concettualmente discrete che si presentano con modalità numerose e vicine fra loro, conviene considerarle continue.

2.4 Popolazione e campione

Un insieme finito di dati statistici tra loro omogenei per quanto riguarda una o più caratteristiche è definita *popolazione statistica*. Rappresenta una popolazione statistica, ad esempio, l'insieme degli alunni di una scuola, o l'insieme delle stature degli alunni di una certa classe. Un *universo* è "una popolazione statistica composta da un numero infinito di elementi", mentre un *campione* è "un complesso di osservazioni effettive sugli elementi di una popolazione o universo". Ad esempio, è un universo l'insieme formato dai risultati di un numero illimitato di lanci di un dado, o l'insieme degli esperimenti. Costituisce un campione, ad esempio, l'insieme formato da un certo numero di studenti estratti dalla popolazione scolastica di una città, oppure l'insieme di un certo numero di unità di un bene economico, scelte dalla produzione in base a determinati criteri. La scelta del campione statistico rappresenta uno degli elementi cruciali dello studio di un fenomeno.

2.5 Le scale di misura

La classificazione delle modalità dei caratteri avviene attraverso uno strumento formale che è la SCALA DI MISURA. Le scale di misure seguono un ordine che è legato alla quantità di informazioni che esse forniscono e perciò sono ordinate da quella che fornisce il minor numero di informazioni a quella che ne fornisce il maggior numero. Inoltre ogni scala contiene sempre tutte le informazioni della precedente.

- *SCALA DICOTOMICA*: Si usa per la classificazione dei caratteri con due sole modalità di cui una è complementare dell'altra. Fra le modalità in questo caso è possibile stabilire una sola relazione e cioè quella di diversità (Esempio sesso).

- *SCALA NOMINALE*: Si usa per caratteri con più di due modalità (in genere qualitativi nominali o sconnessi). Si tratta dei caratteri le cui modalità non possono essere ordinate in alcun modo e per i quali l'unico tipo di studio è connesso alla verifica della presenza o assenza del carattere e alla loro uguaglianza o diversità (Esempio colore degli occhi).

- *SCALA ORDINALE*: Si usa per quei caratteri (qualitativi) le cui modalità o attributi possono essere ordinati secondo un particolare artificio di natura logica o formale in ordine crescente o decrescente. Per questi caratteri è possibile individuare oltre alla loro diversità anche una relazione d'ordine. Non è invece possibile effettuare operazioni né di natura additiva che moltiplicativa (Esempio mesi dell'anno).

- *SCALA DI CONTEGGIO*: Questa scala si usa per i caratteri quantitativi per i quali lo zero è convenzionale (per esempio la temperatura). In tal caso non ha senso rapportare le misure ottenute, ed è, invece, corretto confrontare per differenze.

- *SCALA DI MISURE IN SENSO STRETTO*: anche questa scala si usa per soli caratteri quantitativi, ma per quelli continui. In questo caso si parla di misure in senso stretto poiché esiste uno zero naturale (per esempio il peso, l'altezza eccetera). In tal caso è possibile definire degli intervalli ed anche elaborare le informazioni tramite funzioni.

3. FASI DELL'INDAGINE STATISTICA:

Lo studio di un fenomeno collettivo con metodo statistico è definito *indagine statistica*. Finalità specifica di un'indagine statistica è quella di risolvere un problema sociale specifico e di interesse immediato, volendo con il termine immediato distinguere questo tipo di ricerca da quella cosiddetta pura, volta cioè alla verifica di ipotesi che contengono concetti astratti e non immediatamente utilizzabili. Il lavoro necessario per effettuare una analisi statistica deve essere suddiviso in fasi successive.

- a) Per prima cosa occorre predisporre il **Piano della rilevazione** che definisce chiaramente gli *obiettivi* da raggiungere attraverso l'indagine, l'*oggetto* della rilevazione e il *modo* di raccolta dei dati. Questa fase è la più delicata perché, se non è preparata in modo accurato, può pregiudicare la riuscita dell'intera indagine.
- b) Quindi si procede alla **Rilevazione**, ovvero quel complesso di operazioni finalizzate alla raccolta dei dati, cioè delle modalità o dei caratteri delle unità del collettivo statistico preso in esame.
- c) Al termine della rilevazione, subentra la fase dello **spoglio**, dei dati, che consiste nell'enumerazione e classificazione dei dati raccolti, secondo il piano prestabilito e gli obiettivi da raggiungere. I dati sono quindi raggruppati secondo le modalità di un carattere qualitativo, o secondo le intensità di un carattere quantitativo. Lo spoglio può essere fatto manualmente da personale opportunamente addestrato o con l'aiuto di PC.
- d) Prima di procedere all'elaborazione vera e propria, i dati vanno corretti per eliminare gli **errori** eventualmente commessi nella fase di rilevazione.
- e) I dati così raccolti ed organizzati vengono **elaborati**.
- f) I risultati dell'indagine svolta vanno quindi **interpretati**.

Vediamo le fasi della rilevazione punto per punto:

3.1 Piano di rilevazione

Il soggetto che deve pianificare la rilevazione deve innanzi tutto definire lo scopo della rilevazione fino nei minimi dettagli e i mezzi finanziari e umani di cui disporre in modo che non vi siano ambiguità.

- Quindi occorre definire l'unità statistica e l'unità di rilevazione. Per unità statistica si intende l'unità oggetto del collettivo che si vuole analizzare (m ad esempio se il collettivo da analizzare è la popolazione, l'unità statistica oggetto della rilevazione è ragionevolmente l'individuo, ma si può decidere di considerare come unità la famiglia). Definire l'unità di rilevazione significa invece definire particolari accorgimenti nella scelta delle unità statistiche da considerare (se ad esempio l'unità statistica è rappresentata dalla famiglia occorre ben definire il significato che si vuol dare a questa parola e cioè stabilire se l'indagine dovrà comprendere anche i conviventi, le persone che vivono da sole ecc).

Bisogna inoltre evitare la sovrainformazione, evitando quelle informazioni che non sono strettamente necessarie allo svolgimento della analisi o quelle difficili da rilevare. Bisogna inoltre adeguare la mole di informazioni alle risorse finanziarie e al tempo disponibile

- Definite le scale di misurazione adottabili, il piano di rilevazione dovrà indicare i mezzi tecnici di osservazione e cioè: a) *supporti* da utilizzare per la rilevazione. Va indicato cioè se la stessa sarà effettuata mediante questionari, modelli o schede e se necessario un operatore, la persona o

le persone che si occuperanno della rilevazione. b) *personale* addetto. A tal proposito distinguiamo le rilevazioni automatiche, in cui sono le unità statistiche a "recarsi" dall'operatore (è il caso delle nascite o delle morti) da quelle riflesse in cui i dati sono raccolti da personale addetto e appositamente addestrato. c) gli strumenti necessari.

- A questo punto della formulazione del piano occorrerà, prima di procedere, stabilire i limiti territoriali della rilevazione e quelli temporali. Per quanto riguarda i *limiti di tempo*, questi possono far riferimento ad un istante preciso (come avviene nei censimenti per i quali è necessario che tutte le informazioni forniscano la situazione esistente ad una certa ora di un dato giorno), o a un dato periodo di tempo (anno, stagione, mese, settimana o altro). La scelta deve essere fatta tenendo conto di numerosi fattori: in estate sono notevoli gli spostamenti delle famiglie per le ferie e quindi, per certe rilevazioni di carattere generale, si preferisce fissare l'epoca per la raccolta dei dati in un periodo compreso fra la fine del mese di aprile e i primi di giugno. Riguardo ai *limiti di spazio*, la rilevazione può essere fatta su un territorio più o meno vasto. Ci sono indagini condotte su numerose nazioni contemporaneamente, altre eseguite a livello regionale, provinciale o comunale. La scelta dipende, come affermato più volte, dagli obiettivi che si vogliono raggiungere. Limiti di spazio sono anche quelli riferiti a particolari *categorie* dell'oggetto dell'indagine; per esempio, volendo esaminare le variazioni relative al numero degli iscritti nelle scuole italiane nei vari anni, ci si può limitare agli istituti tecnici, oppure alle scuole pubbliche. Tali statistiche saranno valide limitatamente al territorio per cui sono state fatte; un'indagine nelle scuole pubbliche del Lazio non sarà valida per le altre regioni italiane a causa dei numerosi fattori, diversi in ogni regione, che possono influire sull'andamento delle iscrizioni.

- Nella predisposizione di un piano di rilevazione occorre porre sempre particolare attenzione alla scelta della *ampiezza della rilevazione*, ovvero alla scelta della esecuzione di una rilevazione totale o parziale. Una rilevazione si dice *totale* se è estesa a tutti gli elementi di una popolazione statistica, mentre ridefinisce *parziale* o *per campione* se è fatta solo su alcuni elementi della popolazione, scelti con criteri opportuni.

Campionamento

Indubbiamente una rilevazione totale ha il pregio di fornire risultati completi e precisi sul fenomeno indagato; tuttavia se la popolazione è molto numerosa, comporta costi elevati e tempi di realizzazione molto lunghi. Nella realtà si ricorre molto spesso a indagini parziali, dette *rilevazioni campionarie*, quando occorre contenere i costi (il costo di un'indagine campionaria è minore di quello di un'indagine totale), quando occorre ottenere più tempestivamente i risultati, infine, quando non è possibile procedere diversamente (ad esempio gli studi sulla durata dei prodotti industriali comportano inevitabilmente la distruzione delle unità esaminate).

Le rilevazioni campionarie devono soddisfare l'importante requisito di dare un'immagine fedele, sebbene ridotta, del fenomeno preso in considerazione.

La statistica moderna fa ampissimo uso del campionamento per la circostanza che la rilevazione campionaria consente riduzioni in termini di spesa e di tempo impiegato per lo svolgimento delle analisi, così permettendo di aumentare il numero delle analisi stesse alimentando la produzione statistica nazionale.

La scelta del campione deve essere fatta in modo che esso sia *rappresentativo*, ovvero che sia estratto con un criterio oggettivo e possieda le stesse caratteristiche della popolazione da esaminare. Quando invece la rilevazione tende ad escludere quelle unità che presentano il carattere con intensità inferiore a determinati limiti o entro limiti determinati si parla di campioni non rappresentativi. Ad esempio, se un lotto di 2000 pezzi fosse formato da 300 pezzi difettosi e da 1700 pezzi non difettosi, un campione ideale di 100 pezzi dovrebbe contenere 15 pezzi difettosi e 85 pezzi non difettosi, ma difficilmente un campione estratto da quel lotto ha questa composizione. Altro elemento fondamentale per una rilevazione campionaria è che il

campione sia *sufficientemente ampio*, infatti più il campione è numeroso, più i risultati che fornisce sono significativi.

Le metodologie di campionamento sono raggruppabili in due grandi classi:

- a) campionamento casuale, in cui ciascuna unità della popolazione ha la stessa probabilità di essere estratta
- b) campionamento mediante scelta ragionata. In questo caso le unità sono scelte con metodi non del tutto casuali e pertanto non hanno la stessa probabilità di essere inserite nel campione

3.2 Rilevazione dei dati

La raccolta dei dati si può distinguere:

- riguardo al *modo*, in *automatica* se sono gli stessi interessati che forniscono spontaneamente, spesso in virtù di precise disposizioni di legge, le informazioni sui fenomeni che li riguardano (è quanto accade giornalmente negli uffici anagrafici dove sono denunciate le nascite, le morti, i trasferimenti), o *riflessa* se è fatta direttamente da organi preposti a tale scopo, come avviene per esempio nei censimenti e nelle ricerche di mercato.
- Riguardo alla *durata*, in: *continue*, se sono fatte senza interruzione (ad esempio, rilevazione anagrafica dei nati, dei morti, dei matrimoni), *periodiche*, se sono fatte ad intervalli regolari di tempo (come i censimenti con periodicità decennale, le rilevazioni delle forze di lavoro con periodicità trimestrale) o *occasionali*, se sono fatte una volta tanto per indagini su fenomeni particolari (ad esempio, le rilevazioni dei danni di un terremoto, le indagini di mercato per il lancio di un prodotto).

Anche i fenomeni sono distinguibili in: *fenomeni di stato* se riferiti ad istanti di tempo (ad esempio la popolazione) e *di flusso* se occorre analizzarli per intervalli di tempo (come le nascite, le migrazioni o le morti).

3.3 Spoglio dei dati

Una volta effettuata la rilevazione delle unità statistiche si può procedere alla formazione dei dati statistici, attraverso l'operazione di spoglio delle unità stesse.

- Lo spoglio è la prima fase del processo di elaborazione dei dati statistici e consiste nell'*enumerazione*. Attraverso l'enumerazione le schede, i moduli, i questionari sono contati e, nello stesso tempo, controllati, al fine di fornire i primi dati sull'estensione della rilevazione.
- Successivamente si procede alla *classificazione*, ovvero nel raggruppamento dei dati ottenuti in categorie, classi, a seconda delle intensità o delle qualità dei caratteri rilevati. Ad esempio, in un'indagine relativa ad un certo istituto tecnico potranno in un primo momento formarsi classi diverse comprendenti gli alunni del triennio e quelli del biennio; si potranno ancora distinguere altre due categorie: maschi e femmine e così via.
- Infine si effettua lo *spoglio materiale*, riportando i dati, già enumerati e classificati, in prospetti chiamati *tabelle di spoglio*.

3.4 Correzione degli errori

Prima di concludere la fase di spoglio e passare alla fase di vera e propria analisi occorre depurare i dati da eventuali errori.

(*tipi di errori...*) Va precisato che per errori non si intendono soltanto eventuali errori materiali che possono essere stati compiuti nella fase di rilevazione; affette da errori sono quelle rilevazioni riferite a fenomeni perturbati da cause varie quali imperfezioni nei questionari o degli strumenti, omissioni del soggetto che effettua la rilevazione o del soggetto cui la

rilevazione si riferisce, o ancora, in fase di spoglio dei dati, possono essere dovute ad una erronea interpretazione di talune risposte o dati, ad errori di trascrizione o di arrotondamento dei valori.

Gli errori possono poi essere *accidentali*, *costanti*, *sistematici* o di *distribuzione* a seconda che siano legati a cause fortuite ed occasionali (come nel caso di errori di misura o di stima), che si ripetano con la stessa intensità in tutte le osservazioni, che agiscano secondo particolari leggi o che infine alterino la valutazione dell'esatta distribuzione di un fenomeno.

(...e loro correzione). La correzione a posteriori degli errori di rilevazione può essere effettuata utilizzando diverse metodologie.

In linea generale si cerca di predisporre la rilevazione in maniera quanto più precisa possibile al fine di evitare il presentarsi di errori di qualunque genere

3.5 Elaborazione

Questa fase comprende tutte le possibili analisi statistiche utili per lo studio del fenomeno considerato, attraverso la applicazione di particolari procedimenti matematici che hanno la funzione di trasformare i dati raccolti, considerati *grezzi*, in dati *elaborati* che mettono in evidenza le caratteristiche del fenomeno collettivo da studiare.

3.6 Interpretazione

Interpretare i dati significa esaminarli al fine di verificare se l'obiettivo prefissato è stato raggiunto, se esistono relazioni con altri fenomeni, se è possibile fare una previsione sulla tendenza del fenomeno nel tempo.

4. SERIE STATISTICHE

4.1 Definizione

Una serie statistica è un insieme di dati posti consecutivamente l'uno all'altro rappresentanti ognuno un dato, cioè l'entità del carattere così come osservata sulla singola unità statistica.

Esempio Serie statistica del carattere "numero di componenti della famiglia", riferito ai pazienti afferenti ad un medico di medicina generale:

3 2 4 3 3 5 3 6 ... 3 ... 2

4.2 Rappresentazione di una serie statistica

una serie statistica si indica simbolicamente nel modo seguente

$$\{x_i\} \quad i = 1, 2, \dots, N$$

dove l'indice i serve ad individuare, nella serie dei dati, quello specifico dato che occupa la posizione i indicata.

$$x_1 = 3 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 3 \quad x_5 = 3 \quad x_6 = 5 \quad x_7 = 3 \quad x_8 = 6 \quad \dots \quad x_i = 3 \quad \dots \quad x_N = 2$$

Una serie può essere rappresentata orizzontalmente o verticalmente

Serie non decrescenti

spesso può aversi interesse ad ordinare i dati in ordine non decrescente. Si origina in tal modo una Serie non decrescente che si rappresenta con l'indice racchiuso fra parentesi:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \dots \leq x_{(N)}$$

Dove l'indice (i) sta ad indicare il dato che occupa la i -esima posizione di ordine e dove la presenza del segno uguale sta ad indicare che tutti i dati con lo stesso valore vanno considerati e non presi una sola volta. La serie non decrescente sarà allora indicata come segue:

$$\{x_{(i)}\} \quad i = 1, 2, \dots, N,$$

nel nostro esempio:

$$x_{(1)} = 2 \quad x_{(2)} = 2 \quad x_{(3)} = 3 \quad x_{(4)} = 3 \quad x_{(5)} = 3 \quad x_{(6)} = 4 \quad x_{(7)} = 4 \quad x_{(8)} = 5 \quad \dots \quad x_{(i)} = 7 \quad \dots \quad x_{(N)} = 8$$

Questo tipo di serie è maggiormente utile per fini pratici.

4.3 Tipi particolari di serie

Fra le serie statistiche ve ne sono alcune che rivestono particolare importanza e sono oggetto di specifici settori di ricerca. Si tratta delle serie storiche e delle serie territoriali. Entrambi questi tipi di serie sono caratterizzate dal fatto che i dati sono raccolti per specifici ambiti, rispettivamente temporali o spaziali. queste serie vengono rappresentate attraverso l'utilizzo di una tabella a due colonne, una contenente la suddivisione in tempi o spazi e l'altra contenente i dati osservati sulle unità statistiche.

1) Serie storiche:

Le serie storiche si rappresentano come nella tabella di seguito:

TEMPI	FENOMENO
0	x_0
1	x_1
...	...
i	x_i
...	...
N	x_N

Nelle serie storiche vale il principio di *non scambiabilità* dei tempi: i dati, cioè, si ordinano in base ai tempi e non possono essere scambiati fra di loro.

A seconda del tipo di fenomeno che rappresentano, possono poi essere di due tipi diversi:

- Serie Storiche di flusso o di movimento: per fenomeni che si realizzano in archi di tempo (esempio: le nascite di un anno).
- Serie storiche di stato o di fondo: per fenomeni che accadono e vengono registrati in specifici istanti di tempo (esempio: vendita di biglietti del cinema in una giornata.).
- Serie storica ciclica: per fenomeni registrati in corrispondenza di giorni della settimana o mesi dell'anno.

Le serie storiche si rappresentano, similmente alle serie in generale, con il simbolo:

$$\{i, x_i\} \quad i = 1, 2, \dots, N$$

Dove la i sta ad indicare che il dato i -esimo si è presentato al tempo i -esimo.

2) Serie territoriali:

Anche le Serie territoriali si rappresentano in una tabella del tipo di quella vista per le serie storiche ma con la prima colonna contenente la suddivisione per aree geografiche invece che per tempi.

Similmente a quanto visto per le serie storiche, poi, anche le serie territoriali possono rappresentarsi come segue:

$$\{t, x_t\} \quad t = 1, 2, \dots, N$$

Dove l'indice t associa ciascun dato al territorio dove è stato rilevato.

5. TABELLE STATISTICHE DI FREQUENZA

5.1 Tabelle statistiche

Quando le serie statistiche contengono molte informazioni la loro rappresentazione risulta complessa e si utilizza allora un metodo per classificare i dati e sintetizzarli in modo esaustivo e senza perdita di informazione: le tabelle statistiche di frequenza.

Le tabelle statistiche sono costituite da due colonne; la prima contiene, nelle diverse righe, le varie modalità del carattere considerato e viene definita *colonna madre*, mentre la seconda può contenere diversi valori, come vedremo di seguito. Infine la parte superiore della tabella contiene l'indicazione dei dati contenuti nelle varie colonne e viene definita *testata*.

Una tabella statistica viene definita, in maniera generica *distribuzione statistica semplice o distribuzione di frequenza*, intendendo, con tale termine, qualunque tabella che associa a ogni modalità del carattere una frequenza assoluta, dove per *frequenza assoluta* s'intende il numero delle unità statistiche associate ad una modalità. Le modalità si indicano con

$$x_i, i=1, 2, \dots, k,$$

dove k rappresenta il numero di modalità del carattere rappresentato, mentre le frequenze assolute si indicano con il simbolo

$$n_i, i=1, 2, \dots, k,$$

Esempio: Distribuzione di frequenze assolute relativa alla facoltà di appartenenza degli iscritti all'università di Torino in un dato anno

x_i	n_i
Lettere	8.494
Giurisprudenza	8.519
Scienze politiche	6.737
Magistero	5.674
Agraria	1.731
Scienze naturali	7.422
Economia	9.006
Farmacia	2.482
Medicina	4.554
Veterinaria	944
	55.563

Vi sono vari tipi di tabelle statistiche a seconda del tipo di carattere che si è rilevato sulla popolazione che si sta rappresentando.

Quando le tabelle si riferiscono ad un solo carattere, si distinguono in:

- a) *variabili statistiche*, quando la distribuzione statistica è relativa a caratteri quantitativi. Le variabili statistiche si distinguono, a loro volta, in
- *discrete* o
 - *continue*,
- a seconda che i valori cui esse si riferiscano facciano parte di un insieme discreto di valori isolati, oppure ad un insieme continuo di valori.

Esempio: Variabile statistica discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	n_i
20	6
35	9
50	5
75	10
	30

Esempio: Variabile statistica continua relativa all'altezza di 200 ragazzi misurata in cm

x_i	$ - x_{i+1}$	n_i
140	-150	2
150	-160	20
160	-170	84
170	-180	8
180	-190	77
190	-200	9
		200

Quando il carattere quantitativo è continuo non si considerano tutte le possibili determinazioni del suo insieme universo, ma i valori si raggruppano in **classi**; cioè in intervalli contigui che possono essere di ampiezza eguale o diversa. La prima e l'ultima classe possono poi non contenere un estremo, poiché i valori possono essere molto diradati e in questo caso sono dette classi aperte; è consigliabile evitare di eseguire elaborazioni con classi aperte, occorre chiuderle mediante valori scelti in modo opportuno. Quando si fa una rilevazione è molto importante stabilire l'ampiezza (o modulo), se possibile costante, delle classi, perché se le classi sono troppo

ampie la classificazione e le successive elaborazioni potrebbero essere distorte, mentre se le classi sono troppo ristrette la distribuzione sarebbe dispersa e poco regolare.

Talvolta si raggruppano in classi anche i caratteri quantitativi discreti, soprattutto se le modalità del carattere sono molto numerose. In generale il numero delle classi varia fra un minimo di cinque e un massimo di venti.

In una variabile statistica divisa in classi si definisce *limite inferiore* della classe il valore sinistro dell'intervallo, si definisce invece *limite superiore* della classe l'estremo destro dell'intervallo. L'*Intervallo* di classe è poi l'insieme dei valori contenuti fra i due limiti, laddove invece per *Ampiezza* della classe (o modulo) si intende la differenza fra limite superiore e limite inferiore della stessa. Infine la semisomma dei due limiti è definita *valore centrale* della classe.

Una variabile statistica è dunque definita dall'insieme dei valori osservati di un carattere quantitativo e dalle frequenze a essi associate.

Invece delle frequenze assolute si possono associare ai valori della variabile le *frequenze relative*, che si ottengono dividendo ogni frequenza per la somma totale delle frequenze stesse.

- b) *mutabile statistica*. Quando il carattere che interessa la rilevazione è un carattere qualitativo, la distribuzione configurata prende il nome di *mutabile statistica* ed è definita dall'insieme delle modalità osservate di un carattere qualitativo e dalle frequenze a esse associate.

La m.s. può essere:

- *rettilenea* se le modalità ammettono un ordine di successione naturale, con una modalità iniziale e una finale

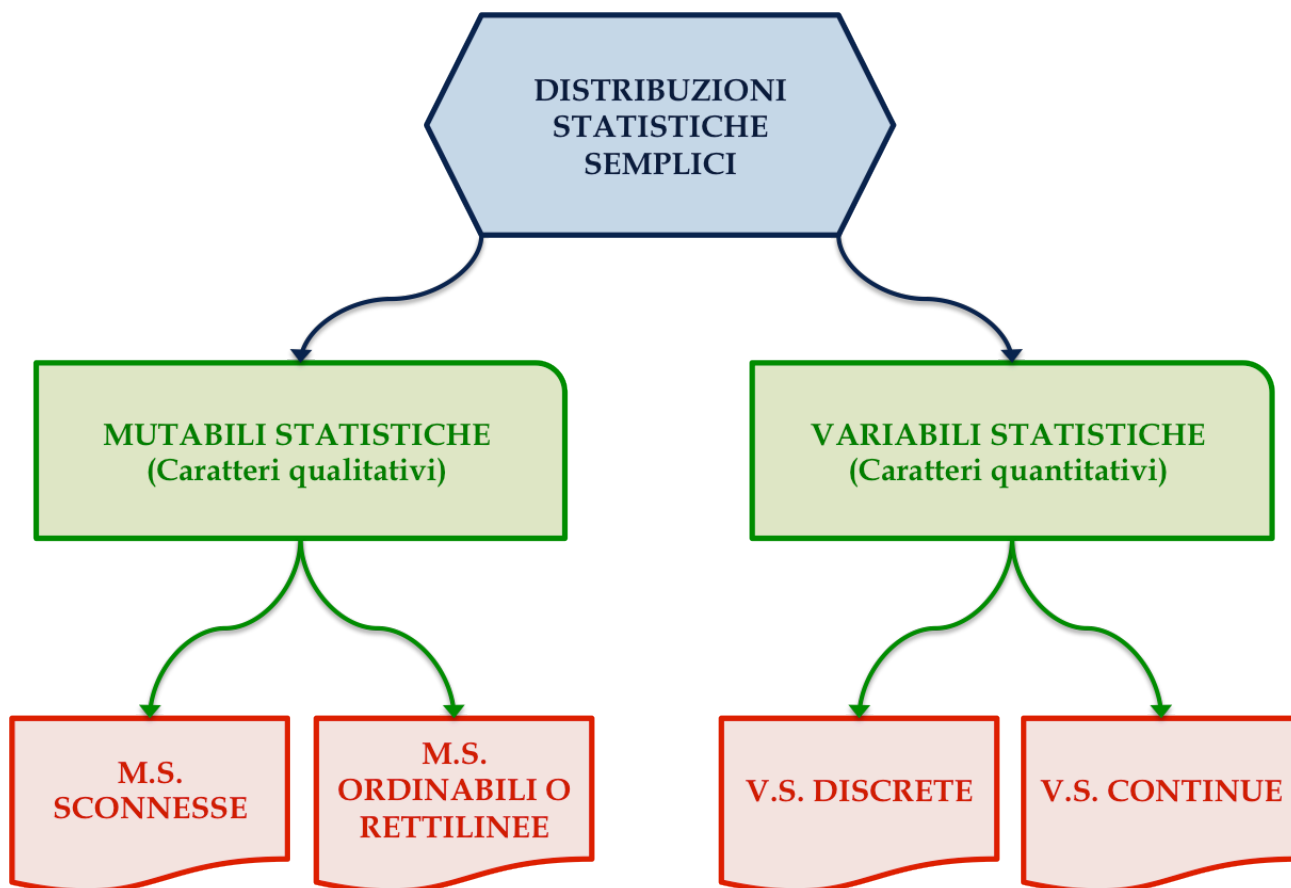
Esempio: *Mutabile rettilenea* relativa al rendimento di 100 iscritti ad un corso di laurea

x_i	n_i
sufficiente	3
buono	37
discreto	24
ottimo	36
	100

- *ciclica* se le modalità ammettono un ordine di successione ma non è possibile, a meno di una convenzione, stabilire quale sia il valore iniziale e quale quello finale (per esempio i giorni della settimana)
- *sconnessa*, se le modalità non ammettono alcun ordine di successione, come ad esempio nel caso delle nazionalità o delle professioni (la distribuzione degli iscritti all'università di Torino, vista in precedenza, è un esempio mutabile sconnessa)

- c) Tipi specifici di variabili statistiche sono quelle nelle quali le modalità sono rappresentate da istanti (o periodi) di tempo e quelle in cui sono rappresentate da divisioni territoriali scelte. Tali tabelle sono rispettivamente le *serie storiche* e le *serie territoriali*.

Dunque le distribuzioni statistiche possono essere schematizzate nel modo seguente.



5.2 Frequenze

Oltre alle *frequenze assolute*, che indicano esattamente il numero di volte in cui una modalità si presenta in una popolazione, esistono altri tipi di frequenze che possono essere associate alle modalità. L'esistenza e il calcolo di varie frequenze è giustificato dal fatto che, ciascuna di esse, fornisce un'informazione aggiuntiva rispetto alle frequenze assolute e permette di eseguire alcune operazioni specifiche:

- a) *frequenze cumulate* - N_i - sono quelle frequenze che associano a ogni valore della variabile discreta (o a ogni classe della variabile continua) la somma della rispettiva frequenza assoluta con le frequenze assolute dei valori precedenti.

$$N_i = \sum_{j=1}^i n_j \quad i=1,2,\dots,k \Rightarrow \forall i N_i = n_1 + n_2 + \dots + n_i$$

Esempio: Variabile discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	n_i	N_i
20	6	6
35	9	6+9=15
50	5	6+9+5=20
75	10	6+9+5+10=30
	30	/

- b) frequenze relative - f_i - sono quelle ottenute dal rapporto fra le frequenze assolute e il numero totale di unità:

$$f_i = \frac{n_i}{N}.$$

Relativizzare un valore, in questo caso una frequenza, consente di rendere le misure delle frequenze confrontabili per vari gruppi, ciò poiché le frequenze relative hanno l'importante proprietà di avere somma pari ad 1.

Esempio: Variabile discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	n_i	f_i
20	6	0,2
35	9	0,3
50	5	0,16
75	10	0,33
	30	1

- c) frequenze relative cumulate - F_i - sono le frequenze che associano a ogni valore della variabile discreta (o a ogni classe della variabile continua) la somma della rispettiva frequenza relativa con le frequenze dei valori precedenti.

$$F_i = \sum_{j=1,2,\dots,i}^i f_j \Rightarrow \forall i F_i = f_1 + f_2 + \dots + f_i$$

Esempio: Variabile discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	f_i	F_i
20	0,2	0,2
35	0,3	0,2+0,3=0,5
50	0,16	0,2+0,3+0,16=0,67
75	0,33	0,2+0,3+0,16+0,33=1
	1	/

- d) frequenze percentuali - p_i - sono date dalle frequenze relative moltiplicate per cento.

$$p_i = f_i \cdot 100$$

Esempio: Variabile discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	f_i	p_i
20	0,2	20%
35	0,3	30%
50	0,16	16%
75	0,33	33%
	1	100%

e) Frequenze percentuali cumulate - P_i -

$$P_i = F_i \cdot 100$$

Esempio: Variabile discreta relativa agli investimenti pubblicitari di 30 aziende in migliaia di euro

x_i	f_i	F_i	P_i
20	0,2	0,2	20%
35	0,3	0,5	50%
50	0,16	0,67	67%
75	0,33	1	100%
	1	/	/

f) Densità di frequenza - h_i - Questo tipo di frequenza si può calcolare solamente in caso di modalità rappresentate da classi di frequenza e usa per la realizzazione degli istogrammi (rappresentazioni grafiche per caratteri quantitativi continui). Le densità di frequenza sono date dal rapporto fra le frequenze assolute e le ampiezze delle classi:

$$h_i = \frac{n_i}{a_i} \text{ con } a_i = x_{i+1} - x_i$$

Talvolta le rilevazioni statistiche sono effettuate considerando per ogni unità statistica l'andamento dei valori riferiti a più di un carattere. La rappresentazione in tabelle di tali rilevazioni da origine alle *tabelle a doppia entrata*, nel caso di due caratteri e alle *tabelle multiple* nel caso di più di due caratteri. Le tabelle a doppia entrata rappresentanti distribuzioni di frequenza possono essere:

- di *contingenza*, se i due caratteri sono entrambi qualitativi (si parla di *mutabile statistica doppia*),
- di *correlazione*, se i due caratteri sono entrambi quantitativi (si parla di *variabile statistica doppia*),
- *miste*, se uno dei due caratteri è qualitativo e l'altro quantitativo.

Esempio: Mutabile statistica doppia (tabella di contingenza) relativa al numero di turisti (in migliaia), in alcune città italiane in un dato mese.

	TURISTI		
CITTA'	Italiani	Stranieri	TOTALI
Venezia	4	2	6
Firenze	5	4	9
Roma	7	6	13
Torino	3	3	6
Napoli	1	1	2
TOTALI	20	16	36

6. RAPPRESENTAZIONI GRAFICHE

Le tabelle statistiche sono un primo strumento di presentazione dei dati. Oltre alle tabelle, i dati possono essere presentati anche attraverso rappresentazioni grafiche. Le rappresentazioni grafiche offrono numerosi *VANTAGGI*:

- descrivono il fenomeno in forma visiva e permettono di esaminare l'andamento in modo globale e di confrontare caratteri diversi dello stesso fenomeno e le sue variazioni nel tempo e nello spazio.
- sono molto più espressive di una tabella di valori. I grafici, essendo di facile lettura, permettono infatti a tutti di capire l'andamento di un fenomeno.

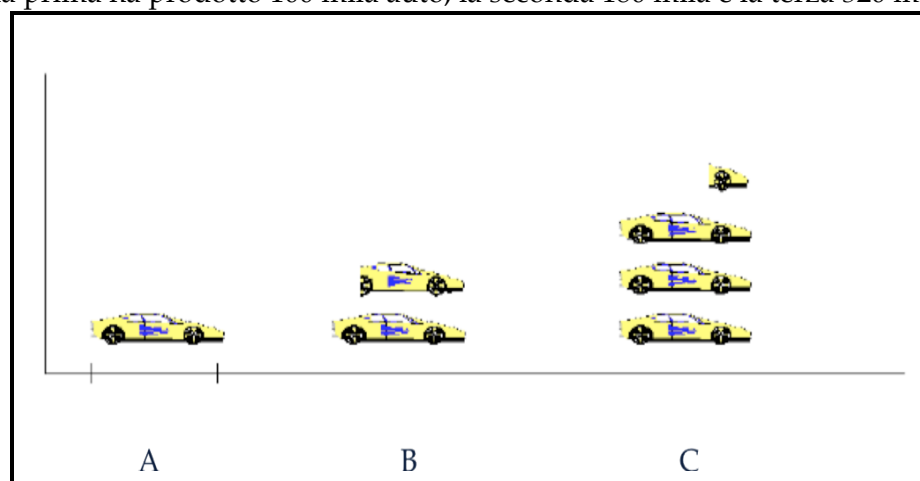
Le rappresentazioni sono utilizzate con un duplice *SCOPO*:

- a) *descrittivo* quando vengono utilizzate, come già detto, per permettere di esaminare il fenomeno in forma visiva
- b) *scientifico* se si utilizzano per ricercare il modello matematico alla base del fenomeno (ossia una funzione che ne esprima l'andamento).

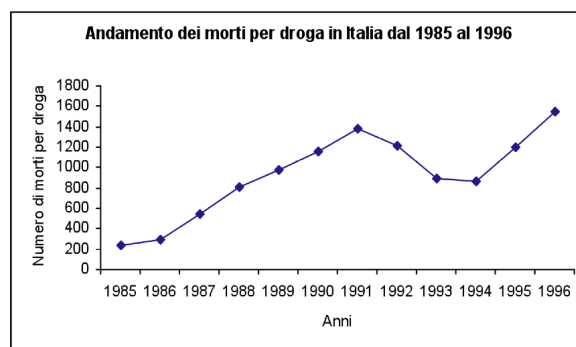
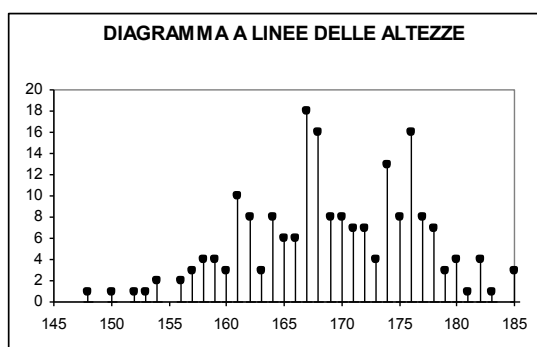
Si utilizzano diversi *TIPI* di rappresentazioni grafiche:

1. **Pittogrammi a figure ripetute**, utilizzati quando il grafico deve servire, più che come metodo di studio, come metodo di divulgazione al pubblico dei dati. Si utilizzano allora dei motivi ideografici che, opportunamente affiancati da una legenda, offrono un'idea del fenomeno in esame. Nei pittogrammi ad ogni immagine corrisponde una data intensità del fenomeno. Occorre tuttavia osservare che questo genere di rappresentazioni non sono sempre corrette. Ciò accade quando nel rappresentare il raddoppiamento di un fenomeno da un anno all'altro si utilizzi una figura che è il doppio di dimensioni della prima. In questo caso, infatti, si avrà un'immagine di dimensione quadrupla rispetto alla precedente e perciò stesso errata. Una corretta rappresentazione dovrebbe raddoppiare l'area della figura e non le sue dimensioni. Peraltro una corretta legenda può essere di aiuto in tal senso, guidando l'osservatore nella interpretazione della rappresentazione.

Esempio: Pittogramma della produzione mensile di auto di 3 case automobilistiche:
la prima ha prodotto 100 mila auto, la seconda 180 mila e la terza 320 mila



2. **Diagrammi cartesiani**. Fanno uso delle coordinate cartesiane ortogonali del piano. Sull'asse delle ascisse si usa rappresentare le modalità del fenomeno, mentre su quello delle ordinate vengono rappresentate le corrispondenti frequenze. Si individuano quindi i punti nel piano corrispondenti alle coppie di valori (modalità e frequenze corrispondenti). A questo punto, nel caso di caratteri quantitativi discreti, si tracciano delle linee verticali in corrispondenza di ciascuna modalità, aventi altezza pari alla frequenza corrispondente a quella modalità. Nel caso di serie storiche, invece, si congiungono i punti successivi con dei segmenti in maniera da formare una spezzata.

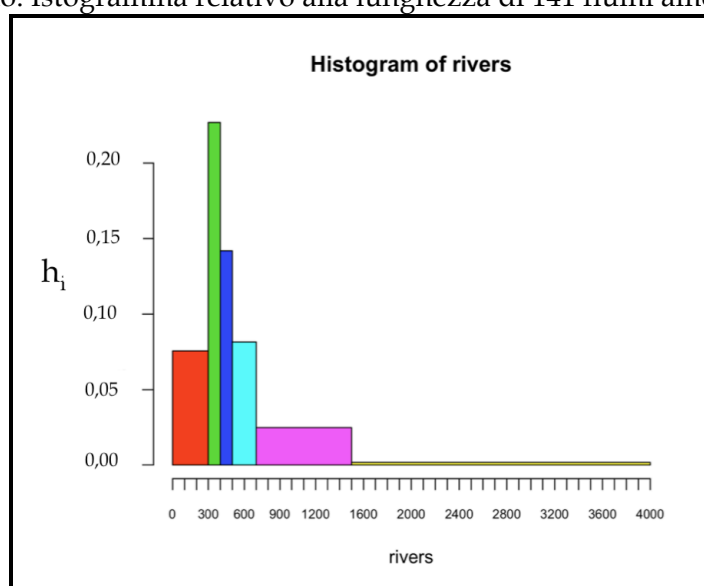


3. **Istogrammi:** anche questo tipo di rappresentazione fa uso del piano cartesiano. Si utilizza nel caso di serie statistiche continue, cioè i cui caratteri sono continui o discreti ma con un gran numero di modalità. Fissato un sistema di assi cartesiani ortogonali, sull'asse delle ascisse si riportano tanti intervalli consecutivi quante sono le classi; su questi intervalli si costruiscono dei rettangoli le cui aree sono proporzionali alle frequenze. Ciò vale a dire che su un segmento di base pari all'ampiezza della classe si costruisce un rettangolo di altezza pari al rapporto fra la frequenza totale di quella classe e l'ampiezza della classe stessa, ovvero la *densità di frequenza*. Solo nel caso in cui le classi abbiano la stessa ampiezza basterà riportare altezze proporzionali alle frequenze. Con gli istogrammi la somma delle aree di tutti i rettangoli è proporzionale alla somma delle frequenze.

Esempio: Variabile statistica continua della lunghezza di 141 fiumi americani in km

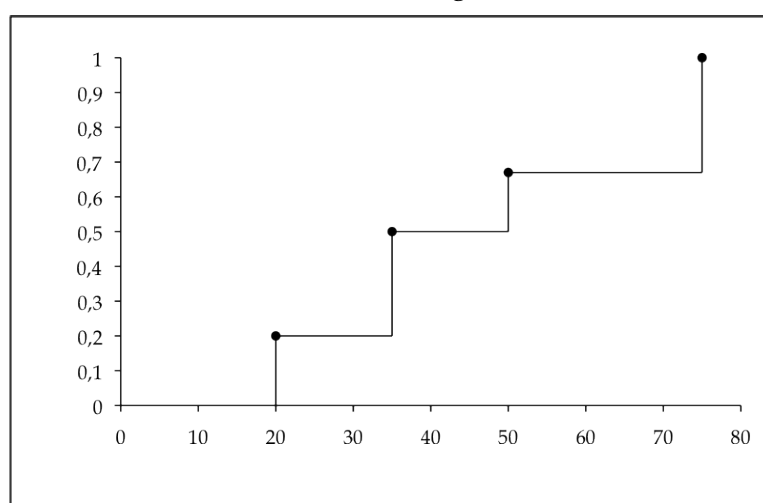
x_i - x_{i+1}	n_i	a_i	h_i
0 - 300	32	300	0,11
300 - 400	32	100	0,32
400 - 500	20	100	0,20
500 - 600	13	100	0,13
600 - 700	10	100	0,10
700 - 1500	28	800	0,04
1500 - 4000	6	2500	0,00
	141	/	/

Esempio: Istogramma relativo alla lunghezza di 141 fiumi americani in km

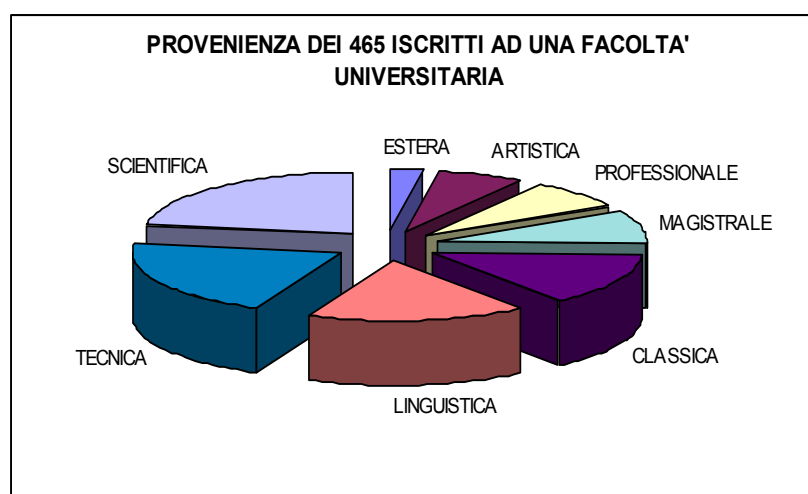


4. **Diagramma integrale.** Si ha a che fare con questo tipo di diagramma quando vengono rappresentate, invece delle frequenze assolute, le frequenze cumulate (cioè le frequenze rispettivamente minori o uguali o maggiori di ciascuna modalità successiva) o ancora possono essere rappresentate le cosiddette le frequenze relative cumulate (in altre parole le frequenze dei valori maggiori o uguale di ogni successiva modalità in relazione alla frequenza totale). L'opportunità della presa in considerazione di questo tipo di frequenza sta nella circostanza che la somma totale delle frequenze relative accumulate (o funzioni di ripartizione) è sempre uguale a 1 e ciò ne facilita la rappresentazione grafica. Il diagramma integrale si costruisce rappresentando le successive frequenze cumulate o cumulate relative, in corrispondenza degli estremi superiori delle classi e congiungendo fra loro i vari punti con una spezzata. Questo stesso genere di rappresentazione è quella che viene utilizzata ad esempio per le serie storiche che hanno la caratteristica di avere sulle ascisse, come modalità, gli anni o i periodi di tempo che sono stati oggetto della rilevazione.

Funzione di ripartizione della variabile Investimenti pubblicitari
di 30 aziende in migliaia di euro

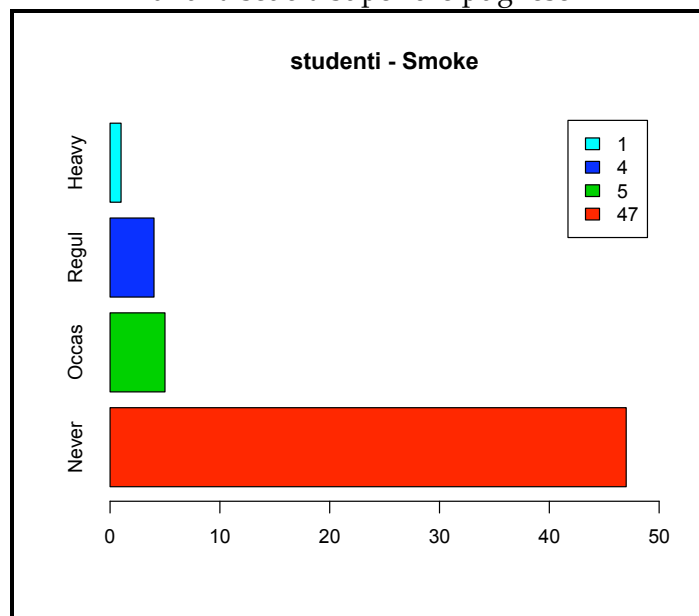


5. **Diagramma a settori circolari.** Questo genere di rappresentazione si utilizza per le m.s. sconnesse. All'interno di un cerchio o un semicerchio si rappresentano tutte le modalità, ognuna con frequenza pari all'area della sezione del cerchio che occupano. L'ampiezza di ogni settore sarà data dalla formula: $N : n = 360^\circ : \alpha$. Dove n è la frequenza della modalità α ed N è la frequenza totale.



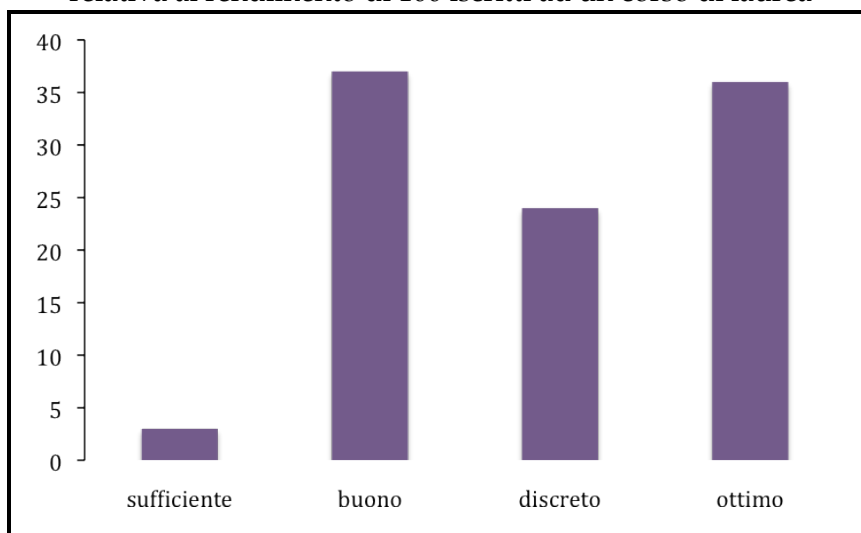
6. **diagramma a nastri.** Per la rappresentazione delle mutabili statistiche sconnesse talvolta si utilizzano dei nastri orizzontali. Questo metodo però necessita del preventivo ordinamento dei dati.

Esempio: Diagramma a nastri del carattere "Abitudine al fumo" di 57 studenti di una scuola superiore pugliese



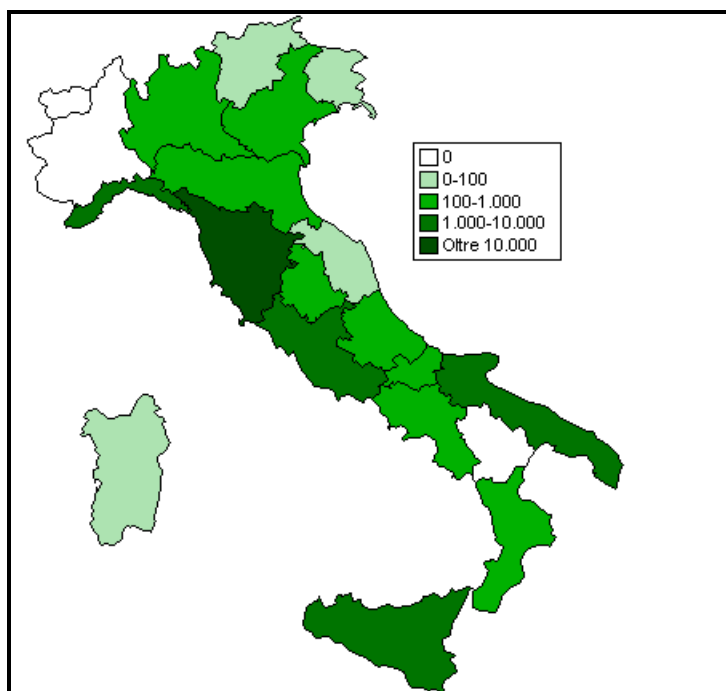
7. **Diagramma a barre verticali.** Per la rappresentazione delle mutabili statistiche ordinabili si possono utilizzare barre verticali aventi altezza pari alle frequenze.

Esempio: Rappresentazione della mutabile rettilinea relativa al rendimento di 100 iscritti ad un corso di laurea

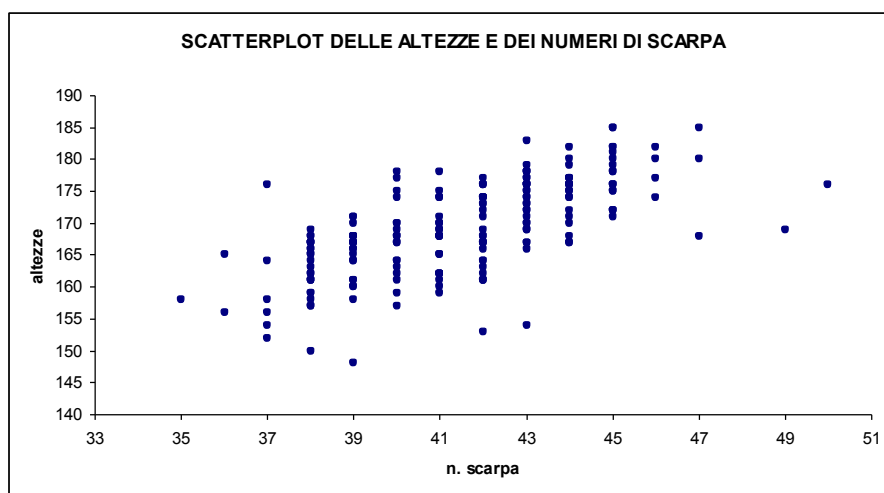


8. **Cartogrammi.** Sono i tipi di rappresentazioni utilizzate per le serie territoriali. Per realizzarle si utilizza lo schema della carta geografica dell'area oggetto di rilevazione, utilizzando dei simboli in ogni area o una colorazione particolare per rappresentare l'intensità del fenomeno.

Esempio: Numero di operatori del settore olii extravergine d'oliva DOP e IGP per regione (anno 2009).



9. **Nuvola di punti.** È questo il primo dei metodi che si utilizzano per rappresentare le variabili o mutabili statistiche doppie o multiple. Quando si devono rappresentare N coppie di valori corrispondenti si utilizza un sistema di assi cartesiani all'interno del quale si individuano tanti punti quante sono le coppie di valori corrispondenti alle intensità o frequenze individuate per ogni coppia di modalità.



Regole importanti da osservare nella rappresentazione grafica:

Quando si realizzano dei grafici è necessario prestare attenzione ad alcune importanti regole:

- 1) La rappresentazione deve essere autonoma dalla tabella che la ha originata, nel senso che deve contenere tutte le informazioni utili per la sua esatta interpretazione.

2) Conseguentemente il titolo deve contenere:

- a. Oggetto della rappresentazione (cosa)
- b. Epoca (quando)
- c. Ambito territoriale (dove)

Esempio: "Studenti residenti in Italia per ripartizione geografica, 1990".

"Consumi finali delle famiglie per tipo di consumo, composizione percentuale, anno 2000".

3) E' necessario indicare nel grafico:

- a. Quali caratteri sono stati osservati sulle unità.
- b. Le unità di misura.

La scelta delle unità di misura è arbitraria e ha un'importanza rilevante se si pensa che scegliendo opportunamente la scala di misura si può dare maggiore o minore risalto a alle variazioni delle frequenze e delle modalità o si può riuscire a risolvere problemi di rappresentazione dovuti alla presenza di intervalli assoluti fra le modalità o le frequenze di ampiezza molto elevata.

Molto di frequente si utilizzano le scale logaritmiche (che consistono nel sostituire a ciascuna modalità o frequenza il suo corrispondente logaritmo naturale). Questa scala fa sì che gli intervalli sugli assi siano rappresentativi di rapporti fra i fenomeni rappresentati e quindi delle loro differenze relative. (ad intervalli uguali sugli assi corrispondono rapporti uguali dei valori)

- c. Gli eventuali troncamenti di scala
- d. Se sono stati inseriti più caratteri della variabile, questi devono essere indicati in maniera diversa
- e. La legenda, quando necessaria, va posta alla base del grafico.

ESERCIZI - 1

ESERCIZIO 1

Di seguito sono riportate le votazioni finali in centesimi di 20 studenti al test di ammissione al corso di Laurea in Biotecnologia

64	71	83	82	64	75	91	68	83	71
64	87	83	62	75	62	68	87	87	64

1. Individuare il tipo di carattere
2. Realizzare la variabile statistica relativa al carattere "voto in centesimi"
3. Calcolare le frequenze cumulate, le frequenze relative, relative cumulate, e percentuali;
4. Rappresentare graficamente la distribuzione

ESERCIZIO 2

Di seguito sono riportate le lunghezze di 40 foglie di platano, registrate al millimetro più prossimo:

138	164	150	132	144	149	157	146	158	140
147	136	148	152	144	168	126	138	176	163
119	154	165	146	173	142	147	135	153	140
135	161	145	135	142	150	156	145	128	150

1. Individuare il tipo di carattere
2. Realizzare la variabile statistica relativa al carattere "lunghezza delle foglie" utilizzando la suddivisione in classi ritenuta opportuna;
3. Calcolare le frequenze cumulate, le frequenze relative, relative cumulate, percentuali e le densità di frequenza;
4. Rappresentare graficamente la distribuzione.

ESERCIZIO 3

Di seguito è riportato il corso di laurea di 16 studenti iscritti alla facoltà di Lettere e filosofia dell'Università di Udine nell'anno 2010

C	C	D	D	D	D	L	D
C	C	D	L	L	D	C	D

C= Conservazione dei beni culturali D= Dams L= Lettere

1. Individuare il tipo di carattere
2. Realizzare la distribuzione statistica relativa al carattere "Corso di Laurea";
3. Calcolare le frequenze cumulate, le frequenze relative, relative cumulate e percentuali;
4. Rappresentare graficamente la distribuzione.

ESERCIZIO 4

Di seguito è riportato il grado di soddisfazione di 16 studenti iscritti alla facoltà di Lettere e filosofia dell'Università di Udine nell'anno 2010

1	3	3	3	4	2	2	2
2	4	3	2	1	1	3	4

1=insoddisfatto 2=poco soddisfatto 3=soddisfatto 4=molto soddisfatto

1. Individuare il tipo di carattere
2. Realizzare la distribuzione statistica relativa al carattere "Corso di Laurea";
3. Calcolare le frequenze cumulate, le frequenze relative, relative cumulate e percentuali;
4. Rappresentare graficamente la distribuzione.

7.A. GLI INDICATORI STATISTICI

7.A.1 Definizioni

Dopo aver effettuato una raccolta di dati e dopo aver sistemato i dati ottenuti in tabelle statistiche è importante effettuare una sintesi sulle distribuzioni al fine di consentire la comparazione del fenomeno oggetto di studio con altri fenomeni dello stesso tipo o di tipo diverso per i quali abbia senso un confronto dei valori.

È quindi necessario trovare un valore che sappia sintetizzare il fenomeno cogliendo il sottofondo costante esistente fra le singole osservazioni. È ovvio però che attraverso la sintesi si chiede all'indice di esprimere la distribuzione secondo pochi aspetti o uno solo, quello che sembra più importante ed è ovvio che il dato di sintesi così costruito porta ad una perdita di informazione.

Lo scopo degli indicatori statistici, dunque, è quello di sintetizzare in un singolo valore numerico l'intera distribuzione di frequenza al fine di ottenere valori confrontabili nel tempo, nello spazio o tra circostanze e fenomeni differenti ma in qualche modo assimilabili.

7.A.2 Requisiti

Gli indicatori statistici (medie) devono soddisfare uno o entrambi i seguenti requisiti:

Condizione di internalità di CAUCHY:

Il valore sintetico ($\bar{x}$) deve essere compreso fra il valore più grande e il più piccolo della distribuzione:

$$x_{(1)} < \bar{x} < x_{(s)}$$

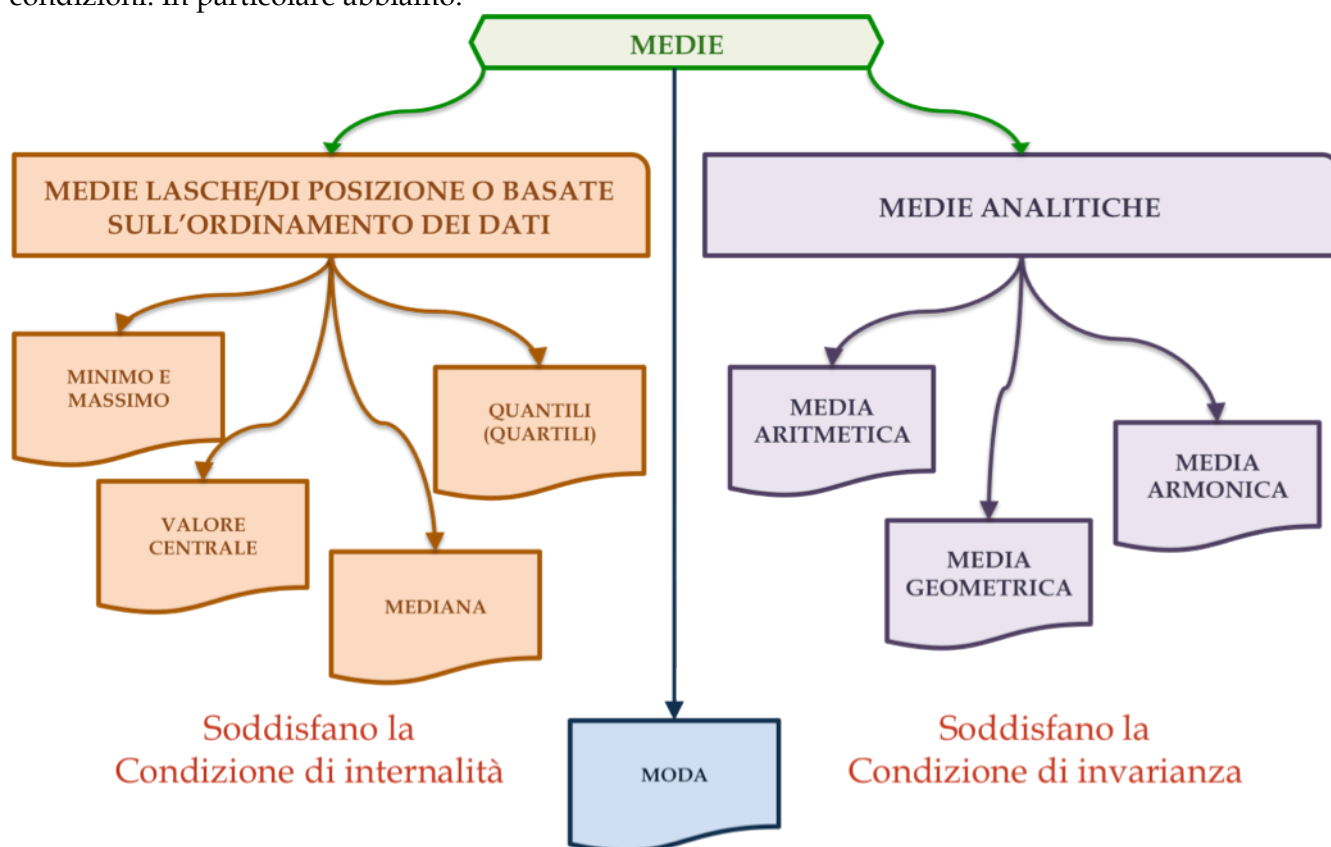
Condizione di invarianza di CHISINI:

Il valore sintetico deve essere tale che, sostituito a ciascuna osservazione, sia rappresentativo della distribuzione, ovvero possa essere sostituito a ciascuna delle osservazioni, lasciando immutata la quantità assunta come invariante e rappresentata da una funzione f . In altre parole deve valere la seguente uguaglianza:

per le serie di valori: $f(x_1, x_2, \dots, x_N) = f(\bar{x}, \bar{x}, \dots, \bar{x})$

per le distribuzioni: $f(x_1, x_2, \dots, x_s) = f(\bar{x}, \bar{x}, \dots, \bar{x})$

Ci sono diversi tipi di "indicatori statistici" e non tutti soddisfano contemporaneamente queste due condizioni. In particolare abbiamo:



7.B. INDICI STATISTICI DI POSIZIONE O MEDIE LASCHE O MEDIE BASATE SULL'ORDINAMENTO DEI DATI:

Questo primo tipo di medie sono definite in modo generale "medie interne" perché soddisfano la condizione di internalità di Cauchy. Per la loro determinazione è necessario preventivamente ordinare i dati in ordine non decrescente e quindi queste medie non possono essere calcolate per mutabili sconnesse), ma possono essere calcolati per tutti gli altri tipi di variabili. Inoltre, sono definite medie "lasche" perché non tengono conto di tutti i valori della serie o distribuzione ma solo di alcuni; medie "di posizione" perché individuano la posizione occupata da una particolare modalità; infine sono definite medie "basate sull'ordinamento dei dati" perché sono, appunto, basate sulla posizione degli elementi della variabile.

Le medie posizionali sono:

- I Valori estremi
- Il Valore Centrale o Campo di Variazione
- La mediana
- I quantili.
- la Moda, che studieremo in questo gruppo anche se in realtà non fa parte di nessuno dei gruppi di medie visti.

7.B.1 Valori estremi

Questi indici posizionali, insieme al valore centrale e al campo di variazione, possono essere calcolati solo nel caso di variabili quantitative, sia continue che discrete. Precisamente:

- $x_{\min}$ è il primo valore della distribuzione (ordinata)
- $x_{\max}$ è l'ultimo valore della distribuzione.

7.B.2 Valore centrale

Anche questo indice è calcolabile soltanto per variabili quantitative sia continue, che discrete ed è dato dalla semisomma del minimo e del massimo: $\frac{x_{\max} + x_{\min}}{2}$.

Questo indice possiede il grosso svantaggio di essere fortemente influenzato da eventuali valori anomali della distribuzione, laddove, per valori anomali si intendono quei valori molto grandi o molto piccoli rispetto alla maggior parte dei valori della distribuzione o della serie.

In genere, nella ricerca di un indicatore, si cerca di evitare che tale valore sia influenzato da dati anomali. Tuttavia il valore centrale è utile in alcune situazioni.

7.B.3 Mediana

La mediana è quella modalità che bipartisce la serie o distribuzione di dati lasciando ugual numero di termini da un lato e dall'altro. Per la sua determinazione, come per le altre medie interne, è necessario ordinare preliminarmente i dati. In sostanza la Mediana bipartisce la distribuzione in due sotto-distribuzioni: la prima a sinistra della Mediana (costituita dalla metà delle unità la cui modalità è minore o uguale alla Mediana) e la seconda a destra della Mediana (costituita dalla metà delle unità la cui modalità è maggiore o uguale alla Mediana). Tecnicamente si afferma che la mediana è il valore/modalità per il quale frequenza relativa cumulata vale 0,5.

La mediana può essere calcolata per caratteri quantitativi e qualitativi ordinabili ma non per caratteri qualitativi sconnessi, poiché questi non possono essere ordinati. Il suo calcolo è differente a seconda che N sia pari o dispari e, anche, a seconda del tipo di distribuzione.

1. MEDIANA PER SERIE STATISTICHE CON CARATTERI QUANTITATIVI DISCRETI

In questo caso le singole osservazioni vanno ordinate in modo non decrescente e il valore mediano è differente nei due seguenti casi:

- Se N è un numero pari $Me = \frac{x_{(N/2)} + x_{(N/2+1)}}{2}$
- Se N è un numero dispari $Me = x_{\left(\frac{N+1}{2}\right)}$

2. MEDIANA PER VARIABILI STATISTICHE DISCRETE

In questo caso la mediana è calcolata utilizzando le frequenze cumulate. La mediana è data dalla modalità che ricopre il posto centrale nella sequenza delle frequenze cumulate N_i e quindi le formule per il calcolo della mediana sono le stesse viste per le serie con la sola differenza che dovremo guardare le frequenze cumulate e non quelle assolute :

- Se N è un numero pari $Me = \frac{x_{(N/2)} + x_{(N/2+1)}}{2}$
- Se N è un numero dispari $Me = x_{\left(\frac{N+1}{2}\right)}$

3. MEDIANA PER VARIABILI STATISTICHE DISCRETE

Se le modalità sono raggruppate in classi, non si definisce un valore univoco, ma una classe mediana $x_i - x_{i+1}$ determinata considerando frequenze cumulate: la classe mediana è, cioè, la classe la cui frequenza cumulata N_i è immediatamente superiore a $N/2$.

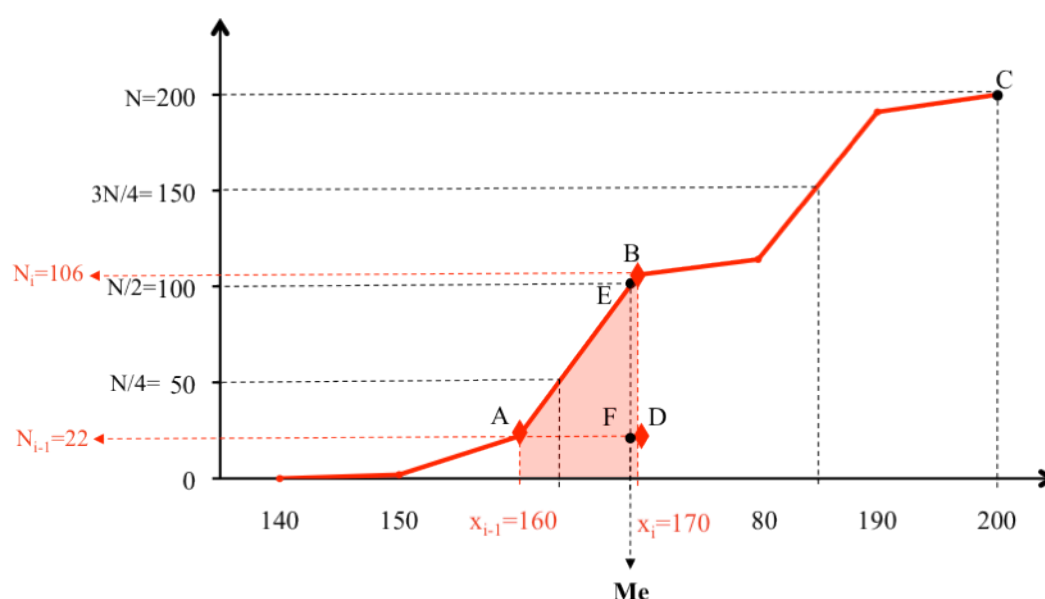
Pur essendo corretto considerare un qualsiasi elemento dell'intervallo mediano un valore mediano si è soliti procedere, al fine di avere una misura unica del valore, ad un'approssimazione della mediana con la una formula calcolata, anch'essa, utilizzando le frequenze cumulate. Precisamente il valore mediano viene ricercato mediante una funzione interpolatrice che si ottiene attraverso una similitudine fra triangoli nel grafico delle frequenze cumulate.

A titolo di esempio, per comprendere come si ottiene la formula, consideriamo la distribuzione delle altezze in cm di 200 ragazzi di una scuola, divisa per classi, nel modo seguente:

Variabile statistica continua relativa all'altezza di 200 ragazzi di una
Scuola media pugliese nell' a.s. 2009-2010

x_i	$- x_{i+1}$	n_i	N_i
140	-150	2	2
150	-160	20	22
160	-170	84	106
170	-180	8	114
180	-190	77	191
190	-200	9	200
TOTALE		200	

Quindi consideriamo il grafico delle frequenze cumulate relativo a tale variabile (la spezzata rossa nel grafico riportato di seguito):



In questo grafico la mediana corrisponderebbe al valore sull'asse delle ascisse in corrispondenza di $N/2=100$. Per ottenere tale punto si fa riferimento al teorema sui triangoli simili, per cui si ha:

$$AF : AD = FE : DB$$

Ovvero, nel nostro caso:

$$a : 10 = 78 : 84$$

da cui:

$$a = \frac{10 \cdot 78}{84}.$$

Ora riflettiamo sul fatto che Me non è altro che $a +$ il valore x_{i-1} , ovvero 160, per cui:

$$Me = 160 + \frac{10 \cdot 78}{84}.$$

Che può essere scritto come

$$Me = 160 + \frac{10}{84} \cdot 78 = x_{i-1} + \frac{x_i - x_{i-1}}{N_i - N_{i-1}} \cdot \left(\frac{N}{2} - N_{i-1} \right)$$

Da cui, se indichiamo d_i l'ampiezza della classe, otteniamo la formula della mediana:

$$Me = x_{i-1} + \frac{d_i}{n_i} \cdot \left(\frac{N}{2} - N_{i-1} \right)$$

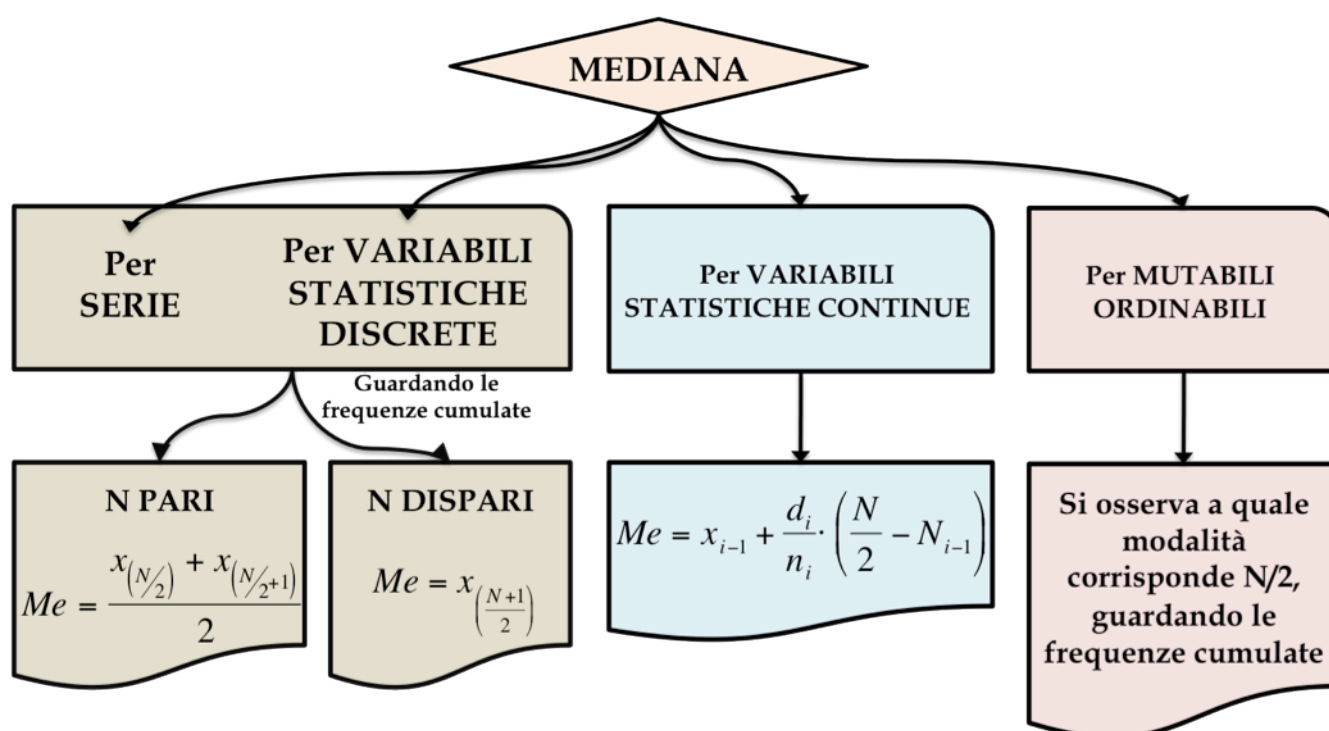
dove

- x_{i-1} è l'estremo superiore della classe $(i-1)$ esima
- d_i è l'ampiezza della classe
- N_{i-1} è la frequenza cumulata della classe precedente
- $\frac{N}{2}$ è il valore della distribuzione della mediana.

4. MUTABILI QUALITATIVE ORDINABILI

La mediana in questo caso non può che essere individuata semplicemente osservando le frequenze cumulate e la modalità a cui corrisponde la frequenza mediana.

Ricapitoliamo quanto visto fino ad ora in uno schema:



7.B.4 Proprietà della mediana

Prima di passare alla analisi di un altro tipo di media ricordiamo brevemente le *proprietà e i vantaggi della mediana*:

- la mediana può sempre essere calcolata e perciò viene utilizzata ogni volta che non è possibile calcolare la media aritmetica.
- La mediana è il valore che occupa la posizione centrale nell'elenco ordinato dei valori della serie o distribuzione e pertanto ha il vantaggio di non essere influenzato dai valori estremi che possono essere valori anomali (outliers). Per tale motivo è indicato come indice robusto.
- La mediana gode di una importante proprietà definita Proprietà di minimo. Per tale proprietà, dopo aver definito la quantità $x_i - Me$ scarto dalla mediana, affermiamo che la somma degli scarti in valore assoluto dalla mediana è un minimo, cioè:

$$\sum_{i=1}^s |x_i - Me| = \min$$

E quindi la somma dei valori assoluti degli scarti da qualunque altro valore generico k , sarà maggiore:

$$\sum_{i=1}^s |x_i - Me| < \sum_{i=1}^s |x_i - k| \neq Me$$

7.B.5 Quantili e quartili

I quantili sono quei valori che dividono la serie o la distribuzione in h parti, ognuna avente la h -esima parte della numerosità del collettivo. Naturalmente se la serie o distribuzione è divisa in h parti allora i quantili saranno $h-1$.

Quando, in particolare, $h=4$ allora i quantili prendono il nome di QUARTILI, dividono la serie o distribuzione in 4 parti e sono 3; li chiameremo Q_1 , Q_2 e Q_3 e diremo che:

- il I quartile Q_1 contiene il 25% della distribuzione (e ne lascia dopo di sé il 75%)
- il II quartile Q_2 contiene il 50% (ne lascia dopo di sé il 50% e coincide con la mediana)
- il III quartile Q_3 contiene il 75% (e ne lascia dopo di sé il 25%)

I quartili non possono essere calcolati per variabili qualitative sconnesse per lo stesso motivo per cui non può essere calcolata la mediana e cioè perché le modalità non sono suscettibili di alcun ordinamento.

I quartili e i quantili in generale si calcolano con una metodologia del tutto analoga a quella con la quale si calcola la mediana (poiché hanno un significato molto simile). Analogamente a quanto visto per la mediana, pertanto, anche per i quartili bisogna distinguere i vari casi:

1. QUARTILI PER SERIE STATISTICHE CON CARATTERI QUANTITATIVI DISCRETI

In questo caso è necessario ordinare i valori e quindi calcolare i quartili in modo differente a seconda che il valore $N/4$ abbia un resto oppure no, a seconda cioè che N sia o meno divisibile per 4.

- Se N è divisibile per 4 naturalmente sarà pari perciò, analogamente alla mediana:

$$Q_1 = \frac{x_{\left(\frac{N}{4}\right)} + x_{\left(\frac{N}{4}+1\right)}}{2}$$

$$Q_3 = \frac{x_{\left(\frac{3}{4}N\right)} + x_{\left(\frac{3}{4}N+1\right)}}{2}$$

- Se N non è divisibile per 4 allora avremo un resto che chiameremo h e i quartili saranno i seguenti:

$$Q_1 = x_{\left(\frac{N-h}{4}+1\right)}$$

$$Q_3 = x_{\left(\frac{3N-h}{4}\right)}$$

2. QUANTILI PER VARIABILI STATISTICHE DISCRETE

Anche questo caso è analogo a quello della mediana. Pertanto occorrerà calcolare le frequenze cumulate e individuare i quartili con le stesse formule viste nel caso di serie di valori, ma guardando questa volta le frequenze cumulate.

- Se N è divisibile per 4 naturalmente sarà pari, perciò:

$$Q_1 = \frac{x_{\left(\frac{N}{4}\right)} + x_{\left(\frac{N}{4}+1\right)}}{2}$$

$$Q_3 = \frac{x_{\left(\frac{3N}{4}\right)} + x_{\left(\frac{3N}{4}+1\right)}}{2}$$

- Se N non è divisibile per 4 allora avremo un resto che chiameremo h e i quartili saranno i seguenti:

$$Q_1 = x_{\left(\frac{N-h}{4}+1\right)}$$

$$Q_3 = x_{\left(\frac{3N-h}{4}+1\right)}$$

3. QUANTILI PER VARIABILI STATISTICHE CONTINUE

Come abbiamo verificato per la mediana, anche qui utilizziamo formule interpolatrici che utilizzano le frequenze cumulate. Riflettendo sul grafico precedentemente illustrato si ottiene quanto segue:

$$\text{- I quartile: } Q_1 = x_{i-1} + \frac{d_i}{n_i} \left(\frac{1}{4} N - N_{i-1} \right)$$

$$\text{- II quartile = mediana: } Q_2 = x_{i-1} + \frac{d_i}{n_i} \left(\frac{N}{2} - N_{i-1} \right)$$

$$\text{- III quartile: } Q_3 = x_{i-1} + \frac{d_i}{n_i} \left(\frac{3}{4} N - N_{i-1} \right)$$

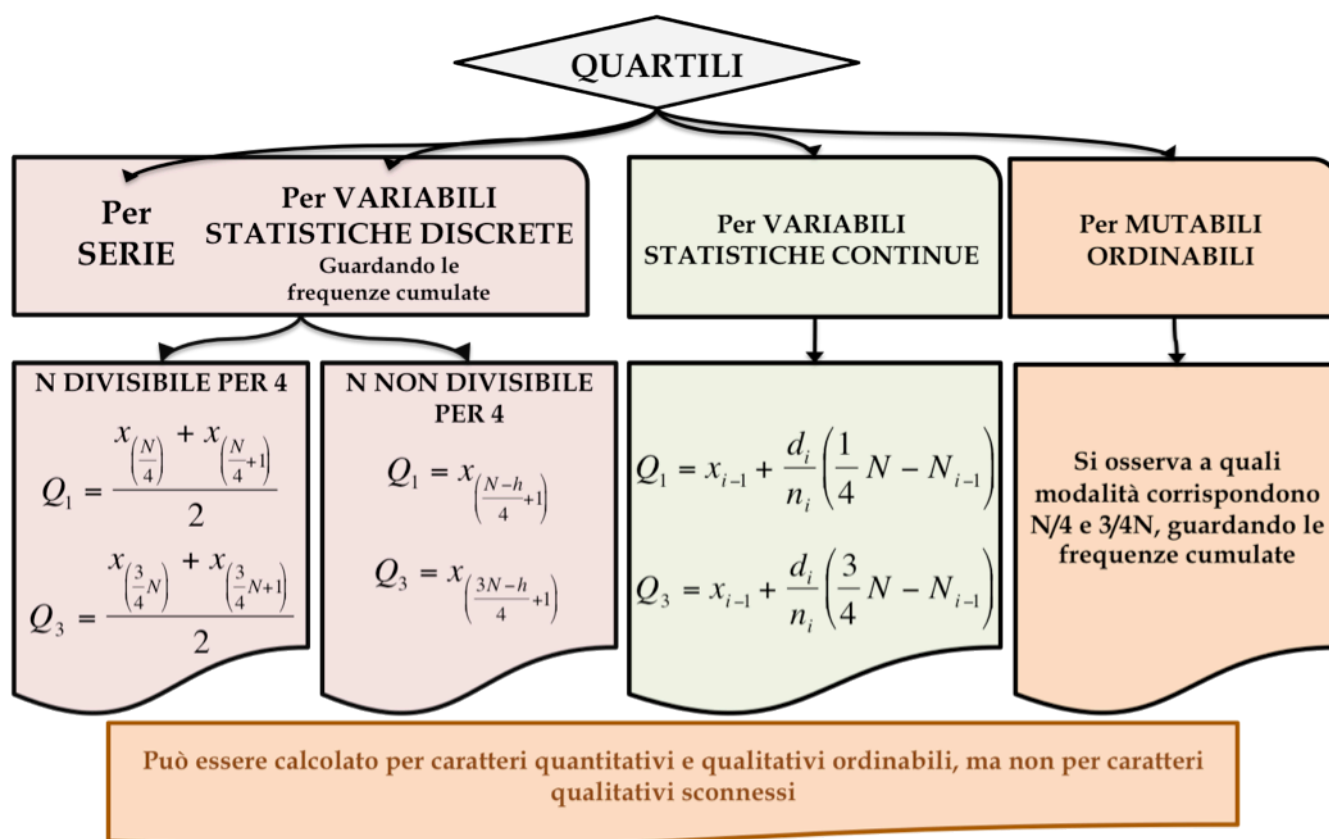
4. QUANTILI PER VARIABILI QUALITATIVE ORDINABILI

Anche in questo caso, come per la mediana, bisognerà fare un calcolo manuale sempre considerando le frequenze cumulate e purché si abbia a che fare con mutabili ordinabili.

Osservazione

I quartili, insieme ai centili, sono il tipo più usato di Quantili. Più in generale, può essere utile suddividere una qualunque distribuzione in q distribuzioni parziali, aventi, ognuna, la q -esima parte della numerosità. La modalità che fa da cardine fra due distribuzioni parziali successive è, appunto il quantile.

Ricapitoliamo quanto visto fino ad ora in uno schema:



7.B.6 Moda

La moda è la modalità che presenta la frequenza più elevata rispetto alle modalità adiacenti. Questo indicatore può essere calcolato sia per caratteri qualitativi che per caratteri quantitativi ma solo per distribuzioni (cioè quando i dati sono raffigurati in tabelle) e non per serie. Per distribuzioni di caratteri quantitativi discreti la moda è:

$$Mo = x_i \ni \begin{cases} n_i > n_{i-1} \\ n_i > n_{i+1} \end{cases}$$

Per le distribuzioni divise in classi in particolare la moda è la frequenza della classe che presenta la maggiore densità di frequenza e viene di solito associato al valore centrale della classe.

$$Mo = x_{i-1} - x_i \ni \begin{cases} h_i > h_{i-1} \\ h_i > h_{i+1} \end{cases}$$

Questo tipo di media non appartiene a nessuna delle due categorie di medie poiché non soddisfa nessuna delle due condizioni viste.

Vi sono due casi particolari di moda:

- Quando in una distribuzione si hanno più modalità con frequenza (o classi di modalità con densità di frequenza) più elevata rispetto a quelle adiacenti si dice che si ha una *DISTRIBUZIONE PLURIMODALE* e la modalità (o classe di modalità) con la frequenza più elevata viene detta *MODA PRINCIPALE*.
- Se la moda invece cade in uno degli estremi della distribuzione si dice che la distribuzione è *ZEROMODALE*.

7.C. MEDIE ANALITICHE.

Le medie analitiche sono quelle medie che soddisfano anche la condizione di invarianza di Chisini, cioè:

per le serie di valori: $f(x_1, x_2, \dots, x_N) = f(\bar{x}, \bar{x}, \dots, \bar{x})$

per le distribuzioni: $f(x_1, x_2, \dots, x_s) = f(\bar{x}, \bar{x}, \dots, \bar{x})$

dove f è un qualunque operatore matematico, al variare del quale si ottiene un tipo diverso di media analitica. In particolare se scegliamo $f = \text{somma}$ otteniamo la media aritmetica.

7.C.1 Media aritmetica

Definizioni:

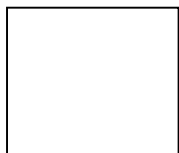
La media aritmetica è quel valore che sintetizza il carattere di un collettivo statistico lasciando invariato l'ammontare totale del carattere stesso, vale a dire che sostituendo la media a tutti i valori del carattere rimane invariato l'ammontare del carattere.

$$x_1 + x_2 + \dots + x_N = \underbrace{M + M + \dots + M}_{N \text{ volte}}$$

Cioè:

$$\sum_{i=1}^N x_i = N \cdot M$$

Da cui la media aritmetica μ per serie di valori:



Esempio: La media aritmetica della seguente serie di valori:

3 5 5 3 4 7 3 9 2

E' data da:

$$\mu = \frac{3 + 5 + 5 + 3 + 4 + 7 + 3 + 9 + 2}{9} = \frac{41}{9} = 4,56$$

Per una v.s. discreta rappresentata da una distribuzione di frequenza, la media aritmetica si dice ponderata ed è pari a:



Esempio: Per calcolare la media aritmetica della seguente distribuzione

Variabile statistica discreta relativa agli investimenti pubblicitari
di 30 aziende in migliaia di euro

x_i	n_i
20	6
35	9
50	5
75	10
	30

E' necessario aggiungere una colonna alla tabella con il valore di prodotti di modalità per frequenze

x_i	n_i	$x_i n_i$
20	6	120
35	9	315
50	5	250
75	10	750
	30	1435

E quindi la media è data da:

$$\mu = \frac{\sum_{i=1}^n x_i n_i}{N} = \frac{120 + 315 + 250 + 750}{30} = \frac{1435}{30} = 47,83$$

Infine quando si abbia a che fare con una *distribuzione statistica divisa in classi*, per calcolare la media aritmetica si parte da alcune ipotesi: Innanzitutto si ipotizza che la classe sia equamente distribuita e pertanto la media si calcolerà prendendo in considerazione il valore centrale della classe e ponendolo pari ad x'_i .

$$\mu = \frac{\sum_{i=1}^n x'_i n_i}{N} .$$

Esempio: La media per la variabile relativa all'altezza, già vista in precedenza:

Variabile statistica continua relativa all'altezza
di 200 ragazzi misurata in cm

x_i - x_{i+1}	n_i	x'_i	$x'_i n_i$
140 -150	2	145	290
150 -160	20	155	3100
160 -170	84	165	13860
170 -180	8	175	1400
180 -190	77	185	14245
190 -200	9	195	1755
	200		34650

$$\mu = \frac{\sum_{i=1}^n x'_i n_i}{N} = \frac{290 + 3.100 + 13.860 + 1.400 + 14.245 + 17.55}{200} = \frac{34.650}{200} = 173,25$$

N.B.: Quando le classi esterne sono aperte, il valore centrale viene scelto arbitrariamente, ipotizzando una ampiezza della classe opportuna.

7.C.2 Proprietà della media aritmetica

La media aritmetica gode di importanti proprietà:

- 1) La somma algebrica degli scarti dalla media aritmetica $x_i - \mu$ è uguale a zero. Cioè in formula:

$$\sum_{i=1}^s (x_i - \mu) = 0$$

Infatti nella formula invariante della media aritmetica:

$$x_1 n_1 + x_2 n_2 + \dots + x_s n_s = \mu (n_1 + n_2 + \dots + n_s)$$

Se portiamo tutto al secondo membro otteniamo proprio:

$$(x_1 - \mu)n_1 + (x_2 - \mu)n_2 + \dots + (x_s - \mu)n_s = \sum_{i=1}^s (x_i - \mu)n_i = 0$$

- 2) La somma dei quadrati degli scarti è un minimo

Ovvero è minore della somma degli scarti di x da un qualunque altro valore k diverso da 0

$$\sum_{i=1}^N (x_i - k)^2 = \sum_{i=1}^N (x_i - \mu)^2 + N(\mu - k)^2$$

dove $N(\mu - k)^2$ è una quantità positiva.

- 3) La media aritmetica non è invariante per traslazioni, cioè aggiungendo a ciascuna modalità della distribuzione una costante, anche la media risulterà essere traslata di un valore corrispondente proprio alla costante:

$$\text{se } x'_i = x_i + a \rightarrow \mu' = a + \mu$$

- 4) La media aritmetica non è invariante per cambiamenti di unità di misura, cioè :

$$x'_i = b x_i \rightarrow \mu' = b \mu$$

Dalle proprietà 3 e 4 derivano le ultime due:

- 5) La media aritmetica è associativa, ovvero la media generale può ottenersi come media ponderata delle medie parziali.

$$\mu = \frac{\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_k N_k}{N_1 + N_2 + \dots + N_k}$$

- 6) La media aritmetica non è invariante per la più generale trasformazione lineare $\alpha + \beta x$, cioè:

$$x' = \alpha + \beta y \rightarrow \mu' = \alpha + \beta \mu.$$

Questa proprietà è detta linearità della media aritmetica ed implica:

- la proprietà traslativa se si pone $\beta=1$
- il cambiamento di scala ponendo $\alpha=0$.

7.C.3 Osservazioni sulla media aritmetica

- La media aritmetica è considerata la media per eccellenza ma non sempre risulta la migliore. Ciò poiché per il suo calcolo si presuppone che l'andamento del fenomeno sia lineare e più precisamente che i dati siano in regressione aritmetica
- La media aritmetica è considerata il BARICENTRO di una serie o distribuzione poiché rappresenta il centro di tutti i valori.
- Viene anche definita VALORE DI TENDENZA CENTRALE cioè il valore verso il quale tendono tutti gli altri valori di una serie o distribuzione, cioè ancora il valore più vicino a tutti gli altri.

7.C.4 Caratteristiche della media aritmetica

- è espressa nella stessa unità di misura delle osservazioni
- soddisfa la condizione di internalità di CAUCHY
- soddisfa la condizione di CHISINI poiché lascia invariato l'ammontare complessivo del fenomeno:

$$\sum_{i=1}^n x_i = \sum_{i=1}^n \mu$$

- Se il valore della media corrisponde ad un valore della distribuzione allora la media si definisce MEDIA EFFETTIVA O MEDIA REALE, si invece non coincide con una delle modalità della distribuzione, la media aritmetica è detta MEDIA DI CONTO.
- Questa media è utilizzata esclusivamente per finalità di tipo additivo lineare.
- La media aritmetica NON è un indice robusto poiché tiene conto e viene influenzata dalla eventuale presenza di valori anomali.

ESERCIZI - 2

ESERCIZIO 5

Utilizzando la variabile statistica discreta costruita nell'ESERCIZIO 1, individuare:

1. il minimo e il massimo,
2. il valore centrale,
3. la moda,
4. la mediana e i quartili,
5. la media aritmetica.

ESERCIZIO 6

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 2, individuare:

1. il minimo e il massimo,
2. il valore centrale,
3. la moda,
4. la mediana e i quartili,
5. la media aritmetica.

ESERCIZIO 7

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 3, calcolare tutti gli indicatori statistici possibili.

ESERCIZIO 8

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 4, calcolare tutti gli indicatori statistici possibili.

8.A. INDICI DI VARIABILITA'

Le informazioni fornite dalle medie non sono sufficienti da sole a caratterizzare una distribuzione. È necessario, cioè, integrare tali informazioni con altre che tengano conto del grado di dispersione delle modalità intorno alla media o di quanto le modalità con cui il carattere è presente differiscano fra loro. Tali indici sono gli *indici di variabilità*.

Definizione: gli indici di variabilità sono quegli indici che danno una misura della distribuzione rispetto ad una media, o di quanto differiscano fra loro le modalità presenti nelle unità, basandosi su misure della diversità fra due quantità.

Caratteristiche: un indice di variabilità deve soddisfare le seguenti condizioni:

- se le x_i sono uguali fra loro la variabilità deve essere nulla:
se $x_1 = x_2 = \dots = x_N$ allora $Va=0$
- una misura di variabilità deve essere tanto più grande quanto più grande è la differenza fra i dati.

Le misure di variabilità si distinguono in:

1) INDICI BASATI SULL'ORDINAMENTO DEI DATI

- a. Campo di variazione
- b. Differenza interquartilica

2) MISURE DI DISPERSIONE

- a. Devianza
- b. Varianza
- c. Scarto quadratico medio

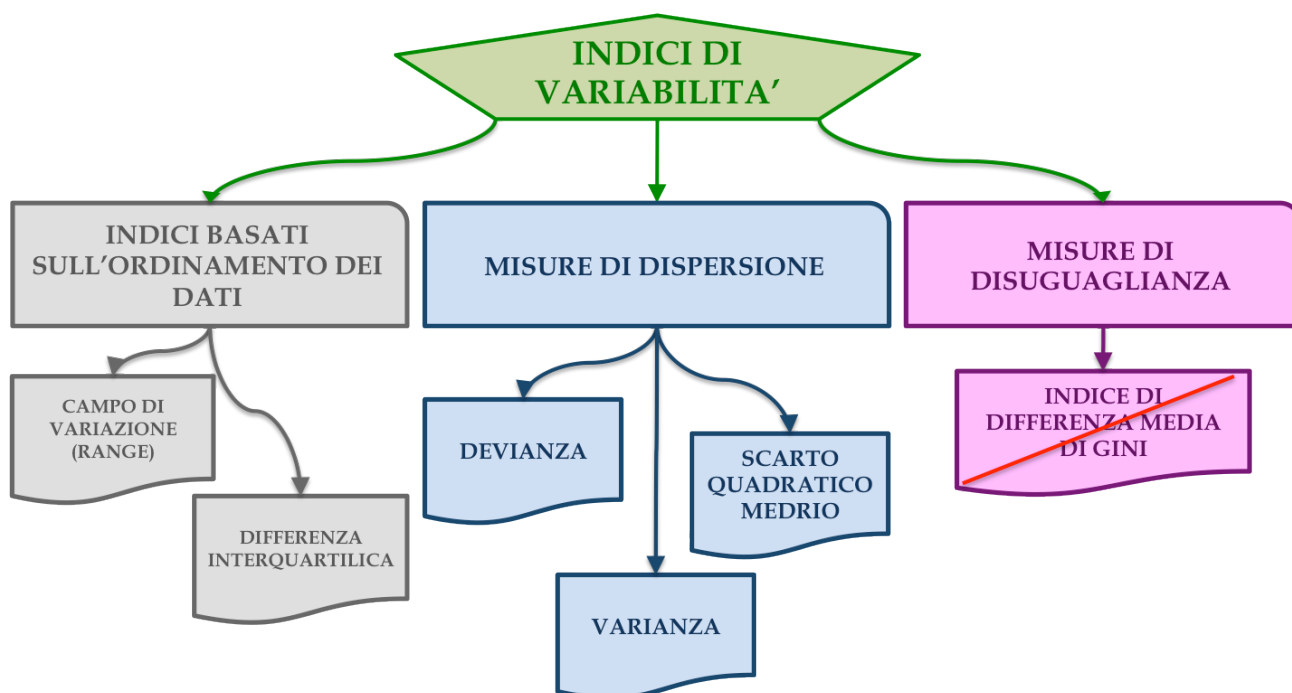
3) MISURE DI DISUGUAGLIANZA

- a. Indice di disuguaglianza di GINI

La seconda categoria è quella più utilizzata. Inoltre, tranne la prima categoria, che può essere nulla anche quando le modalità non sono tutte uguali fra loro, le altre soddisfano entrambe le due condizioni viste prima.

Infine, notiamo che gli indici di variabilità possono essere calcolati soltanto per caratteri quantitativi e nemmeno per i qualitativi ordinabili. Ciò poiché questi ultimi sono su scala ordinale ma non distanziata perciò per essi ha senso solo osservare se sono diversi o se uno è maggiore o minore dell'altro, ma non ha senso fare differenza né tantomeno rapporti. Per i caratteri qualitativi studieremo in seguito delle misure specifiche di calcolo che definiremo di MUTABILITÀ.

Possiamo schematizzare gli indici di variabilità come segue:



8.B. INDICI DI VARIABILITA' BASATI SULL'ORDINAMENTO DEI DATI

8.B.1 Campo di variazione

Si calcola facendo la differenza fra il valore massimo e il minimo della distribuzione:

$$W = x_{(N)} - x_{(1)} \quad \text{per serie} \quad \text{e} \quad W = x_{(s)} - x_{(1)} \quad \text{per distribuzioni.}$$

Notiamo che se $x_N = x_1$ allora $W=0$ (perciò vale la prima condizione). Non è però detto che valga la seconda condizione, infatti, se, ad esempio, nella serie vi sono valori anomali, questi aumentano la differenza fra i dati, ma il campo di variazione è lo stesso rispetto a quello di una distribuzione dove non vi sono valori anomali. Si dice, dunque, che questo indice NON È robusto.

Viene utilizzato principalmente nel campo industriale per controllare la qualità di un prodotto generato da una lavorazione in serie, il quale deve avere prefissati limiti di altezza o di peso. In casi del genere individua rapidamente la variabilità delle caratteristiche del prodotto per poter bloccare in tempo la produzione nel caso che la stessa ecceda i limiti che si sono prefissati.

8.B.2 Differenza interquartilica

È data dalla differenza fra il terzo e il primo quartile:

$$D_q = Q_3 - Q_1.$$

Questo indice ci dice come si comporta il 50% centrale della distribuzione ed è certamente più robusto del precedente, poiché non è influenzato da valori anomali, infatti si eliminano le code della distribuzione. Non soddisfa però la prima caratteristica, poiché il terzo e il primo quartile possono coincidere anche se la distribuzione è variabile.

N.B.: si possono costruire anche differenze interdeciliche o intercentiliche che avranno significato molto simile a quello di D_q .

8.C. MISURE DI DISPERSIONE

Per misure di dispersione si intendono tutti quegli indici che misurano la dispersione di ogni termine della distribuzione della media, sono, cioè, basati sugli scarti $x_i - \mu$. Le deviazioni sono positive per tutti i valori più grandi della media e negative per i valori più piccoli della media, inoltre, per la prima proprietà della media aritmetica, se si sommano gli scarti, questi si annullano. Se, però, si considerano i valori assoluti degli scarti, questi sono indicativi della variabilità di una serie, in particolare, più alti sono gli scarti in valore assoluto di ogni modalità rispetto alla media, maggiore è la variabilità, cioè appunto la dispersione delle modalità intorno al valore rappresentativo che è la media.

Un modo per eludere il problema dell'annullamento degli scarti, è quello di considerarne il quadrato. Gli indici di dispersione sono basati su tali dati. Tale scelta è legata al fatto che per la seconda proprietà della media aritmetica, la somma di tali valori è un minimo, questo garantisce che essi siano in grado di rappresentare la variabilità di una serie in modo sintetico più di qualunque altra metodologia.

Gli indici di questo tipo soddisfano, tutti, sia la prima condizione, che la seconda.

8.C.1 Devianza

La devianza è data da:

$$Dev(x) = \sum_{i=1}^N (x_i - \mu)^2 \quad \text{nel caso di serie}$$

$$Dev(x) = \sum_{i=1}^s (x_i - \mu)^2 n_i \quad \text{nel caso di distribuzioni.}$$

Questo indice presenta l'inconveniente di non essere espresso in termini della numerosità del collettivo N , esso, cioè, aumenta con l'aumentare della numerosità e ciò lo rende un valore non facilmente interpretabile: si dice che la devianza esprime un ammontare globale. Per tale motivo questo è l'indice di questa categoria meno utilizzato. Viene, invece, più spesso usato un altro indice che elimina l'inconveniente appena citato, tale indice è la varianza.

Esempio: considerando la variabile seguente

Variabile statistica discreta relativa agli investimenti pubblicitari
di 30 aziende in migliaia di euro

x_i	n_i
20	6
35	9
50	5
75	10
	30

e sapendo che la media di questa distribuzione è pari a 47,83, per calcolare la devianza dovremo aggiungere una colonna alla distribuzione con il calcolo degli scarti AL QUADRATO

x_i	n_i	$(x_i - \mu)^2 * n_i$
20	6	72600
35	9	81225
50	5	32000
75	10	30250
	30	216075

Quindi la devianza sarà data da:

$$Dev(x) = \sum_{i=1}^N (x_i - \mu)^2 n_i = 216075$$

8.C.3 Varianza

La Varianza si ottiene dalla Devianza, dividendo questa per N :

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad \text{nel caso di serie}$$

$$\sigma^2 = \frac{\sum_{i=1}^s (x_i - \mu)^2 n_i}{N} \quad \text{nel caso di distribuzioni.}$$

La varianza è espressa in termini di N , tuttavia, presenta l'inconveniente di essere espresso non nella unità di misura del carattere ma nella stessa unità al quadrato.

Esempio: Nel caso precedente la varianza è data da:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2 n_i}{N} = \frac{216075}{30} = 7202,5$$

8.C.4 Scarto quadratico medio

Lo SQM è dato dalla radice quadrata della Varianza e si indica con σ

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{nel caso di serie}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^s (x_i - \mu)^2 n_i}{N}} \quad \text{nel caso di distribuzioni.}$$

Questo indice di variabilità è espresso nella stessa unità di misura della distribuzione, pertanto è più sintetico e facilmente interpretabile rispetto alla varianza e alla devianza.

N.B.: Nella terminologia anglosassone lo scarto quadratico medio è definito STANDARD DEVIATION ed è quindi in italiano spesso definito come "deviazione standard".

Esempio: Sempre nel caso precedente, la varianza è data da

$$\sigma = \sqrt{\frac{\sum_{i=1}^s (x_i - \mu)^2 n_i}{N}} = 84,87$$

8.C.5 Proprietà delle misure di dispersione

Tutte le misure di dispersione sono INVARIANTI per traslazioni ma NON sono invarianti per cambiamenti di scala:

se $Y=a+bX$ allora $\sigma_y^2 = b^2 \sigma_x^2$

ESERCIZI - 3

ESERCIZIO 9

Utilizzando la variabile statistica discreta costruita nell'ESERCIZIO 1, calcolare:

1. Il campo di variazione,
2. La differenza interquartilica,
3. La devianza,
4. La varianza,
5. Lo scarto quadratico medio,

ESERCIZIO 10

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 2, calcolare:

1. Il campo di variazione,
2. La differenza interquartilica,
3. La devianza,
4. La varianza,
5. Lo scarto quadratico medio,

ESERCIZIO 11

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 3, calcolare tutti gli indicatori di variabilità possibili.

ESERCIZIO 12

Utilizzando la variabile statistica continua costruita nell'ESERCIZIO 4, calcolare tutti gli indicatori di variabilità possibili.

9. CONCENTRAZIONE

Un caso particolare della variabilità è la *Concentrazione*. La concentrazione è applicabile a tutti e soli i fenomeni cosiddetti *trasferibili*, ovvero i fenomeni per i quali è ipoteticamente possibile trasferire parte dell'ammontare globale da alcuni soggetti ed ipoteticamente dare ad altri fino al limite che uno possieda tutto l'ammontare del carattere e gli altri niente. Esempi di caratteri trasferibili sono il reddito, il numero di passeggeri di una compagnia aerea. Non sono invece trasferibili caratteri come il peso o l'altezza.

La concentrazione si rappresenta attraverso

- 1) un indice che misura il grado di concentrazione del carattere
- 2) una curva, chiamata curva di Lorenz.

9.1 Concetto e misura della concentrazione (nel caso di serie)

Quando i nostri valori sono rappresentati sotto forma di *serie*, per il calcolo della concentrazione occorre:

1. ordinare i valori in senso non decrescente (dai più poveri ai più ricchi):

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(N)}$$

2. determinare le *frazioni relative cumulate dei soggetti*:

$$p_i = \frac{i}{N}.$$

Per cui:

se $i=1$ $p_1=1/N$ cui corrisponde $x_{(1)}$

se $i=2$ $p_2=2/N$ cui corrisponde $x_{(2)}$

...

3. determinare le *frazioni relative cumulate delle intensità*. Si tratta di frazioni relative q_i , con al denominatore l'ammontare globale e al numeratore le intensità cumulate:

$$q_i = \frac{\sum_{j=1}^i x_{(j)}}{\sum_{i=1}^N x_{(i)}}$$

Se $i=1$	allora	$j=1$	e quindi	$\sum_{j=1}^i x_{(j)} = x_{(1)}$
Se $i=2$	allora	$j=1,2$	e quindi	$\sum_{j=1}^i x_{(j)} = x_{(1)} + x_{(2)}$
Se $i=3$	allora	$j=1,2,3$	e quindi	$\sum_{j=1}^i x_{(j)} = x_{(1)} + x_{(2)} + x_{(3)}$
...				
Se $i=N$	allora	$j=1,2,\dots,N$	e quindi	$\sum_{j=1}^i x_{(j)} = x_{(1)} + x_{(2)} + \dots + x_{(N)}$

Si osservi che:

- Se $p_i = q_i \quad \forall i$ vuol dire che le $x_{(i)}$ sono tutte uguali fra di loro. Si dice, allora, che vi è EQUIDISTRIBUZIONE

Esempio: Numero di passeggeri per compagnia aerea:

Compagnie Aeree	Numero di viaggiatori
A	50
B	50
C	50
D	50
E	50
TOT	250

Per prima cosa occorre ordinare le x . Inoltre da questo momento attribuiamo un numero progressivo al posto dei nomi delle compagnie poiché non interessano. Calcoliamo inoltre le p_i e le q_i :

i	x_i	Σx_j	p_i	q_i
1	50	50	0,2	0,2
2	50	100	0,4	0,4
3	50	150	0,6	0,6
4	50	200	0,8	0,8
5	50	250	1	1
TOT	250			

- Se $q_i = 0 \quad \forall i = 1, 2, \dots, N-1$ e $p_i = q_i \quad \forall i = N$ si dice che c'è MASSIMA CONCENTRAZIONE.

Riprendiamo lo stesso esempio:

Compagnie Aeree	Numero di viaggiatori
A	0
B	0
C	0
D	250
E	0
TOT	250

Ordiniamo e calcoliamo le p_i e le q_i :

i	x_i	Σx_j	p_i	q_i
1	0	0	0,2	0
2	0	0	0,4	0
3	0	0	0,6	0
4	0	0	0,8	0
5	250	250	1	1
TOT	250			

- In tutti gli altri casi sarà sempre $p_i \geq q_i \quad \forall i = 1, 2, \dots, N-1$ e $p_i = q_i \quad \forall i = N$. Si parlerà allora di CONCENTRAZIONE e occorrerà individuare un modo per poterla rappresentare e calcolare.

Esempio:

Compagnie Aeree	Numero di viaggiatori
A	50
B	70
C	30
D	100
E	90
TOT	340

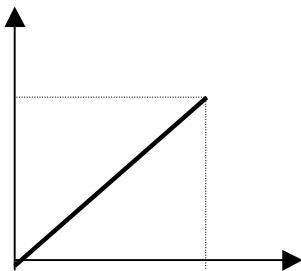
E quindi:

i	x_i	Σx_j	p_i	q_i
1	30	30	0,2	0,088235
2	50	80	0,4	0,235294
3	70	150	0,6	0,441176
4	90	240	0,8	0,705882
5	100	340	1	1
TOT	340			

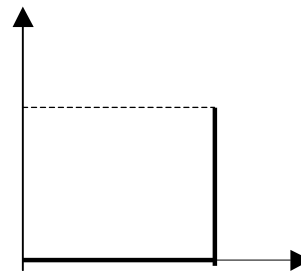
9.2 La curva di Lorenz

Un primo modo per rappresentare la concentrazione è il grafico definito, appunto, curva di Lorenz. Tale curva si ottiene ponendo in ascissa le p_i e in ordinata le q_i e assume la forma seguente nei vari casi:

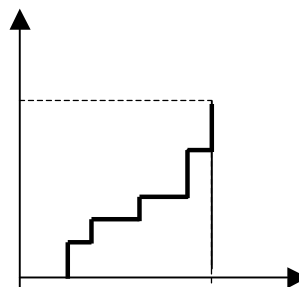
EQUIDISTRIBUZIONE quando $p_i = q_i$



MASSIMA CONCENTRAZIONE



Caso intermedio: SPEZZATA DI CONCENTRAZIONE



9.3 Indici di concentrazione

Indici assoluti di concentrazione

Gli indici di concentrazione saranno basati sulle quantità $p_i - q_i$. infatti:

- in caso di equidistribuzione $p_i - q_i = 0 \quad \forall i = 1, 2, \dots, N-1 \quad e \quad p_i = q_i \quad \forall i = N$
- in caso di massima concentrazione $p_i - q_i = p_i \quad \forall i = 1, 2, \dots, N-1 \quad e \quad p_i = q_i \quad \forall i = N$
- nei casi intermedi $p_i - q_i > 0 \quad \forall i = 1, 2, \dots, N-1 \quad e \quad p_i = q_i \quad \forall i = N$

perciò una prima misura assoluta di concentrazione è data da:

$$\sum_{i=1}^{N-1} (p_i - q_i) = \begin{cases} = \sum p_i & \text{massima concentrazione} \\ = 0 & \text{equidistribuzione} \\ > 0 & \text{concentrazione} \end{cases}$$

Indici relativi di concentrazione (rapporto di concentrazione di gini)

Per ottenere il corrispondente indice relativo di concentrazione basta rapportare l'indice appena calcolato al suo massimo. L'indice che si ottiene viene definito RAPPORTO DI CONCENTRAZIONE DI GINI:

$$R = \frac{\sum_{i=1}^{N-1} (p_i - q_i)}{\sum_{i=1}^{N-1} p_i} = \begin{cases} = 0 & \text{equidistribuzione} \\ = 1 & \text{massima concentrazione} \end{cases}$$

Il rapporto di concentrazione di Gini R può essere anche interpretato nel modo seguente:

ESERCIZI - 4

ESERCIZIO 13

Una banca possiede le seguenti sedi con i relativi profitti (anno 2008):

Sedi	Profitti
Bari	220
Como	320
Forlì	87
Genova	213

Calcolare l'indice di concentrazione e disegnare la curva di Lorenz

$$M_0 \mu_0 - M_e - Q_1 \approx Q_3 - M_e$$

11. SIMMETRIA E CURVA NORMALE

$$Q_1 - x_{(1)} \approx x_{(N)} - Q_3$$

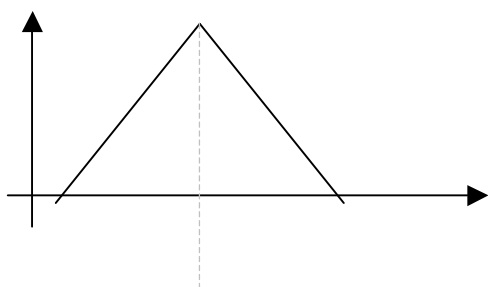
11.1 Simmetria e asimmetria

Dopo aver individuato gli indicatori statistici e gli indici di variabilità, un'altra informazione importante circa il carattere oggetto di analisi è se esso si distribuisce in maniera simmetrica oppure asimmetrica.

La simmetria è un indicatore della forma delle distribuzioni. Una distribuzione si dice simmetrica quando è possibile traslarla intorno alla sua mediana e vedere i due rami, quello destro e quello sinistro, perfettamente sovrapposti. In altri termini una distribuzione è simmetrica se le modalità che sono equidistanti dalla mediana hanno la stessa frequenza.

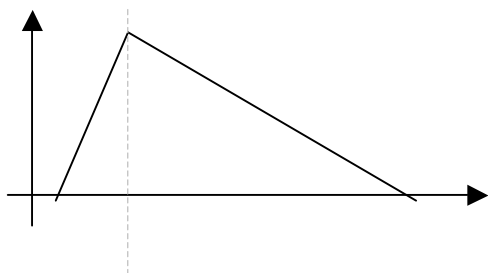
Una distribuzione simmetrica si presenta nel modo seguente e con delle specifiche caratteristiche:

Simmetria

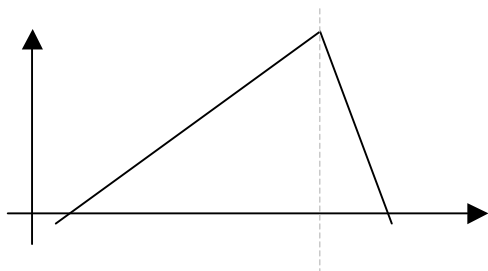


Quando una distribuzione non è simmetrica allora si dice che è asimmetrica e si può presentare con asimmetria positiva o negativa, rispettivamente :

Asimmetria positiva (la distribuzione si allunga verso valori positivi)



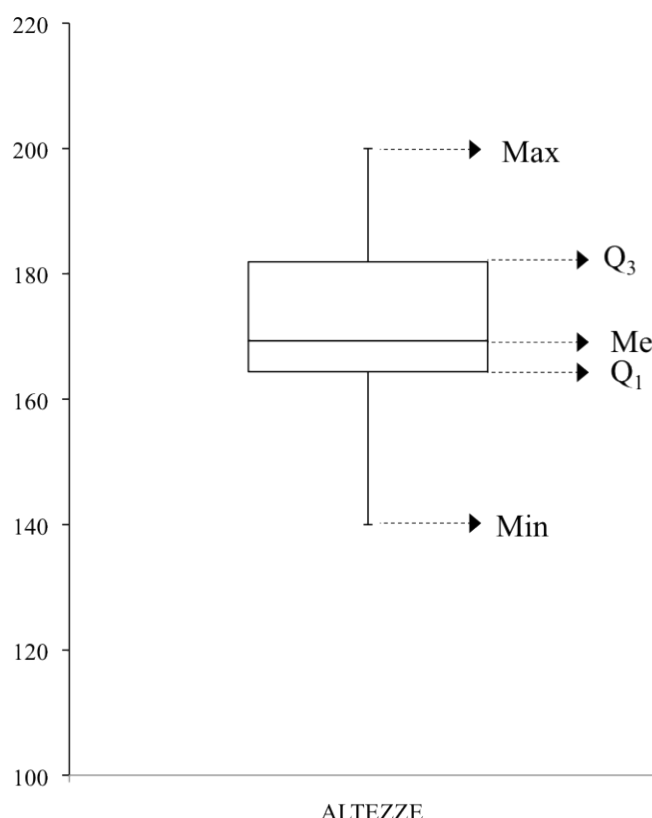
Asimmetria negativa (la distribuzione si allunga verso valori negativi)



11.2 Box plot

Attraverso le informazioni ottenute tramite mediane, quartili, minimo e massimo è possibile costruire una rappresentazione grafica della forma della distribuzione che permette di stabilire se la distribuzione è simmetrica o asimmetrica. Questa rappresentazione è definita box-plot, o diagramma a scatola.

A titolo esemplificativo, si osservi il box plot della variabile continua relativa al numero di altezze, precedentemente osservata:



11.3 La curva normale

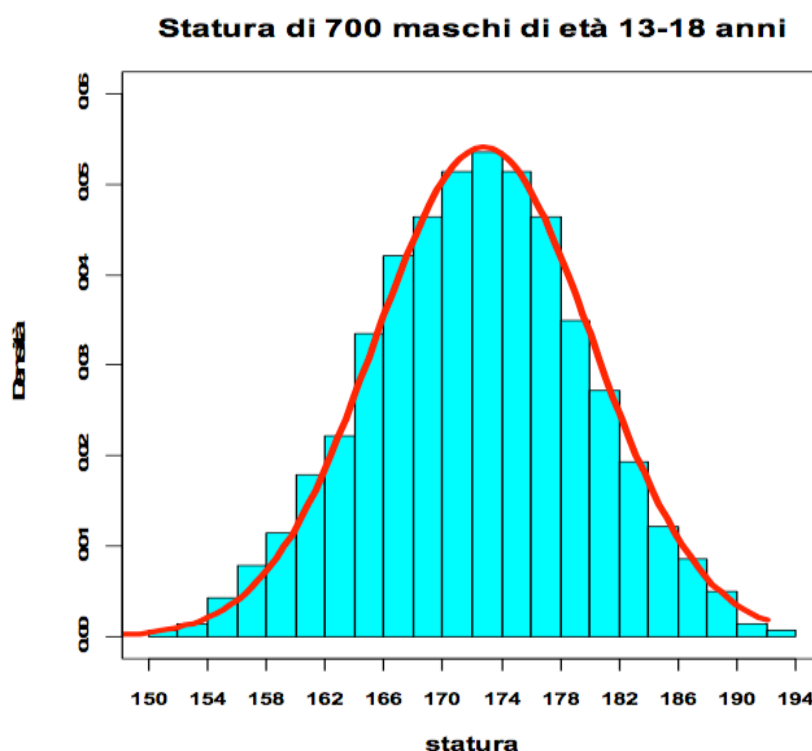
E' necessario accennare come in natura esistano un infinito numero di popolazioni possibili.

Da tempo si è notato che le misurazioni fatte in relazione alla gran parte dei fenomeni collettivi possono, in ultima analisi, essere ricondotte ad una sola distribuzione di frequenze, la cosiddetta *distribuzione normale*.

La distribuzione normale è una distribuzione continua avente una forma simmetrica e campanulare. E' possibile dimostrare che la maggior parte di distribuzioni empiriche, all'aumentare del numero di unità N , sono riconducibili ad una forma normale.

A titolo esemplificativo, consideriamo l'istogramma relativo alla statura di 700 ragazzi di età compresa fra 13-18 anni. Nel grafico si è aggiunta una *curva interpolatrice*, la cui forma è campanulare e simmetrica.

Gauss dimostrò che la maggior parte dei fenomeni collettivi, per grandi numerosità, assumono molto spesso questa forma distributiva. La ragione di questa circostanza è dovuta alle sue caratteristiche. Infatti, Gauss dimostrò che gli errori (ovvero gli scarti dalla media) tendono a distribuirsi in modo simmetrico intorno alla media e tendono a diminuire all'aumentare della distanza dalla media stessa.



A partire dalle ricerche di Gauss, si dimostra che l'espressione algebrica di questa curva dipende solamente dal numero di osservazioni N , dalla media μ e dallo scarto quadratico medio σ .

Senza soffermarci sulla espressione matematica della curva, vale la pena osservare che, al variare dei tre elementi caratterizzanti della curva, essa varia parzialmente in forma e posizione. In particolare:

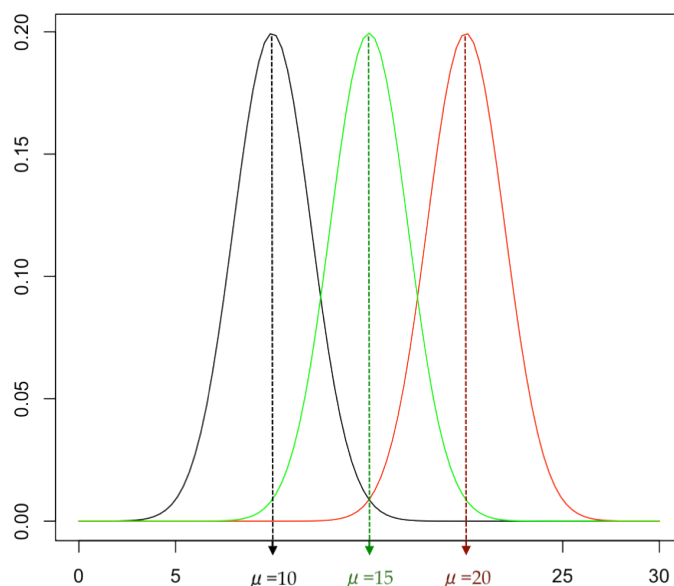
- al variare della media μ , varia la posizione della curva sull'asse x ;
- al variare dello scarto quadratico medio σ , varia la forma della curva, che diviene più o meno aguzza;
- al variare della numerosità N , varia l'area sotto la curva.

A titolo esemplificativo si osservino i grafici riportati alla pagina successiva, i quali mostrano cosa accade facendo variare di μ e σ .

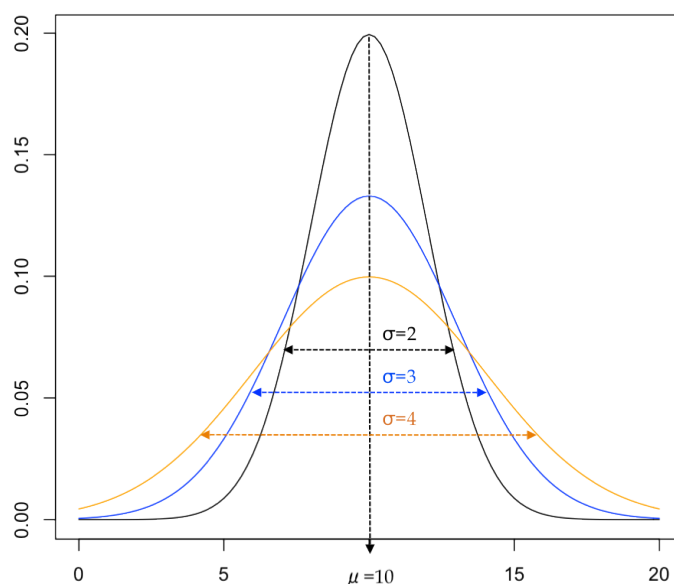
In statistica la curva normale viene utilizzata nel senso che, data una qualunque distribuzione, si cerca di verificare se essa, in effetti, si distribuisce normalmente, infatti, laddove questo accada, le successive analisi su quella distribuzione saranno più precise ed accurate, proprio perché si conoscono alcune caratteristiche di base del fenomeno, cioè quelle della curva normale.

Infine si osservi che non tutte le curve campanulari e simmetriche sono normali. La normalità è data dalla esistenza di una specifica forma funzionale che lega media, scarto quadratico medio e numerosità. Esistono diverse misure della *normalità* di una distribuzione, tali misure, però, esulano dagli scopi di questo corso.

Distribuzione normale



Distribuzione normale



ESERCIZI - 5

ESERCIZIO 14

Con riferimento alla distribuzione dell'esercizio 1, realizzare il box-plot e dire se la distribuzione può essere considerata simmetrica. In caso di asimmetria indicarne anche il tipo.

ESERCIZI RIEPILOGATIVI - A

ESERCIZIO 15

Di seguito si riporta il numero di dipendenti assenti da un'azienda agricola in 8 giorni lavorativi

5 16 10 4 3 14 13 2

Si misuri la concentrazione della serie.

ESERCIZIO 16

La distribuzione di frequenza assoluta degli stipendi mensili in euro dei 65 dipendenti di una determinata azienda è riportata nella seguente tabella

Stipendio	N. di dipendenti
1000 - 1100	8
1100 - 1200	10
1200 - 1300	16
1300 - 1400	14
1400 - 1500	10
1500 - 1800	5
1800 - 2500	2
Totale	65

1. si dica se la variabile X, oggetto di interesse, è qualitativa, sconnessa o ordinale, o quantitativa, discreta o continua;
2. Con riferimento ai risultati della rilevazione, sintetizzati nella tabella di frequenza, si specifichi la associata distribuzione di frequenza relativa e le corrispondenti frequenze cumulate, assolute e relative;
3. Si costruisca la rappresentazione grafica;
4. Si determini la moda, la mediana, i quartili e la media aritmetica della distribuzione;
5. Si determini, inoltre, lo scarto quadratico medio dei valori.

ESERCIZIO 17

Le seguenti temperature sono state rilevate durante il funzionamento di un frigorifero da laboratorio ad intervalli di due ore durante una giornata:

-1	3	0	0	2	-2	1	2	-3	1	0	1
----	---	---	---	---	----	---	---	----	---	---	---

1. Si determini la temperatura mediana e la temperatura media della giornata;
2. Si realizzi la variabile discreta dei valori e si rappresenti graficamente la distribuzione.

ESERCIZIO 18

Sia data la seguente tabella nella quale è riportato il reddito annuo di 20 individui con le frequenze con le quali è stato osservato:

Reddito (migliaia di euro)	N. individui
42,5	2
25	1
36	3
26	4
47	3
22,5	1
Totale	14

1. Si calcoli il reddito mediano e i quartili della distribuzione;
2. Si calcoli il reddito medio;
3. Se tutti i soggetti guadagnassero 1500 euro di più ogni anno, quale sarebbe il reddito medio di questi individui?

ESERCIZIO 19

Una regione delle Alpi è stata suddivisa in 6 sottoaree di uguale dimensione e conformazione. Per ogni sottoarea è stato svolto un censimento delle marmotte presenti. I risultati ottenuti sono stati i seguenti:

Area	n. marmotte
A	24
B	7
C	10
D	4
E	35
F	13
Totale	93

1. Si calcoli un indicatore statistico che sintetizzi questa distribuzione;
2. Si rappresenti graficamente la distribuzione;

ESERCIZIO 20

Di seguito è riportata la serie relativa ai valori di azotemia di 20 soggetti sani in mg/100ml

16	4	37	58	4	16	37	16	16	61
53	4	58	45	45	37	4	53	58	53

1. Si realizzi la distribuzione dei valori di azotemia;
2. Si calcoli il livello mediano, i quartili della distribuzione e il valore modale;
3. Si realizzi la funzione di ripartizione dei valori di azotemia;

ESERCIZIO 21

Con riferimento alla distribuzione dell'esercizio 2, realizzare il box-plot e dire se la distribuzione può essere considerata simmetrica. In caso di asimmetria indicarne anche il tipo.

12.A ANALISI BIVARIATA

12.A.1 Introduzione: serie e tabelle doppie

La ricerca scientifica non si limita alla descrizione dei singoli fenomeni, ognuno considerato indipendentemente dagli altri. Essa si interessa anche, e soprattutto, dell'analisi delle relazioni che ognuno di essi può avere con gli altri.

Quando in ciascuna unità statistica di un collettivo sono state rilevati due caratteri, che indichiamo con X e Y , i corrispondenti valori osservati saranno rappresentati da coppie di valori (x_i, y_i) con $i=1,2,\dots,N$. Quando N non è troppo grande, le osservazioni possono essere lasciate sotto forma di serie, dando origine alle, cosiddette, *serie doppie*:

X	Y
x_1	y_1
x_2	y_2
...	...
x_i	y_i
...	...
x_N	y_N

Quando, invece, N è molto grande e ci sono valori ripetuti, conviene disporre i dati rilevati in una *tabella a doppia entrata*, o *distribuzione doppia* (come è stato già accennato quando si è parlato di distribuzioni statistiche). Una distribuzione doppia si indica nel modo seguente:

X	Y						$TOT.$
	y_1	y_2	...	y_h	...	y_t	
x_1	n_{11}	n_{12}	...	n_{1h}	...	n_{1t}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2h}	...	n_{2t}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ih}	...	n_{it}	$n_{i.}$
...	...	...	...	...	...	...	...
x_s	n_{s1}	n_{s2}	...	n_{sh}	...	n_{st}	$n_{s.}$
$TOT.$	$n_{.1}$	$n_{.2}$	...	$n_{.h}$	...	$n_{.t}$	N

Dove:

- X e Y sono due caratteri che possono essere entrambi quantitativi, entrambi qualitativi, oppure uno quantitativo e uno qualitativo e le cui modalità sono rappresentate, rispettivamente nella *colonna madre* e nella *riga madre*;
- n_{ih} sono le frequenze assolute incrociate, cioè il numero di volte in cui si presentano congiuntamente la generica modalità x_i e la generica modalità y_h , quindi i è l'indice riferito alla X e h è l'indice riferito alla Y
- i rappresenta il numero di modalità della x e varia fra 1 e s ;
- h rappresenta il numero di modalità della y e varia fra 1 e t ;
- $n_{i.}$ e $n_{.h}$ sono le somme delle frequenze, rispettivamente, delle x_i e delle y_h e vengono definite *frequenze marginali*, in particolare ciascuna delle frequenze marginali $n_{i.}$ è data dalla somma delle corrispondenti frequenze per riga e ciascuna delle frequenze marginali $n_{.h}$ è data dalla somma delle corrispondenti frequenze per colonna, cioè:

$$n_{i1} + n_{i2} + \dots + n_{ih} + \dots + n_{it} = n_{i.}$$

$$n_{1h} + n_{2h} + \dots + n_{ih} + \dots + n_{sh} = n_{.h}$$

- N è la numerosità totale del collettivo, per la quale vale la seguente uguaglianza:

$$\sum_{i=1}^s n_{i.} = \sum_{h=1}^t n_{.h} = N;$$

Inoltre:

- la colonna contenente le modalità del carattere X è definita *colonna madre*;
- la riga contenente le modalità del carattere Y è definita *riga madre*;
- la colonna contenente le frequenze $n_{i.}$ è definita *colonna marginale* e rappresenta la colonna delle frequenze associate alla X indipendentemente dalla Y , ovvero la distribuzione semplice del carattere X ;
- la riga contenente le frequenze $n_{.h}$ è definita *riga marginale* e rappresenta la riga delle frequenze associate alla Y indipendentemente dalla X , ovvero la distribuzione semplice del carattere Y ;
- Ciascuna colonna interna rappresenta le frequenze della X condizionatamente ad una specifica modalità della Y , pertanto tali colonne sono definite delle *frequenze condizionate*;
- Ciascuna riga interna rappresenta le frequenze della Y condizionatamente ad una specifica modalità della X , pertanto anche tali colonne sono definite delle *frequenze condizionate*;

Quanto appena detto significa che, in teoria, è possibile estrarre da una tabella mista un certo numero di distribuzioni semplici, precisamente tante *distribuzioni semplici condizionate* quante sono le colonne e le righe delle frequenze condizionate più due *distribuzioni semplici marginali*, una per la X e una per la Y . Cioè, essendo s il numero di modalità della X e t il numero di modalità della Y , da una distribuzione doppia si possono ricavare $s+t+2$ distribuzioni semplici.

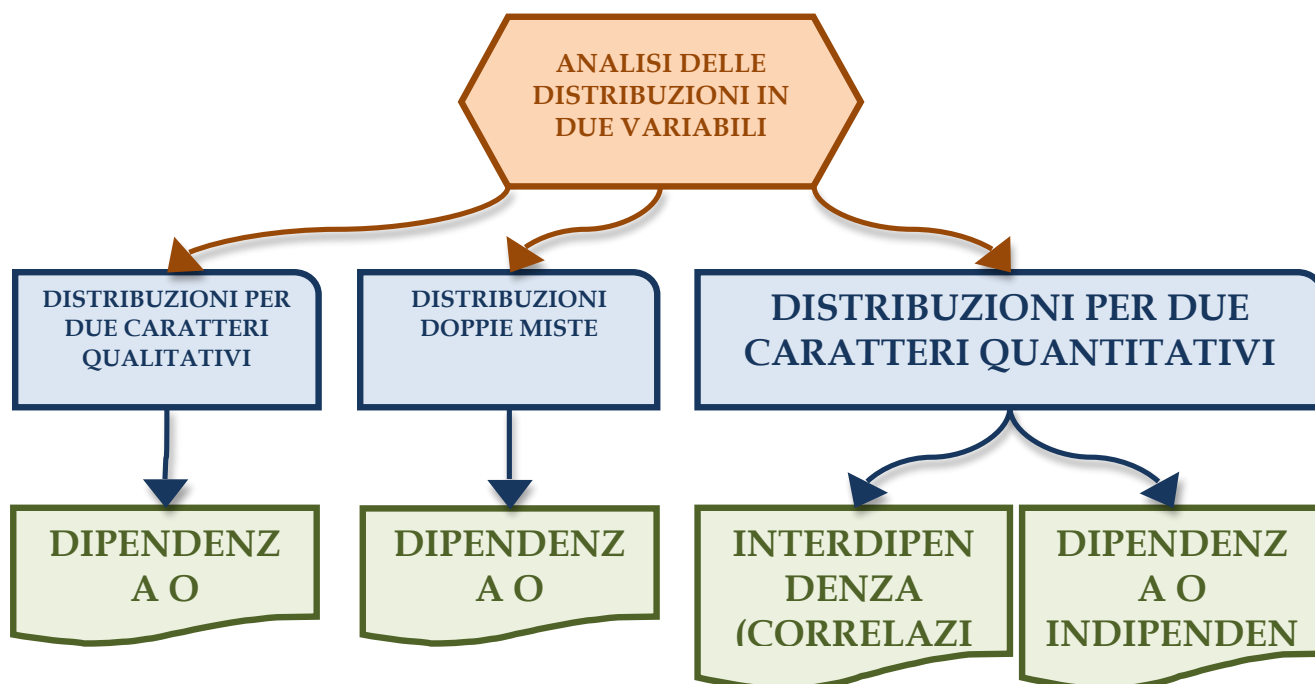
N.B.: Se uno dei due caratteri è qualitativo continuo, allora, al posto dei valori singoli x_i e y_h , ci saranno le classi di modalità $x_i - | x_{i+1}$.

12.A.2 La connessione

Lo studio delle relazioni esistenti fra caratteri è chiamato CONNESSIONE. Nella letteratura si distingue fra *misure di connessione* e *misure di correlazione* :

- Gli indici di *connessione* sono prevalentemente, ma non esclusivamente applicati allo studio dei caratteri qualitativi
- Gli indici di *correlazione* derivano dal confronto fra variabili quantitative e fanno parte di quello studio della statistica definito della REGRESSIONE. Questi indici, a loro volta, si distinguono in:
 - o Indici di dipendenza (o Regressione) in media
 - o Indici di correlazione (o di interdipendenza lineare)
 - o Indici di Regressione lineare.

L'analisi delle distribuzioni in due variabili può essere schematizzata nel modo seguente:



12.B. INDIPENDENZA E DIPENDENZA IN DISTRIBUZIONE (CONNESSIONE) PER CARATTERI QUALITATIVI

Cominciamo dalla analisi di distribuzioni in due variabili quando i due caratteri analizzati, X e Y , sono entrambi di tipo *qualitativo*. La tabella a doppia entrata generata da una situazione di questo genere è definita MUTABILE STATISTICA DOPPIA.

Analizzare contemporaneamente due variabili significa analizzare cosa succede alla Y (definita CONSEQUENTE) al variare della X (ANTECEDENTE). Fra le due variabili può non esservi alcuna relazione, e allora si parlerà di *indipendenza distributiva* (assenza di qualunque relazione), oppure potrà esserci una relazione di *dipendenza distributiva* (connessione in distribuzione).

12.B.1 Indipendenza in distribuzione

Si dice che fra due caratteri qualitativi (rappresentati nella forma di una mutabile statistica doppia, non esiste alcun legame e vi è, dunque, indipendenza, se le distribuzioni parziali condizionate della Y restano invariate al variare dei valori della X , cioè se vale:

$$n_{11}=n_{12}=\dots=n_{1.}$$

$$n_{21}=n_{22}=\dots=n_{2.}$$

...

$$n_{i1}=n_{i2}=\dots=n_{i.}$$

...

$$n_{s1}=n_{s2}=\dots=n_{s.}$$

...

$$n_{.1}=n_{.2}=\dots=N$$

Come sappiamo, le numerosità dei fenomeni possono, però, essere notevolmente variabili e non è perciò corretto parlare in termini assoluti, bensì in termini relativi. Si considera, allora, la generica riga i e si dice che c'è indipendenza se vale la seguente relazione:

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \dots = \frac{n_{ih}}{n_{.h}} = \dots = \frac{n_{it}}{n_{.t}} = \frac{n_{i.}}{N} \quad i = 1 \dots s$$

Cioè:

$$\frac{n_{ih}}{n_{.h}} = \frac{n_{i.}}{N} \quad \forall i = 1 \dots s \text{ e } \forall h = 1 \dots t$$

Si aggiunga che l'indipendenza è reciproca, cioè se Y è indipendente da X, anche X è indipendente da Y essendo le distribuzioni parziali tutte simili alla marginale e, quindi, simili fra loro. Si dice, allora, che fra le due variabili non sussiste alcuna relazione (assenza di connessione). Più precisamente si dice che il carattere Y è indipendente (in distribuzione) dal carattere X, o che Y non è connesso con X, se sono uguali fra loro tutte le distribuzioni parziali, secondo il carattere Y, corrispondenti alle varie modalità di X.

Se invece fra tutte le distribuzioni marginali ve ne sono almeno due fra loro diverse, allora diciamo che Y è connesso con X.

12.B.2 Connessione in distribuzione

Le situazioni di dipendenza vengono analizzate confrontando i valori delle frequenze con quelli delle frequenze stesse in caso di indipendenza, ovvero:

$$n_{ih} = \frac{n_{i.} \cdot n_{.h}}{N} = n_{ih}^*$$

Queste quantità n_{ih}^* sono chiamate FREQUENZE TEORICHE o n_{ih} ATTESE e rappresentano il valore che ciascuna frequenza dovrebbe avere nel caso di indipendenza distributiva.

Esempio: partendo da una mutabile doppia in cui siano date soltanto le distribuzioni marginali:

X	Y			TOT.
	D	E	F	
A				22
B				15
C				13
TOT.	12	18	20	50

e si voglia costruire la mutabile delle frequenze teoriche

X	Y			TOT.
	D	E	F	
A	n_{11}^*	n_{12}^*	n_{13}^*	22
B	n_{21}^*	n_{22}^*	n_{23}^*	15
C	n_{31}^*	n_{32}^*	n_{33}^*	13
TOT.	12	18	20	50

Si otterrà la seguente mutabile doppia:

X	Y			TOT.
	D	E	F	
A	5	8	9	22
B	4	5	6	15
C	3	5	5	13
TOT.	12	18	20	50

Definita **distribuzione teorica**, in cui:

$$n_{11}^* = \frac{12 \cdot 22}{50} = 5$$

$$n_{12}^* = \frac{18 \cdot 22}{50} = 8$$

...

Nella realtà è molto difficile che una distribuzione doppia sia formata dai valori attesi o teorici. Quando ciò non accade, allora, si studia qual è l'allontanamento della distribuzione dalla situazione teorica attraverso un indice costruito a partire da uno scarto fra due quantità significative ed è piuttosto immediato comprendere il motivo per il quale la differenza da utilizzare in questo caso è quella fra frequenze teoriche e frequenze empiriche. Queste differenze vengono definite **contingenze** e si indicano come segue:

$$n_{ih} - n_{ih}^* = c_{ih} \rightarrow \begin{cases} = 0 & \forall i \text{ e } \forall h & \text{INDIPENDENZA DISTRIBUTIVA DI Y DA X E VICEVERSA} \\ \neq 0 & \text{esiste un legame} \rightarrow \begin{cases} > 0 & \forall i \text{ e } \forall h & \text{CONNESSIONE POSITIVA} \\ < 0 & \forall i \text{ e } \forall h & \text{CONNESSIONE NEGATIVA} \end{cases} \end{cases}$$

Si dice che vi è **MASSIMA CONNESSIONE** se ad ogni modalità di X corrisponde solo una modalità delle Y e viceversa. Per esempio:

X	Y			TOT.
	D	E	F	
A	22	-	-	22
B	-	15	-	15
C	-	-	13	13
TOT.	22	15	13	50

oppure

X	Y			TOT.
	D	E	F	
A	-	22	-	22
B	-	-	15	15
C	13	-	-	13
TOT.	13	22	15	50

oppure

X	Y			TOT.
	D	E	F	
A	-	-	22	22
B	15	-	-	15
C	-	13	-	13
TOT.	15	13	22	50

Naturalmente è facile comprendere che la situazione di massima distribuzione è possibile solamente per le mutabili doppie quadrate, in cui, cioè $s=t$.

12.B.3 Indici di connessione: L'indice Chi-quadrato di Pearson

Vediamo come si costruiscono, a partire dalle contingenze, gli indici per il calcolo della connessione. Per prima cosa si noti che, gli indici che vedremo, sono validi nel caso in cui $n_{ih}^* \geq 5$ $\forall i$ e $\forall h$. Se ciò non si verifica sarà necessario assemblare alcune modalità.

Gli indici di connessione sono tutti strutturati in modo da assumere valore pari a :

- zero in caso di connessione nulla o indipendenza.
- Massimo o 1 in caso di perfetta connessione, unilaterale o bilaterale.

L'indice di connessione che andremo a calcolare è un indice basato sui confronti mediante differenza fra le frequenze della distribuzione effettiva e quelle teoriche.

L'indice Chi-quadrato (χ^2) effettua la sintesi tramite le contingenze al quadrato ed è dato da

$$\chi^2 = \sum_{i=1}^s \sum_{h=1}^t \frac{(n_{ih} - n_{ih}^*)^2}{n_{ih}^*} \quad (\chi^2 \text{ assoluto})$$

Il campo di variazione del χ^2 è il seguente:

$$0 \leq \chi^2 \leq N \min(s-1; t-1)$$

Il χ^2 , per come è costruito, non può mai essere minore di 0.

Il corrispondente indice relativo è dato da:

$$\frac{\chi^2}{\max \chi^2} = \begin{cases} 0 = \text{indipendenza} \\ 1 = \text{massima connessione} \end{cases} \quad (\chi^2 \text{ relativo})$$

N.B.: È conveniente calcolare questo indice attraverso una *formula abbreviata* che si ottiene dallo sviluppo del quadrato ed è data da:

$$\chi^2 = \sum_{i=1}^s \sum_{h=1}^t \frac{(n_{ih} - n_{ih}^*)^2}{n_{ih}^*} = N \left(\sum_{i=1}^s \sum_{h=1}^t \frac{n_{ih}^2}{n_{i.} n_{.h}} - 1 \right)$$

Esempio: Si voglia verificare se esiste connessione fra i due caratteri X e Y rappresentati nella seguente mutabile statistica doppia:

X	Y			TOT.
	D	E	F	
A	3	12	5	20
B	7	10	12	29
C	5	0	20	20
TOT.	10	22	37	69

Per prima cosa osserviamo che, poiché più di una frequenza ha valore minore di 5, allora non è possibile calcolare il χ^2 senza aver prima accorpato opportunamente le frequenze. Ad esempio accorpriamo per colonna:

X	Y		TOT.
	D+E	F	
A	15	5	20
B	17	12	29
C	5	20	20
TOT.	32	37	69

Per il calcolo del χ^2 si costruisce la seguente tabella nella quale, al posto delle frequenze vi sono i

valori $\frac{n_{ih}^2}{n_i \cdot n_h}$:

X	Y		TOT.
	E	F	
A	0,352	0,034	0,385
B	0,311	0,134	0,446
C	0,039	0,541	0,580
TOT.			1,411

Quindi il χ^2 è dato da:

$$\chi^2 = 69 (1,411-1) = 69 \cdot 0,411 = 28,359$$

questo valore, assoluto, ci fa capire che esiste una connessione positiva, ma non sappiamo in percentuale quanto vale. A tale scopo calcoliamo il corrispondente indice relativo:

$$\frac{\chi^2}{\max \chi^2} = \frac{28,359}{69 \min[1,2]} = \frac{28,359}{69} = 0,411 \approx 40\%$$

Ciò significa che la connessione è piuttosto bassa.

ESERCIZI - 6

ESERCIZIO 22

Alcuni ricercatori hanno ipotizzato una possibile relazione tra il sesso e l'incidenza dell'epatite C in un gruppo di 400 pazienti di un ospedale. Pertanto su tutti i 400 pazienti rilevano i dati relativi all'incidenza e al sesso. Esiste una relazione tra questi due criteri di classificazione?

Sesso	HCV+	HCV-	Totale
Maschi	31	194	225
Femmine	48	127	175
Totale	79	321	400

12.C. REGRESSIONE PER DISTRIBUZIONI DOPPIE MISTE:

12.C.1 Indipendenza in distribuzione

Quando almeno uno dei due caratteri è quantitativo, allora il χ^2 non è più idoneo a misurare la connessione fra due caratteri, poiché è possibile usare altri indici che sfruttino le informazioni aggiuntive provenienti dalla variabile quantitativa.

Supponiamo che fra i due caratteri si abbia:

X = variabile quantitativa

Y = variabile qualitativa

In questo caso si valuta se ed in quale misura le medie del carattere quantitativo X, variano al variare delle modalità del carattere qualitativo Y. Il carattere X, in questo caso, è definito *conseguente*, poiché è quello che potenzialmente può variare al variare dell'altro carattere, Y, il quale, poiché è quello che causa la dipendenza, è definito *antecedente*.

Diremo che X è indipendente in media da Y se, al variare delle modalità di Y, le medie condizionate di X rimangono costanti, diremo invece che X è dipendente da Y, nel caso contrario.

Per il calcolo della Regressione un primo passo è quello di confrontare le MEDIE CONDIZIONATE. Dunque, in una tabella doppia (mista) in cui si hanno t medie condizionate della X dalla Y, si ha indipendenza in media quando:

$$\mu_1 = \mu_2 = \dots = \mu_h = \dots = \mu_t = \mu$$

Laddove ciascuna media μ_h è una delle medie per colonna della X condizionata ad una specifica modalità della Y.

X	Y						TOT.
	y_1	y_2	...	y_h	...	y_t	
x_1							
x_2							
...							
x_i	μ_1	μ_2	...	μ_h	...	μ_t	μ
...							
x_s							
TOT.	n_1	n_2	...	n_h	...	n_t	N

se anche una sola di queste medie non è uguale alle altre, allora esiste dipendenza in media (connessione) e più precisamente in questo caso, si parla di *regressione*.

12.C.2 Dipendenza in media: il rapporto di correlazione η^2 di Pearson:

Una misura della dipendenza in media è ottenuta confrontando le devianze di ciascuna media μ_h dalla media generale μ . L'indice di dipendenza in media, cioè, è basato su una devianza, definita, *devianza fra i gruppi*, perché misura proprio come si distribuiscono le medie dei vari gruppi della X, ovvero delle distribuzioni della X condizionate alla Y, rispetto alla media generale della X:

$$Dev_{fra} = \sum_{h=1}^t (\mu_h - \mu)^2 n_{.h}$$

L'indice in questione si chiama "eta-quadro" ed è un indice relativo che ha la devianza fra i gruppi al numeratore e il massimo di questa devianza al denominatore, laddove il massimo di questa variabilità è la devianza totale della X:

$$Dev(X) = \sum_i^s \sum_h^t (x_i - \mu)^2 n_{ih}$$

L'indice è il seguente:

$$\eta^2 = \frac{Dev_{fra}}{Dev(X)}.$$

Questo indice varia fra 0 e 1, perché è il rapporto fra una parte non negativa e il totale ed esprime quanta parte della variabilità complessiva di Y è attribuibile alla diversità derivante dalle medie di X che (eventualmente) variano con il variare delle modalità di Y.

In particolare:

- $\eta^2 = 0$ vuol dire *indipendenza in media*
- $\eta^2 \neq 0$ vuol dire *dipendenza*
- $\eta^2 = 1$ vuol dire *massima dipendenza*, cioè ad ogni valore di X corrisponde uno ed un solo valore di Y.

Esempio: Si voglia studiare se esiste una connessione fra i due caratteri "Voto in statistica" e "Corso di Laurea" relativi a 35 studenti iscritti ad un corso di statistica alla Facoltà di Economia.

Voto in statistica	Corso di laurea			Totali
	E.C.	E.A.	M.R.	
18	3	-	-	3
22	2	1	-	3
24	1	9	6	16
28	4	4	1	9
30	2	1	1	4
Totali	12	15	8	35

Poiché:

X = variabile quantitativa = Voto in statistica.

Y = variabile qualitativa = Corso di Laurea

Si tratta di una tabella mista perciò possiamo stabilire se il voto dipende mediamente dal corso di laurea. Useremo allora η^2 . Per poterlo calcolare abbiamo bisogno prima di calcolare tutte le quantità che ci servono. Cioè, per la Devianza fra i gruppi, abbiamo bisogno delle medie per gruppo e della media generale, quindi calcoleremo la devianza fra, poi la devianza totale e, infine, l'indice.

$$\mu_{EC} = \frac{18 \cdot 3 + 22 \cdot 2 + 24 \cdot 1 + 28 \cdot 4 + 30 \cdot 2}{12} = 24,5$$

$$\mu_{EA} = \frac{22 \cdot 1 + 24 \cdot 9 + 28 \cdot 4 + 30 \cdot 1}{15} = 25,3$$

$$\mu_{MR} = \frac{24 \cdot 6 + 28 \cdot 1 + 30 \cdot 1}{8} = 25,2$$

$$\mu = \frac{18 \cdot 3 + 22 \cdot 3 + 24 \cdot 16 + 28 \cdot 9 + 30 \cdot 4}{35} = 25,03$$

$$\begin{aligned} Dev_{fra} &= \sum_{h=1}^t (\mu_h - \mu)^2 n_{.h} = (24,5 - 25,03)^2 12 + (25,3 - 25,03)^2 15 + (25,2 - 25,03)^2 8 = \\ &= 3,371 + 1,093 + 0,387 = 4,850 \end{aligned}$$

Il calcolo della devianza totale della X è un pochino più laborioso, perché occorre calcolare gli scarti ponderati al quadrato per ciascuna delle celle della tabella doppia.

Conviene pertanto realizzare una tabella come quella delle frequenze, ma contenente i valori degli scarti ponderati al quadrato:

Voto in Statistica	Corso di laurea			Totale
	E.C.	E.A.	M.R.	
18	$(18-25)^2 \cdot 3 = 147$	-	-	147
22	$(22-25)^2 \cdot 2 = 18$	$(22-25)^2 \cdot 1 = 9$	-	27
24	$(24-25)^2 \cdot 1 = 1$	$(24-25)^2 \cdot 9 = 9$	$(24-25)^2 \cdot 6 = 6$	16
28	$(28-25)^2 \cdot 4 = 36$	$(28-25)^2 \cdot 4 = 36$	$(28-25)^2 \cdot 1 = 9$	81
30	$(30-25)^2 \cdot 2 = 50$	$(30-25)^2 \cdot 1 = 25$	$(30-25)^2 \cdot 1 = 25$	100
Totale	252	79	40	371

$$Dev(X) = \sum_i^s \sum_h^t (x_i - \mu)^2 n_{ih} = 371$$

Pertanto l'indice di Pearson è dato da:

$$\eta^2 = \frac{Dev_{fra}}{Dev(X)} = \frac{4,850}{371} = 0,013 \approx 1,3\%$$

Possiamo quindi dedurre che, l'appartenenza ad uno specifico corso di laurea (gruppo), non influisce sul voto in statistica.

ESERCIZI - 7

ESERCIZIO 23

Un'azienda produce lampadine in tre diversi stabilimenti (A, B e C). Si decide di valutare la durata in ore delle lampadine prodotte: a tal fine viene misurata la durata di 20 lampadine prodotte nello stabilimento A, di 17 lampadine prodotte nello stabilimento B e di 31 prodotte nello stabilimento C. Si misuri l'entità della dipendenza fra la durata delle lampadine e lo stabilimento.

Durata delle lampadine	Stabilimento			Totale
	STAB.A	STAB.B	STAB.C	
0 - 50	3	2	4	9
50 - 150	6	6	9	21
150 - 250	7	5	10	22
250 - 300	4	4	8	16
Totale	20	17	31	68

N.B.: per effettuare i calcoli si consideri il valore centrale di ciascuna classe.

12.D. CORRELAZIONE (=INTERDIPENDENZA) PER VARIABILI STATISTICHE DOPPIE (DISTRIBUZIONI PER CARATTERI QUANTITATIVI)

Quando si debba studiare il comportamento congiunto di due caratteri X e Y entrambi di natura quantitativa, laddove si supponga che NON sussista indipendenza fra i due, possono essere studiati due diversi tipi di connessione:

- *interdipendenza*, cioè l'eventuale circostanza che i due si influenzino reciprocamente
- *dipendenza di uno dall'altro*.

Nel primo caso, in particolare, per individuare la relazione reciproca esistente fra i due caratteri, occorre realizzare un indice che li tenga in considerazione entrambi, infatti, le informazioni relative ai singoli caratteri separatamente non sono sufficienti, da sole, a fornire informazioni significative. In particolare, una misura della interdipendenza è data dalla variabilità congiunta, cioè da una misura di come variano contemporaneamente i due caratteri rispetto alle proprie medie.

12.D.1 Covarianza (misura assoluta di interdipendenza)

La Covarianza è la varianza comune alle due variabili (quantitative) ed è, quindi, una misura della variabilità congiunta fra la X e la Y. Si indica con

$$\text{Cov}(X;Y) \quad \text{oppure} \quad \sigma_{XY}$$

e vale:

$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^N (y_i - \mu_y)(x_i - \mu_x)}{N} \quad \text{per le serie}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^s \sum_{h=1}^t (y_i - \mu_y)(x_h - \mu_x) n_{ih}}{N} \quad \text{per le distribuzioni.}$$

Per il modo in cui è costruita, la covarianza assume i seguenti valori:

- $Cov(X,Y) > 0$ in caso di relazione positiva, cioè quando i valori della X e della Y crescono entrambi;
- $Cov(X,Y) < 0$ in caso di relazione negativa, cioè quando i valori della X crescono e quella della Y decrescono, o viceversa;
- $Cov(X,Y) = 0$ in caso di assenza di relazione fra le due variabili.

Inoltre si dimostra che il campo di variabilità della covarianza è il seguente:

$$-\sqrt{\sigma_X^2 \cdot \sigma_Y^2} \leq Cov(X,Y) \leq \sqrt{\sigma_X^2 \cdot \sigma_Y^2}$$

12.D.2 Coefficiente di correlazione di Bravais - Pearson (misura relativa di interdipendenza)

La covarianza è una misura assoluta di interdipendenza. Come siamo ormai abituati a fare, calcoliamo la corrispondente misura relativa, rapportando la covarianza al suo massimo. L'indice che ne deriva prende il nome di coefficiente di correlazione di Bravais - Pearson e si indica come segue:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

Si noti che : $-1 \leq \rho_{XY} \leq 1$ e, più precisamente:

$\rho_{XY} = -1$ vuol dire $Cov(X,Y) < 0$ PERFETTA DISCORDANZA fra i caratteri

$\rho_{XY} = 1$ vuol dire $Cov(X,Y) > 0$ PERFETTA CONCORDANZA fra i caratteri

$\rho_{XY} = 0$ vuol dire $Cov(X,Y) = 0$ INCORRELAZIONE o INDIPENDENZA LINEARE.

Per esteso tale indice vale:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{\frac{\sum_{i=1}^N (y_i - \mu_Y)(x_i - \mu_X)}{N}}{\sqrt{\frac{\sum_{i=1}^N (x_i - \mu_X)^2}{N} \frac{\sum_{i=1}^N (y_i - \mu_Y)^2}{N}}} \quad \text{in caso di serie}$$

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{\frac{\sum_{i=1}^s \sum_{h=1}^t (y_i - \mu_Y)(x_h - \mu_X) n_{ih}}{N}}{\sqrt{\frac{\sum_{h=1}^t (x_h - \mu_X)^2 n_{.h}}{N} \frac{\sum_{i=1}^s (y_i - \mu_Y)^2 n_{i.}}{N}}}$$

in caso di distribuzioni.

Infine si osservi che, se si semplificano tutte le N, si ottiene:

$$\rho_{XY} = \frac{\sum_{i=1}^s \sum_{h=1}^t (y_i - \mu_Y)(x_h - \mu_X) n_{ih}}{\sqrt{\sum_{h=1}^t (x_h - \mu_X)^2 n_{.h} \sum_{i=1}^s (y_i - \mu_Y)^2 n_{i.}}} = \frac{Codev(X;Y)}{\sqrt{Dev(X)Dev(Y)}}$$

Dove è stata definita una nuova quantità, la **Codevianza**. Questa, analogamente alla devianza, è pari alla Covarianza per N.

ESERCIZI - 8

ESERCIZIO 24

I seguenti dati rappresentano i pesi (in Kg) e i livelli di glucosio nel sangue (mg/100 ml) appartenenti a 6 maschi adulti apparentemente sani.

Glucosio (X)	Pesi (Y)
108	64
104	73
102	82,1
121	95,7
85	78,9
108	83,9

Si verifichi se esiste correlazione fra i due caratteri.

12.E. DIPENDENZA FRA VARIABILI QUANTITATIVE: REGRESSIONE LINEARE

Quando fra le due variabili quantitative si vuole studiare la relazione di dipendenza di una dall'altra si svolge quella che si definisce una analisi di regressione lineare. Nello studio di dipendenza è necessario definire quale variabile si ritiene sia la causa della dipendenza e quale sia la conseguenza. Generalmente la variabile causa è definita antecedente e la dipendente è definita conseguente. Inoltre di solito si usa porre la X come antecedente e la Y come conseguente.

La regressione lineare cerca di individuare la relazione funzionale che lega la X alla Y nel modo seguente:

$$Y = f(X).$$

Si osservi, però, che non si può considerare la relazione f come una relazione perfettamente deterministica fra la Y e la X. Nella relazione si considera, anche, la presenza di una componente accidentale, o residuale, che viene indicata con il simbolo ε e che entra a far parte della relazione nel modo seguente:

$$Y = f(X) + \varepsilon.$$

 COMPONENTE RESIDUALE, ACCIDENTALE O DI DISTURBO

In tal modo la relazione tiene, quindi conto anche di tutti quei fattori della Y che non possono essere determinati tramite la X e che, quindi, "disturbano" la perfetta relazione fra i due caratteri. Per tale ragione la ε viene anche definita COMPONENTE DI DISTURBO. Dalla relazione appena vista segue che vale anche la seguente:

$$\varepsilon = Y - f(X)$$

se in questa relazione, poniamo:

$$f(X) = \hat{Y}^* = \text{FUNZIONE TEORICA}$$

otteniamo:

$$\varepsilon = Y - \hat{Y}^*.$$

Dove $\hat{Y}^*$ rappresenta la funzione che lega le due variabili e che si vuole studiare.

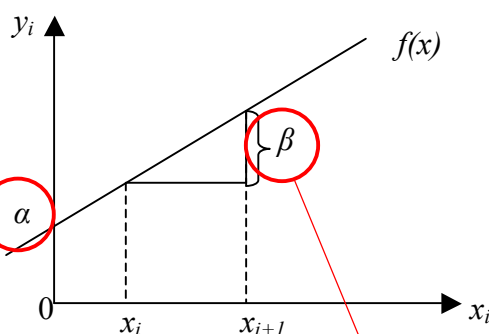
La funzione $f(x)$ che lega la X e la Y viene espressa nel modo seguente:

$$f(x) = \alpha + \beta x + \varepsilon$$

e quindi:

$$y = \alpha + \beta x + \varepsilon$$

dove α e β possono essere interpretati, ricorrendo ad una interpretazione grafica, nel modo seguente:



α è un parametro di posizione e rappresenta l'INTERCETTA DELLA RETTA

β è un parametro di pendenza (matematicamente rappresenta il coefficiente angolare). Viene definito COEFFICIENTE DI REGRESSIONE LINEARE e rappresenta la variazione della y al variare di una unità della x . β è, quindi, espressa nella stessa unità di misura di Y .

Si dimostra che i parametri della retta di regressione si calcolano nel modo seguente:

$$\begin{cases} \alpha = \mu_Y - \beta \mu_X \\ \beta = \frac{\text{Codev}(X, Y)}{\text{Dev}(X)} \end{cases}$$

12.E.1 Indice di determinazione R^2

Un indice che esprime la relazione di dipendenza della X dalla Y è dato dal, cosiddetto, indice di determinazione, il quale è dato da:

$$R^2 = \frac{\text{Codev}(X; Y)^2}{\text{Dev}(X)\text{Dev}(Y)}, \text{ cioè } R^2 = \rho_{XY}^2.$$

Tale indice, detto anche INDICE DI ADATTAMENTO AL MODELLO, esprime quanta parte della devianza totale di Y è determinata o spiegata dalla retta di regressione supposta rappresentativa del fenomeno. Questo indice varia tra 0 e 1, in particolare:

- $R^2 = 0$ quando $\beta = 0$

In questo caso si dice che vi è indipendenza in media e, cioè, tutta la variabilità è di tipo accidentale.

- $R^2 = 1$ quando tutta la variabilità è spiegata dalla retta di regressione.

I punti della Y sono, cioè, tutti appartenenti alla retta di regressione.

Esempio:

Consideriamo la seguente Serie doppia dei caratteri "numero di vani per abitazione" e "numero di componenti della famiglia" per vedere se esiste una relazione fra il numero di vani della casa acquistata e la numerosità della famiglia:

N. vani (X)	N. componenti della famiglia (Y)
3	5
2	2
3	4
5	7
4	5
2	1
1	2
2	2
3	2
1	2
26	32

Per prima cosa siamo interessati a capire se esiste una relazione di tipo lineare fra i due caratteri e perciò calcoliamo il coefficiente di correlazione di Bravais Pearson :

$$\rho_{XY} = \frac{Codev(X, Y)}{Dev(X)Dev(Y)}.$$

Per farlo, calcoliamo dapprima le medie:

$$\mu_x = 26/10 = 2,6 \text{ (vani)}$$

$$\mu_y = 32/10 = 3,2 \text{ (componenti)}$$

Quindi, per il calcolo degli altri valori, utilizzeremo la seguente tabella

N. vani (X)	N. componenti della famiglia (Y)	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x)(y_i - \mu_y)$
3	5	0,4	-18,2	0,16	331,24	-7,28
2	2	-0,6	-21,2	0,36	449,44	12,72
3	4	0,4	-19,2	0,16	368,64	-7,68
5	7	2,4	-16,2	5,76	262,44	-38,88
4	5	1,4	-18,2	1,96	331,24	-25,48
2	1	-0,6	-22,2	0,36	492,84	13,32
1	2	-1,6	-21,2	2,56	449,44	33,92
2	2	-0,6	-21,2	0,36	449,44	12,72
3	2	0,4	-21,2	0,16	449,44	-8,48
1	2	-1,6	-21,2	2,56	449,44	33,92
26	32			14,4	4033,6	18,8

Da cui:

$$\sigma_x^2 = \frac{\sum (x_i - \mu_x)^2}{N} = \frac{14,4}{10} = 1,44$$

$$Dev(X) = N \sigma_x^2 = 14,4$$

$$\sigma_Y^2 = 13,36$$

$$Dev(Y) = N \sigma_Y^2 = 33,6$$

Quindi:

$$Codev(X;Y) = \sum_{i=1}^N (y_i - \mu_y)(x_i - \mu_x) = 18,8.$$

A questo punto possiamo calcolare il coefficiente di correlazione di Bravais-Pearson:

$$\rho_{XY} = \frac{Codev(X,Y)}{Dev(X)Dev(Y)} = \frac{18,8}{\sqrt{14,4 * 33,6}} = 0,85 \approx 85\%$$

Esiste, dunque, un legame abbastanza forte fra il numero di componenti della famiglia e il numero di vani della casa abitata.

Adesso vediamo quanto vale questo legame in termini di regressione calcolando il coefficiente di regressione e la retta di regressione:

$$\beta = \frac{Cov(X;Y)}{Var(X)} = \frac{Codev(X;Y)}{Dev(X)} = \frac{18,8}{14,4} = 1,31(\text{vani})$$

$$\alpha = \mu_Y - \mu_X \beta = 3,2 - 2,6 * 1,31 = -0,106$$

E quindi l'equazione della retta è:

$$\hat{y}^* = -0,106 + 1,31x.$$

A questo punto, per completare l'esercizio calcoliamo anche quanta parte della Y è spiegata dalla retta di regressione:

$$R^2 = \rho_{XY}^2 = 0,85^2 = 0,7225 \approx 72\%$$

Risulta che il 72% del numero dei vani è spiegata dal numero dei componenti della famiglia.

ESERCIZI - 9

ESERCIZIO 25

Considerando i dati dell'esercizio 24, si verifichi se il glucosio (X) dipende dal peso (Y).

ESERCIZI RIEPILOGATIVI - B

ESERCIZIO 26

Di seguito è riportata la distribuzione doppia relativa al tipo di investimento preferito di 330 professionisti classificati in base alla loro professione. Si calcoli l'indice di Associazione Chi-quadro (assoluto e relativo) per la distribuzione data.

Professione	Fondi azionari	Fondi obbligazionari	Azioni	Titoli di stato	Totale complessivo
Medico	30	25	15	0	70
Avvocato	29	34	12	6	81
Commerciante	50	35	29	15	129
Altro	21	14	10	5	50
Totale complessivo	130	108	66	26	330

ESERCIZIO 27

Di seguito è riportata la distribuzione di 400 studenti classificati in base al tipo di maturità conseguito e al voto ottenuto. Si misuri l'entità della dipendenza media fra il tipo di maturità e il voto.

Voto	Tipo di maturita'			Totale
	Classica	Scientifica	Altro	
0 - 70	10	15	31	56
71 - 80	25	35	52	112
81 - 90	30	49	103	182
91 - 100	15	11	24	50
Totale	80	110	210	400

ESERCIZIO 28

Data la seguente serie doppia relativa all'altezza(in pollici) e alla circonferenza del collo (in cm) di 10 soggetti

Altezza in pollici	Circonferenza del collo in cm
67,75	36,2
72,25	38,5
66,25	34
72,25	37,4
71,25	34,4
74,75	39
69,75	36,4
72,5	37,8
74	38,1
73,5	42,1

1. Si verifichi se esiste correlazione fra i due caratteri
2. Si verifichi se la circonferenza del collo (Y) dipende dall'altezza (X), calcolando i parametri della retta di regressione e l'indice di determinazione.

ESERCIZI GENERALI

ESERCIZIO 29

Un ricercatore sta eseguendo uno studio sulle produzioni di un vigneto di Sangiovese. Per questo motivo, ha misurato la produzione unitaria di 500 piante, ottenendo la seguente distribuzione di frequenze assolute.

CLASSI di produzione (kg/pianta)	n_i
2,0 - 2,5	21
2,5 - 3,0	46
3,0 - 3,5	78
3,5 - 4,0	102
4,0 - 4,5	106
4,5 - 5,0	69
5,0 - 5,5	51
5,5 - 6,0	27
Totale	500

Lo studente:

1. rappresenti graficamente la distribuzione,
2. calcoli la produzione media, la produzione mediana e i quartili,
3. verifichi se la distribuzione può essere considerata simmetrica,

ESERCIZIO 30

Un campo di mais è concimato con tre dosi crescenti di azoto e pari a 0, 150 e 300 kg/ha. Le produzioni osservate sono rispettivamente pari a 5, 9 e 12 t/ha. Lo studente:

1. Stabilisca se tra dosi di concimazione e produzione esiste correlazione
2. Stabilisca se la produzione (Y) è dipendente dalla concimazione (X) definendo l'equazione di regressione ed il valore di R^2 .

ESERCIZIO 31

La popolazione delle altezze alla fioritura di 20 piante di mais di una serra è data dalla seguente serie di valori:

150	160	155	172	160	150	148	160	155	148
137	150	172	148	137	137	172	155	172	160

Lo studente:

1. realizzi la distribuzione di frequenza relativa al carattere in esame,
2. rappresenti graficamente la distribuzione,
3. rappresenti la funzione di ripartizione,
4. Qual è la moda di questa distribuzione?

ESERCIZIO 32

Una ditta si rivolge a tre fornitori per l'acquisto di alcuni prodotti. Nella tabella che segue vengono indicati, per ciascun fornitore, le percentuali di prodotti integri e di prodotti con difetti.

Fornitore	Prodotti integri	Prodotti con difetti	Totale
A	97	3	100
B	90	10	100
C	96	4	100
Totale	283	17	300

Lo studente stabilisca se il numero di prodotti difettosi è associato al fornitore.

ESERCIZIO 33

In un campione di 10 soggetti maschi è stata determinata la statura e il peso corporeo ottenendo i seguenti risultati.

Altezza in cm(X)	Peso in kg(Y)
170	70
181	75
175	69
173	68
169	58
172	70
168	67
177	63
178	66
175	70

Lo studente:

1. stabilisca se i due caratteri sono correlati,
2. In caso di risposta affermativa al punto precedente, determini se il peso può essere considerato dipendente dall'altezza, attraverso il calcolo del indice R^2 .

ESERCIZIO 34

Un'azienda rileva su un campione 15 famiglie il numero di volte che è stato acquistato il prodotto "saponetta" nell'arco di 3 anni:

18	14	21	27	3	14	17	28	33	40	36	39	20	12	7
----	----	----	----	---	----	----	----	----	----	----	----	----	----	---

Lo studente determini il livello di concentrazione della variabile oggetto di studio.

ESERCIZIO 35

Su 20 studenti iscritti ad un corso di inglese presso una scuola privata di Bari, si sono rilevate le province di residenza:

BA	BR	FG	FG	BA	BA	BA	BA	TA	BR
BR	LE	BA	BA	BA	FG	MI	BA	TA	FG

Lo studente effettui un'analisi descrittiva completa rispetto a tale variabile.

ESERCIZIO 36

Di seguito si riportano i giudizi, da A a D, dati da un gruppo di valutatori a 15 quadri esposti in una galleria da giovani artisti:

A	D	C	B	B	A	B	B	C	C	B	C	D	C	B
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Lo studente:

1. realizzi la distribuzione di frequenze relativa al carattere in esame,
2. rappresenti graficamente la distribuzione appena realizzata,
3. calcoli gli indicatori statistici che è possibile calcolare per il tipo di carattere,
