LA RILEVAZIONE DEI FENOMENI STATISTICI

La statistica è una scienza che studia metodi per la raccolta ed elaborazione dei dati al fine di sintetizzare informazioni di vario tipo.
Può essere utilizzata per lo studio di fenomeni di vario genere es. studi sull'età della popolazione, sull'espansione geografica di un virus ecc.

Chiamiamo **SERIE STORICA o TEMPORALE** la rappresentazione di un fenomeno nel tempo, un esempio il cambiamento climatico.
Chiamiamo **SERIE TERRITORIALE o SPAZIALE** la rappresentazione di un fenomeno nello spazio.

In ogni caso, perché vi sia bisogno di statistica, è necessario che il fenomeno studiato sia variabile e non statico; per quanto riguarda i fenomeni in ambito socio-economico sono sempre variabili in quanto entra in gioco il fattore umano che è molto vario.
Tra le situazioni di rilevazione in cui lo statistico si trova ad operare per l'acquisizione di dati, distinguiamo:

Rilevazione sperimentale, caratterizzata da: ipotesi di lavoro, costituite da enunciati formalizzabili spesso in termini matematici; possibilità di controllare le condizioni in cui l'esperimento si svolge e le caratteristiche delle unità statistiche da impiegare.

Dunque l'osservatore ha un continuo controllo del fenomeno con possibilità di modifica del corso naturale degli eventi (es. esperimenti fisici, chimici, biologici, psicologici).

Questo controllo, riguarda le variabili che vengono ritenute le più importanti per la determinazione del fenomeno, dette *fattori*; si possono distinguere:

- *fattori sperimentali*, ovvero le variabili su cui l'esperimento verifica il loro diverso effetto e dunque spesso costituiscono l'oggetto della ricerca (es. diverso trattamento farmacologico, diverso programma di educazione sanitaria ecc.);
- *fattori di stratificazione*, riguardano la composizione delle unità sperimentali (es. peso, età, sesso ecc.).

Il controllo effettuato su questi fattori, è di due tipologie:

- Controllo diretto → disegno sperimentale specifica la metodologia da impiegare; es. un'industria farmaceutica sperimenta un nuovo farmaco contro un tumore ma è indecisa sul dosaggio; un fattore di stratificazione influente è il peso; supponendo vi siano 36 cavie allora, un piano sperimentale può essere la divisione in 3 fasce di peso, all'interno di ciascun gruppo da 12 si suddividono le cavie in 3 gruppi a ciascuno dei quali verrà somministrata una dose diversa.

- Controllo indiretto → ci possono essere altri fattori trascurati dal piano sperimentale che influenzano i risultati dell'esperimento; per questo si procede alla randomizzazione ovvero una selezione di unità statistiche. Nell'es. precedente, per evitare che altri fattori come età, sesso ecc. influiscano sulla risposta dell'esperimento, si fa la randomizzazione delle cavie dei sottogruppi, cioè la scelta casuale di queste all'interno di ogni gruppo in modo che l'attribuzione di una cavia ad un certo sottogruppo (e quindi dosaggio) sia casuale.

Rilevazione osservazionale, in cui non si ha la possibilità di controllare le condizioni sotto le quali si svolge l'osservazione e solo in parte si possono controllare le caratteristiche delle unità statistiche. L'osservatore si pone dunque in un ruolo passivo di minima inferenza con i fenomeni osservati (es. indagini statistiche, indagini di mercato, sondaggi, rilevazioni economiche ecc.).

La più importante è l'indagine statistica il cui obbiettivo principale è la conoscenza di una popolazione intesa come insieme di unità su cui si manifesta il fenomeno oggetto di studio; per la raccolta di informazioni si può utilizzare: l'indagine totale (o censuaria) o l'indagine campionaria. In ogni caso un'indagine statistica prevede diverse fasi:

- *Definizione degli obbiettivi*: scelta delle unità da rilevare, delle variabili e del periodo di riferimento (non sempre contemporaneo)

- *Individuazione della popolazione e della lista delle unità statistiche* (elenco degli appartenenti a quella popolazione); le liste non sempre sono affidabili ad es. negli USA non esiste l'anagrafe, oppure se si vuole l'elenco dei malati di tumore si riuscirà ad ottenere la lista di quelli registrati, mentre nella realtà ci sono alcuni che non ne sono consapevoli e alcuni non registrati. In questi casi si usano tecniche che non prevedono l'utilizzo di liste, una di queste è il campionamento areale che prevede l'estrazione casuale di aree territoriali sulle quali vengono intervistate tutte le unità presenti.

- *Raccolta dei dati*: scelta della tecnica di rilevazione tra cui assume particolare rilievo **l'intervista**, che consiste nel rivolgere delle domande (raccolte in apposito questionario) alle unità che compongono la popolazione di interesse. Vi sono varie tecniche:
  - Intervista diretta → l'osservatore interagisce direttamente con l'unità osservata (es. intervista faccia a faccia), questo porta aspetti positivi quali: l'intervistatore controlla l'identità di chi risponde; e può indurre a rispondere in maniera più dettagliata fornendo alcuni dettagli. D'altra parte se l'intervistatore non è ben addestrato può portare effetti negativi sbagliando l'ordine delle domande, formulandole in maniera sbagliata, rapportandosi male con le unità ecc.
  - Auto compilazione → si utilizza nei casi in cui si ritiene che la popolazione sia disposta a collaborare con la ricerca; ha diversi aspetti positivi in quanto permette di ridurre i costi dell'indagine. Può essere inviato il questionario via posta (indagine postale) o consegnato e ritirato da personale, o tramite l'invio di una e-mail che manda al sito web (intervista CAWI)
  - Intervista indiretta o telefonica → ci si avvale di un mezzo di mediazione tra l'osservatore e l'unità osservata, il telefono.

- CATI e CAPI → sono due nuove tecniche informatiche di supporto all'intervista: la prima riduce tempi e costi dell'intervista migliorando la qualità dei dati; la seconda prevede l'inserimento dei dati direttamente su computer ma con intervista faccia a faccia. Le due tecniche possono essere usate combinate in modo da sfruttare entrambi i vantaggi
- Exit poll → è una modalità di intervista diventata popolare negli ultimi anni che permette di stimare i risultati delle elezioni prima dello spoglio delle schede elettorali tramite intervista anonima fuori da alcuni seggi.

- *Registrazione dei dati*: controllo dell'attendibilità dei dati ottenuti (es. un medico non può dichiarare di avere licenza elementare) le mancate risposte e i dati poco attendibili (es. famiglia con 25 figli) con conseguente correzione e registrazione.

- *Elaborazione e analisi dei dati.*

Oltre a poter acquisire dati tramite esperimenti o indagini, si può ricorrere a collezioni di dati predisposti da enti o società esterne, pronti per essere analizzati; importante è l'attendibilità, per questo ogni nazione possiede un istituto nazionale di statistica: in Italia è l'ISTAT.

Inoltre, è stato istituito il sistema statistico nazionale (SISTAN) che vede l'ISTAT come coordinatore degli altri enti ufficiali di rilevazione statistica; particolare rilevanza assumono i dati statistici rilevati da uffici statistici interni ad organi amministrativi (anagrafi, uffici statistici dei comuni, dei ministeri, delle ASL ecc.).

Altre fonti di dati sono le banche dati che raccolgono dati la cui organizzazione e gestione è controllata dalle società o enti tramite appositi software. Tra le organizzazioni più importanti che forniscono la consultazione online di banche dati ricordiamo: ISTAT, OCSE, EUROSTAT, Banca Mondiale e Nazioni Unite.

La statistica analizza, in termini quantitativi, i fenomeni collettivi, ovvero fenomeni il cui studio richiede l'osservazione di un insieme di manifestazioni individuali. L'indagine può essere campionaria o totale: raramente si prendono dati riguardanti l'intera popolazione, questo per la complessità e le tempistiche richieste dall'operazione, infatti il censimento della popolazione è soltanto decennale.

Ogni 2 anni però, la Banca d'Italia seleziona un gruppo di circa 8.000 famiglie e sottopone loro un questionario per poter ricavare dati ed informazioni riguardanti il reddito e la ricchezza di queste.

Questa viene definita **RILEVAZIONE CAMPIONARIA**, consiste nel scegliere un campione di unità statistiche (in questo caso 8000 famiglie italiane) al quale verranno sottoposti gli studi necessari; i risultati che si ottengono sono un'approssimazione di ciò che si sarebbe ottenuto esaminando l'intera popolazione italiana.

Nonostante ciò, in statistica sono sempre più diffuse le indagini campionarie per i vantaggi che ne derivano: diminuiscono i tempi e i costi necessari all'effettuazione dell'indagine; la riduzione dei tempi diventa molto importante qualora si studino fenomeni che, in caso di cambiamento, necessitano di interventi (es. inflazione); le informazioni possono essere più dettagliate.

Inoltre, la statistica si suddivide in: ***statistica descrittiva*** cioè osservazione di un fenomeno e descrizione delle sue caratteristiche; e ***inferenza statistica***, si occupa di misurare e controllare l'attendibilità delle informazioni provenienti da un campione.

IL QUESTIONARIO è lo strumento di rilevazione dei dati più utilizzato nelle indagini statistiche; il primo passo per la progettazione di un questionario è la concettualizzazione: si individuano le entità che entrano in gioco, si descrivono le relazioni esistenti tra esse e si individuano le possibili gerarchie tra relazioni, tutto con l'obbiettivo di ricondurre il fenomeno studiato ad un modello logico-concettuale.

Es. si vuole fare indagine sulle scelte dolo le scuole medie: l'entità principale è lo studente, la scelta però può essere condizionata da altre entità come la famiglia, la scuola, il luogo di residenza ecc. le quali hanno tra loro delle relazioni.

La rilevazione mediante questionario può generare errori non campionari ovvero che non rientrano in quelli dovuti al campionamento, ma imputabili a: il ricercatore (es. errori di formulazione domande, lunghezza questionario ecc.); il rispondente che potrebbe rispondere in maniera non fedele alla realtà per vari motivi; o l'intervistatore che può registrare le risposte sbagliate o condizionare le risposte.

Per limitare questo tipo di errori dunque bisogna:

- evitare che la formulazione delle domande si a troppo generica o troppo dettagliata per evitare fraintendimenti;
- le domande devono facilitare il ricordo di eventi passati ma senza creare stati di imbarazzo o tensione psicologica;
- le domande devono essere poste in modo da non indirizzare il rispondente verso una certa risposta.

Viene chiamata domanda filtro quella che consente di passare direttamente da una batteria di domande ad un'altra evitando di sottoporre l'intervistato a domande non pertinenti (es. la domanda "pratichi sport?" se si, porterà alla parte di questionario riguardante lo sport; se no, si passa alla parte di questionario dove vengono richiesti i motivi ecc.).

Aspetto importante da tenere presente durante la progettazione di un questionario è la sequenza delle domande, possiamo distinguere:

- Successione a imbuto: si formulano inizialmente domande molto generiche per giungere via a via a domande sempre più specifiche; in questo modo l'intervistato risponde in maniera graduale evitando che domande iniziali possano condizionare le risposte successive.
- Successione a imbuto capovolto: vengono poste inizialmente domande specifiche per poi arrivare a domande generali; in questo modo si aiuto l'intervistato a dare giudizi più ponderati su domande riguardanti temi generali.

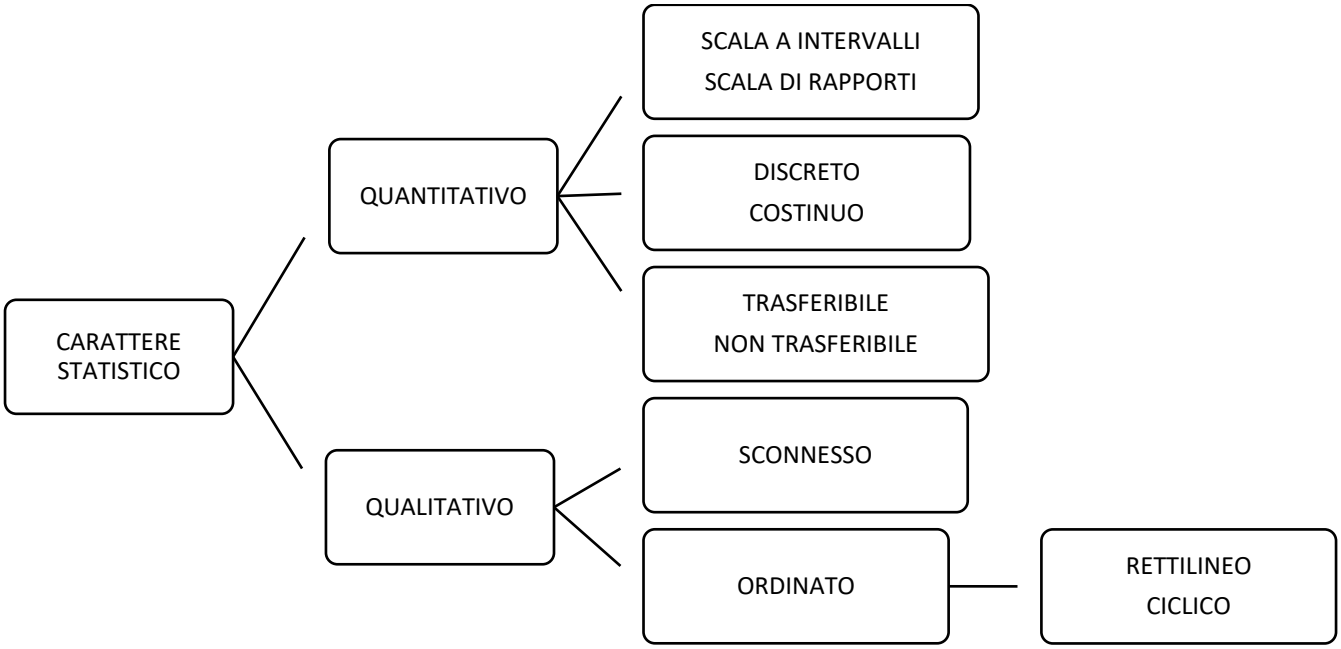
Le domande dei questionari possono essere classificate diversamente a seconda degli aspetti considerati:

- *Domanda diretta/indiretta*: diretta quando si chiama in causa l'intervistato (quel è il tuo voto di laurea?); indiretta quando ci si riferisce ad una generica terza persona o gruppo di persone (secondo te qual è lo stipendio percepito dai professori?).
- *Domande a risposta semplice/multipla*
- Domande aperte/chiuse: aperte danno la possibilità di massima personalizzazione e per questo utilizzate quando si possiedono pochi elementi conoscitivi del fenomeno indagato e si vogliono trarre dettagli dagli intervistati (sondaggi di opinione, ricerche motivazionali); chiuse permettono la riduzione di tempi ed errori di trascrizione ed agevolano l'intervistato a rispondere. Per la formulazione delle domande chiuse è molto importante la scelta delle risposte e il loro eventuale raggruppamento.

Per la raccolta di dati spesso si utilizzano database in Excel:

- Ad ogni riga corrisponde un'**UNITA' STATISTICA** = unità su cui vengono osservati i caratteri oggetto di studio (in tal caso famiglie).
- Ad ogni colonna corrisponde una caratteristica, detta **CARATTERE** o **VARIABILE** (età del capo famiglia, sesso, titolo di studio, reddito…)
- Ogni caratteristica assume una **MODALITA'** ovvero il valore (numerico o non) che può assumere la variabile; le modalità di un carattere devono essere: *esaustive*, ovvero rappresentare tutti i possibili modi di manifestarsi del carattere; e *non sovrapposte*, ovvero ad ogni unità può essere associata una sola modalità.
  Es. il reddito avrà infinite modalità perché può assumere qualsiasi valore, il sesso avrà due modalità perché può essere solo m o f ecc.
- **COLLETTIVO STATISTICO** = insieme di unità statistiche omogenee rispetto ad una o più caratteristiche; si possono distinguere:
  → Collettivi di stato, individuabili solo fissando un preciso istante di tempo (es. popolazione residente a Roma); e collettivi di movimento (es. automobili vendute in Italia in un certo periodo).
  → Collettivi empirico, se tutte le unità sono effettivamente osservabili; e collettivi teorici.
  → Collettivi finiti, se costituito da un numero finito di unità statistiche; e collettivi infiniti.

Le variabili (caratteri) possono essere di natura diversa, ed in base a ciò vengono gestite in maniera differente.



La prima distinzione da effettuare è tra:

- Variabili qualitative → rispondono alla domanda "quale?" ad esempio, titolo di studio, sesso, stato civile ecc.
- Variabili quantitative → rispondono alla domanda "quanto?" ad esempio, età, reddito, numero di percettori ecc.

In realtà vi sono alcune variabili, anche se molto raro, che si trovano nel mezzo tra le due; ad esempio i voti.

Per quanto riguarda le variabili **qualitative**, possono avere scala di misura differente, il che porta ulteriore distinzione:

- ***Scala nominale (o variabile sconnessa)*** → date due modalità è solo possibile dire se sono uguali o diverse senza classificarle in un dato ordine; ad esempio lo stato civile, il sesso, il luogo di residenza ecc.
- ***Scala ordinale (o variabile ordinata)*** → date due o più modalità, è possibile darvi un ordine, specificando che una precede l'altra e potendo dunque fare confronti di maggiore e minore; ad esempio il titolo di studio ecc.
  Questa tipologia di variabile può a sua volta essere classificata in:
  - *variabili ordinate cicliche* → non hanno vera e propria modalità di inizio e fine e per questo possono essere fissate in modo convenzionale; ad esempio il mese di nascita
  - *variabili ordinate rettilinee* → possiedono una modalità iniziale e una finale; sono la gran parte della variabili qualitative ordinate

Per quanto riguarda le variabili **quantitative** invece, possono essere distinte diversamente a seconda degli aspetti presi in considerazione:

- ***Scala a intervalli*** → non esiste uno 0 assoluto ma questo viene fissato su base convenzionale, ha quindi senso considerare la differenza tra le modalità di queste variabili ma non il rapporto tra esse. Queste variabili sono molto rare, un esempio è la temperatura (lo 0 dei gradi centigradi è diverso da 0 dei gradi celsius; non ha infatti senso dire che a 40° fa il doppio caldo di 20°)
- ***Scala di rapporti*** → esiste uno zero assoluto, naturale e non arbitrario, permettono quindi confronti di rapporto; sono la gran parte delle variabili quantitative, ad esempio età, reddito, altezza ecc.

- ***Variabili discrete*** → le modalità assumibili sono un numero finito o un infinità numerabile, ad esempio il numero di percettori
- ***Variabili continue*** → le modalità assumibili possono assumere un'infinità non numerabile di valori, ad esempio il reddito, l'altezza, l'età (ogni secondo siamo sempre più vecchi)

- ***Variabili trasferibili*** → se è possibile che un'unità statistica ceda tutto o parte del carattere posseduto ad un'altra unità statistica, ad esempio tenendo conto di un collettivo di aziende, il fatturato o i dipendenti
- ***Variabili non trasferibili*** → non è possibile che il carattere venga trasferito da un'unità ad un'altra, ad esempio l'età, il sesso ecc.

Vi sono alcune variabili quantitative che hanno moltissime modalità distinte, in questi casi può essere necessario o conveniente la **SUDDIVISIONE IN CLASSI** ovvero un'operazione consistente nel suddividere l'insieme dei possibili valori, in intervalli tra loro disgiunti.
In questo modo si ottiene un'immagine sintetica del fenomeno osservato, ma al tempo stesso si perde una parte di informazione.

Quando si suddivide un carattere in classi, è opportuno che:

- Il loro numero sia abbastanza piccolo da fornire una sintesi adeguata, ma sufficientemente grande da mantenere l'informazione con un livello accettabile di dettaglio
- Siano tra loro disgiunte
- Comprendano tutte le possibili modalità del carattere
- Abbiano uguale o diversa ampiezza (=differenza tra estremo superiore ed estremo inferiore)
- Nel definire gli estremi, è necessario tenere conto che ogni determinazione del carattere deve essere compresa in una sola classe; in base a ciò si possono distinguere:
  - intervalli chiusi a destra e aperti a sinistra (es. da 100 escluso a 200 incluso, ovvero da 99 a 200)
  - intervalli aperti a destra e chiusi a sinistra (es. da 100 incluso a 200 escluso, ovvero da 100 a 199)
- È inoltre possibile utilizzare la forma mista: fino ad un certo numero si procede per unità, da un numero in poi si divide in classi

DISTRIBUZIONI DI FREQUENZA

L'insieme dei dati è costituito da tutte le misurazioni effettuate su ogni unità statistica, quindi:

- ad ogni unità statistica corrisponderanno le modalità osservate per ciascun carattere (ogni riga);
- ad ogni carattere corrisponderà l'insieme delle modalità osservate nel collettivo → = DISTRIBUZIONE UNITARIA SEMPLICE (ogni colonna). Per effettuare una sintesi dei dati, è possibile:
    1. Contare il numero di volte in cui si verifica una certa modalità con riferimento ad un determinato carattere → **FREQUENZA ASSOLUTA**
    2. Rappresentare le frequenze assolute in una nuova tabella → **DISTRIBUZIONE DI FREQUENZA**

Questo è molto efficiente nel caso di variabili aventi un numero ristretto di modalità; per variabili del tipo reddito, età ecc. si andrebbe a sintetizzare troppo poco, per questo motivo è necessario procedere nella divisione per classi. Esempio:

Unità	Sesso	Età
1	M	2
2	F	10
3	F	5
4	M	20
5	M	13

➡

Distribuzione di frequenza dell'età (divisa in classi)

Carattere	Frequenza assoluta
0-9	2
10-19	2
Da 20 in poi	1

Distribuzione di frequenza del sesso

Carattere	Frequenza assoluta
M	3
F	2

Le frequenze assolute dipendono dalla grandezza del collettivo preso in considerazione: se il collettivo è ristretto, le frequenze assolute saranno basse. Dunque se si volesse fare un confronto tra due o più collettivi, queste non sono efficienti. Si provvederà allora a calcolare:

FREQUENZA RELATIVA = rapporto tra frequenza assoluta e numero totale di unità osservate;

FREQUENZA PERCENTUALE = frequenza relativa moltiplicata per 100 (identico valore ma espresso in %).

Queste, non dipendono dalla grandezza del collettivo e dunque permettono confronti tra due o più collettivi in maniera più efficiente.

Nel caso in cui le modalità del carattere in esame sono ordinate, ovvero il carattere è qualitativo ordinato o quantitativo, può essere interessante considerare la frequenza con cui si presentano le modalità di ordine inferiore o uguale ad una certa modalità.

Si calcola allora la **FREQUENZA CUMULATA**, ottenuta dalla somma della corrispondente frequenza e di tutte quelle relative alle classi (o unità) precedenti.

Oltre alle distribuzioni di frequenza si possono costruire le *distribuzioni di quantità* che si ottengono dalle operazioni di classificazione del collettivo rispetto ad un carattere, e di misurazione all'interno di ogni classe, di un carattere quantitativo trasferibile.

Molto spesso risulta utile la rappresentazione delle distribuzioni di frequenza in forma grafica per rendere le caratteristiche più evidenti e di facile lettura. Esistono differenti tipologie di grafico:

- **DIAGRAMMA A BARRE** → è utilizzabile per tutte le distribuzioni di frequenza ed è particolarmente adatto per rappresentare caratteri quantitativi discreti (es. numero di componenti della famiglia). Può essere verticale o orizzontale, ogni barra ha base uguale (se orizz. altezza uguale) e rappresenta la frequenza dunque l'altezza (se orizz. lunghezza) sarà proporzionale.
  È possibile creare grafici a barre multipli nel caso in cui vengano osservate due o più distribuzioni relative ad es. a due collettivi.

- **CARTOGRAMMA** → è utilizzabile solo ed esclusivamente per la rappresentazione di serie territoriali; ha come base una mappa in cui sono visibili i confini delle aree geografiche o territoriali rispetto alle quali vengono analizzate le frequenze di un carattere.
  La rappresentazione della frequenza all'interno delle aree può essere fatta in diversi modi, il più utilizzato è quello delle ripartizioni colorate in cui ogni area è colorata in base alla distribuzione di frequenza; nel caso in cui il carattere è quantitativo o qualitativo ordinato, allora si utilizzerà una scala di intensità del colore (più intenso è il colore, maggiore è la frequenza).

- **DIAGRAMMA A TORTA** → è utilizzato per le distribuzioni di frequenza relative, solitamente per caratteri con numero ristretto di modalità in quanto aumentando le modalità, aumentano i settori circolari che diventano sempre più piccoli risultando più difficili da confrontare. La base è un cerchio, ogni "fetta" rappresenta la distribuzione di frequenza in maniera proporzionale ad essa.

- **ISTOGRAMMA** → utilizzabile per tutte le distribuzioni di frequenza, è molto simile ad un diagramma a barre con la differenza che le barre NON sono distanziate (mentre nel diagramma a barre si), e le basi di ogni barra sono proporzionali all'ampiezza di classe.
  Gli istogrammi sono infatti solitamente utilizzati per caratteri continui, che dunque necessitano di una suddivisione in classi:
    - Se le classi sono di ampiezza uguale tra loro, allora le basi saranno tutte uguali, e le altezze saranno proporzionali alle frequenze; si avrà dunque un diagramma a basi regolari (uguale a diagramma a barre solo con barre attaccate)
    - Se le classi invece sono di ampiezza diversa, ogni base sarà proporzionale all'ampiezza della classe rappresentata. L'altezza di ogni barra invece, viene chiamata densità e si ottiene dal rapporto tra la frequenza e l'ampiezza di classe.

- **GRAFICI AD AREE** → utili per rappresentare l'andamento di un fenomeno nel tempo; è una spezzata che unisce i punti aventi come coordinate i valori delle frequenze corrispondenti ai valori osservati. L'area sotto la spezzata viene colorata e può essere raffigurata tridimensionalmente; se si vogliono confrontare due o più distribuzioni, relative ad uno stesso fenomeno, si possono sovrapporre.

- **GRAFICI RADAR** → sono utili per la rappresentazione di caratteri ciclici (es. nascite per mese o vendite di un prodotto per mese); si costruisce dividendo l'angolo di 360° con tanti raggi quante sono le modalità del carattere, ognuno di ampiezza uguale (es. se le modalità sono i mesi dell'anno, si avranno 12 raggi distanziati da angoli di 30° ciascuno). Infine, su ogni raggio si calcola un segmento di lunghezza proporzionale o uguale alla corrispondete frequenza; è utile a livello visivo colorare l'area del poligono che si forma.

- **DIAGRAMMA CARTESIANO** → per rappresentare un fenomeno nel tempo (serie storica) è più diffuso il diagramma cartesiano (che grafico ad aree) soprattutto se si vogliono confrontare più serie. Il grafico è costituito da una serie di punti individuati sul piano in cui: ascisse = tempo; ordinate = carattere osservato; i punti vengono uniti formando una spezzata che indica l'andamento del fenomeno.

SINTESI DELLA DISTRIBUZIONE DI UN CARATTERE: LE MEDIE

Per descrivere l'insieme delle modalità osservate di un carattere su di un collettivo, si possono impiegare le distribuzioni di frequenza e loro rappresentazioni grafiche; spesso può essere però utile utilizzare degli indici che sintetizzano ed evidenziano alcune caratteristiche essenziali della distribuzione.

Tra questi le medie, che si suddividono in:

~ *Medie analitiche*, calcolate attraverso operazioni algebriche sui valori dei caratteri e di conseguenza applicabili a caratteri di tipo quantitativo (media aritmetica, media geometrica e trimmed mean)
~ *Medie di posizione*, che non prevedono operazioni algebriche sui valori dei caratteri e quindi possono essere determinate anche su caratteri di tipo qualitativo (moda e mediana)

La scelta di una media dipende anche dal tipo di carattere da analizzare:

	Caratteri quantitativi	Caratteri qualitativi ordinati	Caratteri qualitativi sconnessi
Media aritmetica	OK	NO	NO
Media geometrica	OK	NO	NO
Trimmed mean	OK	NO	NO
Mediana	OK	OK	NO
Moda	OK	OK	OK

→ LA MODA

È una media di posizione che può essere calcolata per qualsiasi tipo di carattere, anche qualitativo sconnesso; si tratta della modalità della distribuzione che si presenta con la frequenza più alta (assoluta, relativa o percentuale).
Questo indice ci da poche informazioni riguardo a tutte le altre modalità, infatti, quando si hanno caratteri quantitativi che assumono valori diversi allora la frequenza corrispondente ai valori sarà quasi sempre unitaria; la moda corrisponderà dunque a quel valore che si è osservato più di una volta. Per ovviare a questo problema si possono dividere i caratteri in classi; in tal caso allora si tratterà di trovare la **classe modale** ovvero la classe alla quale corrisponde la frequenza più alta. Se all'interno di essa si vuole individuare un unico valore, si prende quello centrale; se le classi sono di ampiezza diversa, occorre dividerle per la loro ampiezza prima di effettuare il calcolo. Se rappresentiamo la distribuzione di frequenza graficamente, la moda corrisponde al picco della distribuzione; si parla di:

- Distribuzione unimodale, se presenta un solo picco
- Distribuzione bimodale, se presenta due picchi della medesima altezza oppure di altezze diverse (in questo secondo caso va a significare che il collettivo osservato è composto da due gruppi di unità distinti)

→ LA MEDIANA

È una media di posizione che può essere calcolata per caratteri quantitativi o qualitativi ordinabili; si tratta del valore centrale di una distribuzione, ovvero quel valore tale per cui metà dei dati è ≤ e l'altra metà è ≥. Per calcolare la mediana bisogna:

1) Ordinare le n unità in senso crescente rispetto al carattere
2) Individuare la posizione in graduatoria dell'unità centrale
3) Osservare la modalità rappresentata dall'unità centrale

Se il carattere è qualitativo, bisogna utilizzare le frequenze relative cumulate; utilizzate anche nel caso in cui il carattere sia diviso in classi per trovare la **classe mediana** con l'apposita formula. Proprietà della mediana: è robusta, ovvero non risente dei valori estremi.

→ LA MEDIA ARITMETICA

È una media analitica che può essere calcolata solo per caratteri quantitativi; è pari a somma dei valori osservati divisa per il loro numero. Nel caso di caratteri divisi in classi si può ottenere un'approssimazione della media considerando il valore centrale delle classi (=semisomma degli estremi della classe); se c'è una classe non limitata, bisogna chiuderla. In alcuni casi, nel calcolo della media

aritmetica si vuole dare diversa importanza (peso) a ciascuna osservazione del carattere; si calcolerà allora la **media aritmetica ponderata**; se i pesi sono tutti uguali, sarà pari alla media aritmetica non ponderata. Proprietà della media aritmetica:

1. La somma dei valori x assunti da un insieme di n unità statistiche, è uguale al valore medio moltiplicato per il numero di unità
2. La somma delle differenze tra i valori delle x e la loro media aritmetica, è pari a zero
3. La somma degli scarti al quadrato dei valori x da una costante c è minima quando c è uguale alla media aritmetica
4. La media aritmetica di un carattere Y, ottenuto da una trasformazione lineare $Y = aX + b$ di un carattere X è uguale a $\underline{y} = \underline{a}x + b$

➔ **LA MEDIA GEOMETRICA**

È una media analitica che può essere calcolata solo per caratteri quantitativi; è pari alla radice n-esima del prodotto dei valori. Proprietà della media geometrica:

1. Il prodotto dei valori x assunti da un insieme di unità statistiche è uguale alla potenza n-esima della media geometrica
2. Il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi

➔ **LA TRIMMED MEAN**

Il maggior difetto della media aritmetica è che risente fortemente dei valori estremi, cosicché può accadere che il suo valore non sia ben rappresentativo dell'insieme dei valori osservati. Un modo che consente di diminuire l'effetto dei valori estremi è il calcolo dei suoi valori centrali ovvero della trimmed mean che per l'appunto non considera i valori estremi.

SINTESI DELLA DISTRIBUZIONE DI UN CARATTERE: LA VARIABILITA'

La variabilità di un fenomeno è il fattore che richiede l'entrata in gioco della statistica in quanto va a significare che il fenomeno non è statico ma tende a manifestarsi con diverse modalità.

La variabilità esprime la tendenza delle unità di un collettivo ad assumere diverse modalità del carattere; è allora possibile utilizzare degli indici che sintetizzino la diversità tra ogni modalità e una media, oppure tra due particolari valori.

Questi, vengono chiamati indici di variabilità e ciascuno dovrebbe soddisfare almeno 2 requisiti:

- Deve assumere il suo valore minimo se e solo se le unità della distribuzione presentano uguale modalità del carattere
- Deve aumentare all'aumentare della diversità tra le modalità assunte dalle varie unità

Per misurare la variabilità di caratteri quantitativi, si possono utilizzare differenti indici:

- **Range o campo di variazione**

Dato un insieme di n valori osservati x1, x2,…..xn, ordinati in senso crescente, definiamo campo di variazione la differenza tra il più grande e il più piccolo di tali valori.
Il minimo del campo di variazione è 0, che si verifica solo se tutte le unità presentano lo stesso valore (=mancanza di variabilità).
Tenendo conto di solo due valori, ovviamente è un indice molto approssimativo e inoltre presenta il difetto di risentire dei valori estremi e quindi in caso di valori anomali porta informazioni molto grossolane.

- **Scarto interquartile**

Dato un insieme di n valori osservati x1, x2,……xn definiamo differenza interquartile la differenza tra il terzo e il primo quartile.

Bisogna innanzitutto arrivare alla definizione di quartile, con le relative osservazioni:
definiamo percentili quei valori che dividono la distribuzione in 100 parti di uguale numerosità; i percentili più utilizzati sono i quartili, ovvero 3 percentili che dividono la distribuzione in quattro parti uguali. Questi sono:

- il primo quartile ovvero 25-esimo percentile (Q1),
- la mediana ovvero il 50-esimo percentile (secondo quartile Q2)
- il terzo quartile ovvero il 75-esimo percentile (Q3)

Il primo e terzo quartile individuano un intervallo centrale che contiene il 50% delle unità statistiche; questo può essere considerato come misura della dispersione dei valori più frequenti del collettivo osservato. Se la distribuzione di frequenza è divisa in classi, non è possibile trovare l'esatto valore del quartile ma, come per la mediana, possiamo arrivarci tramite una sua approssimazione.

Quindi, la differenza interquartile si può dire essere il campo di variazione del 50% delle unità centrali; in questo modo vengono esclusi i valori estremi evitando dunque la considerazione di eventuali valori anomali.

- **Devianza**

È la sommatoria del quadrato degli scarti dalla media aritmetica; ha la caratteristica di crescere al crescere del numero di addendi.
Inoltre, non è mai negativa in quanto l'elevazione al quadrato trasforma tutte le differenze in quantità positive oltre a mettere in risalto le differenze più grandi (in quanto crescono più che proporzionalmente rispetto a quelle piccole).

- **Varianza**

È la media dei quadrati degli scarti dalla media aritmetica; la varianza rispetta i requisiti iniziali in quanto aumenta all'aumentare della differenza dei valori osservati ed è uguale a 0 solo quando tutte le differenze sono nulle ovvero quando tutte le modalità sono uguali al valore medio ossia tutte uguali tra loro.

- **Scarto quadratico medio (o deviazione standard)**

La varianza (e anche la devianza) ha il difetto di non possedere la stessa unità di misura dei valori della distribuzione, es. considerando la distribuzione delle altezze in cm, la varianza esprimerà un valore in cm quadrati.
Per questo motivo si utilizza la deviazione standard ovvero la radice quadrata della varianza.
Anch'essa aumenta all'aumentare della variabilità dei valori osservati ed è uguale a 0 solo in assenza di variabilità.

- **Indice di variabilità relativo (o coefficiente di variazione)**

La deviazione standard, così come devianza e varianza, sono indici di variabilità assoluti che risentono dell'unità di misura e dell'ordine di grandezza dei dati. Pertanto non consentono confronti:

· tra variabilità di fenomeni con unità di misura differenti (es. cm e m)
· in alcuni casi nemmeno se hanno stessa unità di misura (es. distribuzione peso in kg di bambini confrontato con quello in kg di adulti).

Si calcola allora l'indice d variabilità relativo, dato dal rapporto tra la deviazione standard e la media.

LA STANDARDIZZAZIONE è una trasformazione lineare dei dati, che conduce tutte le variabili ad avere valor medio nullo e varianza unitaria

Un efficace metodo di rappresentazione grafica della variabilità di un carattere, è il **box plot**, o grafico a scatola e baffi; questo si avvale di una media e di un indice di variabilità a scelta, in base alla media e all'indice scelti, si possono creare differenti tipi di box plot.
Il più utilizzato è quello che ha come media la mediana, come altezza del rettangolo la distanza interquartile e come estremi dei segmenti il valore minimo e il valore massimo della distribuzione.

Gli indici appena elencati misurano la variabilità di caratteri quantitativi; per misurare la variabilità di caratteri qualitativi invece, si utilizza il grado di **ETEROGENEITA'** o **OMOGENEITA'**.

· Si è in situazione di *massima omogeneità* (o minima eterogeneità) quando tutte le unità statistiche hanno la stessa modalità (es. se si studia lo stato civile, quando tutti sono coniugati)
· Si è in situazione di *minima omogeneità* (o massima eterogeneità) quando le osservazioni si distribuiscono in misura uguale su tutte le modalità (es. se ho un collettivo di 10 unità, e le possibili modalità sono 10, ogni unità avrà modalità differente)

Nella realtà, raramente si è in una di queste due situazioni estreme, per questo è necessario calcolare il grado di eterogeneità di una distribuzione; per farlo si utilizza ***l'indice di eterogeneità di Gini*** (padre della statistica italiana nonché fondatore dell'ISTAT).
Questo indice può assumere:

- valore minimo pari a 0
- valore massimo che dipende dal numero di modalità (K)

Per far sì che non dipenda da K, si può normalizzare l'indice, trovando dunque l'indice di Gini normalizzato (tramite la procedura della normalizzazione dell'indice, applicabile a tutti gli indici).

Altro indice utilizzato per misurare l'eterogeneità (soprattutto usato per misurare la biodiversità di un habitat), è ***l'indice di entropia*** (o di shannon); anch'esso può essere normalizzato.

Per misurare il grado di variabilità di caratteri quantitativi trasferibili, si utilizza la **CONCENTRAZIONE**; l'esempio più tipico è quello del reddito: per calcolare la distribuzione del reddito in un determinato periodo in una data zona geografica, si utilizza la concentrazione.

· Si è in situazione di massima concentrazione quando l'intero ammontare del carattere (A) è posseduto da una sola unità del collettivo
· Si è in situazione di equidistribuzione (assenza di concentrazione) quando ognuna delle n unità possiede 1/n dell'ammontare complessivo del carattere (A), ovvero ogni unità possiede quantità di carattere pari alla media aritmetica.

Nella realtà queste due situazioni sono rare, e allora bisogna calcolare il grado di concentrazione con l'apposito indice: ***il rapporto di concentrazione di Gini***. Questo ha sempre valore compreso tra 0 e 1:

- Sarà uguale a 0 solo in caso di equidistribuzione
- Sarà uguale a 1 solo in caso di massima concentrazione

Per la rappresentazione grafica della concentrazione, è molto utilizzato il **diagramma di Lorenz**; si tratta di un piano cartesiano con asse delle ascisse F e asse delle ordinate Q.

- In corrispondenza di ogni coppia (F,Q) si avrà un punto sul piano, unendo questi punti si ottiene la *spezzata di concentrazione* (o curva di Lorenz).
- In qualsiasi situazione, qualunque sia il punto corrispondente alle coordinate, si troverà sempre all'interno di un quadrato di lato 1; questo perché abbiamo visto che il rapporto di concentrazione di Gini è sempre compreso tra 0 e 1.
- Tracciando una retta che va dal punto (0,0) al punto (1,1) si ottiene la *linea di equidistribuzione* (infatti in caso di equidistribuzione abbiamo visto che F=Q);
- L'area che sta tra la linea di equidistribuzione e la spezzata di concentrazione è detta area di concentrazione.
- Maggiore è l'area di concentrazione, maggiore sarà la distanza tra le due linee e viceversa; dunque, maggiore sarà il grado di concentrazione.
- L'indice del rapporto di concentrazione di Gini è pari (o quasi) all'area di concentrazione
- Nel caso specifico di massima concentrazione, tutto il carattere è posseduto da una unità mentre le restanti n-1 unità non detengono nulla, dunque, a livello grafico, la spezzata di concentrazione coinciderà con l'asse delle ascisse fino all'n-1 unità, per poi raggiungere il punto (1,1).

SERIE STORICHE e NUMERI INDICI:

Si definisce **SERIE STORICA** una sequenza di osservazioni di un fenomeno Y osservato in n tempi; ne sono degli esempi: la rilevazione annuale degli abitanti di una regione, la rilevazione trimestrale delle retribuzioni, la rilevazione mensile del numero di occupati ecc.

Per analizzare opportunamente l'evoluzione nel tempo di un fenomeno rilevato in serie storica, e dunque si vogliono misurare i mutamenti verificati, si ricorre generalmente alla costruzione dei **NUMERI INDICI** ovvero il rapporto tra due misurazioni opportunamente scelte.

I numeri indici possono essere:

- A **base fissa** sono costruiti in rapporto a un tempo fissato che chiameremo *base*, vengono usati quando è importante rapportare tutte le quantità misurate a quella di un periodo di riferimento (la base).
  Una serie di numeri indici a base fissa esprime l'intensità o frequenza di un fenomeno in ogni periodo di tempo, come quota dell'intensità o frequenza della base. Molto utile per lo studio di fenomeni economici è l'andamento dei prezzi in quanto può portare a variazioni di molti fattori, ad esempio la spesa per il consumo di beni alimentari può variare al variare dei prezzi anche tenendo costanti le quantità acquistate.
  Il periodo di tempo preso come base deve rappresentare una situazione di normalità caratterizzata dall'assenza di eventi esterni che possano aver influito in modo rilevante sull'andamento del fenomeno; è inoltre conveniente scegliere come base uno dei periodi centrali della serie. Passato un certo periodo di tempo è inoltre necessario aggiornare la base; il cambiamento di base è anche necessario qualora si vogliano confrontare due serie con basi diverse.
  Aspetti pratici:
  - Per ogni generico istante t, è possibile calcolare il rapporto tra il prezzo rilevato in quel preciso istante e quello rilevato in un istante preso come riferimento della serie
  - L'indice calcolato al tempo della base è sempre uguale a 1
  - Per ragioni di praticità si usa moltiplicare l'indice per 100, ottenendo dunque gli indici percentuali

- A **base mobile** sono costruiti tenendo conto del rapporto tra misurazioni successive, vengono usati quando è importante valutare l'andamento della quantità misurata nel tempo. Una serie di numeri indici a base mobile esprime l'intensità o frequenza di un fenomeno in ogni periodo di tempo, come rapporto con l'intensità o frequenza del periodo precedente.
  Aspetti pratici:
  - Per ogni generico istante t (a partire dal secondo periodo della serie), è possibile calcolare il rapporto tra il prezzo rilevato in quel preciso istante e il prezzo rilevato nel periodo precedente
  - Se l'indice risulta pari a 1 significa che i prezzi sono invariati rispetto al periodo precedente
  - Per ragioni di praticità si usa moltiplicare l'indice per 100, ottenendo dunque gli indici percentuali

Proprietà dei numeri indici:

a) Proprietà di identità → se si confronta una misurazione al tempo k con se stessa, si ottiene 1
b) Proprietà di reversibilità delle basi → il numero indice $s|t$ è il reciproco (inverso) del numero indice $t|s$
c) Proprietà di circolarità

Variazione di una serie storica

- Variazione assoluta tra due periodi consecutivi, è data dalla differenza dei valori osservati nei due periodi

- Variazione relativa (o tasso di variazione) tra due periodi consecutivi, è data dal rapporto tra variazione assoluta e il valore del periodo precedente (es. var. assoluta calcolata facendo differenza tra valore nel 2014 e 2015; variazione relativa sarà data dal rapporto tra variazione assoluta e il valore del 2014 in quanto precedente a 2015).

NUMERI INDICI COMPLESSI

I numeri indici che abbiamo calcolato si riferiscono all'andamento di una sola serie di misurazioni; in molti casi, tuttavia, è necessario prendere in esame più di una variabile. Questo succede tipicamente nel calcolo degli indici di inflazione periodicamente calcolati dall'ISTAT.

In tal caso infatti viene rilevato l'andamento dei prezzi di una pluralità di beni e tali andamenti vengono poi sintetizzati in un unico indice, si parla in tal caso di numero indice complesso.

L'**inflazione** è un processo di aumento continuo e generalizzato del livello dei prezzi dei beni e servizi destinati al consumo delle famiglie. Un aumento dell'inflazione corrisponde ad una situazione in cui aumenta la velocità di crescita dei prezzi, mentre una riduzione dell'inflazione si verifica nel caso in cui i prezzi, pur essendo in aumento, crescono a una velocità minore.

L'inflazione si misura attraverso la costruzione di un **indice dei prezzi al consumo**, strumento statistico che misura le variazioni nel tempo dei prezzi di un insieme di beni e servizi, chiamato **paniere**, rappresentativo degli effettivi consumi delle famiglie in uno specifico anno.

L'Istat produce tre diversi indici dei prezzi al consumo ciascuno con proprie finalità:

- il **NIC** (per ***l'intera collettività nazionale***) misura l'inflazione a livello dell'intero sistema economico; in altre parole considera l'Italia come se fosse un'unica grande famiglia di consumatori, all'interno della quale le abitudini di spesa sono ovviamente molto differenziate. Per gli organi di governo il NIC rappresenta il parametro di riferimento per la realizzazione delle politiche economiche;
- il **FOI** (per le ***famiglie di operai e impiegati***) si riferisce ai consumi dell'insieme delle famiglie che fanno capo a un lavoratore dipendente (extragricolo). È l'indice usato per adeguare periodicamente i valori monetari, ad esempio gli affitti o gli assegni dovuti al coniuge separato;
- l'**IPCA** (***indice armonizzato europeo***) è stato sviluppato per assicurare una misura dell'inflazione comparabile a livello europeo. Infatti viene assunto come indicatore per verificare la convergenza delle economie dei paesi membri dell'Unione Europea, ai fini dell'accesso e della permanenza nell'Unione monetaria.

I tre indici si basano su un'unica rilevazione svolta periodicamente dagli Uffici Comunali, hanno stessa rappresentatività territoriale (tutti i capoluoghi) e sono prodotti con la stessa metodologia di calcolo, condivisa a livello internazionale, adottando un paniere di 930 prodotti.

NIC e FOI si basano sullo stesso paniere, ma il peso attribuito a ogni bene o servizio è diverso, a seconda dell'importanza che questi rivestono nei consumi della popolazione di riferimento. Per il NIC la popolazione di riferimento è la popolazione presente sul territorio nazionale; per il FOI è l'insieme delle famiglie residenti che fanno capo a un operaio o un impiegato.

L'IPCA ha in comune con il NIC la popolazione di riferimento, ma si differenzia dagli altri due indici perché il paniere esclude, sulla base di un accordo comunitario, le lotterie, il lotto e i concorsi pronostici.

Un'ulteriore differenziazione fra i tre indici riguarda il concetto di prezzo considerato: il NIC e il FOI considerano sempre il prezzo pieno di vendita. L'IPCA si riferisce invece al prezzo effettivamente pagato dal consumatore. Ad esempio, nel caso dei medicinali, mentre per gli indici nazionali viene considerato il prezzo pieno del prodotto, per quello armonizzato europeo il prezzo di riferimento è rappresentato dalla quota effettivamente a carico del consumatore (il ticket). Inoltre, l'IPCA tiene conto anche delle riduzioni temporanee di prezzo (saldi e promozioni).

Attraverso i numeri indici dei prezzi è possibile seguire l'evoluzione del valore di un prodotto riferendosi solo alle quantità fisiche e non ai cambiamenti del prezzo. Tali serie, depurate dall'effetto dell'inflazione, vengono dette ***serie a prezzi costanti***. La trasformazione di una serie nella corrispondente serie a prezzi costanti si chiama **deflazione**.

Data una serie, la corrispondente serie a prezzi costanti si calcola dividendo ciascun valore osservato per l'indice dei prezzi (ad esempio il FOI calcolato dall'ISTAT).
Analogamente alla deflazione, è possibile utilizzare gli indici dei prezzi per rivalutare salari, canoni di affitto ed altre grandezze espresse in termini monetari. La **rivalutazione** si effettua moltiplicando il valore da aggiornare per il coefficiente che misura la variazione dei prezzi a partire dal momento in cui il valore è stato fissato (ossia il numero indice dei prezzi avente come base detto momento).

I RAPPORTI STATISTICI:

In un rapporto statistico si mettono a confronto due termini, o frequenze, o quantità, di cui almeno uno di natura statistica (riferito ad un fenomeno collettivo) e tale che tra i due termini sussista qualche legame logico. Questi rapporti permettono di confrontare l'intensità di un fenomeno su collettivi, tempi o luoghi, diversi e per questo sono largamente impiegati nella descrizione di fenomeni socio-economici.