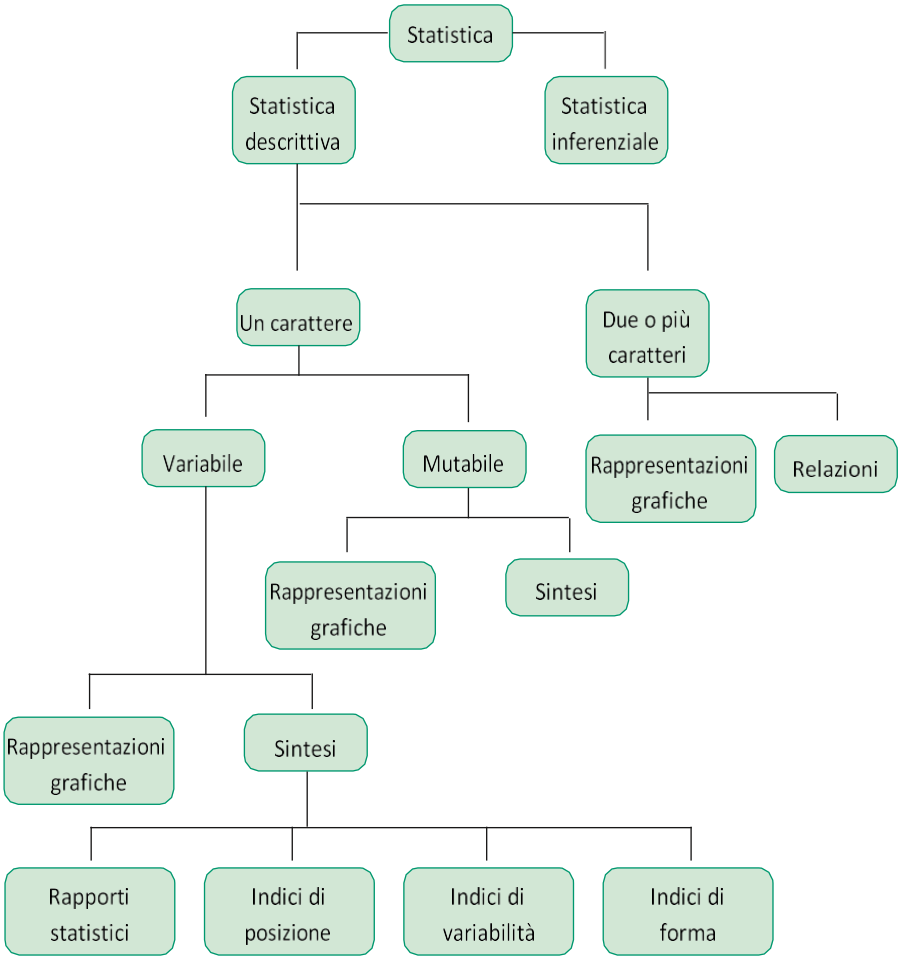
1. L'ANALISI STATISTICA



1) STATISTICA DESCRITTIVA E STATISTICA INFERENZIALE

Nell'ambito della metodologia statistica si distinguono, a fini puramente didattici, due filoni fondamentali: la *statistica descrittiva* e la *statistica inferenziale*.
La **statistica descrittiva** è volta alla *rappresentazione*, attraverso mezzi matematici, di uno o più fenomeni reali conducendo lo studio sull'intera popolazione in cui si palesa il fenomeno o i fenomeni oggetto di studio.
La **statistica inferenziale** è volta all'*induzione probabilistica* circa la struttura incognita di una popolazione. Questo filone della statistica si occupa di risolvere il cosiddetto *problema inverso*, ossia, sulla base di osservazioni su un campione di unità selezionate con date procedure dalla popolazione, perviene a soluzioni valide, entro dati livelli di probabilità, anche per la popolazione stessa.

STATISTICA APPLICATA

Il campo di applicazione della statistica si è notevolmente esteso negli ultimi anni, comprendendo tutte le situazioni in cui sono implicati fenomeni collettivi. A seconda della materia cui la scienza statistica si **applica**, si possono distinguere varie specializzazioni della stessa: statistica economica, statistica demografica, statistica giudiziaria etc.

2) FASI DELL'ANALISI STATISTICA

Si conviene, generalmente, di dividere un'analisi statistica in cinque fasi:

1. LA DEFINIZIONE DEGLI OBIETTIVI

Si tratta di una fase alquanto delicata in cui gli obiettivi prefissati devono essere esattamente individuati delimitando la ricerca in termini spaziali e temporali.

2. LA RILEVAZIONE

È la fase dell'analisi statistica concernente l'osservazione dei caratteri relativi alle unità statistiche mediante opportune tecniche e strumenti.

Essa può essere:

— **completa** (censimento) se è condotta su tutte le unità costituenti la popolazione cui si riferisce il fenomeno in esame;
— **parziale** se è condotta su un campione estratto dalla popolazione, il cui impiego si basa sull'approccio induttivo (dalla parte al tutto) tipico dell'inferenza statistica.

Data la multiforme varietà in cui si manifestano i fenomeni collettivi oggetto dell'analisi statistica, le rilevazioni non sono quasi mai effettuate sull'intera popolazione, ma su un campione rappresentativo della stessa.
La rilevazione dei dati può essere effettuata da enti privati (aziende, società commerciali, studi professionali etc.). In Italia l'organo statistico ufficiale dello Stato è l'**ISTAT** (Istituto Nazionale di Statistica), persona giuridica di diritto pubblico con ordinamento autonomo, sottoposta alla vigilanza della Presidenza del Consiglio dei Ministri e al controllo della Corte dei Conti.
La rilevazione si esegue sulla base di modelli che consistono in formulari completi di domande e risposte, predisposti in modo da ottenere esattamente quei dati che interessano ai fini dell'analisi.

3. LA ELABORAZIONE DEI DATI

In questa fase i dati rilevati sono sintetizzati allo scopo di ottenere dati più significativi.

4. LA PRESENTAZIONE ED INTERPRETAZIONE DEI DATI

Questa fase dell'analisi statistica è particolarmente delicata in quanto consiste nella rappresentazione dei dati attraverso tabelle, grafici e indici, e nella spiegazione dei risultati ottenuti dall'intera analisi statistica.

5. L'APPLICAZIONE DEGLI ESITI DELL'ANALISI

La statistica trova applicazione in diversi campi; è compito dello statistico definire i criteri e i limiti all'impiego degli esiti di un'analisi.

3) LE UNITÀ STATISTICHE

La statistica acquisisce le informazioni su una data popolazione, non necessariamente riferita ad esseri umani. Le componenti elementari della popolazione su cui materialmente è effettuata un'indagine sono denominate **unità statistiche** e si distinguono in:

— **unità semplici** come una singola persona, una singola abitazione etc;
— **unità composte** se sono insiemi di unità semplici simili considerate anche a prescindere dall'unità composta; in questa tipologia rientra una famiglia intesa come insieme dei suoi componenti, un edificio inteso come insieme di abitazioni etc;
— **unità complesse** se sono insiemi di unità semplici diverse considerate, però, nella loro globalità; in questa tipologia rientra il rapporto coniugale di cui sono unità semplici il marito e la moglie, oppure un determinato rapporto di lavoro di cui sono unità semplici il datore di lavoro e i dipendenti etc.

4) IL CARATTERE STATISTICO

Il fenomeno oggetto dell'analisi statistica è il **carattere** (o **caratteristica**), e rappresenta l'elemento che consente di descrivere una popolazione (o un campione). I valori che può assumere un carattere su un'unità statistica sono denominati **modalità**.
Un carattere può essere *qualitativo* o *quantitativo*.

CARATTERE QUALITATIVO

Un **carattere qualitativo** o **mutabile** si manifesta nell'unità statistica mediante modalità, dette **attributi**, e può essere indicato solo con espressioni verbali (aggettivi, sostantivi etc.). Ad esempio, il sesso di una persona si presenta nell'attributo: maschio o femmina; il tipo di lavoro svolto da un certo numero di persone può essere indicato solo con la qualifica verbale di esso: operaio, impiegato etc. A volte è possibile indicare caratteri qualitativi con simboli numerici, ad esempio, gli impiegati dello Stato sono classificati secondo categorie A, B etc. previste dalla legge sul pubblico impiego. Questi simboli numerici, tuttavia, non sono altro che delle qualifiche e restano caratteri qualitativi.

CARATTERE QUANTITATIVO

Un **carattere quantitativo** o **variabile** è indicato mediante espressioni numeriche, in altre parole, per esso è realizzabile una misurazione espressa in cifre, come il reddito delle persone, il loro peso, la loro età etc.

A sua volta, una variabile può essere:

— **continua** quando può assumere come modalità un numero reale qualsiasi, come la temperatura di una stanza, la statura, l'età, il peso di un individuo etc.;
— **discreta** quando può assumere come modalità solo numeri interi, come il numero dei componenti di una famiglia, il numero di addetti di un'azienda etc.

5) FREQUENZE E INTENSITÀ

Il numero di volte in cui una data modalità del carattere si presenta nel collettivo è denominato **frequenza assoluta**. Essa è il risultato di una *enumerazione*. Per esempio, poiché al censimento del 2001 in Italia erano 4.706.206 le famiglie con 3 componenti, allora 4.706.206 è la frequenza assoluta del carattere «*famiglie per numero di componenti*» relativo alla modalità «3».
Il rapporto tra la frequenza assoluta e il numero totale di unità statistiche del collettivo esprime la **frequenza relativa**.
La **frequenza cumulata** rappresenta l'ammontare del carattere posseduto dalle prime *i* modalità, ordinate in senso non decrescente. Essa è anche detta *frequenza cumulata assoluta* per distinguerla dalla **frequenza cumulata relativa** e che rappresenta la frazione del carattere posseduta complessivamente dalle prime *i* modalità, ordinate in senso non decrescente.
L'**intensità** è, invece, l'ammontare o la misura di un carattere additivo posseduto dalle unità statistiche. Sono esempi di intensità il reddito di un individuo, il peso di una persona etc.

6) LA CLASSE DI MODALITÀ

La **classe**, o **classe di modalità**, è ciascuno degli intervalli di prefissata ampiezza in cui risulta suddiviso l'insieme delle modalità di un carattere quantitativo *X*. Ad ogni classe si fa corrispondere una frequenza assoluta o relativa, che indica il numero di unità della popolazione che possiedono un valore del carattere compreso tra i suoi limiti, x_i e

x_{i+1}.

TIPOLOGIE

Una classe di modalità può essere:
— aperta sia a sinistra che a destra, in tal caso i limiti inferiore e superiore sono esclusi dalla classe; essa è del tipo:

$$x_i - x_{i+1}$$

— aperta a sinistra e chiusa a destra, in tal caso il limite inferiore è escluso dalla classe; essa è del tipo:

$$x_i \dashv x_{i+1}$$

— chiusa a sinistra e aperta a destra, in tal caso il limite superiore è escluso dalla classe; essa è del tipo:

$$x_i \vdash x_{i+1}$$

— chiusa sia a sinistra che a destra, in tal caso entrambi i limiti sono inclusi nella classe; essa è del tipo:

$$x_i \dashv\vdash x_{i+1}$$

Talvolta, quando l'insieme dei valori del carattere non è strettamente specificabile, si tende a non precisare il limite inferiore e il limite superiore, rispettivamente, della prima e dell'ultima classe, in altre parole, la prima classe è aperta a sinistra e l'ultima è aperta a destra.

VALORE CENTRALE

Il **valore centrale** di una classe è dato dalla semisomma dei limiti superiore e inferiore della classe. In simboli:

$$\overline{x}_i = \frac{x_i + x_{i+1}}{2}$$

I valori centrali $\overline{x}_i$ sono utilizzati ai fini della determinazione di medie e di indici di variabilità in luogo delle modalità x_i quando i dati sono raggruppati in classi.

CONFINI

I **confini** di una classe sono gli estremi della classe:

— l'*estremo superiore* di una classe si ottiene dalla semisomma del limite superiore della classe data e del limite inferiore della classe successiva;
— l'*estremo inferiore* di una classe si ottiene dalla semisomma del limite inferiore della classe data e del limite superiore della classe precedente.

AMPIEZZA

L'**ampiezza** (α_i) di una classe è la differenza tra il confine superiore e il confine inferiore della classe stessa. Essa è detta anche **modulo** e può essere uguale o diversa per tutte le classi.

7) SCALE DI MISURAZIONE DEI CARATTERI

Dopo aver scelto il carattere da rilevare, lo statistico deve occuparsi della misurazione delle modalità con cui esso si presenta nelle varie unità. A tal fine, è stata introdotta la seguente distinzione dei caratteri in funzione della scala di misurazione:

CARATTERI CON SCALA NOMINALE

Sono quelli per cui non si riscontra alcun ordine di successione tra le modalità (attributi); date due osservazioni su due unità statistiche, si può stabilire se esse sono **uguali** o **diverse**.
Rientrano in questa tipologia la professione, la nazionalità, il sesso, la religione, il partito politico etc.

CARATTERI CON SCALA ORDINALE

Sono quelli per cui si riscontra un ordine logico di successione delle modalità; date due osservazioni su due unità statistiche, si può stabilire una **relazione d'ordine**, ossia se esse sono uguali o l'una maggiore o minore dell'altra.
Rientrano in questa tipologia i giudizi scolastici, i gradi militari etc.

CARATTERI CON SCALA AD INTERVALLO

Sono caratteri quantitativi per cui è possibile operare un confronto, per **differenza**, tra le modalità e per cui si assume una origine arbitraria e un'unità di misura.
Rientrano in questa tipologia la misurazione degli anni, in cui si è convenuto di fissare l'anno zero come l'anno di nascita di Cristo, oppure la misurazione della temperatura in gradi Celsius la cui origine arbitraria, 0°, coincide con il punto di congelamento dell'acqua.

CARATTERI CON SCALA PROPORZIONALE

Sono caratteri quantitativi per i quali si stabilisce oggettivamente un'origine, il cosiddetto **zero assoluto**. Per tali caratteri si effettua una vera e propria operazione di misurazione con strumenti appropriati, o una operazione di computo, ed è possibile effettuare **rapporti tra le misure** nel senso che si stabilisce se l'una è la metà, il doppio … dell'altra. Rientrano in tale tipologia il numero dei componenti di una famiglia, il numero di appartamenti di un edificio, l'età, il peso di un individuo etc.

8) RAPPRESENTAZIONE DELLE RILEVAZIONI STATISTICHE

Dopo aver effettuato la rilevazione, lo statistico deve realizzare una sintesi efficace dei dati, attraverso una classificazione delle modalità del carattere (o dei caratteri) investigato, non potendole presentare sempre in forma enumerativa. A seconda dell'esigenza i dati sono rappresentati in forma tabellare e in forma grafica; ciascun modo presenta dei vantaggi e degli svantaggi e l'uso dell'uno non esclude l'altro. La **distribuzione statistica** è la più importante rappresentazione statistica, essa consiste nella organizzazione dei dati in forma tabellare che ad ogni modalità di un dato carattere fa corrispondere la rispettiva frequenza ed è denominata **distribuzione di frequenza**, oppure che ad un insieme di eventi fa corrispondere un insieme di numeri reali ed è denominata **distribuzione di probabilità**.

La tabella seguente indica la distribuzione di frequenza (semplice) di un carattere *X* discreto:

Tabella 1

MODALITÀ DI X	**FREQUENZE**
x_1	n_1
x_2	n_2
:	:
x_i	n_i
:	:
x_r	n_r
Totale	n

La tabella seguente indica, invece, la distribuzione (semplice) di frequenza di un carattere continuo (con modalità raggruppate in classi):

Tabella 2

CLASSI DI MODALITÀ DI X	**FREQUENZE**
$x_1 \mid\!- x_2$	n_1
$x_2 \mid\!- x_3$	n_2
:	:
$x_i \mid\!- x_{i+1}$	n_i
:	:
$x_r \mid\!- x_{r+1}$	n_r
Totale	n

Nel caso si consideri un carattere qualitativo, nella tabella, invece delle modalità, figurano gli attributi.

9) MISURE SINTETICHE DI DISTRIBUZIONI STATISTICHE

L'analisi statistica fornisce **misure sintetiche** per valutare aspetti complessi e globali di una distribuzione di un fenomeno *X* mediante un solo numero reale costruito in modo da disperdere al minimo le informazioni sui dati originari. In rapporto alle caratteristiche che si misurano si parla di **rapporti statistici** (di cui ci occuperemo nel Capitolo Terzo), **indici di posizione**, **indici di variabilità**, **indici di forma** (di cui ci occuperemo nel Capitolo Quarto).

In rapporto alla natura, gli indici si distinguono in:

— **indici assoluti** che sono introdotti per valutare in modo sintetico un aspetto di una distribuzione e sono espressi nella stessa unità di
misura del fenomeno o in sua funzione;
— **indici relativi** che non dipendono dall'unità di misura del fenomeno, e si ottengono rapportando due misure assolute oppure un indice assoluto al suo massimo.
Infine, gli **indici normalizzati** sono indici relativi che assumono valori in un intervallo finito quasi sempre [0, 1] oppure [–1, +1].

10) LE SERIE

La **serie** è la successione di dati ordinati secondo modalità di caratteri qualitativi.

Si distinguono diverse tipologie di serie:

SERIE STORICHE

Sono quelle serie in cui il fenomeno è studiato in funzione del tempo. Sono anche dette *serie temporali*.

Le serie storiche possono essere:

— *statiche* se presuppongono la costanza (o quasi) dei loro termini nel tempo;
— *dinamiche* se descrivono un fenomeno in continuo mutamento. Esse si dicono *evolutive,* quando il fenomeno si modifica in senso costante, *periodiche,* quando i termini del fenomeno oscillano in modo irregolare.

SERIE TERRITORIALI

Sono quelle serie in cui il fenomeno collettivo è studiato in relazione al territorio.

SERIE RETTILINEE

Sono così chiamate quelle serie di termini in cui è possibile riscontrare un *ordine logico naturale* di *successione* dei termini. Esiste cioè un termine che rappresenta una modalità *iniziale* del fenomeno ed un altro che rappresenta una modalità *finale* di esso.

SERIE CICLICHE

Sono così chiamate quelle serie di termini in cui non è dato stabilire quale sia il termine *iniziale* assoluto e quello *finale* assoluto della serie, pur senza un inizio ed una fine certi dell'andamento del fenomeno.

SERIE SCONNESSE

Sono così chiamate quelle serie in cui non è possibile riscontrare alcun ordine di successione tra le modalità. Sono serie sconnesse quelle che rappresentano la *professione,* la *nazionalità,* la *religione* e il *partito politico* degli intervistati etc.

11) GLI ERRORI

Nell'analisi statistica il concetto di **errore** è più ampio di quello comune. Infatti, considerato che per *errore* si intende la differenza tra il valore stimato e il valore vero o teorico, una prima distinzione si opera tra:

— **errori campionari** costituiti dalla differenza tra il valore stimato su un campione e il valore calcolato sulle unità statistiche della popolazione;
— **errori extracampionari** non dovuti, appunto, al campionamento e che si possono commettere nel corso di un'indagine statistica.

Tra gli errori non campionari si distinguono:

— la *mancata rilevazione dei dati* quando da una unità statistica non si è ottenuta l'informazione cercata e può consistere nella mancata risposta al questionario oppure nella mancata risposta alla singola domanda;
— l'*errore di rilevazione* che è una difformità tra la modalità rilevata e la realtà e può aversi nel caso di inadatta formulazione o cattiva comprensione della domanda, per il contesto in cui si svolge l'intervista etc.

Una ulteriore distinzione si opera tra:

— **errore sistematico** provocato dall'utilizzo di strumenti difettosi o modalità erronee di rilevazione e può essere ridotto o eliminato;
— **errore casuale** provocato da fattori esterni è controllabile con metodi statistici, ma non eliminabile.

Accade spesso che, dopo la raccolta dei dati, si presenti la necessità di correggere le cifre ottenute.