# Statistica

## Statistica

La parola **statistica** ha due significati.

**Numeri** scelti per fornire indicazioni

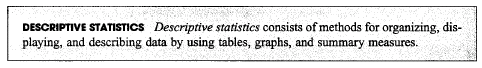
1. Età media degli studenti iscritti al I, II, … anno del corso di laurea di …
2. Numero massimo e minimo di figli per famiglia censiti nell’anno 2001
3. Percentuale di tiri liberi realizzati da Michel Johnson nel campionato NBA …
4. Valutazione media assegnata al test di ingresso del corso di …
5. Sconto medio praticato dai negozi di abbigliamento …
6. Tempo massimo impiegato sul percorso casa …

**Tecniche** e **Metodi** usati per collezionare, analizzare, interpretare e presentare i dati al fine di prendere decisioni e decidere strategie.

* La statistica permette di ricavare indicazioni quantitative di *fenomeni soggetti a fluttuazioni*.
* Gli indicatori forniti dalla statistica permettono di fare *strategie* e prendere decisioni *guidate*.

## Tipi di Statistica

**Statistica descrittiva**.

Siamo interessati ad analizzare gli Esiti dell’esame di statistica. **Data set**:studenti e punteggio. Gli studenti sono gli **Elementi** del data set, il punteggio ottenuto da ciascun studente è una **osservazione** [vedi **Dati Esame.xls**]. Le informazioni presenti nel data set vengono riassunte in **tabelle** e **grafici** o tramite **indicatori numerici** riassuntivi (media, media, varianza …).

**Statistica inferenziale**

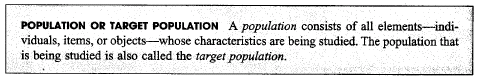
* Tutte le famiglie di una città con figli in età scolaresono una **Popolazione** perchérappresentano *tutti gli elementi* utili per una analisi.
* Le famiglie con figli in età scolare di un quartiere sono un **Campione** che contiene *alcuni elementi* della popolazione.

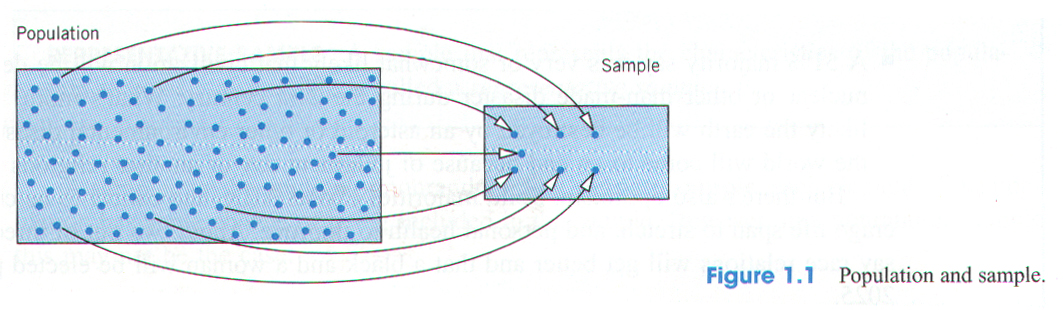
Il compito specifico della Statistica inferenziale è fare inferenze, ricavare predizioni e previsioni su una popolazione a partire dall’analisi di un campione (Decision Making).

[Vedi il Capitolo 1 del Testo Consigliato per un Esempio di Statistica Inferenziale]

## Popolazione e Campione

1. Percentuale degli elettori che voteranno per la lista ….
2. Costo delle case al metro quadro nelle città della regione ….
3. Numero di automobilisti che provocheranno un incidente di gravità pari a ….

Tutti gli elettori, tutte le case e tutti gli automobilisti sono le **Popolazioni Obiettivo (Target)** della nostra indagine.

Le analisi vengono eseguite e le decisioni sono prese sulla base di una numero limitato di oggetti della popolazione che prende il nome di **Campione**.

Affinché dalla **indagini** condotte su un campione sia possibile trarre informazioni utili è importante che il campione sia rappresentativo.Daun **campione rappresentativo** si ricavano risultati molto vicini a quelli che si otterrebbero eseguendo un **censimento**[[1]](#footnote-1)[Exit pool elettorali].

Nella scelta di un campione è importante tener conto dei **fattori** che influenzano la scelta degli elementi che lo compongono.

Esistono **regole precise** che guidano uno statistico nella scelta del campione e nell’esecuzione di una indagine

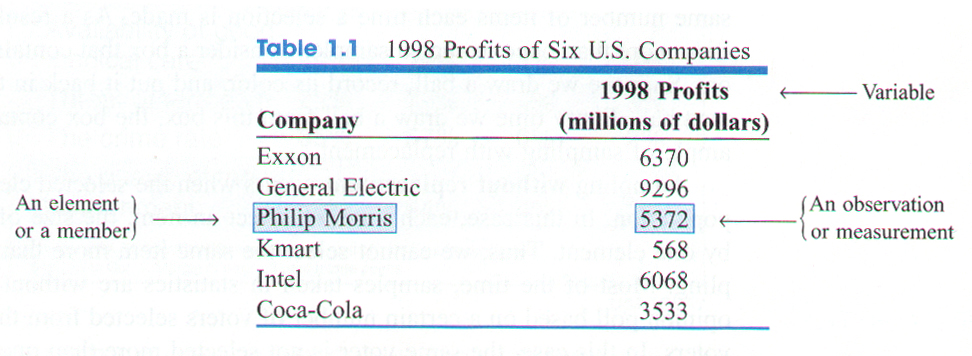
* campione *casuale* o *non* *casuale*
  + interrogazione di uno studente a caso
  + interrogazione degli studenti seduti nei primi banchi
* campionamento con *rimpiazzamento* o senza *rimpiazzamento*
  + lancio di un dado *onesto* – il medesimo numero può uscire più volte
  + estrazione dei numeri al lotto – il numero è estratto una sola volta

**Esempio** **di Popolazione/Campione**: Età dei componenti di una famiglia. Stato civile di 50 abitanti del quartiere.

## Termini della Statistica

**Elementi** o **membri** di un campione, **variabile analizzata**, **osservazione** o **misura** della variabile.

**Esempio.** Profitti dichiarati da sei grandi compagnie con sede negli U.S.A.

Gli **elementi** o membri di un campione (popolazione) sono i soggetti dell’indagine, le compagnie. La **variabile** è il parametro che stiamo analizzando e che può assumere valori diversi, i profitti. La **osservazione** o **misura** è il valore che assume la variabile per un dato elemento del campione (il profitto di una delle sei compagnie analizzate), [vedi Dati Esame.xls].

Le informazioni raccolte nella Tabella 1.1 sono i dati risultato delle nostre analisi o **data set** che riassume il risultato dell’indagine (esperimento) eseguita.

## Tipi di Variabile[[2]](#endnote-1)

* L’altezza delle persone; il reddito di una famiglia; il voto di esame; il costo dei libri di testo; il numero di automobili possedute dalle famiglie; il tempo impiegato nel percorso casa università; numero di carte elettroniche utilizzate da ciascuna famiglia.
* Il diploma di scuola media secondaria (Scientifico, Classico, Tecnico,…); il corso di laurea (Medicina e Chirurgia, Odontoiatria, Scienze Motorie, …); i criteri di valutazione della scuola media: Insufficiente, Sufficiente, Buono, Distinto; Ottimo. L’intensità con la quale si pratica una attività sportiva: nessuna, leggera, moderata, intensa.

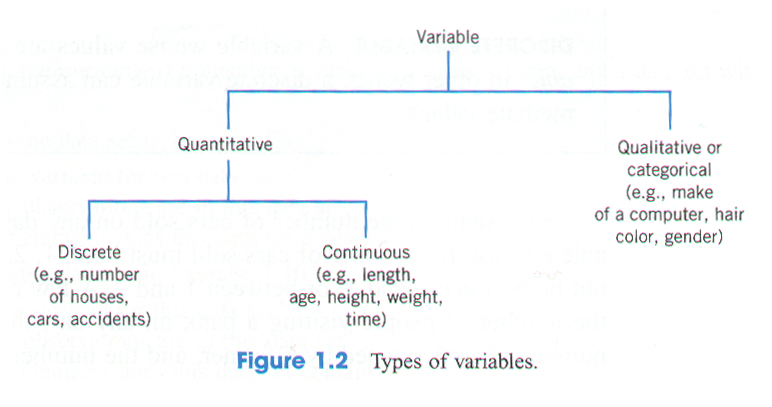
**Variabili quantitative**

Le variabili che possono assumere un valore numerico sono dette **variabili quantitative. Dati discreti.**

* Le **variabili** numeriche che possono assumere solo alcuni valori discreti (0, 1, 2, 3, …) ma non i valori intermedi sono dette **discrete** (dati discreti).
* Le **variabile** che possono assumere tutti i valori numerici compresi in un intervallo sono dette **continue** (dati continui).

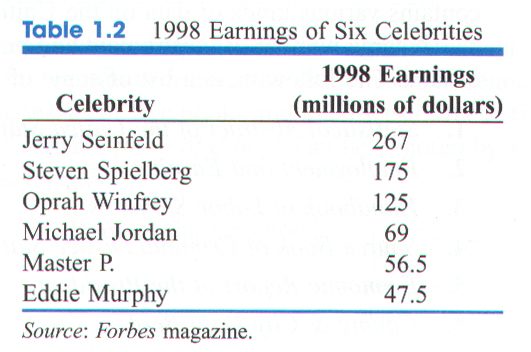
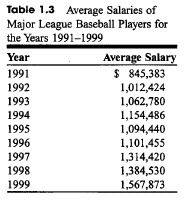
**Variabili qualitative o categoriali**

Le variabili che non possono assumere una valore numerico ma possono essere classificate in due o più categorie distinte sono dette **variabili qualitative o categoriali.** I **dati** così ottenuti sono detti **qualitativi**.

**Esempi di dati Qualitativi e Quantitativi**. Colore di una automobile, alunni per classe, stato civile di una persona

**Esempi di dati Discreti/Continui**. Numero di case possedute, Colore di una automobile, alunni per classe, stato civile di una persona

**Dati in Sezione (Cross-section data) e Serie di dati (Time series)**

* Il reddito delle famiglie italiane nel 2001; il numero dei rigori concessi a ciascuna squadra di serie A nel torneo in corso.
* Il numero degli studenti laureati in corso negli anni 1993-2000; la percentuale di tiri liberi realizzati dalla Fortitudo nei Play-Off.

I **dati in Sezione** contengono le informazioni raccolte nel *medesimo periodo di tempo* sui diversi elementi di un campione (popolazione). Le **Serie di** **dati** contengono le informazioni sul *medesimo elemento* di una campione per diversi periodi di tempo.

## Sommatoria o Notazione Somma

Campione con il prezzo di 5 libri di testo, espresso in Euro:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 25 | 60 | 37 | 53 | 13 |

I libri sono gli elementi del campione, il prezzo la variabile casuale ed i valori riportati in tabella sono le osservazioni della variabile da esaminare. Generalizziamo il discorso indicando con con *x* i valori della variabile e utilizziamo l’indice i-esimo per contraddistinguerli

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x*1 = 25 | *x*2 = 60 | *x*3 = 37 | *x*4 = 53 | *x*5 = 13 |

Sommiamo i valori *x*i

|  |
| --- |
| *x*1 +*x*2 + *x*3 + *x*4 + *x*5 = 25 + 60 + 37+ 53 + 13 = 192 ∈ |

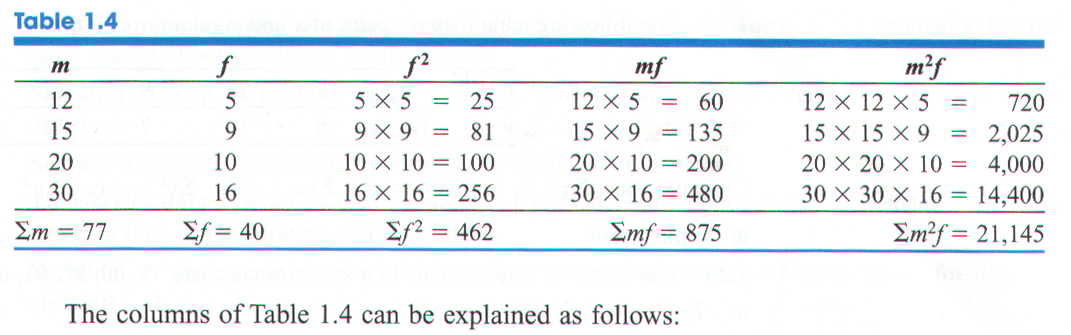
la somma di tutti i valori della variabile *x* è indicato con il simbolo greco *sigma* maiuscolo

|  |
| --- |
|  *x = x*1 +*x*2 + *x*3 + *x*4 + *x*5 = 25 + 60 + 37+ 53 + 13 = 192 ∈ |

**Esempio 1-2**. Dati i delle variabili *m* ed *f* valori in tabella

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *m* | 12 | 15 | 20 | 30 |
| *f* | 5 | 9 | 10 | 16 |

vogliamo calcolare

La soluzione è rappresentata in forma di tabella e può essere facilmente ottenuta utilizzando Excel. L’esercizio è sviluppato in dettaglio nel file <Sommatoria.xls> , dove sono stati inseriti alcuni esercizi aggiuntivi .

**Importante**. Il Capitolo 1 del testo Introductory Statistics è stato riprodotto in un file formato PDF e può essere richiesto direttamente al docente.

1. L’ estrazione di informazioni da un campione prende il nome di **indagine**; la raccolta delle informazioni relative all’intera popolazione prende il nome di **censimento**. [↑](#footnote-ref-1)
2. Le variabili vengono classificate anche come **nominali**, **ordinali**, di **intervallo** e di **rapporto**.

   Sono dette **variabili** **nominali** le variabili che permettono una classificazione di tipo qualitativo. La misura di una variabile qualitativa equivale a classificare ciascun elemento (individuo) del data set come appartenente ad una ben precisa categoria, presa fra un’inseme di categorie che npon hanno fra loro nessura regola di ordine o quantificazione. Per esempio possiamo classificare 2 individui dicendo che sono diversi sulla base di una variabile che indichiamo con A (ad es. appartengono ad una razza diversa: bianca e nera) ma non possiamo assolutamente dire che ciascun individuo ha “maggiore quantità” di uno dei valori di A (il soggetto è più bianco che nero). Tipici esempi di variabile nominale sono il sesso, il colore, la cittadinanza, ecc.

   Le **variabili ordinali** permettono di classificare e ordinare ciascun elemento del data set (individuo, osservazione) sulla base del fatto che possiedono in misura maggiore o minore la qualità rappresentata dalla variabile; tuttavia, le variabili ordinali non consentono di specificare nel dettaglio “quanto grande” sia questa differenza. Un esempio tipico di una variabile ordinale è lo stato socio economico delle famiglie: si è soliti indicare una famiglia appartenente alla classe media o medio-alta, ma non possiamo dire che la classe medio-alta è il 18% più elevata della classe media. La distinzione fra variabili nominali, ordinali, e di intervallo è un esempio di variabile ordinale. Infatti, possiamo dire che una grandezza di tipo nominale contiene meno informazione della ordinale, ma non possiamo dire “quanto di meno“ ne contiene, e analogamente non possiamo confrontare la loro differenza con la differenza fra variabile di intervallo.

   Le **variabili di intervallo** permettono di dare un punteggio alle osservazioni esaminate e al medesimo tempo di confrontare e quantificare la differenza fra ciascuna di esse con valori precisi. Per esempio la temperatura, misurata in gradi Celsius o Fahreneit, è indicata con un intervallo (scala) di valori. Sulla base di questa scala diciamo che la temperatura di 40 gradi è maggiore della temperatura di 30 gradi, così pure diciamo che un aumento di temperatura da 20 a 40 gradi è I doppio di un aumento da 30 a 40 gradi.

   Le **variabili di rapporto** sono molto simili alle variabili di intervallo; in aggiunta a tutte le proprietà delle variabili di intervallo hanno quale caratteristica un ben preciso valore di zero, e per questa ragione permettono di dare definizioni quali “il valore di x e pari al doppio di y” Esempio tipico di varaibili di rapporto sono le misure di spazio e di tempo. Altro esempio è la misura della temperatura in gradi Kelvin; infatti, questa unità di misura al il valore di 0 assoluto e per questa ragione possiamo dire che una temperatura di 200 gradi è maggiore di 100 gradi, ma anche che è il valore pari a due volte il valore 100 gradi. Le variabili o scale di intervallo non hanno proprietà di rapporto La maggior parte delle procedure di analisi statistica non distinguono fra le proprietà delle variabili di intervallo e delle variabili di rapporto nelle rispettive scale di misura [↑](#endnote-ref-1)