CONCETTI FONDAMENTALI DI STATISTICA

La Statistica è una scienza che grosso modo si divide in tre branche:

- statistica descrittiva (analisi dei dati);
- probabilità (elemento chiave);
- inferenza statistica (estensione dei risultati rilevati su un campione all'intera popolazione).

La Statistica si definisce come la **scienza delle decisioni in condizioni di incertezza**. Tale incertezza rende palese il concetto di probabilità come concetto chiave di tutta la disciplina.

Ciò detto, è opportuno evidenziare quelli che sono i concetti base che caratterizzano la branca della statistica descrittiva, fondamentale per comprendere gli sviluppi alla base di argomenti più complessi.

Studiare un fenomeno significa associare allo stesso il concetto di **variabile**, che rappresenta un determinato carattere di una popolazione statistica (altezza, peso, colore degli occhi, età, voto conseguito a un esame e così via).

La **popolazione statistica è il collettivo** su cui si osserva la variabile oggetto di studio.

Ogni manifestazione della variabile si chiama **modalità** della variabile. Ogni soggetto o oggetto su cui si osserva una modalità della variabile si chiama **unità statistica**. L'insieme delle unità statistiche definisce la popolazione statistica.

Una variabile può essere:

- **quantitativa**, se le sue modalità sono espresse da numeri (altezza, peso, età, voto conseguito a un esame);
- **qualitativa**, se le sue modalità non sono espresse da numeri (colore degli occhi, giudizio espresso su un libro, su un film su una canzone, per citarne alcune).

Una variabile quantitativa può essere a sua volta:

- **discreta**, se l'insieme delle sue modalità è elencabile;
- **continua**, nel caso contrario.

  Ad esempio, se io chiedo a 100 studenti il voto conseguito all'esame di Statistica, ho una popolazione statistica di 100 unità, su cui osservo 100 modalità della variabile quantitativa

voto. Tale variabile a sua volta è discreta, in quanto già prima di osservare le risposte dei 100 studenti, so che le stesse possono essere elencabili: 18, 19, 20 e così via fino a 30. Idem per l'età in anni compiuti: 0,1,2,3….Al contrario, se considerò l'età come tempo di vita vissuta, quindi espressa in anni, mesi, giorni e compagnia bella, sono di fronte a una variabile quantitativa continua.

Una volta raccolte le informazioni, posso schematizzarle in una **tabella di frequenza**. Questo ragionamento vale per qualsiasi tipologia di variabile. Da un lato vanno collocate le modalità osservate, dall'altro le **frequenze assolute**, ovvero il numero di volte in cui ciascuna modalità si presenta. **La somma delle frequenze assolute è uguale al numero complessivo di unità statistiche** su cui si studia la variabile di interesse (i 100 studenti di prima per intenderci).

Il rapporto tra ciascuna frequenza assoluta e il totale delle unità statistiche definisce il concetto di **frequenza relativa. La somma delle frequenze relative è sempre uguale a 1.**

Per ogni modalità esiste:

- una **frequenza assoluta cumulata** (pari alla somma tra la frequenza assoluta associata alla stessa modalità e le frequenze assolute associate a tutte le modalità precedenti);
- una **frequenza relativa cumulata** (pari alla somma tra la frequenza relativa associata alla stessa modalità e le frequenze relative associate a tutte le modalità precedenti).

Una volta organizzati i dati, si può procedere al calcolo degli indici di sintesi:

- **Moda**, corrispondente alla modalità che si presenta con maggiore frequenza (individuabile sia per variabili quantitative che per variabili qualitative);
- **Media aritmetica**, corrispondente alla somma di tutte le modalità diviso il numero complessivo di unità statistiche (calcolabile solo per variabili quantitative);
- **Mediana**, corrispondente al valore che distribuisce equamente in due parti la distribuzione ordinata delle modalità (calcolabile per variabili quantitative e qualitative ordinabili).

Per tutti i procedimenti connessi ai tre indici e per i relativi approfondimenti si rimanda alle lezioni registrate e alle altre dispense esemplificative.

Un ulteriore sviluppo della media aritmetica è dato dalla **media ponderata**. La differenza tra le due medie è la seguente: per la media aritmetica ogni modalità ha lo stesso "peso", la stessa "importanza" delle altre; per la media ponderata a ogni modalità è attribuito un "peso" specifico (ad esempio il numero dei crediti formativi associati a ciascun esame universitario; non a caso, il voto di ammissione all'esame di laurea è calcolato sulla base del procedimento tipico della media ponderata).

Ulteriori indici da calcolare sono gli indici di variabilità:

- **Varianza e scarto quadratico medio** per le variabili quantitative;
- **Indici di omogeneità ed eterogeneità** per le variabili qualitative.

Per tutti i procedimenti connessi ai succitati indici e per i relativi approfondimenti si rimanda alle lezioni registrate e alle altre dispense esemplificative.

Per quanto attiene alla forma di una distribuzione, la stessa può essere:

- **Simmetrica**, se media e mediana coincidono;
- **Asimmetrica positiva**, se la media è più grande della mediana;
- **Asimmetrica negativa**, se la media è più piccola della mediana.

Ai fini della forma, un ruolo importante è svolto anche dal cosiddetto **BOXPLOT**, diagramma a scatola che si compone di 5 capisaldi: valore minimo della distribuzione, primo quartile, mediana, terzo quartile, valore massimo della distribuzione. Il **primo quartile** è il valore prima del quale è collocato un quarto della distribuzione ordinata dei dati; il **terzo quartile** è quello prima del quale si collocano i tre quarti della distribuzione.
Per distribuzione ordinata dei dati si intende la successione dei valori in misura non decrescente.
La differenza tra terzo e primo quartile si definisce **campo di variazione interquartile**. Se la distanza tra mediana e primo quartile è uguale a quella tra terzo quartile e mediana, la distribuzione della variabile è simmetrica.