Descrizione grafica dei dati

In seguito ad una raccolta di dati si rende necessario sintetizzarli e rappresentarli in modo da poterli analizzarli in modo efficace. Non esiste un metodo universale per farlo, ma ci sono diverse soluzioni che si prestano in determinati casi.

La statistica descrittiva fa largo uso di:

- Grafici
- Tabelle
- Indici sugli aspetti più importanti

Come si esplorano i dati?

1. si sceglie il grafico più appropriato
2. Descriverne la forma della distribuzione e in base agli indici di forma
3. Calcolare una misura del centro della distribuzione
4. Calcolare un indice di dispersione

Metodi di rappresentazione delle variabili qualitative

- Distribuzione di frequenze
- Grafici a barre
- Grafici a torta
- Diagrammi di Pareto

La distribuzione di frequenza è una tabella in cui organizzare i dati formata da due colonne, nella colonna di sinistra ci sono i possibili valori delle variabili mentre nella colonna di destra sono riportate l'elenco delle frequenze per ogni classe.

Se lo scopo della nostra rappresentazione è quello di attirare l'attenzione sulla frequenza di ogni categoria allora il metodo di rappresentazione consigliato è il diagramma a barre mentre se vogliamo rappresentare la proporzione di ciascuna categoria, allora la scelta migliore è il diagramma a torta.

Diagrammi di Pareto

Vengono utilizzati per tenere conto delle problematiche che si verificano durante l'attività e sono stati teorizzati dallo studioso Alfredo Pareto. Sono composte da una serie di barre decrescenti da sinistra verso destra che rappresentano le frequenze delle cause di difettosità. La barra a sinistra indicherà le cause più frequenti, mentre l'ultima a destra l'esatto contrario. Viene usato per la maggior parte per separare le cause rilevanti da quelle irrilevanti.

Il grafico per serie storiche rappresenta una serie di dati rilevati in istanti di tempo diversi. Il tempo viene rappresentato sull'asse orizzontale mentre la variabile di interesse viene rappresentata sull'asse verticale.

Metodi di rappresentazione delle variabili quantitative

- Istogrammi
- Ogive
- Grafici ramo-foglia
- Distribuzione di frequenze

Nel caso di dati raggruppati in classi possiamo utilizzare istogrammi e ogive, nel caso di dati grezzi conviene utilizzare il diagramma ramo-foglia.

Le distribuzioni di frequenze in questo caso sono più complesse e per rappresentarlo dobbiamo avere degli accorgimenti.

si possono avere due casi:

- Possiamo avere una variabile discreta con poche modalità→ conviene utilizzare una normale distribuzione di frequenza
- Possiamo avere una variabile discreta con molte modalità→ conviene utilizzare una distribuzione delle frequenze di classe

Nel caso di distribuzione delle frequenze di classe per prima cosa dobbiamo determinare il numero delle classi di intervallo (k).

L'ampiezza delle classi (w) viene determinata facendo $\frac{valore\ massimo - valore\ minimo}{k}$ e nel caso sia necessario si arrotondano le classi per eccesso all'intero successivo.

Il numero delle classi normalmente si decide in modo arbitrario. Dobbiamo fare attenzione che ogni osservazione possa essere inserita in una sola classe.

Esistono diversi tipi di distribuzioni di frequenze:

- Relativa→ si ottengono dividendo ogni frequenza per il numero totale di osservazioni
- Cumulate→ si ottiene sommando la frequenza corrente alle frequenze precedenti, ha senso calcolarlo solo nei casi in cui le variabili siano:
  - Numeriche
  - Qualitative, ma ordinali
- Relative cumulate→segue il principio della cumulata, ma i dati si presentano sotto forma di percentuale

L'istogramma e l'ogiva sono grafici utili per rappresentare le distribuzioni di frequenze.

L'istogramma è un grafico composto da rettangoli verticali la cui area è proporzionale al numero di osservazione della classe corrispondente. Sulla linea orizzontale sono delimitate le classi di intervallo individuate nella distribuzione di frequenza, se queste hanno tutte la stessa ampiezza allora sarà l'altezza ad essere proporzionale al numero di osservazioni. La forma degli istogrammi rivela se i dati sono distribuiti simmetricamente rispetto al loro valore centrale, infatti in alcuni istogrammi, che prendono il nome di istogrammi simmetrici, il valore centrale li suddivide in due immagini speculari.

- La forma di un istogramma è detta simmetrica se le osservazioni sono bilanciate in modo approssimativamente regolare intorno al centro dell'istogramma.
- Una distribuzione asimmetrica si ha quando la distribuzione non è simmetrica rispetto al valore centrale della distribuzione, esistono due tipi di asimmetrie:
  - Asimmetria a destra→ si ha quando c'è una coda che si estende verso destra, cioè verso i valori positivi
  - Asimmetria a sinistra→ si ha quando c'è una coda che si estende verso sinistra, cioè verso i valori negativi

L'ogiva, detta anche curva di frequenze cumulate, è una spezzata che rappresenta la distribuzione delle frequenze percentuali cumulate.

Fino ad ora ci siamo limitati a rappresentare insiemi di dati di una sola variabile, tuttavia è possibile rappresentare anche una relazione tra variabili, esempi di queste rappresentazioni sono:

- Diagrammi di dispersione→ servono a studiare le possibili relazioni tra due variabili quantitative. Vengono creati associando un punto del piano cartesiano a ogni coppia di valori che costituiscono un'osservazione congiunta delle due variabili. Questa tipologia di grafico evidenzia:
  - I possibili valori di ogni variabile
  - La distribuzione dei dati all'interno dei valori possibili
  - L'eventuale relazione tra le due variabili
  - La presenza di eventuali valori anomali
- Tabella a doppia entrata →servono a studiare le possibili relazioni tra due variabili qualitative. Elencano la frequenza delle osservazioni per ogni combinazione di classi di misura di due variabili. Il numero di celle di cui è composto questo grafico dipende dal numero possibile di combinazione delle due variabili, se entrambe le variabili sono qualitative queste tabelle vengono chiamate "tabelle di contingenza".

In queste tabelle le variabili prendono il nome di:

- variabile dipendente (y)
- variabile indipendente (x)

i grafici sono strumenti efficaci per la rappresentazione dei dati, a patto che sia fatti con criterio, ecco alcuni errori tipici in cui ci possiamo imbattere:

- istogrammi ingannevoli→ solitamente l'ampiezza delle classi di intervallo dovrebbe essere omogenea, infatti non è consigliabile creare istogrammi con basi diverse nonostante sia da tener conto delle aree dei rettangoli e non la loro altezza in quanto questo potrebbe portare facilmente ad un errore di valutazione.
- Serie storiche ingannevoli→ possono dare l'impressione di una certa stabilità anche se non è così

Come raggruppare i dati in classi

1. Determinare il numero di classi di intervallo, solitamente vanno da 5 a 20 e si scelgono arbitrariamente in base al caso in questione
2. Determinare l'ampiezza degli intervalli→ ampiezza = (massimo-minimo) / numero classi e si arrotonda
3. Determinare la regola di chiusura degli intervalli in modo che non siano presenti valori che possono essere inseriti in due classi, la chiusura può essere a destra o sinistra