Statistica e analisi dei dati

Kevin Muka

Contents

L01 - 25/02/2025

[. . .] **Dispense L01-Introduzione_a_python**

L02 - 27/02/2025

[. . .] **Dispense L02-Pandas**

L03 - 04/03/2025

Dispense L03-Dati_e_frequenze

Dati quantitativi e qualitativi

Una delle principali distinzioni che si possono fare sui dati osservabili riguarda il modo in cui questi sono misurati:

- si parla di dati *quantitativi* se l'esito della misurazione è una quantità numerica;
- si parla invece di dati *qualitativi* (o categorici / nominali) quando la misurazione è fatta scegliendo un'etichetta a partire da un insieme disponibili.

Classificazione dei dati qualitativi I dati qualitativi si distinguono spesso in binari (o booleani), nominali e ordinali. I dati binari o booleani si hanno quando l'osservazione può assumere soltanto due valori tra loro non confrontabili: in questo contesto, si usa talvolta "booleani" per evidenziare la presenza o assenza di una proprietà, mentre "binari" per indicare due etichette possibili. Anche i dati nominali (detti anche sconnessi) non ammettono un confronto d'ordine tra i valori, e i dati binari ne sono un caso particolare. In altre parole, con dati nominali si può solo dire se due osservazioni sono uguali o diverse, senza poter stabilire quale sia "maggiore" o "minore". Nei dati ordinali, invece, esiste una relazione d'ordine tra i valori osservabili, per cui, se due valori sono diversi, è anche possibile stabilire quale sia il più piccolo e quale il più grande.

Classificazione dei dati quantitativi I dati quantitativi sono spesso distinti in discreti e continui a seconda dell'insieme di valori che possono assumere. Tuttavia, poiché i dati vengono memorizzati su un computer, i valori reali risultano approssimati da valori appartenenti a un insieme finito (quindi discreto). Per questo motivo, è più utile ragionare su quei caratteri per cui ha senso attribuire un significato a un singolo valore e su quei caratteri in cui, normalmente, si considerano intervalli di valori. In alcuni casi, inoltre, i caratteri quantitativi si distinguono a seconda che abbia senso o meno valutare il rapporto tra i valori osservati.

Frequenze assolute e relative

La frequenza assoluta rappresenta il conteggio di quante volte una determinata osservazione compaia in un campione. Questa misura è particolarmente facile da analizzare quando il numero di osservazioni differenti non è troppo elevato, come spesso accade con i caratteri qualitativi, e più raramente per quelli quantitativi.

Oltre a questa, si definisce la frequenza relativa come il rapporto tra la frequenza assoluta di un'osservazione e il numero totale di osservazioni, consentendo di esprimere la presenza di ogni valore in termini di proporzione o percentuale rispetto all'intero campione.

Capitolo 1 - RS

1.1 Introduzione

La **statistica** è l'arte di apprendere dei dati. Si occupa della raccolta, della descrizione e dell'analisi dei dati, possibilmente permettendo di trarne delle conclusioni.

La parte della statistica che si occupa di descrivere e riassumere i dati si chiama **statistica descrittiva**.
La parte della statistica che si occupa invece di trarre conclusioni dai dati si chiama **statistica inferenziale**.

1.3 Popolazioni e campioni

Nella statistica è cruciale ottenere delle informazioni su tutto un insieme di elementi, che vengono definiti **popolazione**. Spesso la popolazione però è troppo numerosa per poter analizzare ciascuno dei suoi membri: in questo csaso si sceglie e si esamina un suo sottoinsieme, che viene definito **campione**.

Affinché il campione ci dia informazioni su tutta la popolazione, esso deve essere scelto in modo da essere **rappresentativo** di tutta la popolazione. Rappresentativo significa che il campione deve essere scelto in modo che tutte le parti della popolazione abbiano uguale probabilità di fare parte del campione. Il campione deve quindi riflettere la variabilità reale della popolazione.

Un campione di k membri di una popolazione si dice **campione casuale**, o talvolta *campione casuale semplice*, se i membri sono scleti in modo che tutte le possibili sclete dei k membri siano ugualemente probabili.

Una volta che si sceglie un campione casuale, è possibile usare l'inferenza statistica per giungere a conclusioni sull'intera popolazione studiando gli elementi del campione.

1.3.1 Campionamento casuale stratificato Un metodo più sofisticato del campionamento casuale semplice è il campionamento casuale stratificato. Inizialmente si stratifica la popolazione in sottopopolazioni, ognuno dei quali contiene unità simili secondo determinati criteri. Ins eguito, da ogni strato si estrae un casualmente un numero di unità proporzionale alla sua consistenza nella popolazione totale. In questo modo, le proporzioni di ciascuno strato presenti nel campione rispecchiano esattamente quelle dell'intera popolazione.

La stratificazione è particolarmente efficace per conoscere il membro *medio* della popolazione totale quando ci sono differenze tra le sottopopolazioni rispetto alla questione studiata.

Capitolo 2 - RS

2.2 Tabelle e grafici di frequenza

Una **tabella di frequenza** associa, a ciascun valore distinto osservato in un insieme di dati, il numero (o la **frequenza**) di volte in cui quel valore compare nel campione. Data una variabile statistica X che può assumere vari valori, si elencano i valori distinti di X in una colonna e, a fianco di ognuno, la relativa frequenza di occorrenza nel campione. La somma delle frequenze deve corrispondere al numero totale di osservazioni.

2.2.1 Grafici a bastoncini, a barre e poligonali

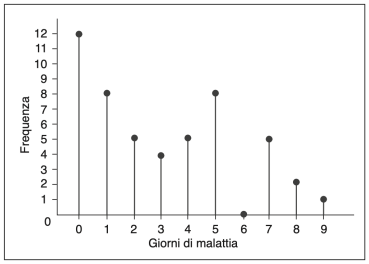


Figura 2.1 Un grafico a bastoncini.

I dati di una tabella di frequenza possono essere rappresentati graficamente in diversi modi. Uno dei più intuitivi è il **grafico a bastoncini**, in cui i valori della variabile statistica sono disposti lungo l'asse orizzontale, mentre le frequenze si riportano sull'asse verticale. Ogni valore viene quindi associato a un semplice segmento che parte dall'asse orizzontale e arriva all'altezza corrispondente alla relativa frequenza.

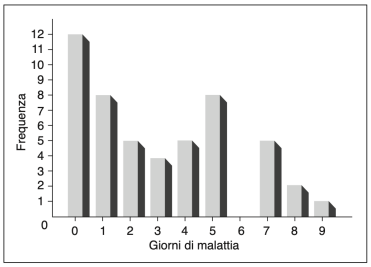


Figura 2.2 Un grafico a barre.

Un secondo tipo di rappresentazione, molto simile concettualmente, è il **grafico a barre**: anche in questo caso i valori si trovano sull'asse orizzontale e le frequenze su quello verticale, ma invece dei singoli segmenti si utilizzano barre di un certo spessore. Ciò permette di mettere in evidenza ciascuna categoria o classe di dati e risulta particolarmente efficace quando si vogliono confrontare categorie di grandezza diversa.

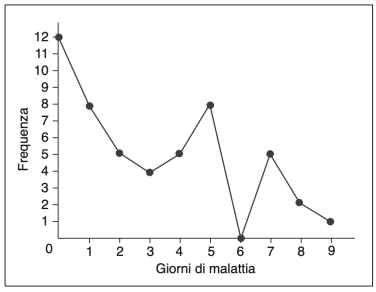


Figura 2.3 Un grafico poligonale.

Infine, esiste il **grafico poligonale**, in cui i valori (sempre disposti sull'asse orizzontale) vengono rappresentati da punti, collocati a un'altezza proporzionale alla loro frequenza, che vengono poi congiunti da segmenti. In questo modo si ottiene una linea spezzata che rende immediata la visualizzazione delle variazioni di frequenza da un valore all'altro, permettendo di apprezzare più facilmente tendenze o andamenti complessivi.

Simmetria dei dati Un insieme di dati si dice **simmetrico** attorno a un valore x_0 se, per ogni scostamento c da x_0, la frequenza dei valori $(x_0 - c)$ è uguale a quella dei valori $(x_0 + c)$. In tal caso, il valore x_0 si definisce "centro di simmetria" della distribuzione. Se i dati non sono perfettamente simmetrici, ma la distribuzione è comunque "quasi" speculare rispetto a un punto centrale, si parla di **quasi simmetria**. Un modo semplice per rendersi conto se una distribuzione è (quasi) simmetrica è rappresentarla graficamente (ad esempio con un grafico a barre) e osservarne la forma.

2.2.2 Grafici delle frequenze relative

Può risultare necessario considerare e rappresentare le **frequenze relative** piuttosto che quelle assolute. Se f è la frequenza di occorenza del valore x di un dato, allora possiamo rappresentare in un grafico la frequenza relativa f/n rispetto ad x, dove n è il numero totale di osservazioni nell'insieme di dati.

Per costruire una tabella delle frequenze relative da un insieme di dati, bisogna innanzitutto disporre i valori dei dati in ordine crescente. Si determinano i valori distinti e quante volte ciascuno di essi compaia. Si elencano questi valori distinti affiacati dalla loro frequenza f e dalla loro frequenza relativa f/n, dove n è il numero totale di osservazioni nell'insieme di dati.

2.2.3 Grafici a torta

[...]

2.3 Raggruppamento dei dati e istogrammi

Utilizzare i grafici presentati precedentemente è un metodo efficace per descrivere un insieme di dati. Tuttavia alcuni di questi insiemi hanno troppi valori distinti per poter usare questo metodo.

Quello che bisogna fare in questi casi è suddividere i valori in gruppi, o **classi**, e poi si rappresenta con un grafico il numero di valori deti dati che cadono in ciascuna classe. Il numero di classi scelte è un compromesso tra:

1. scegliere poche classi al costo di perdere molte informazioni sui valori effettivi in una classe
2. scegliere troppe classi, ottendendo frequenze troppo basse all'interno di ciascuna di esse

I valori al bordo di una classe si chiamano *estremi* della classe. Si adotta la convenzione di inclusione a sinistra, che richiede che una classe includa il suo estremo sinistro ma non quello destro.

Una volta suddivisi i dati in classi, si costruisce la tabella delle frequenze (e delle frequenze relative), e da questa si ottiene l'istogramma, un grafico a barre adiacenti che mostra la distribuzione dei dati in ciascuna classe. L'istogramma offre una visione immediata di come i valori si distribuiscono: per esempio, se sono concentrati in un certo intervallo, se ci sono "vuoti" senza osservazioni o se alcuni valori si distaccano notevolmente dagli altri. Pur non contenendo tutte le informazioni dell'insieme di dati originale, la tabella delle frequenze di classe e l'istogramma illustrano le caratteristiche fondamentali della distribuzione, come la simmetria, la dispersione e i possibili estremi isolati.

2.4 Diagrammi ramo-foglia

[. . .] **Dispense L03-Dati_e_frequenze** in fondo

2.5 Insieme di dati a coppie

Un insieme di dati può consistere in coppie di valori che hanno una relazione di qualche tipo tra di loro. Ne viene che ogni elemento dell'insieme di dati sia costituito da un valore x e da uno y. Si indica con $(x_i, y_i),\ i = 1 \cdots n$ la i-esima coppia.

Un metodo per rappresentare un insieme di dati di questo tipo consiste nel considerare ogni elemento della coppia separatamente, producendo istogrammi (o diagrammi ramo-foglia) separati per ciascuno. Così facendo però, nonostante i due grafici ci diano molte informazioni sulle singole variabili (attributi), non si ha nessun tipo di informazione riguardo al rapporto tra queste due variabili.

Per capirne la relazione è necessario considerare i valori accoppiati di ciascun dato simultaneamente. Si possono allora rappresentare questi dati accoppiati in un diagramma rettangolare e bidimensionale, in cui l'asse x rappresenta il valore x dei dati, e l'asse y il valore y. Così facendo si ottiene un **diagramma di dispersione**.

Una delle ragioni per cui questo tipo di diagramma è utile consiste nella possibilità di **fare previsioni** sul valore y di una futura osservazione, noto il valore x. Per stimare il valore y a partire da x si cerca, in modo intuitivo, di tracciare una "retta media" che approssimi l'andamento dei punti sul diagramma, ovvero una retta che passi "il più vicino possibile" a tutti i dati.

- In pratica, si ricorre a **metodi di regressione lineare**, come il *metodo dei minimi quadrati*, che permette di trovare l'equazione della retta (del tipo$ y = a + bx$) minimizzando la somma delle distanze (al quadrato) tra i valori osservati (x_i, y_i) e i valori $\hat{y}_i$ previsti dalla retta. Una volta trovata questa retta di "miglior adattamento", per un qualunque valore x che possa presentarsi in futuro, si ottiene la stima di y semplicemente sostituendo x nell'equazione $y = a + bx$.

Il diagramma di dispersione, oltre a mostrare il comportamento relativo di due variabili e ad aiutrarci nelle previsoni, è utile per riconoscere i **valori anomali** (outlier) che sono i punti che non sembrano seguire il comportamento degli altri. Una volta identificati questi valori, si può decidere quali di essi siano appropriati e quali siano invece causati da errori nella racolta dei dati.

L04 - 06/03/2025

Dispense L03-Dati_e_frequenze

Frequenze cumulate

Le frequenze cumulate si ottengono quando esiste un ordinamento sui valori di una variabile. Il procedimento consiste nel disporre i valori in ordine crescente, calcolare le loro frequenze individuali e poi sommarle progressivamente: al primo valore si associa la sua frequenza, al secondo la somma della frequenza del primo e del secondo, al terzo la somma delle frequenze dei primi tre, e così via.

È importante notare che l'ultima frequenza cumulata rappresenta il totale dei casi osservati. Inoltre, il concetto di frequenza cumulata vale sia per le frequenze assolute che per quelle relative; nel caso delle frequenze relative, i valori variano da 0 a 1.

Quando i dati sono numerici o comunque ordinabili, un concetto affine alle frequenze relative cumulate è quello della *funzione cumulativa empirica* (nota anche come *funzione di ripartizione empirica*). Data una serie di osservazioni $x_1, \ldots, x_n$, tale funzione $\hat{F} : \mathbb{R} \to [0, 1]$ è definita in modo che per ogni $x \in \mathbb{R}$ essa assuma il valore pari alla frequenza relativa delle osservazioni minori o uguali a x. In altre parole:

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{(-\infty, x]}(x_i)$$

dove $\mathrm{I}_A : \mathbb{R} \to 0, 1$ è la funzione indicatrice dell'insieme A, che restituisce 1 se l'argomento appartiene ad A e 0 altrimenti, e l'intervallo $(-\infty, x]$ include tutti i valori minori o uguali a x. Pertanto, per ogni x, $\hat{F}(x)$ rappresenta la frequenza relativa cumulata del massimo valore osservato che non supera x, e il grafico di questa funzione sarà a tratti costanti.

$$I_A(x) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases} \quad \Rightarrow \quad I_{(-\infty,x]}(x_i) = \begin{cases} 1 & \text{se } x_i \in (-\infty, x] \\ 0 & \text{se } x_i \notin (-\infty, x] \end{cases} = \begin{cases} 1 & \text{se } x_i \leq x \\ 0 & \text{se } x_i > x \end{cases}$$

In pratica rappresenta il numero di osservazioni dei miei campioni che sono minori o uguali di una certa x, diviso per il numero totale di campioni. La divisione per n è fatta per avere dei valori relativi.

Diagrammi di Pareto I diagrammi di Pareto sono grafici a barre ordinati in ordine decrescente di frequenza, ai quali è spesso affiancata una linea che rappresenta la frequenza cumulata. In questo modo, oltre a mostrare il numero di casi per ciascuna categoria, permettono di evidenziare quali categorie contribuiscono maggiormente al totale, facilitando l'individuazione delle cause o delle categorie più rilevanti.

Frequenze congiunte e marginali

Quando si analizza un insieme di osservazioni, può essere particolarmente utile considerare due caratteri contemporaneamente, in modo da verificare se esiste una relazione tra i valori dei due attributi. In questo caso, il concetto di frequenza si adatta contando il numero di occorrenze in cui i due caratteri assumono contemporaneamente determinati valori. Questo conteggio porta alla definizione di *frequenza congiunta assoluta*; se invece si considera la frazione delle osservazioni, si parla di *frequenza congiunta relativa*.

Quando il numero dei possibili valori osservabili per i caratteri non è elevato, è possibile rappresentare visivamente queste frequenze tramite una *tabella delle frequenze congiunte* o *tabella di contingenza*. In tale tabella, le righe sono associate ai valori di uno dei caratteri, mentre le colonne rappresentano i valori del secondo carattere. Gli elementi all'interno della tabella indicano le frequenze congiunte (assolute o relative) per le coppie di valori.

Per facilitare ulteriori analisi, si riportano spesso nelle ultime colonne e nelle ultime righe della tabella le *frequenze marginali*, ottenute sommando rispettivamente i valori per ogni riga e per ogni colonna. Se si desiderano valori relativi, questi totali devono essere normalizzati rispetto al numero complessivo delle osservazioni.

Capitolo 3 - RS

Le quantità numeriche calcolate a partie da un insieme di dati si chiamano **statistiche**.

3.1 Centralità [3.1]

Verranno presentate le statistiche che descrivono la tendenza centrale di un insieme di dati, ossia delle statistiche che descrivono il centro di un insieme di dati. Questa proprietà che si può individuare in un insieme di dati è detta **centralità** o posizione.

Esistono tre indici di posizione: media, mediana e moda. In tutti i tre i casi si parla di campionaria, in quanto sono effettuate su dei campioni.

3.1.1 Media campionaria [3.2]

Si supponga di avere un campione di n dati i cui valori sono $x_1, x_2, \cdots, x_n$. Una statistica per indicare il centro di questo insieme di dati è la media campionaria, definita come la media aritmetica dei valori dati:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Si osserva che $\bar{x}$ può non corrispondere ad uno dei dati x_i con $1 \leq i \leq n$ presi in considerazione.

3.1.1.1 Trasformazioni

Traslazione Si consideri ancora lo stesso insieme di dati. Se ciascun valore viene incrementato di una costante b, allora anche la media campionaria viene incrementata di b:

$$y_i = x_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = \bar{x} + b$$

dove $\bar{y}$ e $\bar{x}$ sono le medie campionarie risettivamente degli y_i e degli x_i.

Dimostrazione: $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} h}_{\frac{1}{n}\cdot nh} = \bar{x} + h$$

Scalatura Se invece ciascun calore dei dati viene moltiplicato per a, lo è anche la media campionaria:

$$y_i = ax_i \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = a\bar{x}$$

Dimostrazione: $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ax_i = a\frac{1}{n}\sum_{i=1}^{n} x_i = a\bar{x}$$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = ax_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = a\bar{x} + b$$

Queste tre proprietà derivano dal fatto che tutte queste trasformazioni siano lineari.

3.1.1.2 Media pesata Quando i dati sono disposti in una tabella delle frequenze, la media campionaria può essere espressa come la somma del prodotto dei valori distinti per le loro frequenze, divisi per la dimensione dell'insieme dei dati.

Per verificarlo, si supponga di disporre di una tabella delle frequenze che elenca k valori distinti $x_1, x_2, \cdots, x_k$ con le rispettive frequenze $f_1, f_2, \cdots f_k$. Ne segue che questo insieme di dati è costituito da n osservazioni, dove $n = \sum_{i=1}^{k} f_i$ e dove il valore x_i comprare f_i volte per $i = 1, 2, \cdots, k$. La media campionaria per questo insieme di dati è:

$$\bar{x} = \frac{x_1 + \cdots + x_1 + x_2 + \cdots + x_2 + \cdots + x_k + \cdots + x_k}{n} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n} \tag{3.1}$$

Ora, se $w_1, w_2, \cdots, w_k$ sono numeri non negativi la cui somma è 1, allora

$$w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$$

prende il nome di **media pesata** dei valori $x_1, x_2, \cdots, x_k$ dove w_i è il peso di x_i.

Scrivendo l'equazione (3.1) come

$$\bar{x} = \frac{f_1}{n} x_1 + \frac{f_2}{n} x_2 + \cdots + \frac{f_k}{n} x_k$$

possiamo vedere che la media campionaria $\bar{x}$ è la media pesata dell'insieme dei valori distinti. Il peso assegnato al valore x_i è f_i/n, ossia rappresenta la frazione di volte in cui il valore x_i comare nell'insieme dei dati.

3.1.1.3 Scarti [3.2.1] Si supponga che l'insieme di dati sia costituito dagli n valori $x_1, \cdots, x_n$ e che $\bar{x} = \sum_{i=1}^{n} x_i/n$ sia la media campionaria. Le differenze tra ciascun valore dei dati e la media campionaria si chiamano **scarti**. Il valore dell'i-esimo scarto è $x_i - \bar{x}$

La somma di tutti gli scarti è sempre 0, ovvero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Questa uguaglianza afferma che la somma degli scarti positivi della media campionaria controbilancia esattamente la somma degli scarti negativi.

Utilizzando un linguaggio fisico, questo significa che se n pesi dotati della stessa massa vengono posti su un'asta nei punti x_i con $i = 1, \cdots, n$, allora $\bar{x}$ è il punto in cui l'asta può essere messa in equilibrio. Questo punto di equilibrio si chiama centro di gravità.

3.1.2 Mediana campionaria [3.3]

La media campionaria presenta un forte punto debole come indicatore del centro di un insieme di dati: il suo valore è infatti ampliamente influenzato da eventuali valori estremi (valori fuori scala).

Si dispongano i valori dei dati in ordine crescente. Se il numero di valori è dispari, allora la mediana campionaria è il valore intermedio della lista ordinata; se è pari, allora la mediana campionaria è la media dei due valori intermedi.

Sia $x_{(i)}$ l'i-esimo dato del campione ordinato in maniera crescente, la mediana m è definita come:

$$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{per } n \text{ dispari} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)/2 & \text{per } n \text{ pari} \end{cases}$$

La media campionaria e la mediana campionaria sono due statistiche utili per descrivere la tendenza centrale di un insieme di dati. Il loro utilizzo è però molto diverso, in quanto la media campionaria (essendo una media aritmetica) prende in considerazione tutti i valori dell'insieme di dati, mentre invece la mediana campionaria, dato che considera solo uno o due valori centrali, non è influenzata dai valori estremi.

Per gli insiemi di dati che sono approssimativamente simmetrici rispetto ai valori centrali, la media campionaria e la mediana campionaria sono vicine. Entrambe le statistiche sono informative, e il loro utilizzo dipende dal contesto.

3.1.2.1 Percentili campionari [3.3.1] La mediana campionaria è un caso particolare di una statistica nota come $100p$-esimo percentile campionario, dove p indica qualunque frazione compresa tra 0 e 1. **[...]**

3.1.3 Moda campionaria [3.4]

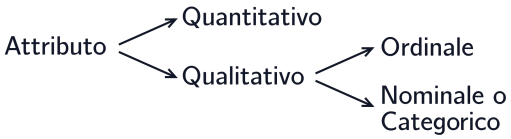
Un altro indicatore della tendenza centrale è la moda campionaria, che è il valore che si verifica con maggiore frequenza nell'insieme di dati.

Se non esiste un singolo valore che si verifica con più frequenza rispetto agli altri, allora tutti i valori con la frequenza più alta sono detti *valori modali*. In questo caso si dice che non c'è un valore unico della moda campionaria.

Questi valori si vedono facilmente in una tabella delle frequenza; sono infatti i valori con la frequenza più alta.

Riepilogo

Si considerino le varie classificazioni degli attributi:



La media si può fare solo per gli attributi quantitativi; la mediana è possibile svolgerla anche sugli attributi qualitativi ordinali con cardinalità del campione dispari; la moda si può fare per qualsiasi tipo di attributo.

3.2 Dispersione [3.5]

Due campioni A e B possono presentare la stessa centralità ma essere molto diversi tra loro. Si considerino:

$$A : 1, 2, 5, 6, 6 \qquad B : -40, 0, 5, 20, 35$$

Entrambi i campioni hanno la stessa media campionaria e la stessa mediana campionaria, però i valori contenuti nell'insieme B sono decisamente più sparsi di quelli nell'insieme A.

Un modo per misurare la dispersione dei dati è considerare gli scarti dei valori dei dati rispetto ad un valore centrale. Il valore centrale più usato per questo scopo è proprio la media campionaria. Se i valori dei dati sono $x_1, \cdots, n_n$ e la media campionaria è $\bar{x} = \sum_{i=1}^{n} x_i/n$, allora lo scarto del valore x_i dalla media campionaria è $x_i - \bar{x}$ con $i = 1, \cdots, n$.

Si potrebbe pensare di misurare la dispersione totale di un insieme di dati calcolando la media aritmetica degli scarti dalla media. Tuttavia, come abbiamo osservato precedentemente, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$; questo significa che la somma degli scarti rispetto alla media campionaria è sempre uguale a 0, e di conseguenza lo è anche la media aritmetica degli scarti.

Questo avviene proprio perché gli scarti positivi e negativi si cancellano tra di loro. Si vogliono quindi considerare i singoli scarti indipendentemente dal segno. Si può ottenere questo risultato sia considerando il valore assoluto degli scarti che, come risulta più utile in pratica, il quadrato.

3.2.1 Varianza campionaria

La varianza campionaria è una misura della media degli scarti quadratici rispetto alla media campionaria. Tuttavia, per ragioni tecniche questa "media" divide la somma di n scarti quadratici per $n-1$, piuttosto che per l'usuale valore n.

La varianza campionaria si può calcolare solo per attributi quantitativi, e a differenza degli indici di centralità presenta un problema: la sua unità di misura è diversa da quella dei singoli dati del campione.

La varianza campionaria s^2 dell'insieme di dati $x_1, \cdots, x_n$ di media $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ è definita come

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

L'identità algebrica che segue è utile per calcolare la varianza campionaria a mano:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 \; - \; n\,\bar{x}^2 \tag{3.2}$$

3.2.1.1 Trasformazioni

Traslazione Si supponga di sommare una costante b a ciascuno dei valori $x_1, \cdots, x_n$ per ottenere un nuovo insieme di dati, la varianza campionaria non cambia:

$$y_i = x_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = s_x^2$$

Si ricordi che $\bar{y} = \bar{x} + b$ e quindi $y_i - \bar{y} = x_i + b - (\bar{x} + b) = x_i - \bar{x}$. Questo significa che gli scarti di y sono uguali agli scarti di x, e quindi anche le somme dei quadrati sono uguali.

La varianza campionaria quindi non cambia se sommiamo una costante a ciascun valore. *Questa proprietà può essere utilizzata insieme all'identità algebrica* (3.2) *per semplificare il calcolo della varianza campionaria.*

Scalatura Se ciascun valore dei dati viene moltiplicato per a, la varianza campionaria viene moltiplicata per il quadrato di a:

$$y_i = a x_i \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = a^2 s_x^2$$

Dimostrazione: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[a(x_i - \bar{x})\right]^2 = a^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = a^2 s_x^2$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = a x_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = a^2 s_x^2$$

3.2.2 Deviazione standard campionaria

La radice quadrata positiva della varianza campionaria si dice deviazione standard campionaria, e si indica con s. Questa è definita come

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

La deviazione standard campionaria, a differenza della varianza campionaria, è espressa nella stessa unità di misura dei dati originali.

3.2.2.1 Trasformazioni

Traslazione Si supponga di sommare una costante b a ciascuno dei valori $x_1, \cdots, x_n$ per ottenere un nuovo insieme di dati, la deviazione standard campionaria non cambia:

$$y_i = x_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = s_x$$

Scalatura Se ciascun valore dei dati viene moltiplicato per a, si ottiene che $s_y^2 = a^2 s_x^2$. Calcolando la radice quadrata di entrambi i membri dell'uguaglianza si ottiene che la deviazione standard dei valori y è uguale al valore assoluto di a moltiplicato per la deviazione standard dei valori in x:

$$y_i = ax_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = |a| \, s_x$$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = ax_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = |a| \, s_x$$

Due altri indicatori della dispersione di un insieme di dati frequentemente utilizzati sono l'**intervallo di variazione**, ossia la differenza fra il più grande e il più piccolo valore, e lo **scarto interquartile**, ossia la lunghezza dell'intervallo in cui troviamo la metà centrale dei dati.