

Riassunto Statistica

Dimix

2025-03-14

Indice

Capitolo 1 – Introduzione alla Statistica	2
Definizione	2
Popolazione e campioni	2
La statistica descrittiva	3
Capitolo 2 - Descrivere insiemi di dati	3
Dati quantitativi e qualitativi	3
Frequenze	3
Grafici	4
Capitolo 3 - Statistiche	7
Centralità	7
Media campionaria	7
Mediana campionaria	8
Percentili campionari	9
Moda campionaria	9
Dispersione	10
Varianza campionaria	10
Deviazione standard campionaria	11
Scarto interquantile	12
Altri grafici	12
Box Plot	12
Q-Q Plot	12
Distribuzioni normali	13
Indici di dipendenza	13
Covarianza campionaria	13
Coefficiente di correlazione di Pearson	14
Indici di eterogeneità	16
Indice di Gini	16
Indice di entropia	17
Formulario	19

Capitolo 1 – Introduzione alla Statistica

Definizione

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 **statistica descrittiva** è quella parte della statistica che si occupa di descrivere e riassumere i dati

La **statistica inferenziale** è quella parte della statistica che si occupa di trarre conclusioni dai dati

La statistica inferenziale si basa sul modello probabilistico che consiste nel fare un insieme di assunzioni sulle probabilità di ottenere un certo valore, per cui essa richiede la conoscenza della teoria della probabilità. L'inferenza statistica si basa sull'assunzione che importanti aspetti del fenomeno in analisi si possano rappresentare in termini di probabilità e giunge a conclusioni usando i dati per fare inferenza su queste probabilità.

Popolazione 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 caso 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, quindi esso deve riflettere la variabilità reale della popolazione.

Un campione di k membri di una popolazione si dice **campione casuale**. o talvolta *campione casuale semplice* quando i membri sono scelti in modo che tutte le possibili scelte dei k membri siano ugualmente probabili.

Una volta scelto il campione casuale, si può utilizzare l'inferenza statistica per giungere a conclusioni sull'intera popolazione studiando gli elementi del campione.

Campione casuale stratificato

Un metodo più sofisticato del campionamento casuale semplice è il **campionamento casuale stratificato**. inizialmente si stratifica la popolazione in *sottopopolazioni*, ognuna delle quali contiene unità simili secondo determinati criteri. in seguito da ogni strato si estrae casualmente un numero di unità proporzionale alla sua consistenza nella popolazione totale. in questo modo, le proporzioni di ciascun 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.

La statistica descrittiva

Capitolo 2 - Descrivere insiemi di dati

Dati quantitativi e qualitativi

Una distinzione che si può fare sui dati osservabili riguarda il modo in cui questi sono misurati:

- **Dati quantitativi:** l'esito della misurazione è una quantità numerica.
- **Dati qualitativi:** l'esito della misurazione è un'etichetta appartenente a un insieme fissato di etichette, vengono anche detti categorici o nominali.

Classificazione dei dati qualitativi

I dati qualitativi si distinguono in dati binari, nominali e ordinali:

- Dati binari o booleani: possono assumere soltanto due valori tra loro non confrontabili, si utilizza *booleani* per evidenziare la presenza/assenza di una proprietà, mentre binari per indicare due etichette possibili.
- Dati nominali: non ammettono un confronto d'ordine tra i valori, ma è possibile stabilire una relazione di equivalenza.
- Dati ordinali: se abbiamo due valori diversi riusciamo a stabilire quale sia il più piccolo e quale il più grande, quindi esiste una relazione d'ordine tra i valori.

Classificazione dei dati quantitativi

I dati quantitativi si distinguono in discreti e continui a seconda dell'insieme di valori che possono assumere:

- Dati discreti: costituiscono variabili che possono assumere un insieme numerabile di valori distinti e separati, ad ogni valore corrisponde un significato specifico.
- Dati continui: in teoria possono assumere un qualsiasi valore all'interno di un intervallo, anche se nella pratica vengono approssimati a una precisione finita, per via della memorizzazione digitale.

Frequenze

La **frequenza assoluta** di un'osservazione x in un insieme di dati $A = \{x_1, \dots, x_n\}$ è definita come il numero di volte in cui x compare in A .

In modo formale possiamo indicarla con f_x la frequenza assoluta di x , si ha che $f_x = \#\{j \in \{1, \dots, n\} \mid x_j = x\}$

La **frequenza relativa** consente di esprimere la presenza di ogni valori in termini di proporzione rispetto all'intero campione. sia $A = \{x_1, \dots, x_n\}$ un insieme di n dati e sia f_i la frequenza assoluta di un osservazione x_i in A , possiamo definire frequenza relativa di x_i il valore f_i/n .

Teniamo presente che la somma di tutte le frequenze relative in un campione è sempre uguale ad 1.

Le **frequenze cumulate** si ottengono quando i valori di una variabile possono essere ordinati. 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.

L'ultima frequenza cumulata rappresenta il totale dei casi osservati. inoltre possiamo applicare il concetto di frequenza cumulata sia alle frequenze assolute che a quelle relative, in caso di frequenze relative i valori cumulati variano da 0 a 1.

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

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

dove $I_A : \mathbb{R} \rightarrow 0, 1$ è la funzione indicatrice dell'insieme A , che restituisce 1 se l'argomento appartiene ad A o 0 altrimenti: di conseguenza 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 grafo di questa funzione sarà a tratti costanti.

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

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

Le **frequenze congiunte e marginali**, quando si analizza un insieme di osservazioni, può essere 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 una frazione delle osservazioni, si parla di *frequenza congiunta relativa*.

Se il numero dei possibili valori osservabili per i caratteri non è elevato, possiamo rappresentare visivamente queste frequenze tramite una *tabella delle frequenze congiunte* o *tabella di contingenza*. In questa 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 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.

Grafici

La **simmetria**, un insieme di dati è detto 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 viene detto **centro di simmetria** della distribuzione.

La **quasi simmetria** si presenta quando i dati non sono perfettamente simmetrici, ma la distribuzione rimane quasi speculare rispetto a un punto centrale.

Un modo semplice per rendersi conto se una distribuzione è (quasi) simmetrica consiste nel rappresentarla graficamente e osservare la sua forma.

Grafici per la frequenza

Se l'insieme di dati contiene un numero ridotto di valori distinti allora è rappresentabile con una tabella delle frequenze. Questa tabella associa a ciascun valore distinto osservato la sua frequenza assoluta. La somma di tutte le frequenze deve corrispondere al numero totale di osservazioni. Data una variabile statistica \mathbf{X} che può assumere vari valori, si elencano i valori distinti di \mathbf{X} in una colonna e, a fianco di ognuno, la relativa frequenza di occorrenza nel campione.

Per costruire la 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 affiancati dalla loro frequenza f e dalla loro frequenza relativa f/n , dove n è il numero totale di osservazioni nell'insieme di dati.

Grafici a bastoncini, a barre e poligonal

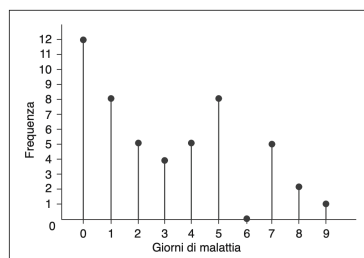


Figura 2.1 Un grafico a bastoncini.

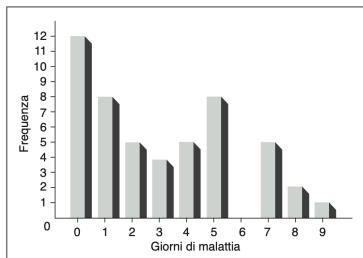


Figura 2.2 Un grafico a barre.

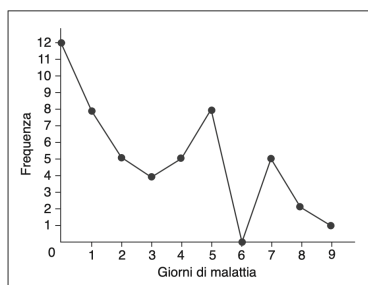


Figura 2.3 Un grafico poligonale.

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.

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.

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.

Diagramma a torta

In caso di dati non sono numerici si utilizza un diagramma a torta, esso consiste in un cerchio suddiviso in settori, uno per ogni valore distinto dei dati. Dato un valore con frequenza relativa f/n , allora l'area del settore corrisponde all'area del cerchio moltiplicata per f/n , ovvero un arco con un angolo di $360 \cdot (f/n)$ gradi.

Diagramma ramo-foglia

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575
26	183, 894, 982
27	671, 711, 744
28	345, 764, 913, 967

Tabella 1: Diagramma a stelo

Un modo efficiente di rappresentare un insieme di dati di dimensioni medie consiste nell'utilizzare il *diagramma ramo-foglia* (o a stelo). Tale grafico si ottiene dividendo ciascun valore dei dati in due parti, chiamati appunto rami e foglie.

La scelta dei rami dovrebbe essere fatta in modo che il diagramma ramo-foglia che ne risulta sia informativo sui dati. Questi diagrammi sono particolarmente adatti a descrivere insiemi di dati dimensioni ridotte.

Questo grafico ha l'aspetto di un istogramma ruotato su un lato, con il vantaggio di contenere tutti i valori dei dati originali in ogni classe. Quando il grafico presenta troppe foglie per ogni riga, si può raddoppiare il numero di rami utilizzando due righe per ogni valore del ramo.

Diagramma di Pareto

I diagrammi di Pareto sono grafici a barre ordinate 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, permettendo di evidenziare quali categorie contribuiscono maggiormente al totale, facilitando l'individuazione delle cause o delle categorie più rilevanti.

Istogrammi e raggruppamento dei dati

Utilizzare i grafici precedenti è un metodo efficace per descrivere un insieme di dati, ma alcuni di questi insiemi hanno troppi valori distinti per poter usare questo metodo. Perciò è necessario suddividere i valori in gruppi, o classi, e poi rappresentare con un grafico il numero di valori dei dati che cadono in ciascuna classe. Il numero di classi scelte è un compromesso tra:

- Scegliere poche classi al costo di perdere molte informazioni sui valori effettivi in una classe.
- Scegliere troppe classi, ottenendo 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, cioè una classe include 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.

Diagramma di dispersione e 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) , $i = 1 \dots n$ 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 nella equazione $y = a + bx$.

Il diagramma di dispersione, oltre a mostrare il comportamento relativo di due variabili e ad aiutarci nelle previsioni, 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 invece siano causati da errori nella raccolta dati.

Capitolo 3 - Statistiche

Una **statistica** è una quantità numerica calcolata a partire da un insieme di dati.

Centralità

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 e tre i casi si parla di campionaria, in quanto vengono effettuate su dei campioni.

Media campionaria

Dato un campione di n dati i cui valori sono x_1, x_2, \dots, x_n . Una statistica per indicare il centro di questo insieme di dati è la media campionaria, cioè la media aritmetica dei valori dati:

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

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

Trasformazioni

La **traslazione** dello stesso insieme di dati, si verifica quando ciascun valore viene incrementato di una costante b , allora anche la media campionaria viene incrementata di b :

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

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

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

La **scalatura** è quando ciascun valore dei dati viene moltiplicato per a , lo è anche la media campionaria:

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

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

La **combinazione** delle due trasformazioni precedentemente illustrate:

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

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

Media pesata

Quindi se abbiamo un insieme di dati organizzato in una tabella delle frequenze, la media campionaria può essere calcolata moltiplicando ciascun valore distinto per la sua frequenza, sommando tutti questi prodotti e poi dividendo il risultato per il numero totale di osservazioni.

In modo formale, supponiamo di avere k valori distinti x_1, x_2, \dots, x_k con frequenze corrispondenti f_1, f_2, \dots, f_k . Il numero totale di osservazioni è dato da: $n = \sum_{i=1}^k f_i$.

La media campionaria \bar{x} è quindi calcolata come:

$$\bar{x} = \frac{x_1 + \dots + x_1 + x_2 + \dots + x_2 + \dots + x_k + \dots + x_k}{n} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n} \quad (3.1)$$

Ora, consideriamo w_1, w_2, \dots, w_k numeri non negativi la cui somma è pari ad 1, allora

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

prende il nome di media pesata dei valori x_1, x_2, \dots, x_k dove w_i è il peso di x_i . Quindi scriviamo l'equazione (3.1) come:

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

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 compare nell'insieme dei dati.

Scarti

Si supponga che l'insieme di dati sia costituito dagli n valori x_1, x_2, \dots, 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, dato che:

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

Questa uguaglianza afferma che la somma degli scarti positivi e la somma degli scarti negativi della media campionaria si controbilanciano.

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

Mediana campionaria

La media campionaria presenta un notevole punto debole come indicatore del centro di un insieme di dati, il suo valore infatti è ampiamente influenzato da eventuali valori fuori scala.

Per calcolare la **mediana campionaria** andiamo a disporre il valore dei dati in ordine crescente. Se il numero di valori è dispari, allora il valore intermedio della lista ordinata è la mediana campionaria, mentre se è pari, la media dei due valori centrali è la mediana campionaria.

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

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

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 prende in considerazione tutti i valori dell'insieme di dati, mentre la mediana campionaria considera solo uno o due valori centrali e quindi non è influenzata dai valori fuori scala.

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

Percentili campionari

La mediana campionaria è un caso particolare di una statistica nota come 100p-esimo percentile campionario, dove p indica un qualunque numero \mathbb{R} dell'intervallo $[0,1]$.

Per calcolare il percentile è necessario definire un ordinamento sulle osservazioni.

il **100p-esimo percentile campionario** è un valore maggiore o uguale di almeno 100p percento dei valori dati, e minore o uguale di almeno 100(1-p) percento dei valori dati. Se due valori dei dati soddisfano questa condizione, allora il 100p-esimo percentile campionario è la media aritmetica di essi.

la mediana campionaria è il 50-esimo percentile, ossia è il percentile campionario 100p quando $p = 0.5$

Supponiamo che i dati di un campione di cardinalità n siano disposti in ordine crescente. Per determinare il 100p-esimo percentile campionario bisogna determinare quale valore sia:

- maggiore o uguale di almeno np valori dei dati
- minore o uguale di almeno $n(p-1)$ valori dei dati

Se np non è un intero, il solo valore dei dati che soddisfa questi requisiti è quello la cui posizione è il più piccolo intero maggiore di np . Se invece np è un intero, allora sia il valore in posizione np che il valore in $np+1$ soddisfano i due requisiti, e quindi il 100p-esimo percentile campionario è la media dei due valori.

Calcolo del 100p-esimo percentile campionario di un insieme di dati di n elementi:

1. Si dispongono i dati in ordine crescente
2. Se np non è un intero, si determina il più piccolo intero maggiore di np . Il valore dei dati in questa posizione è il 100p-esimo percentile campionario.
3. Se np è un intero, allora la media dei valori nelle posizioni np e $np+1$ è il 100p-esimo percentile campionario.

il valore p prende il nome di *quantile di livello*, e a seconda dei valori che può assumere si ottengono statistiche diverse. In particolare si definiscono:

- Decili: i percentili multipli di 10, che dividono il campione in 10 parti uguali.
- Quartili: i percentili multipli di 25, che dividono il campione in 4 parti uguali.

il 25-esimo percentile campionario si chiama *primo quartile*, il 50-esimo percentile campionario si chiama *mediana* o *secondo quartile*, e il 75-esimo percentile campionario si chiama *terzo quartile*.

i quartili suddividono i dati in quattro parti in modo tale che il 25% dei dati sia inferiore del primo quartile, il 25% compreso tra il primo e il secondo, il 25% tra il secondo e il terzo e il restante 25% sia maggiore del terzo quartile.

Moda campionaria

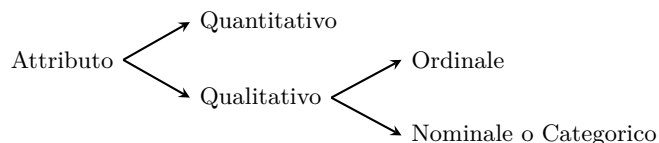
Un ulteriore indicatore della tendenza centrale è la moda campionaria, che è il valore che si verifica con maggiore frequenza nell'insieme di dati.

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.

Tramite una tabella delle frequenze riusciamo a dedurre facilmente questi valori, sono quelli con la maggiore frequenza.

Riepilogo

Si considerino le varie classificazioni degli attributi:



La media si può fare solo per gli attributi quantitativi, la mediana e i percentili si possono svolgere anche sugli attributi qualitativi ordinali in caso di cardinalità del campione dispari, se la cardinalità è pari sarà necessaria la media e quindi possibili solo su attributi quantitativi, la moda si può fare per qualsiasi tipo di attributo.

Dispersione

Due campioni A e B possono presentare la stessa centralità ma essere molto diversi tra loro, per esempio:

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

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

Un modo per misurare la dispersione dei dati è quello di considerare gli scarti dei valori dei dati rispetto ad un valore centrale. Il valore centrale più usato per questo scopo è la media campionaria. Se i valori dei dati sono x_1, \dots, x_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 $i = 1, \dots, n$.

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

Per fare in modo che non valga 0, andiamo a considerare i singoli scarti indipendentemente dal segno, questo possiamo farlo considerando il valore assoluto degli scarti oppure il quadrato.

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, \dots, x_n di media $\bar{x} = (\sum_{i=1}^n x_i) / n$ è definita come

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

La formula algebrica che segue è utile per calcolare la varianza campionaria a mano:

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

Trasformazioni

La **traslazione** cioè quando a ciascun valore dei dati viene sommata una costante b per ottenere un nuovo insieme di dati, la varianza campionaria non cambia:

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

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 alla formula (3.2) per semplificare il calcolo della varianza campionaria.*

La **scalatura** cioè quando ciascun valore dei dati viene moltiplicato per a , la varianza campionaria viene moltiplicata per il quadrato di a :

$$y_i = ax_i \text{ per } i = 1, \dots, n \Rightarrow s_y^2 = a^2 s_x^2$$

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

La **combinazione** consiste nella combinazione delle due trasformazioni precedenti:

$$y_i = ax_i + b \text{ per } i = 1, \dots, n \Rightarrow s_y^2 = a^2 s_x^2$$

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.

Trasformazioni

La **traslazione** cioè la somma di una costante b a ciascuno dei valori x_1, \dots, x_n per ottenere un nuovo insieme di dati, la deviazione standard campionaria non cambia:

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

La **scalatura** cioè quando 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 :

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

La **combinazione** consiste nella combinazione delle due trasformazioni precedenti:

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

La varianza campionaria e la deviazione standard campionaria sono due indici di dispersione che derivano dalla media campionaria.

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**.

Scarto interquantile

Lo scarto interquantile è un indice di dispersione che deriva dalla mediana campionaria, e rappresenta la lunghezza dell'intervallo in cui si trova la metà centrale dei dati. Richiamando i quartili, possiamo dire che si tratta della lunghezza dell'intervallo compreso tra il primo quartile $Q1$ e il terzo quartile $Q3$.

Un IQR piccolo indica che la metà centrale dei dati è relativamente concentrata attorno alla mediana, mentre un IQR ampio suggerisce una maggiore dispersione nella parte centrale della distribuzione.

L'IQR è un indice robusto poichè non è influenzato da valori fuori scala, come la mediana campionaria. Questo lo rende particolarmente utile quando la distribuzione dei dati è asimmetrica o contiene anomalie.

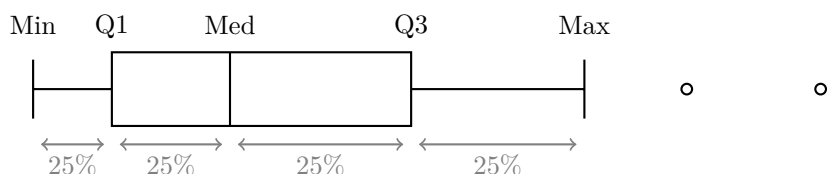
Questo indice è fondamentale per la costruzione dei boxplot perchè viene utilizzato per definire quali valori siano fuori scala e quali no. Generalmente i valori inferiori a $Q1 - 1.5 \cdot IQR$ o superiori a $Q3 + 1.5 \cdot IQR$ sono considerati outlier.

Altri grafici

Box Plot

Se vogliamo visualizzare alcune statistiche riassuntive di un insieme di dati usiamo un **box plot** (diagramma a scatola). Per realizzarlo tracciamo un segmento orizzontale dal minore al maggiore dei dati. A questo segmento sovrapponiamo un rettangolo che si estende dal primo al terzo quartile. Il rettangolo è diviso in due parti da un segmento verticale in corrispondenza della mediana campionaria.

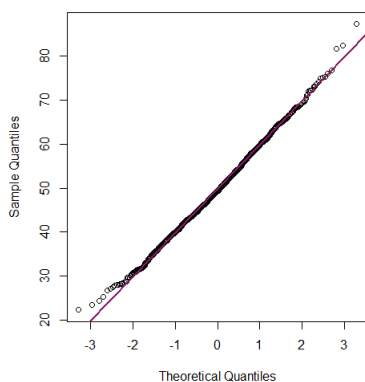
La lunghezza della base del rettangolo corrisponde allo scarto interquantile.



In un box plot ciascuno dei quattro segmenti contiene il 25% delle osservazioni, però la lunghezza di ciascun segmento sulla scala orizzontale dipende dalla distribuzione dei valori. Se i dati sono più concentrati in un certo intervallo, quel tratto sarà più corto. I quartili dividono le osservazioni in parti uguali sul numero dei dati ma non sulla distanza numerica.

I pallini a destra del box plot rappresentano dei valori fuori scala, determinati tramite l'utilizzo dell'IQR.

Q-Q Plot



Un diagramma Q-Q, o diagramma quantile-quantile, è una rappresentazione grafica qualitativa che permette di verificare le similarità tra le distribuzioni di due campioni diversi. Può tornare utile per vedere quindi se seguono una stessa distribuzione.

Questi diagrammi si basano sul fatto che i quantili campionari rappresentino l'approssimazione di quantili teorici, se considerati tutti insieme, individuano la distribuzione dei dati.

Ogni asse cartesiano di questo diagramma contiene i quantili dei due campioni presi in considerazione. Poichè i quantili sono ordinati in modo crescente anche il grafico lo sarà, o perlomeno non decrescente.

Se due campioni hanno una distribuzione uguale, allora estraendo da entrambi il quantile di un livello fissato si dovranno ottenere due numeri vicini. In questo caso i punti del diagramma Q-Q tenderanno ad allinearsi alla bisettrice del I° e III° quadrante.

Distribuzioni normali

Un insieme di dati si dice **normale** se il rispettivo istogramma ha le seguenti proprietà:

- L'istogramma è simmetrico rispetto all'intervallo centrale.
- Ha il punto massimo in corrispondenza dell'intervallo centrale.
- Spostandoci dal centro verso destra o sinistra, l'altezza diminuisce in modo tale che l'intero istogramma è a forma di campana.

Se l'istogramma di un insieme di dati è vicino ad essere un istogramma normale, allora diciamo che l'insieme di dati è approssimativamente normale. Inoltre l'insieme di dati si dice asimmetrico a destra o a sinistra a seconda di quale sia la coda più lunga.

A causa della simmetria dell'istogramma normale, la media e la mediana di un insieme di dati approssimativamente normale sono uguali o molto prossime.

Siano \bar{x} e s rispettivamente la media e la deviazione standard campionarie di un insieme di dati approssimativamente normale. La **regola empirica** specifica le proporzioni approssimate delle osservazioni che si trovano a una distanza di s , $2s$ e $3s$ da \bar{x} :

- circa il 68% delle osservazioni rientrano nell'intervallo $\bar{x} \pm s$
- circa il 95% delle osservazioni rientrano nell'intervallo $\bar{x} \pm 2s$
- circa il 99,7% delle osservazioni rientrano nell'intervallo $\bar{x} \pm 3s$

Un insieme di dati ottenuto campionando una popolazione costituita da sottogruppi eterogenei non è di solito normale. L'istogramma di un insieme di dati di questo tipo spesso assomiglia ad una sovrapposizione di istogrammi normali e quindi spesso ha due o più massimi locali. Questi massimi locali si comportano come mode. Un insieme di dati il cui istogramma ha due massimi locali si dice **bimodale**.

In questi casi, quando nei dati si hanno due popolazioni ben distinte per quanto riguarda un certo attributo, ha senso dividere i dati in base a queste popolazioni e ottenere un insieme normale.

Indici di dipendenza

Consideriamo un insieme composto da dati accoppiati $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Per vedere la relazione relativa di queste due variabili è possibile rappresentarle in un diagramma di dispersione. Questo approccio è però qualitativo e quindi soggetto a interpretazione. Vogliamo trovare un indice quantitativo in grado di rappresentare questa relazione oggettivamente. Questi indici sono detti di dipendenza o associazione e misurano la forza della relazione, ossia forniscono un valore numerico che indica quanto intensamente le variabili siano collegate.

Covarianza campionaria

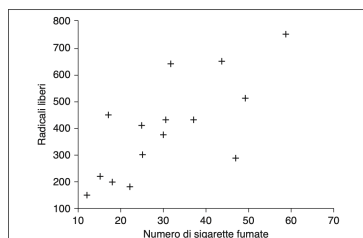
La **covarianza campionaria** è una statistica che quantifica in che misura grandi valori di x corrispondano a grandi valori di y e piccoli valori di x a piccoli valori di y . Questo indice quindi misura la tendenza con cui due variabili si muovono insieme ed è definita come la media dei prodotti degli scostamenti delle variabili dalle loro medie.

Relazione tendenziale

Procediamo considerando una **relazione** di tipo **tendenziale** e non deterministico. Ciò significa che le affermazioni che seguiranno varranno tendenzialmente sempre: ci saranno quindi delle eccezioni ma per lo più saranno valide.

Si supponga che un insieme sia composto dalle coppie di valori (x_i, y_i) con $i = 1, \dots, n$ e calcolando le rispettive medie campionarie \bar{x} e \bar{y} . Per la i -esima coppia di dati, consideriamo $(x_i - \bar{x})$ e $(y_i - \bar{y})$ cioè gli scarti dei valori x, y rispetto alla loro media campionaria.

Possiamo notare che quando grandi valori di x tendono ad essere associati a grandi valori di y , e piccoli valori di x tendono ad essere associati a piccoli valori di y , allora i segni (positivi o negativi) di $(x_i - \bar{x})$ e $(y_i - \bar{y})$ tenderanno a essere gli stessi. Quindi se gli scarti hanno segno concorde, il loro prodotto $(x_i - \bar{x})(y_i - \bar{y})$ sarà positivo. Otteniamo che la sommatoria $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ tenderà ad essere un grande numero positivo.



Sigarette fumate rispetto al numero di radicali liberi.

$$\begin{array}{ll} x \text{ "grande"} & e \quad y \text{ "grande"} \\ x \geq \bar{x} & y \geq \bar{y} \\ (x_i - \bar{x}) \geq 0 & (y_i - \bar{y}) \geq 0 \end{array}$$

Tendenzialmente:

$$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$$

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

$$\begin{array}{ll} x \text{ "piccolo"} & e \quad y \text{ "piccolo"} \\ x < \bar{x} & y < \bar{y} \\ (x_i - \bar{x}) < 0 & (y_i - \bar{y}) < 0 \end{array}$$

Tendenzialmente:

$$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$$

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

Si individua quindi una correlazione positiva tra le due variabili poiché tendenzialmente presentano segno concorde. In questo caso si parla di relazione tra le due variabili di tipo diretta.

Per lo stesso motivo, quando grandi valori di una variabile tendono a verificarsi in corrispondenza a piccoli valori dell'altra, allora i segni di $(x_i - \bar{x})$ e $(y_i - \bar{y})$ saranno discordi e quindi la sommatoria $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ tenderà ad essere un grande numero negativo.

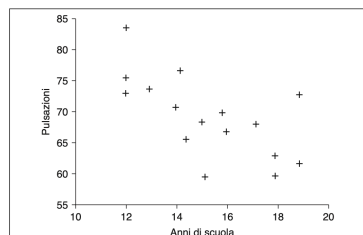


Diagramma di dispersione degli anni di scuola e delle pulsazioni.

$$\begin{array}{ll} x \text{ "grande"} & e \quad y \text{ "piccola"} \\ x \geq \bar{x} & y < \bar{y} \\ (x_i - \bar{x}) \geq 0 & (y_i - \bar{y}) < 0 \end{array}$$

Tendenzialmente:

$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$

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

$$\begin{array}{ll} x \text{ "piccolo"} & e \quad y \text{ "grande"} \\ x < \bar{x} & y \geq \bar{y} \\ (x_i - \bar{x}) < 0 & (y_i - \bar{y}) \geq 0 \end{array}$$

Tendenzialmente:

$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$

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

Si individua quindi una correlazione negativa tra le due variabili poiché tendenzialmente presentano segno discorde. In questo caso si parla di relazione tra le due variabili di tipo indiretta.

Andiamo ad standardizzare la sommatoria dividendo per $n-1$, al fine di evitare che questo indice assuma valori troppo elevati. Possiamo osservare che la formula della covarianza campionaria è riconducibile a quella della varianza campionaria, motivo per il quale si possa intuire perché si vada a dividere per $n-1$ e non direttamente per il numero totale di osservazioni (*correzione sottostima*).

Infine possiamo definire la covarianza campionaria come:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \begin{cases} > 0 & \text{relazione diretta} \\ \approx 0 & \text{assenza di relazione / indipendenza} \\ < 0 & \text{relazione indiretta / inversa} \end{cases}$$

Coefficiente di correlazione di Pearson

La covarianza campionaria non può essere posizionata all'interno di una scala assoluta in quanto non è normalizzata e il suo valore dipende dalle osservazioni coinvolte. Si ricava perciò da questo indice il **coefficiente di correlazione lineare campionaria**, anche detto indice di correlazione di Pearson, che si indica con ρ .

Possiamo standardizzare il valore della covarianza campionaria dividendolo per il prodotto delle due deviazioni standard campionarie delle due variabili:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Il coefficiente di correlazione di Pearson è quindi un numero puro e, proprio come la covarianza campionaria, quando $\rho > 0$ i dati sono correlati positivamente, mentre quando $\rho < 0$ sono correlati negativamente. Non dipendendo dalle unità di misura, questo indice può essere usato per comparare dataset diversi.

Una proprietà importante è che vale $-1 \leq \rho \leq 1$.

Relazione deterministica

Come **primo caso** vogliamo passare da una relazione tendenziale a una deterministica, nella quale la variabile y è una trasformazione lineare della variabile x ; tutti i vari indici statistici variano di conseguenza:

$$\forall i \quad y_i = a + bx_i \quad \Rightarrow \quad \bar{y} = a + b\bar{x} \quad \Rightarrow \quad s_y^2 = b^2 s_x^2 \quad \Rightarrow \quad s_y = |b| s_x$$

Nella relazione deterministica $y = a + bx$, la costante b rappresenta la pendenza della retta che lega le due variabili e indica di quanto varia y all'aumentare di x . possiamo aspettarci:

- Se b è positivo, all'aumento di x corrisponde un incremento di $y \Rightarrow$ relazione diretta
- Se b è negativo, all'aumento di x corrisponde un decremento di $y \Rightarrow$ relazione inversa

Si calcoli ora il coefficiente di correlazione di Pearson:

$$\rho = \frac{b \sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)|b|s_x^2} = \frac{b}{|b|} \cdot \frac{1}{s_x^2} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{b}{|b|} \cdot \frac{1}{s_x^2} s_x^2 = \frac{b}{|b|} = \begin{cases} +1 & \text{se } b > 0 \\ -1 & \text{se } b < 0 \end{cases}$$

Questo significa che:

- l'indice ρ è uguale a $+1$ se b è una costante positiva, e se quindi le due variabili esibiscono una relazione di tipo deterministica diretta.
- l'indice ρ è uguale a -1 se b è una costante negativa, e se quindi le due variabili esibiscono una relazione di tipo deterministica indiretta

Le conclusioni ottenute con i calcoli rispecchiano le nostre attese iniziali.

Come **secondo caso** consideriamo una relazione in cui entrambi le variabili x e y sono soggette ad una trasformazione lineare, i vari indici statistici variano nel seguente modo:

$$\begin{aligned} \forall i \quad x'_i = a + bx_i &\Rightarrow \bar{x}' = a + b\bar{x} &\Rightarrow s_{x'} = |b| s_x &\Rightarrow x'_i - \bar{x}' = b(x_i - \bar{x}) \\ y'_i = c + dy_i &\Rightarrow \bar{y}' = c + d\bar{y} &\Rightarrow s_{y'} = |d| s_y &\Rightarrow y'_i - \bar{y}' = d(y_i - \bar{y}) \end{aligned}$$

Procediamo a calcolare il coefficiente di correlazione di Pearson:

$$\rho' = \frac{\sum_{i=1}^n (x'_i - \bar{x}')(y'_i - \bar{y}')}{(n-1)s_{x'} s_{y'}} = \frac{bd \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)|b||d|s_x s_y} = \frac{bd}{|b||d|} \rho = \begin{cases} +\rho & \text{se } b \text{ concorda con } d \\ -\rho & \text{se } b \text{ discorda con } d \end{cases}$$

Ciò significa che la correlazione tra x' e y' rimane numericamente invariata rispetto a quella tra x e y e può eventualmente cambiare solo di segno:

- se i coefficienti di trasformazione b e d hanno lo stesso segno allora $\rho' = \rho$
- se i coefficienti di trasformazione b e d hanno segni diversi allora $\rho' = -\rho$

Conclusioni

Il coefficiente di correlazione di Pearson è un indicatore fondamentale per valutare la forza e la direzione di una relazione, o associazione, di tipo lineare tra due variabili, con valori che spaziano fra -1 e $+1$.

Relazione deterministica Quando due variabili presentano una relazione lineare deterministica $y = a + bx$, il coefficiente di correlazione assume valore estremo: $\rho = +1$ se $b > 0$ e $\rho = -1$ se $b < 0$, in altre parole se tutti i punti che giacciono esattamente su una retta crescente o decrescente, la correlazione è massima o minima.

Relazione tendenziale Nella maggior parte dei casi reali, le due variabili seguono una relazione lineare tendenziale. In questo contesto, il valore assoluto del coefficiente di correlazione $|\rho|$, fornisce una misura di quanto le osservazioni si dispongano in prossimità di una retta:

- $|\rho| = 1$ evidenzia una perfetta relazione lineare, cioè è possibile collegare tutti i valori (x_i, y_i) con $i = 1, \dots, n$ con una retta.
- Più $|\rho|$ si avvicina a 1 e più i dati esibiscono una relazione lineare forte, anche se non perfetta: ciò significa che se anche non esiste una retta che attraversa tutti i valori dei dati, ce n'è una che passa vicino a tutti.
- Se $|\rho|$ è prossimo allo 0, non c'è evidenza di un legame lineare tra le variabili.

Il segno di ρ indica invece la direzione della relazione, se è positivo allora l'approssimazione lineare è crescente (diretta) mentre è negativo quando l'approssimazione lineare è decrescente (inversa).

È importante tenere a mente che un valore di $\rho = 0$ non implica automaticamente l'assenza di qualsiasi relazione, poichè potrebbero esistere legami non lineari che questo indice non è in grado di cogliere.

Vale inoltre la pena sottolineare che il coefficiente di correlazione di Pearson non implica in alcun modo un rapporto causa-effetto tra le due variabili prese in considerazione, cioè due variabili possono presentare un valore di correlazione elevato senza che una determini o causi l'altra. Spesso infatti può intervenire un terzo fatto, o più fattori, a influenzare entrambe le variabili, generando un legame che in realtà non corrisponde a un meccanismo causale diretto.

Indici di eterogeneità

Sappiamo che per le variabili qualitative nominali non è possibile calcolare la varianza nè gli indici che ne derivano, perchè non esistono una media, una mediana o altri valori numerici di riferimento su cui misurare le distanze. Ma è necessario avere un indice che misuri la dispersione della distribuzione delle frequenze, detta **eterogeneità**. In particolare si dice che una variabile è distribuita in modo eterogeneo quando ogni suo valore compare con la stessa frequenza.

Indice di Gini

Consideriamo un campione x_1, \dots, x_n in cui occorrono i valori distinti v_1, \dots, v_m e indichiamo con f_j la frequenza relativa dell'elemento v_j per $j = 1, \dots, m$. Definiamo l'**indice di eterogeneità di Gini** come:

$$I = 1 - \sum_{j=1}^m f_j^2$$

Una proprietà importante di questo indice è che vale $0 \leq I < 1$ ed inoltre l'omogeneità massima dell'insieme di dati si presenta quando $I = 0$, mentre l'eterogeneità massima quando $I = 1$. Di conseguenza, all'aumentare dell'indice di Gini aumenta l'eterogeneità.

Per dimostrare le limitazioni inferiori e superiori, ricordiamo che trattandosi di frequenze relative, vale $0 \leq f_j \leq 1 \quad \forall j \in [1, m]$ ed inoltre vale $\sum_{j=1}^m f_j = 1$. Di conseguenza si avrà:

- per almeno un j si ha $f_j > 0 \Rightarrow f_j^2 > 0 \Rightarrow \sum_{j=1}^m f_j^2 > 0 \Rightarrow I < 1$

- per ogni j , dato che $0 \leq f_j \leq 1$, si ha che $f_j^2 \leq f_j \Rightarrow \sum_{j=1}^m f_j^2 \leq \sum_{j=1}^m f_j = 1 \Rightarrow I \geq 0$

Notiamo che l'estremo inferiore si presenta quando l'insieme è massimamente omogeneo e l'estremo superiore quando è massimamente eterogeneo, quindi:

- l'eterogeneità minima la si ha quando tutti gli elementi hanno lo stesso valore, quindi $\exists k \in [1, m] \mid f_k = 1, \forall j \neq k \ f_j = 0 \Rightarrow I = 1 - \sum_{j=1}^m f_j^2 = 1 - f_k^2 = 1 - 1 = 0$
- l'eterogeneità massima la si ha quando tutti gli elementi hanno la stessa frequenza, quindi $\forall j \in [1, m] \ f_j = \frac{1}{m} \Rightarrow I = 1 - \sum_{j=1}^m f_j^2 = 1 - \sum_{j=1}^m \frac{1}{m^2} = 1 - \frac{1}{m} = \frac{m-1}{m} \rightarrow 1$ al crescere di m

Indice di Gini normalizzato

Ricordiamo che m è la cardinalità dell'insieme dei valori distinti. Questo indice presenta due problematiche:

- il valore massimo che può assumere, ossia quando l'insieme è massimamente eterogeneo, è $(m-1)/m$. Di conseguenza, specialmente nel caso in cui non si conosca il valore m , non si può sapere quanto questo indice debba tendere a 1 affinché si abbia la massima eterogeneità nell'insieme dei dati.
- il suo valore dipende fortemente dal valore di m . Non è quindi possibile confrontare due attributi qualitativi che presentano intervalli di valori diversi, ovvero m diverso.

Per eliminare questi problemi introduciamo l'**indice di Gini normalizzato**, che si ottiene dividendo l'indice di Gini per il valore massimo $(m-1)/m$ che può assumere:

$$I' = \frac{m \cdot I}{m-1}$$

Questo indice può assumere anche 1 come valore: $0 \leq I' \leq 1$. Consideriamo infatti il caso in cui l'eterogeneità dell'insieme è massima:

$$I = \frac{m-1}{m} \Rightarrow I' = \frac{m \cdot I}{m-1} = \frac{m \cdot (m-1)}{(m-1) \cdot m} = 1$$

Indice di entropia

Consideriamo un campione x_1, \dots, x_n in cui occorrono i valori distinti v_1, \dots, v_m e indichiamo con f_j la frequenza relativa dell'elemento v_j per $j = 1, \dots, m$. Possiamo definire l'**indice di entropia** del campione come:

$$H = \sum_{j=1}^m f_j \log \frac{1}{f_j} = - \sum_{j=1}^m f_j \log f_j$$

La funzione $I(p) = \log 1/p = -\log p$ è detta *autoinformazione* e misura la quantità di informazione ottenuta dal verificarsi di un evento con probabilità p . In altre parole, misura quanto viene ridotta l'incertezza una volta che sappiamo che l'evento si è effettivamente realizzato. Questa funzione è monotona decrescente, vale 0 quando $p = 1$ e tende a infinito per p tendente a 0.

Nel calcolo dell'entropia compare $-f_j \log f_j$. Se $f_j = 0$ questa espressione assume la forma indeterminata $0 \cdot \infty$. Però possiamo estendere la definizione di entropia anche nei casi in cui alcune frequenze relative siano nulle. Valutando il limite $\lim_{f_j \rightarrow 0^+} -f_j \log f_j = 0$ si definisce per convenzione $-0 \log 0 = 0$.

Effettuiamo le seguenti osservazioni:

- $\forall j$ vale $-f_j \log f_j \geq 0 \Rightarrow H \geq 0$
- $\forall j$ si ha che $-f_j \log f_j = 0 \Leftrightarrow f_j = 0 \vee \log f_j = 0$ e quindi $f_j = 1$. Pertanto $H = 0$ se e solo se ci si trova in condizione di massima omogeneità, e quindi tutti i dati del campione assumono lo stesso valore.

- in caso di invece massima eterogeneità si avrà $f_j = 1/m$ e quindi

$$H = \sum_{j=1}^m \frac{1}{m} \log m = m \left(\frac{1}{m} \log m \right) = \log m$$

Si può dimostrare che in questo caso l'entropia assume il valore massimo.

Una proprietà importante di questo indice è che vale $0 \leq H \leq \log m$. Più questo indice cresce più aumenta il grado di eterogeneità dell'insieme e vale il viceversa più decresce più aumenta il grado di omogeneità.

Indice di entropia normalizzato

Definiamo l'**indice di entropia normalizzato** come:

$$H' = \frac{H}{\log m}$$

I valori di questo indice sono compresi tra 0 e 1, infatti nel caso di massima eterogeneità si ha che:

$$H = \log m \quad \Rightarrow \quad H' = \frac{\log m}{\log m} = 1$$

Nel nostro caso è utile avere il logaritmo in base 2 in modo da poter misurare l'entropia di bit, che risulta utile quando bisogna svolgere i calcoli in un computer, ma volendo si può utilizzare una qualsiasi base, come il logaritmo naturale o in base 10.

Formulario

Frequenza cumulata:

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