Statistica e analisi dei dati

Kevin Muka

a.a 2024-2025

Indice

Lezioni

L01 - 25/02/2025

Dispense L01-Introduzione_a_python

Introduzione a Python: panoramica sul linguaggio Python e le sue applicazioni.

Tipi di dati e operatori:

- dati semplici e operatori: numeri, stringhe e booleani; operazioni aritmetiche e logiche.
- dati strutturati: liste, tuple, dizionari e insiemi; metodi e funzioni utili per manipolare le liste (e altri tipi strutturati).

Operazioni sulle liste: creazione e manipolazione delle liste; uso di metodi e operatori specifici per le liste (*list comprehension*); esempi pratici e funzioni integrate.
Import e utilizzo di librerie per specifiche funzionalità: `numpy` per operazioni numeriche; `pandas` per la gestione di dati strutturati; `matplotlib.pyplot` per la visualizzazione grafica; `csv` e `collections` per la manipolazione dei dati.

L02 - 27/02/2025

Dispense L02-Pandas

Introduzione alla libreria `Pandas` e ai dataset: caricamento del file `heroes.csv`.

- Serie: creazione e manipolazione di serie; gestione degli indici e accesso tramite slicing, `loc` e `iloc`; calcolo delle frequenze con `value_counts`; operazioni matematiche e uso del metodo `apply`.
- Visualizzazione grafica: creazione di grafici a barre per rappresentare le frequenze; personalizzazione dell'asse delle ascisse (utilizzo di `numpy.arange` e `xticks`).
- DataFrame: creazione di DataFrame da file CSV tramite `read_csv`; accesso a righe e colonne (utilizzo di `loc`, `iloc`, `at` e `iat`); ordinamento dei dati con `sort_values` e `sort_index`; filtraggio e selezione di sottoinsiemi di dati

Librerie utilizzate: `pandas`, `matplotlib.pyplot`, `numpy`, `csv`.

L03 - 04/03/2025

Dispense L03-Dati_e_frequenze - Capitolo 2 RS

Dati e Frequenze: concetti di base; introduzione alla distinzione tra dati quantitativi e qualitativi.
Classificazione dei dati: dati qualitativi (categorie e modalità); dati quantitativi (tipologie e suddivisione).
Frequenze: calcolo e visualizzazione delle frequenze assolute e relative. Frequenze cumulate: definizione e applicazioni. Frequenze congiunte e marginali. Diagrammi:

- grafici a barre, istogrammi e diagrammi a torta.
- diagrammi di Pareto: ordinamento delle frequenze in ordine decrescente; identificazione della regola dell'80/20 per evidenziare le componenti più rilevanti.
- diagrammi stelo-foglia: rappresentazione della distribuzione dei dati quantitativi; visualizzazione della forma e concentrazione dei valori.

Suggerimenti e tecniche per la generazione dei grafici con l'uso di librerie come `matplotlib` e `pandas`.

L04 - 06/03/2025

Dispense L03-Dati_e_frequenze - Capitolo 3 RS

L05 - 11/03/2025

Dispense L03-Dati_e_frequenze: diagrammi a stelo
Dispense L04-Indici_di_dispersione - Capitolo 3 RS

L06 - 13/03/2025

Dispense L05-Indici_di_eterogeneita

Indici di eterogeneità: indice di Gini, indice di entropia. Alberi di decisione e "machine learning" tramite questi indici.

1 Introduzione alla statistica

1.1 Definizione

Statistica 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.

A volte un'analisi statistica comincia con un insieme di dati prestabilito, in questo caso la statistica si usa per descrivere, riassumere e analizzare i dati. In altre situazioni i dati non sono ancora disponibili e si può usare la statistica per progettare un esperimento che li generi. Se ne occupa la statistica descrittiva.

Statistica descrittiva La statistica descrittiva è la parte della statistica che descrive e riassume i dati.

Una volta che i dati sono stati raccolti e descritti, si vogliono trarre delle conclusioni. Se ne occupa la statistica inferenziale.

Statistica inferenziale La statistica inferenziale è la parte della statistica che trae conclusioni sui dati.

La statistica inferenziale si basa sul modello probabilistico che consiste nel fare un insieme di assunzioni sulle probabilità di ottenere un certo valore. La statistica inferenziale, quindi, 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.

1.2 Popolazioni e campioni

Nella statistica è cruciale ottenere delle informazioni su tutto un insieme di elementi, che viene definito 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.

Popolazione Si definisce popolazione l'insieme di tutti gli elementi di interesse.
Campione Si definisce campione un sottoinsieme della popolazione, utile quando questa è troppo numerosa.

Affinché il campione ci dia informazioni su tutta la popolazione, esso deve essere rappresentativo di tutta la popolazione. Con rappresentativo si intende 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.*

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

1.2.1 Campionamento 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 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.

Parte I

Statistica descrittiva

2 Descrivere insiemi di dati

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

- Dati binari o booleani: l'osservazione può assumere soltanto due valori tra loro non confrontabili. Si utilizza "booleani" per evidenziare la presenza o 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 solo stabilire una relazione di equivalenza.
- Dati ordinali: esiste una relazione d'ordine tra i valori osservabili, e quindi se due valori sono diversi è possibile stabilire quale sia il più piccolo e quale il più grande.

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

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

2.2 Frequenze

Frequenza assoluta La frequenza assoluta di un'osservazione x in un insieme di dati $A = \{x_1, \cdots, x_n\}$ è definita come il numero di volte in cui x compare in A.
Formalmente, se si indica con f_x la frequenza assoluta di x, si ha che $f_x = \#\{j \in \{1, \cdots, n\} \,|\, x_j = x\}$

Frequenza relativa La frequenza relativa consente di esprimere la presenza di ogni valore in termini di proporzione rispetto all'intero campione. Sia $A = \{x_1, \cdots, x_n\}$ un insieme di n dati e sia f_i la frequenza assoluta di un'osservazione x_i in A, si definisce la frequenza relativa di x_i il valore f_i/n. Si osserva che la somma di tutte le frequenze relative in un campione è sempre pari ad 1.

Frequenze cumulate

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.

È importante notare che l'ultima frequenza cumulata rappresenta il totale dei casi osservati. Inoltre, il concetto di frequenza cumulata si applica sia alle frequenze assolute che a quelle relative: nel caso delle frequenze relative, i valori cumulati 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 $x_1, \ldots, x_n$, la funzione cumulativa empirica $\hat{F} : \mathbb{R} \to [0, 1]$ 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: 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 grafico di questa funzione sarà a tratti costanti.

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

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.

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.

2.3 Grafici

2.3.1 Simmetria

Simmetria 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.

Quasi simmetria 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 e osservarne la forma.

2.3.2 Grafici per la frequenza

Se l'insieme di dati contiene un numero ridotto di valori distinti, lo si può rappresentrare 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 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.

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

2.3.3 Grafici a bastoncini, a barre e poligonali

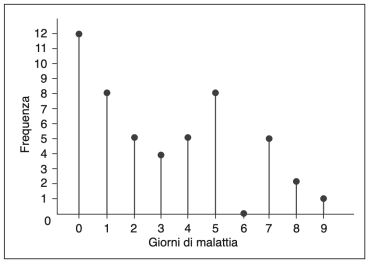


Figura 2.1 Un grafico a bastoncini.

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.

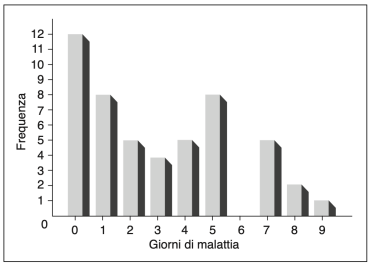


Figura 2.2 Un grafico a barre.

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.

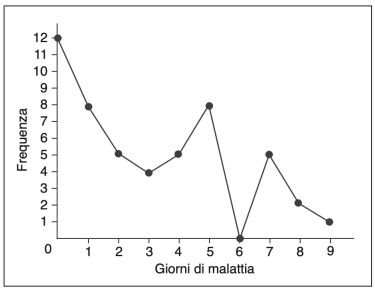


Figura 2.3 Un grafico poligonale.

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.

2.3.4 Diagramma ramo-foglia

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 557
26	183, 894, 982
27	671, 711, 744
28	345, 764, 913, 967

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.

Fisicamente, 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.

2.3.5 Diagramma a torta

Se i dati non sono numerici si utilizza un diagramma a torta. Si costruisce usando 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.

2.3.6 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.

2.3.7 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.

Raggruppamento dei dati In questi casi si suddividono i valori in gruppi, o classi, e poi si rappresenta con un grafico il numero di valori dei 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.

2.3.8 Diagramma di dispersione

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.

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 siano invece causati da errori nella raccolta dei dati.

---

3 Statistiche

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

3.1 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 i tre i casi si parla di campionaria, in quanto sono effettuate su dei campioni.

3.1.1 Media campionaria

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.

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

Scalatura Se invece ciascun valore 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.

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 compare 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.

Scarti

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

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.

3.1.3 Percentili campionari

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

Per poter calcolare il percentile si deve poter definire un ordinamento sulle osservazioni.

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 posizione $np + 1$ soddisfano i due requisti, 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*. 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% sia compreso tra il primo e il secondo quartile, il 25% tra il secondo e il terzo quartile e il restante 25% sia maggiore del terzo quartile.

3.1.4 Moda campionaria

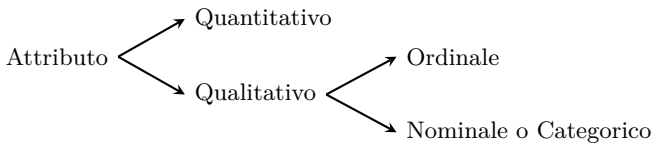
Un altro indicatore della tendenza centrale è la moda campionaria, che è il valore che si verifica con maggiore frequenza nell'insieme di dati.

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 frequenze; 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 e i percentili si possono svolgere anche sugli attributi qualitativi ordinali con cardinalità del campione dispari; la moda si può fare per qualsiasi tipo di attributo.

3.2 Dispersione

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:

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

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.

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:

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

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**.

3.2.3 Scarto interquartile

Lo scarto interquartile 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 concentrato attorno alla mediana, mentre un IQR ampio suggerisce una maggiore dispersione nella parte centrale della distribuzione.

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

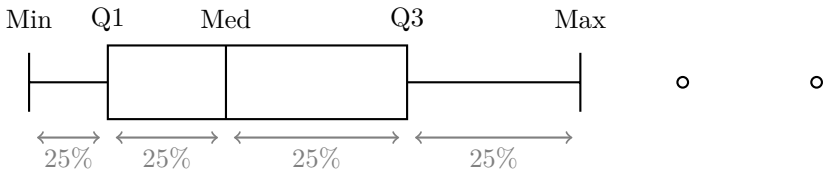
Questo indice è fondamentale per la costruzione dei boxplot perché viene proprio utilizzato per definire quali valori siano fuori scala e quali no. Generalmente, i valori inferiori a $Q1 - 1.5 \cdot \text{IQR}$ o superiori a $Q3 + 1.5 \cdot \text{IQR}$ sono considerati outlier.

3.3 Altri grafici

3.3.1 Box Plot

Per visualizzare alcune statistiche riassuntive di un insieme di dati si usa un *box plot* (diagramma a scatola). Per realizzarlo tracciamo un segmento orizzontale dal minore al maggiore dei dati. Al 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.

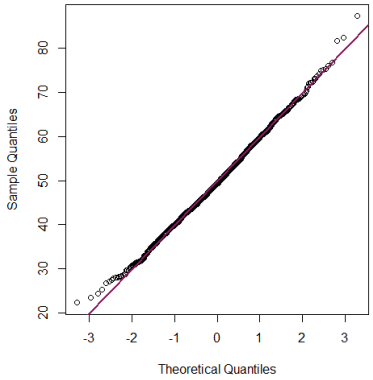
La lunghezza della base del rettangolo corrisponde allo scarto interquartile.



In un box plot ciascuno dei quattro segmenti contiene il 25% delle osservazioni, ossia un quarto dei dati, 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 quindi dividono le osservazioni in parti uguali sul numero dei dati, non sulla distanza numerica.

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

3.3.2 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 (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 che, se considerati tutti insiemi, individuano la distribuzione dei dati.

Ogni asse cartesiano di questo diagramma contiene i quantili dei due campioni presi in considerazione. Poiché i quantili sono ordinati in modo crescente, anche il grafico risultante sarà crescente, 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.

3.4 Distribuzioni normali

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

- L'istogramma è simmetrico rispetto all'intervallo centrale
- Ha il punto massimo in corrispondenza dell'intervallo centrale
- Spostandoci dal centro verso destra o verso 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 a essere un istogramma normale, allora diciamo che l'insieme di dati è approssimativamente normale. Inoltre l'insieme di dati si dice assimmetrico 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.

Regola empirica 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 ritentrano nell'intervallo $\bar{x} \pm 2s$
- circa il 99.7% delle osservazioni ritentrano nell'intervallo $\bar{x} \pm 3s$

Insiemi di dati bimodali Un insieme di dati ottenuto campionando una popolazione costituita da sottogruppi eterogenei non è di solito normale. Piuttosto, l'istogramma di un insieme di dati di questo tipo spesso rassomiglia a 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 quindi *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.

3.5 Indici di dipendenza

Si consideri un insieme composto da dati accoppiati $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. 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.
Si vuole 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.

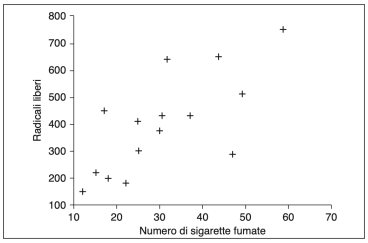
3.5.1 Covarianza campionaria

Si introduce una statistica, detta *covarianza campionaria*, 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 Si procede 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 costituito dalle coppie di valori (x_i, y_i) con $i = 1, \cdots, n$. Si calcolino le rispettive medie campionarie $\bar{x}$ e $\bar{y}$. Per la i-esima coppia di dati, si considerino $(x_i - \bar{x})$ lo scarto del valore x rispetto alla sua media campionaria e $(y - \bar{y})$ lo scarto del valore y rispetto alla sua media campionaria.

Quando grandi valori di x tendono a essere associati con grandi valori di y, e piccoli valori di x tendono a 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. A questo punto, se gli scarti hanno segno concorde, il loro prodotto $(x_i - \bar{x})(y_i - \bar{y})$ sarà positivo. Si ottiene che la sommatoria $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tenderà a essere un grande numero positivo.



Sigarette fumate rispetto al numero di radicali liberi.

x "grande" e y "grande"	x "piccolo" e y "piccolo"
$x \geq \bar{x}$	$x < \bar{x}$
$y \geq \bar{y}$	$y < \bar{y}$
$(x_i - \bar{x}) \geq 0$	$(x_i - \bar{x}) < 0$
$(y_i - \bar{y}) \geq 0$	$(y_i - \bar{y}) < 0$
Tendenzialmente:	Tendenzialmente:
$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$	$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$

Si individua quindi una correlazione 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.

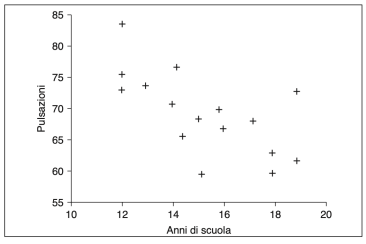


Diagramma di dispersione degli anni di scuola e delle pulsazioni.

x "grande" e y "piccola"	x "piccolo" e y "grande"
$x \geq \bar{x}$	$x < \bar{x}$
$y < \bar{y}$	$y \geq \bar{y}$
$(x_i - \bar{x}) \geq 0$	$(x_i - \bar{x}) < 0$
$(y_i - \bar{y}) < 0$	$(y_i - \bar{y}) \geq 0$
Tendenzialmente:	Tendenzialmente:
$(x_i - \bar{x})(y_i - \bar{y}) < 0$	$(x_i - \bar{x})(y_i - \bar{y}) < 0$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$

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.

Si procede poi standardizzando la sommatoria dividendo per $n-1$, al fine di evitare che questo indice assuma valori troppo elevati. Si osserva che la formula della covarianza campionaria è riconducibile a quello 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.

Ricapitolando, si definisce la covarianza campionaria come:

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

3.5.2 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.

Presa la covarianza campionaria, si standardizza il suo valore 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; invece quando $\rho < 0$ sono correlati negativamente.
Non dipendendo dalle unità di misura, questo indice può essere usato per comparare dataset diversi.

Una proprietà importante, che non verrà dimostrata, è che $-1 \leq \rho \leq 1$

Relazione deterministica

Primo caso Si passi da una relazione tendenziale a una deterministica, in cui la variabile y è una trasformazione lineare della variabile x; tutti i vari indici statistici variano di conseguenza:

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

Nella relazione deterministica $y = a + bx$, la costante b rappresenta la pendenza (ossia il coefficiente angolare) della retta che lega le due variabili e indica di quanto varia y all'aumentare di x. Ci si aspetta che:

- 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 una diminuzione di $y \Rightarrow$ relazione inversa

Si calcoli ora il coefficiente di correlazione di Pearson:

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

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 che abbiamo ottenuto con i calcoli rispecchiano le attese iniziali.

Secondo caso Si consideri ora una relazione in cui entrambe le variabili x e y sono soggette a una trasformazione lineare; i vari indici statistici variano nel seguente modo:

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

Si procede calcolando il coefficiente di correlazione di Pearson:

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

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 opposti 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 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: in altre parole, è possibile collegare tutti i valori (x_i, y_i) con $i = 1, \cdots, n$ 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. Il segno è positivo quando l'approssimazione lineare è crescente (diretta), ed è invece 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, poiché 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 di causa-effetto tra le due variabili prese in considerazione. In altre parole, due variabili possono presentare un valore di correlazione elevato senza che una determini o causi l'altra. Spesso, infatti, può intervenire un terzo fattore (o più fattori) a influenzare entrambe le variabili, generando un legame che in realtà non corrisponde a un meccanismo causale diretto.

3.6 Indici di eterogeneità

Per le variabili qualitative nominali non è possibile calcolare la varianza né gli indici che ne derivano, perché non esistono una media, una mediana o altri valori numerici di riferimento su cui misurare distanze. Risulta comunque 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.

3.6.1 Indice di Gini

Si consideri un campione $\{x_1, \cdots, x_n\}$ in cui occorrono i valori distinti $v_1, \cdots, v_m$, e si indichi con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Si definisce l'*indice di eterogeneità di Gini* come:

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

Una proprietà importante di questo indice è che $0 \leq I < 1$. Inoltre l'omogeneità massima dell'insieme di dati si presenta quando $I = 0$, mentre l'eterogeneità massima la si ha quando $I \to 1$. Di conseguenza, più aumenta il valore dell'indice di Gini e più aumenta il grado di eterogeneità.

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

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

Si noti, come accennato precedentemente, che l'estremo inferiore si presenta quando l'insieme è massimamente omogeneo e l'estremo superiore quando è massimamente eterogeneo:

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

Indice di Gini normalizzato Si ricordi che m è la cardinalità dell'insieme dei valori distinti. Questo indice presenta due problematiche:

1. il valore massimo che può assumere, ossia quando l'insieme di dati 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.
2. il suo valore dipende fortemente dal valore di m. Non è quindi possibile confrontare 2 attributi qualitativi che presentano intervalli di valori diversi, ossia m diverso.

Per ovviare a questi problemi si introduce 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: $0 \leq I' \leq 1$. Si consideri infatti il caso in cui l'eterogeneità di un insieme di dati è massima:

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

3.6.2 Indice di entropia

Si consideri un campione $\{x_1, \cdots, x_n\}$ in cui occorrono i valori distinti $v_1, \cdots, v_m$, e si indichi con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Si definisce 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 descrescente monotona, vale 0 quando $p = 1$ e tende a infinito per p che tende a 0.

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

Si effettuano le seguenti osservazioni:

- $\forall j$ vale $-f_j \log f_j \geq 0 \quad \Rightarrow \quad H \geq 0$
- $\forall j$ si ha che $-f_j \log f_j = 0 \quad \Leftrightarrow \quad f_j = 0 \vee \log f_j = 0$ 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 quindi che $0 \leq H \leq \log m$. Più questo indice cresce, e più aumenta il grado di eterogeneità dell'insieme, viceversa più descresce e più aumenta il grado di omogeneità.

Indice di entropia normalizzato Si definisce 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:

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

Se il logaritmo è in base 2 allora l'entropia si misura in bit: ciò risulta utile quando bisogna svolgere i calcoli in un computer; è comunque possibile usare altre basi, come per sempio il logaritmo naturale e quello in base 10.

3.6.3 Alberi di decisione

Gli indici di eterogeneità non servono solo per misurare la dispersione delle frequenze nelle variabili qualitative, ma trovano anche un'applicazione fondamentale nella costruzione degli alberi di decisione. In un albero di decisione, ogni oggetto da classificare è descritto da un vettore di attributi, e la classificazione avviene valutando, a partire dalla radice dell'albero, condizioni sui valori di tali attributi.

In pratica, ad ogni nodo dell'albero viene associata una condizione (o test) che suddivide il campione in sottoinsiemi: si percorre la freccia corrispondente in base al risultato del test, fino a raggiungere una foglia, la quale indica la classe assegnata. La scelta del test in ciascun nodo è guidata proprio dagli indici di eterogeneità: l'obiettivo è quello di ridurre l'eterogeneità dei dati nei nodi figli rispetto a quella del nodo padre.

Si cerca quindi di porre nei vari nodi domande che permettono di ottenere sottogruppi il più omogenei possibile. Così facendo si riduce il numero di domande da fare.

Ad esempio, utilizzando l'indice di Gini si seleziona la condizione che minimizza l'indice nei gruppi risultanti, cioè quella che porta a sottoinsiemi in cui la distribuzione delle classi è il più possibile concentrata in una sola categoria. Analogamente, se si impiega l'indice di entropia, si cerca la divisione che riduce al minimo l'incertezza (ovvero l'entropia) nei nodi successivi. In entrambi i casi, il criterio adottato assicura che, procedendo lungo l'albero, si raggiungano foglie contenenti gruppi di oggetti omogenei rispetto alla classe di appartenenza.

Così, l'impiego degli indici di eterogeneità consente di valutare quantitativamente la bontà delle suddivisioni, contribuendo a costruire alberi di decisione efficaci per il compito di classificazione.