STATISTICA 1 - Introduzione

Riccardo Corradin, Ludovica De Carolis

Introduzione

Benvenuti a Statistica 1!

- Questo corso vuole fornire una introduzione alla **statistica** ed ai suoi **aspetti descrittivi**.
- Il corso presenta le metodologie principali per **analizzare ed interpretare** un insieme di dati raccolti relativi ad un particolare fenomeno in studio.
- Gli strumenti presentati in questo corso sono **propedeutici ed utili** in diversi corsi di studio che affronterete durante i prossimi anni.

Introduzione

L'**origine etimologica** del termine statistica deriva dalla parola **stato**.

- In principio, la statistica era intesa come disciplina atta a descrivere le quantità e gli elementi che compongono uno stato.
  → Statistica: **scienza dello stato**.

- In tempi moderni, per statistica si indente in maniera più lata la disciplina che studia fenomeni osservabili e misurabili tramite la produzione e l'analisi di dati.
  → Statistica: **scienza dei dati**.

La statistica è ampiamente utilizzata in **svariati settori** dove bisogna estrarre informazioni per **sintetizzare ed interpretare** quanto osservato.

- Ricerche di mercato, marketing quantitativo.
- Finanza quantitativa, analisi del rischio.
- Statistiche ufficiali, demografia.
- Studi macroeconomici e microeconomici quantitativi.

Fondamenti del corso

Fondamenti del corso

Possiamo quindi definire, generalmente, la statistica come la **disciplina** che studia le **metodologie** necessarie per la **raccolta e l'elaborazione delle informazioni**. Tali metodologie sono utili a **esaminare** i fenomeni collettivi che osserviamo e le loro caratteristiche, osservando l'intera **popolazione** o un suo sottoinsieme, detto **campione**.

- In questo senso, la statistica utilizza strumenti matematici per:
  - → **rappresentare** l'informazione osservata, mediante l'uso di strumenti adeguati;
  - → **sintetizzare** ed **interpretare** l'informazione osservata, ed estrarre delle misure, dei risultati numerici e delle quantità utili a descrivere quanto raccolto nel campione;
  - → **identificare** e **descrivere** relazioni tra quantità differenti;
  - → **supportare** processi decisionali.

- Con il termine statistica, nel linguaggio comune, vengono indicati risultati numerici derivanti dall'analisi di alcuni dati osservati.
  - → Una **statistica** è una funzione dei dati osservati.

Un esempio emblematico: Il caso Harold Shipman[1]

Harold Shipman è stato un medico di famiglia della periferia di Manchester nonchè il **serial killer** più prolifico della storia del Regno Unito.

Si stima che tra il 1975 ed il 1998 Shipman abbia ucciso almeno 215 pazienti anziani tramite un'iniezione con una forte **overdose di oppiacei**.

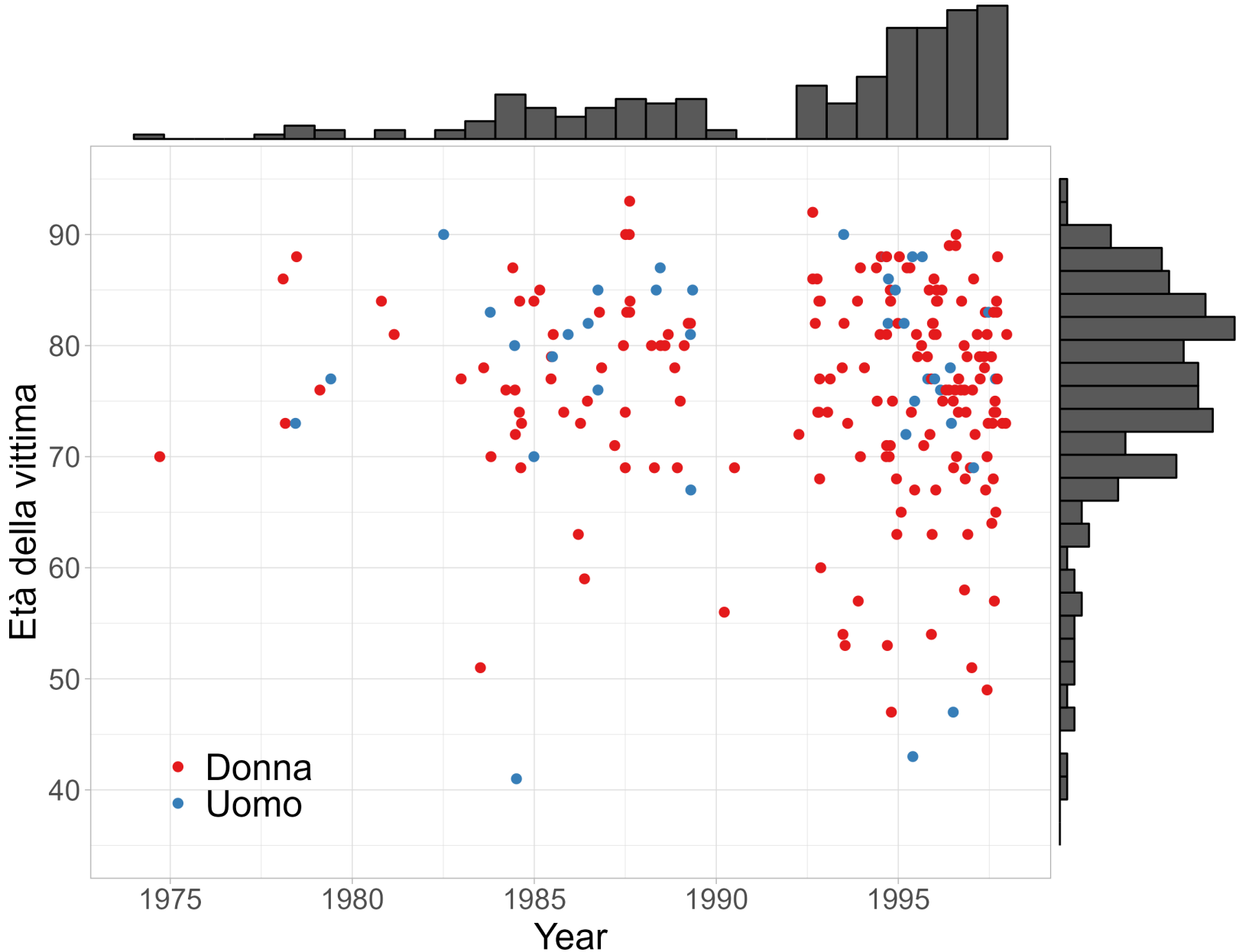
Shipman fu **condannato all'ergastolo** per l'omicidio di 15 persone nel 1999. Non testimoniò al processo, non spiegò mai il movente. Morì suicida in carcere nel 2004.

Un'indagine successiva, basata anche su metodi statistici, concluse che Shipman molto probabilmente uccise più di 215 pazienti.

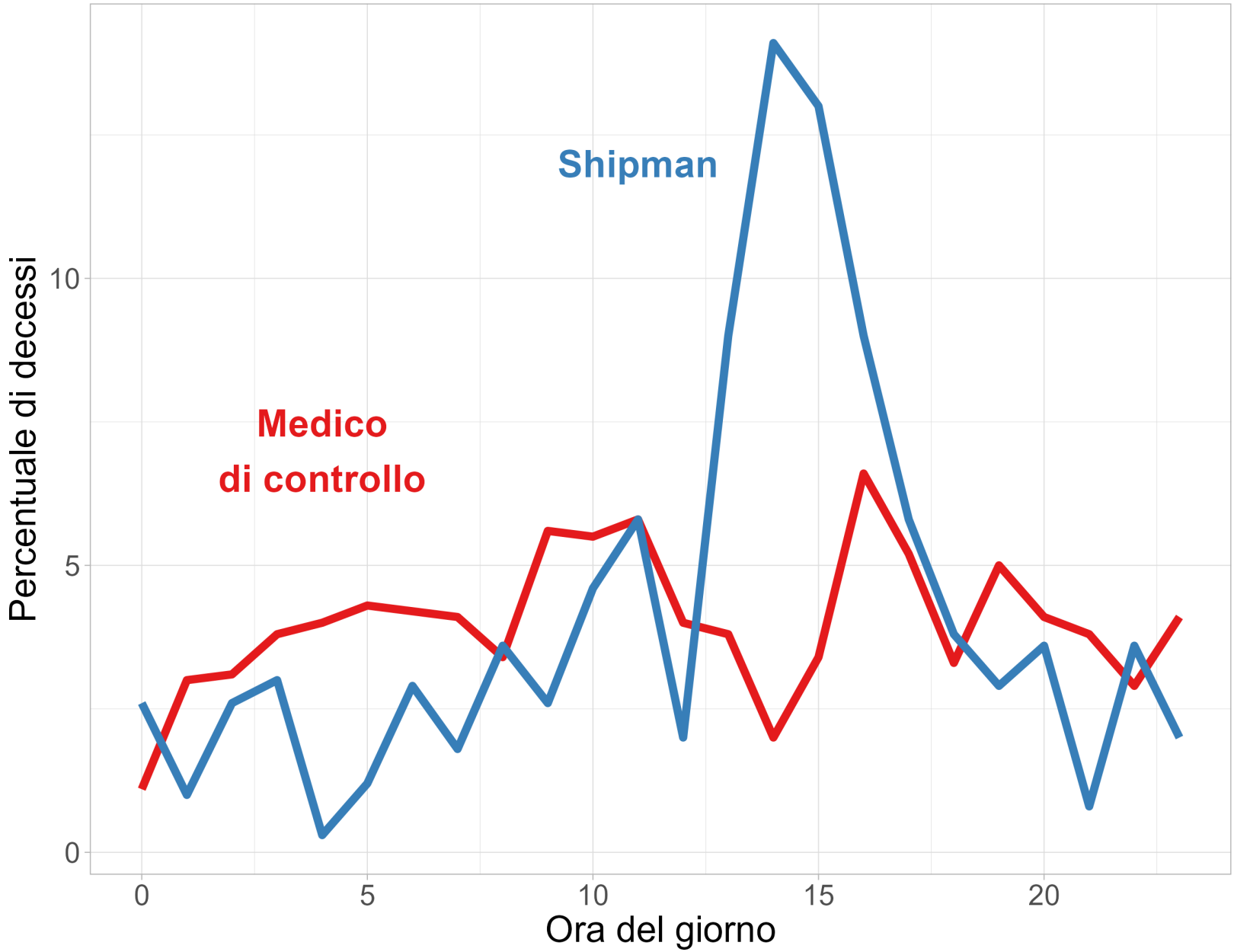
È naturale quindi chiedersi: chi erano le vittime di Harold Shipman? Poteva essere fermato prima?

[1] Fonte: `https://doi.org/10.1111/j.1740-9713.2004.00002.x`

Grafico a dispersione di età, genere, ed anno di morte delle vittime



Serie storica del numero annuale di vittime



Fondamenti del corso

L'oggetto principale delle nostre analisi sono le **unità statistiche**. Un'unità statistica è l'oggetto elementare su cui sono rilevate le variabili che compongono le nostre analisi. Esempi di unità statistiche possono essere le seguenti.

- **Persone**, per le quali possiamo rilevare altezza, età, città di nascita, etc.
- **Automobili**, per le quali rilevare colore, cilindrata, cavalli, produttore, etc.
- ...

Distinguiamo principalmente tra due tipologie di insiemi di unità statistiche.

- **Popolazione**, l'insieme esaustivo di tutte le unità statistiche che vogliamo analizzare e studiare. Ogni unità statistica potenziale fa parte della popolazione.
- **Campione**, un sottoinsieme delle possibili unità statistiche di riferimento, che abbiamo effettivamente osservato e per le quali conosciamo i valori corrispondenti delle variabili in studio.

Unità statistiche e Popolazione - Esempi

Proviamo ad analizzare meglio questi due quesiti apparentemente semplici:

Q1: Quanti alberi ci sono nel Parco Nord a Milano?

A questo punto potremmo domandarci:

- Chi sono le unità statistiche? Esiste una definizione precisa ed univoca?
- Qual è la popolazione? Ed il campione?

Q2: Quanti studenti rimangono disoccupati dopo la Laurea?

- Chi sono le unità statistiche? Esiste una definizione precisa ed univoca? La baseline?
- Qual è la popolazione? Ed il campione?

Fondamenti del corso

Possiamo anche dividere gli approcci statistici che verranno presentati durante l'intero corso di studi in due macro-gruppi.

- **Statistica descrittiva**, con lo scopo di descrivere, tramite l'utilizzo di opportune sintesi, i dati relativi all'intero campione o all'intera popolazione.
  - → Idealmente, vogliamo dare una descrizione esaustiva di quanto osservato, che sia il campione o la popolazione di riferimento.
  - → Gli strumenti della statistica descrittiva usualmente servono a rappresentare in maniera interpretabile e sintetizzare l'informazione osservata.
- **Statistica inferenziale**, con lo scopo di dedurre le caratteristiche dell'intera popolazione analizzando solo un campione di unità statistiche, e di quantificare l'incertezza associata alle nostre deduzioni.
  - In questo caso, osserviamo solo una parte dell'informazione che vogliamo sintetizzare.
  - Vogliamo riportare le conclusioni derivanti dal caso particolare (campione) nel caso generale (popolazione).

Nel corso di **Statistica 1** ci concentreremo sui metodi di **statistica descrittiva**.

Fondamenti del corso

Le analisi vogliono descrivere esaustivamente quanto osserviamo. Tali descrizioni si sviluppano principalmente in due fasi.

- **Formazione dei dati statistici**, in cui vengono definite le quantità di interesse e le metodologie di raccolta delle informazioni.
  - → Individuazione della popolazione di riferimento e dei fenomeni o caratteri di interesse.
  - → Rilevazione dei dati statistici, mediante l'uso di questionari, interviste, rilevazioni automatiche, rilevazioni di specialisti, etc.
  - → Spoglio dei dati e preparazione delle tabelle statistiche.
- **Elaborazione dei dati statistici**, dove vogliamo estrarre informazioni utili da quanto osservato e sintetizzare tali informazioni.
  - → Definizione delle metodologie necessarie, a seconda di quali domande vogliamo rispondere tramite le nostre analisi.
  - → Applicazioni delle metodologie e interpretazione dei risultati ottenuti.

Durante il corso ci concentreremo sulla seconda parte, assumendo di avere già formato i dati necessari per le nostre analisi.

Fondamenti del corso

Esistono diverse **tipologie di dato statistico**. A seconda della tipologia, alcune operazioni e metodologie di analisi sono lecite ed altre inapplicabili.

- **Caratteri qualitativi**, quando le modalità che può assumere non sono numeriche, per esempio parole, aggettivi, o altro genere di valori.
  - → **Caratteri qualitativi su scala nominale**, le modalità non sono ordinabili.
    Esempi: colore degli occhi, professione.
  - → **Caratteri qualitativi su scala ordinale**, le modalità sono ordinabili.
    Esempi: titolo di studio, giorni della settimana.
- **Caratteri quantitativi**, quando le modalità che può assumere sono numeriche.
  - → **Caratteri quantitativi continui**, possono assumere ogni valore all'interno di un intervallo predefinito.
    Esempi: altezza, temperatura.
  - → **Caratteri quantitativi discreti**, possono assumere solo valori interi.
    Esempi: numero di studenti in aula, numero di libri in biblioteca.

Tipologie di tabelle statistiche

Tipologie di tabelle statistiche

Una volta osservate le quantità di interesse, vengono organizzate in **tabelle opportune**, a seconda della loro natura.

Distribuzione di unità: quando vogliamo annotare direttamente i dati osservati per N osservazioni.

osservazione	valore
1	x_1
2	x_2
3	x_3
$\vdots$	$\vdots$
i	x_i
$\vdots$	$\vdots$
N	x_N

Riportiamo l'indice dell'osservazione corrispondente (colonna di sinistra) ed il valore (colonna di destra).

- Non sintetizziamo in alcun modo quanto osservato.
- Adatto quando osserviamo caratteri quantitativi continui e non vogliamo sintetizzare l'informazione in intervalli.

Esempio: per ogni persona in aula, misuriamo la sua altezza. Il valore generico x_i rappresenta l'altezza della i-esima persona.

Tipologie di tabelle statistiche

Distribuzione di frequenze: quando vogliamo riportare un valore osservato ed il numero di volte che osserviamo tale valore (frequenza).

Frequenza quante volte osserviamo una specifica modalità od un valore specifico (quanto frequentemente) all'interno del campione o della popolazione.

valore	frequenza
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_i	n_i
$\vdots$	$\vdots$
x_k	n_k
totale	N

k modalità o valori differenti, riportiamo i valori osservati (colonna di sinistra) ed il numero di volte che vengono osservati (colonna di destra).

- Stiamo facendo una prima sintesi dell'informazione osservata.
- Adatto quando osserviamo caratteri qualitativi o quantitativi discreti.

Esempio: per ogni persona in aula, rileviamo il colore dei capelli. Il generico x_i sarà quindi un colore specifico, mentre n_i sarà il numero di persone con i capelli di colore x_i.

Tipologie di tabelle statistiche

Distribuzione di frequenze in classi (continuo): quando vogliamo riportare un intervallo di valori osservati ed il numero di volte che osserviamo un valore all'interno dell'intervallo.

valore	frequenza
$\ell_1 \dashv u_1$	n_1
$\ell_2 \dashv u_2$	n_2
$\vdots$	$\vdots$
$\ell_i \dashv u_i$	n_i
$\vdots$	$\vdots$
$\ell_p \dashv u_p$	n_p
totale	N

k intervalli disgiunti, riportiamo i valori osservati (colonna di sinistra) ed il numero di volte che vengono osservati (colonna di destra).

- Stiamo facendo una prima sintesi dell'informazione osservata.
- $\ell_i = u_{i-1}$, con la notazione $\ell_i \dashv u_i$ si intende che ℓ_i è escluso dall'intervallo e u_i incluso.
- Adatto quando osserviamo caratteri quantitativi continui e vogliamo compattare l'informazione in intervalli.

Esempio: per ogni persona in aula, rileviamo l'altezza, espressa in centimetri. Vogliamo sintetizzare le osservazioni in una tabella con tre intervalli, $0 \dashv 155$, $155 \dashv 170$ e $170 \dashv 250$.

Tipologie di tabelle statistiche

Distribuzione di frequenze in classi (discreto): quando vogliamo riportare un intervallo di valori osservati ed il numero di volte che osserviamo un valore all'interno dell'intervallo.

valore	frequenza
$\ell_1 - u_1$	n_1
$\ell_2 - u_2$	n_2
$\vdots$	$\vdots$
$\ell_i - u_i$	n_i
$\vdots$	$\vdots$
$\ell_p - u_p$	n_p
totale	N

k intervalli disgiunti, riportiamo i valori osservati (colonna di sinistra) ed il numero di volte che vengono osservati (colonna di destra).

- Stiamo facendo una prima sintesi dell'informazione osservata.
- $\ell_i > u_{i-1} > \ell_{i-1}$
- Adatto quando osserviamo caratteri quantitativi discreti e vogliamo compattare l'informazione in intervalli.

Esempio: per ogni insegnamento erogato all'interno del corso di laurea, osserviamo il numero di studenti. Dividiamo il numero di studenti in tre classi, $0 - 30$, $31 - 100$ e $101 - 300$.

Tipologie di tabelle statistiche

Serie storiche temporali: quando osserviamo un fenomeno che ha dipendenza temporale. Le osservazioni hanno un vincolo di ordinamento.

tempo	valore
1	x_1
2	x_2
$\vdots$	$\vdots$
t	x_t
$\vdots$	$\vdots$
T	x_T

T tempi di osservazione, riportiamo il tempo (colonna di sinistra) ed il valore osservato (colonna di destra).

- Non stiamo facendo sintesi dell'informazione osservata.
- Adatto quando osserviamo caratteri quantitativi che dipendono dal tempo.

Esempio: per ogni giorno a partire dal primo gennaio, misuriamo il numero di studenti che frequentano l'edificio U7. In questo caso t rappresenta un istante temporale (giorno) e x_t il numero di studenti nel giorno t-esimo.

Strumenti matematici che useremo

Sommatoria

Supponiamo di avere **cinque numeri**: 3, 6, 2, 11, 7, indicati come segue

$$a_1 = 3,\ a_2 = 6,\ a_3 = 2,\ a_4 = 11,\ a_5 = 7.$$

La **somma di tutti i valori** osservati è data da

$$a_1 + a_2 + a_3 + a_4 + a_5 = 3 + 6 + 2 + 11 + 7 = 29,$$

che può essere letta come la somma di tutti i valori di a_i, per $i = 1, \ldots, 5$.

Una sommatoria non è altro che un **modo più compatto** per scrivere l'operazione precedente. Viene indicata dalla **lettera greca sigma maiuscola**, come

$$\begin{aligned} \sum_{i=1}^{5} a_i &= a_1 + a_2 + a_3 + a_4 + a_5 \\ &= 3 + 6 + 2 + 11 + 7 = 29, \end{aligned}$$

e si legge **sommatoria delle a_i per i che va da 1 a 5**.

Sommatoria

Proprietà S1. Sia $\alpha \in \mathbb{R}$ una costante nota che non dipende dall'indice i, allora

$$\sum_{i=1}^{n} \alpha a_i = \alpha \sum_{i=1}^{n} a_i.$$

Possiamo infatti vedere che

$$\sum_{i=1}^{n} \alpha a_i = (\alpha a_1 + \alpha a_2 + \cdots + \alpha a_n) = \alpha(a_1 + a_2 + \cdots + a_n) = \alpha \sum_{i=1}^{n} a_i.$$

Proprietà S2. La seguente è valida

$$\sum_{i=1}^{n} 1 = n.$$

Infatti

$$\sum_{i=1}^{n} 1 = \overbrace{1 + 1 + \cdots + 1}^{\text{n volte}} = n.$$

Sommatoria

Proprietà S3. Per una generica costante $\alpha \in \mathbb{R}$, abbiamo

$$\sum_{i=1}^{n} \alpha = n \times \alpha.$$

Applicando la Proprietà S1 e la Proprietà S2, abbiamo

$$\sum_{i=1}^{n} \alpha = \alpha \sum_{i=1}^{n} 1 = \alpha \times n.$$

Nota bene Se abbiamo due sequenze di numeri $\{a_1, \ldots, a_n\}$ e $\{b_1, \ldots, b_n\}$,

$$\sum_{i=1}^{n} (a_i \times b_i) \neq \left(\sum_{i=1}^{n} a_i \right) \left(\sum_{i=1}^{n} b_i \right)$$

Per due sequenze di due numeri, abbiamo

$$(a_1 + b_1) + (a_2 + b_2) \neq (a_1 + a_2) \times (b_1 + b_2) = a_1 b_1 + a_2 b_1 + a_1 b_2 + a_2 b_2.$$

Analogamente, la disuguaglianza vale per più di due numeri.

Sommatoria

Proprietà S4. Siano $\{a_1, \ldots, a_n\}$ e $\{b_1, \ldots, b_n\}$ due sequenze di numeri. La sommatoria di una somma è la somma delle sommatorie,

$$\sum_{i=1}^{n}(a_i + b_i) = \left(\sum_{i=1}^{n} a_i\right) + \left(\sum_{i=1}^{n} b_i\right)$$

Applicando la Proprietà S1 e la Proprietà S2, abbiamo

$$\sum_{i=1}^{n}(a_i + b_i) = [(a_1 + b_1) + \cdots + (a_n + bn)]$$

$$= (a_1 + \cdots + a_n) + (b_1 + \cdots + b_n) = \left(\sum_{i=1}^{n} a_i\right) + \left(\sum_{i=1}^{n} b_i\right)$$

Proprietà S5. Siano $\{a_1, \ldots, a_n\}$ e $\{a_{n+1}, \ldots, a_m\}$ due sequenze di numeri. Allora abbiamo che

$$\sum_{i=1}^{n} a_i + \sum_{i=n+1}^{m} a_i = \sum_{i=1}^{m} a_i.$$

Sommatoria

Proprietà S6. La sommatoria di una trasformazione lineare di una sequenza di numeri

$$\alpha a_i + \beta, \qquad i = 1, \ldots, n,$$

per $\alpha, \beta \in \mathbb{R}$, può essere scritta come

$$\sum_{i=1}^{n} (\alpha a_i + \beta) = \left(\alpha \sum_{i=1}^{n} a_i \right) + n\beta.$$

Applicando la Proprietà S4, abbiamo

$$\sum_{i=1}^{n} (\alpha a_i + \beta) = \sum_{i=1}^{n} \alpha a_i + \sum_{i=1}^{n} \beta.$$

Infine, applicando la Proprietà S1 e la Proprietà S3, otteniamo

$$\sum_{i=1}^{n} \alpha a_i + \sum_{i=1}^{n} \beta = \alpha \sum_{i=1}^{n} a_i + n\beta.$$

Sommatoria

Example

Calcolare esplicitamente il risultato delle seguenti sommatorie

$$\sum_{i=2}^{4}(5+3i+i^2); \qquad \sum_{i=1}^{3}\left(\frac{i^2}{4}+3i\right).$$

1. Si ha che

$$\sum_{i=2}^{4}(5+3i+i^2) = (5+3\cdot 2+2^2)+(5+3\cdot 3+3^2)+(5+3\cdot 4+4^2) =$$

$$= 15+23+33 = 71$$

2. Analogamente si ha che

$$\sum_{i=1}^{3}\left(\frac{i^2}{4}+3i\right) = \left(\frac{1^2}{4}+3\cdot 1\right)+\left(\frac{2^2}{4}+3\cdot 2\right)+\left(\frac{3^2}{4}+3\cdot 3\right) =$$

$$= 3.25+7+11.25 = 21.5$$

Sommatoria

Example

Calcolare esplicitamente il risultato delle seguenti sommatorie

$$\sum_{i=1}^{3} 2^i, \qquad \sum_{i=3}^{4} 3 \times \frac{1}{i}, \qquad \sum_{i=1}^{3} (i-1).$$

Sommatoria

Produttoria

Supponiamo di avere **cinque numeri**: 3, 4, 2, 5, 3, indicati come segue

$$a_1 = 3,\ a_2 = 4,\ a_3 = 2,\ a_4 = 5,\ a_5 = 3.$$

Il **prodotto di tutti i valori** osservati è dato da

$$a_1 \times a_2 \times a_3 \times a_4 \times a_5 = 3 \times 4 \times 2 \times 5 \times 3 = 360,$$

che può essere letto come il prodotto di tutti i valori di a_i, per $i = 1, \ldots, 5$.

Una produttoria non è altro che un **modo più compatto** per scrivere l'operazione precedente. Viene indicata dalla **lettera greca pi maiuscola**, come

$$\begin{aligned} \prod_{i=1}^{5} a_i &= a_1 \times a_2 \times a_3 \times a_4 \times a_5 \\ &= 3 \times 4 \times 2 \times 5 \times 3 = 360, \end{aligned}$$

e si legge **produttoria delle a_i per i che va da 1 a 5**.

Produttoria

Proprietà P1. Sia $\alpha \in \mathbb{R}$ una costante che non dipende dall'indice i. Allora vale la seguente

$$\prod_{i=1}^{n} \alpha a_i = \alpha^n \prod_{i=1}^{n} a_i.$$

Possiamo notare che

$$\prod_{i=1}^{n} \alpha a_i = (\alpha a_1 \times \cdots \times \alpha a_n) = \alpha^n (a_1 \times a_n) = \alpha^n \prod_{i=1}^{n} a_i.$$

Proprietà P2. Siano $\{a_1, \ldots, a_n\}$ e $\{b_1, \ldots, b_n\}$ due sequenze di numeri. Il prodotto delle produttorie è uguale alla produttoria dei prodotti, ovvero

$$\prod_{i=1}^{n} a_i \times \prod_{i=1}^{n} b_i = \prod_{i=1}^{n} (a_i \times b_i).$$

Possiamo osservare che

$$\prod_{i=1}^{n} a_i \times \prod_{i=1}^{n} b_i = (a_1 \times \cdots \times a_n)(b_1 \times \cdots \times b_n) = \prod_{i=1}^{n} (a_i \times b_i).$$

Produttoria

Example

Calcolare esplicitamente il risultato delle seguenti produttorie

$$\prod_{j=0}^{3} \cos(2j\pi); \qquad \prod_{i=1}^{3} \frac{i \cdot (i+1)}{i+2}$$

1. Si ha che

$$\prod_{j=0}^{3} \cos(2j\pi) = \cos(2 \cdot 0 \cdot \pi) \cdot \cos(2 \cdot 1 \cdot \pi) \cdot \cos(2 \cdot 2 \cdot \pi) =$$

$$= \cos(0) \cdot \cos(2\pi) \cdot \cos(4\pi) = 1 \cdot 1 \cdot 1 = 1$$

2. Si ha che

$$\prod_{i=1}^{3} \frac{i \cdot (i+1)}{i+2} = \frac{1 \cdot 2}{3} \cdot \frac{2 \cdot 3}{4} \frac{3 \cdot 4}{5} = \frac{12}{5}$$

Produttoria

Example

Calcolare esplicitamente il risultato delle seguenti produttorie

$$\prod_{i=1}^{3} 2^i, \qquad \prod_{i=1}^{3} 3 \times \frac{1}{i}, \qquad \prod_{i=1}^{3} (i-1).$$

Produttoria

Distribuzioni di frequenze

Distribuzioni di frequenze

Ricordiamo che le frequenze rappresentano quanto **frequentemente** osserviamo un valore o una modalità di un carattere in studio.

Per semplicità, **assumiamo di avere** N osservazioni, dove ogni osservazione può assumere un valore fra k **modalità** $\{x_1, \ldots, x_k\}$.

Possiamo definire diverse tipologie di frequenze.

Frequenze assolute: rappresentano semplicemente quante volte rileviamo una specifica modalità, uno specifico valore o uno specifico intervallo di valori all'interno delle nostre osservazioni. Indichiamo la frequenza assoluta associata alla j-esima modalità come n_j.

Proprietà F1. La somma delle frequenze assolute è pari al numero totale delle osservazioni, ossia $n_1 + \cdots + n_k = N$.

Distribuzioni di frequenze

Frequenze relative: rappresentano l'incidenza relativa di una specifica modalità sul totale delle osservazioni, e sono definite come

$$f_j = \frac{n_j}{N}, \qquad j = 1, \ldots, k,$$

dove n_j è la frequenza assoluta della j-esima modalità e N il numero totale di osservazioni.

Proprietà F2. Per le frequenze relative valgono le seguenti:

- $0 \leq f_j \leq 1$, $j = 1, \ldots, k$ ossia ogni frequenza relativa non può essere minore di 0 o maggiore di 1.
- La somma di tutte le frequenze relative è pari a 1.

Il secondo punto delle precedenti proprietà deriva dalle proprietà delle sommatorie. Infatti abbiamo che

$$\sum_{j=1}^{k} f_j = \sum_{j=1}^{k} \frac{n_j}{N} = \frac{1}{N} \sum_{j=1}^{k} n_j = \frac{N}{N} = 1.$$

Distribuzioni di frequenze

Example

Consideriamo le seguenti osservazioni

$$\{2, 2, 1, 2, 1, 3, 2, 2, 1, 1\}.$$

Calcolare le frequenze assolute e relative e mostrare che le proprietà F1 e F2 sono valide.

Distribuzioni di frequenze

Frequenze cumulate assolute: possiamo calcolarle solo quando le modalità sono ordinabili. Disponiamo le modalità in ordine crescente. Le frequenze cumulate assolute rappresentano il numero di osservazioni che assumono una modalità minore o uguale di quella considerata, ossia

$$C_j = n_1 + n_2 + \cdots + n_j = \sum_{i=1}^{j} n_i,$$

dove n_i è la frequenza assoluta della i-esima modalità.

Frequenze cumulate relative: simili al caso precedente, possiamo calcolarle solo con modalità ordinabili. Rappresentano l'incidenza relativa delle osservazioni che assumono una modalità minore o uguali di quella considerata,

$$F_j = \frac{C_j}{N} = \frac{n_1 + \cdots + n_j}{N} = \sum_{i=1}^{j} \frac{n_i}{N} = \sum_{i=1}^{j} f_i,$$

dove n_i è la frequenza assoluta della i-esima modalità e N il numero totale di osservazioni.

Distribuzioni di frequenze

Frequenze retrocumulate assolute: possiamo calcolarle solo quando le modalità sono ordinabili. Disponiamo le modalità in ordine crescente. Le frequenze retrocumulate assolute rappresentano il numero di osservazioni che assumono una modalità maggiore o uguale di quella considerata, ossia

$$R_j = n_j + n_{j+1} + \cdots + n_k = \sum_{i=j}^{k} n_i,$$

dove n_i è la frequenza assoluta della i-esima modalità.

Frequenze retrocumulate relative: simili al caso precedente, possiamo calcolarle solo con modalità ordinabili. Rappresentano l'incidenza relativa delle osservazioni che assumono una modalità maggiore o uguale di quella considerata,

$$Q_j = \frac{R_j}{N} = \frac{n_j + n_{j+1} + \cdots + n_k}{N} = \sum_{i=j}^{k} \frac{n_i}{N} = \sum_{i=j}^{k} f_i,$$

dove n_i è la frequenza assoluta della i-esima modalità e N il numero totale di osservazioni.

Distribuzioni di frequenze

Example

Continuando l'esempio precedente, dove abbiamo osservato

$$\{2, 2, 1, 2, 1, 3, 2, 2, 1, 1\},$$

calcolare le frequenze cumulate e retrocumulate, assolute e relative.

Distribuzioni di frequenze

Example

Supponiamo di analizzare le recensioni lasciate da $N = 100$ clienti di un nuovo smartphone venduto online da una nota azienda di elettronica. I clienti possono assegnare un voto da 1 a 5 stelle. Sapendo che 10 clienti hanno lasciato una recensione ad una stella, 15 due stelle, 25 tre stelle e 30 quattro stelle, calcolare: frequenze *assolute*, *relative*, *cumulate*, *retrocumulate*.

Punteggio	Freq: Assoluta	Relativa	Cumulata	Retrocumulata
$\star$	10	$\frac{10}{100} = 0.1$	10	100
$\star\star$	15	$\frac{15}{100} = 0.15$	25	90
$\star\star\star$	25	$\frac{25}{100} = 0.25$	50	75
$\star\star\star\star$	30	$\frac{30}{100} = 0.30$	80	50
$\star\star\star\star\star$	20	$\frac{20}{100} = 0.20$	100	20
	100	1		

Distribuzioni di frequenze

Quando le nostre osservazioni sono quantitative (discrete o continue) e misurate in intervalli, una quantità cruciale è data dalle **ampiezze delle classi** definite da ogni singolo intervallo.

Nel caso discreto, l'ampiezza a_j dell'intervallo j-esimo è data dal numero di modalità che appartengono all'intervallo. Nel caso di numeri interi, abbiamo che

$$a_j = u_j - \ell_j + 1, \qquad j = 1, \ldots, k,$$

ossia tutti i valori compresi tra ℓ_j e u_j, inclusi gli estremi dell'intervallo.

Distribuzioni di frequenze

Nel caso continuo, l'ampiezza a_j dell'intervallo j-esimo è data da

$$u_j - \ell_j, \qquad j = 1, \ldots, k,$$

ossia la lunghezza dell'intervallo stesso.

Distribuzioni di frequenze

Nel caso di rappresentazioni in intervalli, possiamo calcolare le seguenti frequenze.

Frequenze specifiche assolute: correggiamo la frequenza dell'intervallo per l'ampiezza dello stesso, ovvero

$$k_j = \frac{n_j}{a_j}, \qquad j = 1, \ldots, k.$$

Frequenze specifiche relative: correggiamo la frequenza relativa di un intervallo per l'ampiezza dello stesso, ovvero

$$d_j = \frac{f_j}{a_j} = \frac{k_j}{N}, \qquad j = 1, \ldots, k.$$

Distribuzioni di frequenze

Example

Supponiamo di avere osservato i seguenti valori continui per $N = 10$ osservazioni

$$x_1 = 1.1,\ x_2 = 1.4,\ x_3 = 0.7,\ x_4 = 1.6,\ x_5 = 0.5,$$
$$x_6 = 2.7,\ x_7 = 3.1,\ x_8 = 1.9,\ x_9 = 2.9,\ x_{10} = 2.6.$$

Vogliamo dividere le osservazioni in tre intervalli distinti, $0 \dashv 1.5$, $1.5 \dashv 2.5$ e $2.5 \dashv 4$. Successivamente, vogliamo calcolare le ampiezze delle varie classi e le frequenze specifiche assolute e relative.

Numeri indice

Numeri indice

I **numeri indice**, spesso chiamati semplicemente indici, sono degli indicatori che misurano come varia un fenomeno rispetto ad una base di riferimento. Tipicamente vengono usati per studiare fenomeni che variano nel tempo. Abbiamo quindi una sequenza di valori $\{x_0, x_1, \ldots, x_N\}$ osservata ad istanti temporali $\{t_0, t_1, \ldots, t_N\}$.

Distinguiamo tra due tipologie di numeri indice.

- **Numeri indice a base mobile**, definiti come

$$\mathrm{I}_{i,i-1} = \frac{x_i}{x_{i-1}}, \qquad i = 1, \ldots, N.$$

  Rappresentano come varia un fenomeno in un istante temporale specifico, rispetto all'istante temporale precedente. Il denominatore cambia nel tempo, la base di riferimento è mutevole.

- **Numeri indice a base fissa**, definiti come

$$\mathrm{I}_{i,0} = \frac{x_i}{x_0}, \qquad i = 1, \ldots, N.$$

  Rappresentano come varia un fenomeno in un istante temporale specifico, ad un'istante temporale specifico, in questo caso al tempo t_0. Il denominatore non cambia nel tempo, la base di riferimento è immutata.

Numeri indice

Example

Supponiamo di avere osservato i seguenti valori a diversi istanti temporali

$$x_0 = 1.2, x_1 = 3.2, x_2 = 4.3, x_3 = 2.1, x_4 = 2.7, x_5 = 4.1.$$

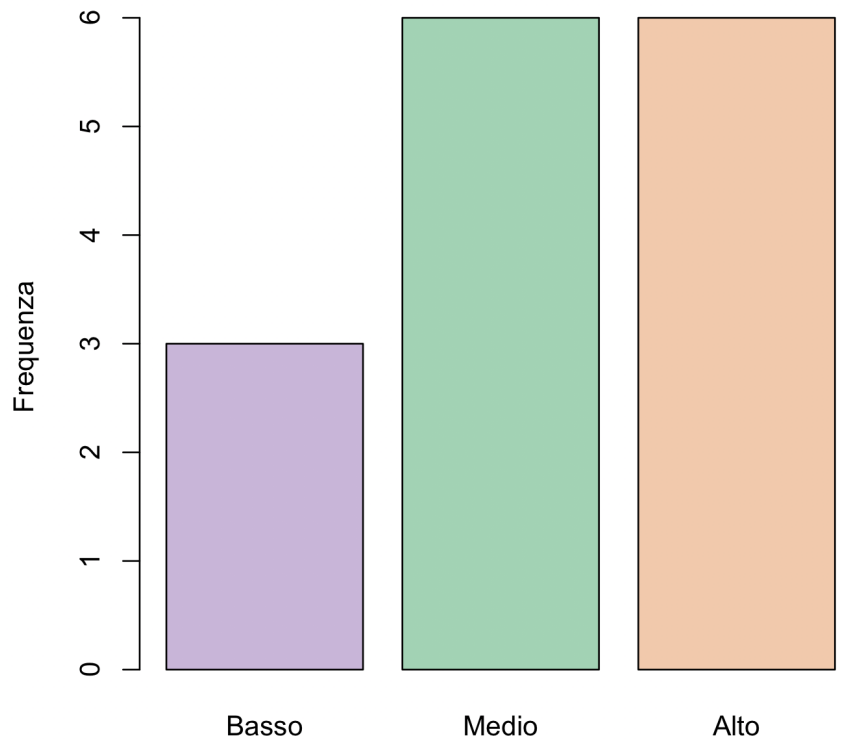
Calcolare i numeri indice a base mobile dei precedenti valori ed i numeri indice a base fissa, utilizzando x_0 come base di riferimento.

Rappresentazioni grafiche

Rappresentazioni grafiche

Diagramma a barre. Presenta una barra in corrispondenza di ogni modalità osservata di altezza pari alla frequenza di tale modalità. Notiamo che:

- rende visibile l'ordinamento delle modalità
- sull'asse orizzontale sono riportate le modalità in ordine crescente ma tale asse non ha alcun significato numerico
- è più facile confrontare due modalità di frequenza simili



Rappresentazioni grafiche

Example

Supponiamo di avere osservato il canale di acquisizione di 12 nuovi clienti di un e-commerce (campagne mail, Search Engine Optimization e pubblicità):

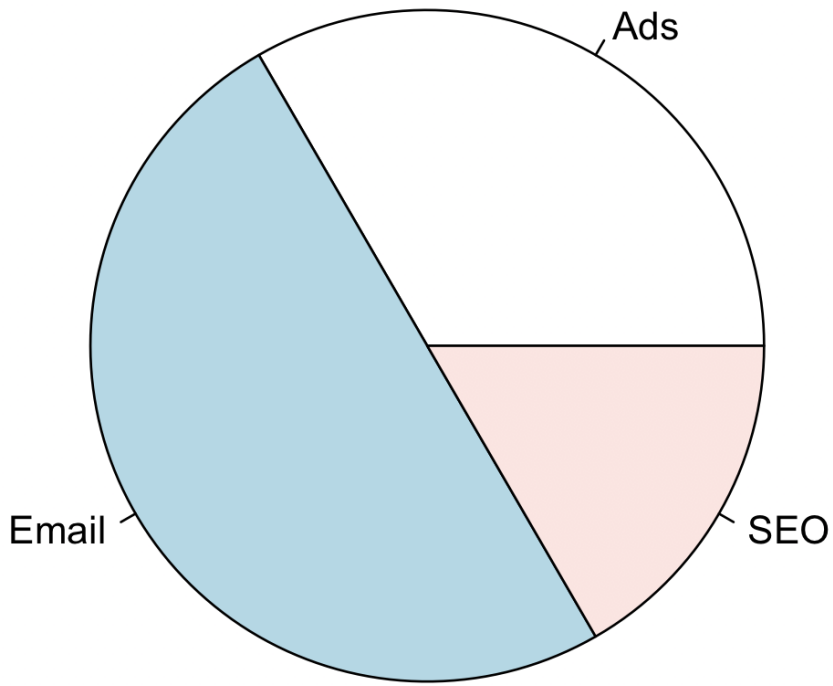
{Ads, Ads, SEO, Ads, SEO, Email, Email, Email, Email, Ads, Email, Email}

Che rappresentazioni grafiche possono essere idonee a illustrare la distribuzione di frequenza di una variabile di questo tipo?

Rappresentazioni grafiche

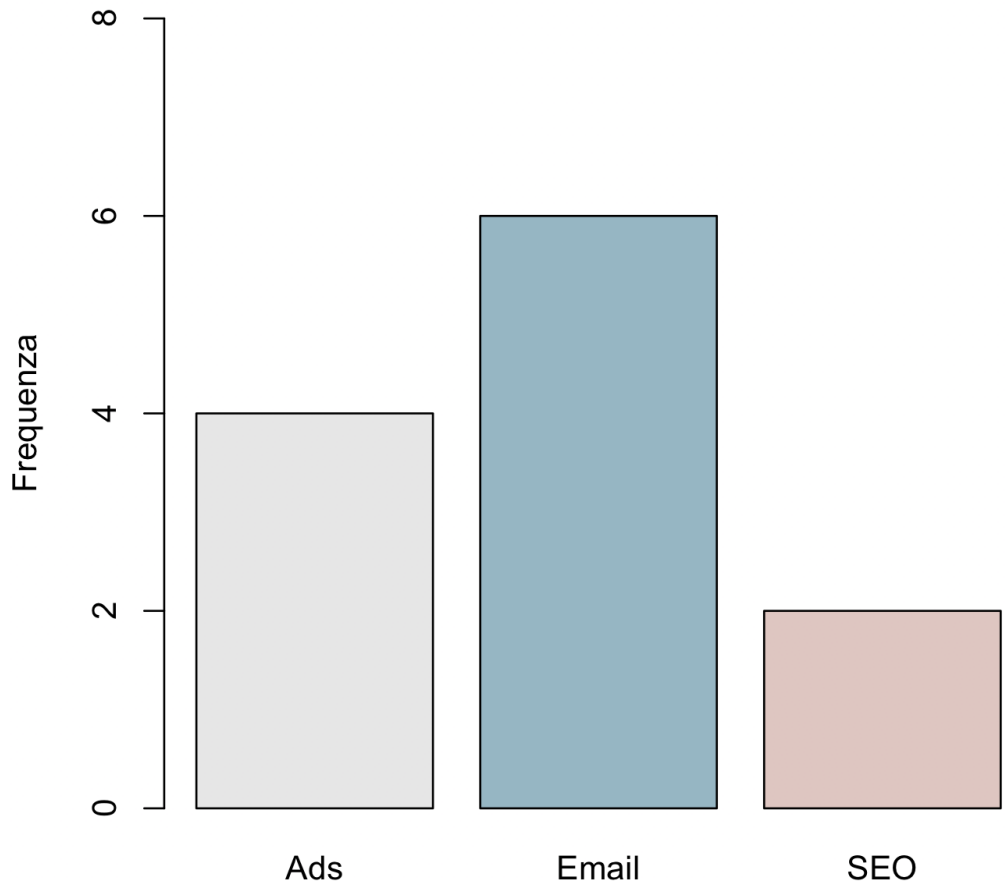
Diagramma a torta È una rappresentazione grafica circolare che suddivide un insieme di dati in spicchi proporzionali alle frequenze relative delle categorie (quote di clienti provenienti da quel canale). Notiamo che:

- fornisce informazioni solo sull'importanza relativa di ogni modalita
- può risultare difficile confrontare frequenze di diverse modalità, se simili



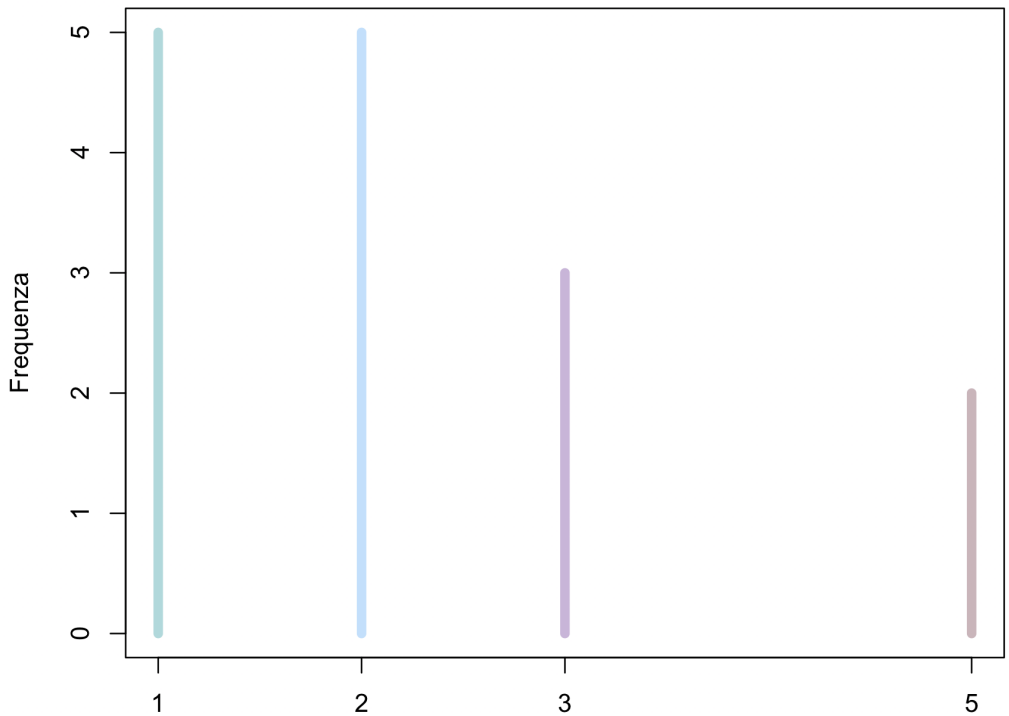
Rappresentazioni grafiche

Diagramma a barre. Può essere usato anche per variabili nominali purché non induca una falsa impressione di ordinamento.



Rappresentazioni grafiche

Diagramma a bastoncino (stick plot): Presenta un bastoncino in corrispondenza di ogni valore osservato la cui altezza è pari alla frequenza della modalità. Notiamo che tiene conto sia delle modalità osservate sia delle loro distanze (il 5 è a una distanza proporzionale da 3).



Rappresentazioni grafiche

Example

Supponiamo di avere osservato il numero di prodotti acquistati da 15 clienti durante una promozione:

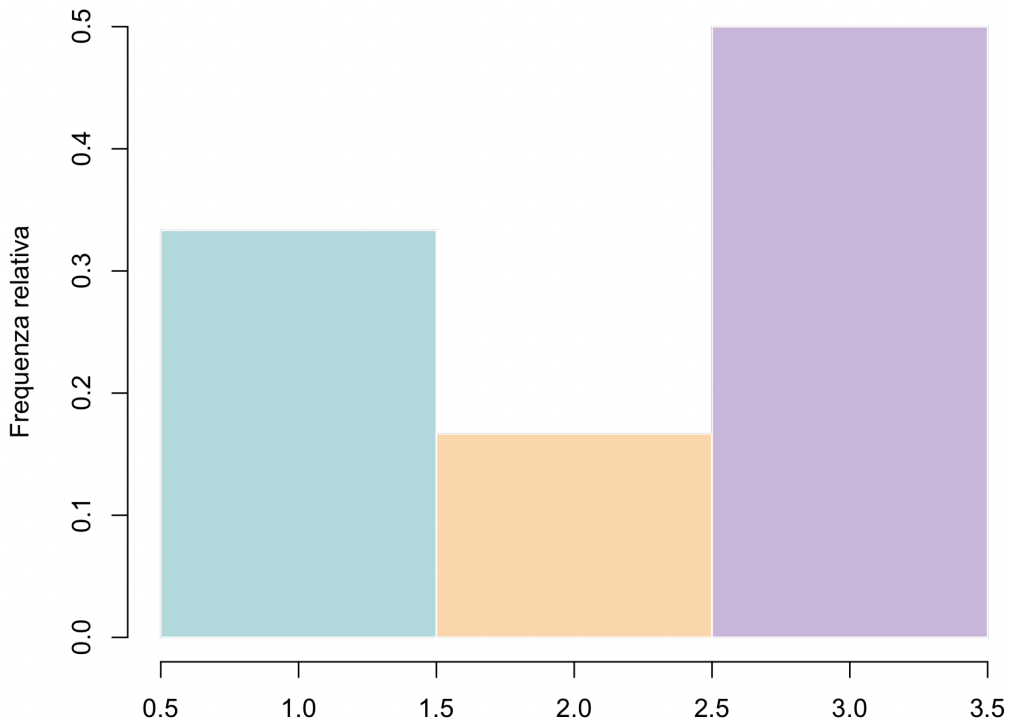
$$\{1, 2, 1, 3, 2, 1, 5, 2, 3, 1, 2, 5, 1, 3, 2\}.$$

Che rappresentazioni grafiche possono essere idonee a illustrare la distribuzione di frequenza di una variabile di questo tipo?

Rappresentazioni grafiche

Istogramma: Se costruiamo delle classi di uguale ampiezza e su ciascuna classe si costruisce un rettangolo, avente:

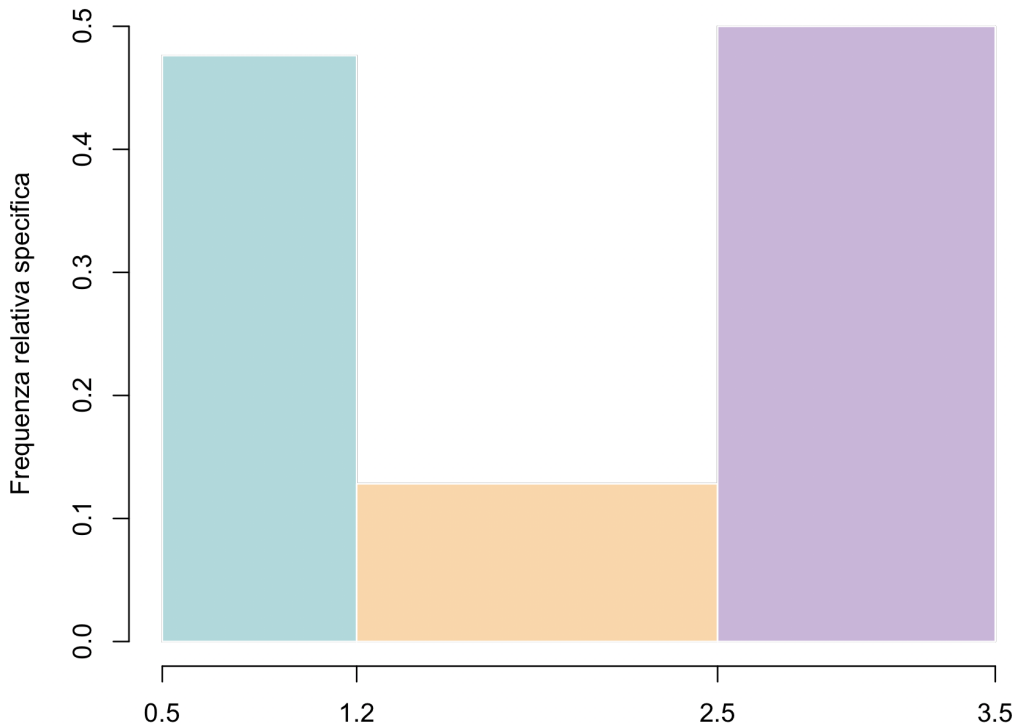
- la base pari all'ampiezza della classe
- l'altezza pari alla frequenza relativa (o assoluta) della classe
- (di conseguenza) l'area pari alla frequenza relativa (o assoluta) per l'ampiezza della classe



Rappresentazioni grafiche

Istogramma: Se costruiamo delle classi di diversa ampiezza, su ciascuna classe si costruisce un rettangolo avente:

- la base pari all'ampiezza della classe
- l'area pari alla frequenza relativa (o assoluta) della classe
- (di conseguenza) l'altezza pari alla frequenza specifica relativa (o assoluta), ovvero frequenza diviso ampiezza della classe



Example

Un e-commerce osserva la spesa (in euro) effettuata da 10 clienti durante una campagna promozionale:

$$\{12, 27, 14, 15, 38, 31, 28, 11, 18, 22\}.$$

Disegnare un istogramma dei valori precedenti, considerando come classi $0 \dashv 15$, $15 \dashv 23$ e $23 \dashv 40$. Interpretare il risultato in termini di segmentazione della clientela (bassa, media e alta spesa).