

Elementi di Statistica descrittiva

per il corso di Laurea in Biologia
a cura della Prof.ssa F. Prinari

October 8, 2019

Informazioni utili per il corso di Matematica

- **Docenti:** Prof.ssa Francesca Prinari, Prof.ssa Cinzia Bisi, Prof. Alberto Calabri (nell'ordine rispetto ai tre periodi di lezione)

Informazioni utili per il corso di Matematica

- **Docenti:** Prof.ssa Francesca Prinari, Prof.ssa Cinzia Bisi, Prof. Alberto Calabri (nell'ordine rispetto ai tre periodi di lezione)
- **Uffici:** via Machiavelli 35 (Dipartimento di Matematica e Informatica)

Informazioni utili per il corso di Matematica

- **Docenti:** Prof.ssa Francesca Prinari, Prof.ssa Cinzia Bisi, Prof. Alberto Calabri (nell'ordine rispetto ai tre periodi di lezione)
- **Uffici:** via Machiavelli 35 (Dipartimento di Matematica e Informatica)
- **e-mail:** cinzia.bisi@unife.it, alberto.calabri@unife.it, francesca.prinari@unife.it

Informazioni utili per il corso di Matematica

- **Docenti:** Prof.ssa Francesca Prinari, Prof.ssa Cinzia Bisi, Prof. Alberto Calabri (nell'ordine rispetto ai tre periodi di lezione)
- **Uffici:** via Machiavelli 35 (Dipartimento di Matematica e Informatica)
- **e-mail:** cinzia.bisi@unife.it, alberto.calabri@unife.it, francesca.prinari@unife.it
- **Ricevimenti in ottobre:** compilare google-form <https://forms.gle/CasCy1qMgzPeHAaRA>

Informazioni utili per il corso di Matematica

- **Docenti:** Prof.ssa Francesca Prinari, Prof.ssa Cinzia Bisi, Prof. Alberto Calabri (nell'ordine rispetto ai tre periodi di lezione)
- **Uffici:** via Machiavelli 35 (Dipartimento di Matematica e Informatica)
- **e-mail:** cinzia.bisi@unife.it, alberto.calabri@unife.it, francesca.prinari@unife.it
- **Ricevimenti in ottobre:** compilare google-form <https://forms.gle/CasCy1qMgzPeHAaRA>
- **Programma, libri consigliati e materiale del corso:** <http://www.unife.it/scienze/biologia/Insegnamenti/matematica-a-a-2018-2019>

Esami e date

- **Date degli esami della sessione invernale:** martedì 21 gennaio e giovedì 13 febbraio; l'iscrizione alla prova scritta è obbligatoria, così come la presentazione di un documento di identità.

Esami e date

- **Date degli esami della sessione invernale:** martedì 21 gennaio e giovedì 13 febbraio; l'iscrizione alla prova scritta è obbligatoria, così come la presentazione di un documento di identità.
- **Modalità di esame:** una prova scritta composta da 12 domande (10 a risposta multipla e 2 a risposta aperta). Le 2 domande aperte prevedono uno studio di funzione e la risoluzione di un integrale.

Esami e date

- **Date degli esami della sessione invernale:** martedì 21 gennaio e giovedì 13 febbraio; l'iscrizione alla prova scritta è obbligatoria, così come la presentazione di un documento di identità.
- **Modalità di esame:** una prova scritta composta da 12 domande (10 a risposta multipla e 2 a risposta aperta). Le 2 domande aperte prevedono uno studio di funzione e la risoluzione di un integrale.
- **Valutazione:** ogni domanda a risposta multipla corretta vale 1 punto, le domande aperte valgono insieme un massimo di 22 punti. Le risposte alle domande aperte, però, verranno corrette se e solo se il punteggio ottenuto nelle 10 domande a risposta multipla è almeno 6. Il voto dell'esame è la somma dei punteggi ottenuti ed è sufficiente se è almeno 18.
- **Materiale:**
<http://www.unife.it/scienze/biologia/Insegnamenti/matematica-a-a-2018-2019/materiale-didattico/>

Introduzione

La Statistica é un settore della Matematica Applicata di supporto a varie discipline. Lo studio di un fenomeno (per esempio di tipo biologico) conduce spesso all'analisi di informazioni espresse sotto forma di dati.

Introduzione

La Statistica é un settore della Matematica Applicata di supporto a varie discipline. Lo studio di un fenomeno (per esempio di tipo biologico) conduce spesso all'analisi di informazioni espresse sotto forma di dati. La Statistica fornisce concetti e strumenti per

- **organizzare** e **riassumere** in modo significativo i dati raccolti per evidenziare gli aspetti rilevanti ivi contenuti e descrivere quindi le caratteristiche del campione o dell'intera popolazione. Questo é il compito della **Statistica descrittiva**;

Introduzione

La Statistica é un settore della Matematica Applicata di supporto a varie discipline. Lo studio di un fenomeno (per esempio di tipo biologico) conduce spesso all'analisi di informazioni espresse sotto forma di dati. La Statistica fornisce concetti e strumenti per

- **organizzare** e **riassumere** in modo significativo i dati raccolti per evidenziare gli aspetti rilevanti ivi contenuti e descrivere quindi le caratteristiche del campione o dell'intera popolazione. Questo é il compito della **Statistica descrittiva**;
- **fare delle previsioni** o dire qualcosa (ossia **inferire**) sull'intera popolazione. Questo é il compito della **Statistica inferenziale**.

ESEMPIO

Le elezioni richiedono l'operazione di contare voti, calcolare percentuali, descrivere la composizione dell'organo eletto. Questi compiti sono di competenza della Statistica descrittiva.

ESEMPIO

Le elezioni richiedono l'operazione di contare voti, calcolare percentuali, descrivere la composizione dell'organo eletto. Questi compiti sono di competenza della Statistica descrittiva. Elaborare i risultati di un sondaggio per capire come si orienterà il voto dell'intera popolazione é di competenza della Statistica inferenziale.

Popolazione

L'indagine statistica viene compiuta su una **popolazione**: essa può essere **reale** (per esempio un insieme di oggetti o di persone) o può essere **virtuale**, per esempio l'insieme di tutte le possibili repliche di un esperimento.

Popolazione

L'indagine statistica viene compiuta su una **popolazione**: essa può essere **reale** (per esempio un insieme di oggetti o di persone) o può essere **virtuale**, per esempio l'insieme di tutte le possibili repliche di un esperimento. Chiameremo **unità statistica** ogni singolo elemento che costituisce la popolazione.

Popolazione

L'indagine statistica viene compiuta su una **popolazione**: essa può essere **reale** (per esempio un insieme di oggetti o di persone) o può essere **virtuale**, per esempio l'insieme di tutte le possibili repliche di un esperimento. Chiameremo **unità statistica** ogni singolo elemento che costituisce la popolazione.

Un sottoinsieme della popolazione si dice **campione** ed é scelto mediante una selezione detta **campionamento**.

Popolazione

L'indagine statistica viene compiuta su una **popolazione**: essa può essere **reale** (per esempio un insieme di oggetti o di persone) o può essere **virtuale**, per esempio l'insieme di tutte le possibili repliche di un esperimento. Chiameremo **unità statistica** ogni singolo elemento che costituisce la popolazione.

Un sottoinsieme della popolazione si dice **campione** ed é scelto mediante una selezione detta **campionamento**. Quando si esaminano tutte le unità si parla invece di **censimento**. Attenzione ai campionamenti!!!

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

- **qualitativa** se le modalità sono espresse "**in forma verbale**": per esempio

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

- **qualitativa** se le modalità sono espresse "**in forma verbale**": per esempio
 - **livello di istruzione**: diploma di scuola media, diploma di scuola media superiore, laurea triennale, laurea magistrale, master, dottorato...;

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

- **qualitativa** se le modalità sono espresse "**in forma verbale**": per esempio
 - **livello di istruzione**: diploma di scuola media, diploma di scuola media superiore, laurea triennale, laurea magistrale, master, dottorato...;
 - **gruppo sanguigno**: A,B,0,AB;

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

- **qualitativa** se le modalità sono espresse "**in forma verbale**": per esempio
 - **livello di istruzione**: diploma di scuola media, diploma di scuola media superiore, laurea triennale, laurea magistrale, master, dottorato...;
 - **gruppo sanguigno**: A,B,0,AB;
 - **colore degli occhi**: celeste, verde, grigio, castano, nero...;

Tipi di variabili

Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

DEFINIZIONE

*Una **variabile** é una caratteristica delle unità statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...

Una variabile può assumere una molteplicità di valori detti **modalità** ed in base alle modalità una variabile si dice

- **qualitativa** se le modalità sono espresse **"in forma verbale"**: per esempio
 - **livello di istruzione**: diploma di scuola media, diploma di scuola media superiore, laurea triennale, laurea magistrale, master, dottorato...;
 - **gruppo sanguigno**: A,B,0,AB;
 - **colore degli occhi**: celeste, verde, grigio, castano, nero...;
- **quantitativa** se le modalità sono espresse **in forma numerica**: **altezza, peso, lunghezza**...

Una variabile **qualitativa** si dice

- **ordinale** se esiste un ordinamento naturale delle modalità (e.g.: livello di istruzione);

Una variabile **qualitativa** si dice

- **ordinale** se esiste un ordinamento naturale delle modalità (e.g.: livello di istruzione);
- **sconnessa** in caso contrario (e.g.: gruppo sanguigno, colore degli occhi).

Una variabile **qualitativa** si dice

- **ordinale** se esiste un ordinamento naturale delle modalità (e.g.: livello di istruzione);
- **sconnessa** in caso contrario (e.g.: gruppo sanguigno, colore degli occhi).

Una variabile **quantitativa** si dice

- **discreta** se le possibili modalità sono un insieme numerico finito o al più numerabile (e.g.: voto di laurea, voto di maturità...);
- **continua** se le possibili modalità sono un insieme numerico continuo (e.g.: l'altezza, il peso).

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

DEFINIZIONE

Chiameremo **variabile statistica** ogni singola rilevazione della variabile di interesse su un campione.

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

DEFINIZIONE

Chiameremo **variabile statistica** ogni singola rilevazione della variabile di interesse su un campione.

Quindi la stessa variabile rilevata su campioni diversi dà luogo, in genere, a variabili statistiche differenti.

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

DEFINIZIONE

Chiameremo **variabile statistica** ogni singola rilevazione della variabile di interesse su un campione.

Quindi la stessa variabile rilevata su campioni diversi dà luogo, in genere, a variabili statistiche differenti. Per esempio scegliamo come variabile l'altezza dei ragazzi di terza media; in una classe potrebbero esserci molti ripetenti oppure molti anticipatari; i dati raccolti differirebbero molto.

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

DEFINIZIONE

Chiameremo **variabile statistica** ogni singola rilevazione della variabile di interesse su un campione.

Quindi la stessa variabile rilevata su campioni diversi dà luogo, in genere, a variabili statistiche differenti. Per esempio scegliamo come variabile l'altezza dei ragazzi di terza media; in una classe potrebbero esserci molti ripetenti oppure molti anticipatari; i dati raccolti differirebbero molto. Indicheremo con $Y = (y_1, y_2, \dots, y_N)$ la generica variabile statistica discreta (dove N è la dimensione del campione).

ESEMPIO

A 20 studenti iscritti al primo anno di università si chiede il voto riportato alla maturità. La variabile é il *Voto di maturità* e supponiamo che i valori rilevati sul campione di studenti siano

(76, 90, 70, 74, 80, 84, 78, 80, 70, 94, 76, 82, 74, 80, 70, 76, 76, 94, 100, 82).

Frequenza ass

Frequenza rel

media

ESEMPIO

A 20 studenti iscritti al primo anno di università si chiede il voto riportato alla maturità. La variabile é il *Voto di maturità* e supponiamo che i valori rilevati sul campione di studenti siano

(76, 90, 70, 74, 80, 84, 78, 80, 70, 94, 76, 82, 74, 80, 70, 76, 76, 94, 100, 82).

Frequenza ass

Frequenza rel

media

Questa stringa é la nostra variabile statistica Y . $N = 20$ é la dimensione del campione.

ESEMPIO

A 20 studenti iscritti al primo anno di università si chiede il voto riportato alla maturità. La variabile é il *Voto di maturità* e supponiamo che i valori rilevati sul campione di studenti siano

(76, 90, 70, 74, 80, 84, 78, 80, 70, 94, 76, 82, 74, 80, 70, 76, 76, 94, 100, 82).

Frequenza ass

Frequenza rel

media

Questa stringa é la nostra variabile statistica Y . $N = 20$ é la dimensione del campione. La variabile di interesse può assumere potenzialmente tutti i valori da 60 a 100 mentre la nostra variabile statistica Y (ossia la variabile "voto" testata sul nostro campione) assume i valori distinti

$$\{z_1, z_2, \dots, z_{10}\} = \{70, 74, 76, 78, 80, 82, 84, 90, 94, 100\}.$$

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti.

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

- si prende una sola volta ogni singola modalità;

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

- si prende una sola volta ogni singola modalità;
- si conta "quante volte compare" ogni singola modalità.

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

- si prende una sola volta ogni singola modalità;
- si conta "quante volte compare" ogni singola modalità.

DEFINIZIONE

*Sia Y una variabile statistica. Si dice **frequenza assoluta** di una modalità il numero di volte che tale modalità viene osservata.*

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

- si prende una sola volta ogni singola modalità;
- si conta "quante volte compare" ogni singola modalità.

DEFINIZIONE

*Sia Y una variabile statistica. Si dice **frequenza assoluta** di una modalità il numero di volte che tale modalità viene osservata.*

Quindi se indichiamo le modalità tutte tra loro distinte riscontrate sul campione

$$\{z_1, z_2, \dots, z_k\}$$

con $k \leq N$,

Frequenza assoluta

Quando il campione é molto numeroso (ossia N é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile Y con l'intera stringa $(y_1, y_2, \dots, y_N)$,

- si prende una sola volta ogni singola modalità;
- si conta "quante volte compare" ogni singola modalità.

DEFINIZIONE

*Sia Y una variabile statistica. Si dice **frequenza assoluta** di una modalità il numero di volte che tale modalità viene osservata.*

Quindi se indichiamo le modalità tutte tra loro distinte riscontrate sul campione

$$\{z_1, z_2, \dots, z_k\}$$

con $k \leq N$, usando le frequenze assolute associate a tali modalità, si costruiscono tabelle del tipo

Modalità	z_1	$\dots$	z_j	$\dots$	z_k
Frequenza ass.	f_1	$\dots$	f_j	$\dots$	f_k

Esempio.

In una indagine in cui esaminiamo il tipo di gruppo sanguigno di 200 persone, abbiamo che $N = 200$. Essendo però i gruppi sanguigni possibili solo 4, nella nostra indagine le modalità distinte riscontrate sul campione possono essere al più 4. Quindi

- se sul campione di $N = 200$ persone troviamo distribuiti tutti e 4 i gruppi sanguigni allora

$$\{z_1, z_2, z_3, z_k\} = \{A, B, AB, 0\};$$

- in teoria, in un campione potremmo anche trovare presente un solo gruppo sanguigno, per esempio tutti hanno il gruppo 0. In tal caso $k = 1$ e abbiamo solo

$$\{z_1\} = \{0\}.$$

- Se invece sul nostro campione troviamo presenti solo i gruppi 0 e A allora $k = 2$ e

$$\{z_1, z_2\} = \{A, 0\}.$$

Così via.

ESEMPIO

Considerando la variabile statistica dell' Esempio1 abbiamo

Voto	70	74	76	78	80	82	84	90	94	100	moda
Freq. ass	3	2	4	1	3	2	1	1	2	1	

ESEMPIO

Considerando la variabile statistica dell' Esempio1 abbiamo

Voto	70	74	76	78	80	82	84	90	94	100	moda
Freq. ass	3	2	4	1	3	2	1	1	2	1	

Se non avessimo saputo qual era la dimensione N del campione, da questa tabella ricaviamo $N = 3 + 2 + 4 + 1 + 3 + 2 + 1 + 1 + 2 + 1 = 20$ ossia

ESEMPIO

Considerando la variabile statistica dell' Esempio1 abbiamo

Voto	70	74	76	78	80	82	84	90	94	100	moda
Freq. ass	3	2	4	1	3	2	1	1	2	1	

Se non avessimo saputo qual era la dimensione N del campione, da questa tabella ricaviamo $N = 3 + 2 + 4 + 1 + 3 + 2 + 1 + 1 + 2 + 1 = 20$ ossia **la somma di tutte le frequenze assolute é uguale alla dimensione N del campione.**

Frequenza relativa e percentuale

DEFINIZIONE

*Sia Y una variabile statistica. Se f é la frequenza assoluta di una modalità ed N é la dimensione del campione, si dice **frequenza relativa** il rapporto*

$$p = \frac{f}{N}.$$

*La **frequenza percentuale** é la frequenza relativa moltiplicata per 100%*

Frequenza relativa e percentuale

DEFINIZIONE

*Sia Y una variabile statistica. Se f é la frequenza assoluta di una modalità ed N é la dimensione del campione, si dice **frequenza relativa** il rapporto*

$$p = \frac{f}{N}.$$

*La **frequenza percentuale** é la frequenza relativa moltiplicata per 100%*

Ovviamente la somma di tutte le frequenze relative é uguale ad 1 mentre la somma delle frequenze percentuali é 100%.

ESEMPIO

La variabile statistica dell'Esempio1 aveva frequenze assolute

Voto	70	74	76	78	80	82	84	90	94	100
Freq.ass	3	2	4	1	3	2	1	1	2	1

ESEMPIO

La variabile statistica dell' **Esempio1** aveva frequenze assolute

Voto	70	74	76	78	80	82	84	90	94	100
Freq.ass	3	2	4	1	3	2	1	1	2	1

Allora dividendo le frequenze assolute per

$$N = 3 + 2 + 4 + 3 + 2 + 1 + 1 + 2 + 1 = 20$$

si ottiene la tabella delle frequenze relative:

Voto	70	74	76	78	80	82	84	90	94	100
Freq.rel	0.15	0.1	0.2	0.05	0.15	0.1	0.05	0.05	0.1	0.05

moda

ESEMPIO

Data la tabella

Gruppo sanguigno	A	B	AB	0
Frequenza ass.	2	3	1	4

si dica lunghezza del campione e tabella di distribuzione delle frequenze relative.

ESEMPIO

Data la tabella

Gruppo sanguigno	A	B	AB	0
Frequenza ass.	2	3	1	4

si dica lunghezza del campione e tabella di distribuzione delle frequenze relative.

Il campione ha dimensione $N = 2 + 3 + 1 + 4 = 10$.

ESEMPIO

Data la tabella

Gruppo sanguigno	A	B	AB	0
Frequenza ass.	2	3	1	4

si dica lunghezza del campione e tabella di distribuzione delle frequenze relative.

Il campione ha dimensione $N = 2 + 3 + 1 + 4 = 10$.

Quindi la tabella di distribuzione delle frequenze relative é

Gruppo sanguigno	A	B	AB	0
Frequenza ass.	0.2	0.3	0.1	0.4

Come visto negli esami sopra, una distribuzione di frequenza si può rappresentare mediante una tabella del tipo

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

oppure

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

Come visto negli esami sopra, una distribuzione di frequenza si può rappresentare mediante una tabella del tipo

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

oppure

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

In alcuni casi (per esempio con variabili continue) é conveniente "**raggruppare**" le modalità in "classi contigue e disgiunte" e contare le unità che appartengono a ciascuna classe.

Classi

ESEMPIO

Da un'indagine condotta su 200 laureati in Biologia si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. ass.	42	50	60	32	12	4

moda

Classi

ESEMPIO

Da un'indagine condotta su 200 laureati in Biologia si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. ass.	42	50	60	32	12	4

moda Dividendo tutte le frequenze assolute per il numero $N = 200$ si ottiene la tabella:

Classi

ESEMPIO

Da un'indagine condotta su 200 laureati in Biologia si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. ass.	42	50	60	32	12	4

moda Dividendo tutte le frequenze assolute per il numero $N = 200$ si ottiene la tabella:

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

Frequenze usando le classi

- Si sceglie il **numero** di classi. Si tenga conto che se tale numero é troppo grande, le frequenze relative alle classi possono diventare troppo piccole; se é troppo piccolo, si perde troppa informazione.

Frequenze usando le classi

- Si sceglie il **numero** di classi. Si tenga conto che se tale numero é troppo grande, le frequenze relative alle classi possono diventare troppo piccole; se é troppo piccolo, si perde troppa informazione. La **regola di Sturges** suggerisce per N osservazioni un numero di classi pari a $1 + \frac{\log N}{\log 2}$.

Frequenze usando le classi

- Si sceglie il **numero** di classi. Si tenga conto che se tale numero é troppo grande, le frequenze relative alle classi possono diventare troppo piccole; se é troppo piccolo, si perde troppa informazione. La **regola di Sturges** suggerisce per N osservazioni un numero di classi pari a $1 + \frac{\log N}{\log 2}$.
- si scelgono **gli estremi delle classi** e si scrivono intervalli del tipo $[z_{j-1}, z_j)$;

Frequenze usando le classi

- Si sceglie il **numero** di classi. Si tenga conto che se tale numero é troppo grande, le frequenze relative alle classi possono diventare troppo piccole; se é troppo piccolo, si perde troppa informazione. La **regola di Sturges** suggerisce per N osservazioni un numero di classi pari a $1 + \frac{\log N}{\log 2}$.
- si scelgono **gli estremi delle classi** e si scrivono intervalli del tipo $[z_{j-1}, z_j)$;
- **si contano il numero di volte** in cui la variabile statistica assume valori compresi tra z_{j-1} (incluso) e z_j (escluso) per trovare la tabella di frequenza assolute di questo tipo:

Classi	$[z_1, z_2)$	$\cdots$	$[z_{j-1}, z_j)$	$\cdots$	$[z_k, z_{k+1}]$
Frequenza ass.	f_1	$\cdots$	f_{j-1}	$\cdots$	f_k

Per trovare la tabella di frequenze relative, occorre dividere le frequenze assolute per la dimensione N del campione che é data da

$$N = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^k f_i.$$

Per trovare la tabella di frequenze relative, occorre dividere le frequenze assolute per la dimensione N del campione che é data da

$$N = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^k f_i.$$

Quindi

$$p_j = \frac{f_j}{N} = \frac{f_j}{\sum_{i=1}^k f_i}.$$

Rappresentazione dei dati

Un primo approccio all'analisi di un insieme di dati é costituito dai metodi grafici. Per rappresentare le distribuzione di frequenza (di variabili quantitative o qualitative), si possono utilizzare

- i **grafici a bastoncini**: sull'asse delle x si posizionano le varie modalità e in corrispondenza di esse si tracciano dei bastoncini di altezza proporzionale o pari alla frequenza assoluta o relativa delle modalità; così si mettono in evidenza le modalità che hanno la frequenza (assoluta o relativa) più grande.

Grafici a bastoncini

Altezza ragazzi	157	160	165	168	173	176	184
Freq. ass.	1	2	6	3	4	3	1

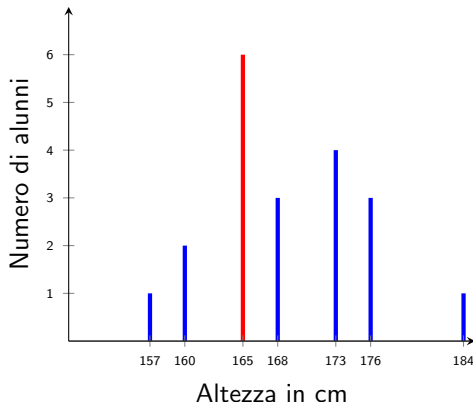


FIGURE: Altezza degli alunni di una classe di terza media

Grafici a poligonali

Quando si tracciano delle **poligonali** che uniscono gli estremi superiori dei bastoncini, si ottiene un grafico a poligonali:

Grafici a poligonal

Quando si tracciano delle **poligonal** che uniscono gli estremi superiori dei bastoncini, si ottiene un grafico a poligonal:

Altezza alunni	157	160	165	168	173	176	184
Freq. ass.	1	2	6	3	4	3	1

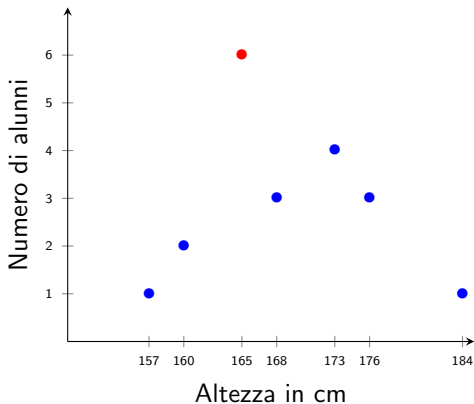


FIGURE: Altezza degli alunni di una classe di terza media

Grafici a poligonal

Quando si tracciano delle **poligonal** che uniscono gli estremi superiori dei bastoncini, si ottiene un grafico a poligonal:

Altezza alunni	157	160	165	168	173	176	184
Freq. ass.	1	2	6	3	4	3	1

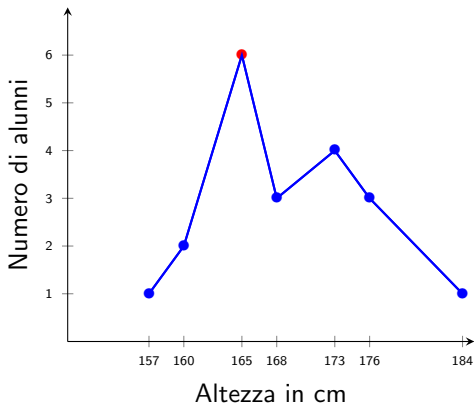


FIGURE: Altezza degli alunni di una classe di prima media

Grafici a barre o a colonne

Quando alle linee viene dato spessore fino a diventare dei rettangoli, si ottengono i **grafici a barre**: sull'asse delle x si posizionano le varie modalità e in corrispondenza ad esse si tracciano dei rettangoli di altezza proporzionale o pari alla frequenza assoluta o relativa delle modalità:

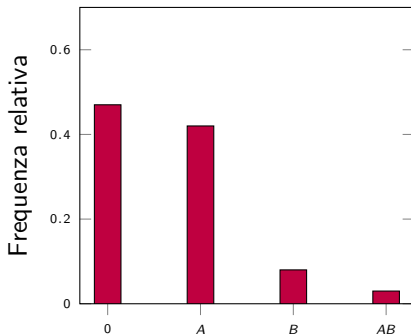


FIGURE: Diffusione dei gruppi sanguigni in Italia

Areogramma

Quando si tratta di visualizzare le diverse parti in cui un tutto é stato suddiviso, si ricorre spesso a un tipo di rappresentazione detto **areogramma** (o torta!).

Areogramma

Quando si tratta di visualizzare le diverse parti in cui un tutto é stato suddiviso, si ricorre spesso a un tipo di rappresentazione detto **areogramma** (o torta!).

Altezza alunni	158	160	165	168	173	176	184
Freq. ass.	1	2	6	3	4	3	1

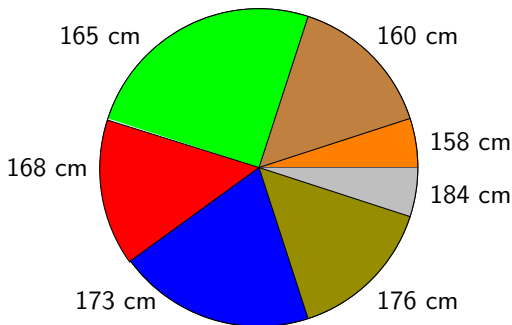


FIGURE: **Altezza alunni di III media**

Per convenzione le ampiezze dei settori circolari devono essere proporzionali alle grandezze delle corrispondenti parti.
La rappresentazione risulta efficace per mettere in evidenza i mutui rapporti tra le parti.

Supponiamo di avere delle frequenze percentuali $p_1, p_2, \dots, p_n$ con $p_1 + p_2 + \dots + p_n = 100\%$. Allora l'areogramma che rappresenta questi dati é un cerchio suddiviso in n settori circolari di ampiezza rispettivamente $p_1/100 \cdot 360^\circ, p_2/100 \cdot 360^\circ, \dots, p_n/100 \cdot 360^\circ$ in modo da riempire l'intero cerchio.

Istogrammi

Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**.

Istogrammi

Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**.

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. ass.	42	50	60	32	12	4

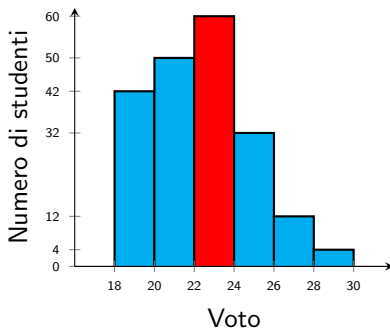


FIGURE: **Voto all'esame di Matematica**

Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**. Le altezze sono scelte in modo che le aree dei rettangoli costruiti siano uguali o proporzionali alle frequenze.

Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**. Le altezze sono scelte in modo che le aree dei rettangoli costruiti siano uguali o proporzionali alle frequenze.

Si osservi che se si considera il rettangolo R_j di base $[z_j, z_{j+1}]$ e altezza uguale a f_j , allora la sua area é data da

$$\text{base} \cdot \text{altezza} = (z_{j+1} - z_j) \cdot f_j.$$

Quindi se le classi $[z_j, z_{j+1})$ hanno la stessa ampiezza $(z_{j+1} - z_j) = b$, le aree dei rettangoli sono uguali a $b \cdot f_j$ ossia sono proporzionali alle frequenze assolute.

Se invece le classi $[z_j, z_{j+1})$ hanno ampiezze diverse, allo scopo di fare in modo che le aree rimangano proporzionali alle frequenze, si scelgono le altezze dei rettangoli uguali (o proporzionali) a $f_j/(z_j - z_{j-1})$. In tal modo le aree dei rettangoli sono uguali (o proporzionali) a

$$\frac{f_j}{(z_j - z_{j-1})} \cdot (z_j - z_{j-1}) f_j = f_j.$$

Stesso discorso se al posto delle frequenze assolute si considerano le frequenze relative.

A partire dall'istogramma si può costruire il **poligono di frequenza**:
basta tracciare la spezzata che congiunge i punti medi dei lati superiori
dei rettangoli dell'istogramma.

La moda.

Questo tipo di rappresentazioni permette subito di individuare la modalità piú "riscontrata" sul campione, la cosiddetta **moda**.

La moda.

Questo tipo di rappresentazioni permette subito di individuare la modalità più "riscontrata" sul campione, la cosiddetta **moda**.

DEFINIZIONE

*Sia Y una variabile qualitativa o quantitativa. La **moda** é la modalità con frequenza (relativa o assoluta) più alta, ossia la modalità rilevata maggiormente sulla popolazione. Se le modalità sono raggruppate in classi della stessa ampiezza si chiama **classe modale** la classe con il numero maggiore di elementi.*

La moda.

Questo tipo di rappresentazioni permette subito di individuare la modalità più "riscontrata" sul campione, la cosiddetta **moda**.

DEFINIZIONE

*Sia Y una variabile qualitativa o quantitativa. La **moda** é la modalità con frequenza (relativa o assoluta) più alta, ossia la modalità rilevata maggiormente sulla popolazione. Se le modalità sono raggruppate in classi della stessa ampiezza si chiama **classe modale** la classe con il numero maggiore di elementi.*

ESEMPIO

Considerando la variabile statistica

- dell'Esempio **Frequenze ass** la moda é 76;
- dell'Esempio **Classi contigue** la classe modale é la classe [18, 20).

Allo scopo di riassumere i dati, si introducono degli indici sintetici che

- descrivono intorno a quale valore si concentrano i dati (**indici di posizione o di centralità**);

Allo scopo di riassumere i dati, si introducono degli indici sintetici che

- descrivono intorno a quale valore si concentrano i dati (**indici di posizione o di centralità**); un indice di posizione che abbiamo già visto é la **moda** moda.

Allo scopo di riassumere i dati, si introducono degli indici sintetici che

- descrivono intorno a quale valore si concentrano i dati (**indici di posizione o di centralità**); un indice di posizione che abbiamo già visto é la **moda** moda. Altri indici di posizione sono la **media** e la **mediana** che ora introdurremo;

Allo scopo di riassumere i dati, si introducono degli indici sintetici che

- descrivono intorno a quale valore si concentrano i dati (**indici di posizione o di centralità**); un indice di posizione che abbiamo già visto é la **moda** moda. Altri indici di posizione sono la **media** e la **mediana** che ora introdurremo;
- come si disperdono i dati (**indici di dispersione**): introdurremo la **varianza** e lo **deviazione standard**.

Media

Supponiamo di avere raccolto un numero finito di dati $y_1, y_2, \dots, y_N$. Vogliamo trovare un numero $\bar{y}$ che li riassuma. Vediamo se esiste un $\bar{y}$ tale che renda nulla la somma degli errori ossia tale che

$$\sum_{i=1}^N (\bar{y} - y_i) = 0.$$

Media

Supponiamo di avere raccolto un numero finito di dati $y_1, y_2, \dots, y_N$. Vogliamo trovare un numero $\bar{y}$ che li riassuma. Vediamo se esiste un $\bar{y}$ tale che renda nulla la somma degli errori ossia tale che

$$\sum_{i=1}^N (\bar{y} - y_i) = 0.$$

Otteniamo così

$$N\bar{y} - \sum_{i=1}^N y_i = 0$$

da cui

Media

Supponiamo di avere raccolto un numero finito di dati $y_1, y_2, \dots, y_N$. Vogliamo trovare un numero $\bar{y}$ che li riassuma. Vediamo se esiste un $\bar{y}$ tale che renda nulla la somma degli errori ossia tale che

$$\sum_{i=1}^N (\bar{y} - y_i) = 0.$$

Otteniamo così

$$N\bar{y} - \sum_{i=1}^N y_i = 0$$

da cui

$$\bar{y} = \frac{\sum_{i=1}^N y_i}{N}.$$

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica discreta. Si chiama **media o media aritmetica o valor medio** di Y la quantità

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{y_1 + y_2 + \dots + y_N}{N}.$$

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica discreta. Si chiama **media o media aritmetica o valor medio** di Y la quantità

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{y_1 + y_2 + \dots + y_N}{N}.$$

ESEMPIO

Supponiamo di misurare la lunghezza dell'assone della cellula neurale di un ratto. Ripetiamo 10 volte l'operazione di misurare e otteniamo i seguenti risultati espressi in micrometri:

70.78, 74.22, 74.03, 71.71, 70.97, 73.47, 69.28, 69.62, 72.31, 72.76.

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica discreta. Si chiama **media o media aritmetica o valor medio** di Y la quantità

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{y_1 + y_2 + \dots + y_N}{N}.$$

ESEMPIO

Supponiamo di misurare la lunghezza dell'assone della cellula neurale di un ratto. Ripetiamo 10 volte l'operazione di misurare e otteniamo i seguenti risultati espressi in micrometri:

70.78, 74.22, 74.03, 71.71, 70.97, 73.47, 69.28, 69.62, 72.31, 72.76.

La media delle lunghezze dell'assone della cellula neurale é

$$\frac{70.78+74.22+74.03+71.71+70.97+73.47+69.28+69.62+72.31+72.76}{10} = 71.195. \quad \text{mediana}$$

Calcolo della media usando le frequenze

Se é nota la distribuzione delle freq. assolute di Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Calcolo della media usando le frequenze

Se é nota la distribuzione delle freq. assolute di Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

allora la dimensione del campione é

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

Calcolo della media usando le frequenze

Se é nota la distribuzione delle freq. assolute di Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

allora la dimensione del campione é

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

e quindi

$$\bar{Y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum_{j=1}^k f_j z_j}{\sum_{j=1}^k f_j}.$$

ESEMPIO

Da un'indagine condotta su un gruppo di studenti in Biologia si riscontra il seguente grado di soddisfazione (espresso con voto da 0 a 5) sul corso di Matematica:

Voto	0	1	2	3	4	5
Frequenza ass.	2	8	40	24	10	16

ESEMPIO

Da un'indagine condotta su un gruppo di studenti in Biologia si riscontra il seguente grado di soddisfazione (espresso con voto da 0 a 5) sul corso di Matematica:

Voto	0	1	2	3	4	5	Allora
Frequenza ass.	2	8	40	24	10	16	

$$N = 2 + 8 + 40 + 24 + 10 + 16 = 100$$

e il grado medio di soddisfazione é:

ESEMPIO

Da un'indagine condotta su un gruppo di studenti in Biologia si riscontra il seguente grado di soddisfazione (espresso con voto da 0 a 5) sul corso di Matematica:

Voto	0	1	2	3	4	5	Allora
Frequenza ass.	2	8	40	24	10	16	

$$N = 2 + 8 + 40 + 24 + 10 + 16 = 100$$

e il grado medio di soddisfazione é:

$$\frac{0 \cdot 2 + 1 \cdot 8 + 2 \cdot 40 + 3 \cdot 24 + 4 \cdot 10 + 5 \cdot 16}{100} = 2.8.$$

Se é nota la distribuzione delle freq. relative della Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Freq. rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

Se é nota la distribuzione delle freq. relative della Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Freq. rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

se N la dimensione del campione (che in generale potrebbe essere non nota),

Se é nota la distribuzione delle freq. relative della Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Freq. rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

se N la dimensione del campione (che in generale potrebbe essere non nota), si ha che

$$p_i = \frac{f_i}{N} \implies f_i = Np_i$$

Se é nota la distribuzione delle freq. relative della Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Freq. rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

se N la dimensione del campione (che in generale potrebbe essere non nota), si ha che

$$p_i = \frac{f_i}{N} \implies f_i = Np_i$$

e quindi

$$\bar{Y} = \frac{\sum_{j=1}^k f_j z_j}{N} = \frac{\sum_{j=1}^k Np_j z_j}{N}$$

Se é nota la distribuzione delle freq. relative della Y

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Freq. rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

se N la dimensione del campione (che in generale potrebbe essere non nota), si ha che

$$p_i = \frac{f_i}{N} \implies f_i = Np_i$$

e quindi

$$\bar{Y} = \frac{\sum_{j=1}^k f_j z_j}{N} = \frac{\sum_{j=1}^k Np_j z_j}{N}$$

ossia

$$\bar{Y} = \sum_{j=1}^k p_j z_j.$$

ESEMPIO

Se consideriamo la tabella relativa ad Esempio1

Voto	70	74	76	78	80	82	84	90	94	100
Freq.rel	0.15	0.1	0.2	0.05	0.15	0.1	0.05	0.05	0.1	0.05

per calcolare la media

ESEMPIO

Se consideriamo la tabella relativa ad Esempio1

Voto	70	74	76	78	80	82	84	90	94	100
Freq.rel	0.15	0.1	0.2	0.05	0.15	0.1	0.05	0.05	0.1	0.05

per calcolare la media

$$\bar{Y} = 70 \cdot 0.15 + 74 \cdot 0.1 + 76 \cdot 0.2 + 78 \cdot 0.05 + 80 \cdot 0.15 +$$

$$82 \cdot 0.1 + 84 \cdot 0.05 + 90 \cdot 0.05 + 94 \cdot 0.1 + 100 \cdot 0.05 = 79.94$$

Formule nel caso di classi

Se le modalità sono state raggruppate in classi e si ha una delle tabelle

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Formule nel caso di classi

Se le modalità sono state raggruppate in classi e si ha una delle tabelle

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

per definire $\bar{Y}$ si sostituisce ad ogni classe il valore medio $\bar{z}_j = \frac{z_j + z_{j+1}}{2}$.

Formule nel caso di classi

Se le modalità sono state raggruppate in classi e si ha una delle tabelle

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

per definire $\bar{Y}$ si sostituisce ad ogni classe il valore medio $\bar{z}_j = \frac{z_j + z_{j+1}}{2}$. Si applicano quindi le formule già viste alla tabelle

Classi	$\bar{z}_1$	$\cdots$	$\bar{z}_j$	$\cdots$	$\bar{z}_k$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Formule nel caso di classi

Se le modalità sono state raggruppate in classi e si ha una delle tabelle

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

per definire $\bar{Y}$ si sostituisce ad ogni classe il valore medio $\bar{z}_j = \frac{z_j + z_{j+1}}{2}$. Si applicano quindi le formule già viste alla tabelle

Classi	$\bar{z}_1$	$\cdots$	$\bar{z}_j$	$\cdots$	$\bar{z}_k$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

Formule nel caso di classi

Se le modalità sono state raggruppate in classi e si ha una delle tabelle

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

Classi	$[z_1, z_2)$	$\cdots$	$[z_j, z_{j+1})$	$\cdots$	$[z_k, z_{k+1})$
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

per definire $\bar{Y}$ si sostituisce ad ogni classe il valore medio $\bar{z}_j = \frac{z_j + z_{j+1}}{2}$. Si applicano quindi le formule già viste alla tabelle

Classi	$\bar{z}_1$	$\cdots$	$\bar{z}_j$	$\cdots$	$\bar{z}_k$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k
Frequenza rel.	p_1	$\cdots$	p_j	$\cdots$	p_k

$$\Rightarrow \bar{Y} = \frac{\sum_{i=1}^k f_j \bar{z}_j}{\sum_{j=1}^k f_i} \text{ oppure } \bar{Y} = \sum_{i=1}^k p_j \bar{z}_j.$$

ESEMPIO

Con riferimento all'indagine su un gruppo di laureati in Biologia sul voto Y conseguito all'esame di Matematica si aveva la tabella

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

ESEMPIO

Con riferimento all'indagine su un gruppo di laureati in Biologia sul voto Y conseguito all'esame di Matematica si aveva la tabella

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

Per calcolare il voto medio $\bar{Y}$ conseguito all'esame di Matematica, sostituiamo ad ogni classe il suo valore medio:

Voto	19	21	23	24	27	29
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

ESEMPIO

Con riferimento all'indagine su un gruppo di laureati in Biologia sul voto Y conseguito all'esame di Matematica si aveva la tabella

Voto	[18, 20)	[20, 22)	[22, 24)	[24, 26)	[26, 28)	[28, 30]
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

Per calcolare il voto medio $\bar{Y}$ conseguito all'esame di Matematica, sostituiamo ad ogni classe il suo valore medio:

Voto	19	21	23	24	27	29
Freq. rel.	0.21	0.25	0.3	0.16	0.06	0.02

Quindi, usando le frequenze relative, si ottiene che:

$$\begin{aligned}\bar{Y} &= 0.21 \cdot 19 + 0.25 \cdot 21 + 0.3 \cdot 23 + 0.16 \cdot 25 + 0.06 \cdot 27 + 0.02 \cdot 29 = \\ &= 3.99 + 5.25 + 6.9 + 4 + 1.62 + 0.58 = 22.34\end{aligned}$$

Proprietà della media:

- 1 Se Y é una variabile statistica discreta e $\{z_1, z_2, \dots, z_k\}$ sono le sue modalità con $z_1 < z_2 < \dots < z_k$ allora

$$z_1 \leq \bar{Y} \leq z_k$$

ossia la media é compresa tra il più piccolo e il più grande dei valori osservati.

Proprietà della media:

- Se Y é una variabile statistica discreta e $\{z_1, z_2, \dots, z_k\}$ sono le sue modalità con $z_1 < z_2 < \dots < z_k$ allora

$$z_1 \leq \bar{Y} \leq z_k$$

ossia la media é compresa tra il più piccolo e il più grande dei valori osservati. Basta infatti osservare che

$$z_1 = \left(\sum_{j=1}^k p_j\right)z_1 = \sum_{j=1}^k p_j z_1 \leq \sum_{j=1}^k p_j z_j = \bar{Y} \leq \sum_{j=1}^k p_j z_k = z_k.$$

Proprietà della media:

- ① Se Y é una variabile statistica discreta e $\{z_1, z_2, \dots, z_k\}$ sono le sue modalità con $z_1 < z_2 < \dots < z_k$ allora

$$z_1 \leq \bar{Y} \leq z_k$$

ossia la media é compresa tra il più piccolo e il più grande dei valori osservati. Basta infatti osservare che

$$z_1 = \left(\sum_{j=1}^k p_j\right)z_1 = \sum_{j=1}^k p_j z_1 \leq \sum_{j=1}^k p_j z_j = \bar{Y} \leq \sum_{j=1}^k p_j z_k = z_k.$$

- ② Se $X = (x_1, x_2, \dots, x_n)$ e $Y = (z_1, z_2, \dots, z_k)$ sono due variabile statistiche discrete e

$$Z = (x_1, x_2, \dots, x_n, z_1, z_2, \dots, z_k)$$

é quella ottenuta riunendo i due campioni allora

$$\bar{Z} = \frac{n\bar{X} + k\bar{Y}}{n + k}.$$

Esempio

Supponiamo di avere un campione con $n = 2$ studenti che hanno preso in media 24 all'esame di Matematica, mentre in un campione di $k = 10$ studenti tutti hanno preso 30.

Esempio

Supponiamo di avere un campione con $n = 2$ studenti che hanno preso in media 24 all'esame di Matematica, mentre in un campione di $k = 10$ studenti tutti hanno preso 30. Qual é la media del voto all'esame di Matematica sul campione di $n + k = 2 + 10 = 12$ studenti?

Esempio

Supponiamo di avere un campione con $n = 2$ studenti che hanno preso in media 24 all'esame di Matematica, mentre in un campione di $k = 10$ studenti tutti hanno preso 30. Qual é la media del voto all'esame di Matematica sul campione di $n + k = 2 + 10 = 12$ studenti? La formula corretta non é

$$\frac{24 + 30}{2} = 27$$

Esempio

Supponiamo di avere un campione con $n = 2$ studenti che hanno preso in media 24 all'esame di Matematica, mentre in un campione di $k = 10$ studenti tutti hanno preso 30. Qual é la media del voto all'esame di Matematica sul campione di $n + k = 2 + 10 = 12$ studenti? La formula corretta non é

$$\frac{24 + 30}{2} = 27$$

ma

$$\frac{2 \cdot 24 + 10 \cdot 30}{12} = 29.$$

Media ponderata!!!

In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \dots, y_N$ che tenga conto che ogni elemento y_i ha un'"importanza" diversa, un "peso" diverso p_i .

Media ponderata!!!

In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \dots, y_N$ che tenga conto che ogni elemento y_i ha un'"importanza" diversa, un "peso" diverso p_i . In tal caso la media si chiama **ponderata** si calcola come

$$\frac{\sum_{i=1}^N p_i \cdot y_i}{\sum_{i=1}^N p_i}.$$

Media ponderata!!!

In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \dots, y_N$ che tenga conto che ogni elemento y_i ha un'"importanza" diversa, un "peso" diverso p_i . In tal caso la media si chiama **ponderata** si calcola come

$$\frac{\sum_{i=1}^N p_i \cdot y_i}{\sum_{i=1}^N p_i}.$$

Per esempio la media degli esami all'università é una media ponderata, ossia calcolata dando ad ogni esame un peso che é dato dal numero di crediti dell'esame.

Media ponderata!!!

In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \dots, y_N$ che tenga conto che ogni elemento y_i ha un'"importanza" diversa, un "peso" diverso p_i . In tal caso la media si chiama **ponderata** si calcola come

$$\frac{\sum_{i=1}^N p_i \cdot y_i}{\sum_{i=1}^N p_i}.$$

Per esempio la media degli esami all'università é una media ponderata, ossia calcolata dando ad ogni esame un peso che é dato dal numero di crediti dell'esame. I crediti di un esame non rappresentano solo quante ore sono dedicate alla frequenza dell'esame ma sono quindi anche il "peso" nel calcolo del voto di presentazione alla tesi di laurea.

Media ponderata!!!

In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \dots, y_N$ che tenga conto che ogni elemento y_i ha un'"importanza" diversa, un "peso" diverso p_i . In tal caso la media si chiama **ponderata** si calcola come

$$\frac{\sum_{i=1}^N p_i \cdot y_i}{\sum_{i=1}^N p_i}.$$

Per esempio la media degli esami all'università é una media ponderata, ossia calcolata dando ad ogni esame un peso che é dato dal numero di crediti dell'esame. I crediti di un esame non rappresentano solo quante ore sono dedicate alla frequenza dell'esame ma sono quindi anche il "peso" nel calcolo del voto di presentazione alla tesi di laurea.

Per esempio Matematica = 6CFU. La media (per il calcolo del voto di presentazione alla laurea) é data da

$$\frac{\sum_{i=1}^N (\text{CFU dell'i-mo esame}) \cdot (\text{voto dell'i-mo esame})}{\text{numero totale dei crediti degli esami}}$$

dove

N = numero di esami

La mediana.

La media risente della presenza di osservazioni anomale. Un solo dato sensibilmente diverso dagli altri può spostare la media in modo significativo. Quando i dati non sono equamente distribuiti attorno alla media, può essere utile riassumerli con un altro valore: la mediana.

La mediana.

La media risente della presenza di osservazioni anomale. Un solo dato sensibilmente diverso dagli altri può spostare la media in modo significativo. Quando i dati non sono equamente distribuiti attorno alla media, può essere utile riassumerli con un altro valore: la mediana. Essa si può calcolare per una variabile qualitativa ordinale o quantitativa Y e si indica con $y_{0.5}$.

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica con

$$y_1 \leq y_2 \leq \dots \leq y_N.$$

La **mediana** corrisponde a quel valore $y_{0.5}$ tale che esattamente **metà dei dati siano minori o uguale a $y_{0.5}$** e **metà maggiori o uguale a $y_{0.5}$** .

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica con

$$y_1 \leq y_2 \leq \dots \leq y_N.$$

La **mediana** corrisponde a quel valore $y_{0.5}$ tale che esattamente **metà dei dati siano minori o uguale a $y_{0.5}$** e **metà maggiori o uguale a $y_{0.5}$** .

Quindi $y_{0.5}$ é il dato che si trova "a metà " (non é la metà!!).

Per calcolarla, vanno messi in ordine crescente gli N dati raccolti sul campione e

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica con

$$y_1 \leq y_2 \leq \dots \leq y_N.$$

La **mediana** corrisponde a quel valore $y_{0.5}$ tale che esattamente **metà dei dati siano minori o uguale a $y_{0.5}$** e **metà maggiori o uguale a $y_{0.5}$** .

Quindi $y_{0.5}$ é il dato che si trova "a metà " (non é la metà!!).

Per calcolarla, vanno messi in ordine crescente gli N dati raccolti sul campione e

- se N é **dispari** si prende il dato centrale (corrispondente al posto $\frac{N+1}{2}$)

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica con

$$y_1 \leq y_2 \leq \dots \leq y_N.$$

La **mediana** corrisponde a quel valore $y_{0.5}$ tale che esattamente **metà dei dati siano minori o uguale a $y_{0.5}$** e **metà maggiori o uguale a $y_{0.5}$** .

Quindi $y_{0.5}$ é il dato che si trova "a metà" (non é la metà!!).

Per calcolarla, vanno messi in ordine crescente gli N dati raccolti sul campione e

- se N é **dispari** si prende il dato centrale (corrispondente al posto $\frac{N+1}{2}$)
- se N é **pari** si prende la media aritmetica dei dati centrali ossia si fa la media dei dati che occupano il posto $\frac{N}{2}$ e $\frac{N}{2} + 1$.

DEFINIZIONE

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica con

$$y_1 \leq y_2 \leq \dots \leq y_N.$$

La **mediana** corrisponde a quel valore $y_{0.5}$ tale che esattamente **metà dei dati siano minori o uguale a $y_{0.5}$** e **metà maggiori o uguale a $y_{0.5}$** .

Quindi $y_{0.5}$ é il dato che si trova "a metà" (non é la metà!!).

Per calcolarla, vanno messi in ordine crescente gli N dati raccolti sul campione e

- se **N é dispari** si prende il dato centrale (corrispondente al posto $\frac{N+1}{2}$)
- se **N é pari** si prende la media aritmetica dei dati centrali ossia si fa la media dei dati che occupano il posto $\frac{N}{2}$ e $\frac{N}{2} + 1$.

La mediana non é influenzata dai valori estremi anomali (anche se dipende dal numero di dati raccolti) ed é poco sensibile ad errori dei dati. Nel gergo statistico si dice che é un indice di posizione *robusto*.

Esempi

- Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

Esempi

- Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03

quattro dati quinto posto quattro dati

Esempi

- Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03

quattro dati quinto posto quattro dati

Quindi la mediana é data dal valore 71.71.

Esempi

- Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

$69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03$
 quattro dati quinto posto quattro dati

Quindi la mediana é data dal valore 71.71.

- Se facciamo un'altra misurazione e otteniamo il dato 76, N diventa 10 (pari). Quindi

$69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03, 76.$
 quattro dati quinto posto sesto posto quattro dati

Esempi

- Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

$69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03$
 quattro dati quinto posto quattro dati

Quindi la mediana é data dal valore 71.71.

- Se facciamo un'altra misurazione e otteniamo il dato 76, N diventa 10 (pari). Quindi

$69.28, 69.62, 70.78, 70.97, 71.71, 72.31, 72.76, 73.47, 74.03, 76.$
 quattro dati quinto posto sesto posto quattro dati

Quindi la mediana é data dalla media

$$\frac{71.71 + 72.31}{2} = 72,01$$

e anche se l'ultimo dato é molto grande, la mediana non subisce grosse variazioni.

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il 101^{mo} dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra.

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il 101^{mo} dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra. Il 18 compare 30 volte, il 19 compare 25 volte, il 20 compare 25 volte e il 21 compare 20 volte.

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il 101^{mo} dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra. Il 18 compare 30 volte, il 19 compare 25 volte, il 20 compare 25 volte e il 21 compare 20 volte. Poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il 101^{mo} dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra. Il 18 compare 30 volte, il 19 compare 25 volte, il 20 compare 25 volte e il 21 compare 20 volte. Poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

il 101^{mo} dato "cade" nella modalità 22.

Se sono note le freq. assolute con N dispari.

Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	3

Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il 101^{mo} dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra. Il 18 compare 30 volte, il 19 compare 25 volte, il 20 compare 25 volte e il 21 compare 20 volte. Poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

il 101^{mo} dato "cade" nella modalità 22. Quindi la mediana è 22.

Se sono note le freq. assolute con N pari.

ESEMPIO

Da un'indagine condotta su $N = 200$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F. ass.	30	25	25	22	21	19	10	18	12	8	5	3	2

Se sono note le freq. assolute con N pari.

ESEMPIO

Da un'indagine condotta su $N = 200$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F. ass.	30	25	25	22	21	19	10	18	12	8	5	3	2

Per calcolare la mediana, dobbiamo prendere il 100^{mo} e il 101^{mo} dato e farne la media.

Se sono note le freq. assolute con N pari.

ESEMPIO

Da un'indagine condotta su $N = 200$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F. ass.	30	25	25	22	21	19	10	18	12	8	5	3	2

Per calcolare la mediana, dobbiamo prendere il 100^{mo} e il 101^{mo} dato e farne la media. Poiché

$$30 + 25 + 25 = 80 \text{ e } 30 + 25 + 25 + 22 = 102$$

il dato 100^{mo} e il 101^{mo} hanno modalità 21.

Se sono note le freq. assolute con N pari.

ESEMPIO

Da un'indagine condotta su $N = 200$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F. ass.	30	25	25	22	21	19	10	18	12	8	5	3	2

Per calcolare la mediana, dobbiamo prendere il 100^{mo} e il 101^{mo} dato e farne la media. Poiché

$$30 + 25 + 25 = 80 \text{ e } 30 + 25 + 25 + 22 = 102$$

il dato 100^{mo} e il 101^{mo} hanno modalità 21. Quindi la mediana é

$$\frac{21 + 21}{2} = 21$$

Se la tabella delle frequenze é invece

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	2

Se la tabella delle frequenze é invece

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30
F.ass.	30	25	25	20	24	18	10	18	12	8	5	3	2

poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

il dato 100^{mo} ha modalità 21 mentre il dato 101^{mo} ha modalità 22.
Quindi la mediana é $\frac{21+22}{2} = 21.5$.

La varianza

Se $Y = (y_1, y_2, \dots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

La varianza

Se $Y = (y_1, y_2, \dots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

DEFINIZIONE

*Si chiama **varianza** di Y la media degli scarti quadratici $(y_i - \bar{Y})^2$.*

La varianza

Se $Y = (y_1, y_2, \dots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

DEFINIZIONE

*Si chiama **varianza** di Y la media degli scarti quadratici $(y_i - \bar{Y})^2$.*

La varianza di Y viene indicata con σ_Y^2 o più semplicemente con σ^2 .

La varianza

Se $Y = (y_1, y_2, \dots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

DEFINIZIONE

*Si chiama **varianza** di Y la media degli scarti quadratici $(y_i - \bar{Y})^2$.*

La varianza di Y viene indicata con σ_Y^2 o più semplicemente con σ^2 .
Dalla definizione si ha

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N \underbrace{(y_i - \bar{Y})^2}_{\text{scarti quadratici}} . \quad (1)$$

La varianza

Se $Y = (y_1, y_2, \dots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

DEFINIZIONE

*Si chiama **varianza** di Y la media degli scarti quadratici $(y_i - \bar{Y})^2$.*

La varianza di Y viene indicata con σ_Y^2 o più semplicemente con σ^2 .
Dalla definizione si ha

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N \underbrace{(y_i - \bar{Y})^2}_{\text{scarti quadratici}} . \quad (1)$$

Tanto più piccola é la varianza di Y tanto più i dati sono concentrati intorno al suo valor medio $\bar{Y}$.

DEFINIZIONE

Si definisce **deviazione standard** o **scarto quadratico medio** di Y la radice quadrata della sua varianza.

DEFINIZIONE

Si definisce **deviazione standard** o **scarto quadratico medio** di Y la radice quadrata della sua varianza.

Pertanto la deviazione standard viene indicata con σ_Y .

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N}}.$$

DEFINIZIONE

Si definisce **deviazione standard** o **scarto quadratico medio** di Y la radice quadrata della sua varianza.

Pertanto la deviazione standard viene indicata con σ_Y .

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N}}.$$

Nell'ambito della statistica inferenziale, a volte si rimpiazza il denominatore N con $N - 1$ ottenendo:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N - 1}}.$$

Noi useremo sempre la formula con N a denominatore.

Esempio

In un esperimento vengono pesate 15 cavia ottenendo i seguenti pesi in grammi:

28, 32, 37, 29, 31, 30, 26, 32, 27, 29, 30, 32, 28, 31, 31.

Costruire la tabella della frequenze assolute e calcolare media, mediana, varianza e deviazione standard.

Esempio

In un esperimento vengono pesate 15 cavia ottenendo i seguenti pesi in grammi:

28, 32, 37, 29, 31, 30, 26, 32, 27, 29, 30, 32, 28, 31, 31.

Costruire la tabella della frequenze assolute e calcolare media, mediana, varianza e deviazione standard.

Risoluzione. Costruiamo la tabella delle frequenze assolute:

Peso (g)	26	27	28	29	30	31	32	37
F.ass.	1	1	2	2	2	3	3	1

Esempio

In un esperimento vengono pesate 15 cavie ottenendo i seguenti pesi in grammi:

28, 32, 37, 29, 31, 30, 26, 32, 27, 29, 30, 32, 28, 31, 31.

Costruire la tabella della frequenze assolute e calcolare media, mediana, varianza e deviazione standard.

Risoluzione. Costruiamo la tabella delle frequenze assolute:

Peso (g)	26	27	28	29	30	31	32	37
F.ass.	1	1	2	2	2	3	3	1

Quindi

$$\bar{Y} = \frac{26 \cdot 1 + 27 \cdot 1 + 28 \cdot 2 + 29 \cdot 2 + 30 \cdot 2 + 31 \cdot 3 + 32 \cdot 3 + 37 \cdot 1}{15} = 30.2g$$

Per la mediana i dati ordinati sono

$\underbrace{26, 27, 28, 28, 29, 29, 30}_{\text{sette dati}}, \underbrace{30}_{\text{ottavo posto}}, \underbrace{31, 31, 31, 32, 32, 32, 37}_{\text{sette dati}}$

la mediana é 30.

Per usare la formula (1) (senza usare le frequenze) dobbiamo calcolare tutti gli scarti quadratici:

$$(28 - 30.2)^2, (32 - 30.2)^2, (37 - 30.2)^2, \dots (31 - 30.2)^2, (31 - 30.2)^2$$

ossia

Per usare la formula (1) (senza usare le frequenze) dobbiamo calcolare tutti gli scarti quadratici:

$$(28 - 30.2)^2, (32 - 30.2)^2, (37 - 30.2)^2, \dots (31 - 30.2)^2, (31 - 30.2)^2$$

ossia

$$4.84, 3.24, 46.24, \dots, 0.64, 0.64$$

Per usare la formula (1) (senza usare le frequenze) dobbiamo calcolare tutti gli scarti quadratici:

$$(28 - 30.2)^2, (32 - 30.2)^2, (37 - 30.2)^2, \dots (31 - 30.2)^2, (31 - 30.2)^2$$

ossia

$$4.84, 3.24, 46.24, \dots, 0.64, 0.64$$

da cui segue che

$$\sigma^2 = \frac{4.84 + 3.24 + 46.24 \dots + 0.64 + 0.64}{15} g^2 = 6.56 g^2.$$

Per usare la formula (1) (senza usare le frequenze) dobbiamo calcolare tutti gli scarti quadratici:

$$(28 - 30.2)^2, (32 - 30.2)^2, (37 - 30.2)^2, \dots (31 - 30.2)^2, (31 - 30.2)^2$$

ossia

$$4.84, 3.24, 46.24, \dots, 0.64, 0.64$$

da cui segue che

$$\sigma^2 = \frac{4.84 + 3.24 + 46.24 \dots + 0.64 + 0.64}{15} g^2 = 6.56 g^2.$$

Quindi la deviazione standard é $\sigma = 2.56g$.

Osservazioni sulle formula

Se usiamo la tabella con le frequenze assolute

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

dopo aver calcolato la media $\bar{Y}$, possiamo costruire la tabella

Scarti quadratici	$(z_1 - \bar{Y})^2$	$\cdots$	$(z_j - \bar{Y})^2$	$\cdots$	$(z_k - \bar{Y})^2$
Frequenza ass.	f_1	$\cdots$	f_j	$\cdots$	f_k

e quindi calcolare la varianza con la formula

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (z_i - \bar{Y})^2}{N}.$$

Nell'esempio precedente

Peso	26	27	28	29	30	31	32	37
Scarti quadratici	17.64	10.24	4.84	...	...	...	3.24	46.24
F.ass.	1	1	2	2	2	3	3	1

Nell'esempio precedente

Peso	26	27	28	29	30	31	32	37
Scarti quadratici	17.64	10.24	4.84	...	...	...	3.24	46.24
F.ass.	1	1	2	2	2	3	3	1

e quindi

$$\sigma^2 = \frac{1 \cdot 17.64 + 10.24 \cdot 1 + 4.84 \cdot 2 + \dots + 3.24 \cdot 3 + 46.24 \cdot 1}{15} = 6.56.$$

Formula alternativa

Sviluppando i quadrati in (1) si trova:

$$\frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2$$

Formula alternativa

Sviluppando i quadrati in (1) si trova:

$$\frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2 = \frac{1}{N} \sum_{i=1}^N (y_i^2 - 2y_i \bar{Y} + \bar{Y}^2)$$

Formula alternativa

Sviluppando i quadrati in (1) si trova:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2 &= \frac{1}{N} \sum_{i=1}^N (y_i^2 - 2y_i \bar{Y} + \bar{Y}^2) \\ &= \frac{1}{N} \left(\sum_{i=1}^N y_i^2 - 2 \sum_{i=1}^N y_i \bar{Y} + \sum_{i=1}^N \bar{Y}^2 \right) = \frac{1}{N} \left(\sum_{i=1}^N y_i^2 - 2N\bar{Y} \cdot \bar{Y} + N\bar{Y}^2 \right) \end{aligned}$$

Formula alternativa

Sviluppando i quadrati in (1) si trova:

$$\begin{aligned}
 \frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y})^2 &= \frac{1}{N} \sum_{i=1}^N (y_i^2 - 2y_i \bar{Y} + \bar{Y}^2) \\
 &= \frac{1}{N} \left(\sum_{i=1}^N y_i^2 - 2 \sum_{i=1}^N y_i \bar{Y} + \sum_{i=1}^N \bar{Y}^2 \right) = \frac{1}{N} \left(\sum_{i=1}^N y_i^2 - 2N\bar{Y} \cdot \bar{Y} + N\bar{Y}^2 \right) \\
 \sigma^2 &= \overline{Y^2} - \bar{Y}^2 (\geq 0). \tag{2}
 \end{aligned}$$

Quindi la varianza é la **media dei quadrati meno il quadrato della media**.

Quindi, allo scopo di usare la formula (2), si devono calcolare

- la media di Y per poi farne il quadrato: otteniamo quindi $\overline{Y^2}$
- scrivere la tabella di Y^2 calcolando i quadrati delle singole modalità
- calcolare la media di Y^2 , ossia $\overline{Y^2}$.

Quindi, allo scopo di usare la formula (2), si devono calcolare

- la media di Y per poi farne il quadrato: otteniamo quindi $\overline{Y}^2$
- scrivere la tabella di Y^2 calcolando i quadrati delle singole modalità
- calcolare la media di Y^2 , ossia $\overline{Y^2}$.

Nell'esempio precedente:

$Y = \text{Peso}$	26	27	28	29	30	31	32	37
$Y^2 = \text{Peso}^2$	676	729	784	841	900	961	1024	1369
F.ass.	1	1	2	2	2	3	3	1

Quindi, allo scopo di usare la formula (2), si devono calcolare

- la media di Y per poi farne il quadrato: otteniamo quindi $\bar{Y}^2$
- scrivere la tabella di Y^2 calcolando i quadrati delle singole modalità
- calcolare la media di Y^2 , ossia $\overline{Y^2}$.

Nell'esempio precedente:

$Y = \text{Peso}$	26	27	28	29	30	31	32	37
$Y^2 = \text{Peso}^2$	676	729	784	841	900	961	1024	1369
F.ass.	1	1	2	2	2	3	3	1

- la media di Y era $\bar{Y} = 30.2g$

Quindi, allo scopo di usare la formula (2), si devono calcolare

- la media di Y per poi farne il quadrato: otteniamo quindi $\bar{Y}^2$
- scrivere la tabella di Y^2 calcolando i quadrati delle singole modalità
- calcolare la media di Y^2 , ossia $\overline{Y^2}$.

Nell'esempio precedente:

$Y = \text{Peso}$	26	27	28	29	30	31	32	37
$Y^2 = \text{Peso}^2$	676	729	784	841	900	961	1024	1369
F.ass.	1	1	2	2	2	3	3	1

- la media di Y era $\bar{Y} = 30.2g$
- la media di Y^2 é allora

$$\overline{Y^2} = \frac{676 \cdot 1 + 729 \cdot 1 + \dots + 1024 \cdot 3 + 1369 \cdot 1}{15} = 918.6g^2$$

Quindi, allo scopo di usare la formula (2), si devono calcolare

- la media di Y per poi farne il quadrato: otteniamo quindi $\bar{Y}^2$
- scrivere la tabella di Y^2 calcolando i quadrati delle singole modalità
- calcolare la media di Y^2 , ossia $\overline{Y^2}$.

Nell'esempio precedente:

$Y = \text{Peso}$	26	27	28	29	30	31	32	37
$Y^2 = \text{Peso}^2$	676	729	784	841	900	961	1024	1369
F.ass.	1	1	2	2	2	3	3	1

- la media di Y era $\bar{Y} = 30.2g$
- la media di Y^2 é allora

$$\overline{Y^2} = \frac{676 \cdot 1 + 729 \cdot 1 + \dots + 1024 \cdot 3 + 1369 \cdot 1}{15} = 918.6g^2$$

Quindi

$$\sigma^2 = \overline{Y^2} - (\bar{Y})^2 = 918.06 - (30.2)^2 = 918.06 - 912.04 = 6.56g^2$$

Se invece usiamo la tabella con le frequenze relative

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	p_1	$\cdots$	p_j	$\cdots$	p_k

si ottiene

$$\sigma^2 = \sum_{i=1}^k p_i (z_i - \bar{Y})^2.$$

Se invece usiamo la tabella con le frequenze relative

Modalità	z_1	$\cdots$	z_j	$\cdots$	z_k
Frequenza ass.	p_1	$\cdots$	p_j	$\cdots$	p_k

si ottiene

$$\sigma^2 = \sum_{i=1}^k p_i (z_i - \bar{Y})^2.$$

Se infine Y é una variabile statistica continua e si dispone della tabella di frequenza con le modalità raggruppate in classi $[z_i, z_{i+1})$ per definire σ^2 é necessario calcolare il punto medio $\bar{z}_i = \frac{z_i + z_{i+1}}{2}$ di ogni classe. Quindi

$$\sigma^2 = \sum_{i=1}^k p_i (\bar{z}_i - \bar{Y})^2$$

dove p_i é la frequenza associata alla classe $[z_i, z_{i+1})$.

DEFINITION

Una variabile Y tale che $\bar{Y} = 0$ si dice **centrata**. Una variabile Y tale che $\bar{Y} = 0$ e $\sigma^2 = 1$ si dice **standardizzata**.

DEFINITION

Una variabile Y tale che $\bar{Y} = 0$ si dice **centrata**. Una variabile Y tale che $\bar{Y} = 0$ e $\sigma^2 = 1$ si dice **standardizzata**.

Data la variabile Y allora $Z = \frac{Y - \bar{Y}}{\sigma_Y}$ é una variabile standardizzata.