Statistica Descrittiva

Giorgio Bertolazzi, Ph.D.

Ricercatore dell'Universita degli Studi di Enna Kore

UNIVERSITÀ
DEGLI STUDI
DI ENNA "KORE"

Argomenti del corso:

Statistica descrittiva:

- Variabili casuali
- Misure e scale di misura
- Distribuzioni di frequenze
- Misure di sintesi
- Misure di variabilità
- Misure di associazione
- Test diagnostici

Variabili casuali

Una **variabile** è una **caratteristica** di un'unità (ad esempio, una persona, un oggetto o un evento) che può assumere diversi valori possibili (*realizzazioni*).

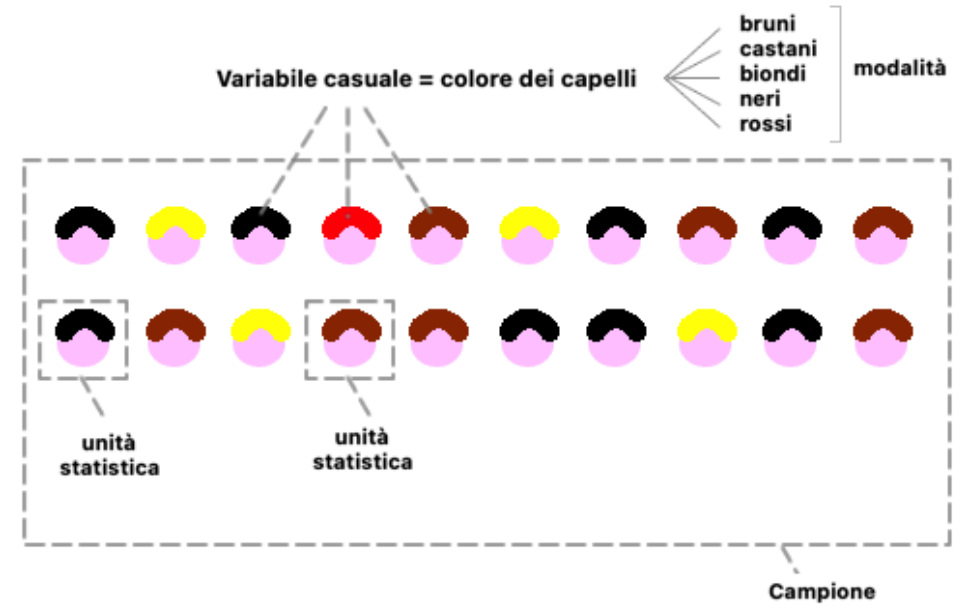
Parliamo di **variabile casuale** quando i valori di una variabile non possono essere conosciuti prima della loro rilevazione.

Variabili casuali

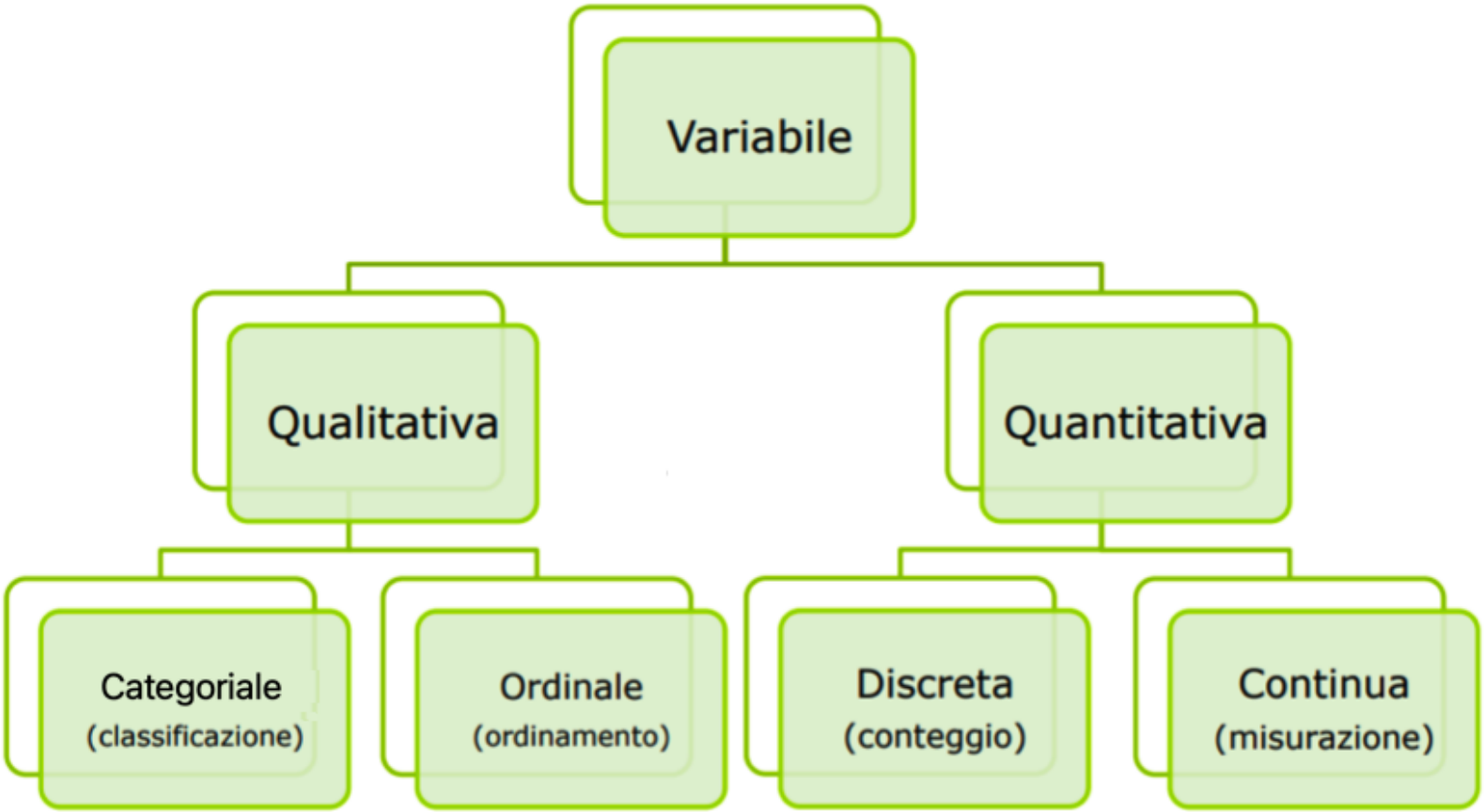
Una **variabile** è una **caratteristica** di un'unità (ad esempio, una persona, un oggetto o un evento) che può assumere diversi valori possibili (*realizzazioni*).

Nel caso di variabili non numeriche, le possibili realizzazioni della variabile sono chiamate **modalità**.

Le **unità statistiche** sono i singoli elementi sui quali vengono osservate le realizzazioni delle variabili casuali.



Variabili casuali



Variabili casuali

Una variabile si dice **qualitativa** quando le sue determinazioni sono espresse mediante delle modalità non numeriche.

- Una variabile qualitativa si dice **categoriale** se le sue modalità non sono ordinabili;

  Esempi: *Gruppo sanguigno, abitudine al fumo, tipo di trattamento*

- Una variabile qualitativa si dice **ordinale** se è possibile stabilire un ordinamento delle sue modalità;

  Esempi: *Stadio della malattia, livello di istruzione, rischio cardiovascolare*

X = *Genere*

$$X = \begin{cases} 0 & \textit{se maschio} \\ 1 & \textit{se femmina} \end{cases}$$

Y = *Stadio della malattia*

$$Y = \begin{cases} 1 & \textit{se precoce} \\ 2 & \textit{se intermedio} \\ 3 & \textit{se avanzato} \end{cases}$$

Variabili casuali quantitative

Una variabile si dice **quantitativa** quando le sue determinazioni sono espresse mediante valori numerici.

Esempi: *Età del paziente, peso corporeo, livelli di emoglobina nel sangue*

Possibili classificazioni delle variabili quantitative:

- **Classificazione basata sul dominio**
  - **Variabile discreta** ($X \in \mathbb{N}$)

    Esempi: *numero di figli, numero di eventi di infarto, numero di pazienti ricoverati*
  - **Variabile continua** ($X \in \mathbb{R}$)

    Esempi: *temperatura corporea, livello di glucosio, indice di massa corporea*

Variabili casuali quantitative

- **Classificazione basata sulla scala di misura**
  - **Scala di intervalli**

    Permette di ordinare i dati e di misurare le differenze tra i valori tramite calcoli di addizione e sottrazione.

    Esempi: *temperatura in gradi celsius, livello di pH*

  - **Scala di rapporti**

    Permettere l'ordinamento dei dati e la misurazione delle differenze. Il valore zero rappresenta l'assenza della caratteristica misurata, questo permette di confrontare i valori in termini di rapporti.

    Esempi: *numero di figli, livello di glucosio, indice di massa corporea*

Variabili casuali

Poniamo di avere rilevato il numero medio mensile di nascite di tutti gli ospedali siciliani:

- Quali sono le unità statistiche in esame?
- Qual'è la variabile di interesse?
- Di che tipologia di variabile si tratta? Quali valori può assumere?

Poniamo di avere rilevato la pressione sanguigna da un gruppo di pazienti:

- Quali sono le unità statistiche in esame?
- Qual'è la variabile di interesse?
- Di che tipologia di variabile si tratta? Quali valori può assumere?

Variabili casuali

Poniamo di avere rilevato il numero medio mensile di nascite di tutti gli ospedali siciliani:

- Quali sono le unità statistiche in esame? (ospedali siciliani)
- Qual'è la variabile di interesse? (numero medio mensile di nascite)
- Di che tipologia di variabile si tratta? Quali valori può assumere? (variabile quantitativa discreta su scala di rapporti)

Poniamo di avere rilevato la pressione sanguigna da un gruppo di pazienti:

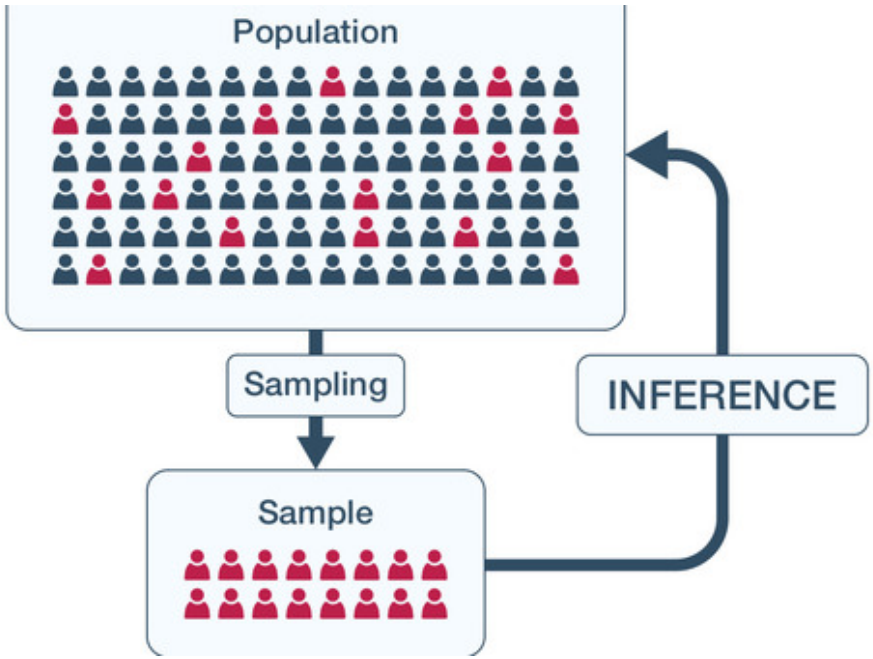
- Quali sono le unità statistiche in esame? (pazienti)
- Qual'è la variabile di interesse? (pressione sanguigna)
- Di che tipologia di variabile si tratta? Quali valori può assumere? (variabile quantitativa continua su scala di rapporti)

Campionamento e inferenza statistica

La **popolazione** è l'insieme di tutte le unità che condividono la caratteristica oggetto di studio.

Il **campione** è un sottoinsieme della popolazione, selezionato per rappresentarla.

L'**inferenza** è il processo mediante il quale si traggono conclusioni riguardanti una popolazione sulla base dei dati raccolti da un campione.



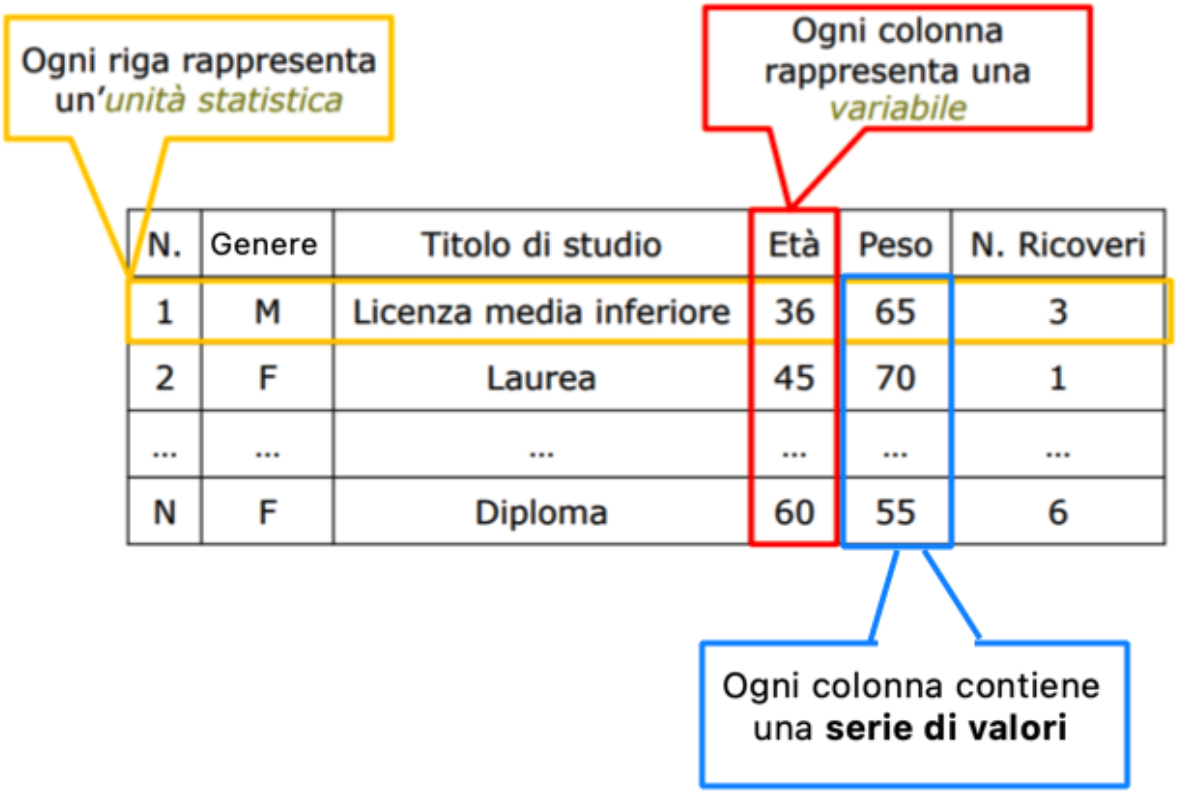
Matrice dei dati (dataset)

N.	Genere	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
...	...	...	...	...	...
N	F	Diploma	60	55	6

Un dataset è costituito da un insieme di dati organizzati in forma tabellare:

- Ogni riga corrisponde ad un unità statistica
- Ogni colonna corrisponde a una variabile

Matrice dei dati (dataset)



N.	Genere	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
...	...	...	...	...	...
N	F	Diploma	60	55	6

L'insieme delle realizzazioni di una variabile casuale costituisce una **serie di valori**.

Distribuzione di frequenze

Ogni riga rappresenta un'*unità statistica*

Ogni colonna rappresenta una *variabile*

N.	Genere	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
...	...	...	...	...	...
N	F	Diploma	60	55	6

Distribuzione di frequenze della variabile "titolo di studio"

Titolo di studio	Frequenze assolute
Licenza media inferiore	5
Diploma	32
Laurea	20
Dottorato	3

Ogni riga corrisponde a una modalità o a una classe

Ogni tabella di frequenze corrisponde a una singola variabile

Distribuzione di frequenze

Ogni riga rappresenta un'*unità statistica*

Ogni colonna rappresenta una *variabile*

N.	Genere	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
...	...	...	...	...	...
N	F	Diploma	60	55	6

Distribuzione di frequenze della variabile "titolo di studio"

Titolo di studio	Frequenze assolute	Freq. relative	Freq. %
Licenza media inferiore	5	0.07	7 %
Diploma	32	0.46	46 %
Laurea	20	0.43	43 %
Dottorato	3	0.04	4 %
	70	1	100 %

Totale delle frequenze

Distribuzione di frequenze

X	Frequenza assoluta	Frequenze relativa	Frequenza percentuale
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
x_3	n_3	f_3	p_3
x_4	n_4	f_4	p_4
	n	1	100

La **distribuzione di frequenze** associa alle modalità che può assumere una variabile le corrispondenti frequenze assolute e relative.

La **frequenza assoluta** n_j della modalità x_j è il numero di volte in cui questa modalità viene osservata nel dataset in esame.

Il parametro n indica il totale delle osservazioni: $n = \sum_{j=1}^{K} n_j$

Probabilità

La **probabilità** è una misura quantitativa della possibilità che un evento si verifichi.

Se si considera un evento E, la probabilità di E è definita come:

$$P(E) = \frac{\text{Numero di casi favorevoli}}{\text{Numero di casi possibili}}$$

Lancio un dado. Qual'è la probabilità che esca il numero 2?

Lancio una moneta. Qual'è la probabilità che esca testa?

Pesco una carta da un mazzo siciliano. Qual'è la probabilità che esca un re?

Probabilità

La **probabilità** è una misura quantitativa della possibilità che un evento si verifichi.

Se si considera un evento E, la probabilità di E è definita come:

$$P(E) = \frac{\text{Numero di casi favorevoli}}{\text{Numero di casi possibili}}$$

Lancio un dado. Qual'è la probabilità che esca il numero 2?

Lancio una moneta. Qual'è la probabilità che esca testa?

Pesco una carta da un mazzo siciliano. Qual'è la probabilità che esca un re?

Probabilità

La **probabilità** è una misura quantitativa della possibilità che un evento si verifichi.

Se si considera un evento E, la probabilità di E è definita come:

$$P(E) = \frac{\text{Numero di casi favorevoli}}{\text{Numero di casi possibili}}$$

Lancio un dado. Qual'è la probabilità che esca il numero 2? (1/6)

Lancio una moneta. Qual'è la probabilità che esca testa? (1/2)

Pesco una carta da un mazzo siciliano. Qual'è la probabilità che esca un re? (4/40)

Distribuzione di frequenze

La **frequenza relativa** di una caratteristica è il rapporto tra il numero di volte in cui questa caratteristica è stata osservata nel campione e il numero totale di osservazioni.

$$f_j = \frac{n_j}{n}$$

Esempio: Ho osservato un campione di 70 individui, 32 dei quali hanno un grado di istruzione pari al diploma. La Frequenza relativa di diplomati è pari a:

$$f_2 = \frac{n_2}{n} = \frac{32}{70} = 0.46$$

La frequenza relativa corrisponde alla **probabilità stimata** di osservare una caratteristica (**stima empirica** della probabilità ottenuta a partire dai dati osservati).

Distribuzione di frequenze

La **frequenza relativa** di una caratteristica è il rapporto tra il numero di volte in cui questa caratteristica è stata osservata nel campione e il numero totale di osservazioni.

$$f_j = \frac{n_j}{n}$$

Esempio: Ho osservato un campione di 70 individui, 32 dei quali hanno un grado di istruzione pari al diploma. La Frequenza relativa di diplomati è pari a:

$$f_2 = \frac{n_2}{n} = \frac{32}{70} = 0.46$$

La frequenza relativa corrisponde alla **probabilità stimata** di osservare una caratteristica (**stima empirica** della probabilità ottenuta a partire dai dati osservati).

Distribuzione di frequenze

Calcolare la distribuzione di frequenze della variabile $X =$ *gravità della malattia* considerando i seguenti dati:

Unità	Valore
1	lieve
2	lieve
3	lieve
4	lieve
5	moderata
6	moderata
7	moderata
8	moderata
9	moderata
10	moderata
11	moderata
12	moderata
13	grave
14	grave
15	grave
16	grave
17	grave
18	grave
19	molto grave
20	molto grave

Distribuzioni di frequenza

Calcolare la distribuzione di frequenze della variabile $X =$ *gravità della malattia* considerando i seguenti dati:

Unità	Valore
1	lieve
2	lieve
3	lieve
4	lieve
5	moderata
6	moderata
7	moderata
8	moderata
9	moderata
10	moderata
11	moderata
12	moderata
13	grave
14	grave
15	grave
16	grave
17	grave
18	grave
19	molto grave
20	molto grave

X	Frequenza assoluta	Frequenze relativa	Frequenza percentuale
Lieve	4	0.2	20%
Moderata	8	0.4	40%
Grave	6	0.3	30%
Molto grave	2	0.1	10%
	$n = 20$	1	100%

Discretizzazione di variabili continue

La distribuzione di frequenze di una variabile continua è ottenuta ripartendo i valori numerici in **classi**.

Per esempio, considerando la variabile $X = età$ avente i seguenti valori osservati:

$X = (30, 25, 18, 16, 14, 28, 32, 48, 43, 42, 36, 32, 41, 51, 68, 66, 55, 92, 81, 75)$

Posso ottenere la seguente distribuzione di frequenze:

Classi di Età	Freq. assoluta	Freq. relativa	Freq. percentuale
≤30	6	0.30	30%
31-50	7	0.35	35%
51-70	4	0.20	20%
≥71	3	0.15	15%
Totale	$n = 20$	1	100%

Distribuzione cumulata

Classi di Età	Freq. assoluta	Freq. assoluta cumulata	Freq. relativa cumulata	Freq. percentuale cumulata
≤30	6	6	0.30	30%
31-50	7	13	0.65	56%
51-70	4	17	0.85	85%
≥71	3	20	1	100%
Totale	$n = 20$			

La **distribuzione cumulata delle frequenze assolute** mostra il numero totale di osservazioni che cadono al di sotto di un certo valore.

Esempio: 13 osservazioni presentano un'età inferiore a 51 anni.

Distribuzione cumulata

Classi di Età	Freq. assoluta	Freq. assoluta cumulata	Freq. relativa cumulata	Freq. percentuale cumulata
≤30	6	6	0.30	30%
31-50	7	13	0.65	56%
51-70	4	17	0.85	85%
≥71	3	20	1	100%
Totale	$n = 20$			

La **distribuzione cumulata delle frequenze relative/percentuali** mostra la frazione/percentuale di osservazioni che cadono al di sotto di un certo valore.

Esempio: il 56% osservazioni presentano un'età inferiore a 51 anni.

La stima della probabilità di osservare un'età inferiore a 51 è pari a 0.56.

Distribuzione cumulata

X	Freq. assoluta	Freq. assoluta cumulata	Freq. relativa cumulata	Freq. percentuale cumulata
x_1	n_1	N_1	F_1	P_1
x_2	n_2	N_2	F_2	P_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_j	n_j	N_j	F_j	P_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K	n_K	$N_K = n$	$F_K = 1$	$P_K = 100$

Freq. assoluta cumulata: $$N_j = \sum_{i=1}^{j} n_i = n_1 + n_2 + \cdots + n_j$$

Freq. relativa cumulata: $$F_j = \frac{N_j}{n}$$

Distribuzione cumulata

Completare la seguente distribuzione di frequenze:

X	Freq. assoluta	Freq. relativa	Freq. percentuale	Freq. cum. assoluta	Freq. cum. relativa	Freq. cum. percentuale
Lieve	4	0.2	20%			
Moderata	8	0.4	40%			
Grave	6	0.3	30%			
Molto grave	2	0.1	10%			
	$n = 20$	1	100%			

Qual'è la frequenza assoluta di pazienti che presentano una malattia la cui gravità è inferiore o uguale alla modalità "moderata"?

Qual'è la percentuale di pazienti con una malattia la cui gravità è *al più* "moderata"?

Qual'è la probabilità si osservare un paziente con una malattia la cui gravità è inferiore o uguale alla modalità "grave"?

Distribuzione cumulata

Distribuzione di frequenze della variabile *gravità della malattia*:

X	Freq. assoluta	Freq. relativa	Freq. percentuale	Freq. cum. assoluta	Freq. cum. relativa	Freq. cum. percentuale
Lieve	4	0.2	20%	4	0.2	20%
Moderata	8	0.4	40%	12	0.6	60%
Grave	6	0.3	30%	18	0.9	90%
Molto grave	2	0.1	10%	20	1	100%
	$n = 20$	1	100%			

Qual'è la frequenza assoluta di pazienti che presentano una malattia la cui gravità è inferiore o uguale alla modalità "moderata"? (12)

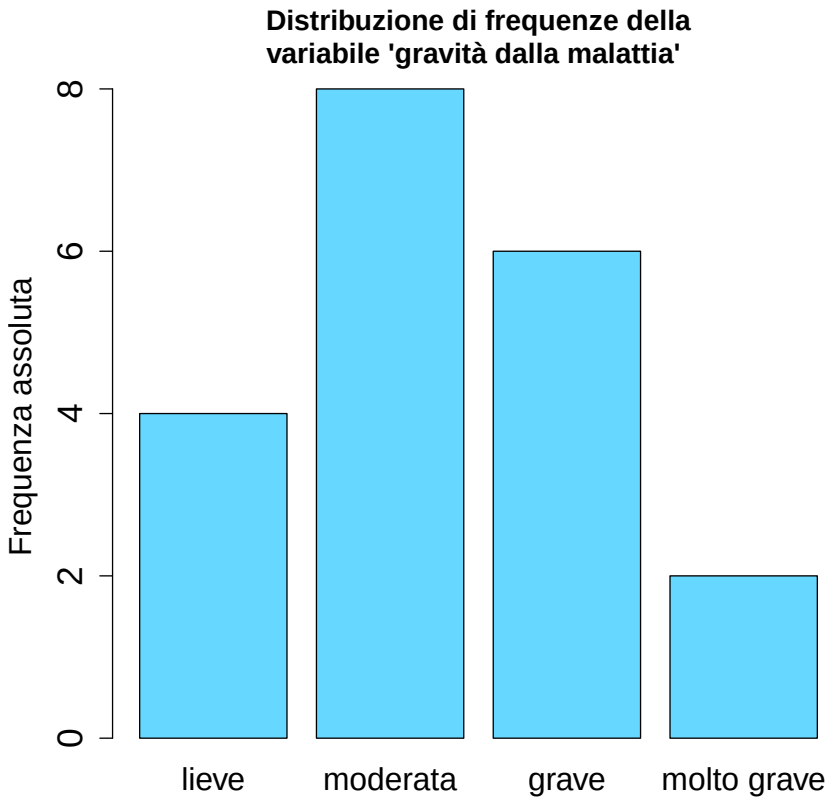
Qual'è la percentuale di pazienti con una malattia la cui gravità è *al più* "moderata"? (60%)

Qual'è la probabilità si osservare un paziente con una malattia la cui gravità è inferiore o uguale alla modalità "grave"? (0.9)

Grafico a barre

Il **grafico a barre** (**barplot**) è utilizzato per rappresentare la distribuzione di frequenza di una variabile categoriale.

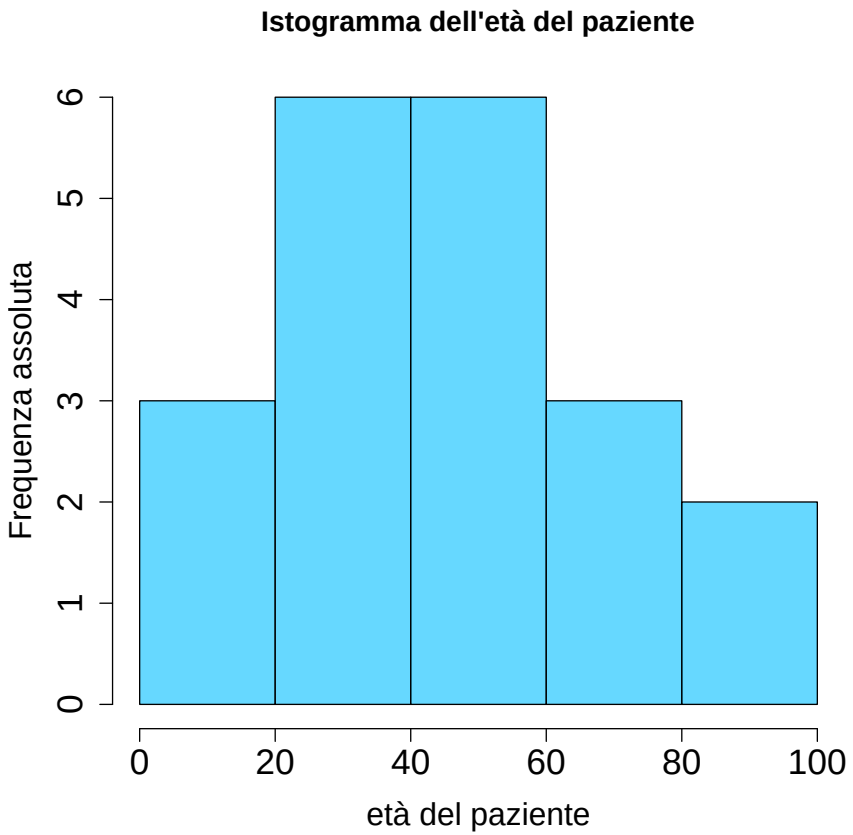
X	Freq. assoluta
Lieve	4
Moderata	8
Grave	6
Molto grave	2
	$n = 20$



Istogramma

L'**istogramma** rappresenta la distribuzione di frequenze di una variabile quantitativa.

Se la variabile quantitativa possiede in dominio continuo, la rappresentazione tramite l'istogramma richiede una ripartizione dei valori in classi.

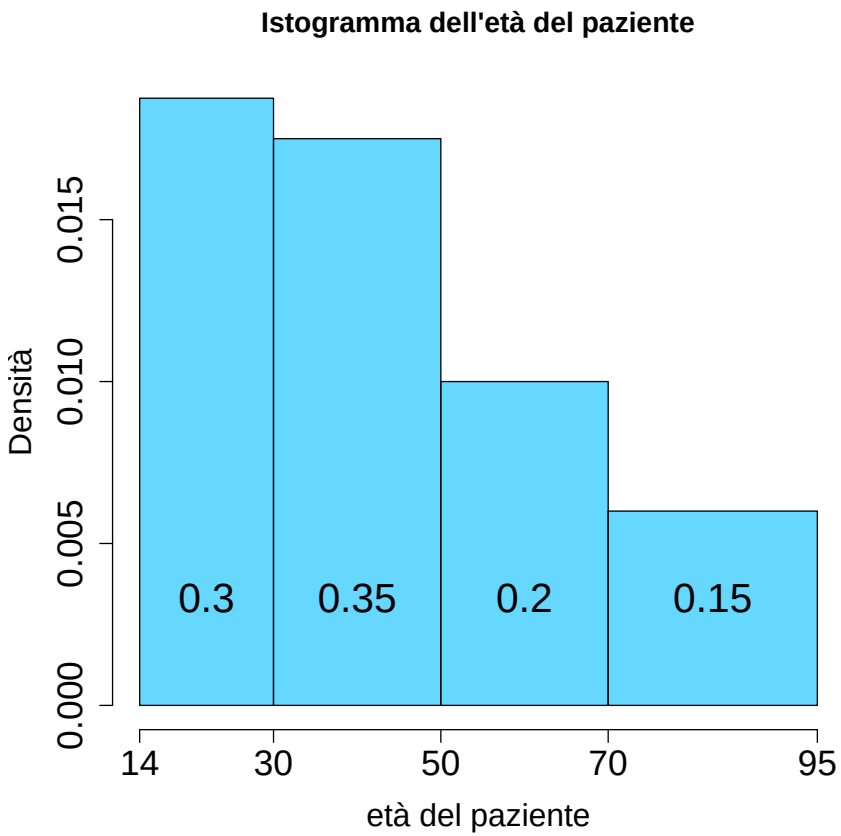


Istogramma e densità

In un istogramma con classi di ampiezza diversa, l'altezza dei rettangoli corrisponde alla **densità**.
La densità si ottiene come rapporto tra la frequenza n_j e l'ampiezza Δ_j della classe.

$$d_j = \frac{n_j}{n \cdot \Delta_j} = \frac{f_j}{\Delta_j}$$

L'**area dei rettangoli** corrisponde alle frequenze relative.



Istogramma e densità

Classi di Età	n_j	f_j	a_j	b_j	Δ_j	d_j
[14-30]	6	0.30	14	30	16	0.0187
]30-50]	7	0.35	30	50	20	0.0175
]50-70]	4	0.20	50	70	20	0.0100
]70-95]	3	0.15	70	95	25	0.0060
	$n = 20$	1				

Calcolo dell'ampiezza nel caso di variabile quantitativa continua;

Esempio di ampiezza della prima classe (14-30):

$$\Delta_1 = b_1 - a_1 = 30 - 14 = 16$$

$$d_1 = \frac{n_1}{n \cdot \Delta_1} = \frac{6}{20 \cdot 16} = 0.0187$$

Istogramma e densità

Numero di figli	n_j	f_j	a_j	b_j	Δ_j	d_j
1-2	10	0.71	1	2	2	0.35
3-4	3	0.21	3	4	2	0.12
5	1	0.07	5	5	1	0.03
	$n = 14$	1				

Nel caso di variabile quantitativa discreta, l'ampiezza si calcola sommando il valore 1 alla differenza tra gli estremi:

Esempio di ampiezza della prima classe:

$$\Delta_1 = b_1 - a_1 + 1 = 2 - 1 + 1 = 2$$

$$d_1 = \frac{n_1}{n \cdot \Delta_1} = \frac{10}{14 \cdot 2} = 0.35$$