2 La sintesi dei dati

2.1 Serie di dati e distribuzioni di frequenze

Una distribuzione di frequenze consente di avere una rappresentazione più compatta e più informativa rispetto alla serie di dati osservati e tanto più quanto più la serie è numerosa.

In particolare consente di:

- disporre dell'elenco di tutte le modalità (valori) distinte/i;
- individuare le modalità (valori) più frequenti;
- determinare l'intervallo di variazione, se si dispone di valori, della serie originaria;
- ipotizzare particolari modelli teorici, atti a descrivere l'andamento delle frequenze.

Le "*frequenze assolute*" n_i indicano il numero di volte con cui ciascuna modalità (valore) si presenta nella serie.

Oltre alle frequenze assolute possono essere calcolate le "*frequenze relative*" f_i, date dal rapporto fra ciascuna frequenza assoluta e il totale delle osservazioni ed esprimibili anche in termini percentuali. Le frequenze relative consentono di confrontare due variabili rilevate su collettivi di numerosità diversa. Si pensi ad esempio di voler confrontare il peso di un gruppo di soggetti di sesso maschile con il peso di un gruppo di soggetti di sesso femminile.

A partire da una variabile qualitativa ordinabile, per costruire una distribuzione di frequenze, è necessario anzitutto disporre le modalità/valori in ordine crescente.

Ha senso, in tal caso, calcolare anche le "*frequenze cumulate*", assolute N_i o relative F_i, date dalla somma di ciascuna frequenza assoluta, o relativa, con le precedenti.

I valori di una variabile quantitativa discreta, se numerosi, possono essere raggruppati in classi; tuttavia, in questo caso, le classi non hanno lo stesso significato che hanno per la descrizione di un fenomeno continuo e necessariamente l'estremo superiore di una classe non coincide con l'estremo inferiore della classe successiva.

La descrizione in classi per i fenomeni quantitativi continui ha appunto lo scopo di recuperare la natura continua del carattere, che al momento della rilevazione viene a cadere, a causa degli arrotondamenti.

Se il carattere è di tipo continuo, le distribuzioni di frequenze per valori singoli possono risultare poco utili o addirittura inutili per suggerire l'eventuale modello teorico atto a descrivere il fenomeno. Si rende pertanto necessario raggruppare i valori osservati in opportune classi di ampiezza costante o variabile.

Il criterio di raggruppamento in classi comporta sempre una perdita di informazioni rispetto alla serie originaria e tanto più quanto più sono ampie le classi. La perdita di informazioni influisce sulla correttezza delle costanti sintetiche calcolate sulla distribuzione di frequenze.

Purtroppo, le procedure con cui le classi possono essere formate sono assolutamente arbitrarie e possono condurre a distribuzioni di frequenze sensibilmente diverse, sebbene determinate sulla stessa serie di dati. Si auspica, pertanto, che vengano rispettate le seguenti regole generali:

- gli estremi delle classi siano arrotondati all'intero più prossimo o abbiano almeno il minor numero possibile di cifre decimali;
- le ampiezze delle classi siano costanti e piccole (l'ampiezza determina il numero delle classi e viceversa);

- l’estremo inferiore della prima classe e l’estremo superiore dell’ultima differiscano il meno possibile, rispettivamente, dal valore più piccolo e dal valore più grande osservato;
- nessuna classe abbia frequenza nulla;
- ci sia un solo massimo o al più due;
- l’andamento sia crescente e poi decrescente o comunque monotono;
- scegliendo intervalli aperti (chiusi) sia a destra che a sinistra, si inseriscano, se è possibile, casi uguali in egual numero nelle classi contigue.

2.2 Rappresentazioni grafiche

Da una tabella di frequenze possono dedurre informazioni solo gli esperti del settore, o comunque chi ha un minimo di conoscenze statistiche, mentre un grafico è immediatamente interpretabile da chiunque. Ciò perché la mente umana percepisce e memorizza con maggiore rapidità figure piuttosto che cifre.

Un grafico, d’altra parte, rappresenta una fonte d’informazione meno ricca, in quanto non consente di evidenziare piccole differenze tra frequenze.

Grafico e tabella, dunque, vanno utilizzati entrambi, cioè sono complementari.

Tuttavia, una rappresentazione grafica deve essere autonoma dalla tabella, ovvero deve contenere tutte le informazioni necessarie per la sua interpretazione: va riportata la fonte da cui sono ricavati i dati, vanno specificate le variabili rilevate e le modalità o i valori assunti, vanno indicate le unità di misura. Le indicazioni devono essere leggibili e il grafico non deve apparire confuso se si rappresentano più fenomeni. E’ necessario, inoltre, scegliere la rappresentazione più semplice, se vi è la possibilità di una gamma di alternative.

2.3 Esempi

All'inizio dell'anno accademico 2002/03 è stato sottoposto il seguente questionario agli studenti del corso di Statistica 1 – Corso di laurea in Economia e finanza, Facoltà di Economia di Palermo:

Corso di laurea in Economia e Finanza
Disciplina: STATISTICA 1
A.A. 2002/03
QUESTIONARIO

1) Cognome……Nome……
2) Sesso
   - ☐ F
   - ☐ M
3) Data di nascita……
4) Comune di residenza……
5) Provincia di residenza……
6) Altezza (in cm) ……
7) Peso (in Kg) ……
8) Scuola media superiore
   - ☐ Liceo classico
   - ☐ Liceo scientifico
   - ☐ Istituto tecnico commerciale
   - ☐ Istituto tecnico per geometri
   - ☐ Istituto tecnico industriale
   - ☐ Altro……
9) Voto di maturità
   - ☐ ……/100
   - ☐ ……/60
10) Matricola
    - ☐ Si
    - ☐ No A.A. di immatricolazione……/……
11) N. di esami sostenuti……
12) Ha sostenuto l'esame di Matematica
    - ☐ Si voto……
    - ☐ No
13) Difficoltà incontrate nei corsi di I semestre
    - ☐ Scarse
    - ☐ Medie
    - ☐ Elevate
    - ☐ Molto elevate
14) E' soddisfatto per la scelta del Corso di studi ?
    - ☐ Si
    - ☐ No
15) N. di componenti del nucleo familiare……
16) Titolo di studio del capofamiglia
    - ☐ Nessun titolo
    - ☐ Licenza elementare
    - ☐ Licenza media
    - ☐ Maturità
    - ☐ Laurea

Raccolti tutti i questionari, è stato effettuato lo spoglio. I dati sono stati organizzati sotto forma di matrice di dimensione $n \times k$, che per motivi di spazio non riportiamo, dove n=140 è il numero delle matricole frequentanti il corso e k sono le variabili rilevate. I dati riguardanti le variabili rilevate (sesso, provincia di residenza, altezza, peso, scuola di provenienza, ecc…) sono stati elaborati e sintetizzati. Di seguito riportiamo alcune di queste variabili, una per ogni tipologia.

Si consideri la variabile "scuola superiore di provenienza". Se si suppone che le diverse scuole abbiano pari importanza, tale variabile può essere considerata una variabile qualitativa sconnessa, poiché considerati due soggetti è possibile dire solo se questi provengono dallo stesso tipo di scuola o meno.

Per ragioni di spazio, le osservazioni riguardanti i 140 soggetti sono riportate sotto forma di tabella, ma nella matrice dei dati, rappresenterebbero una singola colonna. Ovviamente, è conveniente attribuire un'etichetta, o meglio un codice, a ciascuna modalità della variabile, per velocizzare l'immissione dei dati:

- Liceo classico $\rightarrow$ LC
- Liceo scientifico $\rightarrow$ LS
- Istituto tecnico commerciale $\rightarrow$ ITC
- Istituto tecnico per geometri $\rightarrow$ ITG
- Istituto tecnico industriale $\rightarrow$ ITI
- Altro $\rightarrow$ A

LC	LS	LS	ITC	ITC	LS	LS	LC	LS	LS
ITC	LS	ITC	ITC	ITC	LC	LC	ITC	ITC	LS
LC	ITC	ITG	ITC	A	ITC	ITC	ITC	ITC	LC
ITC	ITC	A	LS	LC	ITC	ITC	LS	A	ITC
ITC	LC	LS	ITC	LC	ITC	LS	ITC	ITC	LS
ITC	LS	A	LS	LS	ITC	ITC	LS	LS	ITC
LS	LS	A	LS	ITC	LS	ITC	LS	A	LS
ITC	LS	ITC	LS	LS	ITC	ITC	LS	LS	ITC
ITC	LS	ITC	ITI	LS	ITC	ITC	LS	LS	ITC
ITC	ITC	ITC	LS	LS	ITC	LS	LS	LS	LC
ITC	LS	LS	LS	LS	ITC	LS	LS	ITC	LC
LS	ITC	LS	ITC	A	LC	LS	LS	LS	LS
LS	LS	LS	A	LS	A	ITC	ITC	ITC	LC
ITC	LS	ITC	A	ITG	ITC	ITC	LS	ITC	ITC

La tabella sopra contiene la serie dei dati che, come è evidente, non è per nulla informativa; costruiamo, dunque la distribuzione di frequenza, ovvero contiamo quante volte ciascuna modalità si ripete nella serie. Di seguito, oltre alle frequenze assolute n_i, si riportano anche le frequenze relative f_i e le frequenze relative percentuali f_i*100:

x_i	n_i	f_i	f_i*100
A	10	0,07	7
ITC	58	0,41	41
ITG	2	0,01	1
ITI	1	0,01	1
LC	13	0,09	9
LS	56	0,40	40
totale	140	1	100

Dalla tabella si evince immediatamente quali sono le modalità più frequenti. In particolare, la maggior parte degli studenti, rispettivamente il 41% e il 40%, provengono dall'ITC e dal LS.

La modalità cui è associata la frequenza più alta viene definita "*moda*". In questo caso la moda è "ITC".

Le rappresentazioni grafiche tipiche di una variabile qualitativa sconnessa sono il **grafico a colonne**, il **grafico a barre o a nastri** e, se il numero delle modalità non è elevato, come in questo caso, gli **areogrammi**.

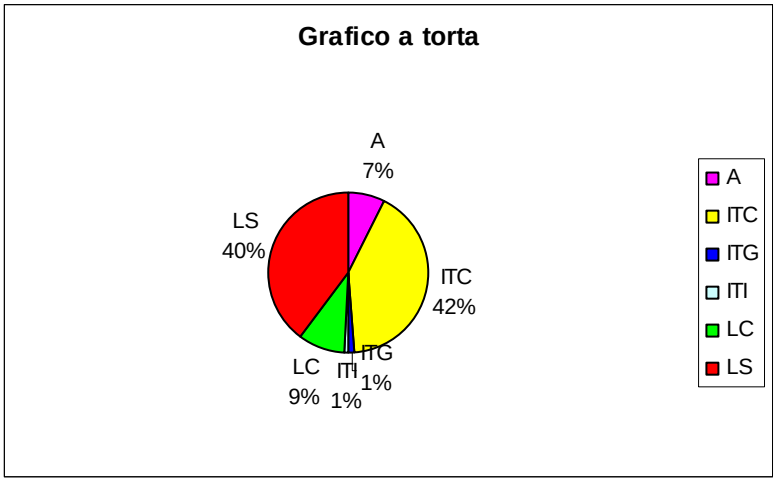
In un areogramma le frequenze sono rappresentate da superfici di figure piane (quadrati, rettangoli, cerchi), poste l'una accanto all'altra, oppure da parti di una stessa figura. L'areogramma, rispetto al grafico a colonne e al grafico a barre, dà meno possibilità di apprezzare piccole differenze fra le frequenze, perché l'occhio umano è più abituato a confrontare lunghezze piuttosto che aree.

Il grafico a settori circolari o **grafico a torta** è un areogramma. Si costruisce un cerchio di area uguale o proporzionale al totale delle frequenze e si ripartisce in tanti settori quante sono le modalità. Ciascun settore ha area uguale o proporzionale alla frequenza della modalità cui è associato, per cui l'angolo α di ciascun settore si può ricavare dalla proporzione:

$$360 : \alpha = n : n_i \qquad \Rightarrow \qquad \alpha = \frac{360 \cdot n_i}{n} = 360 \cdot f_i .$$

Oggi, in realtà, esistono diversi software statistici che consentono di costruire tabelle e grafici tramite procedure molto semplici e automatiche.

In genere, il grafico a torta è accompagnato da una legenda, che associa colori o tratteggi diversi a ciascun settore. In alternativa, si possono specificare le modalità su ciascun settore:



Analizziamo adesso la variabile "titolo di studio del capofamiglia". Questa variabile è una variabile qualitativa ordinabile poiché, considerati due soggetti, è

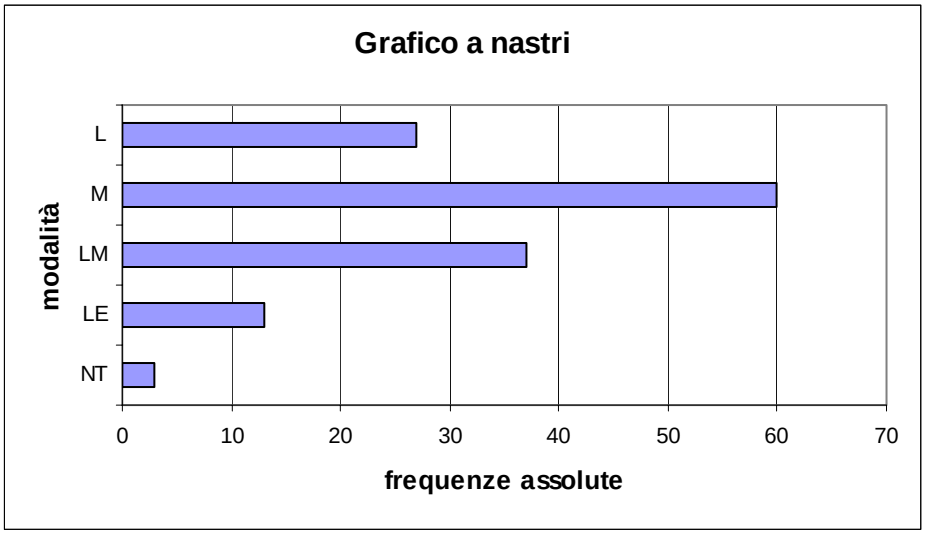
possibile dire non solo se hanno un titolo di studio diverso, ma anche chi possiede un titolo più importante. Si riporta di seguito direttamente la distribuzione di frequenza. Nel costruirla è necessario ricordare di ordinare le modalità. Le modalità sono state codificate nel seguente modo:

- Nessun titolo → NT
- Licenza elementare → LE
- Licenza media → LM
- Maturità → M
- Laurea → L

x_i	n_i	f_i	f_i*100	N_i	F_i	F_i*100
NT	3	0,02	2	3	0,02	2
LE	13	0,09	9	16	0,11	11
LM	37	0,26	26	53	0,38	38
M	60	0,43	43	113	0,81	81
L	27	0,19	19	140	1,00	100
totale	140	1	100			

Dalla tabella si evince che la maggioranza dei genitori ha conseguito la maturità (43%). Per questo tipo di variabile ha un senso calcolare anche le frequenze cumulate, assolute N_i, relative F_i o percentuali F_i*100. La frequenza assoluta cumulata N_3 sta ad indicare, ad esempio, che 53 genitori su 140 hanno un titolo di studio inferiore o uguale alla LM. La frequenza relativa cumulata F_4 sta ad indicare che l'81% dei genitori ha un titolo di studio inferiore o uguale alla maturità, e così via.

Le rappresentazioni grafiche tipiche di una variabile qualitativa ordinabile sono uguali a quelle di una variabile qualitativa sconnessa. Se il carattere è ordinabile, è preferibile disporre i nastri o le colonne secondo l'ordine con cui si susseguono le modalità. Scegliamo il grafico a nastri. I grafici a nastri sono rappresentati da rettangoli aventi tutti la stessa altezza e basi uguali o proporzionali alle frequenze relative alle singole modalità:



Può accadere che le dimensioni del disegno non siano contenute nel foglio. In tal caso, si può assumere un'unità di misura diversa oppure si possono troncare i rettangoli, ovvero si può spostare l'origine di riferimento; così facendo, però, ci si può non rendere conto delle effettive variazioni nelle frequenze. D'altra parte raddoppiando o dimezzando l'unità di misura si possono amplificare o attenuare le oscillazioni di un fenomeno. L'arbitrarietà nella scelta dell'unità di misura e lo spostamento dell'origine degli assi può fornire impressioni totalmente diverse del fenomeno rappresentato; si parla di manipolazione delle informazioni mediante lo strumento statistico. Si pensi, ad esempio, alle rappresentazioni grafiche riguardanti l'andamento dei mercati finanziari.

Quando le dimensioni di un rettangolo (in questo caso di una base, ma potrebbe riguardare l'altezza nel caso di un grafico a colonne) si discostano di molto rispetto alle dimensioni degli altri, un buon metodo potrebbe essere quello di amputare il rettangolo e specificare nella parte amputata la frequenza ad esso associata.

Consideriamo adesso una variabile quantitativa discreta, qual è ad esempio il "numero dei componenti del nucleo familiare".

Di seguito si riporta la serie dei dati già ordinata e la distribuzione delle frequenze assolute, relative e relative cumulate:

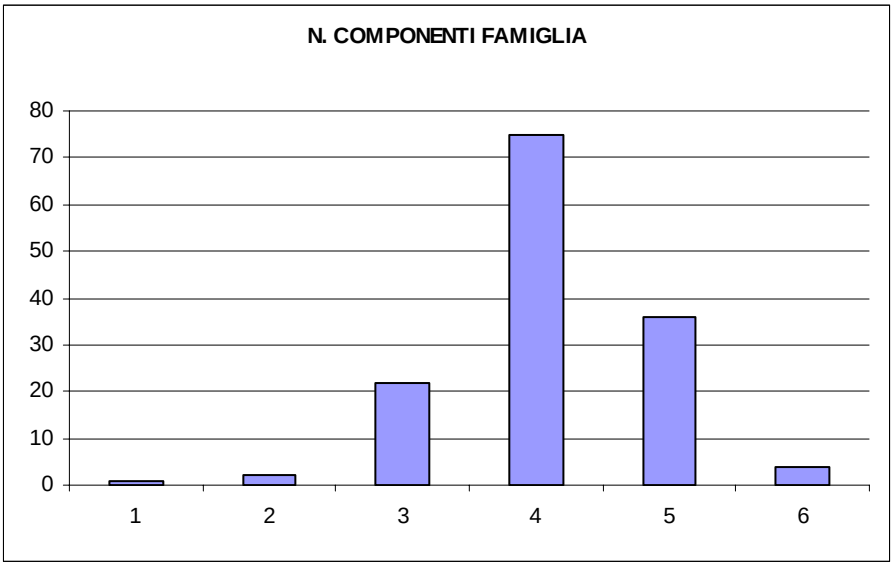
1	3	3	4	4	4	4	4	4	4	5	5	5	5
2	3	3	4	4	4	4	4	4	4	5	5	5	5
2	3	3	4	4	4	4	4	4	4	5	5	5	5
3	3	3	4	4	4	4	4	4	4	5	5	5	5
3	3	3	4	4	4	4	4	4	4	5	5	5	5
3	3	4	4	4	4	4	4	4	4	5	5	5	5
3	3	4	4	4	4	4	4	4	4	5	5	5	6
3	3	4	4	4	4	4	4	4	4	5	5	5	6
3	3	4	4	4	4	4	4	4	4	5	5	5	6
3	3	4	4	4	4	4	4	4	4	5	5	5	6

x_i	n_i	f_i	F_i	f_i*100	F_i*100
1	1	0,01	0,01	1	1
2	2	0,01	0,02	1	2
3	22	0,16	0,18	16	18
4	75	0,54	0,72	54	72
5	36	0,26	0,97	26	97
6	4	0,03	1,00	3	100
tot	140	1		100	

Dalla tabella si evince che la famiglia media è costituita per lo più da 4 componenti (54%).

La rappresentazione grafica tipica di una variabile di conteggio è il **diagramma cartesiano** o anche il grafico a colonne. Nei grafici a colonne, o a rettangoli, ogni modalità viene rappresentata sull'asse delle ascisse con segmenti uguali ed equidistanti. Si costruisce poi, su ciascun segmento, un rettangolo di altezza uguale o proporzionale alla frequenza associata a quella determinata modalità.

Il diagramma cartesiano differisce dal grafico a colonne in quanto anziché rettangoli considera segmenti di retta di lunghezza pari o proporzionali alle singole frequenze.



Si consideri adesso una variabile quantitativa continua, ad esempio la statura.

In questa fase dell'elaborazione non c'è differenza fra variabili misurabili su scala di intervalli o su scala di rapporti.

Si riporta dunque la serie delle stature:

178	173	176	190	173	172	180	153	164	170	165	172	163	164
175	185	178	178	173	183	174	155	168	165	173	160	168	160
180	175	163	170	170	178	173	174	156	170	167	168	166	160
175	176	180	178	175	170	177	167	163	165	170	177	160	160
186	176	170	185	181	181	175	170	165	170	168	163	170	160
170	175	180	165	175	178	175	170	157	160	153	160	160	160
180	191	180	171	180	178	180	170	150	164	172	168	160	160
174	176	182	181	173	177	170	179	160	169	160	165	160	170
184	170	170	182	180	167	163	163	158	170	165	152	165	161
182	177	182	178	171	183	160	170	165	165	165	158	168	158

Costruire una distribuzione di frequenza per valori singoli non porterebbe per tale variabile ad una sintesi significativa delle informazioni; come si può notare, infatti, la tabella che segue è troppo lunga per dare informazioni immediate sull'andamento delle misure; molti sono i valori diversi e con frequenza pari a 1 o comunque con frequenza molto bassa:

x_i	n_i	x_i	n_i
150	1	172	3
152	1	173	6
153	2	174	3
155	1	175	8
156	1	176	4
157	1	177	4
158	3	178	8
160	16	179	1
161	1	180	9
163	6	181	3
164	3	182	4
165	11	183	2
166	1	184	1
167	3	185	2
168	6	186	1
169	1	190	1
170	19	191	1
171	2	*totale*	140

Occorre, pertanto, costruire una distribuzione di frequenza per classi. Scegliamo otto classi di ampiezza costante e pari a *5 cm*, chiuse a destra. Spesso, è conveniente lasciare aperte la prima e l'ultima classe, in modo tale da poter inserire nuove osservazioni, rilevate in tempi successivi:

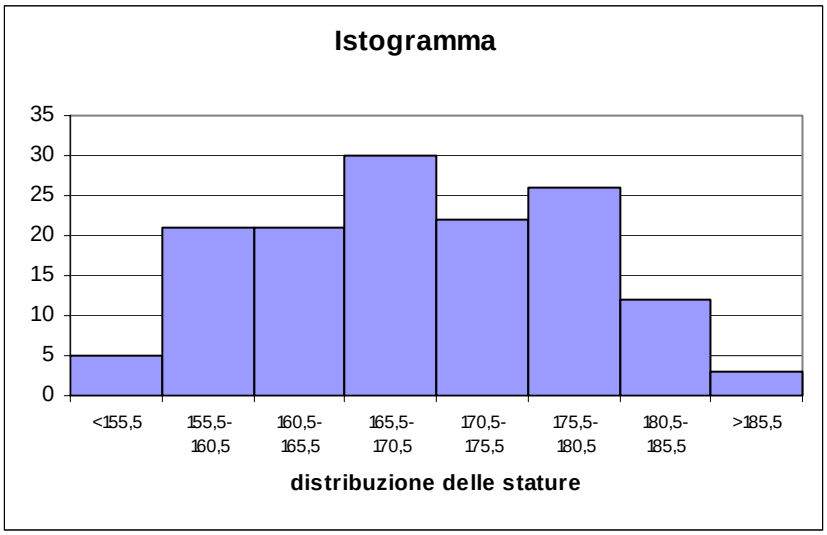
x_i -| x_i	n_i
≤155,5	5
155,5-|160,5	21
160,5-|165,5	21
165,5-|170,5	30
170,5-|175,5	22
175,5-|180,5	26
180,5-|185,5	12
>185,5	3
totale	140

Le rappresentazioni grafiche tipiche di una variabile quantitativa continua sono l'**istogramma** e il **poligono di frequenza**. L'istogramma è costituito da tanti rettangoli adiacenti quante sono le classi e hanno area A_i uguale o proporzionale alle frequenze n_i:

$$A_i = b_i \cdot h_i \cong n_i$$

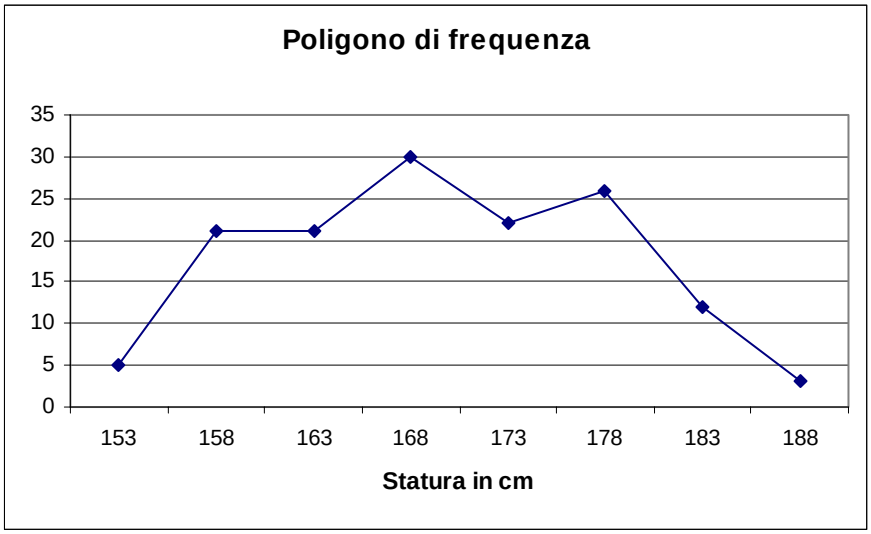
Ciascun rettangolo ha dunque base b_i pari all'ampiezza della classe e altezza h_i pari alla *densità di frequenza*, ossia $h_i = \frac{n_i}{b_i}$. Ovviamente, se le classi hanno tutte la stessa ampiezza, le basi agiscono solo come fattore di scala quindi, in tal caso, riportare in ordinata frequenze o densità di frequenze è in pratica la stessa cosa.

Nell'esempio considerato, per poter rappresentare l'istogramma, la prima e l'ultima classe si considerano di ampiezza pari a quella delle altre:



In relazione alla diversa ampiezza delle classi, c'è un cambiamento nella rappresentazione grafica; l'arbitrarietà nella scelta delle classi modifica, dunque, la visualizzazione del fenomeno in esame.

Il poligono di frequenza viene in genere sovrapposto all'istogramma. Si tratta di una spezzata che passa per i punti medi delle basi superiori dei rettangoli:



Se le classi sono tutte della stessa ampiezza, l'area sottesa dal poligono di frequenza è uguale all'area dell'istogramma.