Statistica

Dati e tabelle di frequenza

Domenico De Stefano

a.a. 2025/2026

Indice

Indice

Distribuzione statistica disaggregata

Si consideri un collettivo statistico di n unità, dove si sia osservata la variabile X. Si chiama distribuzione statistica disaggregata secondo la variabile X l'insieme delle osservazioni (rappresentate da numeri o da espressioni verbali a seconda della natura della variabile) relative alle n unità del collettivo (più semplicemente questi sono i cosiddetti **dati grezzi**).
In simboli, la distribuzione disaggregata sarà indicata come

$$x_1, x_2, \ldots, x_n$$

dove x_1 è l'osservazione relativa all'unità identificata dal numero 1, x_2 l'osservazione relativa all'unità identificata dal numero 2 e così via
(NB: attenzione il minuscolo non è messo a caso: la variabile in se si indica con la X maiuscola, le sue **modalità** osservate sulle unità statistiche con le x minuscole!)

I dati grezzi non consentono una facile visione d'insieme!

Distribuzione di frequenza assoluta

Si consideri ancora la variabile X. Si chiama distribuzione di frequenza assoluta la lista delle modalità osservate di X accompagnata dal numero di volte in cui queste vengono osservate, ossia accompagnata dalle rispettive frequenze assolute.

È molto facile ottenere distribuzioni di frequenza assoluta per caratteri qualitativi e quantitativi discreti. In presenza di caratteri quantitativi continui (o anche discreti, se assumono tantissime modalità), abbiamo bisogno di qualche operazione preliminare per trattarli (vedremo in seguito...).

Esempio: dataset `vets`

Distribuzione di frequenza del luogo di servizio dei veterani

VETERAN	frequenza assoluta
VIETNAM	646
OTHER	97

Esempio: dataset babies

Distribuzione di frequenza del fumo

fumo	frequenza assoluta
S	16
N	16

Esempio: dataset babies

Distribuzione di frequenza della durata della gravidanza

durata	frequenza assoluta
34	1
35	3
36	3
37	2
38	5
39	7
40	3
41	3
42	5

Esempio: dataset babies

Per il peso alla nascita è conveniente definire classi di modalità (o intervalli) contigue ed effettuare il conteggio delle unità che appartengono a ciascuna classe.

peso	frequenza assoluta
$(2400, 2600]$	5
$(2600, 2800]$	5
$(2800, 3000]$	5
$(3000, 3200]$	6
$(3200, 3400]$	5
$(3400, 3600]$	6

NB: la scelta delle classi è condizionata dal livello di disaggregazione con cui i dati sono stati rilevati. In altre parole è un'operazione arbitraria (decidete voi numero e ampiezza classi!) sulla base di come sono "disperse" le modalità della variabile in questione

Classi di differenti lunghezze

Può capitare, o per scelta (si vuole fornire informazioni più dettagliate su parte della distribuzione),
o per necessità (quando i dati sono già stati raggruppati in classi da qualcuno... nel caso ad es. delle classi di età in cui talvolta le classi estreme sono lasciate aperte usando le paroline "...e oltre", es. 20–39; 40–59; 60–79; 80 e oltre),
di costruire delle classi utilizzando intervalli di lunghezza differente.

In questo caso è conveniente definire anche la densità di frequenza.

La densità è definita come:

$$\left(\begin{array}{c}\text{densità}\\ \text{di una classe}\end{array}\right) = \frac{\text{frequenza assoluta di } Y \text{ sull'intervallo}}{\text{lunghezza dell'intervallo}}.$$

Esempio: dataset babies

peso	frequenza assoluta	densità
$(2400, 2600]$	5	$5/200{=}0.025$
$(2600, 2800]$	5	$5/200{=}0.025$
$(2800, 3000]$	5	$5/200{=}0.025$
$(3000, 3200]$	6	$6/200{=}0.030$
$(3200, 3600]$	11	$11/400{=}0.0275$

La densità ci dice il numero atteso di unità statistiche per ogni unità di misura della variabile. Nella prima classe, per esempio, ci aspettiamo di osservare 2,5 neonati ogni 100 grammi di peso (ovvero, 2,5 neonati con peso tra 2400 e 2500 e 2,5 neonati con peso tra 2500 e 2600).

Distribuzioni di frequenza per gruppi: dataset `babies`

Distribuzione di frequenza della durata della gravidanza nel gruppo di madri non fumatrici e nel gruppo di madri fumatrici.

Fumo=N	
durata	frequenza assoluta
34	1
35	2
36	1
37	2
38	2
39	3
40	3
41	1
42	1

Fumo=S	
durata	frequenza assoluta
34	0
35	1
36	2
37	0
38	3
39	4
40	0
41	2
42	4

Esempio: dataset babies

Peso alla nascita da madri non fumatrici e da madri fumatrici.

Fumo=N

durata	frequenza assoluta
(2400, 2600]	2
(2600, 2800]	2
(2800, 3000]	2
(3000, 3200]	3
(3200, 3400]	3
(3400, 3600]	4

Fumo=S

durata	frequenza assoluta
(2400, 2600]	3
(2600, 2800]	3
(2800, 3000]	3
(3000, 3200]	3
(3200, 3400]	2
(3400, 3600]	2

Distribuzione condizionata

Le distribuzioni della durata della gravidanza e del peso alla nascita per una fissata modalità della condizione rispetto al fumo (non fumo/fumo) sono distribuzioni condizionate.

Se indichiamo in modo generico con X la variabile che stiamo studiando (la durata della gravidanza, per esempio) e con Y il carattere tramite cui estraiamo le unità statistiche da considerare nell'analisi (la condizione rispetto al fumo, nel nostro caso), si dice variabile X condizionata a $Y = y$ e si indica $X|Y = y$ la restrizione di X al sottoinsieme $Y = y$.

Distribuzione condizionata (cont)

La distribuzione della variabile $X|Y = y$ viene normalmente detta la distribuzione di X condizionata a $Y = y$ o, equivalentemente, la distribuzione di X dato $Y = y$.

Si osservi che esiste una distribuzione condizionata (di X dato Y) per ogni modalità di Y.

La distribuzione della variabile X senza distinzione per condizione rispetto a Y è detta distribuzione marginale.

Esempio: dataset babies

Distribuzioni condizionate

durata	fumo	
	N	S
34	1	0
35	2	1
36	1	2
37	2	0
38	2	3
39	3	4
40	3	0
41	1	2
42	1	4

Distribuzione marginale

durata	frequenza assoluta
34	1+0=1
35	2+1=3
36	1+2=3
37	2+0=2
38	2+3=5
39	3+4=7
40	3+0=3
41	1+2=3
42	1+4=5

Frequenze relative

Dividendo una frequenza assoluta per il numero totale di unità statistiche nel collettivo analizzato (n nel nostro caso) otteniamo le cosiddette frequenze relative (o proporzioni), ovvero

$$\left(\begin{array}{c}\text{frequenze}\\ \text{relative}\end{array}\right) = \frac{\left(\begin{array}{c}\text{frequenze}\\ \text{assolute}\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

Hanno il vantaggio, rispetto alle frequenze assolute, di permettere di confrontare distribuzioni di frequenza basate su numeri differenti di unità statistiche.

Esempio: effetti del fumo sul peso dei neonati

peso	frequenza relativa
$(2400, 2600]$	$5/32 = 0.15625$
$(2600, 2800]$	$5/32 = 0.15625$
$(2800, 3000]$	$5/32 = 0.15625$
$(3000, 3200]$	$6/32 = 0.18750$
$(3200, 3400]$	$5/32 = 0.15625$
$(3400, 3600]$	$6/32 = 0.18750$

Distribuzioni di frequenza: notazione

Se la nostra variabile si chiama X allora...

- x_i indicherà la **generica** modalità i/ classe $(c_{i-1}, c_i]$ della variabile X, dove $i = 1, 2, \ldots, k$ (e k è il numero delle modalità/classi);
- n_i frequenza assoluta numero di unità statistiche che possiedono la modalità (o classe) x_i (c_i);
- n numero totale di osservazioni nel collettivo $(n = n_1 + n_2 + \cdots + n_k)$;
- f_i frequenza relativa $(f_i = n_i/n)$.

modalità/classe	frequenze assolute	frequenze relative
x_1	n_1	$f_1 = n_1/n$
x_2	n_2	$f_2 = n_2/n$
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = n_k/n$
Totale	n	1

Avviso generale

Ogni libro usa una propria notazione, magari diversa da quella appena introdotta. Un'altra notazione comune è, per esempio, la seguente

- x_i modalità/classe i della variabile X, $i = 1, 2, \ldots, k$ (k modalità/classi)
- f_i frequenza assoluta numero di unità statistiche che possiedono la modalità/classe x_i
- n numero totale di osservazioni ($n = f_1 + f_2 + \cdots + f_k$)
- p_i frequenza relativa ($p_i = f_i/n$)

Qualunque scelta va bene: basta definire cosa si intende con ciascun simbolo ed essere coerenti!

Esercizio: esiti ammissione a Berkeley, 1973

I seguenti dati rappresentano gli esiti dell'ammissione all'Università di California, Berkeley (USA) nel 1973. È riportato l'esito dell'ammissione (`Admit`), il sesso dei candidati (`Gender`) e il Dipartimento erogante il corso di studi scelto dai candidati (`Dept`).

Admit	Gender	Dept	Frequenza assoluta
Admitted	Male	A	512
Rejected	Male	A	313
Admitted	Female	A	89
Rejected	Female	A	19
Admitted	Male	B	353
Rejected	Male	B	207
Admitted	Female	B	17
Rejected	Female	B	8
Admitted	Male	C	120
Rejected	Male	C	205
Admitted	Female	C	202
Rejected	Female	C	391
Admitted	Male	D	138
Rejected	Male	D	279
Admitted	Female	D	131
Rejected	Female	D	244
Admitted	Male	E	53
Rejected	Male	E	138
Admitted	Female	E	94
Rejected	Female	E	299
Admitted	Male	F	22
Rejected	Male	F	351
Admitted	Female	F	24
Rejected	Female	F	317

È una matrice dei dati? Quante sono le variabili rilevate? Di che tipo sono? Quante sono le unità statistiche?

Il simbolo $\sum$ (sommatoria)

Cosa intendiamo per

$$n = \sum_{i=1}^{k} n_i$$

ovvero per 'Somma per i che va da 1 a k' ?

$$n = n_1 + n_2 + \cdots + n_k$$

Alcune proprietà

1. $\sum_{i=1}^{k}(y_i + x_i) = \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} x_i$
2. $\sum_{i=1}^{k} a y_i = a \sum_{i=1}^{k} y_i$
3. Fate attenzione: $\sum_{i=1}^{k} a = ak$

Esercizio: $\sum_{i=1}^{k} f_i = ?$

Frequenze cumulate

- La frequenza cumulata ha senso se la variabile X è almeno ordinata, quindi

$$x_1 < x_2 < \ldots < x_k$$

- La frequenza assoluta (o anche relativa, perchè no?) cumulata per la modalità/classe x_i è la somma delle frequenze assolute (relative) per le modalità/classi $\leq x_i$

$$F_i = f_1 + \ldots + f_i = \sum_{h=1}^{i} f_h$$

modalità/classe	frequenze cumulate assolute	frequenze cumulate relative
x_1	n_1	$F_1 = f_1$
x_2	$n_1 + n_2$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_1 + \ldots + n_i$	$F_i = f_1 + \ldots + f_i$
$\vdots$	$\vdots$	$\vdots$
x_k	n	?

Esercizio: dataset `babies`

Si costruisca la distribuzione di frequenze cumulate per la durata della gravidanza nel dataset `babies` (v. slides precedenti).

Partendo dalla distribuzione di frequenze assolute, abbiamo

durata	frequenza assoluta	frequenza cumulata
34	1	1
35	3	4
36	3	7
37	2	9
38	5	14
39	7	21
40	3	24
41	3	27
42	5	32

Esercizio: dataset `trees`

Si costruisca la distribuzione di frequenze cumulate per il volume degli alberi di ciliegio nero nel dataset `trees` (v. slides precedenti).

I dati sono i seguenti

```
10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 24.2 21.0 21.4 21.3 19.1
22.2 33.8 27.4 25.7 24.9 34.5 31.7 36.3 38.3 42.6 55.4 55.7 58.3 51.5 51.0
77.0
```

Immaginando di fare 7 classi, otteniamo la seguente distribuzione di frequenze assolute e frequenze assolute cumulate.

Classe	Frequenza assoluta	Frequenza assoluta cumulata	Frequenza relativa cumulata
(10, 20]	10	10	10/31
(20, 30]	9	19	19/31
(30, 40]	5	24	24/31
(40, 50]	1	25	25/31
(50, 60]	5	30	30/31
(60, 70]	0	30	30/31
(70, 80]	1	31	31/31

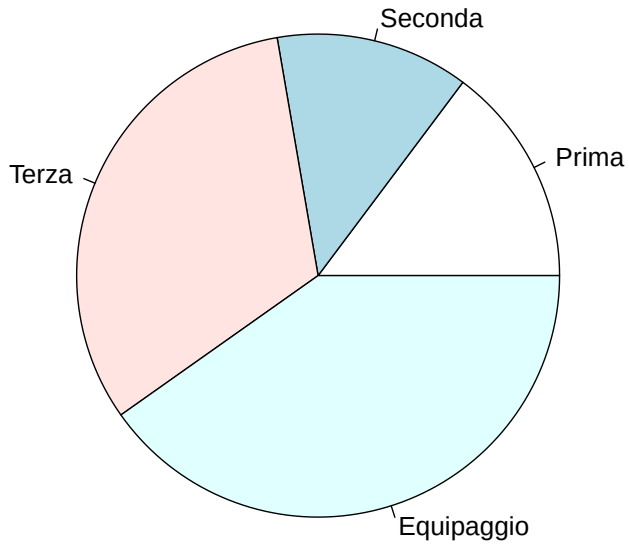
Indice

Finalmente un grafico!

Possiamo cercare di visualizzare le distribuzioni di frequenza, rappresentando in qualche modo ciascuna modalità del carattere con la relativa frequenza.

Esempio: disastro del Titanic.

Passeggero	frequenza assoluta	%
Prima	325	14.77
Seconda	285	12.95
Terza	706	32.08
Equipaggio	885	40.21



Il disastro del Titanic

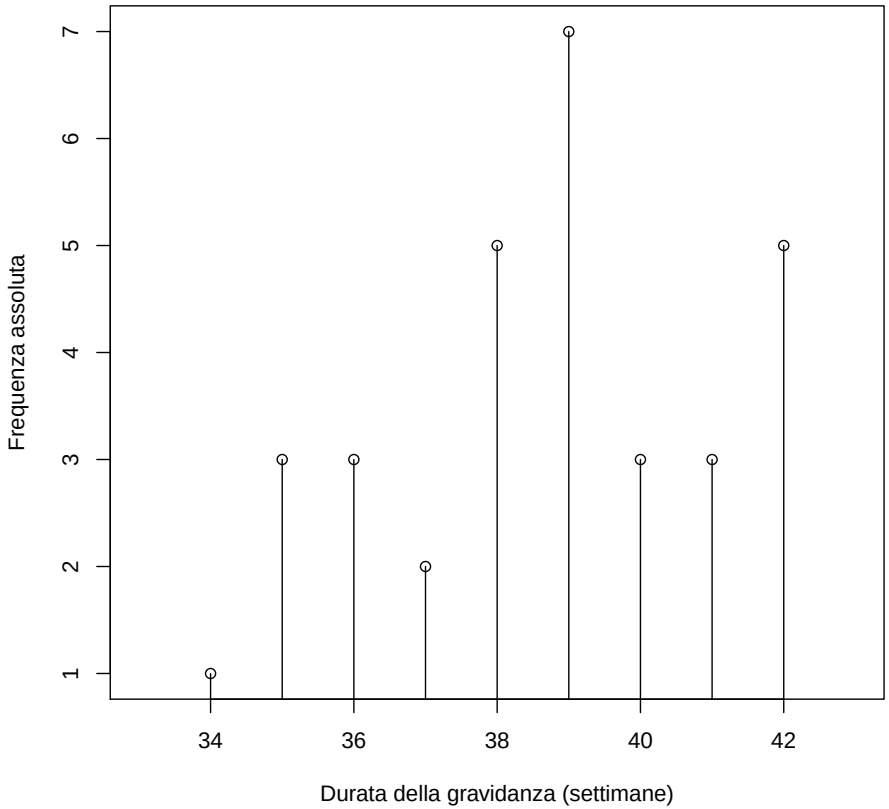
Il grafico è stato costruito ponendo rappresentando ogni modalità con una fetta di torta proporzionale di superficie pari alla sua frequenza:

$$\textit{angolo} = 360 \cdot \textit{frequenza assoluta}/n$$

o

$$\textit{angolo} = 360 \cdot \textit{frequenza relativa}$$

Esempio: dataset `babies`

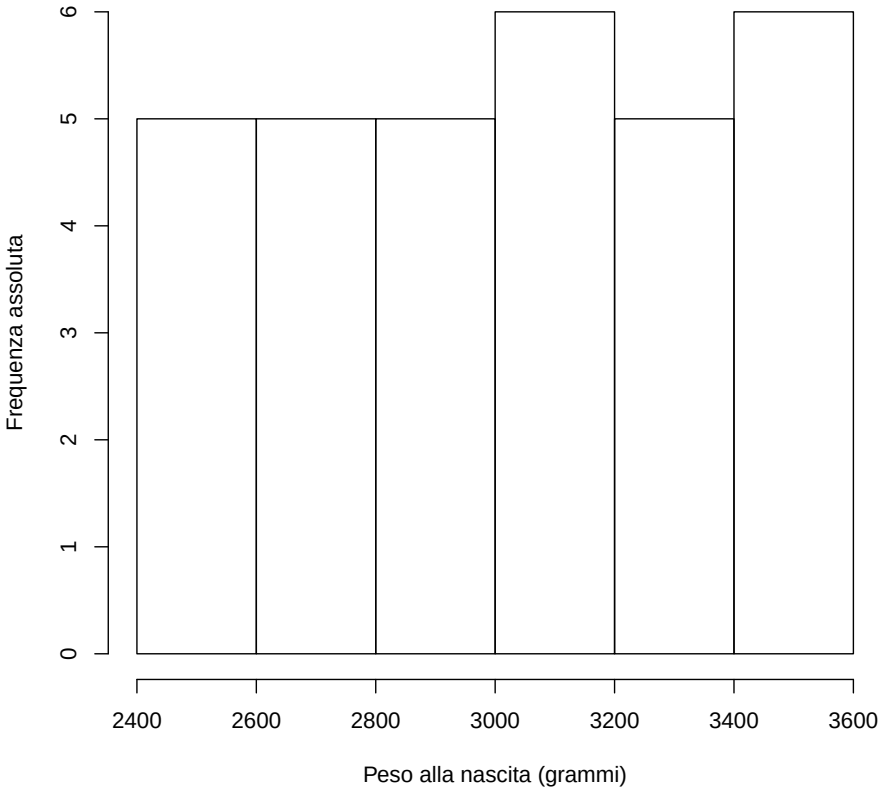


Il grafico è stato costruito ponendo

$$\text{asse}x = \begin{pmatrix}\text{modalità riportate}\\ \text{nella distribuzione}\\ \text{di frequenza}\end{pmatrix}$$

$$(\text{altezza barre}) = (\text{frequenze assolute})$$

Esempio: dataset babies



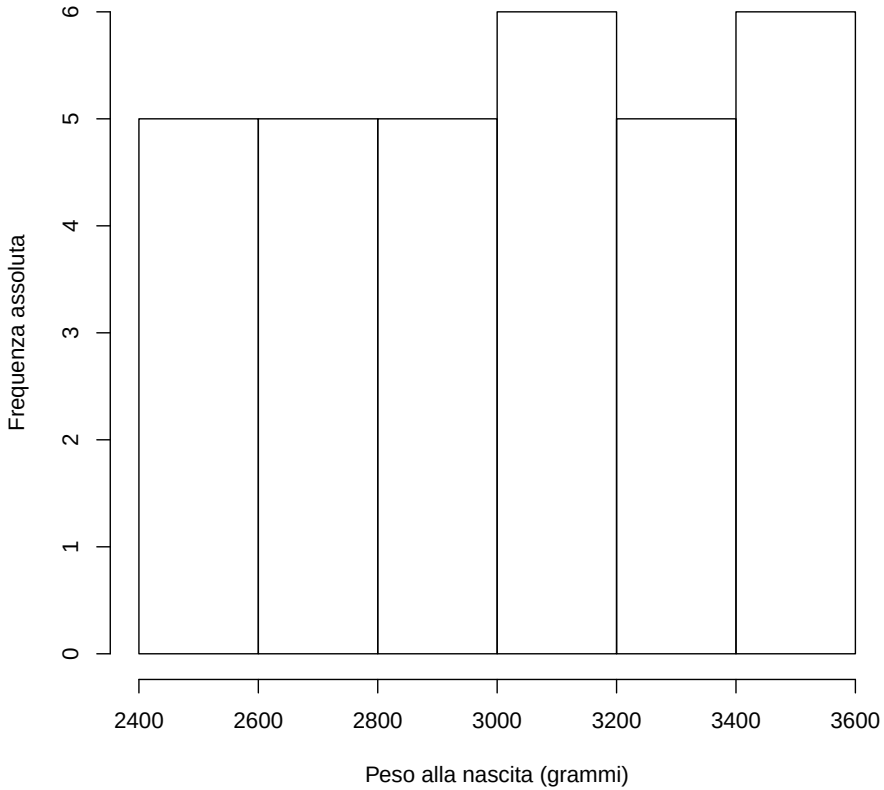
Il grafico è stato costruito ponendo

$$\text{(base rettangoli)} = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

$$\text{(area rettangoli)} \propto \text{(frequenze assolute)}$$

Il simbolo $\propto$ significa "proporzionale a".

Esempio: dataset babies



Il grafico è stato costruito ponendo

$$(\text{base rettangoli}) = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

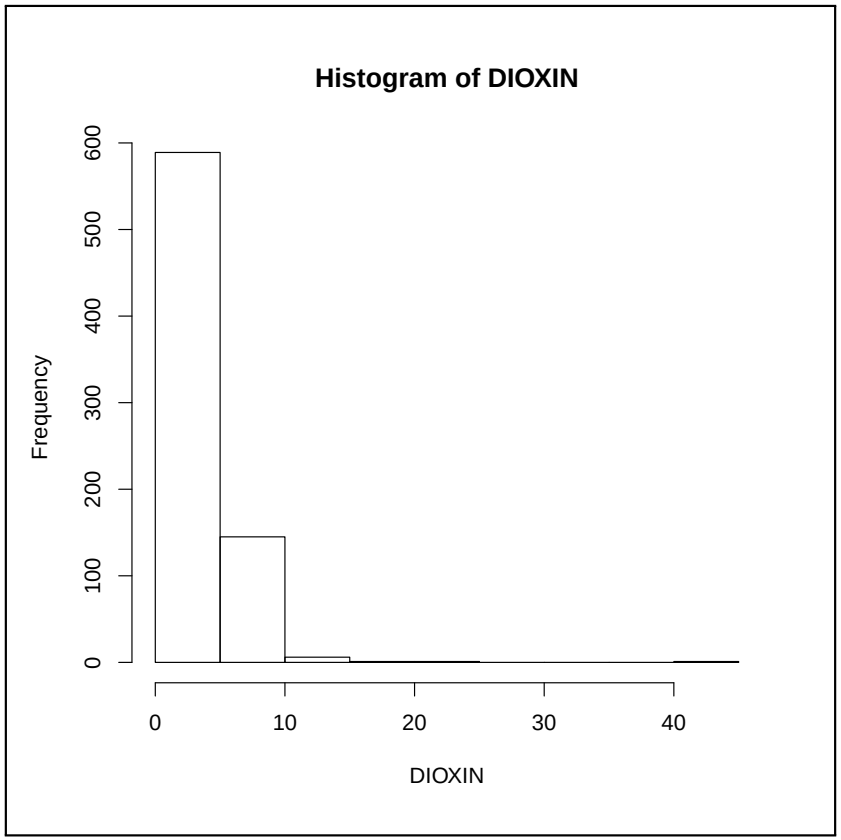
$$(\text{area rettangoli}) \propto (\text{frequenze assolute})$$

Il simbolo $\propto$ significa "proporzionale a".

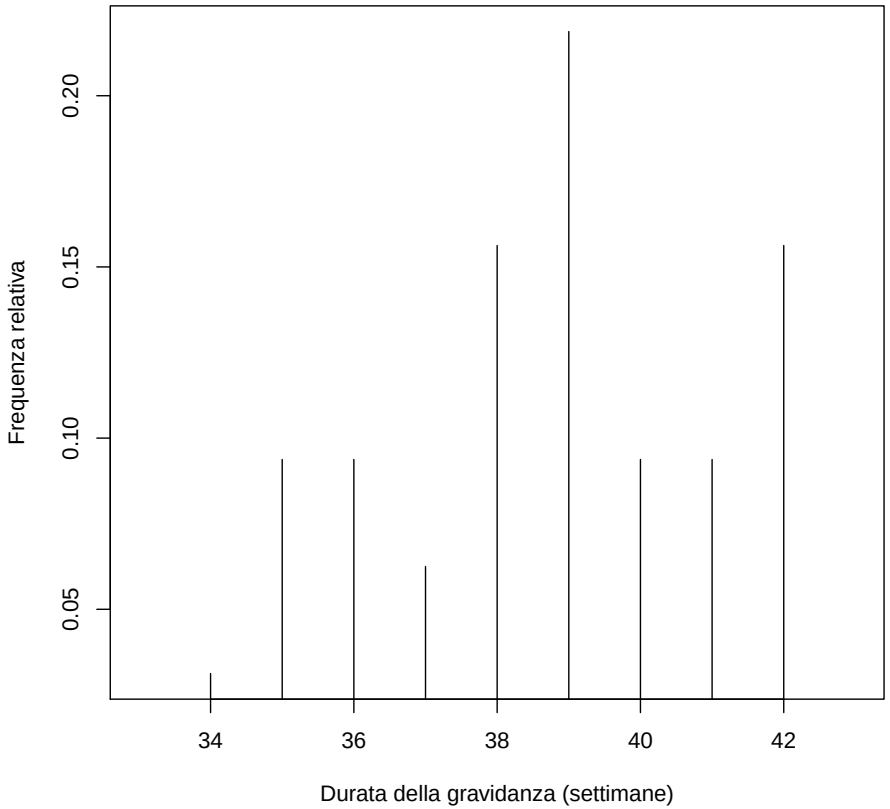
Essendo l'area dei rettangoli uguale a base$\times$altezza, se le gli intervalli hanno uguale ampiezza, di fatto l'altezza coincide con (o è proporzionale a) la frequenza assoluta:

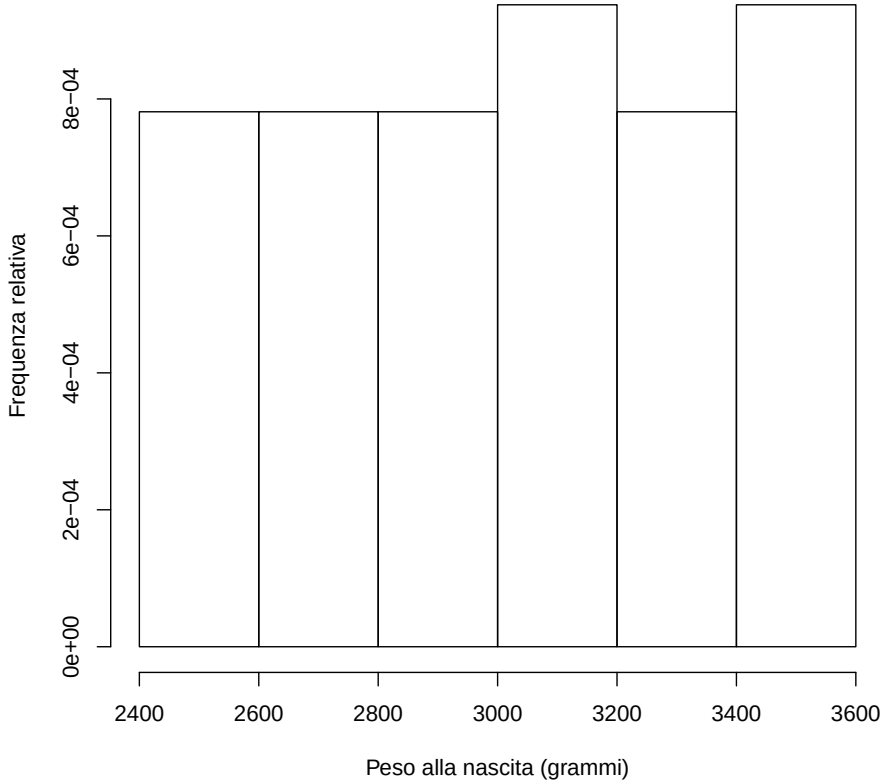
$$(\text{altezza rettangoli}) = (\text{frequenze assolute})$$

Esempio: dataset `vets`



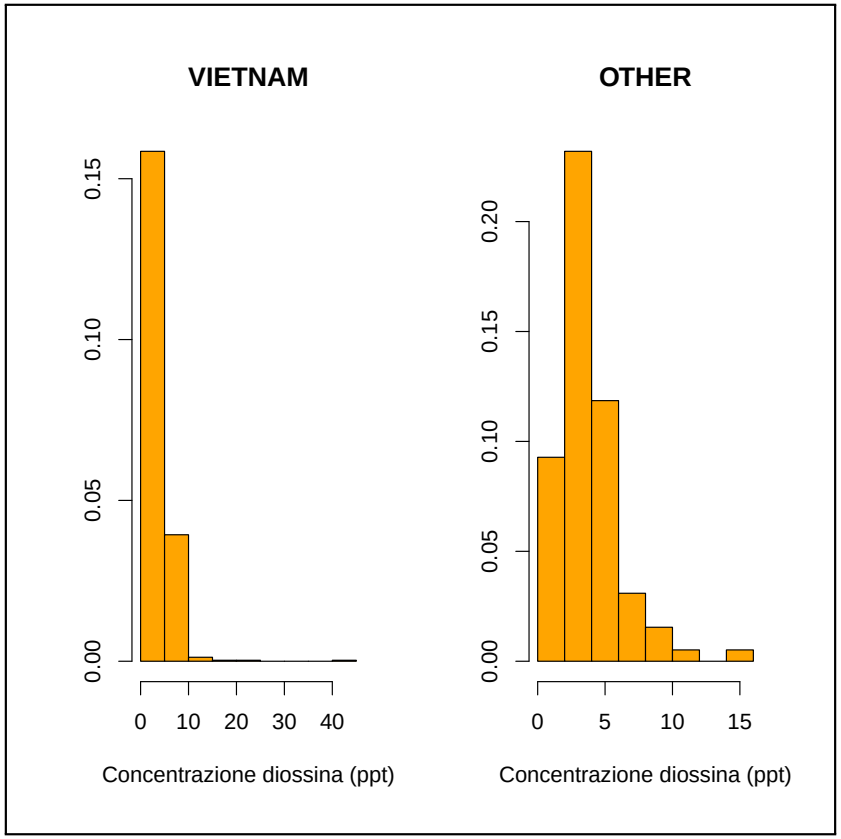
Vale anche per le frequenze relative





Vale anche per le distribuzioni condizionate

Esempio: dataset `vets`



Terminologia

- Per variabili categoriali, la rappresentazione prende il nome di *diagramma a torta* o *diagramma a barre*.
- Per variabili discrete, la rappresentazione prende il nome di *diagramma a barre*.
- Per variabili continue, la rappresentazione prende il nome di *istogramma*.

Osservazioni

Le rappresentazioni grafiche di distribuzioni di frequenza

- forniscono una immagine della distribuzione dei dati: barre o scatole più alte rappresentano modalità più frequenti;

Osservazioni

Le rappresentazioni grafiche di distribuzioni di frequenza

- forniscono una immagine della distribuzione dei dati: barre o scatole più alte rappresentano modalità più frequenti;
- aiutano a decrivere la *forma* della distribuzione dei dati;

Osservazioni

Le rappresentazioni grafiche di distribuzioni di frequenza

- forniscono una immagine della distribuzione dei dati: barre o scatole più alte rappresentano modalità più frequenti;
- aiutano a decrivere la *forma* della distribuzione dei dati;
- sono fortemente comunicative;

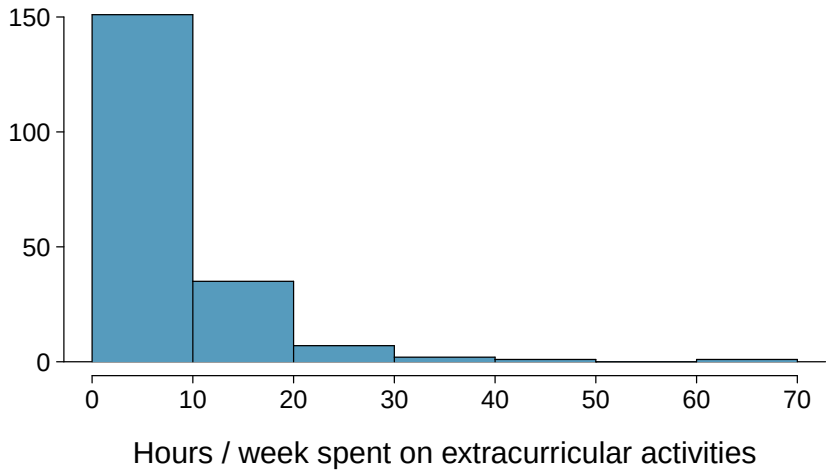
Osservazioni

Le rappresentazioni grafiche di distribuzioni di frequenza

- forniscono una immagine della distribuzione dei dati: barre o scatole più alte rappresentano modalità più frequenti;
- aiutano a decrivere la *forma* della distribuzione dei dati;
- sono fortemente comunicative;
- ma devono essere ben costruite!

Osservazioni: ampiezza delle classi degli istogrammi (cont)

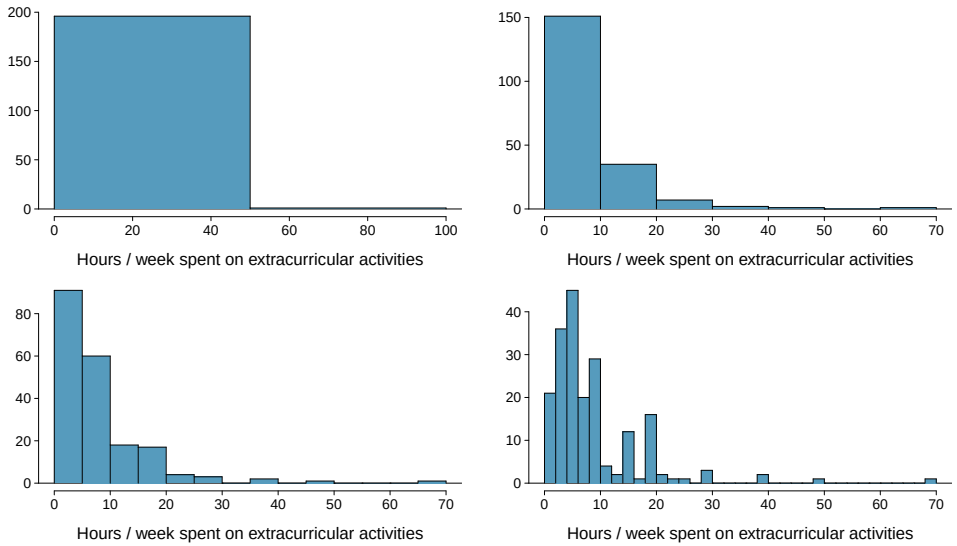
Esempio: ore impiegate settimanalmente da studenti americani in attività extra curricolari.



Osservazioni: ampiezza delle classi degli istogrammi (cont)

Esempio: ore impiegate settimanalmente da studenti americani in attività extra curricolari.

Quale di questi istogrammi è utile? Quale fornisce troppi dettagli? Quale nasconde troppo?



Osservazioni: ampiezza delle classi degli istogrammi (cont)

- Pochi intervalli, pochi dettagli.
- Troppi intervalli, troppi dettagli, probabilmente peculiari del campione a disposizione.
- È conveniente fare più di un grafico: provare differenti lunghezze per gli intervalli e poi scegliere.
- Il numero degli intervalli deve dipendere da come sono distribuiti i valori della variabile!