

Statistica I

Esercitazione 1: distribuzioni di frequenza

Tommaso Rigon

Università Milano-Bicocca



Descrizione del problema



- Il **diametro** del fusto di un albero viene misurato attraverso uno strumento chiamato "Cavalletto".
- La misura viene effettuata tenendo il cavalletto in posizione perpendicolare al fusto ad una altezza dal terreno di circa 1.30 m, con una precisione non superiore a 1 cm.
- Nell'autunno del 1999 sono stati misurati i diametri di $n = 1887$ **abeti rossi** presenti in una zona boschiva a San Vito di Cadore.

I dati grezzi

- Le misure sono elencate nella tabella seguente, in cui D rappresenta il **diametro** mentre n_j sono le corrispondenti **frequenze assolute**.

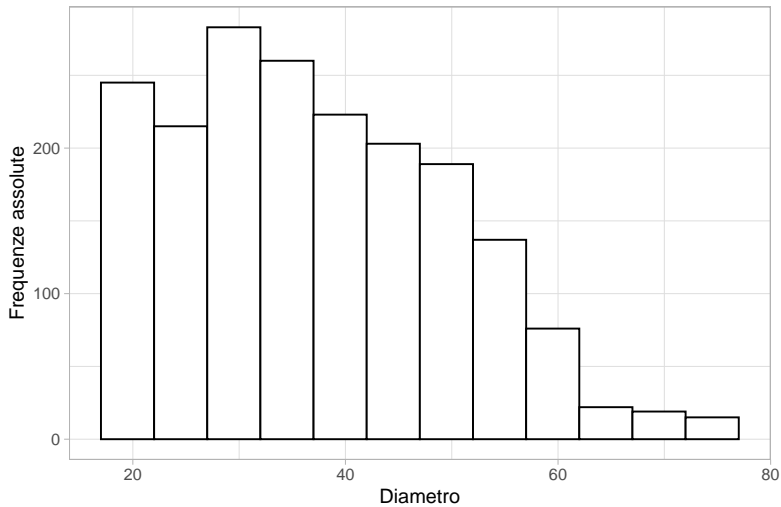
D	n_j	D	n_j	D	n_j	D	n_j	D	n_j	D	n_j
18	21	28	43	38	45	48	38	58	17	68	7
19	47	29	48	39	41	49	26	59	23	69	1
20	34	30	51	40	43	50	31	60	12	70	5
21	69	31	65	41	49	51	46	61	14	71	0
22	74	32	76	42	45	52	48	62	10	72	6
23	52	33	64	43	42	53	23	63	11	73	9
24	46	34	72	44	39	54	39	64	4	74	0
25	28	35	33	45	40	55	30	65	3	75	4
26	49	36	32	46	47	56	29	66	0	76	0
27	40	37	59	47	35	57	16	67	4	77	2

- Qual è l'**unità statistica** di questo problema? Qual è la **numerosità campionaria**?
- Per i dati riferiti al diametro del fusto si è soliti sintetizzare l'informazione in **classi equispaziate** di 5 cm.
- Si costruisca tale tabella, usando pertanto gli intervalli $(17, 22]$, $(22, 27]$, \dots , $(72, 77]$ e si ottengano:
 - le frequenze assolute,
 - le frequenze relative,
 - le frequenze cumulate assolute e relative.
- Si disegni l'**istogramma** dei dati utilizzando gli intervalli del punto precedente.
- Si faccia un disegno (anche approssimativo) della **funzione di ripartizione**.
- Quanto vale la funzione di ripartizione calcolata in 52, ovvero $F(52)$?
- Quanti sono gli alberi aventi diametro più piccolo di 42.5 cm?

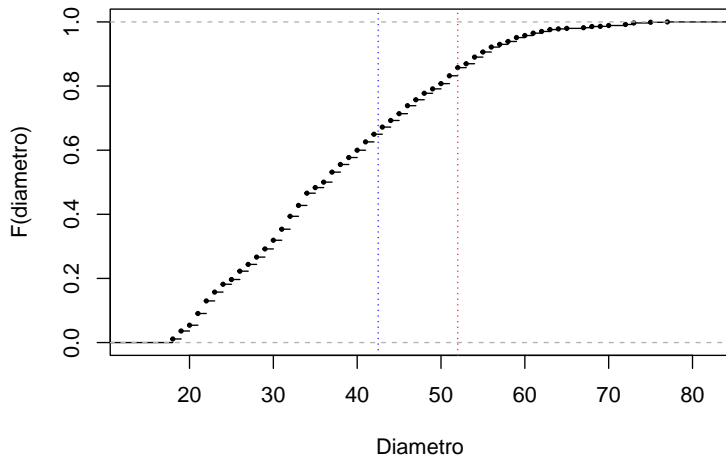
Tabella di frequenze

Classe	Freq. assolute	Freq. relative	Freq. cum. assolute	Freq. cum. relative
(17, 22]	245	0.130	245	0.130
(22, 27]	215	0.114	460	0.244
(27, 32]	283	0.150	743	0.394
(32, 37]	260	0.138	1003	0.532
(37, 42]	223	0.118	1226	0.650
(42, 47]	203	0.108	1429	0.757
(47, 52]	189	0.100	1618	0.857
(52, 57]	137	0.073	1755	0.930
(57, 62]	76	0.040	1831	0.970
(62, 67]	22	0.012	1853	0.982
(67, 72]	19	0.010	1872	0.992
(72, 77]	15	0.008	1887	1.000

Istogramma



Funzione di ripartizione



Commenti ai risultati

- Sia l'istogramma che la funzione di ripartizione mostrano che i diametri degli abeti rossi sono **circa compresi** tra i 20cm e gli 80cm.
- Il **centro della distribuzione**, a giudicare dai grafici, si attesta attorno ai 40cm.
- Tuttavia, la distribuzione dei diametri **non è uniforme**.
- I diametri sono uniformemente distribuiti fino ai 50-55 cm circa, salvo poi decrescere in frequenza. In altre parole, gli **alberi di grandi dimensioni** sono via via **più rari**.

Descrizione del problema



- Simon Newcomb ha misurato nel 1882 la **velocità della luce**.

Descrizione del problema

- Nello specifico, Newcomb ha misurato il **tempo** necessario alla luce per percorrere una **distanza** di 7442 metri. Newcomb ha ripetuto l'esperimento $n = 66$ volte.
- I dati x_1, \dots, x_{66} rappresentano le misurazioni della **velocità della luce** espresse in microsecondi.

Velocità della luce (μs). Dati di Newcomb

[1]	24.828	24.826	24.833	24.824	24.834	24.756	24.827	24.816	24.840
[10]	24.798	24.829	24.822	24.824	24.821	24.825	24.830	24.823	24.829
[19]	24.831	24.819	24.824	24.820	24.836	24.832	24.836	24.828	24.825
[28]	24.821	24.828	24.829	24.837	24.825	24.828	24.826	24.830	24.832
[37]	24.836	24.826	24.830	24.822	24.836	24.823	24.827	24.827	24.828
[46]	24.827	24.831	24.827	24.826	24.833	24.826	24.832	24.832	24.824
[55]	24.839	24.828	24.824	24.825	24.832	24.825	24.829	24.827	24.828
[64]	24.829	24.816	24.823						

Dati trasformati

- Risulta comodo **cambiare** la **scala dei dati**, per semplificarne la loro analisi.
- In particolare, consideriamo le differenze da 24.8 microsecondi ed esprimiamo queste **deviazioni** (o scarti) in nanosecondi. In simboli, vogliamo ottenere

$$y_i = 1000 \times (x_i - 24.8), \quad i = 1, \dots, 66.$$

- I dati y_1, \dots, y_{66} sono espressi nella **scala trasformata**.

Velocità della luce (scarti da 24.8, in ns). Dati di Newcomb

[1]	28	26	33	24	34	-44	27	16	40	-2	29	22	24	21	25	30	23
[18]	29	31	19	24	20	36	32	36	28	25	21	28	29	37	25	28	26
[35]	30	32	36	26	30	22	36	23	27	27	28	27	31	27	26	33	26
[52]	32	32	24	39	28	24	25	32	25	29	27	28	29	16	23		

Domande

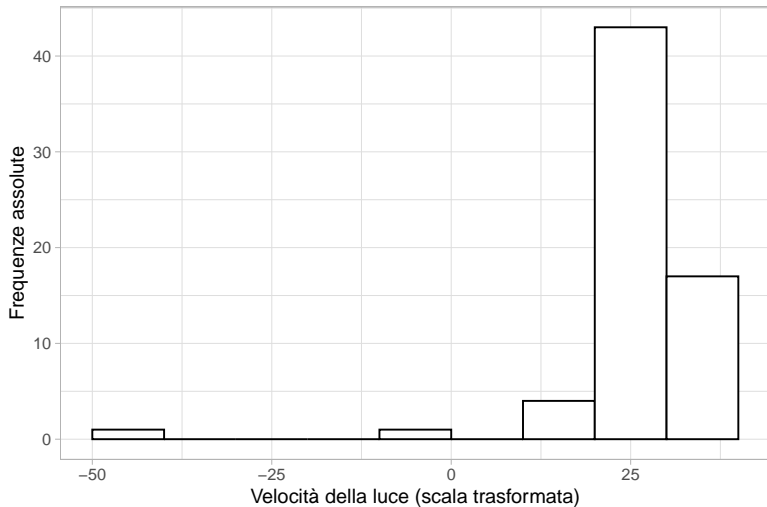
- Si ottengano il **minimo** ed il **massimo** delle misurazioni di Newcomb nella scala trasformata y_1, \dots, y_{66} .
- Si costruisca una **tabella** per la scala trasformata, utilizzando gli intervalli $(-50, -40]$, $(-40, -30]$, \dots , $(30, 40]$. Si ottengano quindi:
 - le frequenze assolute,
 - le frequenze relative,
 - le frequenze cumulate assolute e relative.
- Nella scala trasformata, si calcoli la **frazione di misurazioni** contemporaneamente maggiore di 20 e minore o uguale di 30.
- Nella scala trasformata, si disegni un **istogramma** usando gli intervalli $(-50, -40]$, $(-40, -30]$, \dots , $(30, 40]$.
- Si ottenga un grafico approssimativo della **funzione di ripartizione** nella scala originale ed in quella trasformata. Si **commentino i risultati**.

Tabella di frequenze

- Il **minimo** dei dati y_1, \dots, y_{66} vale -44 , il **massimo** vale 40 .

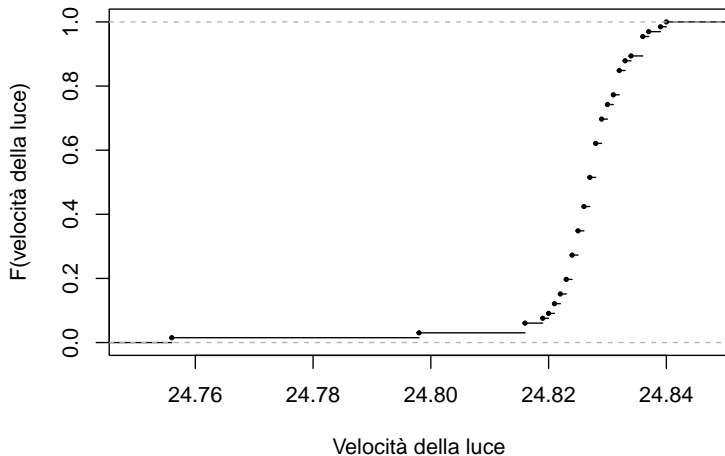
Classe	n_j	f_j	N_j	F_j
$(-50, -40]$	1	0.015	1	0.015
$(-40, -30]$	0	0.000	1	0.015
$(-30, -20]$	0	0.000	1	0.015
$(-20, -10]$	0	0.000	1	0.015
$(-10, 0]$	1	0.015	2	0.030
$(0, 10]$	0	0.000	2	0.030
$(10, 20]$	4	0.061	6	0.091
$(20, 30]$	43	0.652	49	0.742
$(30, 40]$	17	0.258	66	1.000

Istogramma



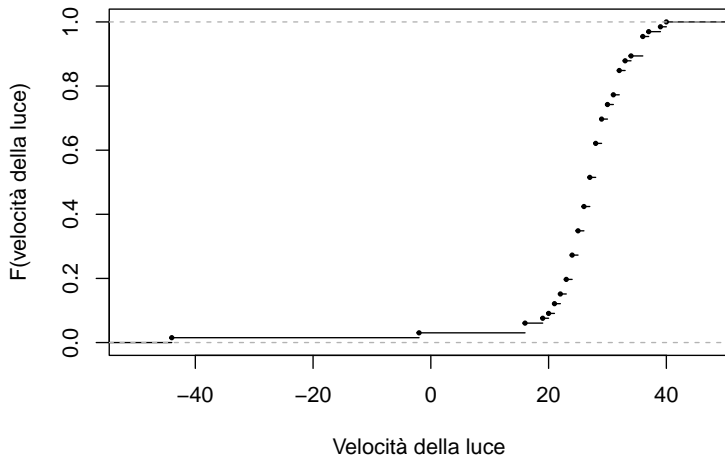
Funzione di ripartizione I

Scala originaria (microsecondi)



Funzione di ripartizione II

Scala trasformata: differenze in nanosecondi da 24.800



Commenti ai risultati

- Sia l'istogramma che la funzione di ripartizione mettono in chiara evidenza che ci sono due esperimenti **anomali**, corrispondenti alle osservazioni $y_6 = -44$ e $y_{10} = -2$.
- La quasi totalità dei dati si concentra in valori compresi tra 20 e 40. Questo porta ad **ipotizzare** che qualcosa sia andato storto nell'esecuzione di questi due esperimenti.
- Una possibilità è **escludere** questi valori dall'analisi. Rimosse tali osservazioni, è evidente che il **centro della distribuzione** sia attorno al valore 30 (e non attorno a 0!).
- In alternativa, potremmo utilizzare delle tecniche **robuste**, come la mediana, che sono affrontate nell'Unità C.
- Sulla base delle recenti misurazioni, il **vero valore** della velocità della luce, espresso nella scala trasformata, è di 33.0, che sembra confermare questo ragionamento.