Esercizi di Statistica Descrittiva

a cura di
Alessandra Dalla Valle e Carlo Gaetan

Indice

Capitolo 1

Rappresentazione grafica

Il diametro del fusto di una pianta viene misurato attraverso uno strumento chiamato "Cavalletto". Esso è simile ad un grosso calibro. La misura viene effettuata tenendo il cavalletto in posizione perpendicolare al fusto ad una altezza dal terreno di $1.30m$ con una precisione non superiore al cm. Nell'autunno del 1999 sono stati misurati i diametri di 1887 Abeti rossi (*Picea Abiens*) presenti in una zona di bosco a San Vito di Cadore. Le misure sono le seguenti:

D	f	D	f	D	f	D	f	D	f	D	f
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

D: diametro in cm, f: frequenza assoluta.

- Si svolga un'analisi preliminare dei dati in modo da sintetizzare l'informazione raccolta.
- Di solito, per dati riferiti al diametro del fusto, l'informazione disponibile è già parzialmente sintetizzata attraverso l'uso di classi (chiamate

classi diametriche) di ampiezza $5cm$ centrate nei valori: 20,25,30,$\cdots$,65,70, 75. Si costruisca questo nuovo insieme di dati e si svolga l'analisi come al punto precedente.

- Quali sono le differenze riscontrate nell'analisi tra i due insiemi di dati (originale e parzialmente sintetizzato). Si cerchi di spiegarne i motivi.

Capitolo 2

Indici di posizione

2.1 ACAP

Lungo una strada rettilinea sono collocati cinque condomini: A, B, C, D ed E. Il comune desidera determinare la posizione ottimale per un supermercato. Conoscendo i seguenti dati:

Condominio	A	B	C	D	E
N^{O} di inquilini	6	6	20	12	8

Distanza di A da B	=	1000 m
Distanza di B da C	=	1000 m
Distanza di C da D	=	100 m
Distanza di D da E	=	50 m

si dica qual è la soluzione che minimizza il disagio per raggiungere il supermercato quando:

- il disagio cresce linearmente con la distanza;
- il disagio cresce con il quadrato della distanza.

2.2 Foglie di Platano

È riportata di seguito una tabella che riassume le lunghezze di 100 foglie di platano, misurate al millimetro più prossimo, dopo 10 giorni di siccità.

Lunghezza	Freq.	Freq. cumulata
$120 \dashv 135$	10	10
$135 \dashv 145$	20	30
$145 \dashv 150$	60	90
$150 \dashv 165$	10	100

È noto che l'effetto di un giorno di pioggia continuata è di aumentare la lunghezza delle foglie di una quota percentuale pari al 10% più una quota fissa pari a 0.5 mm.

- Si rappresentino le osservazioni tramite un istogramma e si calcolino mediana, scarto interquartile, media e varianza.
- Si calcolino mediana, scarto interquartile, media e varianza dopo un giorno di pioggia continuata.

2.3 Gran Premio del Canada

Il Gran premio del Canada viene interrotto per la pioggia dopo 42 giri, quando Ayrton Senna era in terza posizione con una velocità media di $204 km/h$. I restanti 21 giri vengono percorsi da Senna ad una media di $126 km/h$. Al termine si sommano i tempi delle due frazioni di gara e Senna risulta il vincitore.

- Calcolare la velocità media del vincitore della gara.
- Calcolare la durata delle due frazioni di gara e quella complessiva, sapendo che il circuito misura $5 km$. Verificare la correttezza del valore medio calcolato al punto precedente.

2.4 Velocità di crescita dei batteri

È stato osservato il processo di crescita di una popolazione di batteri, rilevando il numero di individui, ad intervalli regolari di un'ora ciascuno.

t_0	t_1	t_2	t_3	t_4
10	170	3290	42230	173510

Si vuole conoscere il tasso medio di crescita della popolazione di batteri.

2.5 Blocchi di marmo

Da una cava di marmo vengono ricavati 25 blocchi cubici, la cui distribuzione secondo la dimensione del lato (l) misurata in metri risulta la seguente:

Dimensione	$1 < l \leq 2$	$2 < l \leq 4$	$4 < l \leq 5$	$5 < l \leq 6$
Blocchi	8	9	6	2

Si calcoli la dimensione media del lato di tali blocchi di marmo.

Capitolo 3

Indici di variabilità

3.1 Un esperimento di misurazione

Simon Newcomb misurò il tempo necessario affiché la luce andasse dal suo laboratorio sul fiume Potomac ad uno specchio alla base del Washington Monument e ritornasse indietro, per una distanza totale di circa 7400 metri. Egli fece 66 misurazioni qui sotto riportate.

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

1. Si rappresenti la distribuzione dei dati mediante un diagramma a scatola.

2. Si stimi la velocità della luce.

3.2 Strani indici ?

Dire se gi indici elencati di seguito possono o non possono essere interpretati come una misura di posizione oppure di variabilità di un certo insieme di dati tutti positivi, $y_1, \ldots, y_n$ (Q_i indica il quartile i-simo):

1. $\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - y_j)$;
2. $Q_2 - Q_1 + |Q_2 - Q_3|$;
3. $\exp\left\{\frac{1}{n} \sum_{i=1}^{n} \log(y_i)\right\}$

3.3 Tubi catodici

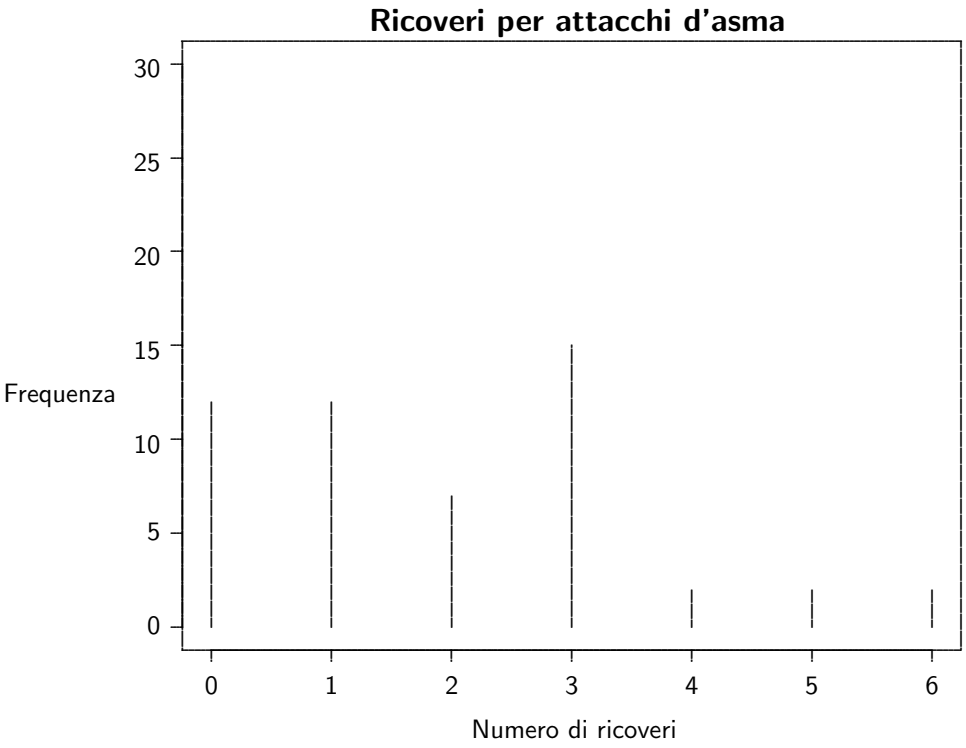
Una fabbrica di televisori produce due tipi di tubi catodici: il tipo e il tipo B. I tubi catodici hanno tempi di durata media, rispettivamente di 1495 ore e 1875 ore e scarti quadratici mdei rispettivamente di 280 ore e 310 ore. In generale, è preferibile il tubo catodico che ha la durata più alta e variabilità più bassa. Quale dei due tubi catodici è preferibile?

3.4 Lampadine

Un'azienda produttrice di lampadine controlla la durata dei suoi prodotti facendo funzionare 200 lampadine ininterrottamente fino a che si rompono. A determinati istanti di tempo si effettua un controllo e si verifica quante in totale non sono più funzionanti. I dati sono riportati nella tabella seguente

Tempo	10	30	1000	150	200	400	$\cdots$
Rotture	2	10	40	80	120	170	200

1. Si valuti la durata media e si dia un indice di posizione appropriato per sintetizzare la distribuzione delle durate.
2. Si calcoli un opportuno indice di variabilità.



3.5 Ricoveri

In un ospedale si registrano il numero di ricoveri per attacco d'asma nel corso di un anno. La sola informazione disponibile è data dal seguente istogramma

1. Si dica quali dei valori $F(0) = 0.231$, $F(1.5) = 0.346$, $F(1) = 0.462$, $F(8) = 1.000$, $F(9) = 1.200$ della funzione di ripartizione empirica F sono compatibili con l'istogramma.

2. Si calcoli (in maniera approssimata un indice di posizione e uno di variabilità.

3.6 Un po' di algebra ...

Per 4 osservazioni, la somma dei quadrati degli scarti da un valore K = media + 3 è uguale a 100. Se il coefficiente di variazione è uguale a 0.20, calcolare la media e la varianza di queste 4 osservazioni.

Capitolo 4

Variabili qualitative

4.1 Marmotte

Una regione delle alpi è stata suddivisa in 6 sottoaree di uguale dimensione e conformazione. Per ogni sottoarea è stato svolto un censimento delle marmotte presenti. I risultati ottenuti sono stati:

Sottoarea	A	B	C	D	E	F
Frequenza	24	7	10	4	35	13

Si valuti se le marmotte sono equamente presenti nelle 6 sottoaree.

4.2 Morbo di Hodgkin

Nella tabella seguente sono riportati i risultati relativi ad uno studio clinico per 200 pazienti affetti morbo di Hodgkin. Sono stati rilevati due tipi istologici di linfonodi, PL (predominanza di linfociti) e SN (sclerosi nodulare). Dopo tre mesi di di trattamento è stato assegnato un giudizio clinico sull'efficacia tramite la classificazione *positivo*, *parziale*, *nessuno*.

	ESITO			
TIPO ISTOLOGICO	positivo	parziale	nessuno	totale
PL	74	18	12	104
SN	68	16	12	96

Si chiede di valutare se l'efficacia del trattamento è stata la stessa per i due tipi di linfonodi.

Capitolo 5

Regressione lineare semplice

5.1 Alimenti

La seguente tabella a doppia entrata riporta la quantità di acqua in grammi (X) e di energia in kcal. (Y) per 100 grammi di prodotto in 8 alimenti (cereali e derivati).

alimenti	cont.acqua	energia
pane	31.0	276
grissini	8.5	433
crackers	6	428
fette biscottate	4	410
biscotti	2.2	418
pasta	12.4	356
riso	12.9	362
pizza	40.5	247

1. dire se esiste tra energia e contenuto in acqua vi è una qualche relazione;
2. calcolare la varianza della variabile $X - Y$;
3. calcolare la correlazione esistente tra energia e contenuto di acqua.

5.2 Statura Padre-Figli

La tabella che segue mostra le stature per un gruppo di padri e figli. Presupponendo una relazione lineare tra le stature dei padri e le stature dei figli, si dica che statura ci si aspetta per il figlio di un padre alto 170.5 cm.

Padri	Figli
165	167
170	169
180	181
172	171
179	180
174	176
176	180
168	171
181	179
173	174
170	173
178	176
176	178

5.3 Pino Silvestre

Sono stati misurati i diametri (in cm) e le altezze (in m) di alcuni pini silvestri:

D	Altezza				
27	18	17.5	17	20	16.5
29	19	20	23	23	
30	22	21	24		
33	22	20	26	23	
34	25	23	25	26	

Si costruisca un modello di regressione appropriato.

5.4 Crescita dei batteri

Nella tabella che segue sono riportate 20 misurazioni eseguite su diversi campioni riguardanti il numero di batteri (in milioni) e il tempo (in ore). Si vuole individuare un appropriato modello di regressione.

Batteri	Tempo
212.048	7.598
45.316	5.760
15.437	4.606
53.508	5.893
60.429	5.697
73.180	6.507
1677.691	10.893
625.179	9.326
918.117	10.879
489.051	9.713

5.5 Battiti cardiaci e temperatura

Nel tabella che segue troviamo alcune misurazioni riguardanti la temperatura a riposo in gradi centigradi, il genere e il numero di battiti cardiaci.

Maschi		Femmine	
Temperatura	Numero di battiti	Temperatura	Numero di battiti
35.7	70	36.0	75
36.2	82	36.5	61
36.3	78	36.6	79
36.6	58	36.7	81
36.7	78	36.8	57
36.7	73	36.9	73
36.8	86	37.0	86
36.9	68	37.1	65
37.0	70	37.1	69
37.1	78	37.3	80
37.3	83	37.7	79

Si vuole conoscere se un aumento del numero di battiti cardiaci porta ad un aumento della temperatura corporea. A tale scopo si costruisca un modello di regressione opportuno.

5.6 Regressione con variabili standardizzate

Si considerino due variabili statistiche X ed Y e siano $\hat{\alpha}$ e $\hat{\beta}$ le stime secondo i minimi quadrati dei parametri del modello

$$Y = \alpha + \beta X + \text{errore}.$$

Si consideri la variabile $Z = \frac{X-\mu_x}{\sigma_x}$ ed il modello

$$Y = a + bZ + \text{errore}^*.$$

- Si esprimano i coefficienti $\hat{a}$ e $\hat{b}$ ottenibili tramite i minimi quadrati in funzione di $\hat{\alpha}$ e $\hat{\beta}$.
- Cosa si può dire circa le devianze residue dei due modelli?

5.7 Dati di Anscombe

Sono date le seguenti osservazioni da 4 regressori X_1, X_2, X_3, X_4 e 4 variabili dipendenti Y_1, Y_2, Y_3, Y_4.

X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

1. Si calcoli il coefficiente di correlazione e la retta di regressione tra (X_i, Y_i) $i = 1, \ldots, 4$.
2. In base ai risultati ottenuti al punto precedente, si può ritenere che le coppie di variabili siano simili ?

3. Si disegni il grafico di dispersione per ciascuna coppia e si rappresenti i valori della corrispondente retta di regressione. Cosa dimostrano questi grafici ?