

ESERCIZI DI STATISTICA

ES1

Data la seguente serie di dati su Sesso e Altezza di 8 pazienti:

- 1) Riempire opportunamente due tabelle per rappresentare le distribuzioni di frequenze dei due caratteri, secondo il sottostante modello (per le classi di Altezza, considerare le classi 140-|160; 160-|170; 170-|200). Aggiungere poi un grafico.
- 2) Successivamente: compilare una tabella doppia, ed evidenziare l'eventuale relazione fra Sesso e Altezza attraverso opportune sintesi statistiche.
- 3) Verificare la significatività statistica al livello del 5% della relazione fra Sesso e Altezza usando la tabella doppia
- 4) Verificare la significatività statistica al livello del 5% della relazione fra Sesso e Altezza usando un T-test

id	Altezza, cm	Sesso: 1=M, 2=F
1	165	M
2	157	F
3	168	F
4	178	M
5	171	F
6	182	M
7	182	M
8	153	F

modalità	freq. assoluta	freq. percentuale	freq. cumulata

ES 2

186 pazienti hanno fatto una terapia per una certa malattia; 122 hanno seguito la terapia A, gli altri 64 hanno seguito la terapia B. Nel gruppo A, hanno risposto 37 soggetti. Nel gruppo B, hanno risposto 32 soggetti.

- 1) Qual è la probabilità complessiva di risposta? Fra i rispondenti, quanti avevano fatto il trattamento B?
- 2) Quale trattamento sembra migliore? Di quanto?
- 3) La differenza è statisticamente significativa al livello del 5%?

ES 3

Calcolare la media e la classe mediana della seguente distribuzione del numero di infermieri in 23 strutture di ricovero e cura private:

infermieri	n
0 – 10	6
10 – 20	13
20 – 40	4
	23

ES 4

Per 6 pazienti sono noti i valori dell'emoglobina registrati prima e dopo una chemioterapia: si desidera conoscere la riduzione media (calcolata facendo la differenza prima-dopo).

- 1) Calcolare il valore medio "prima" e "dopo", e il valore medio della riduzione. Che relazione c'è fra queste medie?
- 2) Calcolare la deviazione standard dell'emoglobina "prima" e "dopo" la terapia, e della riduzione dell'emoglobina: vale la stessa relazione?

prima	dopo
13.0	9.4
12.8	11.5
11.0	11.5
13.2	13.1
12.5	10.2
11.9	12.0

ES 5

Un certo trattamento è utilizzato in due centri diversi, A e B; i soggetti del centro A sono 25 e hanno in media 54 anni; i soggetti trattati nel centro B sono 62 e hanno in media 58 anni. Qual'è l'età media fra tutti i soggetti che fanno uso del trattamento?

Es 6

Le donne in gravidanza (entro il 4° mese) che vengono seguite in un centro dietologico pesano rispettivamente (pesi in kg): 64.3; 65.2; 70.0; 54.5; 58.8; 81.5; 61.0; 62.0. Qual'è la media? e la mediana? I dati suggeriscono una forte asimmetria della distribuzione del Peso?

ES 7

La seguente serie di dati riguarda una casistica di 10 soggetti adulti maschi; consideriamo l'età, il valore della FEV1 (Forced Espiratory volume in 1 second) e la pressione diastolica. Calcolare media e deviazione standard dei tre caratteri. Dire poi quale è il carattere più variabile, fornendo una valutazione quantitativa della differenza.

età	FEV1	pressione
25	2.5	85
32	1.8	71
28	1.5	92
21	2.5	80
33	4.5	87
33	2.1	83
34	3.4	70
24	1.2	101
41	2.8	90
26	3.9	83

ES 8

I quartili dell'età di un collettivo di partecipanti ad un trial clinico erano nell'ordine 27, 41 e 59.

a) Vuol dire che:

- ☐ 1 su 4 era più giovane di ... anni
- ☐ 1 su 4 era più vecchio di ... anni
- ☐ 2 su 4 erano fra ... e ... anni
- ☐ la metà aveva più di ... anni

b) Si sa inoltre che media e deviazione standard erano rispettivamente pari a 42 e 12. Secondo questi dati, si può capire se la distribuzione sembra Normale o no?

c) quale indice di posizione è adatto per descrivere sinteticamente la distribuzione?

ES 9

La distribuzione del peso di un gruppo di soggetti con disabilità motorie è approssimativamente Normale, con media 72 e deviazione standard 8. Individuare un intervallo di valori centrato sulla media tale che:

- a) contiene il 95% dei valori osservati
- b) contiene praticamente tutti i valori osservati (e quindi coincide con il range)
- c) contiene il 50% dei valori osservati

ES 10

Un medico dietologo propone un nuovo tipo di dieta (A) per facilitare la riduzione di peso. Decide di confrontarne l'efficacia rispetto al tipo di dieta che prescriveva precedentemente (B). A tal fine, usa i dati dei risultati di due gruppi di pazienti, 39 trattati con la vecchia dieta B, che hanno perso in media 2.9 kg (std=1.2kg), e 34 trattati con la nuova dieta A, che hanno avuto una riduzione di peso media di 3.5 kg (std=1.1kg).

E' corretto affermare che la dieta A è migliore della dieta B? Effettuare un test di ipotesi (T-test) per verificare la significatività della differenza: usare sia il metodo delle regioni di rifiuto (livello $\alpha=5\%$) che il calcolo del p-value.

SOLUZIONI

ES 1

Carattere Sesso:

modalità	freq. assoluta	freq. percentuale	freq. cumulata*
M	4	50%	
F	4	50%	
tot	8	100%	

* no: il carattere Sesso è qualitativo sconnesso, non è appropriato calcolare le cumulate. Lo facciamo invece sotto, essendo il carattere Altezza quantitativo (e quindi ordinato), continuo .

Un grafico adatto è il grafico a colonne, costituito da due rettangoli separati, uno per M e uno per F, con altezza proporzionali alle percentuali. In generale, è bene che l'asse verticale vada da 0 a 100, per non distorcere la percezione delle frequenze.

Carattere Altezza:

modalità	freq. assoluta	freq. percentuale	freq. cumulata	ampiezza della classe	densità di frequenza*
140 - 160	2	25%	25%	20	$=2/20=0.1$
160 - 170	2	25%	50%	10	$=2/10=0.2$
170 - 200	4	50%	100%	30	$=4/30=0.13$
tot	8	100%			

Un grafico adatto è l'istogramma, costituito da tre rettangoli contigui, ciascuno disegnato in corrispondenza degli estremi della relativa classe, e con altezza proporzionale alla sua densità di frequenza: l'area del rettangolo deve corrispondere alla frequenza della classe.

Tabella doppia:

	Altezza			
Sesso	140 - 160	160 - 170	170 - 200	Tot
M	0	1	3	4
F	2	1	1	4
tot	2	2	4	8

Per evidenziare la relazione tra Sesso e Altezza, calcoliamo separatamente per M e F le percentuali relative alle diverse classi di altezza: sono i profili riga, o distribuzioni condizionate dell'altezza:

	Altezza		
Sesso	140 - 160	160 - 170	170 - 200
M	0%	25%	75%
F	50%	25%	25%

Questa tabella suggerisce che i M sono più alti delle F. Osserviamo anche che per i M la Moda è la classe 170 - | 200, mentre per le F la Moda è 140 - | 160.

Per verificare la significatività sulla tabella, calcoliamo il Chi-Quadrato e facciamo il test (NOTA: facciamo questo solo allo scopo di esercitarci, ma il test NON E' VALIDO quando nelle celle della tabella si hanno frequenze <5)

.

Freq. attese:

1	1	2
1	1	2

Addendi:

1	0	0.5
1	0	0.5

Chi-Quadrato=3

Gdl=1*2=2

Soglia al livello 5%: 5.991

Conclusione: l'associazione non è statisticamente significativa

Per verificare la significatività della differenza fra le medie (di nuovo, solo per esercizio!! Ma il T-test NON è valido in quanto i campioni non sono grandi), raggruppiamo le osservazioni a seconda del Sesso:

	M	F
	165	157
	178	168
	182	171
	182	153
media	176.75	162.25
std	8.057	8.617
var	64.917	74.250

Calcoliamo la statistica test:

num = (176.75-162.25) = 14.5

per il den: s = $\text{radq}[(3 \cdot 64.917 + 3 \cdot 74.250)/(4+4-2)] = \text{radq}(69.69.583)=8.342$

den = 8.342 · $\text{radq}(1/4 + 1/4) = 5.898$

t = 14.5 / 5.898 = 2.458

Soglia al livello 5%: 1.96

Conclusione: l'associazione è statisticamente significativa al livello 5%

ES 2

Inseriamo i **dati** del problema in una tabella, e completiamola:

	Risposta		
Trattamento	no	si	Tot
A	85	37	122
B	32	32	64
tot	117	69	186

Ora ricaviamo alcune percentuali. Quelle di riga sono:

	Risposta		
Trattamento	no	si	Tot
A	85/122=69.7%	37/122=30.3%	100%
B	32/64=50.0%	32/64=50.0%	100%
tot	117/186=62.9%	69/186=37.1%	100%

Dunque la probabilità complessiva di risposta è pari a 37.1%.

La percentuale di pazienti rispondenti provenienti dal gruppo di trattamento B sono 32/69=46.4% (questa è una percentuale individuata guardando al profilo colonna, ovvero alla distribuzione del Trattamento condizionata a Risposta=sì).

Il trattamento migliore sembra essere il trattamento B: 50% prob. di risposta vs. 30.3%. Di quanto è migliore? Possiamo confrontare le percentuali di risposta facendone il rapporto, ossia calcolando il Risk Ratio:

$$RR=50/30.3=1.65$$

Dunque B ha una percentuale di risposta superiore del 65% rispetto a quella del trattamento A.

Per valutare se questa differenza è statisticamente significativa al livello del 5%, calcoliamo il Chi-Quadrato e facciamo il test.

Freq. attese	
76.74	45.26
40.26	23.74

Addendi

0.889	1.507
1.694	2.872

chi2=	6.962
-------	-------

Soglia al livello 5% per 1 gdl: 3.841

Conclusione: l'associazione è statisticamente significativa: il trattamento B è significativamente superiore al trattamento A.

Osserviamo che anche scegliendo un livello di significatività più basso pari a 0.01 – e quindi un test più prudente, che richiede maggiore “forza” dell'evidenza contraria ad H0 per poterla rigettare – abbiamo comunque il rifiuto di H0 (il valore soglia è 6.635).

ES 3

Il carattere Numero di infermieri relativo al campione di 23 strutture (unità statistiche) è di tipo

quantitativo, discreto, ma assimilabile a un continuo. La distribuzione viene data per classi di numero di addetti.

Per calcolare la media, dobbiamo prendere un valore rappresentativo per ciascuna classe: prendiamo il valore centrale, che si trova facendo (estremo inferiore + estremo superiore)/2. L'ammontare di infermieri per classe si trova poi moltiplicando questo valore centrale per la frequenza. La media è l'ammontare totale diviso per il numero di unità statistiche, 23.

Per individuare la classe che contiene la mediana, ci sono utili le frequenze cumulate.

infermieri	n	Valore x_i	$x_i \cdot n_i$	Freq. cumulata
0 - 10	6	5	30	6
10 - 20	13	15	195	19
20 - 40	4	30	120	23
	23		345	

$$\text{Media} = 345 / 23 = 15$$

Mediana: modalità di posto 12. Guardando alle freq. cumulate, capiamo che essa si trova nella classe 10-|20.

ES 4

La Riduzione è la differenza tra valore Prima e valore Dopo; in qualche caso può essere negativa poiché vi è stato invece un aumento di X. Possiamo calcolare tutte le 6 riduzioni, e farne una semplice media aritmetica.

Svolgendo l'esercizio, possiamo verificare che vale una proprietà chiamata di linearità per la media aritmetica: $\text{media}(\text{Prima} - \text{Dopo}) = \text{media}(\text{Prima}) - \text{media}(\text{Dopo})$.

Questa proprietà è SEMPRE vera, cioè non vale solo in questo esercizio. Più in generale, date alcune quantità, la media di una loro trasformazione lineare è uguale alla media delle quantità trasformata allo stesso modo: $\text{media}(a + bx) = a + b\bar{x}$

Questa proprietà è utile ad esempio se i dati devono essere sottoposti a cambiamento di scala e unità di misura, ad esempio per trasformare un dato relativo alla media di alcune temperature espresse in gradi Fahrenheit passando a gradi Celsius.

	prima	dopo	riduz	riduz^2
	13	9.4	3.6	12.96
	12.8	11.5	1.3	1.69
	11	11.5	-0.5	0.25
	13.2	13.1	0.1	0.01
	12.5	10.2	2.3	5.29
	11.9	12	-0.1	0.01
somma	74.4	67.7	6.7	20.21
somma/6	12.4	11.28333	1.116667	3.36833
		12.4-11.3=	1.116667	

Applicando il procedimento "rapido" di calcolo della deviazione standard, aggiungiamo i calcoli

nell'ultima colonna. La varianza è:

$$\text{var} = (3.36833 - (1.11667^2)) \cdot \frac{6}{6-1} = 2.545666$$

e la deviazione standard è ottenuta estraendo la radice quadrata: 1.595514

Osserviamo – e lo svolgimento dei calcoli è lasciato allo Studente per esercizio: la dev. st. dei valori “prima” è pari a 0.822912, quella dei valori “dopo” è 1.313646 – che per la deviazione standard non vale la linearità, in quanto il suo calcolo richiede operazioni di elevamento al quadrato e estrazione della radice che non godono della proprietà matematica di linearità: $(a + bx)^2 \neq a^2 + bx^2$

Dunque ad esempio in presenza di conversione di dati di temperatura da gradi Fahrenheit a gradi Celsius, non si potrà calcolare la deviazione standard trasformando allo stesso modo la dev. st. calcolata in Fahrenheit.

ES 5

Bisogna calcolare una media ponderata, cioè la media delle due medie (54 e 58) pesata per la numerosità dei due gruppi (25 e 62).

$$\text{Media} = (54 \cdot 25 + 58 \cdot 62) / (25 + 62) = 4946 / 87 = 56.85$$

ES 6

Disponiamo per comodità i dati in tabella; il calcolo della media è elementare, per la mediana dobbiamo attribuire i ranghi e individuare le modalità di posto 4 e 5 (avendo n=8 unità statistiche, donne in gravidanza):

valore x_i	rango r_i
64.3	5
65.2	6
70	7
54.5	1
58.8	2
81.5	8
61	3
62	4

$$\text{Somma valori} = 517.3 \rightarrow \text{Media} = 517.3 / 8 = 64.66$$

$$\text{Modalità centrali: } 62 \text{ e } 64.3 \rightarrow \text{Mediana} = (62 + 64.3) / 2 = 63.15$$

Visto che Media e Mediana non sono fra loro molto distanti, non sembra che i dati suggeriscano una forte asimmetria della distribuzione del Peso.

ES 7

Si tratta di 3 caratteri quantitativi continui. Media aritmetica e deviazione standard ne sintetizzano posizione e variabilità. La media è pari alla somma dei valori divisa per 10 (n=10 numerosità del campione). Per il calcolo della deviazione standard usiamo la ‘formula breve’. I calcoli sono riportati in tabella.

Per confrontare i tre caratteri in termini di variabilità, NON è sufficiente guardare alle 3 deviazioni

standard, che peraltro sono in unità di misura diverse e attengono a caratteri di natura diversa!

Dobbiamo calcolare la variabilità come misura relativa rispetto alla media, mediante il coefficiente di variazione. Il carattere più variabile risulta essere FEV1, 4 volte più variabile della pressione e 2 volte più variabile dell'età (aveva invece la deviazione standard più piccola ...)

id	età	FEV1	pressione	età^2	FEV1^2	pressione^2
1	25	2.5	85	625	6.25	7225
2	32	1.8	71	1024	3.24	5041
3	28	1.5	92	784	2.25	8464
4	21	2.5	80	441	6.25	6400
5	33	4.5	87	1089	20.25	7569
6	33	2.1	83	1089	4.41	6889
7	34	3.4	70	1156	11.56	4900
8	24	1.2	101	576	1.44	10201
9	41	2.8	90	1681	7.84	8100
10	26	3.9	83	676	15.21	6889
somma	297	26.2	842	9141	78.70	71678
somma/10	29.7	2.62	84.2	914.1	7.87	7167.8
				varianza	35.57	1.12
				dev.st.	5.96	1.06
				cv	20%	11%

ES 8

Punto a):

1 su 4 era più giovane di 27 anni: è la def. di primo quartile, $\frac{1}{4}=25\%$ delle osservazioni è minore di Q1

1 su 4 era più vecchio di 59 anni: analogamente, è la def. di terzo quartile, $\frac{3}{4}=75\%$ delle osservazioni è minore, il restante 25% è maggiore di Q3

2 su 4 erano fra ... e ... anni: ad esempio, fra Q1 e Q3, quindi fra 27 e 59; ma anche fra 0 (il minimo teorico) e 41, che è la mediana, oppure fra 41 e ... il massimo, che non conosciamo ... La risposta più appropriata è la prima, sebbene queste ultime due non siano errate.

la metà aveva più di 41 anni: definizione di Mediana

b) La media pari a 42 è molto vicina alla mediana, pari a 41, anche osservando che la differenza (pari a 1) è $\frac{1}{12}$ della deviazione standard, dunque piccola. Quindi la distribuzione osservata è simmetrica. Guardiamo però la posizione dei quartili osservati rispetto alla media: per avere una curva tipo Normale dovrebbero trovarsi a distanza di $0.67 \cdot 12 = 8$. Dunque dovrebbero essere pari a 34 e 50. I quartili osservati sono molto più distanti di quelli attesi sotto l'ipotesi di Normalità. Dunque no, la distribuzione osservata non era di tipo Normale; pur essendo simmetrica, la sua forma non era a campana; potrebbe trattarsi di una distribuzione con code molto alte e poche osservazioni nella parte centrale – eventualmente di tipo bimodale.

c) in considerazione delle osservazioni appena fatte, né media né mediana sono adeguatamente rappresentative della distribuzione; se la distribuzione fosse di tipo bimodale, si dovrebbero calcolare le due mode, ovvero le medie (o meglio le mediane) delle due sottopopolazioni.

ES 9

Dobbiamo utilizzare le proprietà della Normale.

Nell'intervallo $\text{media} \pm 2 \cdot \text{dev.st.}$ cade all'incirca il 95% dei valori (per un valore teorico più esatto, si dovrebbe usare 1.96 al posto del fattore 2) . Questo risponde al quesito a) . Analogamente, per il quesito b) costruiamo l'intervallo di raggio $3 \cdot \text{dev.st.}$, che contiene il 99.7% dei valori:

$$\text{a) } 72 \pm 2 \cdot 8 = (56,88)$$

$$\text{b) } 72 \pm 3 \cdot 8 = (48,96)$$

Per l'ultimo punto, osserviamo che l'intervallo centrato sulla media (=mediana) che contiene il 50% delle osservazioni è, per definizione dei quartili, l'intervallo (Q1,Q3), dunque calcoliamo i due quartili con la nota formula:

$$\text{c) } 72 \pm 0.67 \cdot 8 = (66.64, 77.36)$$

ES 10

Applichiamo le formule del t-test:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(34 - 1)1.1^2 + (39 - 1)1.2^2}{34 + 39 - 2}} = 1.156$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{3.5 - 2.9}{1.156 \sqrt{\frac{1}{34} + \frac{1}{39}}} = 2.22$$

Applicando il metodo della regione rifiuto, possiamo rigettare l'ipotesi di base di assenza di differenza fra la media di A e la media di B al livello di significatività del 5% (il valore soglia per il test bilaterale è 1.96).

Volendo valutare la significatività statistica calcolando il p-value, andiamo sulle tavole e in corrispondenza di 2.22 leggiamo 0.987; dunque l'area in una delle due code esterne è $1 - 0.987 = 0.013$, e la probabilità complessiva di andare in una delle due code esterne (il p-value del test bilaterale) è 0.026.

Dunque la differenza osservata fra le due medie risulta abbastanza significativa, i dati supportano l'ipotesi che la dieta A sia più efficace in termini di riduzione di peso.

Osservazione: Questa associazione "statistica" può indicare un nesso di causalità SE i due gruppi sono simili per composizione – rispetto a tutte le caratteristiche potenzialmente influenti sulla riduzione del peso, sesso, età, attività fisica etc – ovvero in assenza di fattori di confondimento, e se anche le metodiche della misurazione sono le stesse nel gruppo A e nel gruppo B (assenza di bias da osservazione o altra forma di distorsione).