

Università di Torino

QUADERNI DIDATTICI

del

Dipartimento di Matematica

V. DEMICHELIS, A. ZIGGIOTO

Esercizi di Biostatistica

Quaderno # 37 - Aprile 2006



Indice

Prefazione	5
1 Probabilità	7
1.1 Esercizi	7
1.2 Soluzioni	11
2 Distribuzione della media campionaria	23
2.1 Esercizi	23
2.2 Soluzioni	24
3 Intervalli di confidenza	29
3.1 Esercizi	29
3.2 Soluzioni	31
4 Test d'ipotesi	35
4.1 Esercizi	35
4.2 Soluzioni	37
5 Confronto fra 2 medie	45
5.1 Esercizi	45
5.2 Soluzioni	48
6 Analisi della varianza	59
6.1 Esercizi	59
6.2 Soluzioni	62
7 Inferenza sulle proporzioni	73
7.1 Esercizi	73
7.2 Soluzioni	74

8	Tabelle di contingenza	77
8.1	Esercizi	77
8.2	Soluzioni	79
9	Correlazione	91
9.1	Esercizi	91
9.2	Soluzioni	93
10	Regressione	101
10.1	Esercizi	101
10.2	Soluzioni	103
	Bibliografia	111
	Appendice 1: Tabelle delle distribuzioni	111

Prefazione

Queste note raccolgono le esercitazioni di un corso semestrale tenuto dagli autori negli ultimi cinque anni presso la Facoltà di Farmacia dell'Università degli Studi di Torino e rivolto agli studenti del secondo anno della laurea specialistica in Farmacia e del primo anno della laurea triennale in Informazione Scientifica sul Farmaco.

Questi esercizi fanno riferimento al quaderno didattico Lezioni di Biostatistica degli stessi autori. Pertanto si rimanda alla prefazione del quaderno citato per avere delucidazioni in merito ai contenuti specifici.

Gli esercizi si concludono con una appendice contenente le tabelle delle distribuzioni teoriche di probabilità.

Il contenuto dell'appendice è ripreso dal testo di M.Pagano e K.Gauvreau (Biostatistica); gli autori ringraziano la casa editrice Idelson-Gnocchi che ne ha permesso la pubblicazione.

Febbraio 2006

Gli autori

Vittoria Demichelis, Andrea Ziggoto

Capitolo 1

Probabilità

1.1 Esercizi

1. Sia A l'evento che un particolare individuo è esposto ad elevati livelli di monossido di carbonio e B l'evento che è esposto ad elevati livelli di biossido di azoto.
 - (a) Qual è l'evento $A \cap B$?
 - (b) Qual è l'evento $A \cup B$?
 - (c) Gli eventi A e B sono mutuamente esclusivi?
2. Si considerino le statistiche relative alla natalità della popolazione degli Stati Uniti nel 1987. In accordo con questi dati, sono di seguito riportate le probabilità dell'età al momento del parto nel 1987 di una donna selezionata casualmente.

Età	Probabilità
< 15	0.0027
15-19	0.1214
20-24	0.2824
25-29	0.3192
30-34	0.1997
35-39	0.0651
40-44	0.0091
45-49	0.0004
Totale	1.0000

- (a) Qual è la probabilità che una donna che ha partorito nel 1987 avesse un'età minore o uguale a 24 anni?

- (b) Qual è la probabilità che avesse un'età maggiore o uguale a 40 anni?
 - (c) Dato che la madre di un determinato bambino è al di sotto dei 30 anni, qual è la probabilità che non abbia ancora 20 anni?
3. Si consideri un gruppo di 7 soggetti selezionati dalla popolazione degli Stati Uniti di età compresa fra 65 e 74 anni. Il numero di soggetti diabetici in questo campione è una variabile casuale binomiale con parametri $n = 7$ e $p = 0.125$.
- (a) Se si volesse fare un elenco dei 7 soggetti selezionati, in quanti modi essi potrebbero essere ordinati?
 - (b) Senza considerare l'ordine, in quanto modi è possibile selezionare 4 soggetti da questo gruppo di 7?
 - (c) Qual è la probabilità che esattamente 2 soggetti del campione siano diabetici?
 - (d) Qual è la probabilità che 4 soggetti siano diabetici?
4. In accordo con la National Health Survey, il 9.8% della popolazione di soggetti degli Stati Uniti di età compresa fra 18 e 24 anni è mancina.
- (a) Supponete di selezionare 10 soggetti da questa popolazione. In quanti modi possono essere ordinati?
 - (b) Senza considerare l'ordine, in quanto modi è possibile selezionare 4 soggetti da questo gruppo di 10?
 - (c) Qual è la probabilità che esattamente 3 di questi 10 soggetti siano mancini?
 - (d) Qual è la probabilità che almeno 6 dei 10 soggetti siano mancini?
 - (e) Qual è la probabilità che al massimo 2 soggetti siano mancini?
5. Si è trovato che il 25% dei bambini esposti ad un particolare agente infettivo contraggono quella particolare malattia. In un gruppo di 4 bambini che solitamente giocano insieme, con uguale esposizione all'agente infettivo
- (a) si dimostri che le probabilità che 0,1,2,3 e 4 bambini si ammalino sono rispettivamente $81/256$, $108/256$, $54/256$, $12/256$ e $1/256$;
 - (b) qual è la probabilità che almeno un bambino si ammali? Qual è la probabilità che almeno uno si ammali e uno sia sano nel gruppo dei 4 bambini?

6. Un produttore farmaceutico asserisce che un particolare farmaco dà miglioramento dei sintomi di *angina pectoris* nell'80% dei pazienti. Un medico prescrive questo farmaco a 5 dei suoi pazienti affetti da angina e trova che soltanto 2 mostrano un miglioramento dei sintomi. Assumendo che l'affermazione del produttore sia vera qual è la chance che si ottengano risultati così poco soddisfacenti o ancora peggiori di quelli osservati?
7. In accordo con il Behavioral Risk Factor Surveillance System, il 58% degli Americani ha uno stile di vita sedentario. Supponete di selezionare un campione di 12 individui e di trovare che 10 di essi non svolgono una regolare attività fisica. Assumendo che il Surveillance System sia corretto, qual è la probabilità di ottenere risultati uguali o peggiori di quelli che avete osservato?
8. In accordo con il Massachusetts Department of Health, 224 donne che hanno partorito nello Stato del Massachusetts nel 1998 erano positive al test per gli anticorpi anti virus HIV. Assumete che, con il tempo, il 25% dei bambini nati da queste madri risultino positivi al virus HIV.
 - (a) Se campioni di dimensione uguale a 224 fossero ripetutamente selezionati dalla popolazione di bambini nati da madri anti HIV positive, quale sarebbe il numero medio di bambini infetti per campione?
 - (b) Quale sarebbe la deviazione standard?
9. Il numero di casi di tetano registrati negli Stati Uniti durante un singolo mese nel 1989 ha una distribuzione di Poisson con parametro $\lambda = 4.5$.
 - (a) Qual è la probabilità che esattamente un caso di tetano sia registrato durante un determinato mese?
 - (b) Qual è la probabilità che siano registrati al massimo 2 casi di tetano?
 - (c) Qual è la probabilità che siano registrati 5 o più casi?
10. In una particolare contea, il numero medio di suicidi registrati ogni mese è 2.75. Assumete che il numero di suicidi segua una distribuzione di Poisson.
 - (a) Qual è la probabilità che non si registri alcun suicidio durante un determinato mese?

- (b) Qual è la probabilità che siano registrati al massimo 4 suicidi?
 - (c) Qual è la probabilità che siano registrati 6 o più suicidi?
11. Tra le donne degli Stati Uniti di età compresa fra 18 e 74 anni, la pressione diastolica è normalmente distribuita con media $\mu = 77$ mmHg e deviazione standard $\sigma = 11.6$ mmHg.
- (a) Qual è la probabilità che una donna selezionata casualmente abbia una pressione diastolica inferiore a 60 mmHg?
 - (b) Qual è la probabilità che questa donna abbia una pressione diastolica superiore a 90 mmHg?
 - (c) Qual è la probabilità che la donna abbia una pressione diastolica compresa fra 60 e 90 mmHg?
12. La distribuzione di peso della popolazione maschile degli Stati Uniti è approssimativamente normale con media $\mu = 172.2$ libbre e deviazione standard $\sigma = 29.8$ libbre.
- (a) Qual è la probabilità che un soggetto selezionato casualmente pesi meno di 130 libbre?
 - (b) Qual è la probabilità che questo soggetto pesi più di 210 libbre?
 - (c) Qual è la probabilità che tra 5 soggetti maschi selezionati casualmente dalla popolazione, almeno uno abbia un peso non compreso nell'intervallo 130-210 libbre?
13. Negli esami a livello nazionale, i punteggi sono dati in funzione di una scala tale per cui tra tutti gli studenti della Facoltà di Medicina che prendono parte a questo esame, i punteggi siano distribuiti in modo approssimativamente normale con media 500 e deviazione standard 100.
- (a) Quale percentuale degli studenti avrà punteggio superiore a 700?
 - (b) Quale punteggio è superato dal 10% degli studenti?
 - (c) Quale percentuale di studenti avrà un punteggio compreso fra 400 e 700?
 - (d) Precedentemente i punteggi avevano una scala tale per cui essi erano distribuiti in maniera normale con media 80 e deviazione standard 5. Qual è il punteggio migliore, 670 nel presente sistema o 89 nel precedente?

14. In un ampio gruppo di pazienti coronarici si trovò che i loro livelli di colesterolo serico presentavano approssimativamente una distribuzione normale. Si trovò inoltre che il 10% del gruppo aveva livelli di colesterolo al di sotto di 182.3 mg per 100 ml, mentre il 5% aveva valori superiori a 359.0 mg per 100 ml. Quali sono la media e la deviazione standard della distribuzione?
15. La quantità di sostanza anestetica rettale generale richiesta per determinare l'anestesia chirurgica fu trovata essere distribuita da paziente a paziente in modo approssimativamente normale con media di 50 mg ed una deviazione standard di 10 mg. La dose letale media era 110 mg con una deviazione standard di 20 mg ed anche questa media era distribuita in maniera approssimativamente normale. Se si usasse una dose che porta il 90% dei pazienti all'anestesia chirurgica, quale percentuale di pazienti sarebbe uccisa da questa dose?
16. Presupposto che i livelli di acido urico serico in maschi normali siano distribuiti in maniera approssimativamente normale con media di 5.4 mg per 100 ml e deviazione standard di 1 mg per 100 ml:
 - (a) qual è la probabilità che un maschio sano selezionato a caso abbia un livello di acido urico serico al di fuori dell'intervallo da 4.0 a 7.0 mg per 100 ml?
 - (b) qual è la chance che tra 4 maschi sani selezionati a caso ve ne sia almeno uno il cui livello di acido urico serico sia al di fuori dell'intervallo da 4.0 a 7.0 mg per 100 ml?
 - (c) quanti maschi sani devono essere selezionati in modo tale che vi sia una chance del 95% o ancora maggiore che almeno uno abbia i livelli di acido urico serico al di fuori dell'intervallo compreso fra 4.0 e 7.0 per 100 ml?

1.2 Soluzioni

1. (a) $A \cap B = \{\text{individuo esposto ad elevati livelli di monossido di carbonio e di biossido di azoto}\}$
- (b) $A \cup B = \{\text{individuo esposto ed elevati livelli di monossido di carbonio o di biossido di carbonio o di entrambi}\}$
- (c) ${}^cA = \{\text{individuo immune da elevati livelli di monossido di carbonio}\}$

(d) A e B *non* sono mutuamente esclusivi ma son indipendenti.

2. (a) A = avere un'età inferiore a 24 anni

$$P(A) = P(< 15) + P(15-19) + P(20-24) = 0.0027 + 0.1214 + 0.2824 = 0.4065$$

(sono mutuamente esclusivi)

(b) A = avere un'età superiore a 40 anni

$$P(A) = P(40-44) + P(45-49) = 0.0091 + 0.0004 = 0.0095.$$

(c) A = avere un'età inferiore a 20 anni, B = essere al di sotto dei 30 anni

$$\begin{aligned} P(B) &= P(< 15) + P(15-19) + P(20-24) + P(25-29) = \\ &= 0.0027 + 0.1214 + 0.2824 + 0.3192 = 0.7257 \end{aligned}$$

$$\begin{aligned} P(A \cap B) &= P(\text{avere meno di 20 anni}) = \\ &= P(< 15) + P(15-19) = 0.0027 + 0.1214 = 0.1241 \end{aligned}$$

Quindi

$$P(A|B) = \frac{0.1241}{0.7257} = 0.1710$$

3. (a) Sono le *permutazioni* di 7 elementi e quindi $7! = 5040$.

(b) Sono le *disposizioni* di 7 elementi presi a 4 a 4 e quindi

$$\binom{7}{4} = \frac{7!}{4!(7-4)!} = \frac{7 \cdot 6 \cdot 5 \cdot 4!}{4!3!} = 35.$$

(c) Qui la variabile casuale X conta il numero di soggetti diabetici. E' una variabile casuale bernoulliana di parametri $n = 5$ e $p = 0.125$.
Quindi

$$P(X = 2) = \binom{7}{2} 0.125^2 (1 - 0.125)^5 = 0.168 \approx 17\%.$$

$$(d) P(X = 4) = \binom{7}{4} 0.125^4 0.875^3 = 0.005 = 0.5\%.$$

4. (a) Sono esattamente $10! = 3628800$.

$$(b) \text{ Sono esattamente } \binom{10}{4} = \frac{10!}{4!(10-4)!} = 210.$$

- (c) Qui abbiamo una variabile casuale bernoulliana X che conta il numero di mancini. I parametri sono $n = 10$ e $p = 9.8\% = 0.098$. Allora

$$P(X = 3) = \binom{10}{3} 0.098^3 (1-0.098)^{10-3} = \frac{10!}{3!7!} 0.098^3 0.902^7 = 5.4\%.$$

- (d) Qui abbiamo

$$\begin{aligned} P(X \geq 6) &= 1 - P(X < 6) = 1 - P(X = 0) - P(X = 1) - P(X = 2) - \\ &\quad - P(X = 3) - P(X = 4) - P(X = 5) = \\ &= 1 - \binom{10}{0} 0.098^0 0.902^{10} - \binom{10}{1} 0.098 0.902^9 - \\ &\quad - \binom{10}{2} 0.098^2 0.902^8 - \binom{10}{3} 0.098^3 0.902^7 - \\ &\quad - \binom{10}{4} 0.098^4 0.902^6 - \binom{10}{5} 0.098^5 0.902^5 = \\ &= 0.3\%. \end{aligned}$$

- (e) Qui abbiamo

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) = \\ &= \binom{10}{0} 0.098^0 0.902^{10} + \binom{10}{1} 0.098 0.902^9 + \binom{10}{2} 0.098^2 0.902^8 = \\ &= 93\%. \end{aligned}$$

5. (a) Abbiamo qui la variabile casuale bernoulliana X che conta il numero di bambini che si ammalano, con parametri $n = 4$ e $p = 25\% = 0.25$. Quindi

$$\begin{aligned} P(X = 0) &= \binom{4}{0} 0.25^0 (1 - 0.25)^4 = \frac{81}{256}, \\ P(X = 1) &= \binom{4}{1} 0.25 0.75^3 = \frac{108}{256}, \\ P(X = 2) &= \binom{4}{2} 0.25^2 0.75^2 = \frac{54}{256}, \\ P(X = 3) &= \binom{4}{3} 0.25^3 0.75 = \frac{12}{256}, \\ P(X = 4) &= \binom{4}{4} 0.25^4 0.75^0 = \frac{1}{256}. \end{aligned}$$

- (b) Avremo che la probabilità che almeno un bambino si ammali è

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{81}{256} \approx 0.68 = 68\%.$$

Invece la probabilità che almeno un bambino si ammali e uno sia sano è

$$P(X \geq 1) - P(X = 4) = 1 - \frac{81}{256} - \frac{1}{256} = \frac{174}{256} \approx 0.67\%.$$

6. La variabile casuale X che conta il numero di pazienti che mostrano un miglioramento dei sintomi è una variabile bernoulliana di parametri $n = 5$ e $p = 0.8$. Allora

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) = \\ &= \binom{5}{0} 0.8^0 (1 - 0.8)^5 + \binom{5}{1} 0.8^1 0.2^4 + \binom{5}{2} 0.8^2 0.2^3 = \\ &\approx 5.7\%. \end{aligned}$$

7. In questo esercizio abbiamo la variabile casuale bernoulliana X che conta il numero di sedentari, con parametri $n = 12$ e $p = 58\%$. Quindi la risposta alla domanda del problema consiste nel calcolare

$$\begin{aligned} P(X \geq 10) &= P(X = 10) + P(X = 11) + P(X = 12) = \\ &= \binom{12}{10} 0.58^{10} (1 - 0.58)^2 + \binom{12}{11} 0.58^{11} 0.42 + \binom{12}{12} 0.58^{12} = \\ &= 6.4\%. \end{aligned}$$

8. La variabile casuale bernoulliana X conta il numero di bambini infetti e ha parametri $n = 224$ e $p = 0.25$.

(a) Il numero medio è $np = 0.25 \cdot 224 = 56$.

(b) La deviazione standard è $\sigma = \sqrt{np(1-p)} = \sqrt{0.25 \cdot (1 - 0.25) \cdot 224} = 6.48$.

9. La variabile casuale X che conta il numero di casi di tetano è una variabile di Poisson con parametro $\lambda = 4.5$.

(a) Si ha che

$$P(X = 1) = \frac{e^{-4.5} \cdot 4.5^1}{1!} \approx 5\%.$$

(b) Qui calcoliamo

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) = \\ &= \frac{e^{-4.5} \cdot 4.5^0}{0!} + \frac{e^{-4.5} \cdot 4.5}{1!} + \frac{e^{-4.5} \cdot 4.5^2}{2!} \\ &\approx 17\%. \end{aligned}$$

(c) Infine

$$\begin{aligned} P(X \geq 5) &= 1 - P(X < 5) = 1 - P(X = 0) - P(X = 1) - \\ &\quad - P(X = 2) - P(X = 3) - P(X = 4) = \\ &= 1 - 0.17 - \frac{e^{-4.5} \cdot 4.5^3}{3!} - \frac{e^{-4.5} \cdot 4.5^4}{4!} \\ &\approx 48\%. \end{aligned}$$

10. La variabile di Poisson X conta il numero medio di suicidi e ha parametro $\lambda = 2.75$.

(a) Si ha

$$P(X = 0) = \frac{e^{-2.75} \cdot 2.75^0}{0!} \approx 6.4\%.$$

(b) Si ha

$$\begin{aligned} P(X \leq 4) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \\ &= 0.064 + e^{-2.75} \cdot 2.75 + \frac{e^{-2.75} \cdot 2.75^2}{2} + \\ &\quad + \frac{e^{-2.75} \cdot 2.75^3}{6} + \frac{e^{-2.75} \cdot 2.75^4}{24} = \\ &\approx 85.5\%. \end{aligned}$$

(c) Si ha

$$\begin{aligned} P(X \geq 6) &= 1 - P(X < 6) = 1 - P(X = 0) - P(X = 1) - \\ &\quad - P(X = 2) - P(X = 3) - P(X = 4) - P(X = 5) = \\ &= 1 - 0.855 - \frac{e^{-2.75} \cdot 2.75^5}{5!} = \\ &= 1 - 0.855 - 0.083 \approx 6.2\%. \end{aligned}$$

11. La variabile casuale X che misura la pressione diastolica è normalmente distribuita con media $\mu = 77$ mmHg e deviazione standard $\sigma = 11.6$ mmHg. Quindi la variabile normale standardizzata Z è data da

$$Z = \frac{X - 77}{11.6}.$$

Utilizziamo la Tabella A.3 di pag. 397 del testo Pagano-Gauvreau.

(a) Abbiamo

$$P(X < 60) = P\left(Z < \frac{60 - 77}{11.6}\right) = P(Z < -1.46) = P(Z > 1.46) = 7.2\%.$$

(b) Abbiamo

$$P(X > 90) = P\left(Z > \frac{90 - 77}{11.6}\right) = P(Z > 1.12) = 13.1\%.$$

(c) Abbiamo

$$\begin{aligned} P(60 < X < 90) &= P(-1.46 < Z < 1.12) = \\ &= 1 - P(Z < -1.46) - P(Z > 1.12) = \\ &= 1 - 0.072 - 0.131 = 79.7\%. \end{aligned}$$

12. La variabile casuale X che rappresenta la distribuzione di peso è una variabile Gaussiana di media $\mu = 172.2$ libbre e deviazione standard $\sigma = 29.8$ libbre. La variabile standardizzata è dunque

$$Z = \frac{X - 172.2}{29.8}.$$

(a) Abbiamo

$$\begin{aligned} P(X < 130) &= P\left(Z < \frac{130 - 172.2}{29.8}\right) = \\ &= P(Z < -1.41) = P(Z > 1.41) = 7.9\%. \end{aligned}$$

(b) Abbiamo

$$P(X > 210) = P\left(Z > \frac{210 - 172.2}{29.8}\right) = P(Z > 1.26) = 0.104 = 10.4\%.$$

(c) Abbiamo che

$$\begin{aligned} P(130 < X < 210) &= P(-1.41 < Z < 1.26) = \\ &= 1 - 0.079 - 0.104 = 81.7\%. \end{aligned}$$

Introduciamo ora una nuova variabile casuale Y che conta il numero di soggetti con peso *non* compreso fra 130 e 210 libbre. Essa è una variabile casuale bernoulliana di parametri $n = 5$ e $p = 1 - 0.817 = 0.183$. Quindi

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - \binom{5}{0} 0.183^0 0.817^5 \approx 93\%.$$

13. Sia X la variabile casuale normale che rappresenta la distribuzione dei voti, con media $\mu = 500$ e deviazione standard $\sigma = 100$. Allora la normale standardizzata è

$$Z = \frac{X - 500}{100}.$$

(a) Si ha

$$P(X \geq 700) = P\left(Z \geq \frac{700 - 500}{100}\right) = P(Z \geq 2) = 2.3\%.$$

(b) Si ha

$$P(Z \geq z) = 0.1$$

quando $z = 1.28$. Quindi risolvendo l'equazione

$$1.28 = \frac{x - 500}{100}$$

otteniamo

$$x = 1.28 \cdot 100 + 500 = 628,$$

che è il punteggio superato dal 10% degli studenti.

(c) Abbiamo

$$\begin{aligned} P(400 \leq X \leq 700) &= P\left(\frac{400 - 500}{100} \leq Z \leq \frac{700 - 500}{100}\right) = \\ &= P(-1 \leq Z \leq 2) = 1 - P(Z \geq 2) - P(Z \leq -1) = \\ &= 1 - P(Z \geq 2) - P(Z \geq 1) = \\ &= 1 - 0.023 - 0.159 = 81.8\%. \end{aligned}$$

- (d) Standardizziamo i due punteggi 670 e 89 in modo da poterli confrontare:

$$z_{\text{prima}} = \frac{89 - 80}{5} = 1.8,$$

$$z_{\text{dopo}} = \frac{670 - 500}{100} = 1.7.$$

Concludiamo quindi che il punteggio migliore è 89 nella precedente scala perchè lascia una *coda* di dimensioni minori a destra.

14. Sia X la variabile casuale normale che rappresenta i livelli di colesterolo serico. Noi sappiamo che

$$P(X \leq 182.3) = 10\% = 0.1.$$

Consultando la Tabella A.3 della normale standardizzata vediamo che una coda di sinistra con area uguale al 10% corrisponde ad un valore $z = -1.28$. Dunque, dette μ e σ la media e la deviazione standard rispettivamente della variabile X , avremo la seguente relazione, ottenuta standardizzando il valore 182.3:

$$\frac{182.3 - \mu}{\sigma} = -1.28,$$

da cui

$$182.3 - \mu = -1.28\sigma.$$

D'altronde, sappiamo anche che

$$P(X \geq 359) = 5\% = 0.05.$$

Consultando la Tabella A.3 della normale standardizzata vediamo che una coda di destra con area uguale al 5% corrisponde ad un valore $z = 1.64$. Dunque, standardizzando il valore 359 avremo:

$$\frac{359 - \mu}{\sigma} = 1.64$$

da cui

$$359 - \mu = 1.64\sigma.$$

Otteniamo allora il seguente sistema di due equazioni in due incognite μ e σ :

$$\begin{cases} 359 - \mu = 1.64\sigma \\ 182.3 - \mu = -1.28\sigma \end{cases}$$

Sottraendo membro a membro le due equazioni otteniamo

$$176.7 = 2.92\sigma$$

da cui

$$\sigma = 60.5 \text{ mg per } 100 \text{ ml.}$$

Sostituendo questo valore di σ , ad esempio, nella prima delle due equazioni del sistema otteniamo

$$\mu = 359 - 1.64 \cdot 60.5 = 259.78 \text{ mg per } 100 \text{ ml.}$$

15. Sia X la variabile casuale normale che rappresenta la quantità di anestetizzante in questione. Sia

$$Z = \frac{X - 50}{10}$$

la variabile casuale standardizzata di X . Dire che usiamo una dose che porta il 90% dei pazienti all'anestesia chirurgica significa dire che

$$P(Z \leq z) = 90\% = 0.9$$

per un certo valore di z . Allora

$$1 - P(Z > z) = 0.9$$

da cui

$$P(Z > z) = 0.1$$

e, consultando la Tabella A.3, otteniamo che $z = 1.28$. Questo valore, in termini della variabile casuale X da cui siamo partiti, corrisponde a

$$\frac{x - 50}{10} = 1.28 \Leftrightarrow x = 1.28 \cdot 10 + 50 = 62.8 \text{ mg.}$$

Sia ora Y la variabile casuale, anch'essa normale, che rappresenta la *dose letale* media di anestetizzante. Sappiamo che la media μ_Y e la deviazione standard σ_Y di Y sono rispettivamente 110 mg e 20 mg. Sapere quale percentuale di pazienti sarebbe uccisa dalla dose che porta il 90% di pazienti all'anestesia chirurgica (che abbiamo trovato essere uguale a 62.8 mg), equivale a richiedere di calcolare, standardizzando la variabile casuale Y ,

$$P\left(\frac{Y - 110}{20} \leq y\right)$$

dove

$$y = \frac{62.8 - 110}{20} = -2.36$$

è il valore che porta il 90% dei pazienti all'anestesia chirurgica, standardizzato secondo la nuova variabile aleatoria Y . Consultando la Tabella A.3 della normale standardizzata troviamo che

$$P\left(\frac{Y - 110}{20} \leq -2.36\right) = 0.009 = 0.9\%,$$

che è dunque il risultato richiesto.

16. Sia X la variabile casuale normale che rappresenta la distribuzione di acido urico serico nei maschi. Sia Z la variabile normale standardizzata. Sappiamo che la media è $\mu = 5.4$ mg per 100 ml e che la deviazione standard è $\sigma = 1$ mg per 100 ml.

(a) Abbiamo

$$\begin{aligned} P(X \leq 4) + P(X \geq 7) &= P\left(Z \leq \frac{4 - 5.4}{1}\right) + P\left(Z \geq \frac{7 - 5.4}{1}\right) = \\ &= P(Z \leq -1.4) + P(Z \geq 1.6) = \\ &= 0.081 + 0.055 = 0.136 = 13.6\%. \end{aligned}$$

- (b) Sia Y la variabile casuale che conta il numero di maschi aventi un livello di acido urico serico al di fuori dell'intervallo da 4 a 7 mg per 100 ml. Essa è una variabile casuale bernoulliana di parametri $n = 4$ e $p = 13.6\%$. Allora

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y < 1) = \\ &= 1 - P(Y = 0) = 1 - \binom{4}{0} 0.136^0 (1 - 0.136)^4 = \\ &= 1 - 0.864^4 = 0.443 = 44.3\%. \end{aligned}$$

- (c) Sia Y la variabile casuale come al punto (b). I parametri questa volta sono $p = 0.136$ e n sconosciuto. Noi vogliamo che

$$P(Y \geq 1) \geq 95\%.$$

Allora

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) = \\ &= 1 - \binom{n}{0} 0.136^0 (1 - 0.136)^n = \\ &= 1 - 0.864^n \geq 0.95, \end{aligned}$$

da cui

$$0.864^n \leq 0.05.$$

Questa disuguaglianza è verificata per

$$n \geq \frac{\log 0.05}{\log 0.864} = 20.5.$$

Essendo n un numero *intero*, avremo dunque che dovremo selezionare almeno 21 maschi per poter avere una chance maggiore del 95% che almeno uno di essi abbia un livello di acido urico serico al di fuori dell'intervallo compreso fra 4 e 7 mg per 100 ml.

Capitolo 2

Distribuzione della media campionaria

2.1 Esercizi

1. Si consideri la distribuzione normale standardizzata.
 - (a) Calcolare un intervallo che comprenda il 95% delle medie di campioni di dimensione uguale a 10.
 - (b) Quale proporzione delle medie giace al di sopra di $z = 1.60$?
 - (c) Quale proporzione delle medie giace al di sotto di $z = -0.85$?
 - (d) Quale valore di z delimita il 20% inferiore della distribuzione delle medie?
2. La distribuzione dei livelli di albumina nel liquido cerebrospinale degli adulti degli Stati Uniti ha una media $\mu = 29.5$ mg/100 ml ed una deviazione standard $\sigma = 9.25$ mg/100 ml.
 - (a) Quale proporzione dei campioni di dimensione uguale a 20 ha medie maggiori di 35 mg/100 ml?
 - (b) Quale proporzione ha medie minori di 22 mg/100 ml?
 - (c) Quale valore di $\bar{x}$ delimita il 5% superiore della distribuzione delle medie di campioni di dimensione uguale a 20?
3. Per la popolazione di donne di età compresa fra 3 e 74 anni che hanno partecipato alla National Health Interview Survey, la distribuzione dei livelli di emoglobina ha una media $\mu = 13.3$ g/100 ml ed una deviazione standard $\sigma = 1.12$ g/100 ml.

- (a) Se si selezionano da questa popolazione campioni ripetuti di dimensione uguale a 15, quale proporzione dei campioni avrà un livello medio di emoglobina compresa fra 13.0 e 13.6 g/100 ml?
 - (b) Se i campioni sono di dimensione uguale a 30, quale proporzione avrà una media compresa tra 13.0 e 13.6 g/100 ml?
 - (c) Quale deve essere la dimensione dei campioni affinché il 95% delle loro medie siano comprese entro ± 0.1 g/100 ml dalla media della popolazione?
4. Nei Paesi Bassi, la popolazione maschile sana di età compresa fra 65 e 79 anni ha una distribuzione dei livelli di acido urico serico approssimativamente normale con media $\mu = 341 \mu\text{mol/l}$ e deviazione standard $\sigma = 79 \mu\text{mol/l}$.
- (a) Quale proporzione di soggetti ha un livello di acido urico serico compreso fra 300 e 400 $\mu\text{mol/l}$?
 - (b) Quale proporzione dei campioni di dimensione uguale a 5 ha un livello medio di acido urico serico compreso fra 300 e 400 $\mu\text{mol/l}$?
5. In un esperimento a ciascun topo di un campione casuale di 25 unità deve essere iniettato un farmaco ad un livello di dose di 0.004 mg per grammo di peso corporeo. Per questo ceppo di topi è noto che il peso è distribuito in modo approssimativamente *normale* con una media di 19 g ed una deviazione standard di 4 g.
- (a) Se il ricercatore possiede un totale di 2 mg del farmaco, qual è la probabilità che questo non sia sufficiente per trattare tutti i topi?
 - (b) Quanto farmaco dovrebbe possedere al fine di correre un rischio dell' 1% di non trattare tutti gli animali?

2.2 Soluzioni

1. (a) Qui abbiamo $\mu = 0$, $\sigma = 1$ e $n = 10$. Allora la variabile standardizzata Z della media campionaria $\bar{X}$ è

$$Z = \frac{\bar{X} - 0}{\frac{1}{\sqrt{10}}} = \sqrt{10} \bar{X}.$$

Dalla Tabella A.3, sappiamo che

$$P(-1.96 \leq Z \leq 1.96) = 95\% = 0.95.$$

Quindi deve essere

$$-1.96 \leq z \leq 1.96.$$

Ma $z = \sqrt{10} \bar{x}$. Pertanto

$$\frac{-1.96}{\sqrt{10}} \leq \bar{x} \leq \frac{1.96}{\sqrt{10}},$$

cioè

$$-0.62 \leq \bar{x} \leq 0.62.$$

Quindi il 95% delle media dei campioni di dimensione uguale a 10 è compreso nell'intervallo $(-0.62, 0.62)$.

(b) Dalla Tabella A.3 otteniamo

$$P(Z \geq 1.60) = 5.5\%.$$

(c) Dalla Tabella A.3 otteniamo

$$P(Z \leq -0.85) = 19.8\%.$$

(d) Vogliamo trovare z tale che

$$P(Z \leq z) = 20\% = 0.2.$$

Dalla Tabella A.3 otteniamo $z = -0.84$.

2. (a) In questo esercizio $\mu = 29.5$ mg/100 ml, $\sigma = 9.25$ mg/200 ml e $n = 20$. Vogliamo calcolare $P(\bar{X} \geq 35)$. Consideriamo la variabile standardizzata

$$Z = \frac{\bar{X} - 29.5}{\frac{9.25}{\sqrt{20}}}.$$

Allora

$$P(\bar{X} \geq 35) = P\left(Z \geq \frac{35 - 29.5}{\frac{9.25}{\sqrt{20}}}\right) = P(Z \geq 2.66) = 0.4\%.$$

(b) Vogliamo calcolare $P(\bar{X} \leq 22)$. Allora

$$P(\bar{X} \leq 22) = P\left(Z \leq \frac{22 - 29.5}{\frac{9.25}{\sqrt{20}}}\right) = P(Z \leq -3.63) = P(Z \geq 3.63) \approx 0\%.$$

(c) Vogliamo trovare per quale $\bar{x}$ si ha che

$$P(\bar{X} \geq \bar{x}) = 5\% = 0.05.$$

Dalla Tabella A.3 otteniamo che il 5% superiore è individuato dal valore $z = 1.65$. Allora, usando la standardizzazione:

$$\frac{\bar{x} - 29.5}{\frac{9.25}{\sqrt{20}}} = 1.65$$

da cui otteniamo

$$\bar{x} = 1.65 \cdot \frac{9.25}{\sqrt{20}} + 29.5 = 32.9 \text{ mg/100 ml.}$$

3. (a) Qui abbiamo $\mu = 13.3$ g/100 ml, $\sigma = 1.12$ g/100 ml e $n = 15$. Vogliamo calcolare

$$P(13 \leq \bar{X} \leq 13.6).$$

Passando alla variabile standardizzata Z avremo

$$\begin{aligned} P\left(\frac{13 - 13.3}{\frac{1.12}{\sqrt{15}}} \leq Z \leq \frac{13.6 - 13.3}{\frac{1.12}{\sqrt{15}}}\right) &= P(-1.04 \leq Z \leq 1.04) = 1 - 2 \cdot P(Z \geq 1.04) = \\ &= 1 - 2 \cdot 0.149 = 0.702 = 70.2\%. \end{aligned}$$

(b) Ora $n = 30$. Quindi

$$\begin{aligned} P(13 \leq \bar{X} \leq 13.6) &= P\left(\frac{13 - 13.3}{\frac{1.12}{\sqrt{30}}} \leq Z \leq \frac{13.6 - 13.3}{\frac{1.12}{\sqrt{30}}}\right) = P(-1.46 \leq Z \leq 1.46) = \\ &= 1 - 2 \cdot P(Z \geq 1.46) = 1 - 2 \cdot 0.072 = 85.6\%. \end{aligned}$$

(c) Vogliamo calcolare n tale che

$$P(13.3 - 0.1 \leq \bar{X} \leq 13.3 + 0.1) = 95\% = 0.95.$$

Allora, passando alla standardizzazione:

$$P\left(\frac{-0.1}{\frac{1.12}{\sqrt{n}}} \leq Z \leq \frac{0.1}{\frac{1.12}{\sqrt{n}}}\right) = 0.95.$$

Dalla Tabella A.3 ricaviamo che il 95% è individuato dal valore $z = 1.96$. Dunque

$$1.96 = \frac{0.1}{\frac{1.12}{\sqrt{n}}}$$

da cui

$$\sqrt{n} = \frac{1.96 \cdot 1.12}{0.1} \approx 22.$$

Pertanto n deve essere uguale *almeno* a 484.

4. (a) Qui abbiamo $\mu = 341 \text{ } \mu\text{mol/l}$ e $\sigma = 79 \text{ } \mu\text{mol/l}$. Quindi

$$\begin{aligned} P(300 \leq X \leq 400) &= P\left(\frac{300 - 341}{79} \leq Z \leq \frac{400 - 341}{79}\right) = P(-0.52 \leq Z \leq 0.75) = \\ &= 1 - P(Z \geq 0.52) - P(Z \geq 0.75) = \\ &= 1 - 0.302 - 0.227 = 0.471 = 47.1\%. \end{aligned}$$

- (b) Sapendo che i campioni hanno dimensione 5, avremo

$$\begin{aligned} P(300 \leq \bar{X} \leq 400) &= P\left(\frac{300 - 341}{\frac{79}{\sqrt{5}}} \leq Z \leq \frac{400 - 341}{\frac{79}{\sqrt{5}}}\right) = \\ &= P(-1.16 \leq Z \leq 1.67) = 1 - P(Z \geq 1.16) - P(Z \geq 1.67) = \\ &= 1 - 0.123 - 0.047 = \\ &= 0.83 = 83\%. \end{aligned}$$

5. (a) 2 mg di farmaco sono sufficienti per $\frac{2}{0.004} = 500$ g di peso corporeo. Con 25 topi il peso medio è dunque $\frac{500}{25} = 20$ g. Allora

$$z = \frac{20 - 19}{\frac{4}{\sqrt{25}}} = 1.25$$

Dalla Tabella della normale ricaviamo che

$$P(Z \geq 1.25) = 0.106 = 10.6\%$$

- (b) $P(Z \geq z) = 0.01$ corrisponde a $z = 2.33$. Quindi il peso medio per fare previsioni è

$$2.33 = \frac{\bar{x} - 19}{\frac{4}{\sqrt{25}}} \Rightarrow \bar{x} = \frac{4}{5} \cdot 2.33 + 19 = 20.864 \text{ g}$$

Questo corrisponde ad un peso totale di $20.864 \cdot 25 = 521.6$ g che dunque richiedono $521.6 \cdot 0.004 = 2.09$ mg di farmaco.

Capitolo 3

Intervalli di confidenza

3.1 Esercizi

1. La media μ_s della distribuzione della pressione sistolica delle donne diabetiche di età compresa fra i 30 e i 34 anni non è nota, mentre la deviazione standard è $\sigma_s = 11.8$ mmHg.
 - (a) Un campione casuale di 10 donne è selezionato dalla popolazione. La pressione sistolica media del campione è $\bar{x}_s = 130$ mmHg. Calcolare l'intervallo di confidenza bilaterale per μ_s al livello $\alpha = 95\%$.
 - (b) Calcolare l'intervallo di confidenza per μ_s al 90%.
 - (c) Quanto grande deve essere la dimensione affinché il campione abbia un'ampiezza dell'intervallo di confidenza al 95% pari a 5 mmHg?
2. La tabella seguente
Velocità di pulsazione (battiti/minuto)
59,72,58,65,77,83,72,77,62,62
riporta le frequenze di pulsazioni di un campione di 10 studenti maschi del primo anno della facoltà di medicina. Costruire l'intervallo di confidenza per la media della popolazione da cui il campione è estratto al livello $\alpha = 95\%$.
3. Per la popolazione delle donne in gravidanza i cui feti sono sottoposti a trattamento chirurgico, la distribuzione della durata delle degenza ospedaliera ha media e deviazione standard sconosciute. Un campione

casuale di 17 madri ha una degenza media di $\bar{x} = 9.5$ giorni e una deviazione standard $s = 3.6$ giorni.

- (a) Calcolare l'intervallo di confidenza al 95% per la media reale della popolazione.
 - (b) Qual è l'ampiezza di questo intervallo?
 - (c) Quanto deve essere grande un campione affinché l'intervallo di confidenza al 95% abbia un'ampiezza di 2 giorni? Assumere che il valore di s non cambi.
4. Le medie delle distribuzioni della pressione sistolica e diastolica delle donne diabetiche di età compresa fra 30 e 34 anni non sono note. Tuttavia, le loro deviazioni standard sono rispettivamente $\sigma_s = 11.8$ mmHg e $\sigma_d = 9.1$ mmHg.
 - (a) Un campione casuale di 10 donne è selezionato da questa popolazione. La pressione sistolica media del campione è $\bar{x}_s = 130$ mmHg. Calcolare l'intervallo di confidenza bilaterale al 95% per μ_s , la reale pressione sistolica media.
 - (b) La pressione diastolica media di un campione di dimensione uguale a 10 è $\bar{x}_d = 84$ mmHg. Calcolare l'intervallo di confidenza bilaterale al 90% per μ_d , la reale pressione diastolica media della popolazione.
 - (c) Calcolare l'intervallo di confidenza bilaterale al 99% per μ_d .
 - (d) Quanto differisce l'intervallo di confidenza al 99% dall'intervallo di confidenza al 90%?
5. E' stata determinata la percentuale di peso ideale per un campione casuale di 18 diabetici insulino-dipendenti. I risultati sono di seguito riportati:

107, 11, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 100, 125, 114, 95, 117(%).

 - (a) Calcolare l'intervallo di confidenza bilaterale al 95% per la percentuale media di peso ideale della popolazione.
 - (b) Questo intervallo di confidenza comprende il valore 100%? Che cosa indica?

3.2 Soluzioni

1. Sappiamo che μ_s non è nota e che $\sigma_s = 11.8$ mmHg.

- (a) Si ha che $n = 10$ e che $\bar{x}_s = 130$ mmHg. Poichè la deviazione standard è nota, utilizziamo la distribuzione normale. Consultando la Tabella A.3 si riscontra che il valore z standardizzato che lascia il 2.5% in ciascuna coda è $z = 1.96$. Dunque l'intervallo di confidenza richiesto per μ_s è

$$\left(130 - 1.96 \cdot \frac{11.8}{\sqrt{10}}, 130 + 1.96 \cdot \frac{11.8}{\sqrt{10}}\right) = (123, 137).$$

- (b) Consultando la Tabella A.3, si riscontra che il valore z standardizzato che lascia il 5% in ciascuna coda è $z = 1.645$. Dunque l'intervallo di confidenza richiesto per μ_s è

$$\left(130 - 1.645 \cdot \frac{11.8}{\sqrt{10}}, 130 + 1.645 \cdot \frac{11.8}{\sqrt{10}}\right) = (124, 136).$$

- (c) L'intervallo di confidenza al 95% è

$$\left(130 - 1.96 \cdot \frac{11.8}{\sqrt{n}}, 130 + 1.96 \cdot \frac{11.8}{\sqrt{n}}\right),$$

la cui ampiezza è chiaramente

$$130 + 1.96 \cdot \frac{11.8}{\sqrt{n}} - \left(130 - 1.96 \cdot \frac{11.8}{\sqrt{n}}\right) = 2 \cdot 1.96 \cdot \frac{11.8}{\sqrt{n}}.$$

Affinchè tale ampiezza sia uguale a 5 deve essere

$$2 \cdot 1.96 \cdot \frac{11.8}{\sqrt{n}} = 5,$$

da cui otteniamo

$$\sqrt{n} = \frac{2 \cdot 1.96 \cdot 11.8}{5} = 9.25$$

e quindi $n = 86$.

2. Calcoliamo innanzitutto $\bar{x}$ e s per il campione estratto:

$$\bar{x} = \frac{59 + 722 + 58 + 65 + 77 + 83 + 72 + 77 + 62 + 62}{10} = 68.7$$

$$s = \sqrt{\frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 8.66.$$

Poichè in questo caso la deviazione standard σ della popolazione non è nota, utilizziamo la distribuzione t di Student. Utilizzando la Tabella A.4 osserviamo che per $10 - 1 = 9$ gradi di libertà il 95% delle osservazioni è compreso fra -2.262 e 2.262. Quindi l'intervallo di confidenza richiesto è

$$\left(68.7 - 2.262 \cdot \frac{8.66}{\sqrt{10}}, 68.7 + 2.262 \cdot \frac{8.66}{\sqrt{10}} \right) = (62.5, 74.9).$$

Se avessimo voluto l'intervallo di confidenza al 90%, avremmo dovuto cercare sulla Tabella A.4 il valore di t per 9 gradi di libertà che lascia il 5% in una coda e avremmo trovato il valore 1.833. L'intervallo di confidenza al 90% sarebbe stato dunque

$$\left(68.7 - 1.833 \cdot \frac{8.66}{\sqrt{10}}, 68.7 + 1.833 \cdot \frac{8.66}{\sqrt{10}} \right) = (63.68, 73.72).$$

Ovviamente l'intervallo al 90% è più ristretto rispetto all'intervallo al 95%.

Supponiamo poi di voler sapere quanti individui deve contenere il campione affinché l'intervallo di confidenza al 90% sia ampio 5 pulsazioni al minuto.

L'intervallo in questione è

$$\left(68.7 - 1.833 \cdot \frac{8.66}{\sqrt{n}}, 68.7 + 1.833 \cdot \frac{8.66}{\sqrt{10}} \right),$$

la cui ampiezza è

$$2 \cdot 1.833 \cdot \frac{8.66}{\sqrt{10}}.$$

Affinchè quest'ultima sia uguale a 5 deve essere

$$2 \cdot 1.833 \cdot \frac{8.66}{\sqrt{10}} = 5$$

da cui

$$\sqrt{n} = \frac{2 \cdot 1.833 \cdot 8.66}{5} = 6.35.$$

Pertanto il campione deve contenere (almeno) 41 soggetti.

3. In questo esercizio abbiamo $n = 17$, $\bar{x} = 9.5$ giorni e $s = 3.6$ giorni e la media μ e deviazione standard σ della popolazione sono sconosciute.

- (a) Utilizziamo dunque la distribuzione t di Student. Consultando la Tabella A.4, notiamo che in corrispondenza di 16 gradi di libertà il valore che lascia in una coda il 2.5% è $t = 2.120$. Dunque l'intervallo di confidenza richiesto è

$$\left(9.5 - 2.12 \cdot \frac{3.6}{\sqrt{17}}, 9.5 + 2.12 \cdot \frac{3.6}{\sqrt{17}} \right) = (7.6, 11.3).$$

- (b) L'ampiezza di tale intervallo è $11.3 - 7.6 = 3.7$.
 (c) L'intervallo di confidenza al 95% è

$$\left(9.5 - 2.12 \cdot \frac{3.6}{\sqrt{n}}, 9.5 + 2.12 \cdot \frac{3.6}{\sqrt{n}} \right).$$

Affinchè esso sia ampio 2 giorni deve essere

$$2 \cdot 2.12 \cdot \frac{3.6}{\sqrt{n}} = 2$$

da cui

$$\sqrt{n} = 7.632.$$

Dunque il campione deve contenere (almeno) 58 individui.

4. In questo esercizio conosciamo le deviazioni standard delle popolazioni in questione. Dunque utilizziamo il test z .

- (a) Consultando la Tabella A.3 otteniamo che il valore che lascia il 2.5 in ogni coda è $z = 1.96$. Dunque l'intervallo richiesto è

$$\left(130 - 1.96 \cdot \frac{11.8}{\sqrt{10}}, 130 + 1.96 \cdot \frac{11.8}{\sqrt{10}} \right) = (123, 137).$$

- (b) Consultando la Tabella A.3, osserviamo che il valore z che lascia il 5% in ogni coda è $z = 1.645$. Dunque l'intervallo richiesto è

$$\left(84 - 1.645 \cdot \frac{9.1}{\sqrt{10}}, 84 + 1.645 \cdot \frac{9.1}{\sqrt{10}} \right) = (79, 89).$$

- (c) Utilizzando la Tabella A.3, otteniamo che il valore z che lascia il 0.5% in ogni coda è $z = 2.57$. Dunque l'intervallo richiesto è

$$\left(84 - 2.57 \cdot \frac{9.1}{\sqrt{10}}, 84 + 2.57 \cdot \frac{9.1}{\sqrt{10}} \right) = (77, 91).$$

- (d) L'intervallo di confidenza al 99% è ampio $91 - 77 = 14$ mmHg, quello al 90% è ampio $79 - 89 = 10$ mmHg. Dunque vi è una differenza uguale a 4 mmHg.
5. Poichè non conosciamo la deviazione standard della popolazione in questione, utilizzeremo la distribuzione t di Student.
- (a) Osserviamo dalla Tabella A.4 che il valore t che in corrispondenza di $18 - 1 = 17$ gradi di libertà lascia il 2.5% in ogni coda è $t = 2.11$. L'intervallo richiesto pertanto è

$$\left(112.7 - 2.11 \cdot \frac{14.4}{\sqrt{18}}, 112.7 + 2.11 \cdot \frac{14.4}{\sqrt{18}} \right) = (105.5, 119.9).$$

- (b) Come osserviamo, l'intervallo di confidenza richiesto NON contiene il 100%. Questo significa che siamo confidenti al 95% che il peso *ideale* deve essere superiore al loro peso *reale*.

Capitolo 4

Test d'ipotesi

4.1 Esercizi

1. La distribuzione delle pressioni diastoliche delle donne della popolazione di donne diabetiche di età compresa fra i 30 e i 34 anni ha una media μ_D non nota, mentre la deviazione standard è $\sigma_D = 9.1$ mmHg. Può essere utile ai medici sapere se la media di questa popolazione è uguale alla pressione diastolica media di 74.4 mmHg della popolazione generale delle donne di questa fascia di età.
 - (a) Qual è l'ipotesi nulla del test?
 - (b) Qual è l'ipotesi alternativa?
 - (c) Si seleziona un campione casuale di 10 donne diabetiche, la loro pressione diastolica media è $\bar{x} = 84$ mmHg. Utilizzando questa informazione, eseguire un test bilaterale ad un livello di significatività $\alpha = 0.05$.
 - (d) Quale conclusione si può trarre dai risultati del test?
 - (e) La conclusione sarebbe stata diversa con $\alpha = 0.01$ invece di 0.05?

2. La tabella seguente

Velocità di pulsazione (battiti/minuto)

59,72,58,65,77,83,72,77,62,62

riporta le frequenze di pulsazioni di un campione di 10 studenti maschi del primo anno della facoltà di medicina. Un manuale di valori clinici normali indica che la frequenza media di pulsazioni per i maschi giovani è di 72 battiti al minuto. Questi dati sui 10 studenti sono

compatibili con la norma riportata sul manuale? La media campionaria che si ottiene dai dati della tabella è verosimilmente derivata da una popolazione la cui media è di 72 battiti al minuto?

3. Una casa farmaceutica sta testando un farmaco per aumentare la frequenza del battito cardiaco. In un campione di 100 individui l'effetto è stato un incremento medio di 1.4 battiti al minuto, con una deviazione standard di 3.6. Dal momento che la ditta vuole evitare di mettere in commercio un farmaco non efficace, essa propone un livello di significatività di 0.005. Potrebbe mettere in commercio il farmaco? (Suggerimento: se il farmaco non funziona, l'aumento della frequenza cardiaca sarà 0).
4. L'infezione da *E.canis* è una malattia parassitaria dei cani che talvolta viene contratta dagli uomini. Tra gli uomini infetti, la distribuzione dei valori dei globuli bianchi ha una media μ ed una deviazione standard σ non note. Nella popolazione generale, i globuli bianchi sono in media 7250 per mm^3 . Si ritiene che i soggetti infetti abbiano, in media, un numero minore di globuli bianchi.
 - (a) Quali sono le ipotesi nulla ed alternativa per un test unilaterale?
 - (b) Per un campione casuale di 15 soggetti infetti, il numero medio di globuli bianchi è $\bar{x} = 4767$ per mm^3 e la deviazione standard è $s = 3204$ per mm^3 . Eseguire il test ad un livello di significatività $\alpha = 0.05$.
 - (c) Che cosa si può concludere?
5. L'indice di massa corporea è calcolato dividendo il peso di un soggetto per il quadrato della sua altezza. Esso è una misura del grado di sovrappeso di un soggetto. Per la popolazione di uomini di mezza età che svilupperanno il diabete mellito, la distribuzione degli indici di massa corporea basali ha una media μ ed una deviazione standard σ non note. Un campione casuale di 58 soggetti selezionati da questo gruppo ha una media $\bar{x} = 25 \text{ kg}/m^2$ ed una deviazione standard $s = 2.7 \text{ kg}/m^2$.
 - (a) Calcolare l'intervallo di confidenza al 95% per la media μ della popolazione.
 - (b) Ad un livello di significatività $\alpha = 0.05$ testare se l'indice medio di massa corporea basale della popolazione di soggetti di mezza età che svilupperanno il diabete mellito è uguale a $24 \text{ kg}/m^2$,

cioè il valore medio della popolazione che non sarà affetta da tale patologia.

- (c) Che cosa si può concludere?

4.2 Soluzioni

1. (a) L'ipotesi nulla del test è

$$H_0 : \mu = 74.4 \text{ mmHg.}$$

- (b) L'ipotesi alternativa del test è

$$H_A : \mu \neq 74.4 \text{ mmHg.}$$

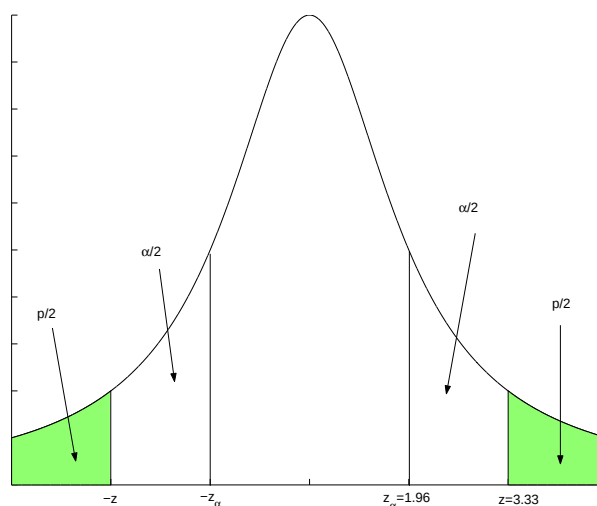
- (c) Poichè conosciamo la deviazione standard $\sigma_d = 9.1$ mmHg, utilizziamo il test z. Poichè nel nostro esempio $n = 10$ e $\bar{x} = 84$ mmHg, avremo che il valore di z è

$$z = \frac{\bar{x} - \mu}{\frac{\sigma_d}{\sqrt{n}}} = \frac{84 - 74.4}{\frac{9.1}{\sqrt{10}}} = 3.33.$$

Per trovare il valore z_α che corrisponde al livello di significatività $\alpha = 5\%$, dobbiamo guardare sulla Tabella A.3 il valore di z che lascia in *entrambe* le code il 2.5%, perchè stiamo eseguendo un test *bilaterale*. Troviamo dunque che

$$z_\alpha = 1.96.$$

Pertanto abbiamo questa situazione:

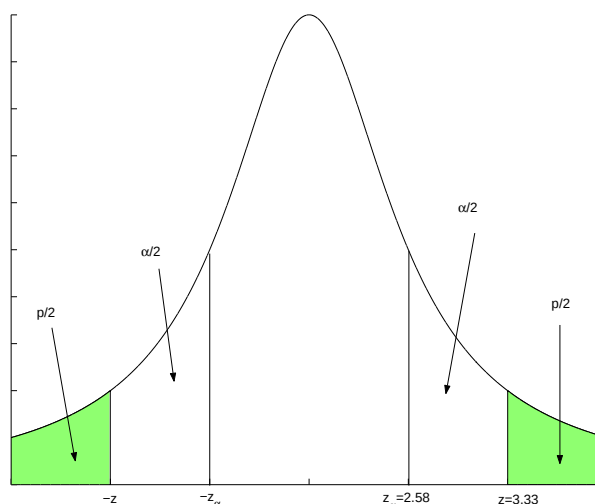


Dunque il valore p del test è *inferiore* ad α : $p < \alpha$.

- (d) Essendo $p < \alpha$, concludiamo che dobbiamo *rifiutare* l'ipotesi nulla H_0 e dire che la pressione diastolica media della popolazione delle donne diabetiche di età compresa fra 30 e 34 anni NON è uguale a 74.4 mmHg. Infatti, essa è significativamente più alta.
- (e) Se $\alpha = 0.01$, per cercare il valore z_α dobbiamo guardare sulla Tabella A.3 il valore di z che lascia il 0.5% in entrambe le code. Troviamo il valore

$$z_\alpha = 2.58.$$

Abbiamo allora la seguente situazione:



Anche in questo caso, dunque, si ha che il valore p del test è *inferiore* ad α e quindi anche in questo caso *rifiutiamo* l'ipotesi nulla H_0 .

2. Dobbiamo chiaramente eseguire un test d'ipotesi in cui l'ipotesi nulla è

$$H_0 : \mu = 72 \text{ battiti al minuto}$$

e l'ipotesi alternativa è

$$H_A : \mu \neq 72 \text{ battiti al minuto.}$$

Eseguiamo il test ad un livello di significatività $\alpha = 5\%$. Poichè non conosciamo il valore della deviazione standard σ della popolazione in

questione, dobbiamo utilizzare il test t di Student. Osserviamo che la media campionaria è

$$\bar{x} = 68.7 \text{ battiti al minuto}$$

e che la deviazione standard campionaria è

$$s = 8.66 \text{ battiti al minuto.}$$

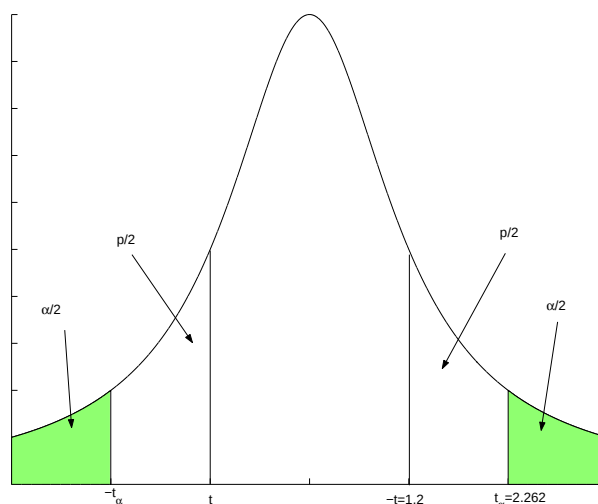
Il valore t del nostro test è dunque

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{68.7 - 72}{\frac{8.66}{\sqrt{10}}} = \frac{-3.3}{\frac{8.66}{\sqrt{10}}} = -1.20.$$

il valore critico t_α che corrisponde al livello di significatività $\alpha = 0.05$ lo troviamo consultando la Tabella A.4. Scopriamo così che il valore t che in corrispondenza di $10 - 1 = 9$ gradi di libertà lascia il 2.5% in ogni coda è

$$t_\alpha = 2.262.$$

Dunque abbiamo la seguente situazione:



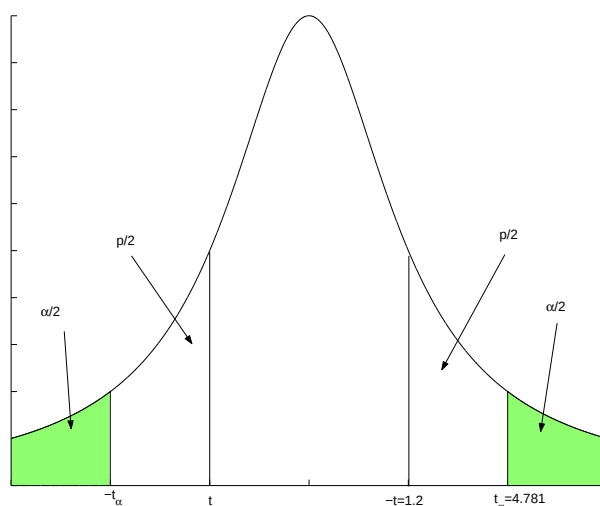
Pertanto il valore p del test è *superiore* ad α : $p > \alpha$. Quindi non possiamo rifiutare l'ipotesi nulla H_0 . Il test *non* è statisticamente significativo.

Vediamo se cambia la situazione fissando un livello di significatività più basso: $\alpha = 0.01$. Allora in questo caso, consultando la Tabella

A.4, troviamo che il valore t che in corrispondenza di 9 gradi di libertà lascia il 0.05% in ciascuna coda è

$$t_{\alpha} = 4.781.$$

Allora abbiamo la seguente situazione:



Pertanto anche con un livello di significatività più basso rispetto al precedente, concludiamo che *non* possiamo rifiutare l'ipotesi nulla e quindi anche in questo caso il test non è statisticamente significativo.

3. Dobbiamo eseguire un test d'ipotesi sull'incremento medio del battito cardiaco, testando l'ipotesi nulla

$$H_0 : \mu = 0,$$

come detto nel suggerimento. L'ipotesi alternativa, poichè a noi interessa un *incremento* del battito cardiaco, è

$$H_A : \mu > 0.$$

Dunque stiamo eseguendo un test d'ipotesi *unilaterale*.

Poichè non conosciamo la deviazione standard della popolazione in questione, ricorreremo al test t di Student. Sappiamo che la media campionaria è

$$\bar{x} = 1.4 \text{ battiti al minuto}$$

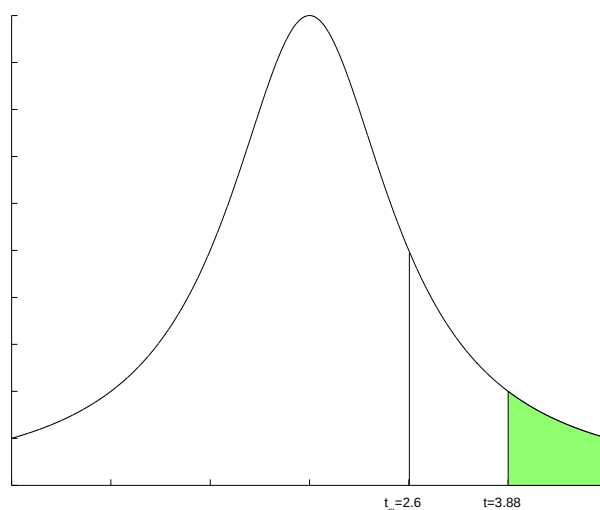
e che la deviazione standard campionaria è

$$s = 3.6 \text{ battiti al minuto.}$$

Il valore t del nostro test dunque è

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1.4 - 0}{\frac{3.6}{\sqrt{100}}} = 3.88.$$

Consultando la Tabella A.4, troviamo che il valore t_α corrispondente al livello di significatività $\alpha = 0.005$ per 99 gradi di libertà è circa uguale a $t_\alpha = 2.6$. Pertanto abbiamo la seguente situazione:



Concludiamo dunque che il valore p del test è *inferiore* ad α : $p < \alpha$. Pertanto possiamo rifiutare l'ipotesi nulla e concludere che il farmaco può essere messo in commercio.

4. (a) L'ipotesi nulla del test è

$$H_0 : \mu = 7250/mm^3$$

mentre l'ipotesi alternativa è

$$H_A : \mu < 7250/mm^3$$

in quanto si ritiene che i soggetti infetti abbiano in media un numero *inferiore* di globuli bianchi.

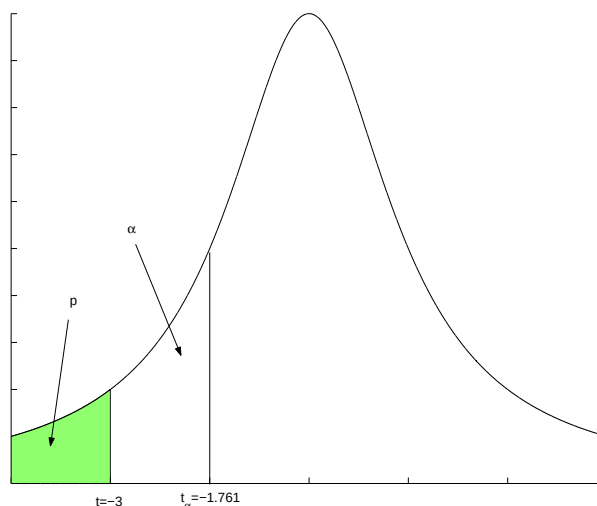
- (b) Poichè non conosciamo la deviazione standard della popolazione in questione, eseguiamo un test t di Student. Essendo $\bar{x} = 4767/mm^3$ e $s = 3204/mm^3$, il valore t del test nel nostro caso è

$$t = \frac{4767 - 7250}{\frac{3204}{\sqrt{15}}} = \frac{-2483}{827} \approx -3.$$

Consultando la Tabella A.4, troviamo che il valore t_α che lascia una coda *sinistra* di 0.05 in corrispondenza di $15 - 1 = 14$ gradi di libertà è

$$t_\alpha = -1.761.$$

La situazione dunque è la seguente:



Pertanto concludiamo che il valore p del test è *inferiore* ad α e quindi possiamo rifiutare l'ipotesi nulla e concludere che si può ritenere che il numero di globuli bianchi in soggetti infetti diminuisca.

5. Osserviamo che, dal momento che non si conosce la deviazione standard della popolazione in questione, per calcolare l'intervallo di confidenza richiesto dobbiamo ricorrere alla Tabella A.4 del test t di Student. Il valore che lascia in ciascuna coda il 2.5% è circa uguale a $t = 2.00$. Quindi l'intervallo di confidenza richiesto è

$$\left(25 - 2 \cdot \frac{2.7}{\sqrt{58}}, 25 + 2 \cdot \frac{2.7}{\sqrt{58}} \right) = (24.3, 25.7).$$

6. L'ipotesi nulla del test è

$$H_0 : \mu = 24 \text{ kg/m}^2$$

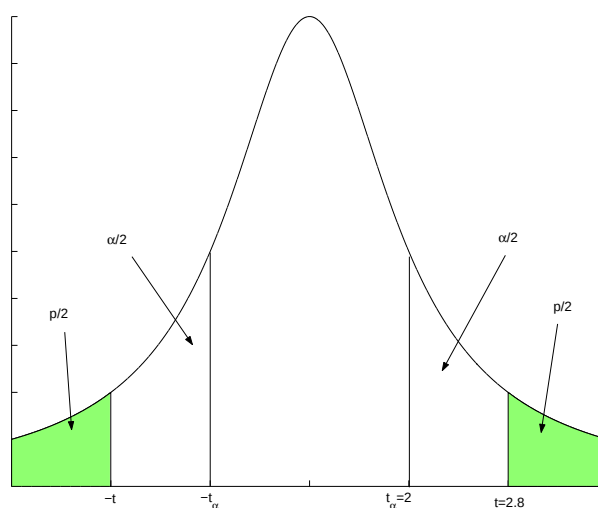
mentre l'ipotesi alternativa è

$$H_A : \mu \neq 24 \text{ kg/m}^2.$$

Stiamo dunque eseguendo un test *bilaterale*. Essendo $\bar{x} = 25 \text{ kg/m}^2$ e $s = 2.7 \text{ kg/m}^2$, il valore t del test nel nostro caso è

$$t = \frac{25 - 24}{\frac{2.7}{\sqrt{58}}} = \frac{1}{\frac{2.7}{\sqrt{58}}} = 2.8.$$

Il valore t_α che lascia in ogni coda il 2.5%, come già trovato al punto (a), è $t_\alpha = 2$. Pertanto la situazione è la seguente:



Concludiamo dunque che il valore p del test è *inferiore* ad α e quindi che possiamo rifiutare l'ipotesi nulla: è verosimile che chi è in situazione di sovrappeso sia maggiormente esposto al rischio di contrarre il diabete mellito.

Capitolo 5

Confronto fra 2 medie

5.1 Esercizi

1. I seguenti pesi, misurati in once, sono stati rilevati in un campione di topi da laboratorio durante lo studio di una particolare dieta.
 - (a) Si usi $\alpha = 0.05$ e si verifichi se la dieta è stata efficace per la riduzione del peso.
 - (b) Si calcoli un intervallo di confidenza *unilaterale* ad livello $\alpha = 0.05$.

Topo	1	2	3	4	5	6	7	8	9	10
Prima	14	27	19	17	19	12	15	15	21	19
Dopo	16	18	17	16	16	11	15	12	21	18

2. E' stato condotto uno studio per valutare l'efficacia della cotinina nella saliva come indicatore dell'esposizione al fumo di tabacco. In una parte dello studio, a 7 soggetti - nessuno dei quali era un forte fumatore e tutti avevano smesso di fumare almeno una settimana prima dell'inizio dello studio - è stato chiesto di fumare una sola sigaretta. Sono stati poi prelevati dei campioni di saliva da tutti i soggetti 2,12,24 e 48 ore dopo aver fumato la sigaretta. I livelli di cotinina a 12 e 24 ore sono di seguito riportati:

Soggetto	1	2	3	4	5	6	7
Dopo 12 ore	73	58	67	93	33	18	147
Dopo 24 ore	24	27	49	59	0	11	43

Sia μ_{12} il livello medio di cotinina della popolazione 12 ore dopo aver

fumato la sigaretta e μ_{24} il livello medio di cotinina 24 ore dopo aver fumato la sigaretta. Si ritiene che μ_{24} sia inferiore a μ_{12} .

- (a) Calcolare l'intervallo di confidenza unilaterale al 95% per la reale differenza nelle medie delle popolazioni $\mu_{12} - \mu_{24}$.
 - (b) Testare l'ipotesi nulla che le medie delle popolazioni sono uguali ad un livello di significatività $\alpha = 0.05$.
3. E' stato condotto uno studio per determinare se il fumo materno ha effetto sul contenuto minerale dei neonati sani. Un campione casuale di 77 neonati le cui madri avevano fumato durante la gravidanza ha un contenuto minerale osseo medio $\bar{x}_1 = 0.098$ g/cm e una deviazione standard $s_1 = 0.026$ g/cm; un campione casuale di 161 neonati le cui madri non avevano fumato ha una media $\bar{x}_2 = 0.095$ g/cm e una deviazione standard $s_2 = 0.025$ g/cm. Si assuma che le varianze delle popolazioni originarie siano uguali.
- (a) Formulare l'ipotesi nulla e l'ipotesi alternativa del test bilaterale.
 - (b) Eseguire il test ad un livello di significatività $\alpha = 0.05$.
 - (c) Che cosa si può concludere?
4. E' stato condotto uno studio al fine di determinare se la crusca d'avena aiuta a ridurre il colesterolo sierico in pazienti maschi ipercolesterolemici. Un campione casuale di 14 soggetti è stato sottoposto ad una dieta con crusca d'avena o fiocchi di granturco; dopo 2 settimane, sono stati registrati i livelli di colesterolo legati a lipoproteine a bassa densità. Ciascun soggetto è stato poi sottoposto all'altra dieta. Dopo 2 settimane, è stato di nuovo registrato il livello di colesterolo LDL di ciascun soggetto. I dati di questo studio sono di seguito riportati:

Soggetto	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Granturco	4.61	6.42	5.40	4.54	3.98	3.82	5.01	4.34	3.80	4.56	5.35	3.89	2.25	4.24
Avena	3.84	5.57	5.85	4.80	3.68	2.96	4.41	3.72	3.49	3.84	6.26	3.73	1.84	4.14

- (a) Quali sono le ipotesi nulla ed alternativa per un test bilaterale?
 - (b) Eseguire un test ad un livello di significatività di 0.05.
 - (c) Che cosa si può concludere?
5. In uno studio sull'ipertensione gravida, un gruppo di donne è stato trattato con basse dosi di aspirina ed un secondo gruppo con placebo. Un campione casuale di 23 donne che avevano ricevuto aspirina ha una

pressione arteriosa media $\bar{x}_1 = 11$ mmHg ed una deviazione standard $s_1 = 8$ mmHg; un secondo campione casuale di 24 donne che avevano ricevuto placebo ha una pressione arteriosa media $\bar{x}_2 = 109$ mmHg ed una deviazione standard $s_2 = 8$ mmHg. Si suppone che le due popolazioni abbiano uguale varianza.

- (a) Ad un livello di significatività $\alpha = 0.05$, testare l'ipotesi nulla che le due popolazioni di donne abbiano la stessa pressione arteriosa media.
 - (b) Calcolare l'intervallo di confidenza al 90% per la reale differenza nelle medie delle popolazioni.
6. La seguente tabella confronta i livelli di carbossiemoglobina per un gruppo di non fumatori ed un gruppo di fumatori di sigarette. Sono riportate le medie e le deviazioni standard del campione. Si ritiene che il livello medio di carbossiemoglobina dei fumatori sia più elevato di quello dei non fumatori. Si assuma che le varianze delle 2 popolazioni siano uguali.

Gruppo	n	Carbossiemoglobina
Non fumatori	121	$\bar{x} = 1.3, s=1.3$
Fumatori	75	$\bar{x} = 4.1, s=2.0$

- (a) Quali sono le ipotesi nulla ed alternativa del test?
 - (b) Eseguire un test ad un livello di significatività $\alpha=0.01$.
 - (c) Che cosa si può concludere?
7. Supponete di voler confrontare le caratteristiche della meningite tubercolare in pazienti affetti da virus dell'immunodeficienza umana e soggetti non infetti. In particolare, volete determinare se le due popolazioni abbiano la stessa età media. Un campione di 37 pazienti infetti ha un'età media $\bar{x}_1 = 27.9$ anni ed una deviazione standard $s_1 = 5.6$ anni; un campione casuale di 19 pazienti non infetti ha un'età media $\bar{x}_2 = 38.8$ anni ed una deviazione standard $s_2 = 21.7$ anni.
- (a) Testare l'ipotesi nulla che le 2 popolazioni di pazienti abbiano la stesse età media ad un livello di significatività $\alpha = 0.05$.
 - (b) Calcolare l'intervallo di confidenza al 95% per la reale differenza nelle medie delle popolazioni.

5.2 Soluzioni

1. (a) Si tratta di una chiara situazione in cui i campioni risultano *appaiati*. Dette μ_p e μ_d rispettivamente le medie dei pesi dei topi prima e dopo la dieta, posto

$$\delta = \mu_p - \mu_d,$$

l'ipotesi nulla del nostro test è

$$H_0 : \delta = 0$$

mentre l'ipotesi alternativa, dal momento che ci interessa verificare l'efficacia della dieta, è

$$H_A : \delta > 0 \quad (\mu_d < \mu_p).$$

Eseguiamo dunque un *test unilaterale* ad un livello di significatività $\alpha = 0.05$.

Costruiamo la tabella delle differenze:

Topo	1	2	3	4	5	6	7	8	9	10
Differenza	-2	9	2	1	3	1	0	3	0	1

La *media* delle differenze dunque è

$$\bar{d} = \frac{-2 + 9 + 2 + 1 + 3 + 1 + 0 + 3 + 0 + 1}{10} = 1.8,$$

mentre la *deviazione standard* delle differenze è

$$s_d = \sqrt{\frac{\sum_{i=1}^{10} (d_i - \bar{d})^2}{10 - 1}} = 2.93.$$

Calcoliamo allora il test statistico t con 9 *gradi di libertà*:

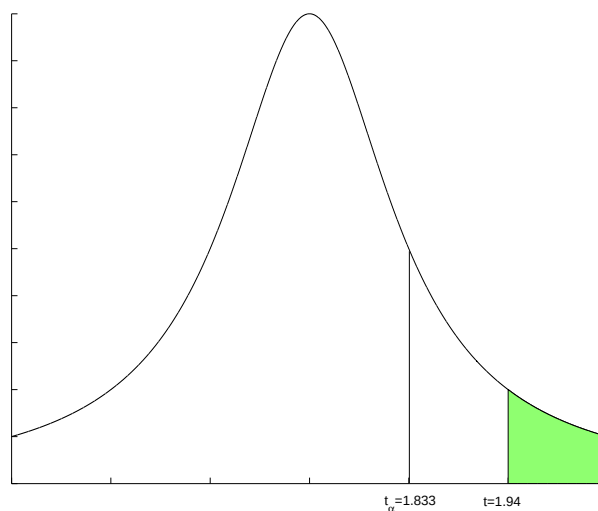
$$t_9 = \frac{\bar{d} - \delta}{\frac{s_d}{\sqrt{n}}} = \frac{1.8 - 0}{\frac{2.93}{\sqrt{10}}} = 1.94.$$

Consultando la Tabella A.4 del testo, troviamo che il valore critico t_α per una coda con $\alpha = 0.05$ è

$$t_\alpha = 1.833.$$

Dunque la situazione è la seguente:

Università di Torino



Troviamo quindi che il valore p del test è inferiore ad α : $p < \alpha$. Pertanto *rifiutiamo* l'ipotesi nulla H_0 e concludiamo che il test è statisticamente significativo: la dieta eseguita è stata efficace.

- (b) Calcoliamo ora un intervallo di confidenza unilaterale per δ ad un livello di confidenza $\alpha = 0.05$. Ovviamente, è lo stesso problema del punto (a), ma visto sotto un'ottica diversa.

Noi sappiamo, per quanto già visto al punto precedente, che

$$P(t_9 \leq 1.833) = 0.95.$$

Dunque deve essere

$$t_9 = \frac{\bar{d} - \delta}{\frac{s_d}{\sqrt{n}}} \leq 1.833,$$

da cui

$$\bar{d} - \delta \leq 1.833 \cdot \frac{s_d}{\sqrt{10}},$$

e quindi

$$\delta \geq \bar{d} - 1.833 \cdot \frac{s_d}{\sqrt{10}} = 1.8 - 1.833 \cdot \frac{2.93}{\sqrt{10}} = 0.101.$$

Abbiamo dunque trovato l'intervallo di confidenza unilaterale richiesto per δ : $\delta \geq 0.101$.

2. (a) Anche in questo caso siamo in presenza di *campioni appaiati*. Sia $\delta = \mu_{12} - \mu_{24}$.

Calcoliamo subito la tabella delle differenze:

Soggetto	1	2	3	4	5	6	7
Differenza	49	31	18	34	33	7	104

La media delle differenze è

$$\bar{d} = \frac{49 + 31 + 18 + 34 + 33 + 7 + 104}{7} = 39.43,$$

mentre la varianza delle differenze è

$$s_d^2 = \frac{\sum_{i=1}^7 (d_i - \bar{d})^2}{7 - 1} = 985.62.$$

Vogliamo un intervallo di confidenza *unilaterale* al 95%. Il valore critico t_α per $\alpha = 0.05$ in una coda con 6 gradi di libertà è, secondo la Tabella A.4 del testo:

$$t_\alpha = 1.943.$$

Calcoliamo ora il test t a 6 gradi di libertà:

$$t_6 = \frac{\bar{d} - \delta}{\frac{s_d}{\sqrt{7}}} = \frac{39.43 - \delta}{\frac{\sqrt{985.62}}{\sqrt{7}}}.$$

Vogliamo trovare allora δ tale che

$$P\left(\frac{39.43 - \delta}{\frac{\sqrt{985.62}}{\sqrt{7}}} \leq 1.943\right) = 0.95.$$

Deve essere

$$\frac{39.43 - \delta}{\frac{\sqrt{985.62}}{\sqrt{7}}} \leq 1.943,$$

da cui

$$\delta \geq 39.43 - 1.943 \cdot \frac{\sqrt{985.62}}{\sqrt{7}} = 16.37.$$

Confidiamo pertanto al 95% che la reale decrescita del livello media di cotinina sia (almeno) 16.37.

(b) L'ipotesi nulla è

$$H_0 : \delta = 0.$$

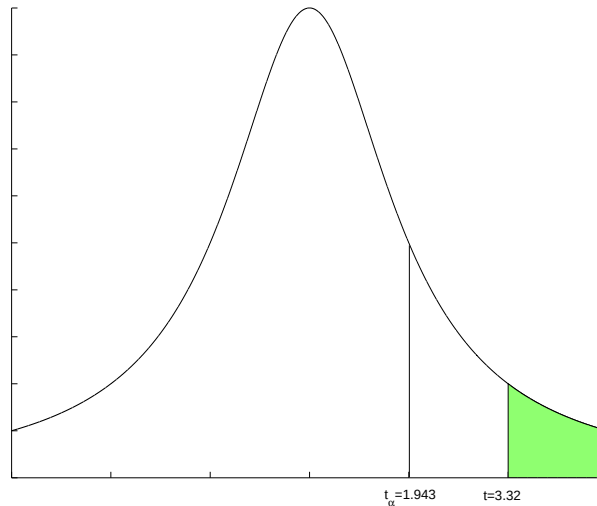
Poichè si ritiene che $\mu_{24} < \mu_{12}$, eseguiamo un test *unilaterale* in cui l'ipotesi alternativa è dunque

$$H_A : \delta > 0.$$

Il valore t_6 del test è, in base ai calcoli fatti al punto (a):

$$t_6 = \frac{\bar{d} - 0}{\frac{\sqrt{s_d^2}}{\sqrt{n}}} = \frac{35.43}{\frac{\sqrt{985.62}}{\sqrt{7}}} = 3.32.$$

La situazione allora è la seguente:



Troviamo che il valore p del test è inferiore ad $\alpha = 0.05$: $p < \alpha$. Pertanto possiamo rifiutare l'ipotesi nulla e concludere che il livello di cotinina dopo 24 è inferiore al livello di cotinina dopo 12 ore.

3. (a) Siamo in presenza di 2 *campioni indipendenti* con *uguale varianza*. Dette μ_1 la media della popolazione di neonati con madre fumatrice e μ_2 la media della popolazioe di neonati con madre non fumatrice, l'ipotesi nulla del test è

$$H_0 : \mu_1 = \mu_2$$

mentre l'ipotesi alternativa è

$$H_A : \mu_1 \neq \mu_2.$$

(b) Calcoliamo la *varianza pooled*:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{76 \cdot (0.026)^2 + 160 \cdot (0.025)^2}{77 + 161 - 2} = 0.00064.$$

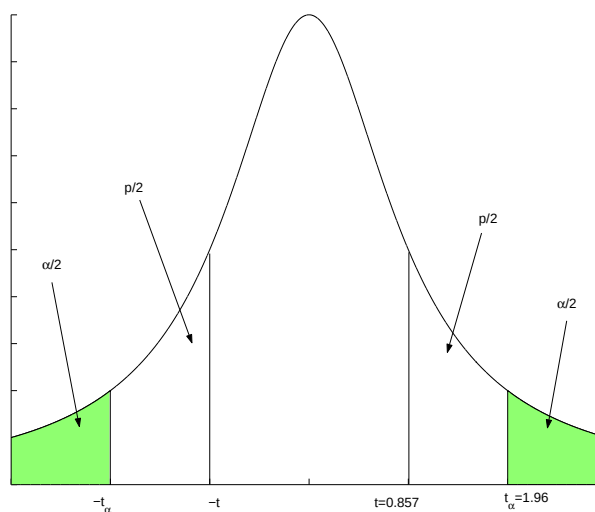
Calcoliamo ora il test t con $77 + 161 - 2 = 236$ gradi di libertà:

$$t_{236} = \frac{\mu_1 - \mu_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.003}{\sqrt{0.00064 \left(\frac{1}{77} + \frac{1}{161} \right)}} = 0.857.$$

Il valore critico per t con $\alpha/2 = 0.025$ in ciascuna delle 2 code (in quanto è un test bilaterale) con 236 gradi di libertà è

$$t_{\alpha/2} = 1.96.$$

La situazione dunque è la seguente:



Pertanto troviamo che il valore del test p è maggiore di $\alpha = 0.05$. Il test non è statisticamente significativo: sulla base dei dati a disposizione, non possiamo negare il fatto che non ci sia alcuna differenza tra madri fumatrici e madri non fumatrici.

4. (a) Siano μ_g e μ_a rispettivamente le medie delle 2 popolazioni per i fiocchi di granturco e la crusca d'avena. Sia $\delta = \mu_g - \mu_a$. L'ipotesi nulla allora è

$$H_0 : \delta = 0$$

mentre l'ipotesi alternativa è

$$H_A : \delta > 0,$$

poichè siamo interessati a vedere se la dieta con crusca d'avena *riduce* il livello di colesterolo sierico.

- (b) Eseguiamo pertanto un test *unilaterale* ad un livello di confidenza $\alpha = 0.05$. Siamo nel caso di *campioni appaiati*. Calcoliamo le differenze:

Soggetto	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Differenza	0.77	0.85	-0.45	-0.26	0.30	0.86	0.60	0.62	0.31	0.72	-0.91	0.16	0.41	0.10

La media delle differenze è

$$\bar{d} = \frac{\sum_{i=1}^{14} d_i}{14} = 0.29,$$

mentre la deviazione standard delle differenze è

$$s_d = \sqrt{\frac{\sum_{i=1}^{14} (d_i - \bar{d})^2}{14 - 1}} = 0.53.$$

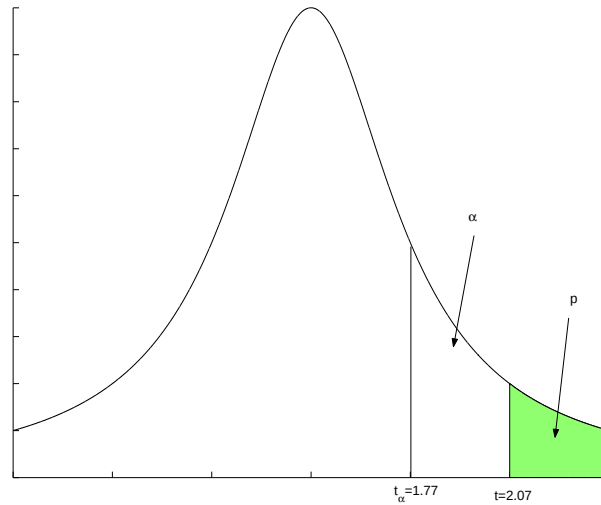
Il valore t del test per $14 - 1 = 13$ gradi di libertà è allora

$$t_{13} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{0.29}{\frac{0.53}{\sqrt{14}}} = 2.07.$$

Il valore critico del test t per 13 gradi di libertà per 1 coda con un livello $\alpha = 0.05$ è, come risulta dalla Tabella A.4:

$$t_\alpha = 1.77.$$

La situazione allora è la seguente:



- (c) Riscontriamo che il valore p del test è *inferiore* ad α : $p < \alpha$. Dunque il test è statisticamente significativo: la dieta con crusca d'avena aiuta a ridurre il livello di colesterolo sierico.
5. (a) Siano μ_a e μ_p le medie delle 2 popolazioni in questione. Sia $\delta = \mu_a - \mu_p$. L'ipotesi nulla è

$$H_0 : \delta = 0$$

mentre l'ipotesi alternativa è

$$H_A : \delta \neq 0.$$

Stiamo dunque eseguendo un test *bilaterale* per campioni *indipendenti* con varianza uguale.

Calcoliamo la varianza pooled:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(23 - 1) \cdot 8^2 + (24 - 1) \cdot 8^2}{23 + 23 - 2} = 64.$$

Eseguiamo ora il test a $n_1 + n_2 - 2 = 45$ gradi di libertà:

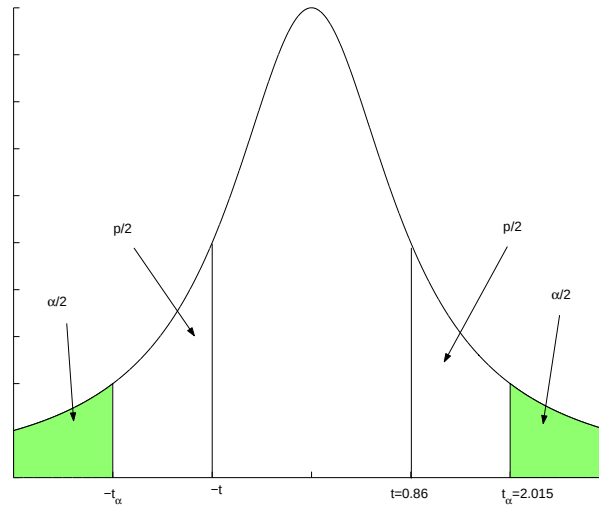
$$t_{45} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{111 - 109 - 0}{\sqrt{64 \cdot \left(\frac{1}{23} + \frac{1}{24} \right)}} = \frac{1}{4\sqrt{\frac{1}{23} + \frac{1}{24}}} = 0.86.$$

Il valore critico t del test per 2 code con 45 gradi di libertà è

$$t_{\alpha/2} = 2.015.$$

La situazione dunque è la seguente:

Università di Torino



Pertanto vediamo che il valore p del test è *maggiore* di α : il test dunque non è statisticamente significativo e non possiamo negare una differenza fra il trattamento con aspirina e quello con placebo in donne gravide.

- (b) Il valore critico di t per 45 gradi di libertà in corrispondenza di un'area in una coda pari al 5% è $t = 1.68$. Noi dunque sappiamo che

$$P\left(-1.68 \leq \frac{111 - 109 - \delta}{\sqrt{64 \cdot \left(\frac{1}{23} + \frac{1}{24}\right)}} \leq 1.68\right) = 0.90.$$

Allora deve essere

$$-1.68 \leq \frac{111 - 109 - \delta}{\sqrt{64 \cdot \left(\frac{1}{23} + \frac{1}{24}\right)}}$$

da cui

$$-1.68 \cdot 8 \cdot \sqrt{\frac{1}{23} + \frac{1}{24}} \leq 2 - \delta \leq 1.68 \cdot 8 \cdot \sqrt{\frac{1}{23} + \frac{1}{24}}$$

e quindi

$$-1.9 \leq \delta \leq 5.9.$$

Siamo confidenti al 90% che la reale differenza fra le medie delle 2 popolazioni sia nell'intervallo $(-1.9, 5.9)$. Osserviamo che questo intervallo contiene anche il valore 0 (cioè possibile uguaglianza fra le 2 medie).

6. (a) Siano μ_A e μ_N rispettivamente le medie delle 2 popolazioni in questione. Sia $\delta = \mu_A - \mu_N$. L'ipotesi nulla è

$$H_0 : \delta = 0$$

mentre l'ipotesi alternativa è

$$H_A : \delta \neq 0.$$

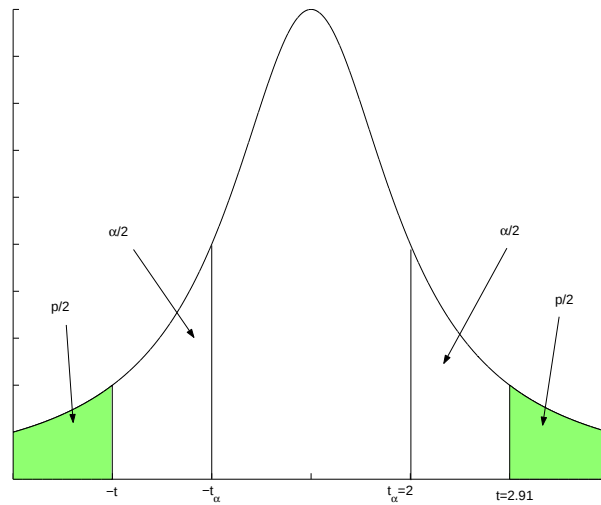
Stiamo dunque eseguendo un test bilaterale ad un livello di significatività $\alpha = 0.05$ per 2 campioni *indipendenti* con uguale varianza. Calcoliamo la varianza pooled:

$$s_p^2 = \frac{(37 - 1) \cdot 5.6^2 + (19 - 1) \cdot 21.7^2}{37 + 19 - 2} = 177.87.$$

Il valore del test t è allora:

$$t_{54} = \frac{38.8 - 27.9}{\sqrt{177.87 \cdot (\frac{1}{37} + \frac{1}{19})}} = \frac{10.9}{3.748} = 2.91.$$

Il valore critico di t per 54 gradi di libertà in corrispondenza di una coda del 2.5% è $t_{\alpha/2} = 2.00$. Dunque abbiamo la seguente situazione:



Vediamo pertanto che il valore p del test è inferiore ad α : $p < \alpha$. Quindi il test è statisticamente significativo: rifiutiamo l'ipotesi nulla e concludiamo che le 2 popolazioni hanno età medie differenti.

(b) Noi sappiamo che

$$P\left(-2.00 \leq \frac{38.8 - 27.9 - \delta}{\sqrt{177.87 \cdot (\frac{1}{37} + \frac{1}{19})}} \leq 2.00\right) = 0.95.$$

Allora deve essere

$$-2 \cdot 3.748 \leq 10.9 - \delta \leq 2 \cdot 3.748$$

da cui

$$3.4 \leq \delta \leq 18.4.$$

Come vediamo, l'intervallo di confidenza al 95% per δ non contiene lo 0 e questo è in linea con il risultato trovato al punto (a) eseguendo il test d'ipotesi ad un livello di significatività $\alpha = 0.05$.

7. (a) Siano μ_f e μ_n rispettivamente i livelli medi di carbossiemoglobina della popolazione dei fumatori e dei non fumatori. Sia $\delta = \mu_f - \mu_n$. Si ritiene che $\mu_f > \mu_n$. Dunque l'ipotesi nulla del test è

$$H_0 : \delta = 0$$

mentre l'ipotesi alternativa è

$$H_A : \delta > 0.$$

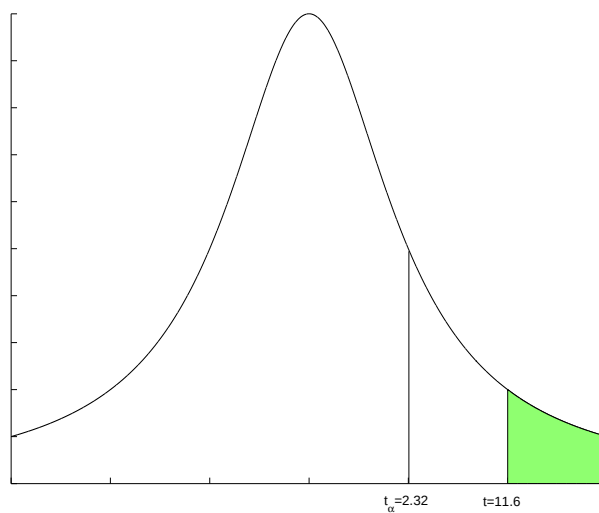
- (b) Siamo in presenza di 2 campioni *indipendenti* con *uguale* varianza. Calcoliamo allora la varianza pooled:

$$s_p^2 = \frac{120 \cdot 1.3^2 + 74 \cdot 2^2}{121 + 75 - 2} = 2.57.$$

Il valore t del test è dunque

$$t_{121+75-2} = t_{194} = \frac{4.1 - 1.3}{\sqrt{2.57 \cdot (\frac{1}{121} + \frac{1}{75})}} = \frac{2.8}{0.24} = 11.6.$$

Il valore critico di t per 194 gradi di libertà con una coda corrispondente all'1% è $t_\alpha = 2.32$. Come vediamo, visto l'elevato numero di gradi di libertà, esso coincide con il corrispondente valore della normale. La situazione pertanto è la seguente:



- (c) Constatiamo allora che il valore p del test è *inferiore* ad α . Pertanto il test è statisticamente significativo e possiamo rifiutare l'ipotesi nulla: il livello medio di carbossiemoglobina è più elevato nei fumatori.

Capitolo 6

Analisi della varianza

6.1 Esercizi

1. E' stato condotto un esperimento per confrontare il raccolto di 4 varietà di riso. Ognuno dei 16 appezzamenti della fattoria sottoposta al test è stato trattato in modo simile per quanto concerne l'acqua e il fertilizzante. Quattro appezzamenti sono stati assegnati casualmente ad ognuna delle 4 varietà di riso. Il raccolto di ogni appezzamento è stato annotato in libbre per acro nella seguente tabella:

Varietà	Raccolti			
1	934	1041	1028	935
2	880	963	924	946
3	987	951	976	840
4	992	1143	1140	1191

I dati della tabella indicano una differenza nel raccolto medio delle 4 varietà? Usare un'analisi della varianza con $\alpha = 0.05$.

2. Uno studio dei vigili del fuoco in un'estesa area urbana si è occupato delle condizioni fisiche degli ingegneri impiegati del Dipartimento per lo lotta contro il fuoco. Per misurare tali caratteristiche, una fisioterapista ha campionato 5 ingegneri con un'esperienza nel dipartimento di 5,10,15 e 20 anni. Quindi ha registrato il numero di piegamenti che ogni persona riusciva a fare in 60 secondi. I risultati sono raccolti nella tabella sottostante:

Anni trascorsi nel dipartimento	5	10	15	20
	56	64	45	42
	55	61	46	39
	62	50	45	45
	59	57	39	43
	60	55	43	41

Eseguire un'analisi della varianza per determinare se vi siano delle differenze nelle condizioni fisiche tra i gruppi di ingegneri dovute all'anzianità di servizio. Usare $\alpha = 0.05$.

3. Uno degli obiettivi della Edinburgh Artery Society è di valutare i fattori di rischio delle patologie arteriose periferiche in soggetti di età compresa fra i 55 e i 74 anni. Si vogliono confrontare i livelli medi di colesterolo LDL, misurati in mmol/litro, in 4 diverse popolazioni di soggetti: pazienti con claudicazione intermittente o interruzione del movimento, pazienti con patologie asintomatiche maggiori, pazienti con patologie asintomatiche minori, e soggetti asintomatici. E' stato selezionato da ciascuna popolazione un campione casuale; di seguito sono riportate le statistiche di sintesi:

	n	$\bar{x}$	s
claudicazione intermittente	73	6.22	1.62
patologie asintomatiche maggiori	105	5.81	1.43
patologie asintomatiche minori	240	5.77	1.24
assenza di malattia	1080	5.47	1.31

- (a) Testare l'ipotesi nulla che i livelli di colesterolo LDL son gli stessi per ciascuna delle 4 popolazioni.
- (b) Quali sono i gradi di libertà associati con questo test?
- (c) Che cosa si può concludere?
4. In un esperimento per determinare l'efficacia dei farmaci contro l'insonnia, sono stati assegnati casualmente 18 pazienti sofferenti di insonnia a 3 diversi trattamenti:
- (a) placebo
- (b) un farmaco standard
- (c) un nuovo farmaco sperimentale

Le risposte riportate nella tabella sottostante sono la durata media (in ore) di sonno per notte in una settimana. Effettuare un'analisi della varianza con $\alpha = 0.05$.

1	2	3
5.6	8.4	10.6
5.7	8.2	6.6
5.1	8.8	8.0
3.8	7.1	8.0
4.6	7.2	6.8
5.1	8.0	6.6

5. L'assorbimento da parte del suolo dei metalli che fuoriescono nell'aria durante alcuni processi industriali produce gravi danni ambientali. Per accertarsi se le percentuali di assorbimento variano tra i tipi di terreno, sono stati casualmente scelti 6 campioni di terre coltivate, aventi 5 tipi di suolo differenti (1, 2, 3, 4, 5) in un'area nota per avere un'esposizione relativamente uniforme ai metalli osservati. I 30 campioni di terreno sono stati analizzati per contenuto di cadmio (Cd). I risultati sono presentati nella seguente tabella. Eseguire un'analisi della varianza per determinare se vi siano differenze nel contenuto di cadmio tra i terreni.

1	2	3	4	5
0.54	0.56	0.39	0.26	0.32
0.63	0.56	0.28	0.13	0.33
0.73	0.52	0.29	0.19	0.34
0.58	0.41	0.32	0.28	0.34
0.66	0.50	0.30	0.10	0.36
0.70	0.60	0.27	0.20	0.32

6. E' stato condotto uno studio su pazienti insulino-dipendenti per esaminare gli effetti del fumo di sigaretta sulle complicanze renali e retiniche. Prima di esaminare i risultati dello studio, si vogliono confrontare le misure iniziali di pressione sistolica in 4 diversi gruppi: non fumatori, fumatori, ex fumatori e masticatori di tabacco. E' stato selezionato da ciascun gruppo un campione; i dati relativi sono di seguito riportati. Le medie e le deviazioni standard sono misurate in millimetri di mercurio.

	n	$\bar{x}$	s
Non fumatori	269	115	13.4
Fumatori	53	114	10.1
Ex fumatori	28	118	11.6
Masticatori di tabacco	9	126	12.2

- (a) Calcolare la stima della varianza entro i gruppi.
- (b) Calcolare la stima della varianza tra i gruppi.
- (c) Ad un livello di significatività di 0.05, testare l'ipotesi nulla che le pressioni sistoliche medie dei 4 gruppi siano uguali.
- (d) Che cosa si può concludere?

6.2 Soluzioni

1. Dette $\mu_1, \mu_2, \mu_3, \mu_4$ le medie (sconosciute) delle 4 varietà di riso, l'ipotesi nulla del test è

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Il campione della prima varietà di riso ha una media (in libbre per acro)

$$\bar{x}_1 = \frac{934 + 1041 + 1028 + 935}{4} = 984.5$$

ed una varianza campionaria

$$s_1^2 = \frac{(934 - 984.5)^2 + (1041 - 984.5)^2 + (1028 - 984.5)^2 + (935 - 984.5)^2}{3} = 3361.667.$$

Il campione della seconda varietà di riso ha una media

$$\bar{x}_2 = \frac{880 + 963 + 924 + 946}{4} = 928.25$$

ed una varianza campionaria

$$s_2^2 = \frac{(880 - 928.25)^2 + (963 - 928.25)^2 + (924 - 928.25)^2 + (946 - 928.25)^2}{3} = 1289.583.$$

Il campione della terza varietà di riso ha una media

$$\bar{x}_3 = \frac{987 + 951 + 976 + 840}{4} = 938.5$$

ed una varianza campionaria

$$s_3^2 = \frac{(987 - 938.5)^2 + (951 - 938.5)^2 + (976 - 938.5)^2 + (840 - 938.5)^2}{3} = 4539.$$

Infine, il campione della quarta varietà di riso ha una media

$$\bar{x}_4 = \frac{992 + 1143 + 1140 + 1191}{4} = 1116.5$$

ed una varianza campionaria

$$s_4^2 = \frac{992 - 1116.5)^2 + (1143 - 1116.5)^2 + (1140 - 1116.5)^2 + (1191 - 1116.5)^2}{3} = 7435.$$

La *media globale* dei dati è

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n_1 + n_2 + n_3 + n_4} = \frac{4 \cdot 984.5 + 4 \cdot 928.25 + 4 \cdot 938.5 + 4 \cdot 1116.5}{4 + 4 + 4 + 4} = 991.9375.$$

Calcoliamo ora la *varianza tra gruppi*:

$$\begin{aligned} s_B^2 &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2}{4 - 1} = \\ &= \frac{4 \cdot (984.5 - 991.9375)^2 + 4 \cdot (928.25 - 991.9375)^2 + 4 \cdot (938.5 - 991.9375)^2 +}{3} \\ &\quad + \frac{4 \cdot (1116.5 - 991.9375)^2}{3} = \\ &= 29977.06. \end{aligned}$$

Calcoliamo ora la *varianza entro gruppi*:

$$\begin{aligned} s_W^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2}{n_1 + n_2 + n_3 + n_4 - 4} = \\ &= 3 \cdot \frac{3361.667 + 1289.583 + 4539 + 7435}{12} = \\ &= \frac{16625.25}{4} = \\ &= 4156.313. \end{aligned}$$

Il test statistico F è allora

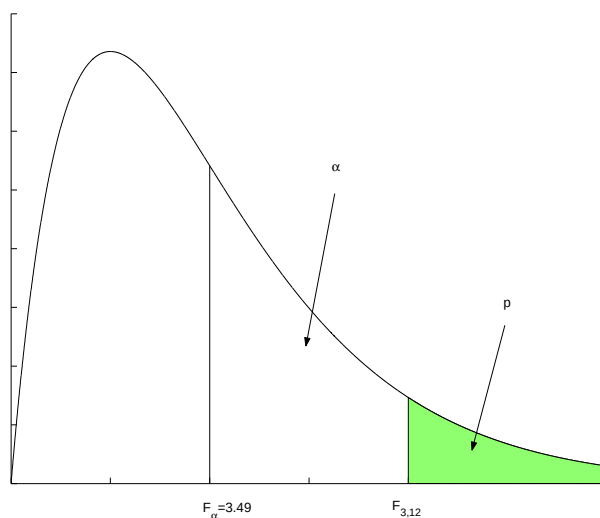
$$F_{4-1, 16-4} = F_{3, 12} = \frac{s_B^2}{s_W^2} = \frac{29977.06}{4156.313} = 7.21.$$

Il valore critico di F per 3 e 12 gradi libertà in modo da avere una coda superiore $\alpha = 0.05$ è (consultando la Tabella A.5 del testo)

$$F_\alpha = 3.49.$$

Dunque la situazione è la seguente:

Quaderni Didattici del Dipartimento di Matematica



Pertanto concludiamo che il valore p del test è *inferiore* ad α : $p < \alpha$. Dunque il test è statisticamente significativo e possiamo rifiutare l'ipotesi nulla e quindi la 4 varietà di riso presentano raccolti medi *differenti*.

2. Siano $\mu_1, \mu_2, \mu_3, \mu_4$ i numeri medi (sconosciuti) di piegamenti degli ingegneri rispettivamente con 5, 10, 15 e 20 anni di anzianità di servizio. L'ipotesi nulla è

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Il primo campione (con 5 anni di esperienza) ha una media

$$\bar{x}_1 = 58.4$$

ed una varianza

$$s_1^2 = 8.3.$$

Il secondo campione (con 10 anni di esperienza) ha una media

$$\bar{x}_2 = 57.4$$

ed una varianza

$$s_2^2 = 29.3.$$

Il terzo campione (con 15 anni di esperienza) ha una media

$$\bar{x}_3 = 43.6$$

ed una varianza

$$s_3^2 = 7.8.$$

Infine il quarto campione (con 20 anni di esperienza) ha una media

$$\bar{x}_4 = 42$$

ed una varianza

$$s_4^2 = 5.$$

Pertanto la *media globale* dei campioni è

$$\bar{x} = \frac{58.4 + 57.4 + 43.6 + 42}{4} = 50.35.$$

La *varianza entro gruppi* è

$$\begin{aligned} s_W^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2}{n_1 + n_2 + n_3 + n_4 - 4} = \\ &= 4 \cdot \frac{8.3 + 29.3 + 7.8 + 5}{5 + 5 + 5 + 5 - 4} = \\ &= 12.6. \end{aligned}$$

La *varianza tra gruppi* invece è

$$\begin{aligned} s_B^2 &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + (n_3 - 1)s_3^2 + n_4(\bar{x}_4 - \bar{x})^2}{4 - 1} = \\ &= 5 \cdot \frac{(58.4 - 50.35)^2 + (57.4 - 50.35)^2 + (43.6 - 50.35)^2 + (42 - 50.35)^2}{3} = \\ &= 382.983. \end{aligned}$$

Pertanto il test statistico è

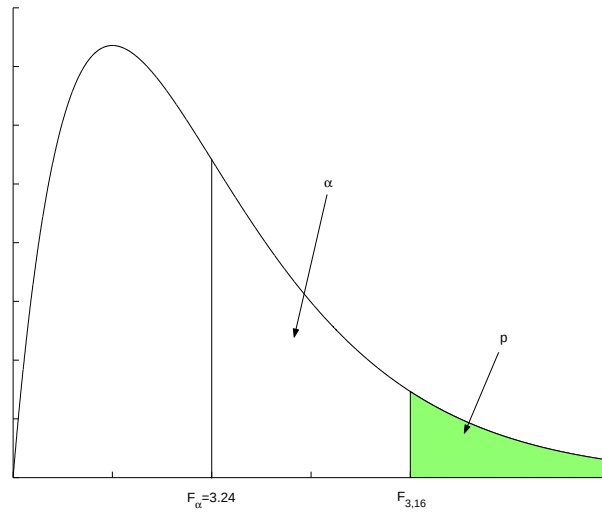
$$F_{4-1, 20-4} = F_{3, 16} = \frac{s_B^2}{s_W^2} = \frac{382.983}{12.6} = 30.4.$$

Il valore critico del test F per 3 e 16 gradi di libertà per avere una coda $\alpha = 0.05$ è (come si constata dalla Tabella A.5)

$$F_\alpha = 3.24.$$

Pertanto la situazione è la seguente:

Quaderni Didattici del Dipartimento di Matematica



Dunque il valore p del test è inferiore ad α : $p < \alpha$. Pertanto possiamo rifiutare l'ipotesi nulla: vi sono differenze nelle condizioni fisiche degli ingegneri dovute all'anzianità di servizio.

3. Detti $\mu_1, \mu_2, \mu_3, \mu_4$ i livelli di colesterolo medi delle 4 popolazioni, l'ipotesi nulla è

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

La *media globale* è

$$\bar{x} = \frac{73 \cdot 6.22 + 105 \cdot 5.81 + 240 \cdot 5.77 + 1080 \cdot 5.47}{73 + 105 + 240 + 1080} = 5.58.$$

La *varianza tra gruppi* è

$$s_B^2 = \frac{73 \cdot (6.22 - 5.58)^2 + 105 \cdot (5.81 - 5.58)^2 + 240 \cdot (5.77 - 5.81)^2 + 1080 \cdot (5.47 - 5.58)^2}{4 - 1} = 19.06.$$

La *varianza entro gruppi* è

$$s_W^2 = \frac{72 \cdot 1.62^2 + 104 \cdot 1.43 + 241 \cdot 1.24 + 1079 \cdot 1.31}{73 + 105 + 240 + 1080 - 4} = 1.73.$$

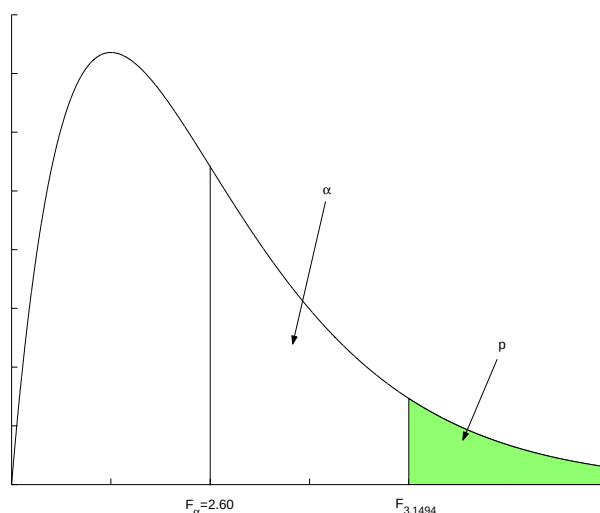
Il test statistico F è allora

$$F_{4-1, 73+105+240+1080-4} = F_{3, 1494} = \frac{s_B^2}{s_W^2} = \frac{19.06}{1.73} = 11.02.$$

Il valore critico F per 3 e 1494 gradi di libertà in modo da avere una coda $\alpha = 0.05$ è (dalla Tabella A.5)

$$F_{\alpha} = 2.60.$$

Pertanto la situazione è la seguente:



Essendo il valore del test p inferiore ad α , rifiutiamo l'ipotesi nulla e concludiamo che i livelli medi di colesterolo sierico sono differenti nelle 4 popolazioni.

4. Dette μ_1, μ_2, μ_3 le ore medie (sconosciute) di sonno nelle 3 popolazioni, l'ipotesi nulla è

$$H_0 : \mu = \mu_2 = \mu_3.$$

La media del campione relativo al placebo è (in ore di sonno)

$$\bar{x}_1 = 4.98,$$

mentre la varianza campionaria è

$$s_1^2 = 0.49.$$

La media del campione relativo al farmaco standard è

$$\bar{x}_2 = 7.95,$$

mentre la varianza campionaria è

$$s_2^2 = 0.45.$$

La media del campione relativo al nuovo farmaco è

$$\bar{x}_3 = 7.76,$$

mentre la varianza campionaria è

$$s_3^2 = 2.35.$$

La *media globale* è allora

$$\bar{x} = \frac{4.98 + 7.95 + 7.76}{3} = 6.9$$

ore di sonno.

La *varianza tra gruppi* è

$$s_B^2 = 6 \cdot \frac{(4.98 - 6.9)^2 + (7.95 - 6.9)^2 + (7.76 - 6.9)^2}{3 - 1} = 16.58.$$

La *varianza entro gruppi* è invece

$$s_W^2 = 5 \cdot \frac{0.49 + 0.45 + 2.35}{6 + 6 + 6 - 3} = 1.10.$$

Il test statistico F pertanto è

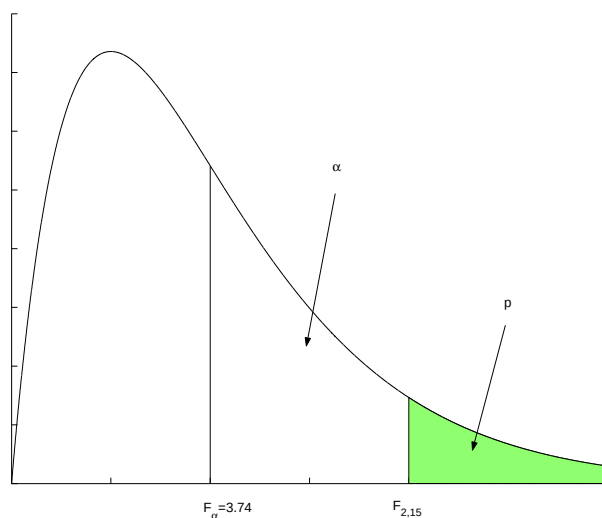
$$F_{3-1, 6+6+6-3} = F_{2, 15} = \frac{s_B^2}{s_W^2} = \frac{16.58}{1.10} = 15.04.$$

Il valore critico di F per 2 e 15 gradi di libertà in modo da avere una coda $\alpha = 0.05$ è (dalla Tabella A.5)

$$F_\alpha = 3.74.$$

Dunque la situazione è la seguente:

Università di Torino



Concludiamo allora che il valore p del test è inferiore ad α : $p < \alpha$. Pertanto possiamo rifiutare l'ipotesi nulla: vi è differenza nell'efficacia dei 3 farmaci testati sulle ore medie di sonno.

5. Dette $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ i contenuti medi (sconosciuti) di cadmio nei 5 suoli differenti, l'ipotesi nulla è

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5.$$

Il campione relativo al primo terreno ha una media

$$\bar{x}_1 = 0.64$$

ed una varianza campionaria

$$s_1^2 = 0.00516.$$

Il campione relativo al secondo terreno ha una media

$$\bar{x}_2 = 0.525$$

ed una varianza campionaria

$$s_2^2 = 0.00439.$$

Il campione relativo al terzo terreno ha una media

$$\bar{x}_3 = 0.308$$

ed una varianza campionaria

$$s_3^2 = 0.00189.$$

Il campione relativo al quarto terreno ha una media

$$\bar{x}_4 = 0.193$$

ed una varianza campionaria

$$s_4^2 = 0.00494.$$

Il campione relativo al quinto terreno ha una media

$$\bar{x}_5 = 0.335$$

ed una varianza campionaria

$$s_5^2 = 0.00023.$$

La *media globale* allora è

$$\bar{x} = \frac{0.64 + 0.525 + 0.308 + 0.193 + 0.335}{5} = 0.4.$$

La *varianza tra gruppi* è

$$s_B^2 = 6 \cdot \frac{(0.64 - 0.4)^2 + (0.525 - 0.4)^2 + (0.308 - 0.4)^2 + (0.193 - 0.4)^2 + (0.335 - 0.4)^2}{5 - 1} = 0.192845.$$

La *varianza entro gruppi* è

$$s_W^2 = 4 \cdot \frac{0.00516 + 0.00439 + 0.00189 + 0.00494 + 0.00023}{6 + 6 + 6 + 6 + 6 - 5} = 0.003325.$$

Il test statistico F pertanto è

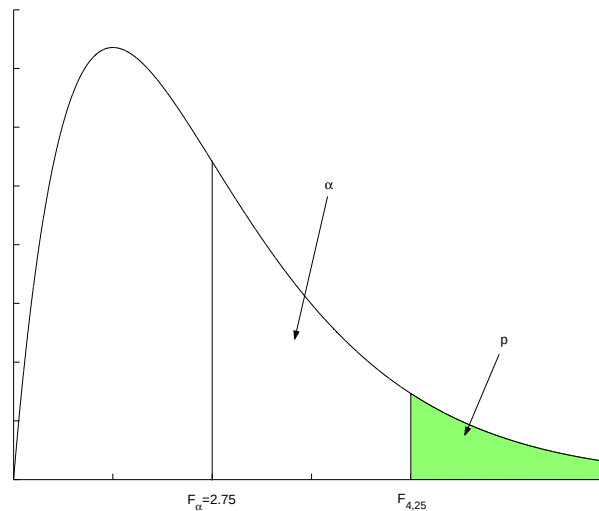
$$F_{5-1, 6+6+6+6+6-5} = F_{4, 25} = \frac{s_B^2}{s_W^2} = \frac{0.192845}{0.003325} = 58.$$

Il valore critico F per 4 e 25 gradi di libertà in modo da avere una coda $\alpha = 0.05$ è (dalla Tabella A.5)

$$F_\alpha = 2.75.$$

Pertanto la situazione è la seguente:

Università di Torino



Quindi possiamo concludere che il valore p del test è inferiore ad α e dunque possiamo rifiutare l'ipotesi nulla: vi è differenza nel contenuto medio di cadmio nei 5 tipi di terreni esaminati.

6. Siano $\mu_1, \mu_2, \mu_3, \mu_4$ le medie (sconosciute) delle pressioni sistoliche delle 4 popolazioni. L'ipotesi nulla allora è

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

La *media globale* è

$$\bar{x} = \frac{269 \cdot 115 + 53 \cdot 114 + 28 \cdot 118 + 9 \cdot 126}{269 + 53 + 28 + 9} = 115.36 \text{ mmHg.}$$

La *varianza tra gruppi* è

$$s_B^2 = \frac{269 \cdot (11 - 115.36)^2 + 53 \cdot (114 - 115.36)^2 + 28 \cdot (118 - 115.36)^2 + 9 \cdot (126 - 115.36)^2}{4 - 1} = 448.975.$$

La *varianza entro gruppi* è

$$s_W^2 = \frac{268 \cdot 13.4^2 + 52 \cdot 10.1^2 + 27 \cdot 11.6^2 + 8 \cdot 12.2^2}{269 + 53 + 28 + 9 - 4} = 164.08.$$

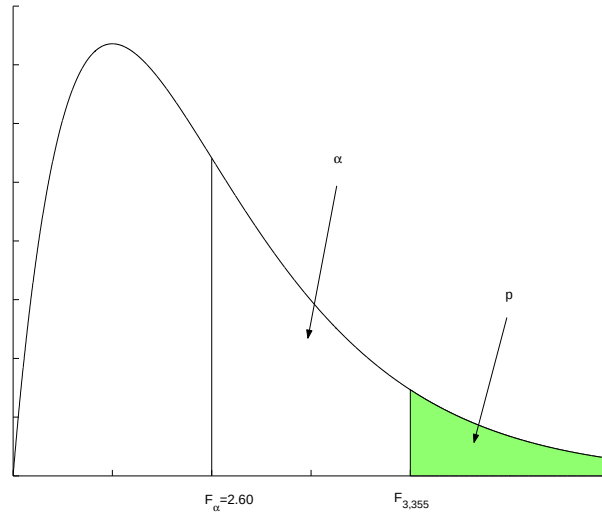
Il test statistico F è allora

$$F_{4-1, 269+53+28+9-4} = F_{3, 355} = \frac{s_B^2}{s_W^2} = \frac{448.975}{164.08} = 2.73.$$

Il valore critico del test F per 3 e 355 gradi di libertà in modo da avere una coda $\alpha = 0.05$ è (dalla Tabella A.5)

$$F_{\alpha} = 2.60.$$

Pertanto la situazione è la seguente:



Vediamo quindi che il valore p del test è inferiore ad α e possiamo rifiutare l'ipotesi nulla: la pressioni sistoliche medie nelle 4 popolazioni sono differenti.

Capitolo 7

Inferenza sulle proporzioni

7.1 Esercizi

1. Supponete di selezionare dalla popolazione dei neonati del Messico un campione casuale di 40 bambini. La probabilità che un bambino di questa popolazione pesi non più di 2500 grammi è 0.15.
 - (a) Per il campione di dimensione 40, qual è la probabilità che al massimo 4 bambini pesino non più di 2500 grammi? Calcolare l'esatta probabilità.
 - (b) Calcolare la stessa probabilità del punto (a), utilizzando l'approssimazione normale.
2. Un'inserzione pubblicitaria dichiara che più del 60% dei medici preferisce una particolare marca di analgesico. Un'agenzia che vuole verificare la validità di tale affermazione conduce un'indagine consistente di un campione casuale di 120 medici. Dei 120 intervistati, 82 dicono di preferire quella marca. E' giustificata l'affermazione dell'inserzione pubblicitaria?
3. E' stato condotto uno studio per esaminare la relazione tra fumo materno durante la gravidanza e presenza di malformazioni congenite nel bambino. Tra i bambini che soffrono di anomalie diverse dalla sindrome di Down o dalla palatoschisi, il 32.8% sono figli di madri che hanno fumato durante la gravidanza. Questa proporzione è omogenea per i bambini con i diversi tipi di difetti.
 - (a) Se selezionate da questa popolazione ripetuti campioni di dimen-

sione 25, che cosa si può dire della distribuzione campionaria delle proporzioni? Elencare 3 proprietà.

- (b) Tra i campioni di dimensione 25, quale frazione ha una proporzione campionaria maggiore o uguale a 0.45?
- (c) Quale frazione ha una proporzione campionaria maggiore o uguale a 0.20?
- (d) Quale valore p delimita il 10% inferiore della distribuzione?

7.2 Soluzioni

1. (a) Abbiamo $n = 40$ e $p = 0.15$. Quindi

$$\begin{aligned}
 P(X \leq 4) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + \\
 &\quad + P(X = 4) = \\
 &= \binom{40}{0} (0.15)^0 (1 - 0.15)^{40} + \binom{40}{1} (0.15)^1 (1 - 0.15)^{39} + \\
 &\quad + \binom{40}{2} (0.15)^2 (1 - 0.15)^{38} + \binom{40}{3} (0.15)^3 (1 - 0.15)^{37} + \\
 &\quad + \binom{40}{4} (0.15)^4 (1 - 0.15)^{36} = \\
 &= 0.26332.
 \end{aligned}$$

- (b) Poichè

$$\begin{aligned}
 np &= 40 \cdot 0.15 = 6 > 5 \\
 n(1 - p) &= 40 \cdot 0.85 = 34 > 5
 \end{aligned}$$

possiamo usare l'approssimazione normale alla binomiale:

$$P(X \leq 4) = P\left(Z \leq \frac{4 - 40 \cdot 0.15}{\sqrt{40 \cdot 0.15 \cdot 0.85}}\right) = P(Z \leq -0.8856) = 0.18792.$$

Usiamo la correzione per la continuità (con +0.5 perchè $4 < np = 6$):

$$P(X \leq 4) = P\left(Z \leq \frac{4 - 40 \cdot 0.15 + 0.5}{\sqrt{40 \cdot 0.15 \cdot 0.85}}\right) = P(Z \leq -0.66) = 0.255.$$

2. Poichè $n = 120$ e $x = 82$ avremo che

$$\hat{p} = \frac{82}{120} = 0.68.$$

Eseguiamo un test d'ipotesi sulla proporzione p ad un livello di significatività $\alpha = 0.05$.

L'ipotesi nulla è

$$H_0 : p = 0.60$$

mentre l'ipotesi alternativa è

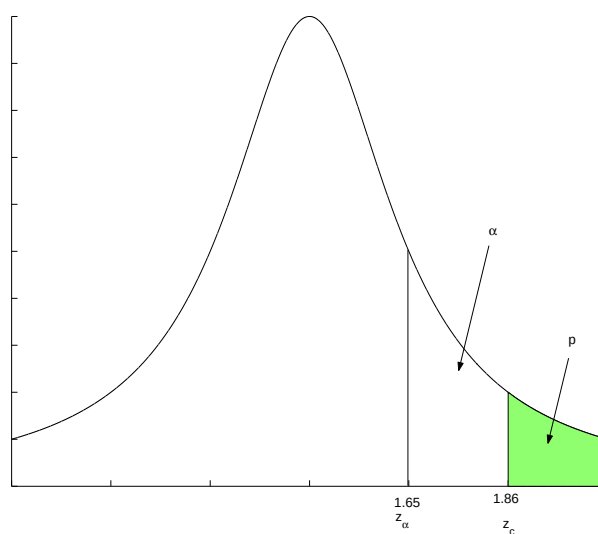
$$H_A : p > 0.60.$$

Quindi eseguiamo un test *unilaterale*.

Abbiamo:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.68 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{120}}} = \frac{0.08}{0.044} = 1.86.$$

Consultando la Tabella A.3 si trova che $z_\alpha = 1.65$. Abbiamo allora la seguente situazione:



Pertanto concludiamo che il valore p del test è inferiore ad α : $p < \alpha$. Quindi rifiutiamo l'ipotesi nulla e concludiamo che l'affermazione dell'inserzione pubblicitaria è corretta.

3. (b) Abbiamo $n = 25$ e $p = 0.328$. Quindi

$$\begin{aligned} P(\hat{p} \geq 0.45) &= P\left(Z \geq \frac{0.45 - 0.328}{\sqrt{\frac{0.328(1-0.328)}{25}}}\right) = P(Z \geq 1.2993) = \\ &= 0.0979 = 9.79\%. \end{aligned}$$

- (c) In questo caso avremo

$$\begin{aligned} P(\hat{p} \geq 0.20) &= P\left(Z \geq \frac{0.20 - 0.328}{\sqrt{\frac{0.328(1-0.328)}{25}}}\right) = P\left(Z \geq \frac{-0.128}{0.094}\right) = \\ &= P(Z \geq -1.36) = 1 - P(Z \leq -1.36) = \\ &= 1 - P(Z \geq 1.36) = 1 - 0.087 = 0.913 = 91.3\%. \end{aligned}$$

- (d) Noi dobbiamo avere che

$$P(\hat{p} \leq p) = 0.1.$$

Allora

$$P\left(Z \leq \frac{p - 0.328}{\sqrt{\frac{0.328(1-0.328)}{25}}}\right) = 0.1.$$

Quindi, consultando la Tabella A.3, si trova che

$$z = \frac{p - 0.328}{0.094} = -1.28$$

da cui

$$p = 0.094 \cdot (-1.28) + 0.328 = 0.20 = 20\%.$$

Capitolo 8

Tabelle di contingenza

8.1 Esercizi

1. Un candidato ad una carica pubblica vuole stabilire se c'è differenza nella sua popolarità tra gli uomini e le donne. Per stabilire l'esistenza di tale differenza, egli conduce un'indagine campionaria tra i votanti. Il campione contiene 250 uomini e 250 donne, di cui il 42% degli uomini ed il 51% delle donne sono favorevoli alla sua candidatura. Questi valori indicano una differenza nella popolarità? Costruire una tabella di contingenza.
2. I sondaggi di opinione forniscono spesso informazioni su come le opinioni di diversi gruppi varino quando si trattano temi controversi. Un campione casuale di 102 elettori statunitensi registrati alle liste elettorali è stato preso in esame da una determinata autorità garante. A ciascun elettore sono state rivolte 2 domande:
 - (a) Qual è il tuo partito?
 - (b) Sei favorevole all'aumento delle spese destinate agli armamenti?

I risultati del sondaggio sono sintetizzati nella tabella sottostante. Verificare l'ipotesi nulla che l'opinione individuale sugli aumenti di spese militari sia indipendente dal partito politico.

	PARTITO			
OPINIONE	Democratico	Repubblicano	Nessuno	Totale
Favorevole	16	21	11	48
Contrario	24	17	13	54
Totale	40	38	24	102

3. I seguenti dati sono di uno studio che esamina i problemi legati al consumo di bevande alcoliche tra gli studenti universitari. Nel 1983, ad un gruppo di studenti fu chiesto se avevano mai guidato in stato di ubriachezza. Nel 1987, dopo l'aumento dell'età minima consentita dalla legge per il consumo di alcolici, ad un altro gruppo di studenti universitari è stata posta la stessa domanda.

	Anno		
Guida in stato di ubriachezza	1983	1987	Totale
Sì	1250	991	2241
No	1387	1666	3053
Totale	2637	2657	5294

Testare l'ipotesi nulla che non esista alcuna associazione tra guida in stato di ubriachezza ed anno.

4. La seguente tabella illustra i dati di 6 studi che esaminano l'accuratezza dei certificati di morte. I risultati di 5373 autopsie sono stati confrontati con le cause di morte riportate nei certificati. Di questi, 3726 erano accurati, 783 carenti di informazioni o inesatti anche se non era necessaria la riformulazione della causa di morte e 864 erano inesatti ed era necessario riformulare la causa di morte.

	Certificato di morte			
Anno dello studio	Accurato	Carente.Nessuna riformulazione	Inesatto.Riformulazione	Totale
1955-1965	2040	367	327	2734
1970	149	60	48	257
1970-1971	288	25	70	383
1975-1977	703	197	252	1152
1977-1978	425	62	88	575
1980	121	72	79	272
Totale	3726	783	864	5373

Pensate che i risultati siano omogenei tra i vari studi?

5. In uno studio sui fattori di rischio per cancro invasivo della cervice condotto in Germania, sono stati raccolti i seguenti dati che mettono in relazione l'abitudine al fumo e la presenza o meno di cancro della cervice.

	Fumatore	Non fumatore	Totale
Cancro	108	117	225
Non cancro	163	268	431
Totale	271	385	656

Testare l'ipotesi nulla che non esiste alcuna associazione tra l'abitudine al fumo e la presenza di cancro della cervice ad un livello di significatività $\alpha = 0.05$.

6. Un articolo di giornale ha fornito le frequenze assolute dei furti a mano armata e dei furti d'auto, relativamente a 3 aree cittadine confinanti. In base ai dati raccolti nella tabella sottostante, è possibile ipotizzare che la tipologia del crimine sia diversa nelle 3 aree cittadine esaminate?

	TIPO DI CRIMINE		
CITTA'	FURTI D'AUTO	RAPINE	Totale
A	175	54	229
B	97	11	108
C	80	10	90
Totale	352	75	427

8.2 Soluzioni

1. Cominciamo a scrivere la tabella delle frequenze *osservate*:

	Favorevoli	Non favorevoli	Totale
Uomini	105 (=42% di 250)	145	250
Donne	128(=51% di 250)	122	250
Totale	233	267	500

Scriviamo ora la corrispondente tabella delle frequenze *attese*:

	Favorevoli	Non favorevoli	Totale
Uomini	$\frac{233 \cdot 250}{500} = 116.5$	$\frac{267 \cdot 250}{500} = 133.5$	250
Donne	$\frac{233 \cdot 250}{500} = 116.5$	$\frac{267 \cdot 250}{500} = 133.5$	250
Totale	250	250	500

Osserviamo che nessuna delle frequenze attese è inferiore a 1 e sono tutte superiori a 5.

Essendo una tabella di contingenza 2x2, vi è 1 grado di libertà. Pertanto, calcolando il test del chi quadro, teniamo conto anche del *fattore di correzione di Yates*:

$$\begin{aligned}\chi^2 &= \frac{(|105 - 116.5| - 0.5)^2}{116.5} + \frac{(|128 - 116.5| - 0.5)^2}{116.5} + \frac{(|145 - 133.5| - 0.5)^2}{133.5} + \frac{(|122 - 133.5| - 0.5)^2}{133.5} = \\ &= 1.04 + 0.91 + 1.04 + 0.91 = \\ &= 3.9.\end{aligned}$$

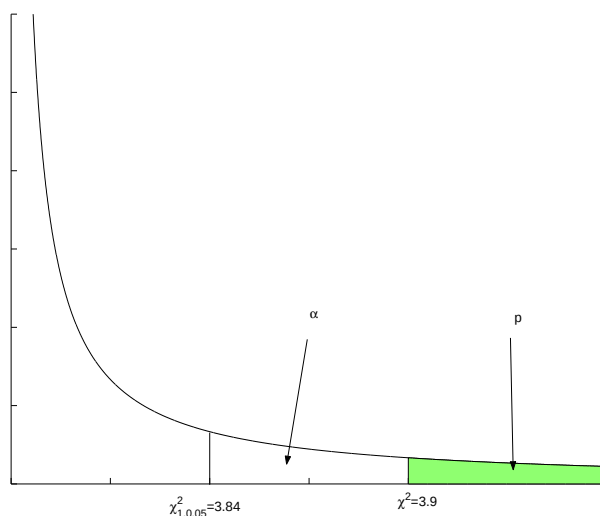
L'ipotesi nulla H_0 che stiamo testando è

H_0 : non esiste alcuna differenza tra uomini e donne nella popolarità del candidato

Fissiamo un livello di significatività $\alpha = 0.05$. Dalla Tabella A.8 del testo risulta che il valore critico del chi quadro per 1 grado di libertà e $\alpha = 0.05$ è

$$\chi^2_{1,0.05} = 3.84.$$

La situazione pertanto è la seguente:



Vediamo allora che il valore p del test è inferiore ad α : $p < \alpha$. Pertanto rifiutiamo l'ipotesi nulla e concludiamo che c'è differenza tra uomini e donne nella popolarità del candidato.

2. Cominciamo a ricopiare la tabella delle frequenze *osservate*:

	PARTITO			
OPINIONE	Democratico	Repubblicano	Nessuno	Totale
Favorevole	16	21	11	48
Contrario	24	17	13	54
Totale	40	38	24	102

Costruiamo ora la tabella delle frequenze *attese*:

	PARTITO			
OPINIONE	Democratico	Repubblicano	Nessuno	Totale
Favorevole	$\frac{40 \cdot 48}{102} = 18.82$	$\frac{38 \cdot 48}{102} = 17.88$	$\frac{24 \cdot 48}{102} = 11.29$	48
Contrario	$\frac{40 \cdot 54}{102} = 21.18$	$\frac{38 \cdot 54}{102} = 20.12$	$\frac{24 \cdot 54}{102} = 12.71$	54
Totale	40	38	24	102

Osserviamo che tutte le frequenze attese sono superiori a 5. I gradi di libertà sono $(2 - 1) \cdot (3 - 1) = 2$. Calcoliamo il test del chi quadro:

$$\begin{aligned}\chi^2 &= \frac{(16 - 18.82)^2}{18.82} + \frac{(24 - 21.18)^2}{21.18} + \frac{(21 - 17.88)^2}{17.88} + \frac{(17 - 20.12)^2}{20.12} + \frac{(11 - 11.29)^2}{11.29} + \frac{(13 - 12.71)^2}{12.71} = \\ &= 0.42 + 0.38 + 0.54 + 0.48 + 0.01 + 0.01 = \\ &= 1.84.\end{aligned}$$

Stiamo testando lipotesi nulla

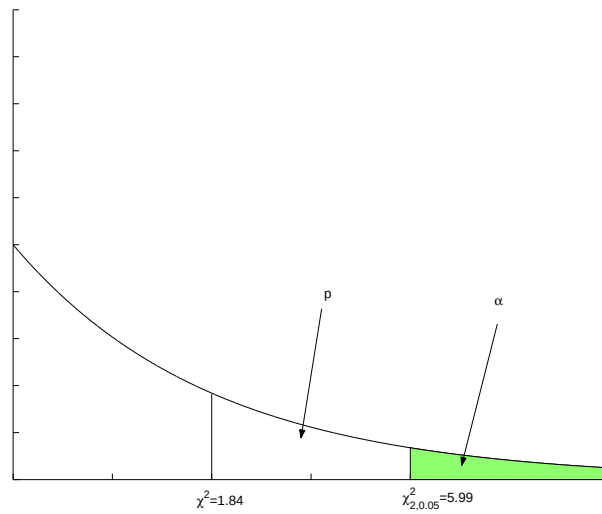
H_0 : Non esiste alcun legame opinione politica e opinione sulle spese militari.

Fissiamo un livello di significatività $\alpha = 0.05$. Dalla Tabella A.8 ricaviamo che il valore critico del chi quadro per 2 gradi di libertà e $\alpha = 0.05$ è

$$\chi_{2,0.05}^2 = 5.99.$$

La situazione dunque è la seguente:

Quaderni Didattici del Dipartimento di Matematica



Pertanto concludiamo che il valore p del test è *superiore* ad α . Non possiamo quindi rifiutare l'ipotesi nulla e dobbiamo concludere che, con i dati a nostra disposizione, non possiamo affermare che esista un legame tra opinione politica e opinione sulle spese militari.

3. La tabella delle frequenze *osservate* è la seguente:

	Anno		
Guida in stato di ubriachezza	1983	1987	Totale
Sì	1250	991	2241
No	1387	1666	3053
Totale	2637	2657	5294

La corrispondente tabella delle frequenze *attese* allora è

Università di Torino

	Anno		
Guida in stato di ubriachezza	1983	1987	Totale
Sì	$\frac{2637 \cdot 2241}{5294} = 1116.27$	$\frac{2657 \cdot 2241}{5294} = 1124.73$	2241
No	$\frac{2637 \cdot 3053}{5294} = 1520.73$	$\frac{2657 \cdot 3053}{5294} = 1532.27$	3053
Totale	2637	2657	5294

Tutte le frequenze attese sono maggiori di 5. E' una tabella di contingenza 2x2, quindi vi è un grado di libertà. Calcoliamo il test del chi quadro con il fattore di correzione di Yates:

$$\begin{aligned}
 \chi^2 &= \frac{(|1250 - 1116.27| - 0.5)^2}{1116.27} + \frac{(|991 - 1124.73| - 0.5)^2}{1124.73} + \frac{(|1387 - 1520.73| - 0.5)^2}{1520.73} + \\
 &\quad + \frac{(|1666 - 1532.27| - 0.5)^2}{1532.27} = \\
 &= 15.90 + 15.78 + 11.67 + 11.58 = \\
 &= 54.93.
 \end{aligned}$$

L'ipotesi nulla da testare è

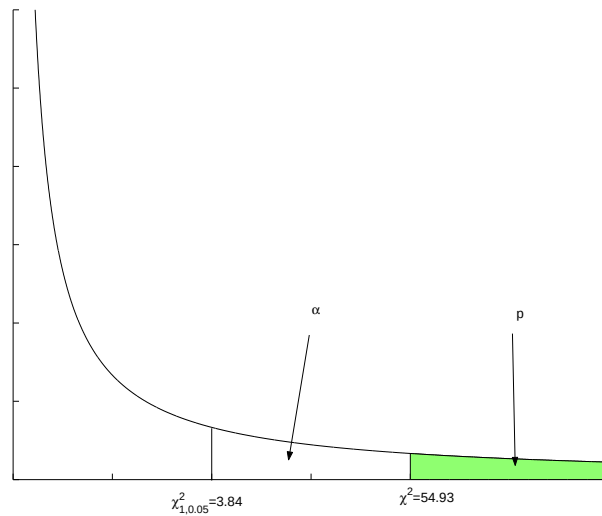
H_0 : Non esiste alcuna relazione tra guida in stato di ubriachezza ed anno.

Fissiamo il livello di significatività $\alpha = 0.05$. Allora dalla Tabella A.8 si riscontra che il valore critico del chi quadro con 1 grado di libertà è

$$\chi_{1,0.05}^2 = 3.84.$$

La situazione pertanto è la seguente:

Quaderni Didattici del Dipartimento di Matematica



Pertanto concludiamo che il valore p del test è *inferiore* ad α . Posso dunque rifiutare l'ipotesi nulla e dire che esiste un legame tra guida in stato di ubriachezza ed anno.

4. La tabella delle frequenze *osservate* è la seguente:

	Certificato di morte			
Anno dello studio	Accurato	Carente.Nessuna riformulazione	Inesatto.Riformulazione	Totale
1955-1965	2040	367	327	2734
1970	149	60	48	257
1970-1971	288	25	70	383
1975-1977	703	197	252	1152
1977-1978	425	62	88	575
1980	121	72	79	272
Totale	3726	783	864	5373

La tabella delle frequenze *attese* allora è la seguente:

Università di Torino

	Certificato di morte			
Anno dello studio	Accurato	Carente.Nessuna riformulazione	Inesatto.Riformulazione	Totale
1955-1965	1895.94	398.42	439.64	2374
1970	178.22	37.45	41.33	257
1970-1971	265.60	55.81	61.59	383
1975-1977	798.87	167.88	185.25	1152
1977-1978	398.74	83.79	92.46	575
1980	188.62	39.64	43.74	272
Totale	3726	783	864	5373

Tutte le frequenze attese sono dunque maggiori di 5. I gradi di libertà sono $(6 - 1) \cdot (3 - 1) = 10$. L'ipotesi nulla da testare è

H_0 : Non esiste alcuna relazione tra anno dello studio e precisione del certificato di decesso.

Calcoliamo il test del chi quadro:

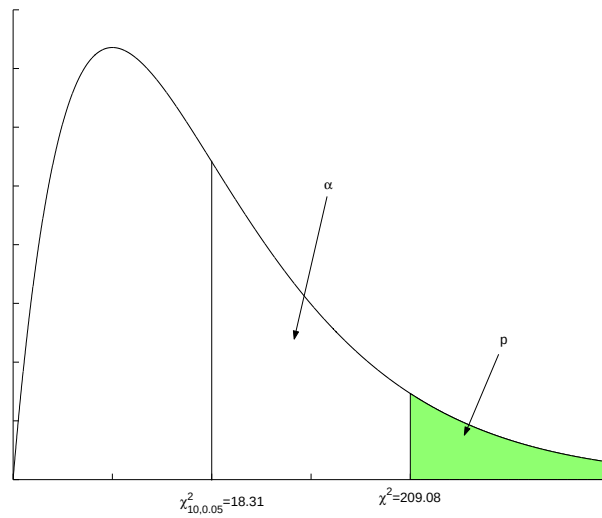
$$\begin{aligned}\chi^2 &= 10.95 + 2.48 + 28.86 + 4.79 + 13.58 + 1.08 + 1.89 + 17.01 + 1.15 + 11.50 + 5.05 + \\ &\quad + 24.05 + 1.73 + 5.67 + 0.21 + 24.24 + 26.42 + 28.42 = \\ &= 209.08.\end{aligned}$$

Fissiamo il livello di significatività $\alpha = 0.05$. Dalla Tabella A.8 troviamo che il valore critico del chi quadro per $\alpha = 0.05$ e 10 gradi di libertà è

$$\chi_{10,0.05}^2 = 18.31.$$

Dunque la situazione è la seguente:

Quaderni Didattici del Dipartimento di Matematica



Concludiamo che il valore p del test è *inferiore* ad α . Rifiutiamo l'ipotesi nulla e diciamo che esiste un legame tra anno dello studio e precisione del certificato di morte: i risultati non sono omogenei.

5. Riportiamo la tabella delle frequenze *osservate*:

	Fumatore	Non fumatore	Totale
Cancro	108	117	225
Non cancro	163	268	431
Totale	271	385	656

Costruiamo la tabella delle frequenze *attese*:

	Fumatore	Non fumatore	Totale
Cancro	92.95	132.05	225
Non cancro	178.05	252.95	431
Totale	271	385	656

Tutte le frequenze attese sono maggiori di 5. E' una tabella 2x2, quindi

Università di Torino

abbiamo 1 solo grado di libertà. Calcoliamo il chi quadro con il fattore di correzione di Yates:

$$\begin{aligned}\chi^2 &= \frac{(|108 - 92.95| - 0.5)^2}{92.95} + \frac{(|163 - 178.05| - 0.5)^2}{178.05} + \frac{(|117 - 132.05| - 0.5)^2}{132.05} + \\ &+ \frac{(|268 - 252.95| - 0.5)^2}{252.95} = \\ &= 2.28 + 1.19 + 1.60 + 0.84 = \\ &= 5.91.\end{aligned}$$

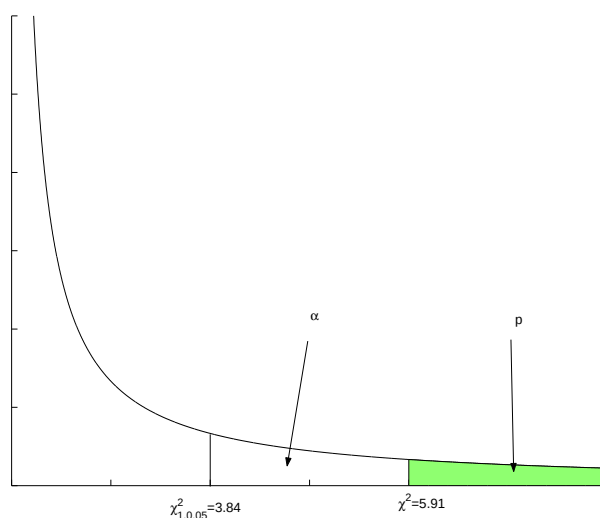
L'ipotesi nulla da testare è

H_0 : Non esiste relazione tra abitudine al fumo e cancro della cervice.

Dalla Tabella A.8 si trova che il valore critico del chi quadro con $\alpha = 0.05$ e 1 grado di libertà è

$$\chi_{1,0.05}^2 = 3.84.$$

Pertanto abbiamo la seguente situazione:



Essendo $p < \alpha$, rifiutiamo l'ipotesi nulla e concludiamo che esiste una relazione tra abitudine al fumo e cancro della cervice.

6. Riportiamo la tabella delle frequenze *osservate*:

Quaderni Didattici del Dipartimento di Matematica

	TIPO DI CRIMINE		
CITTA'	FURTI D'AUTO	RAPINE	Totale
A	175	54	229
B	97	11	108
C	80	10	90
Totale	352	75	427

La corrispondente tabella delle frequenze *attese* allora è

	TIPO DI CRIMINE		
CITTA'	FURTI D'AUTO	RAPINE	Totale
A	188.78	40.22	229
B	89.03	18.97	108
C	74.19	15.81	90
Totale	352	75	427

Osserviamo che tutte le frequenze attese sono maggiori di 5. I gradi di libertà sono $(3 - 1) \cdot (2 - 1) = 2$. Calcoliamo il test del chi quadro:

$$\begin{aligned}
 \chi^2 &= \frac{(175 - 188.78)^2}{188.78} + \frac{(54 - 40.22)^2}{40.22} + \frac{(97 - 89.03)^2}{89.03} + \frac{(11 - 18.97)^2}{18.97} + \\
 &+ \frac{(80 - 74.19)^2}{74.19} + \frac{(10 - 15.81)^2}{15.81} = \\
 &= 1.01 + 0.71 + 0.45 + 4.72 + 3.35 + 2.14 = \\
 &= 12.38.
 \end{aligned}$$

L'ipotesi nulla da testare è

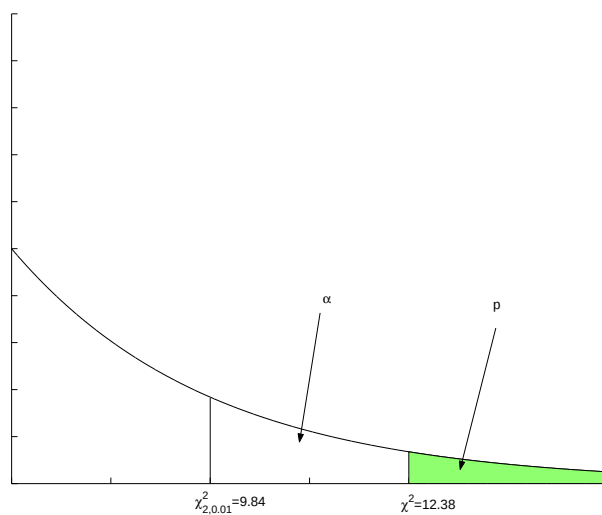
H_0 : Non esiste alcun legame tra tipologia del crimine commesso e città in cui è stato commesso.

Fissiamo il livello di significatività $\alpha = 0.01$. Nella Tabella A.8 troviamo che il valore critico del chi quadro con 2 gradi di libertà e $\alpha = 0.01$ è

$$\chi_{2,0.01}^2 = 9.21.$$

La situazione allora è la seguente:

Università di Torino



Pertanto constatiamo che il valore p del test è *inferiore* ad α : rifiutiamo l'ipotesi nulla e concludiamo che esiste un legame tra città e tipo di crimine commesso.

Capitolo 9

Correlazione

9.1 Esercizi

1. In uno studio condotto in Italia, 10 pazienti affetti da ipertrigliceridemia sono stati sottoposti ad una dieta povera di grassi e ricca di carboidrati. Prima della dieta, sono state registrate per ciascun soggetto le misurazioni di colesterolo e di trigliceridi.

Paziente	Livello di colesterolo (μ mol/l)	Livello di trigliceridi (μ mol/l)
1	5.12	2.30
2	6.18	2.54
3	6.77	2.95
4	6.65	3.77
5	6.36	4.18
6	5.90	5.31
7	5.48	5.53
8	6.02	8.83
9	10.34	9.48
10	8.51	14.20

- (a) Calcolare il coefficiente di correlazione di Pearson.
 - (b) Ad un livello di significatività $\alpha = 0.05$, testare l'ipotesi nulla che la correlazione ρ sia uguale a 0.
2. Supponete di essere interessati a determinare l'esistenza di una correlazione tra il contenuto di fluoro nell'acqua potabile e la carie dentaria in bambini che bevono questa acqua. I dati in uno studio che esamina 7257 bambini in 11 città sono riportati nella tabella sottostante. Il

contenuto di fluoro nelle acque in ciascuna città è misurato in parti per milione (ppm); il numero di carie dentarie per 100 bambini è ottenuto sommando il numero di denti otturati, denti cariati, denti da estrarre e denti estratti per carie.

- (a) Qual è la correlazione tra il numero di carie dentarie per 100 bambini ed il contenuto di fluoro nell'acqua?
- (b) Ad un livello di significatività $\alpha = 0.05$ testare l'ipotesi nulla che la correlazione ρ sia uguale a 0.

Fluoro	1.9	2.6	1.8	1.2	1.2	1.2	1.3	0.9	0.6	0.5	0.4
Carie	236	246	252	258	281	303	323	343	412	444	556

3. Nove pazienti con cardiopatia ischemica hanno partecipato ad una serie di test per valutare la percezione del dolore. In una parte dello studio, i pazienti hanno eseguito degli esercizi fino all'insorgenza dell'angina; è stato registrato il tempo necessario per l'insorgenza dell'angina e la durata dell'attacco. I dati sono riportati nella seguente tabella:

Tempo di insorgenza dell'angina (sec)	105	118	130	150	178	180	192	200	201
Durata dell'angina (sec)	640	670	560	510	212	260	228	335	460

- (a) Calcolare il coefficiente di correlazione di Pearson.
 - (b) Ad un livello di significatività $\alpha = 0.05$, testare l'ipotesi nulla che la correlazione ρ sia uguale a 0.
4. La seguente tabella riporta i voti conseguiti alla scuola medica nel corso di fisiopatologia e all'esame nazionale nel corso di patologia per 10 studenti.
 - (a) Indicare se vi è o meno evidenza di una relazione lineare tra i 2 punteggi.
 - (b) Ad un livello di significatività $\alpha = 0.05$, testare l'ipotesi nulla che la correlazione ρ sia uguale a 0.

Punteggio corso di fisiopatologia	Punteggio esame nazionale di patologia
1.7	86
1.7	83
2.3	88
3.3	87
2.0	78
4.0	94
3.0	87
3.0	85
2.7	86
2.3	88

9.2 Soluzioni

1. (a) Poste X la variabile casuale relativa alle misurazioni di colesterolo e Y la variabile casuale relativa alle misurazioni di trigliceridi, abbiamo che la medie campionarie sono:

$$\bar{x} = \frac{5.12 + 6.18 + 6.77 + 6.65 + 6.36 + 5.90 + 5.48 + 6.02 + 10.34 + 8.51}{10} = 6.73$$

e

$$\bar{y} = \frac{2.30 + 2.54 + 2.95 + 3.77 + 4.18 + 5.31 + 5.53 + 8.83 + 9.48 + 14.20}{10} = 5.91.$$

Allora:

$x_i - \bar{x}$	$y_i - \bar{y}$
-1.61	-3.61
-0.55	-3.37
0.04	-2.96
-0.08	-2.14
-0.37	-1.73
-0.83	-0.60
-1.25	-0.38
-0.71	2.92
3.61	3.57
1.78	8.29

Pertanto:

$$\begin{aligned}\sum_{i=1}^{10} (x_i - \bar{x})^2 &= 21.99 \\ \sum_{i=1}^{10} (y_i - \bar{y})^2 &= 131.44 \\ \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) &= 34.90.\end{aligned}$$

Possiamo calcolare il coefficiente di correlazione:

$$r = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{10} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{10} (y_i - \bar{y})^2}} = \frac{34.90}{\sqrt{21.99 \cdot 131.44}} = 0.65.$$

(b) L'ipotesi nulla è

$$H_0 : \rho = 0$$

mentre l'ipotesi alternativa è

$$H_A : \rho \neq 0.$$

Calcoliamo il test t con $10 - 2 = 8$ gradi di libertà:

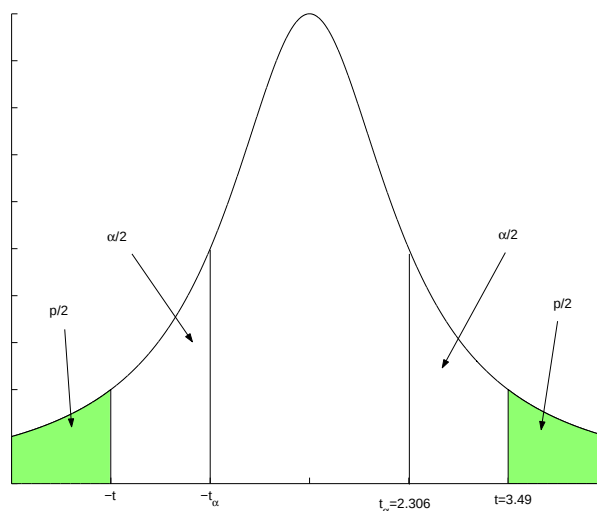
$$t_8 = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = r \sqrt{\frac{n-2}{1-r^2}} = 0.65 \cdot \sqrt{\frac{8}{1-0.65^2}} = 3.49.$$

Analizzando la Tabella A.4 osserviamo che il valore critico di t per un test bilaterale con $\alpha = 0.05$ e 8 gradi di libertà è

$$t_{8,0.05} = 2.306.$$

Pertanto abbiamo la seguente situazione:

Università di Torino



Osserviamo dunque che il valore p del test è *inferiore* ad α : $p < \alpha$. Quindi possiamo rifiutare l'ipotesi nulla e concludere che esiste una relazione *lineare* tra le due variabili casuali.

2. (a) Dette X la variabile casuale relativa al fluoro e Y quella relativa alla carie, abbiamo che le medie campionarie sono:

$$\bar{x} = 1.24$$

e

$$\bar{y} = 332.18.$$

Allora

$$\sum_{i=1}^{11} (x_i - \bar{x})^2 = 4.29$$

$$\sum_{i=1}^{11} (y_i - \bar{y})^2 = 101264.92$$

$$\sum_{i=1}^{11} (x_i - \bar{x})(y_i - \bar{y}) = -558.7.$$

Il coefficiente di correlazione è

$$r = \frac{-558.7}{\sqrt{4.29 \cdot 101264.92}} = -0.85.$$

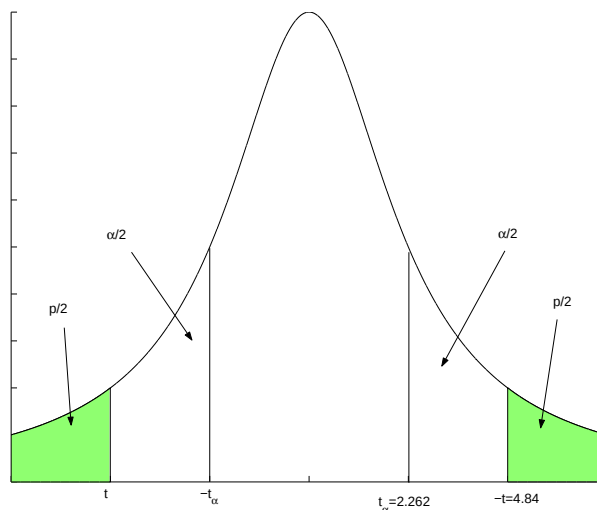
(b) Calcoliamo il test t con $11 - 2 = 9$ gradi di libertà:

$$t_9 = r \cdot \sqrt{\frac{n-2}{1-r^2}} = -0.85 \cdot \sqrt{\frac{9}{1-0.85^2}} = -4.84.$$

Con la Tabella A.4 calcoliamo il valore critico di t per un test bilaterale con $\alpha = 0.05$ e 9 gradi di libertà:

$$t_{0.05,9} = 2.262.$$

La situazione dunque è la seguente:



Vediamo allora che il valore p del test è inferiore ad α : $p < \alpha$. Dunque possiamo rifiutare l'ipotesi nulla e concludere che esiste una relazione lineare tra le 2 variabili.

3. (a) Dette X la variabile casuale relativa al tempo di insorgenza dell'angina e Y quella relativa alla durata dell'angina, abbiamo che le medie campionarie sono:

$$\bar{x} = 161.55$$

e

$$\bar{y} = 430.55$$

Allora:

$$\begin{aligned}\sum_{i=1}^9 (x_i - \bar{x})^2 &= 10796.22 \\ \sum_{i=1}^9 (y_i - \bar{y})^2 &= 252150.22 \\ \sum_{i=1}^9 (x_i - \bar{x})(y_i - \bar{y}) &= -42695.80\end{aligned}$$

Il coefficiente di correlazione è

$$r = \frac{-42695.80}{\sqrt{10796.22 \cdot 252150.22}} = -0.81.$$

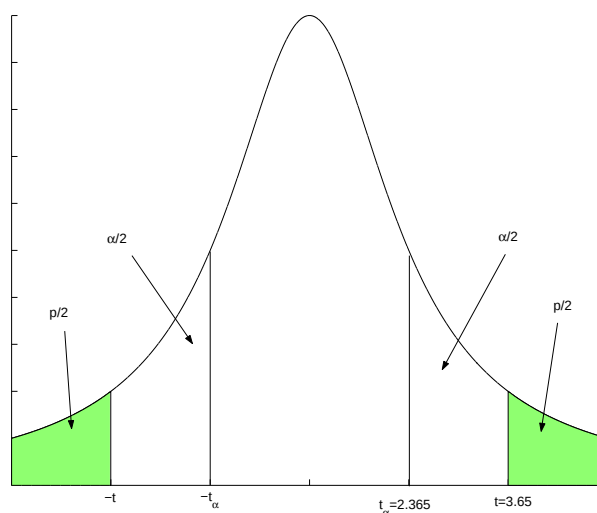
(b) Calcoliamo il test t con $9 - 2 = 7$ gradi di libertà:

$$t_7 = r \sqrt{\frac{n-2}{1-r^2}} = -0.81 \cdot \sqrt{\frac{9-2}{1-0.81^2}} = -3.65.$$

Il valore critico di t per un test bilaterale con $\alpha = 0.05$ e 7 gradi di libertà è

$$t_{0.05,7} = 2.365.$$

La situazione è la seguente:



Pertanto $p < \alpha$: possiamo rifiutare l'ipotesi nulla e concludere che esiste una relazione lineare tra le due variabili casuali.

4. (a) Dette X la variabile casuale relativa al punteggio del corso di fisiopatologia e Y quella relativa al punteggio conseguito all'esame di patologia, le medie campionarie sono:

$$\bar{x} = 2.6$$

e

$$86.2$$

Allora

$$\begin{aligned}\sum_{i=1}^{10} (x_i - \bar{x})^2 &= 4.94 \\ \sum_{i=1}^{10} (y_i - \bar{y})^2 &= 147.6 \\ \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) &= 18.2.\end{aligned}$$

Il coefficiente di correlazione è

$$r = \frac{18.2}{\sqrt{4.94 \cdot 147.6}} = 0.67.$$

- (b) Calcoliamo il test t con $10 - 2 = 8$ gradi di libertà:

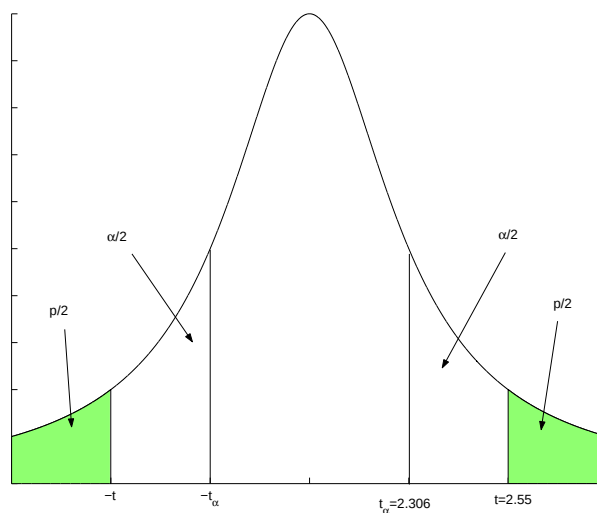
$$t_8 = r \sqrt{\frac{n-2}{1-r^2}} = 0.67 \cdot \sqrt{\frac{10-2}{1-0.67^2}} = 2.55.$$

Il valore critico di t con $\alpha = 0.05$ per un test bilaterale con 8 gradi di libertà è

$$t_{0.05,8} = 2.306.$$

La situazione dunque è la seguente:

Università di Torino



Osserviamo che il valore p del test è inferiore ad α : $p < \alpha$. Quindi possiamo rifiutare l'ipotesi nulla e concludere che esiste una relazione lineare tra le due variabili casuali del problema.

Capitolo 10

Regressione

10.1 Esercizi

1. La tabella seguente riporta i dati relativi ad un campione di 10 neonati con basso peso alla nascita.

età gestazionale	29	31	33	31	30	25	27	29	28	29
pressione sistolica	43	51	42	39	48	31	31	40	57	64

- (a) Utilizzando la pressione sistolica come variabile di risposta, calcolare la retta di regressione dei minimi quadrati.
 - (b) Supponete di selezionare casualmente dalla popolazione un nuovo neonato e di rilevare che la sua età gestazionale è di 31 settimane. Qual è la pressione sistolica predetta?
2. Le misurazioni della lunghezza e del peso di un campione di 20 neonati con basso peso alla nascita sono riportate nella tabella seguente.

lunghezza	41	40	38	38	38	32	33	38	30	34	32	39	38	39	37	39	38	42	39	38
peso	1360	1490	1490	1180	1200	680	620	1060	1320	830	880	1130	1140	1350	950	1220	980	1480	1250	1250

- (a) Disegnare un grafico del peso in funzione della lunghezza per i 20 neonati.
- (b) Calcolare la retta di regressione dei minimi quadrati.

Dal diagramma dei punti è chiaro che un punto si trova al di fuori del range dei dati. Questo punto corrisponde al nono neonato del campione. Per illustrare l'effetto che il valore atipico ha sul modello, eliminare questo punto dai dati.

- (a) Calcolare la nuova retta di regressione dei minimi quadrati.

(b) Confrontare i coefficienti di determinazione delle due rette.

3. Nella tabella sottostante sono riportati i tassi di letalità dei minatori per gli anni 1970-1981. Per semplificare i calcoli gli anni sono stati convertiti in una scala che parte da 1.

anno	1	2	3	4	5	6	7	8	9	10	11	12
tasso di letalità	0.2419	0.1732	0.1361	0.1108	0.0996	0.0952	0.0904	0.0792	0.0701	0.0890	0.0799	0.1084

- (a) Disegnare un diagramma a punti del tasso di letalità in funzione dell'anno.
- (b) Per modellizzare l'andamento del tasso di letalità, adattare la retta di regressione dei minimi quadrati $Y = a + bX$ dove X rappresenta l'anno. Utilizzando il coefficiente di determinazione ed un grafico dei residui, commentare l'adattabilità del modello.
- (c) Trasformare X in $\ln X$. Disegnare un diagramma a punti del tasso di letalità in funzione del logaritmo naturale dell'anno.
- (d) Adattare il modello $Y = a + b \ln X$. Utilizzando ancora il coefficiente di determinazione ed un grafico dei residui, confrontare l'adattabilità di questo modello al modello del punto (b).
- (e) Trasformare X in $\frac{1}{X}$ ed adattare il modello $Y = a + b \frac{1}{X}$. Commentare l'adattabilità di questo modello e confrontarlo con i modelli precedenti.
4. Un tossicologo è interessato all'effetto di una certa sostanza tossica sull'incidenza di tumore in animali da laboratorio. Ad un campione di animali vengono somministrate diverse concentrazioni della sostanza e, successivamente, si esamina su di essi l'eventuale presenza o assenza di tumore. La variabile esplicativa è la concentrazione di sostanza tossica. I risultati sono costituiti dalla percentuale di quelli che hanno il tumore e sono riportati nella seguente tabella:

concentrazione	0.0	2.1	5.4	8.0	15.0	19.5
percentuale	0.04000	0.09259	0.10870	0.19608	0.80000	0.80769

Usando la trasformazione logistica, determinare la retta di regressione dei minimi quadrati, facendo un'inferenza sulla pendenza.

5. Uno degli argomenti di interesse in uno studio di differenti regimi nel trattamento del diabete era la perdita di peso durante la terapia. I seguenti dati si riferiscono al problema riguardante l'eventuale associazione tra l'ammontare di perdita di peso ed il peso iniziale. La

seguente tabella riporta per 16 pazienti adulti, trattati con fenformina, il peso basale (in libbre) all'inizio della terapia e la perdita di peso un anno dopo l'inizio della terapia stessa.

peso basale	225	235	173	223	200	199	129	242	140	156	146	195	155	185	150	149
perdita di peso	15	44	31	39	6	16	21	44	5	12	-3	19	10	24	-3	10

- (a) Si saggi l'ipotesi nulla che non vi è stato alcun cambiamento di peso dopo un anno.
- (b) Si calcolino la pendenza e l'intercetta della retta di regressione dei minimi quadrati per la perdita di peso ad un anno rispetto al peso basale, facendo un'inferenza sulla pendenza.

10.2 Soluzioni

1. (a) Detta X l'età gestazionale e Y la pressione sistolica, si trova

$$\bar{x} = 29.2, \quad \bar{y} = 44.6.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{10} (x_i - 29.2)(y_i - 44.6)}{\sum_{i=1}^{10} (x_i - 29.2)^2} = 1.40$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 44.6 - 1.40 \cdot 29.2 = 3.75.$$

Pertanto l'equazione della retta di regressione è

$$Y = 3.75 + 1.40X.$$

- (b) Avremo

$$Y_{\text{previsto}}(31) = 3.75 + 1.40 \cdot 31 = 47.15.$$

2. Posto X la lunghezza e Y il peso, si trova

$$\bar{x} = 37.15, \quad \bar{y} = 1143.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{20} (x_i - 37.15)(y_i - 1143)}{\sum_{i=1}^{20} (x_i - 37.15)^2} = 51.24$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1143 - 51.24 \cdot 37.15 = -760.52.$$

La retta di regressione quindi ha equazione

$$Y = -760.52 + 51.24X.$$

Togliendo il nono neonato, invece, troviamo

$$\bar{x} = 37.53, \quad \bar{y} = 1133.68.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{19} (x_i - 37.53)(y_i - 1133.68)}{\sum_{i=1}^{19} (x_i - 37.53)^2} = 79.89$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1133.68 - 79.89 \cdot 37.53 = -1864.27.$$

La nuova retta di regressione quindi è

$$Y = -1864.27 + 79.89X.$$

Il coefficiente di determinazione nel primo caso è

$$R_1^2 = \frac{(\sum_{i=1}^{20} (x_i - 37.15)(y_i - 1143))^2}{\sum_{i=1}^{20} (x_i - 37.15)^2 \sum_{i=1}^{20} (y_i - 1143)^2} = 0.42,$$

mentre nel secondo caso è

$$R_2^2 = \frac{(\sum_{i=1}^{19} (x_i - 37.53)(y_i - 1133.68))^2}{\sum_{i=1}^{19} (x_i - 37.53)^2 \sum_{i=1}^{19} (y_i - 1133.68)^2} = 0.76.$$

3. (b) Detto X l'anno e Y il tasso di letalità, abbiamo

$$\bar{x} = 6.5, \quad \bar{y} = 0.114.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{12} (x_i - 6.5)(y_i - 0.114)}{\sum_{i=1}^{12} (x_i - 6.5)^2} = -0.01$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 0.114 + 0.01 \cdot 6.5 = 0.181.$$

Quindi l'equazione della retta di regressione è

$$Y = 0.181 - 0.10X.$$

Il coefficiente di determinazione è

$$R^2 = \frac{(\sum_{i=1}^{12} (x_i - 6.5)(y_i - 0.114))^2}{\sum_{i=1}^{12} (x_i - 6.5)^2 \sum_{i=1}^{12} (y_i - 0.114)^2} = 0.56.$$

Vi è quindi una buona adattabilità del modello.

(c) Ecco la tabella in scala logaritmica:

$\ln(\mathbf{X})$	$\mathbf{Y}$
Anno	Tasso di letalità
0	0.2419
0.693	0.1732
1.099	0.1361
1.386	0.1108
1.609	0.0996
1.792	0.0952
1.946	0.0904
2.079	0.0792
2.197	0.0701
2.303	0.089
2.398	0.0799
2.485	0.1084

(d) Posto $\tilde{X} = \ln(X)$ l'anno e Y il tasso di letalità, abbiamo

$$\bar{\tilde{x}} = 1.67, \quad \bar{y} = 0.11.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{12} (x_i - 1.67)(y_i - 0.11)}{\sum_{i=1}^{12} (x_i - 1.67)^2} = -0.059$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{\tilde{x}} = 0.11 + 0.059 \cdot 1.67 = 0.214.$$

L'equazione della retta di regressione è ora

$$Y = 0.214 - 0.059\tilde{X}.$$

Il coefficiente di determinazione è

$$R^2 = \frac{(\sum_{i=1}^{12} (x_i - 1.67)(y_i - 0.11))^2}{\sum_{i=1}^{12} (x_i - 1.67)^2 \sum_{i=1}^{12} (y_i - 0.11)^2} = 0.84.$$

L'adattabilità del modello è dunque migliorata rispetto a prima.

(e) Ecco la nuova tabella dei valori:

Quaderni Didattici del Dipartimento di Matematica

1/X	Y
Anno	Tasso di letalità
1	0.2419
0.5	0.1732
0.33	0.1361
0.25	0.1108
0.2	0.0996
0.17	0.0952
0.14	0.0904
0.125	0.0792
0.11	0.0701
0.1	0.089
0.09	0.0799
0.08	0.1084

Posto $\tilde{X} = 1/X$ l'anno e Y il tasso di letalità, abbiamo

$$\bar{\tilde{x}} = 0.26, \quad \bar{y} = 0.11.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{12} (x_i - 0.26)(y_i - 0.11)}{\sum_{i=1}^{12} (x_i - 0.26)^2} = 0.1808$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{\tilde{x}} = 0.11 - 0.1808 \cdot 0.26 = 0.0677.$$

L'equazione della retta di regressione è ora

$$Y = 0.0677 + 0.1808\tilde{X}.$$

Il coefficiente di determinazione è

$$R^2 = \frac{(\sum_{i=1}^{12} (x_i - 0.26)(y_i - 0.11))^2}{\sum_{i=1}^{12} (x_i - 0.26)^2 \sum_{i=1}^{12} (y_i - 0.11)^2} = 0.94.$$

L'adattabilità del modello è dunque ancora migliore rispetto ai 2 modelli precedenti.

4. Ecco la tabella secondo la trasformazione logistica:

Università di Torino

$\ln(X)$	Y
Percentuale	Concentrazione
-3.219	0
-2.380	2.1
-2.219	5.4
-1.629	8
-0.223	15
-0.214	19.5

Posto $\tilde{X} = \ln(X)$ la percentuale e Y la concentrazione, abbiamo

$$\bar{\tilde{x}} = -1.65, \quad \bar{y} = 8.33.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^6 (x_i + 1.65)(y_i - 8.33)}{\sum_{i=1}^6 (x_i + 1.65)^2} = 6.04$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{\tilde{x}} = 8.33 + 6.04 \cdot 1.65 = 18.29.$$

L'equazione della retta di regressione è ora

$$Y = 18.29 + 6.04\tilde{X}.$$

Il coefficiente di determinazione è

$$R^2 = \frac{(\sum_{i=1}^6 (x_i + 1.65)(y_i - 8.33))^2}{\sum_{i=1}^6 (x_i + 1.65)^2 \sum_{i=1}^6 (y_i - 8.33)^2} = 0.95.$$

Facciamo ora un'inferenza sulla pendenza, con ipotesi nulla

$$H_0 : \beta = 0.$$

Eseguiamo il test statistico

$$t_4 = \frac{\hat{\beta}}{\hat{es}(\hat{\beta})},$$

dove

$$\hat{es}(\hat{\beta}) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^6 (x_i + 1.65)^2}}$$

e

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^6 (y_i - \hat{y}_i)^2}{4}}$$

con $\hat{y}_i = 18.29 + 6.04\tilde{x}_i$ valori adattati.

Si trova

$$\hat{es}(\hat{\beta}) = 0.71.$$

Allora

$$t_4 = \frac{6.04}{0.71} = 8.5.$$

Fissiamo un livello di significatività $\alpha = 0.05$. Dalla Tabella A.4 troviamo che il valore critico di T con 4 gradi di libertà per un test bilaterale ad un livello di significatività $\alpha = 0.05$ è

$$t_c = 2.776.$$

Concludiamo quindi che il valore p del test è minore di α : $p < \alpha$. Quindi possiamo rifiutare l'ipotesi nulla H_0 .

5. (a) Eseguiamo un test T per 2 campioni appaiati.

La media delle differenze è

$$\bar{d} = \frac{15 + 44 + 31 + 39 + 6 + 16 + 21 + 44 + 5 + 12 - 3 + 19 + 10 + 24 - 3 + 10}{16} = 18.13.$$

La deviazione standard delle differenze è

$$s_d = \sqrt{\frac{\sum_{i=1}^{16} (d_i - 18.13)^2}{15}} = 15.01.$$

Il valore t del test allora è

$$t_{15} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{18.13}{15.01/\sqrt{16}} = 4.83.$$

Fissiamo un livello di significatività $\alpha = 0.05$. Dalla Tabella A.4 troviamo che il valore critico di T è

$$t_c = 2.131.$$

Concludiamo quindi che il valore p del test è inferiore ad α . Pertanto possiamo rifiutare l'ipotesi nulla che asseriva che non esistesse differenza di peso.

- (b) Detto X il peso basale e Y la perdita di peso dopo un anno, abbiamo

$$\bar{x} = 181.38, \quad \bar{y} = 18.13.$$

Pertanto

$$\hat{\beta} = \frac{\sum_{i=1}^{16}(x_i - 181 - 38)(y_i - 18.13)}{\sum_{i=1}^{16}(x_i - 181 - 38)^2} = 0.29$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 18.13 - 0.29 \cdot 181.38 = -33.85.$$

L'equazione della retta di regressione è

$$Y = -33.85 + 0.29X.$$

Il coefficiente di determinazione è

$$R^2 = \frac{(\sum_{i=1}^{16}(x_i - 181.38)(y_i - 18.13))^2}{\sum_{i=1}^{16}(x_i - 181.38)^2 \sum_{i=1}^{16}(y_i - 18.13)^2} = 0.49.$$

Testiamo ora l'ipotesi nulla

$$H_0 : \beta = 0.$$

Calcoliamo il test T con 14 gradi di libertà:

$$t_{14} = \frac{\hat{\beta}}{\hat{es}(\hat{\beta})}$$

dove

$$\hat{es}(\hat{\beta}) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{16}(x_i - 181.38)^2}}$$

e

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{16}(y_i - \hat{y}_i)^2}{14}}$$

con $\hat{y}_i = -33.85 + 0.29x_i$ valori adattati.

Si trova

$$\hat{es}(\hat{\beta}) = 0.078.$$

Allora

$$t_{14} = \frac{0.29}{0.078} = 3.71.$$

Fissiamo un livello di significatività $\alpha = 0.05$. Dalla Tabella A.4 troviamo che il valore critico di T per un test bilaterale con 14 gradi di libertà ad un livello di significatività $\alpha = 0.05$ è

$$t_c = 2.145.$$

Allora concludiamo che il valore p del test è inferiore ad α . Pertanto possiamo rifiutare l'ipotesi nulla H_0 , in accordo con quanto trovato al punto (a).

Bibliografia

1. P.Armitage, G. Berry, Statistica medica, McGraw-Hill
2. T. Colton, Statistica in medicina, Piccin Editore
3. M. Pagano, K. Gauvreau, Biostatistica, Idelson-Gnocchi

Appendice 1: Tabelle delle distribuzioni

Tabella A.3 Aree nella coda superiore della curva normale standardizzata

[illegible]

Tabella A.4 Percentili della distribuzione t

Area nella coda superiore						
gl	0,10	0,05	0,025	0,01	0,005	0,0005
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,599
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,768
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
50	1,299	1,676	2,009	2,403	2,678	3,496
60	1,296	1,671	2,000	2,390	2,660	3,460
70	1,294	1,667	1,994	2,381	2,648	3,435
80	1,292	1,664	1,990	2,374	2,639	3,416
90	1,291	1,662	1,987	2,368	2,632	3,402
100	1,290	1,660	1,984	2,364	2,626	3,390
110	1,289	1,659	1,982	2,361	2,621	3,381
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,327	2,576	3,291

Tabella A.5 Percentili della distribuzione F

Gradi di libertà (gl) per il denominatore	Area nella coda superiore	Gradi di libertà (gl) per il numeratore										
		1	2	3	4	5	6	7	8	12	24	∞
2	0,100	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,41	9,45	9,49
	0,050	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,41	19,45	19,50
	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,41	39,46	39,50
	0,010	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,42	99,46	99,50
	0,005	198,5	199,0	199,2	199,3	199,3	199,3	199,4	199,4	199,4	199,5	199,5
	0,001	998,5	999,0	999,2	999,3	999,3	999,3	999,4	999,4	999,4	999,5	999,5
3	0,100	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,22	5,18	5,13
	0,050	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,74	8,64	8,53
	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,34	14,12	13,90
	0,010	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,05	26,60	26,13
	0,005	55,55	49,80	47,47	46,19	45,39	44,84	44,43	44,13	43,39	42,62	41,83
	0,001	167,0	148,5	141,1	137,1	134,6	132,9	131,6	130,6	128,3	125,9	123,5
4	0,100	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,90	3,83	3,76
	0,050	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	5,91	5,77	5,63
	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,75	8,51	8,26
	0,010	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,37	13,93	13,46
	0,005	31,33	26,28	24,26	23,15	22,46	21,97	21,62	21,35	20,70	20,03	19,32
	0,001	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	47,41	45,77	44,05
5	0,100	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,27	3,19	3,10
	0,050	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,68	4,53	4,36
	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,52	6,28	6,02
	0,010	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	9,89	9,47	9,02
	0,005	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,38	12,78	12,14
	0,001	47,18	37,12	33,20	31,09	29,75	28,83	28,16	27,65	26,42	25,13	23,79
6	0,100	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,90	2,82	2,72
	0,050	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,00	3,84	3,67
	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,37	5,12	4,85
	0,010	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,72	7,31	6,88
	0,005	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,03	9,47	8,88
	0,001	35,51	27,00	23,70	21,92	20,80	20,03	19,46	19,03	17,99	16,90	15,75
7	0,100	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,67	2,58	2,47
	0,050	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,57	3,41	3,23
	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,67	4,41	4,14
	0,010	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,47	6,07	5,65
	0,005	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,18	7,64	7,08
	0,001	29,25	21,69	18,77	17,20	16,21	15,52	15,02	14,63	13,71	12,73	11,70
8	0,100	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,50	2,40	2,29
	0,050	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,28	3,12	2,93
	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,20	3,95	3,67
	0,010	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,67	5,28	4,86
	0,005	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,01	6,50	5,95
	0,001	25,41	18,49	15,83	14,39	13,48	12,86	12,40	12,05	11,19	10,30	9,33

(continua)

Tabella A.5 (continua)

Gradi di libertà (gl) per il denominatore	Area nella coda superiore	Gradi di libertà (gl) per il numeratore										
		1	2	3	4	5	6	7	8	12	24	∞
9	0,100	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,38	2,28	2,16
	0,050	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,07	2,90	2,71
	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	3,87	3,61	3,33
	0,010	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,11	4,73	4,31
	0,005	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,23	5,73	5,19
	0,001	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	9,57	8,72	7,81
10	0,100	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,28	2,18	2,06
	0,050	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	2,91	2,74	2,54
	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,62	3,37	3,08
	0,010	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,71	4,33	3,91
	0,005	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,66	5,17	4,64
	0,001	21,04	14,91	12,55	11,28	10,48	9,93	9,52	9,20	8,45	7,64	6,76
12	0,100	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,15	2,04	1,90
	0,050	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,69	2,51	2,30
	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,28	3,02	2,72
	0,010	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,16	3,78	3,36
	0,005	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	4,91	4,43	3,90
	0,001	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,00	6,25	5,42
14	0,100	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,05	1,94	1,80
	0,050	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,53	2,35	2,13
	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,05	2,79	2,49
	0,010	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	3,80	3,43	3,00
	0,005	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,43	3,96	3,44
	0,001	17,14	11,78	9,73	8,62	7,92	7,44	7,08	6,80	6,13	5,41	4,60
16	0,100	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	1,99	1,87	1,72
	0,050	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,42	2,24	2,01
	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	2,89	2,63	2,32
	0,010	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,55	3,18	2,75
	0,005	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,10	3,64	3,11
	0,001	16,12	10,97	9,01	7,94	7,27	6,80	6,46	6,19	5,55	4,85	4,06
18	0,100	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	1,93	1,81	1,66
	0,050	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,34	2,15	1,92
	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,77	2,50	2,19
	0,010	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,37	3,00	2,57
	0,005	10,22	7,21	6,03	5,37	4,96	4,66	4,44	4,28	3,86	3,40	2,87
	0,001	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,13	4,45	3,67
20	0,100	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,89	1,77	1,61
	0,050	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,28	2,08	1,84
	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,68	2,41	2,09
	0,010	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,23	2,86	2,42
	0,005	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,68	3,22	2,69
	0,001	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	4,82	4,15	3,38

(continua)

Tabella A.5 (continua)

Gradi di libertà (gl) per il denominatore	Area nella coda superiore	Gradi di libertà (gl) per il numeratore										
		1	2	3	4	5	6	7	8	12	24	∞
30	0,100	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,77	1,64	1,46
	0,050	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,09	1,89	1,62
	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,41	2,14	1,79
	0,010	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	2,84	2,47	2,01
	0,005	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,18	2,73	2,18
	0,001	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,00	3,36	2,59
40	0,100	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,71	1,57	1,38
	0,050	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,00	1,79	1,51
	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,29	2,01	1,64
	0,010	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,66	2,29	1,80
	0,005	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	2,95	2,50	1,93
	0,001	12,61	8,25	6,59	5,70	5,13	4,73	4,44	4,21	3,64	3,01	2,23
60	0,100	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,66	1,51	1,29
	0,050	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	1,92	1,70	1,39
	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,17	1,88	1,48
	0,010	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,50	2,12	1,60
	0,005	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	2,74	2,29	1,69
	0,001	11,97	7,77	6,17	5,31	4,76	4,37	4,09	3,86	3,32	2,69	1,89
80	0,100	2,77	2,37	2,15	2,02	1,92	1,85	1,79	1,75	1,63	1,48	1,24
	0,050	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	1,88	1,65	1,32
	0,025	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,35	2,11	1,82	1,40
	0,010	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,42	2,03	1,49
	0,005	8,33	5,67	4,61	4,03	3,65	3,39	3,19	3,03	2,64	2,19	1,56
	0,001	11,67	7,54	5,97	5,12	4,58	4,20	3,92	3,70	3,16	2,54	1,72
100	0,100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,61	1,46	1,21
	0,050	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,85	1,63	1,28
	0,025	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,08	1,78	1,35
	0,010	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,37	1,98	1,43
	0,005	8,24	5,59	4,54	3,96	3,59	3,33	3,13	2,97	2,58	2,13	1,49
	0,001	11,50	7,41	5,86	5,02	4,48	4,11	3,83	3,61	3,07	2,46	1,62
120	0,100	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,60	1,45	1,19
	0,050	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,83	1,61	1,25
	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,05	1,76	1,31
	0,010	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,34	1,95	1,38
	0,005	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,54	2,09	1,43
	0,001	11,38	7,32	5,78	4,95	4,42	4,04	3,77	3,55	3,02	2,40	1,54
∞	0,100	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,55	1,38	1,00
	0,050	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,75	1,52	1,00
	0,025	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	1,94	1,64	1,00
	0,010	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,18	1,79	1,00
	0,005	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,36	1,90	1,00
	0,001	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	2,74	2,13	1,00

Tabella A.6 Funzione della distribuzione T , il test di Wilcoxon dei ranghi con segno[illegible]

Tabella A.8 Percentili della distribuzione chi-quadrato

gl	Area nella coda superiore				
	0,100	0,050	0,025	0,010	0,001
1	2,71	3,84	5,02	6,63	10,83
2	4,61	5,99	7,38	9,21	13,82
3	6,25	7,81	9,35	11,34	16,27
4	7,78	9,49	11,14	13,28	18,47
5	9,24	11,07	12,83	15,09	20,52
6	10,64	12,59	14,45	16,81	22,46
7	12,02	14,07	16,01	18,48	24,32
8	13,36	15,51	17,53	20,09	26,12
9	14,68	16,92	19,02	21,67	27,88
10	15,99	18,31	20,48	23,21	29,59
11	17,28	19,68	21,92	24,72	31,26
12	18,55	21,03	23,34	26,22	32,91
13	19,81	22,36	24,74	27,69	34,53
14	21,06	23,68	26,12	29,14	36,12
15	22,31	25,00	27,49	30,58	37,70
16	23,54	26,30	28,85	32,00	39,25
17	24,77	27,59	30,19	33,41	40,79
18	25,99	28,87	31,53	34,81	42,31
19	27,20	30,14	32,85	36,19	43,82
20	28,41	31,41	34,17	37,57	45,31
21	29,62	32,67	35,48	38,93	46,80
22	30,81	33,92	36,78	40,29	48,27
23	32,01	35,17	38,08	41,64	49,73
24	33,20	36,42	39,36	42,98	51,18
25	34,38	37,65	40,65	44,31	52,62

Appendice 2: Figure

Tabella 1.1 Le 10 principali cause di morte negli Stati Uniti, 1988

Rango	Causa di morte	Numero totale di morti
1	Malattie cardiache	765.156
2	Neoplasie maligne	485.048
3	Malattie cerebrovascolari	150.517
4	Incidenti	97.100
5	Pneumopatie croniche ostruttive	82.853
6	Polmonite ed influenza	77.662
7	Diabete mellito	40.368
8	Suicidiq	30.407
9	Epatopatia cronica e cirrosi	26.409
10	Nefrite, sindrome nefrosica e nefrosi	22.392

Tabella 1.2 Frequenze assolute dei livelli di colesterolo sierico in 1.067 soggetti della popolazione maschile degli Stati Uniti di età compresa tra 25 e 34 anni, 1976-1980

Livello di colesterolo (mg/100 ml)	Numero di soggetti
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5
Totale	1.067

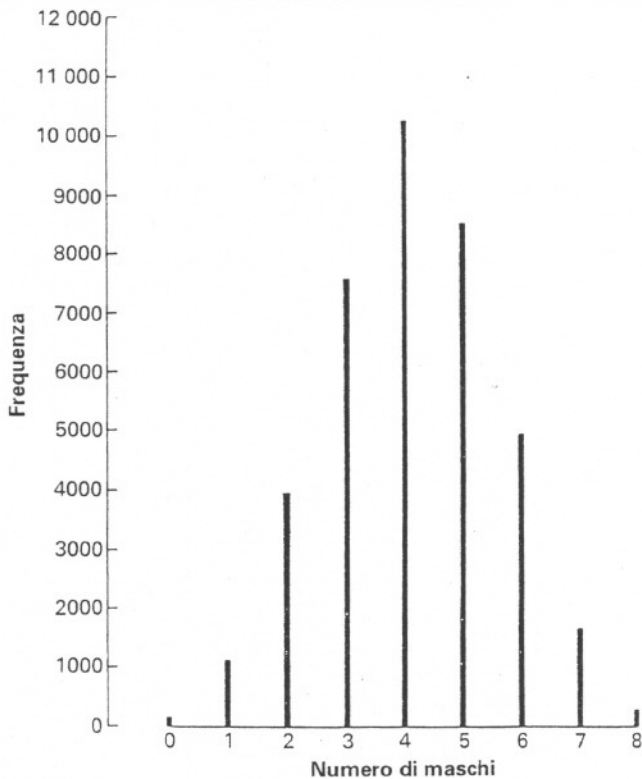
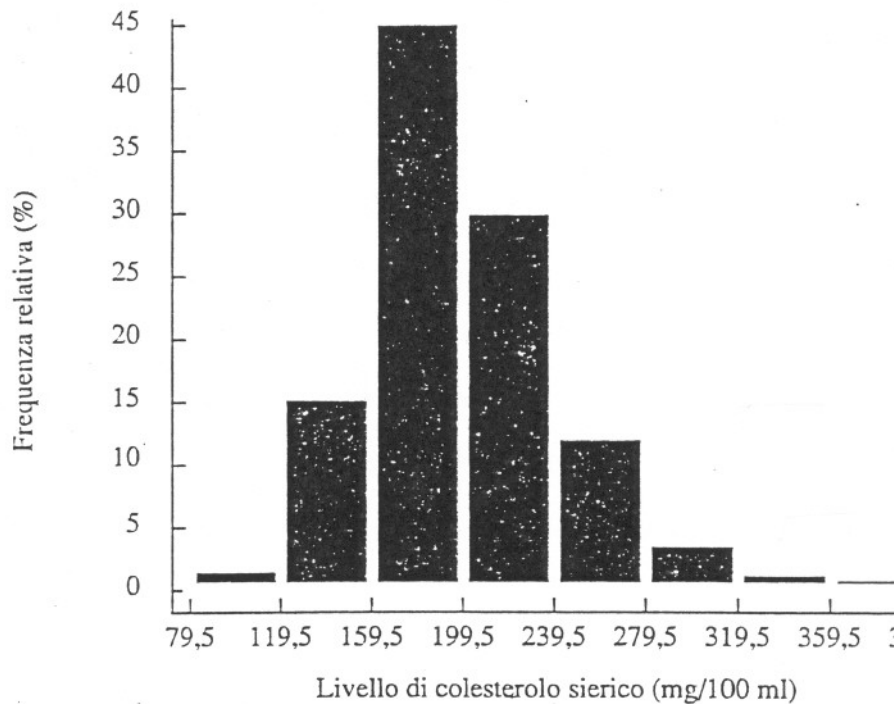


Fig. 1.1

Figura 1.2 Istogramma: frequenze relative dei livelli di colesterolo sierico in 1.067 soggetti della popolazione maschile degli Stati Uniti di età compresa tra 25 e 34 anni, 1976-1980



Numero di soggetti



Figura 1.3 Poligono di frequenza: frequenze assolute dei livelli di colesterolo sierico in 1.067 soggetti della popolazione maschile degli Stati Uniti di età compresa tra 25 e 34 anni, 1976-1980

○ Età 25 - 34
△ Età 55 - 64

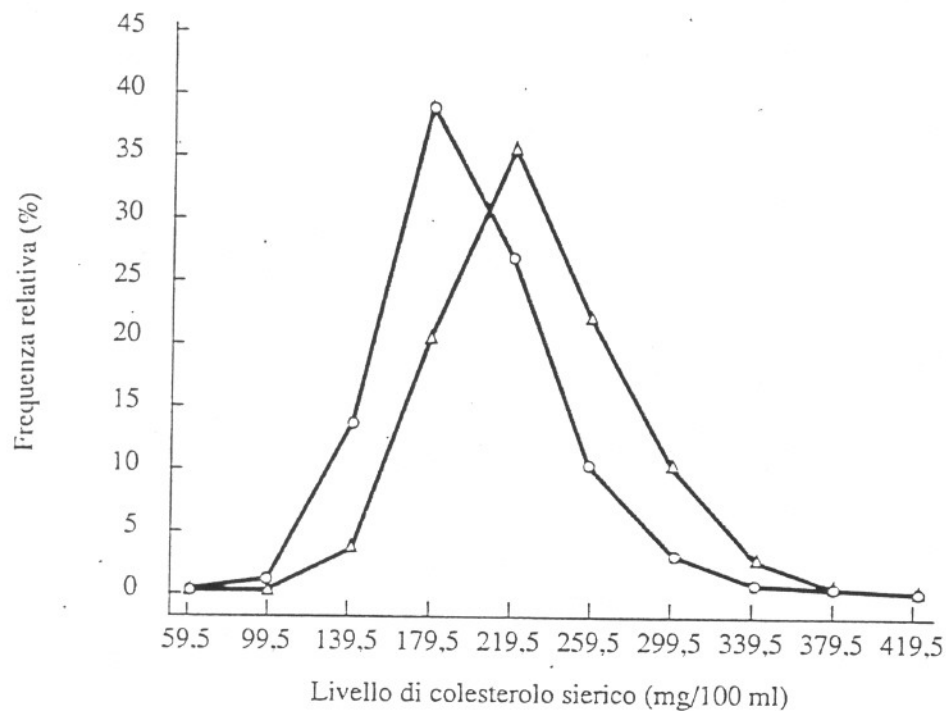


Figura 1.4 Poligono di frequenza: frequenze relative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

Figura 4.5 Poligono di frequenza cumulativa: frequenze relative cumulative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

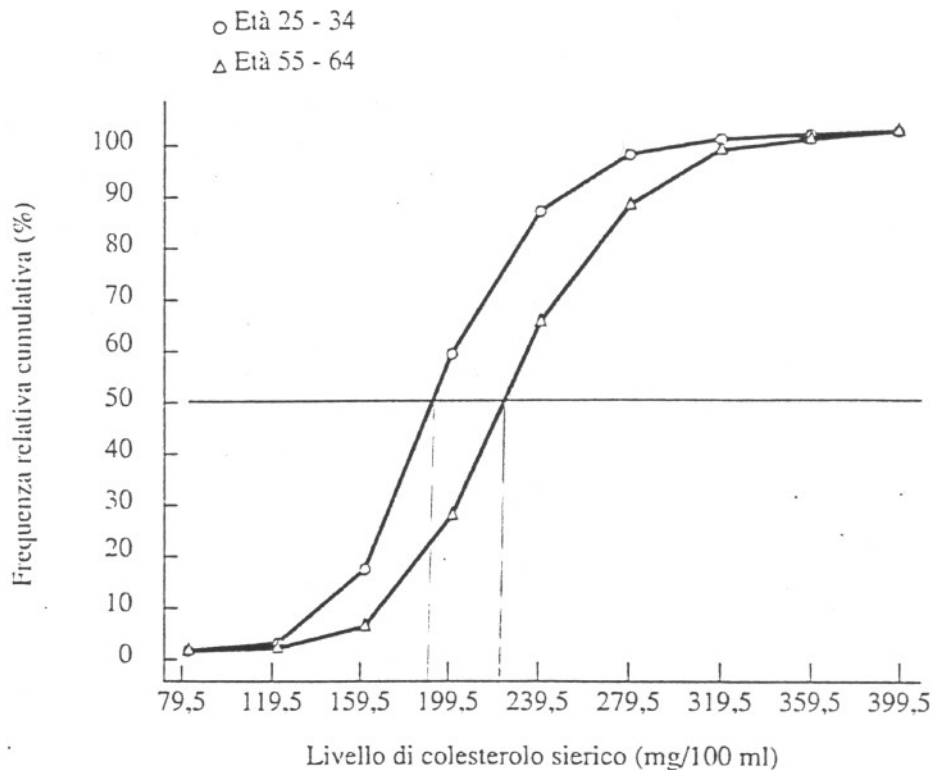


Figura 1.6 Diagramma a punti: capacità vitale forzata in funzione del volume espiratorio forzato in un secondo in 19 soggetti asmatici

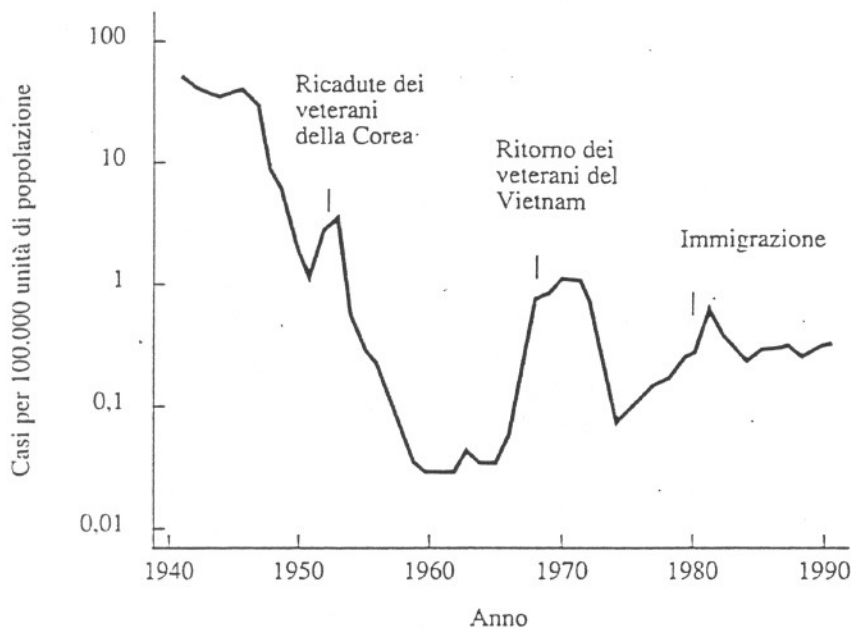
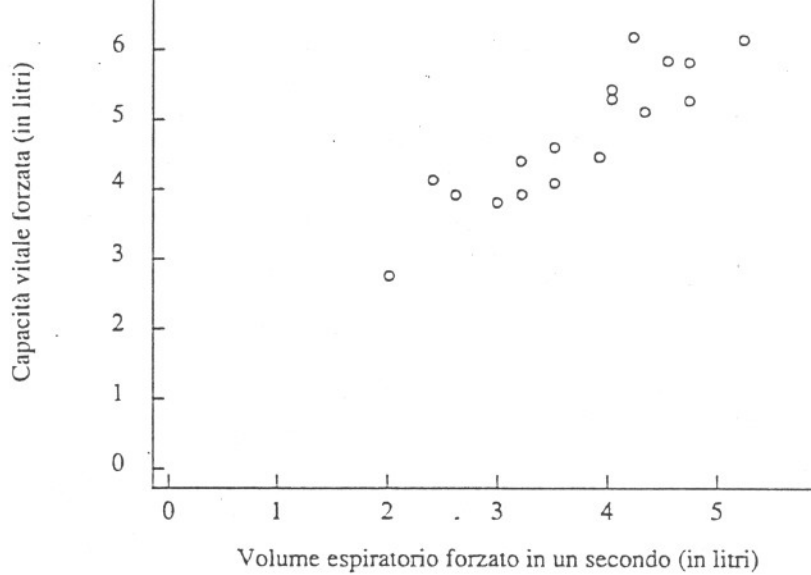
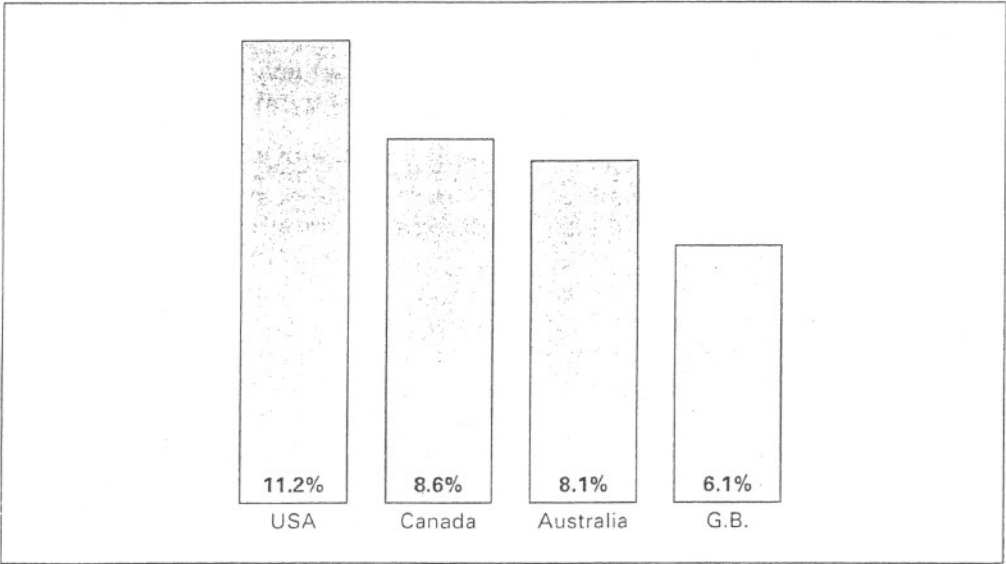
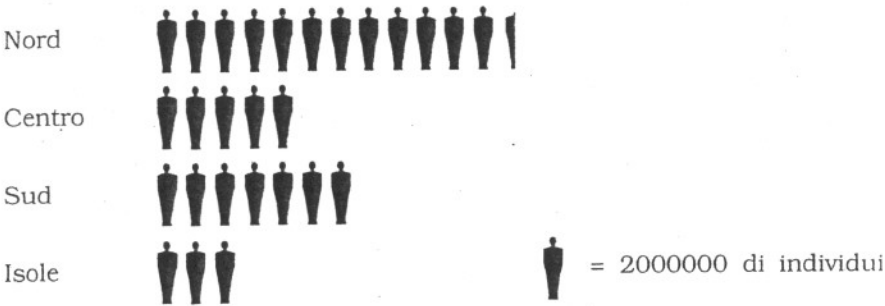


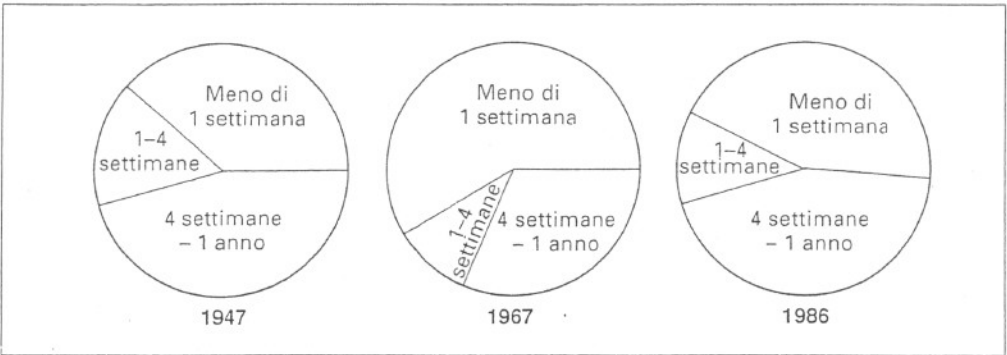
Figura 1.7 Diagramma lineare: tassi di malaria registrati per anno Stati Uniti, 1940-1989

Secondo il *Compendio Statistico* ISTAT 1988 la popolazione residente nel territorio italiano (al 1° Gennaio 1988) risulta così distribuita per gruppi di regioni (il *Nord* comprende Piemonte, Valle d'Aosta, Lombardia, Liguria, Trentino-Alto Adige, Friuli-Venezia Giulia, Emilia-Romagna; il *Centro* comprende Toscana, Umbria, Marche, Lazio; il *Sud* comprende Abruzzi, Molise, Campania, Puglia, Basilicata, Calabria; le *Isole* comprendono Sicilia e Sardegna):

<i>Nord</i>	25 518 866
<i>Centro</i>	10 952 361
<i>Sud</i>	14 135 320
<i>Isole</i>	6 792 561
<i>Totale</i>	57 399 108



“Diagramma a barre” che indica le percentuali di prodotto interno lordo devolute alle spese sanitarie in quattro paesi nel 1987 (riportate con il permesso di Macklin, 1990).



“Diagramma a torta” che indica, per tre anni diversi, le proporzioni della mortalità infantile in Inghilterra e Galles in tre diversi momenti del primo anno di vita.

Tab 2.1 Distribuzione di probabilità di una variabile casuale X che rappresenta l'ordine di nascita dei neonati negli Stati Uniti

x	$P(X = x)$
1	0,416
2	0,330
3	0,158
4	0,058
5	0,021
6	0,009
7	0,004
≥ 8	0,004
Totale	1,000

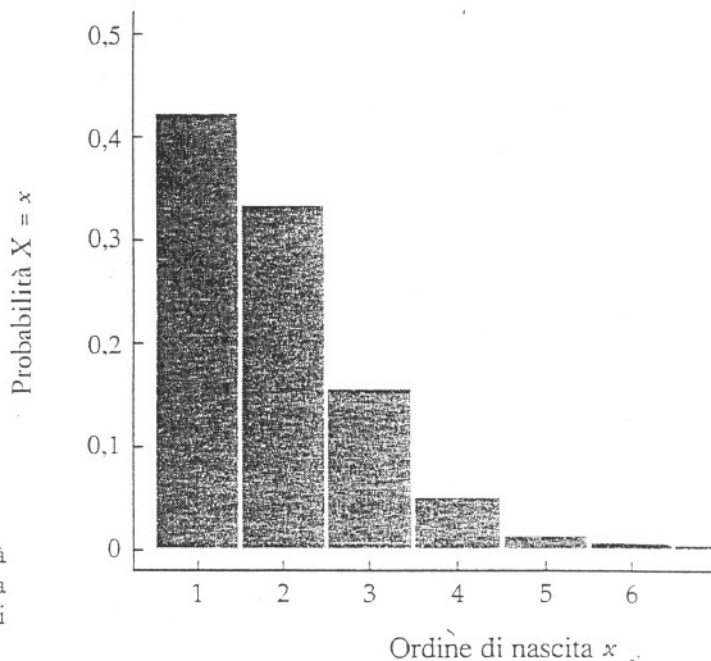


Fig 2.1 Distribuzione di probabilità di una variabile casuale che rappresenta l'ordine di nascita di neonati negli Stati Uniti

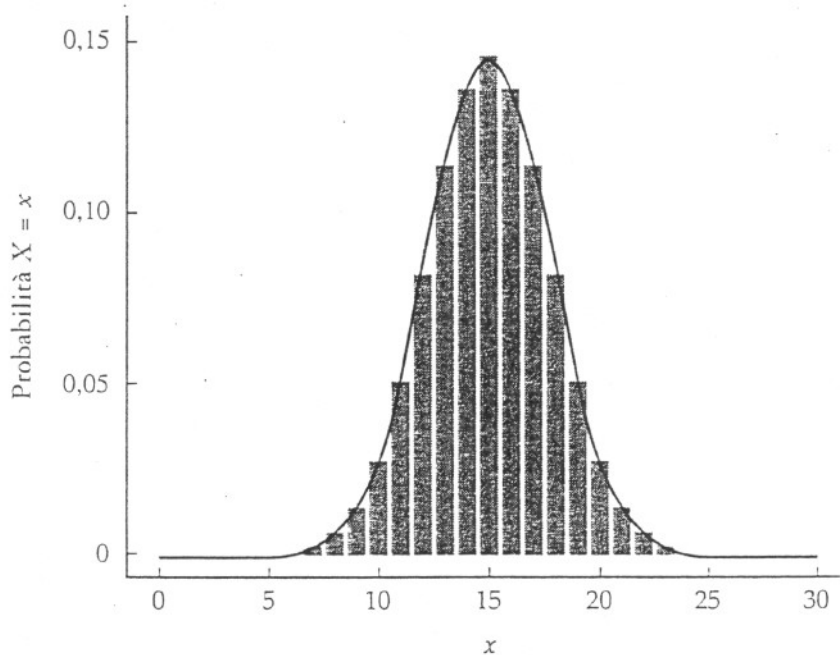


Fig. 2,2

Volume espiratorio forzato in un secondo
in pazienti con patologia coronarica provenienti da tre
diversi centri medici

Johns Hopkins	Rancho Los Amigos	St. Louis
3,23	3,22	2,79
3,47	2,88	3,22
1,86	1,71	2,25
2,47	2,89	2,98
3,01	3,77	2,47
1,69	3,29	2,77
2,10	3,39	2,95
2,81	3,86	3,56
3,28	2,64	2,88
3,36	2,71	2,63
2,61	2,71	3,38
2,91	3,41	3,07
1,98	2,87	2,81
2,57	2,61	3,17
2,08	3,39	2,23
2,47	3,17	2,19
2,47		4,06
2,74		1,98
2,88		2,81
2,63		2,85
2,53		2,43
		3,20
		3,53
$\bar{x}_1 = 2,63$ litri $s_1 = 0,496$ litri	$\bar{x}_2 = 3,03$ litri $s_2 = 0,523$ litri	$\bar{x}_3 = 2,88$ litri $s_3 = 0,498$ litri

Figura 4.1

Tabella 5.1 Riduzione della capacità vitale forzata per un campione di pazienti con fibrosi cistica

Soggetto	Riduzione della capacità vitale forzata (ml)		Differenza	Rango	Rango con segno	
	Placebo	Farmaco				
1	224	213	11	1	1	
2	80	95	- 15	2		- 2
3	75	33	42	3	3	
4	541	440	101	4	4	
5	74	- 32	106	5	5	
6	85	- 28	113	6	6	
7	293	445	-152	7		- 7
8	- 23	-178	155	8	8	
9	525	367	158	9	9	
10	- 38	140	-178	10		-10
11	508	323	185	11	11	
12	255	10	245	12	12	
13	525	65	460	13	13	
14	1.023	343	680	14	14	
					<u>86</u>	<u>-19</u>

Paese	Percentuale vaccinati	Tasso di mortalità per 1.000 nati vivi
Bolivia	40	165
Brasile	54	85
Canada	85	9
Cina	95	43
Egitto	81	94
Etiopia	26	226
Finlandia	90	7
Francia	95	9
Giappone	83	6
Grecia	83	12
India	83	145
Italia	85	11
Iugoslavia	91	27
Messico	65	51
Polonia	98	18
Regno Unito	75	10
Senegal	47	189
Stati Uniti	97	12
Turchia	74	90
URSS	79	35

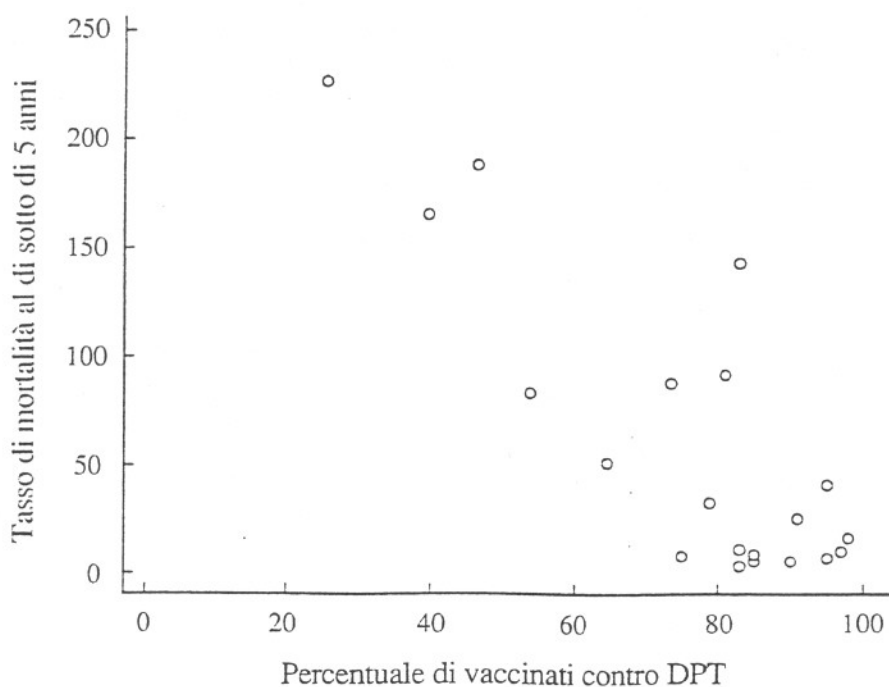
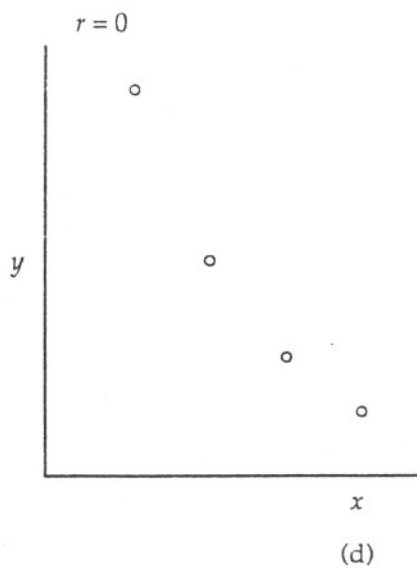
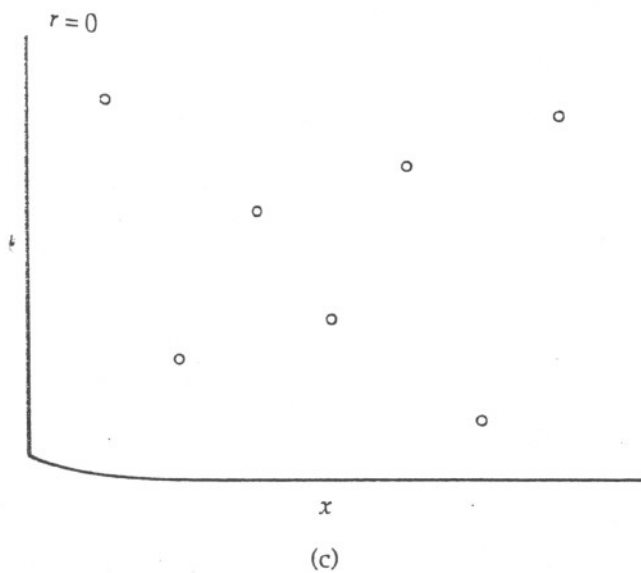
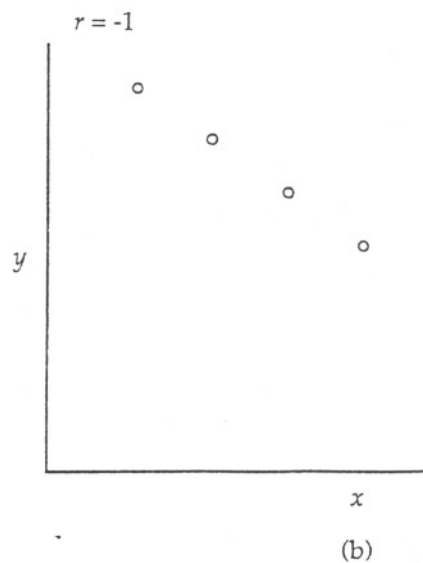
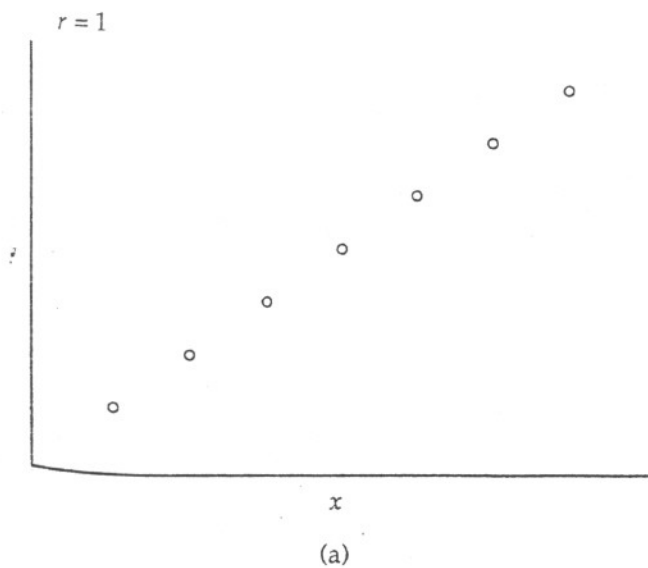


Fig. 8.1



Diagrammi a punti che illustrano possibili relazioni tra X e Y

Fig. 8.2

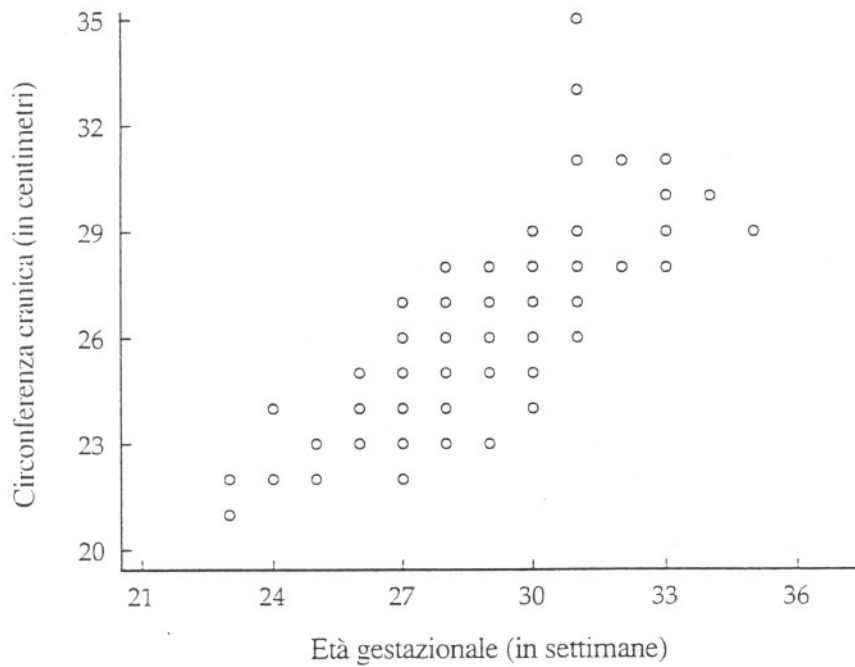


Fig. 9.1

Età e pressione sistolica di 33 donne

Età (anni)	Pressione sistolica (mm Hg)	Età (anni)	Pressione sistolica (mm Hg)	Età (anni)	Pressione sistolica (mm Hg)
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

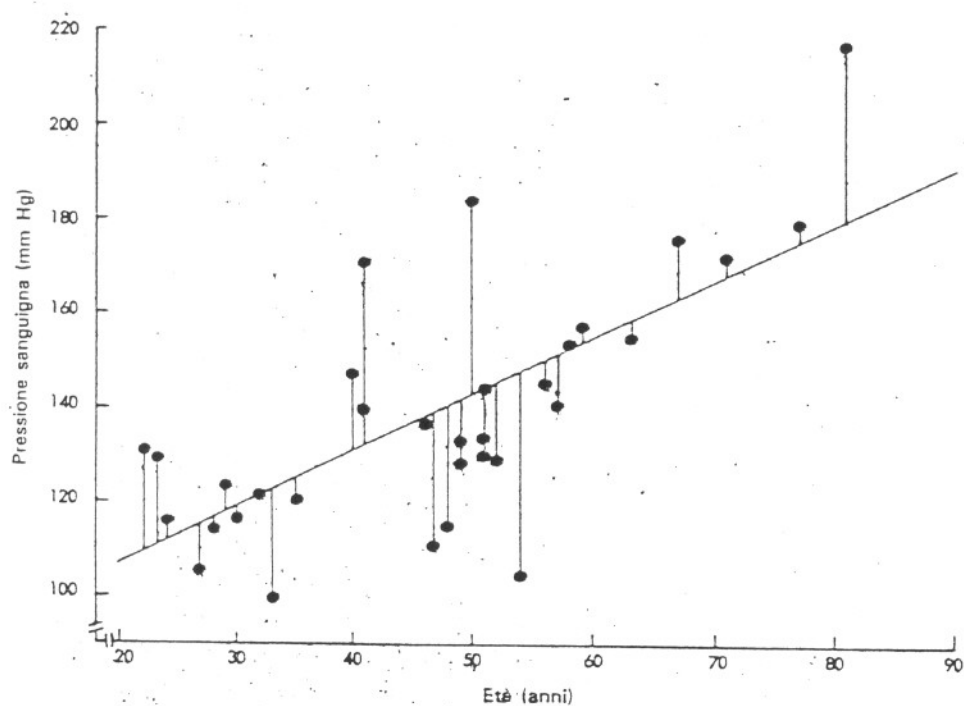
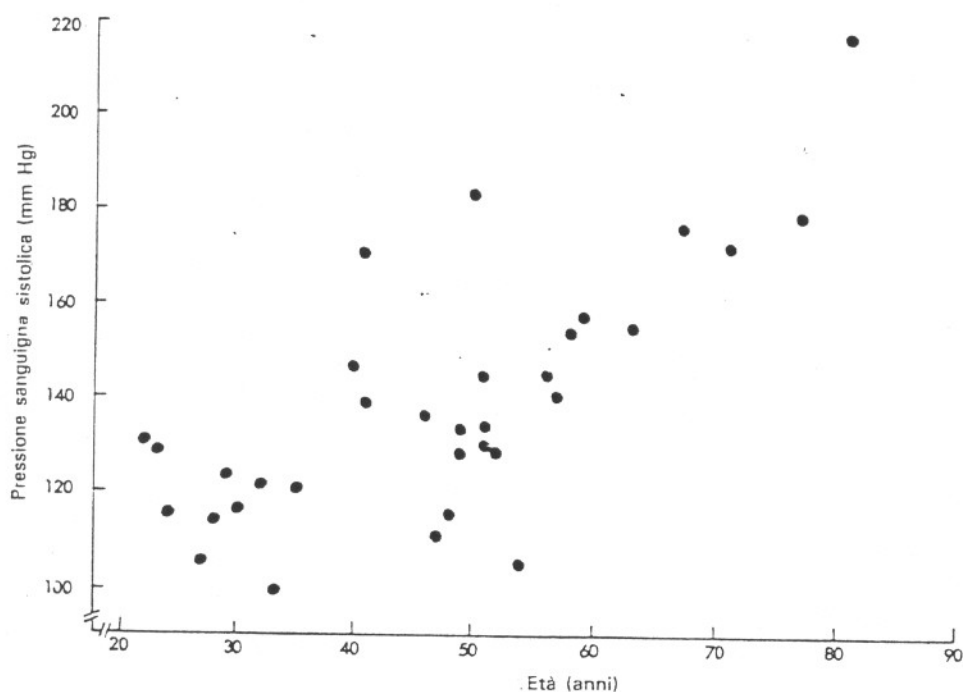


Fig. 9.2

Circonferenza cranica: maschi

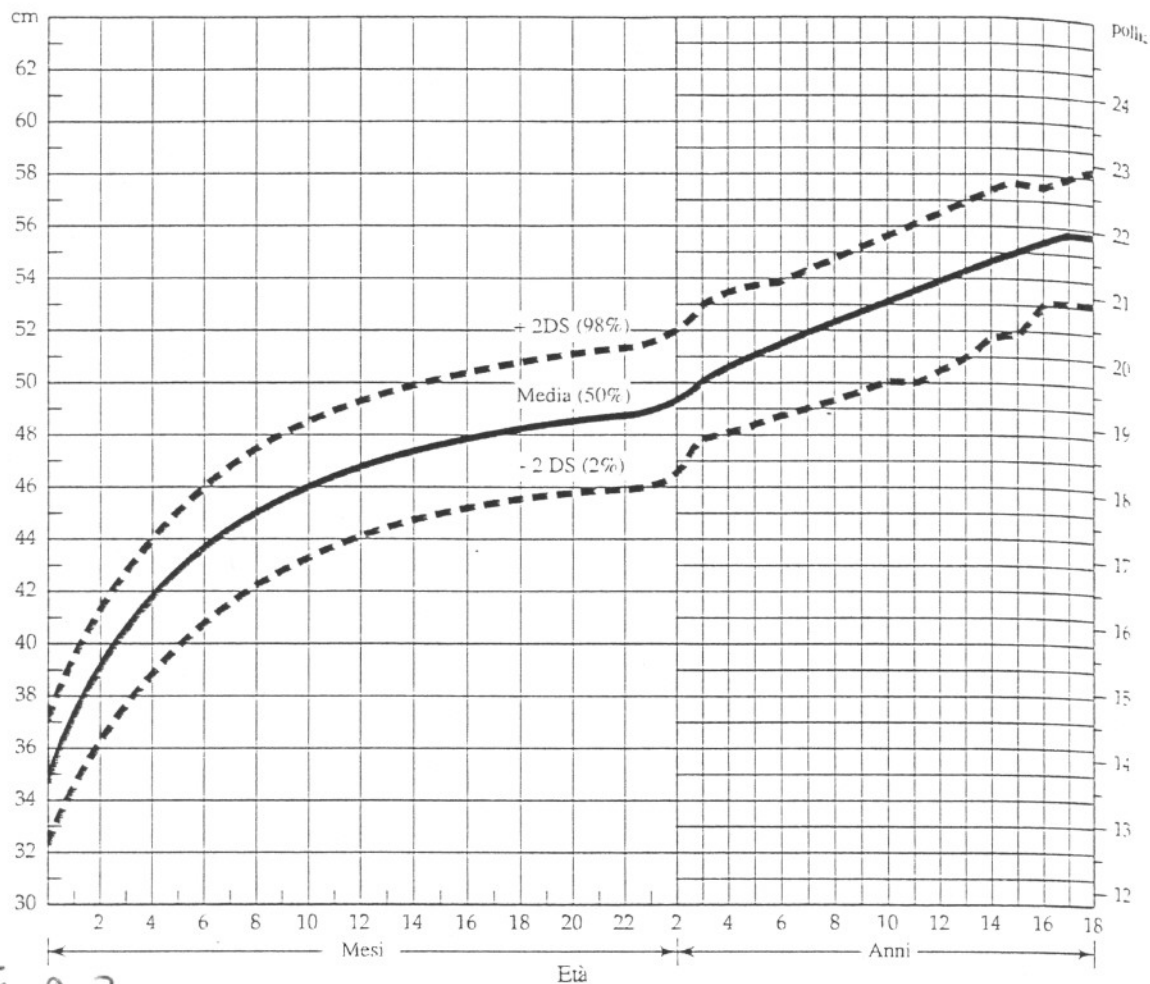


Fig. 9.3

Circonferenza cranica in funzione dell'età nei maschi (Da: G. Nellhaus, *Pediatrics*, 41:106, 1968. Copyright 1968 American Academy of Pediatrics)

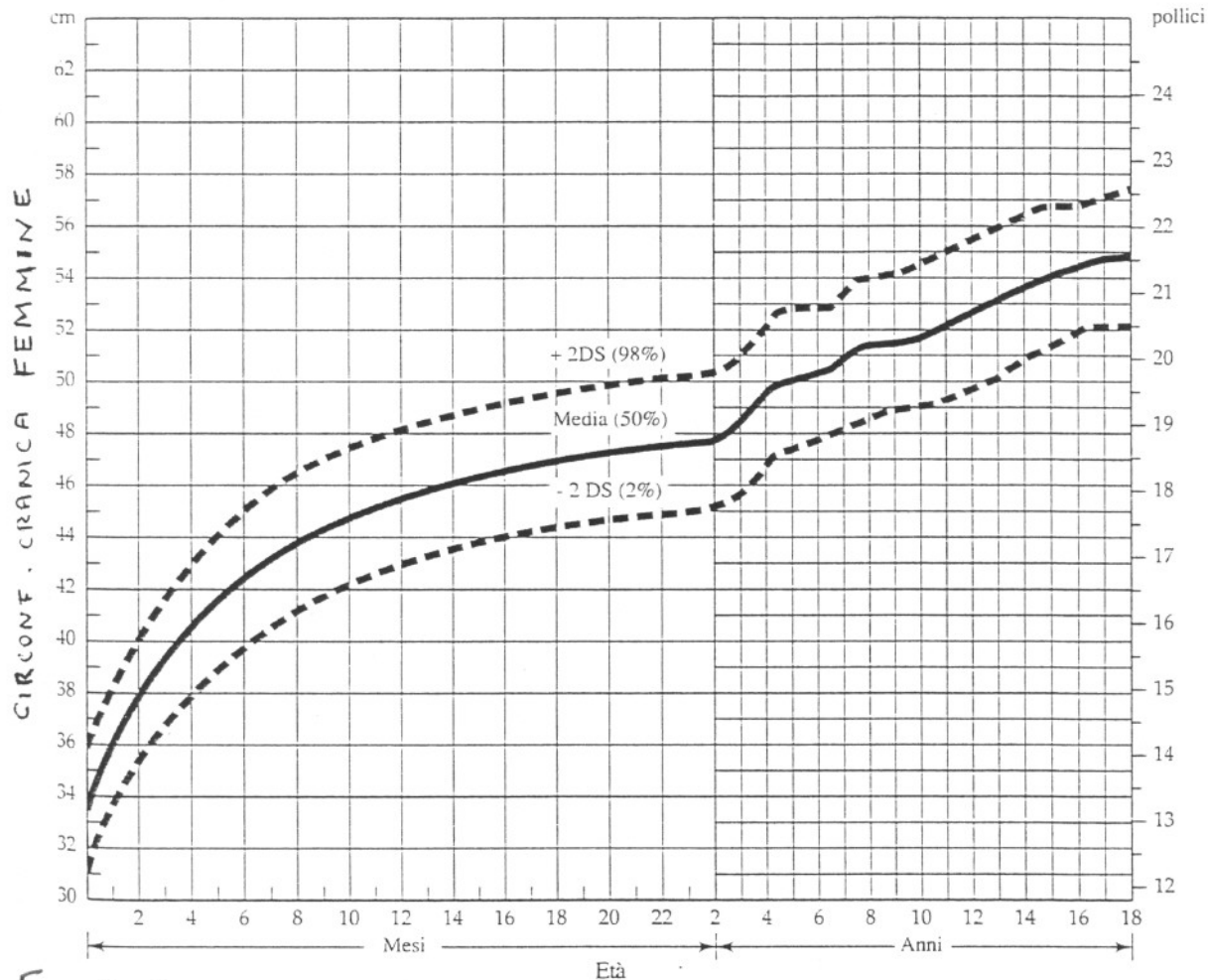


Fig. 9.4 Circonferenza cranica in funzione dell'età nelle femmine (Da: G. Nellhaus, *Pediatrics*, 41:106, 1968. Copyright 1968 American Academy of Pediatrics)

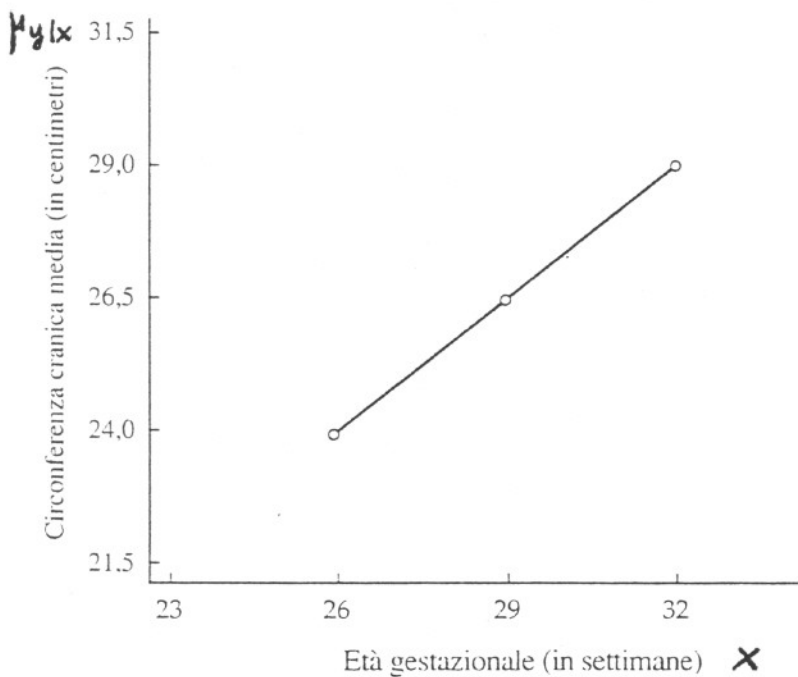


Fig. 9.5

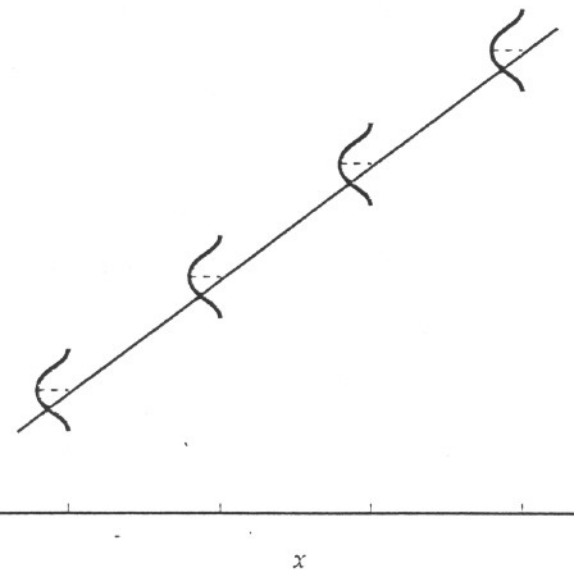


Fig. 9.6

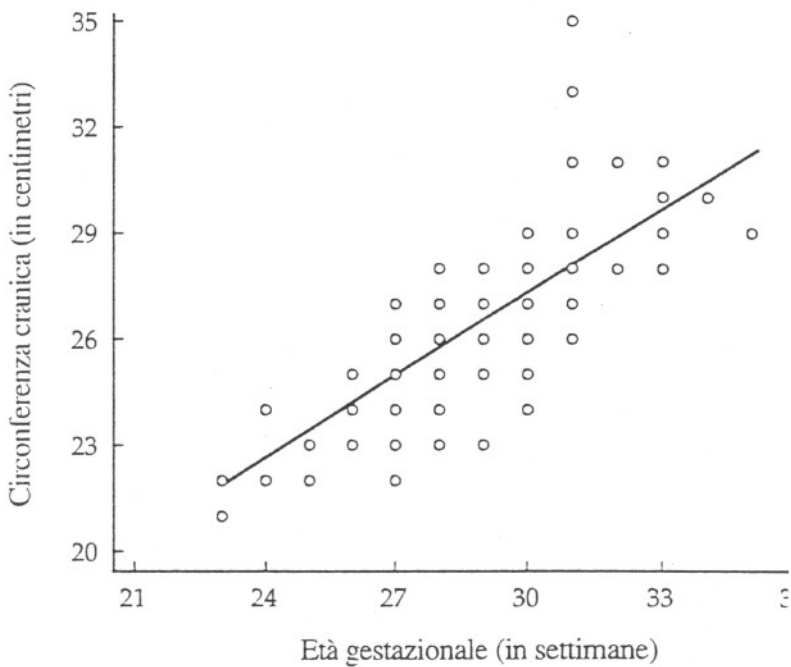
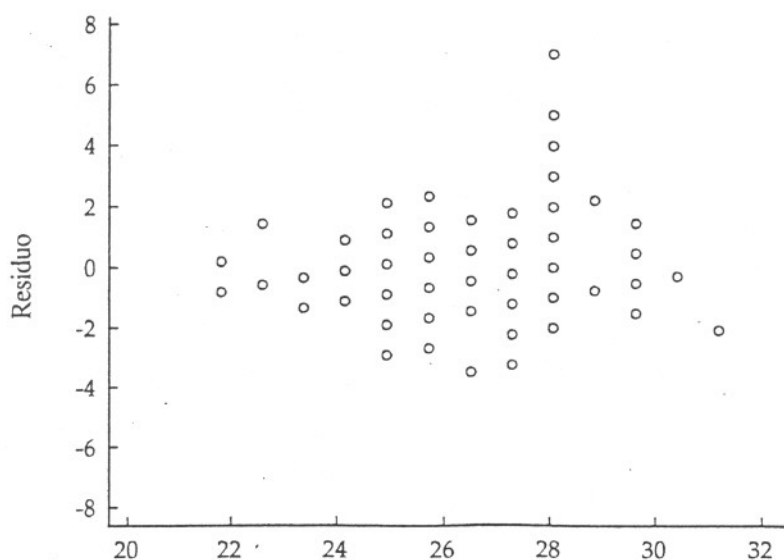
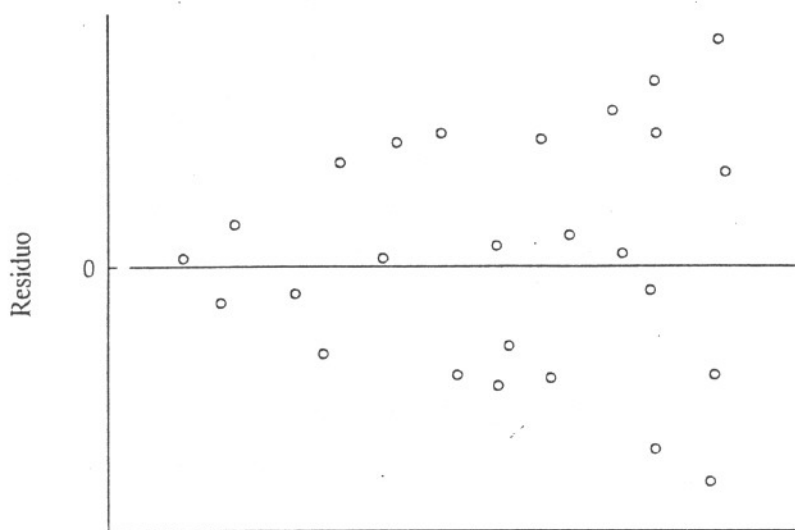


Fig. 9.7

RESIDUI IN FUNZIONE DEI VALORI SULLA RETTA AI MINIMI QUADRATI



(a) Valore adattato della circonferenza cranica



(b) Valore adattato di y

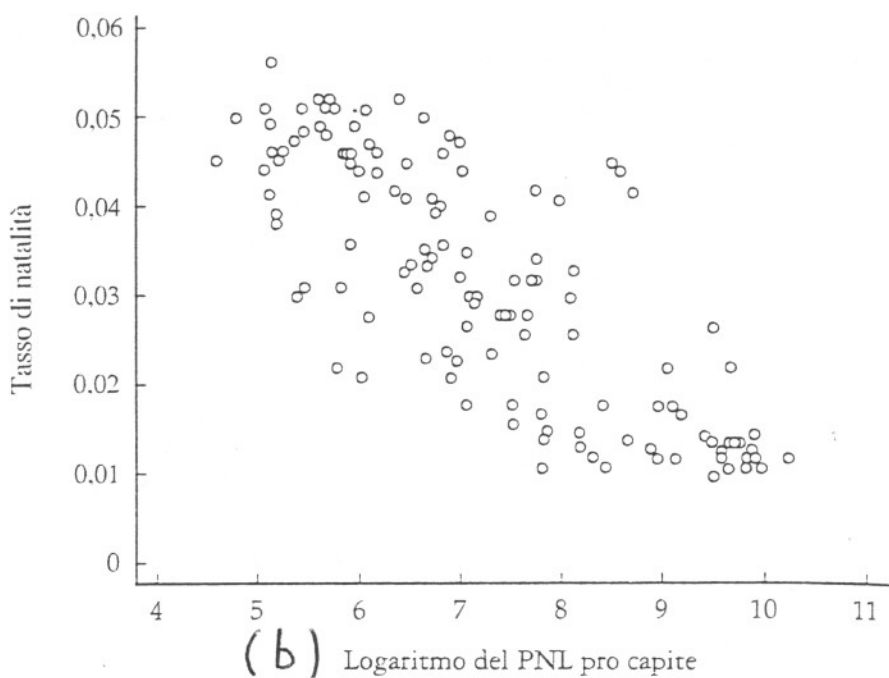
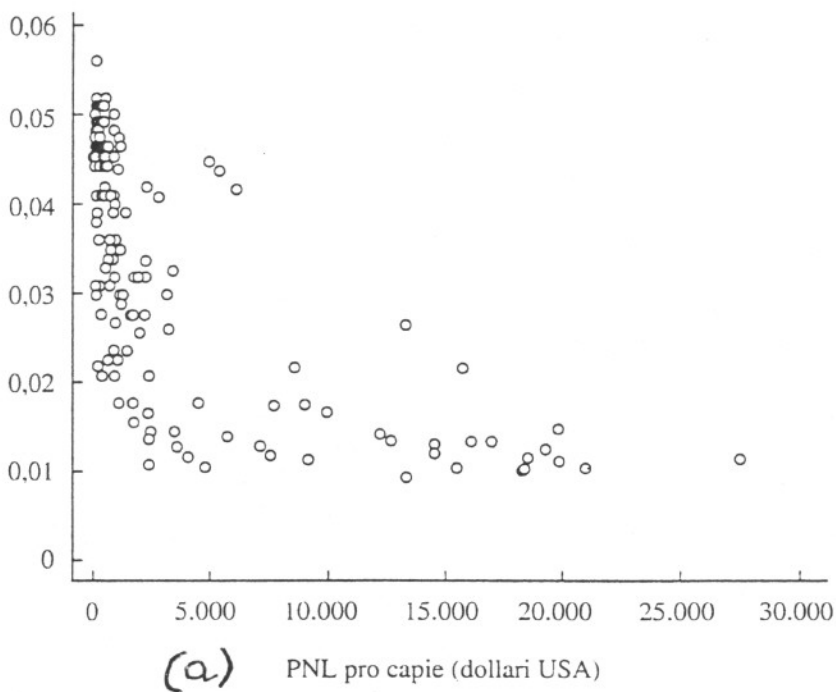


Fig. 9.9

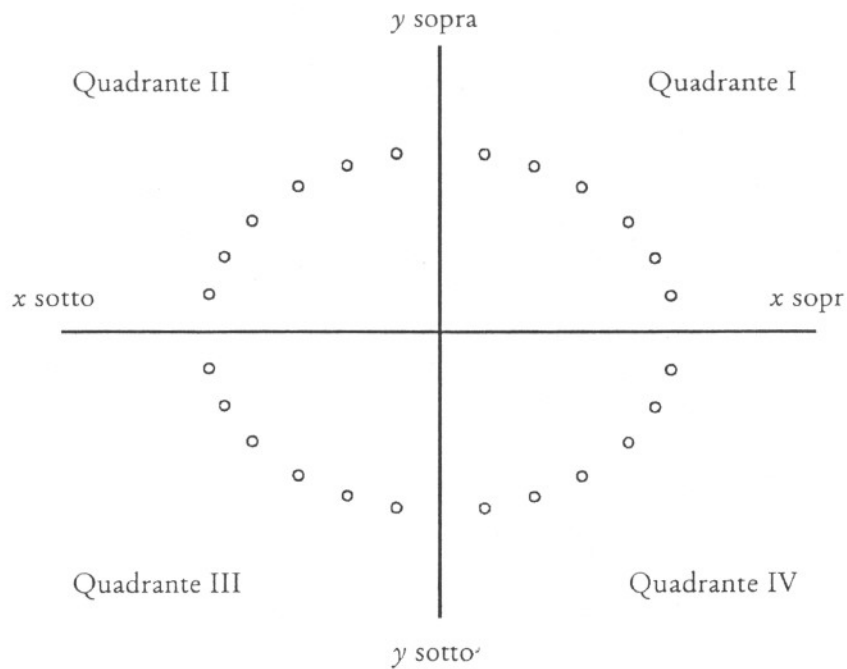


Figura 9.10 Il cerchio delle potenze

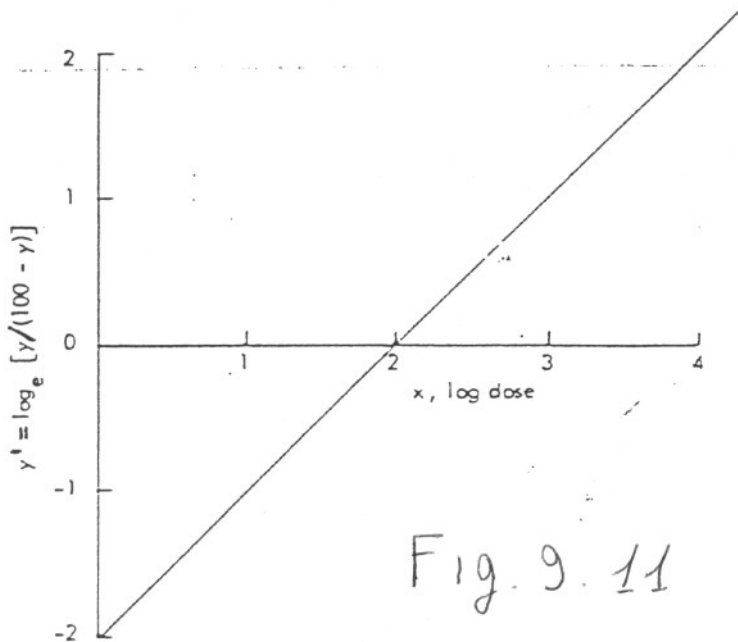
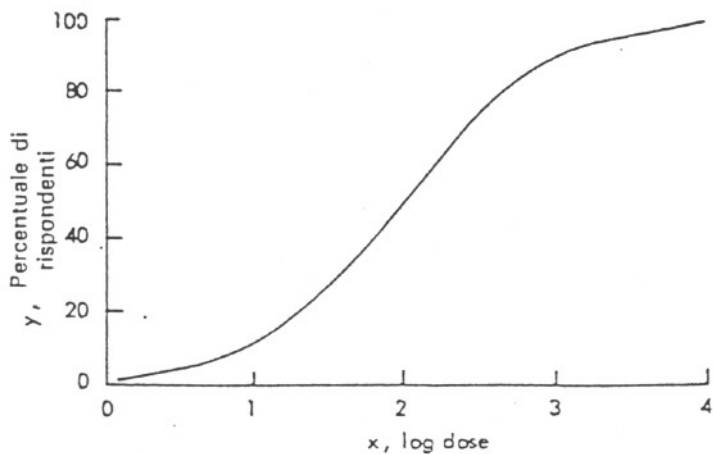


Fig. 9. 11