Università di Torino

QUADERNI DIDATTICI

del

Dipartimento di Matematica

V. Demichelis, A. Ziggioto

Lezioni di Biostatistica

Quaderno # 36 - Aprile 2006

Indice

Prefazione

Queste note raccolgono le lezioni 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.

Il quaderno presenta alcuni metodi di base della Biostatistica in modo semplice e chiaro, senza fare uso del calcolo differenziale ed integrale, con una particolare attenzione alle applicazioni ad insiemi di dati di tipo clinico e farmacologico.

Il primo capitolo tratta argomenti di Statistica descrittiva e quindi la sintesi, la rappresentazione grafica ed i parametri di posizionamento e dispersione di dati sperimentali. Il secondo capitolo introduce il concetto di probabilità e di distribuzione teorica di probabilità, analizzando le distribuzione di Bernoulli e di Poisson nel caso discreto e la distribuzione di Gauss nel caso continuo. Il terzo capitolo è dedicato alla distribuzione della media campionari. Il quarto riguarda l'inferenza sulla media e sul confronto fra due o più medie per mezzo di intervalli di confidenza e test d'ipotesi. Il quinto capitolo contiene un esempio di test d'ipotesi non parametrico per il confronto fra due mediane. Il sesto capitolo riguarda l'inferenza sulle proporzioni. Il settimo presenta il confronto fra due o più proporzioni mediante il test chiquadro applicato a dati organizzati in tabella di contingenza. L'ottavo e nono capitolo sono dedicati alla correlazione ed alla regressione lineare.

Il quaderno si conclude con due appendici: la prima contiene le tabelle delle distribuzioni teoriche di probabilità, la seconda presenta i grafici e le figure che servono per illustrare le applicazioni dei metodi. Il contenuto delle due appendici 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 Ziggioto

Capitolo 1

Statistica descrittiva

La statistica descrittiva organizza e sintetizza le osservazioni, utilizza tabelle, grafici e misure di sintesi numerica per illustrare una serie di dati.

1.1 Dati

I dati raccolti per un'indagine statistica possono essere di vario tipo:

- nominali: i valori rientrano in categorie non ordinate (es. maschio/femmina). Spesso si utilizzano i numeri per rappresentare le categorie, ma l'ordine e la grandezza di questi numeri non sono importanti. I dati nominali che assumono uno di due distinti valori sono detti *dicotomici*. Non tutti i dati nominali sono dicotomici; possono esistere 3 o più possibili categorie in cui possono rientrare le osservazioni (es. gruppi sanguigni).

- ordinali: esiste un ordine predeterminato fra le categorie (es. gravità di una ferita)

- ordinati in ranghi: disponiamo in ordine decrescente le osservazioni in ordine alla grandezza e poi a ciascuna osservazione è assegnato un numero che corrisponde alla relativa posizione nella sequenza (es. le 10 principali cause di decesso negli Stati Uniti nel 1988)

- numerici: sono quelli per cui hanno valore sia l'ordine che la grandezza. I numeri non sono più semplici simboli ma rappresentano quantità realmente misurabili. I dati numerici si dividono in

    1. discreti: possono assumere solo valori specifici che differiscono l'uno dall'altro per quantità fisse; spesso sono numeri interi (es. numero di nascite)
    2. continui: rappresentano quantità misurabili che possono assumere qualunque valore. Teoricamente, ciascuna osservazione cade in un

certo punto lungo un asse continuo. In questo caso la differenza tra 2 possibili valori può essere arbitrariamente piccola (es. altezze, peso, pressione, concentrazione). Il solo fattore restrittivo per un'osservazione continua è il grado di accuratezza dello strumento di misura (es. il tempo è approssimato al secondo, il peso al grammo più vicino).

1.1.1 Sintesi dei dati

I dati raccolti possono essere sintetizzati in vari modi. Due di questi sono i seguenti:

- TABELLE

  Una tabella è forse il modo più semplice per sintetizzare una serie di osservazioni e può essere utilizzata per tutti i tipi di dati.

- DISTRIBUZIONE DI FREQUENZA

  La distribuzione di frequenza è un tipo di tabella comunemente utilizzato per rappresentare i dati facendo una sintesi dei dati stessi.

1. per dati ordinali e nominali: è una serie di classi o categorie con somma numerica relativa a ciascuna di esse;

2. per dati discreti o continui dobbiamo scomporre l'intervallo in cui variano i dati in una serie di sottointervalli distinti e non sovrapposti che si chiamano *classi di frequenza*

Una volta selezionati i limiti inferiore e superiore di ciascuna classe di frequenza, si calcola il numero di osservazioni (=frequenza) i cui valori cadono nella classe ed i risultati sono organizzati in una tabella (Tabella 1.2).

NOTA BENE

Se ci son troppi intervalli, la sintesi non è un reale miglioramento rispetto ai dati grezzi. Se ce ne sono troppo pochi, si perde una grande quantità di informazione. Anche se non è necessario, gli intervalli hanno spesso la stessa ampiezza; questo facilita il confronto fra di essi.

Esistono alcune *regole*, basate sull'esperienza e sul buon senso, per costruire correttamente la suddivisione in classi di frequenza:

1. Posto N il numero di dati, sia n il più piccolo numero naturale tale che $2^n > N$. Allora il numero delle classi è n o $n + 1$;

2. I limiti di ciascuna classe devono essere in accordo con l'accuratezza con cui sono state misurate le osservazioni;

3. Classi di frequenza di ampiezza uguale sono convenienti e facilitano i calcoli successivi;

4. Le classi di frequenza devono essere mutuamente esclusive. Se, ad esempio, devo suddividere in classi di frequenza dell'età, gli intervalli da 5 a 10 anni e da 10 a 15 anni non si escludono a vicenda (come classifico un individuo di 10 anni?);

5. Evitare intervalli aperti su uno dei 2 estremi;

6. Occorre calcolare il punto centrale di ciascuna classe. Per fare questo bisogna distinguere fra limiti tabulati e limiti veri della classe di frequenza. Il punto centrale di una classe è il punto medio dell'intervallo che ha come estremi i limiti veri.

Definiamo ora i concetti chiave di *frequenza relativa* e *frequenza relativa cumulativa*.

- **Frequenza relativa**

  La frequenza relativa di una classe si calcola suddividendo il numero di osservazioni che cadono all'interno della classe per il numero totale delle osservazioni. Se moltiplichiamo la frequenza relativa per 100 otteniamo la *frequenza relativa percentuale*, cioè la percentuale del numero totale di osservazioni che appartiene alla classe. Le frequenze relative sono utili per confrontare serie di dati che contengono numeri diversi di osservazioni.

- **Frequenza relativa cumulativa**

  La frequenza relativa cumulativa di una classe di frequenza è la percentuale del numero totale di osservazioni che ha un valore inferiore o uguale al limite superiore della classe stessa. Si calcola sommando la frequenza relativa della classe stessa con quelle di tutte le classi di frequenza precedenti.

1.2 Rappresentazione grafica di dati statistici

I dati possono essere sintetizzati ed illustrati anche attraverso l'uso di *grafici*, o rappresentazioni figurate di dati numerici. I grafici devono essere realizzati in modo tale da comunicare al primo sguardo l'andamento generale di una serie di dati.

La lettura di un grafico deve garantirmi una maggior semplicità di interpretazione con la descrizione di un minor numero di dettagli, ai fine di ottenere una migliore comprensione dei dati

I diagrammi statistici servono a **2 scopi**:

1. presentazione di informazioni statistiche in articoli e relazioni, nella supposizione che il lettore apprezzi un'illustrazione semplice e suggestiva;

2. aiuto personale alla ricerca statistica. Lo statistico ricorre spesso ai diagrammi per intuire la struttura dei dati e controllare gli assunti possibili per l'analisi. L'utilizzo informale dei diagrammi spesso può svelare nuovi aspetti dei dati e suggerire ipotesi per ricerche successive.

Vediamo ora alcune delle principali rappresentazioni grafiche di dati statistici.

1.2.1 Diagramma a barre

E' utilizzato per illustrare una distribuzione di frequenza per dati nominali, ordinali o numerici discreti non raggruppati in classi di frequenza. Le diverse categorie in cui rientrano le osservazioni sono presentate lungo un asse orizzontale. Una barra verticale è tracciata al di sopra di ogni categoria e l'altezza della barra rappresenta la frequenza assoluta o la frequenza relativa delle osservazioni appartenenti a quella classe. Le barre devono avere uguale ampiezza ed essere separate l'una dall'altra per non implicare alcuna continuità. Come esempio possiamo vedere la Figura 1.1, che riporta il numero di maschi in famiglie con 8 figli.

1.2.2 Istogramma

Illustra una distribuzione di frequenza per dati numerici discreti o continui. L'asse orizzontale indica i limiti reali delle diverse classi di frequenza, cioè i punti che separano ciascun intervallo dagli intervalli contigui. L'asse verticale illustra la frequenza assoluta o relativa delle osservazioni in ciascun intervallo. Se la variabile è discreta e non raggruppata in classi, le frequenze possono essere rappresentate da linee verticali o bastoncini (diagramma a barre). Il metodo più generale se la variabile è raggruppata in classi è disegnare rettangoli che abbiano come basi i singoli intervalli di classe.

Prima di tutto bisogna tracciare le scale degli assi, fissando le unità di misura per ciascun asse. La scala verticale deve iniziare da 0. Su ciascuna classe è posta una barra verticale centrata nel punto medio della classe.

L'area della barra indica la frequenza associata a quella classe. La porzione di area totale dell'istogramma corrispondente ad una classe è pari alla frequenza relativa o ass. della classe stessa (Figura 1.2). Perciò, un istogramma che rappresenti le frequenze relative ha la stessa forma di un istogramma che rappresenti le frequenze ass.

NOTA BENE

La frequenza di una classe è proporzionale all'*area* piuttosto che all'altezza del rettangolo. Questo tiene conto del fatto che non sempre la lunghezza delle classi è costante. Ovviamente se la lunghezza dell'intervallo non varia da classe a classe, le aree sono naturalmente proporzionali alle altezze, e le frequenze sono rappresentate dalle altezze come dalle aree.

Quante classi creare a partire dai dati?

Sia N il numero totale dei dati. Sia n il più piccolo numero naturale tale che $N < 2^n$. Il numero ottimale di suddivisioni in intervalli è n oppure $n+1$. In genere tra n e $n+1$ si preferisce il numero dispari perchè l'istogramma viene ad avere una classe centrale particolarmente significativa per ragioni di simmetria.

1.2.3 Poligono di frequenza

Utilizza i 2 stessi assi dell'istogramma. E' costruito considerando i punti che hanno per ascissa il punto centrale di ciascuna classe e per ordinata la frequenza o la frequenza relativa associata alla classe. Sono posti dei punti anche sull'asse orizzontale nel punto medio degli intervalli che immediatamente precedono o seguono gli intervalli che contengono le osservazioni. I punti sono poi uniti tra loro da segmenti di retta.

Poichè possono essere facilmente sovrapposti, i poligoni di frequenza sono più idonei degli istogrammi per confrontare serie di dati (Figure 1.3 e 1.4).

1.2.4 Poligono di frequenza cumulativa

Rappresenta graficamente le frequenze relative cumulative. Un punto viene posto al limite superiore vero di ciascuna classe, l'altezza del punto rappresenta la frequenza relativa cumulativa associata a quella classe.

Anche questi poligoni possono essere usati per confrontare serie di dati. Essi si possono usare per ottenere i *percentili* di una serie di dati (Figura 1.5). Ad esempio, il 95mo percentile è il valore maggiore o uguale al 95% delle osservazioni e minore o uguale al restante 5%.

1.2.5 Diagramma a dispersione a 2 dimensioni

E' utilizzato per illustrare la relazione tra 2 diverse misure continue. Ogni punto del grafico rappresenta una coppia di valori. Esso dà una completa descrizione della distribuzione delle singole variabili e della relazione tra di esse (Figura 1.6). Inoltre permette di:

1. dare una misura numerica di alcune caratteristiche fondamentali della relazione;

2. predire il valore di una variabile, noto il valore dell'altra;

3. valutare la significatività della direzione di una tendenza apparente.

1.2.6 Diagramma lineare

E' simile ad un diagramma a punti poichè può essere utilizzato per illustrare la relazione tra quantità continue. Ciascun punto sul grafico rappresenta una coppia di valori. In questo caso, però, ciascun valore sull'asse x ha un'unica misurazione corrispondente sull'asse y; i punti adiacenti sono collegati tra loro da linee rette (Figura 1.7).

Viene utilizzato per descrivere l'andamento di un fenomeno variabile in un certo intervallo di tempo (o di spazio).

1.2.7 Areogramma

Si utilizza quando si tratta di visualizzare la diverse *parti* in cui *un tutto* è stato suddiviso. Per convenzione, le ampiezze dei settori circolari devono essere proporzionali alle grandezza delle corrispondenti parti.

Questa rappresentazione è particolarmente efficace quando interessa mettere in evidenza, più che le misure effettive delle singole grandezza in gioco, i loro mutui rapporti (es. composizione del Parlamento, composizione chimica di un medicinale...). Non è invece adatta a rappresentare le temperature registrare in alcune città italiane, nè le temperature di un malato rilevate in diverse ore del giorno.

Il calcolo delle ampiezze dei settori circolari di un areogramma (espresse in gradi) risulta da una semplice proporzione, tenendo presente che il totale complessivo delle quantità considerate deve corripondere all'intera torta, per un'ampiezza, quindi, di 360°.

1.2.8 Ideogramma

Utilizza file di simboli che si ripetono; per esempio, le popolazioni di diversi paesi possono essere disegnate come file di omini, ognuno dei quali rappresenta un certo numero di individui.

Si utilizza per confrontare tra loro 2 o più frequenze.

1.3 Misure di sintesi numerica

1.3.1 Misure di tendenza centrale

La caratteristica di una serie di dati più comunemente studiata è il suo **centro** o il punto in cui le osservazioni tendono a raccogliersi. Le misure di tendenza centrale sono sostanzialmente 3:

1. **Media**

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

dove $x_1, x_2, \ldots, x_n$ sono le osservazioni nel campione che sto analizzando.

- Può essere usata come misura di sintesi per misurazioni *discrete* o *continue*.
- Non è adatta a dati nominali o ordinali, ad eccezione dei dati dicotomici.
- Quando un'osservazione ha un valore molto diverso dagli altri, la media varia molto, cioè è estremamente sensibile a valori insoliti.

2. **Mediana**

- E' il cinquantesimo percentile di una serie di n misurazioni. Se n è dispari, è il valore centrale, se n è pari, è la media dei due valori centrali.
- Può essere usata per dati ordinali, discreti e continui.
- E' *robusta*, cioè è poco sensibile ai valori estremi. Ad esempio, se ordiniamo le 4 misurazioni seguenti

$$2,\ 2.5,\ 3.5,\ 4$$

la mediana è

$$\frac{2.5 + 3.5}{2} = 3.$$

Se sostituiamo a 2 il valore 20 e ordiniamo i valori

$$2.5,\ 3.5,\ 4,\ 20$$

la mediana è

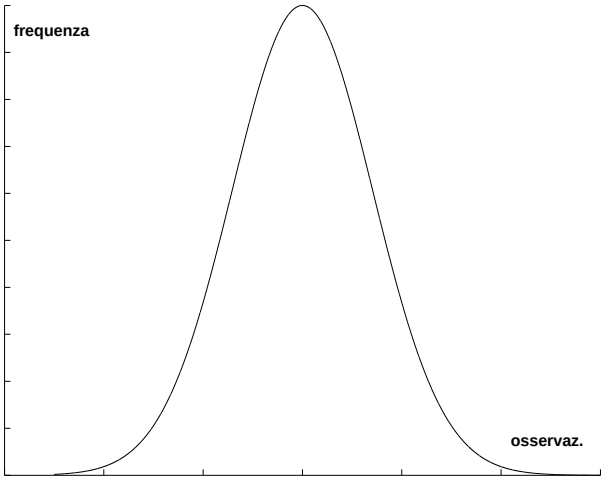
$$\frac{3.5 + 4}{2} = 3.75.$$

3. **Moda**

- E' l'osservazione che si verifica con maggior frequenza. Non è detto che sia unica.
- Può essere usata per qualsiasi tipo di dati.

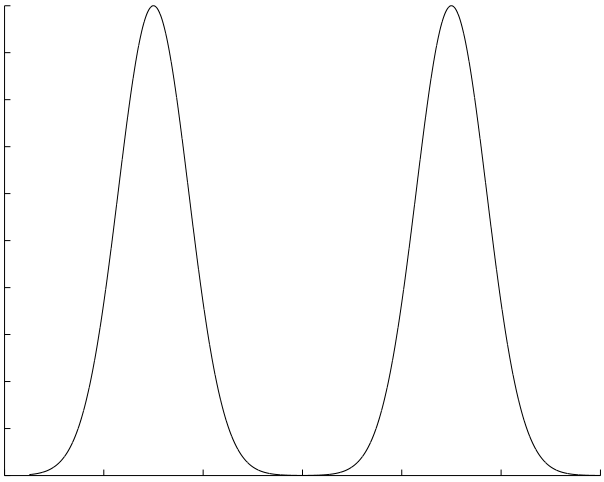
La miglior misura di tendenza centrale per una serie di dati dipende da come sono distribuiti i singoli valori:

- Dati simmetrici **unimodali**



Media, moda e mediana sono approssimativamente uguali.

- Dati simmetrici **bimodali**

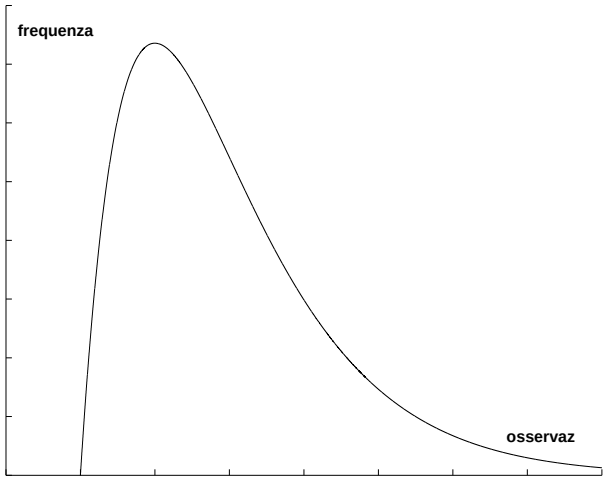


Media e mediana sono approssimativamente uguali. Però questo valore comune potrebbe trovarsi tra 2 picchi ed essere quindi una misurazione che si verifica difficilmente.

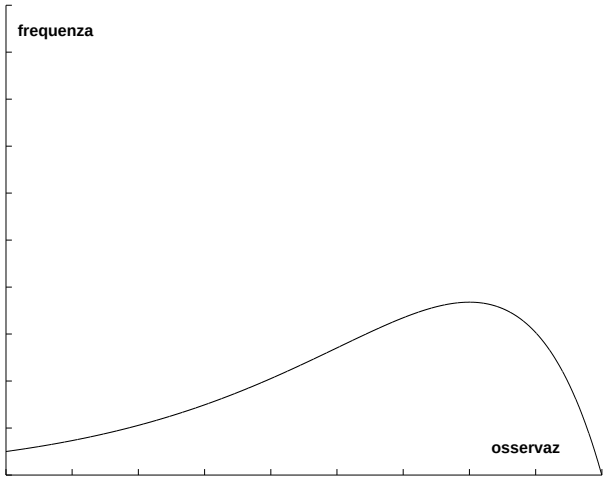
La popolazione consiste in 2 gruppi distinti che differiscono per la caratteristica misurata. E' preferibile qui riportare 2 mode piuttosto che media e mediana.

- Dati **asimmetrici**

I dati possono essere asimmetrici a destra verso i valori più bassi delle misurazioni, quando la media è a destra rispetto alla mediana.



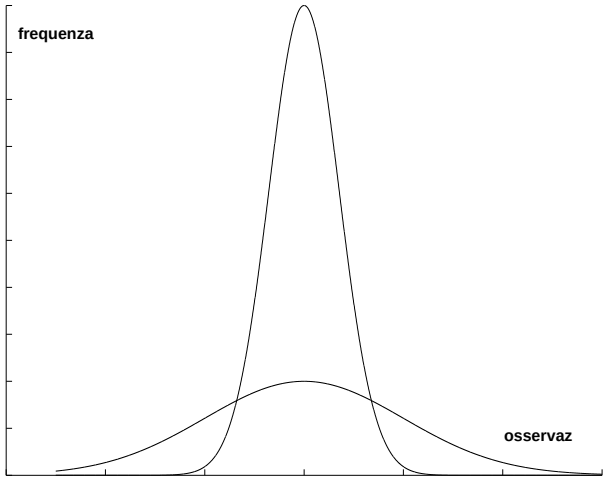
Oppure possono essere asimmetrici a sinistra verso i valori più alti delle misurazioni, quando la media è a sinistra della mediana.



In entrambi i casi, la mediana è la miglior misura di tendenza centrale. Poichè la media è sensibile alle osservazioni estreme, essa è spostata nella direzione dei valori delle osservazioni atipiche e pertanto può risultare essenzialmente aumentata o ridotta.

Per sapere quanto sia realmente valida la nostra misura di tendenza centrale, dobbiamo avere un'idea della **variabilità** tra i valori dei dati.

Tutte le osservazioni tendono ad essere simili e perciò si situano vicino al centro o sono distribuite su un ampio intervallo di valori?



1.3.2 Indici di dispersione

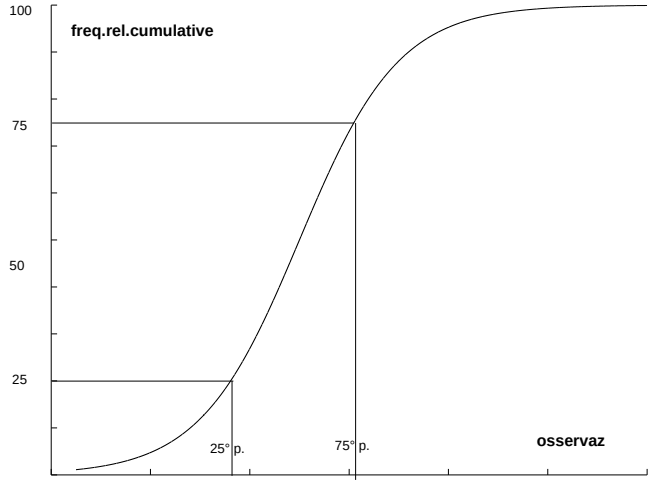
I principali indice di dispersione sono i seguenti:

1. **Campo di variazione**

   E' la differenza (in valore assoluto) fra l'osservazione più grande e quella più piccola. Ha un'utilità molto limitata. E' molto sensibile a valori molto grandi o molto piccoli.

2. **Campo di variazione interquartile**

   E' la differenza fra il 75mo percentile e il 25mo percentile e comprende il 50% delle osservazioni centrali. E' meno sensibile ai valori estremi rispetto al campo di variazione.



3. **Deviazione standard e varianza**

La *varianza* dà la misura dell'entità delle variabilità o dispersione dalla media di un campione:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Dividendo per $n-1$ invece che per n nella definizione appena data per s^2, teniamo conto del fatto che, come si può facilmente verificare,

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

che lega le osservazioni alla loro media.

La *deviazione standard* è la radice quadrata della varianza:

$$s = \sqrt{s^2}.$$

La deviazione standard ha le stesse unità di misura delle osservazioni e della loro media.

Se la deviazione standard è *piccola* allora abbiamo una maggior omogeneità.

Se la deviazione standard è *grande* allora abbiamo una maggior variabilità.

4. **Coefficiente di variazione**

   Il coefficiente di variazione di un insieme di osservazioni è definito dal rapporto fra deviazione standard e media, moltiplicato per 100:

$$CV = \frac{s}{\overline{x}} \cdot 100.$$

   Esso è un numero adimensionale e può essere usato per confrontare la dispersione relativa di due diverse serie di dati.

1.3.3 Dati raggruppati

La tecnica di raggruppare le misurazioni che hanno ugual valore prima di calcolare la media offre vantaggi rispetto al metodo standard perchè si può applicare a dati che sono stati rappresentati sotto forma di distribuzione di frequenza. In questo caso non conosciamo più le singole osservazioni; sappiamo però quanti dati cadono in una certa classe di frequenza.

L'**ipotesi** che si formula è che tutti i valori che rientrano in un determinato intervallo sono uguali al *punto medio* dell'intervallo stesso.

Si ha che la media e la varianza dei dati raggruppati sono

$$\overline{x} = \frac{\sum_{i=1}^{k} m_i\, f_i}{\sum_{i=1}^{k} f_i}, \quad s^2 = \frac{\sum_{i=1}^{k} (m_i - \overline{x})^2 f_i}{\left[\sum_{i=1}^{k} f_i\right] - 1}$$

dove

- k è il numero di intervalli
- m_i è il punto centrale dell'i-esimo intervallo
- f_i è la frequenza associata all'i-esimo intervallo

Nell' esempio di Tabella 1.2, la media raggruppata dei dati è

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{8} m_i\, f_i}{\sum_{i=1}^{8} f_i} = \\
&= \frac{1}{1067}[99.5(13) + 139.5(150) + 179.5(442) + \\
&+ 219.5(299) + 259.5(115) + 299.5(34) + \\
&+ 339.5(9) + 379.5(5)] = \\
&= 198.8\text{mg}/100\text{ ml}.
\end{aligned}$$

La varianza invece è

$$\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{8} (m_i - 198.8)^2 f_i}{[\sum_{i=1}^{8} f_i] - 1} = \\
&= \frac{1}{1067 - 1}[(-99.3)^2(13) + (-59.3)^2(150) + \\
&+ (-19.3)^2(442) + (20.7)^2(299) + \\
&+ (60.7)^2(115) + (100.7)^2(34) + \\
&+ (140.7)^2(9) + (180.7)^2(5)] = \\
&= 1930.9(\text{mg}/100\text{ ml})^2
\end{aligned}$$

La deviazione standard è

$$s = \sqrt{1930.9} = 43.9\text{mg}/100\text{ ml}.$$

1.3.4 Diseguaglianza di Chebychev

La media ci indica dove sono centrate le osservazioni, la deviazione standard indica quanto sono disperse rispetto alla media. Questo concetto può essere reso più preciso dalla *diseguaglianza di Chebychev*: per qualunque numero $k > 1$ almeno

$$1 - \left(\frac{1}{k}\right)^2$$

delle osservazioni in una serie di dati è compresa nell'intervallo

$$[(\overline{x} - ks), (\overline{x} + ks)].$$

Ad esempio, dato $k = 2$, almeno $1 - (\frac{1}{2})^2 = \frac{3}{4}$ delle osservazioni cadono nell'intervallo

$$[(\overline{x} - 2s), (\overline{x} + 2s)].$$

Possiamo quindi dire che $\overline{x} \pm 2s$ comprende almeno il 75% delle osservazioni. Questa affermazione è vera indipendentemente dai valori di $\overline{x}$ e di s.

Se $k = 3$ almeno $1 - (\frac{1}{3})^2 = \frac{8}{9}$ delle osservazioni cadono nell'intervallo

$$[(\overline{x} - 3s), (\overline{x} + 3s)],$$

quindi $\overline{x} \pm 3s$ contiene almeno l'88.9% delle osservazioni.

1.3.5 La scala logaritmica

Supponiamo di studiare un fenomeno rappresentato da una funzione esponenziale del tipo

$$y = Ka^x.$$

Passando ai logaritmi (decimali) la relazione divenata

$$\log y = \log(Ka^x) = \log K + x \log a.$$

Nel sistema di riferimento

$$\begin{cases} X = x \\ Y = \log y \end{cases}$$

il fenomeno è rappresentato da una funzione lineare di coefficiente angolare $m = \log a$.

Esempio. Sia $N(t)$ il numero di atomi radioattivi al tempo t. Si osserva sperimentalmente che gli atomi decadono secondo la legge

$$N(t) = N(0)a^{-t}, \quad t \geq 0, \quad a > 1,$$

dove a è una costante che dipende dal tipo di isotopo. Misurando il numero di atomi radioattivi agli istanti $t = 0, t = 1, t = 2, \ldots, t = 10$ e riportando sul grafico i valori di $N(t)$ si ottiene un andamento esponenziale. Riportando nel grafico i valori di

$$\log N(t)$$

si ottiene un andamento lineare. Infatti

$$\log N(t) = \log(N(0)a^{-t}) = \log N(0) - t \log a.$$

La retta $Y = \log N(t)$ ha coefficiente angolare $-\log a$.

Supponiamo di non conoscere il tipo di isotopo che stiamo analizzando, cioè supponiamo di non conoscere a.

Dal grafico della retta $Y = \log N(t)$ è facile ricavare il coefficiente angolare $-m\,(m > 0)$. Dalla relazione

$$-m = -\log a$$

ricaviamo la costante a che caratterizza il tipo di isotopo:

$$a = 10^m.$$

Se studiamo un fenomeno descritto dalla funzione potenza

$$y = Kx^n,$$

passando ai logaritmi (decimali) la relazione diventa

$$\log y = \log(Kx^n) = \log K + n \log x.$$

Nel sistema di riferimento

$$\begin{cases} X = \log x \\ Y = \log y \end{cases}$$

il fenomeno è rappresentato da una funzione lineare.

Capitolo 2

Probabilità

2.1 Generalità

La probabilità è il fondamento dell'*inferenza statistica*.

Il concetto base da cui parte la probabilità è quello di **evento**: esso è il risultato di un'osservazione o di un esperimento, o descrizione di un potenziale risultato (es. uscita di testa nel lancio di una moneta, uscita del numero 1 nel lancio di un dado, infarto a 50 anni)

Un evento si verifica oppure non si verifica. Gli eventi sono rappresentati con lettere maiuscole A,B,C....

- evento *intersezione* $A \cap B$: A e B si verificano *contemporaneamente* (es. A =essere alti più di 1.75 **e** B =essere italiani)
- evento *unione* $A \cup B$: *almeno uno* dei 2 eventi si verifica (es. A =essere alti più di 1.75 **o** B =essere italiani)
- evento *complementare* ${}^cA$: è l'evento *non* A (es. A =uscita del numero 1 nel lancio di un dado, ${}^cA$ =uscita di un numero *diverso* da 1)

Definiamo la **probabilità** di un evento A come la frequenza relativa con cui l'evento si verifica in una lunga serie di esperimenti tutti ripetuti in condizioni virtualmente identiche (definizione *frequentista*).

$$\frac{m}{n} \to P(A), \quad \text{per} \quad n \to +\infty,$$

dove m è il numero di volte in cui si verifica l'evento A, n è il numero di esperimenti eseguiti, $P(A)$ è la probabilità dell'evento A.

Vediamo quali sono le **proprietà della probabilità**:

1. $0 \leq P(A) \leq 1$.
2. Se l'evento A è *certo* allora $P(A) = 1$.

3. Se l'evento A è *impossibile* allora $P(A) = 0$. Indichiamo con $\emptyset$ l'evento impossibile.

4. Se un esperimento viene ripetuto n volte in condizioni identiche e l'evento A si verifica m volte, l'evento complementare ${}^cA$ si verifica $n - m$ volte. Allora

$$P({}^cA) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(A).$$

Definiamo **eventi mutuamente esclusivi** gli eventi che *non* possono verificarsi contemporaneamente (es. piove e non piove). Se A e B sono mutuamente esclusivi allora

$$A \cap B = \emptyset \quad \text{e} \quad P(A \cap B) = P(\emptyset) = 0.$$

Abbiamo i seguenti due principi per la probabilità:

1. **principio della somma della probabilità**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Se A e B sono mutuamente esclusivi allora $P(A \cap B) = 0$ e quindi

$$P(A \cup B) = P(A) + P(B).$$

In generale, se $A_1, A_2, \ldots, A_n$ sono tali che $A_i \cap A_j = \emptyset$, $i \neq j$, $i, j = 1, \ldots, n$, allora

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots P(A_n).$$

Esempio. Se in un lancio di un dado abbiamo A_1 =uscita del numero 1, A_2 =uscita del 2, A_3 =uscita del 3, allora

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) = \\ = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

2. **principio del prodotto delle probabilità**:

$$P(A \cap B) = P(A)\, P(B|A) = P(B)\, P(A|B),$$

dove $P(B|A)$ è la probabilità che si verifichi B *dato che* si è già verificato A (probabilità *condizionata*).

Esempio. A=soggetto ha 60 anni, B= tale soggetto vive fino a 65 anni

$A \cap B$=il soggetto è vivo sia a 60 che a 65 anni=il soggetto sopravvive fino a 65 anni

Secondo la tavola di sopravvivenza del 1998 per la popolazione degli USA si ha che

$$P(A) = 0.85331, \quad P(A \cap B) = 0.79123$$

Pertanto

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.79123}{0.85331} = 0.9272$$

Pertanto se un soggetto è vivo a 60 anni, la sua possibilità di sopravvivere fino a 65 anni è *maggiore* di quanto fosse alla nascita.

Diciamo che due sono **eventi indipendenti** quando il verificarsi di un evento non ha influenza sul verificarsi o non verificarsi dell'altro.

Se A e B sono indipendenti allora

$$P(A|B) = P(A)$$

e

$$P(B|A) = P(B).$$

$$\longrightarrow P(A \cap B) = P(A)\,P(B)$$

Esempio. Scegliamo a caso il nome di un farmacista dall'Albo professionale.

$$P(\text{maschio}) = P(m) = 0.6$$
$$P(\text{laureato in Italia}) = P(LI) = 0.8$$
$$\text{m e LI indipendenti} \leftarrow \begin{cases} P(m|LI) = P(m) \\ P(LI|m) = P(LI) \end{cases}$$
$$P(m \cup LI) = P(m) + P(LI) - P(m \cap LI) =$$
$$= 0.6 + 0.8 - (0.6 \cdot 0.8) =$$
$$= 1.4 - 0.48 = 0.92$$

NOTA BENE

Dire che 2 eventi sono mutuamente esclusivi NON EQUIVALE a dire che essi sono indipendenti.

Infatti, se A e B sono indipendenti e si verifica A allora l'evento B può verificarsi o no e $P(B|A) = P(B)$.

Se A e B sono mutuamente esclusivi e si verifica A allora B *non* può verificarsi e quindi $P(B|A) = 0$.

2.2 Distribuzioni di probabilità

Definiamo **variabile casuale (aleatoria)** una qualsiasi caratteristica che può essere misurata o categorizzata e che è soggetta alle leggi della probabilità. Una variabile aleatoria può essere di 2 tipi:

- **discreta**,quando può assumere solo un numero finito o un'infinità numerabile di valori (es. sesso, stato civile, regioni...)
- **continua**, quando può assumere qualunque valore nell'ambito di un certo intervallo (es. altezza, peso, pressione, concentrazione...)

Ogni variabile casuale ha una corrispondente *distribuzione di probabilità* che, utilizzando la teoria della probabilità, descrive il comportamento della variabile casuale stessa.

- *var. discreta*: la distribuzione di probabilità associa a *tutti* i possibili risultati della variabile casuale la probabilità che ciascuno di essi ha di verificarsi. La somma delle probabilità associate a tutti i possibili valori della variabile casuale deve essere uguale a 1.

  Esempio. $X =$ uscita di una faccia nel lancio di un dado

  Possibili valori di X: 1,2,3,4,5,6

  distribuzione di probabilità: $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, $P(X = 3) = \frac{1}{6}$,$P(X = 4) = \frac{1}{6}$,$P(X = 5) = \frac{1}{6}$,$P(X = 6) = \frac{1}{6}$

  La distribuzione di probabilità per una variabile discreta si può rappresentare con una tabella o graficamente con un diagramma a barre.

  Esempio. Supponiamo di voler conoscere la probabilità che un neonato selezionato casualmente sia il quartogenito. Dalla distribuzione di probabilità in Tabella 2.1 abbiamo

  $$P(X = 4) = 0.058 = 5.8\%.$$

  Supponiamo invece di voler conoscere la probabilità che un neonato sia il primogenito o il secondo genito. Allora applichiamo il principio della somma di eventi mutuamente esclusivi:

  $$\begin{aligned} P(X = 1 \cup X = 2) &= P(X = 1) + P(X = 2) = \\ &= 0.416 + 0.330 = \\ &= 0.746 = 74.6\%. \end{aligned}$$

- *var. continua*: la distribuzione di probabilità consente di determinare le probabilità associate a determinati range di valori.

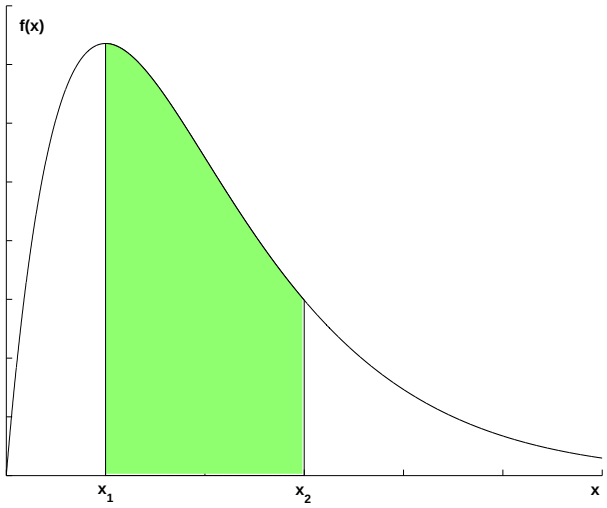
  Una distribuzione di probabilità discreta si può rappresentare graficamente con un diagramma a barre.

  Al crescere del numero di barre, l'ampiezza di ogni barra diventa più piccola e il diagramma tende a diventare una regione piana sottesa da una curva.

L'area totale sotto la curva è sempre uguale a 1.

La curva che sottende l'area unitaria rappresenta la distribuzione di probabilità di una variabile casuale *continua* e viene chiamata **densità di probabilità**. La si indica con $f(x)$.

La probabilità che la variabile continua X assuma un valore nell'intervallo compreso fra 2 risultati x_1 e x_2 è uguale all'**area** sottesa dalla densità di probabilità $f(x)$ nell'intervallo $[x_1, x_2]$.



Pertanto la probabilità associata ad un singolo valore di X è uguale a 0.

Esempio. $X =$ peso alla nascita di un neonato in grammi

La distribuzione di probabilità ci permette, ad esempio, di determinare la probabilità che un neonato abbia un peso compreso fra 2500 e 3000 grammi: $P(2500 < X < 3000)$ oppure la probabilità che un neonato pesi meno di 2400 grammi: $P(X \leq 2400)$ e così via.

2.2.1 Funzione di distribuzione cumulativa

La funzione di distribuzione cumulativa si indica con $F(x)$ e rappresenta la probabilità che la variabile casuale X (discreta o continua) assuma un valore minore o uguale a x:

$$F(x) = P(X \leq x)$$

Esempio. Sia X una variabile casuale continua con funzione di densità di probabilità $f(x)$. Allora

$$F(x) = P(X \leq x)$$

è l'area sottesa dalla densità di probabilità f dall'estremo sinistro della distribuzione fino al valore x di X. Se $x_1 \leq x_2$ abbiamo

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1).$$

Ricordando che $f(x)$ sottende un'area pari a 1, si ha che

$$P(X \geq x) = 1 - F(x).$$

2.2.2 Distribuzioni di frequenza e distribuzioni di probabilità

Abbiamo i seguenti fatti:

- Per un campione di osservazioni una distribuzione di frequenza mostra ogni risultato e la sua frequenza
- Per una variabile casuale discreta, una distribuzione di probabilità discreta elenca ogni valore possibile con la relativa probabilità
- La probabilità rappresenta la frequenza relativa di un evento in numerosi esperimenti ripetuti in condizioni sostanzialmente identiche. La distribuzione di probabilità dell'ordine di nascita dei bambini in Tabella 2.1 è stata ricavata in base a dati rilevati su una popolazione in un dato anno.
- Come per le distribuzioni di frequenza, possiamo descrivere una distribuzione di probabilità utilizzando una misura di tendenza centrale e una misura di dispersione.
- Il valor medio di una variabile casuale è la media μ della popolazione, la dispersione dei valori rispetto a tale media è la varianza σ^2 della popolazione.
- Si possono determinare le distribuzioni di probabilità di molte variabili di interesse in base a considerazioni teoriche. Queste distribuzioni di probabilità si dicono *distribuzioni teoriche di probabilità*.

2.2.3 Distribuzione binomiale o di Bernoulli

Consideriamo una variabile casuale discreta dicotomica. Indichiamo i due possibili valori mutuamente esclusivi (es: vita-morte, maschio-femmina, testa-croce) come **successo** e **insuccesso**. Si chiama *variabile casuale di Bernoulli*.

Chiamiamo *esperimento di Bernoulli* un esperimento che dà due possibili risultati (successo o insuccesso).

Consideriamo una successione di n esperimenti *indipendenti* di Bernoulli, ciascuno dei quali avente una probabilità di *successo* p, e la variabile casuale

X definita come il numero di successi in n esperimenti di Bernoulli. La distribuzione di probabilità della variabile casuale X è la *distribuzione di Bernoulli*.

I numeri n e p sono detti i *parametri* della distribuzione di Bernoulli.

Esempio. Lanciamo 2 volte un dado ($n = 2$) e sia *successo*=uscita del numero 6. Detta X la variabile casuale che conta il numero di successi,

$$P_2(X = 0) = P(\neq 6 \cap \neq 6) = (1-p)^2 = \left(\frac{5}{6}\right)^2 = \frac{25}{36}$$

(applicando il principio del prodotto per eventi indipendenti).

Poi

$$P_2(X = 1) = P((6 \cap \neq 6) \cup (\neq 6 \cap 6)) =$$
$$= p(1-p) + (1-p)p = \frac{1}{6}\frac{5}{6} + \frac{5}{6}\frac{1}{6} = \frac{10}{36}$$

(applicando il principio del prodotto per eventi indipendenti e il principio della somma per eventi mutuamente esclusivi).

Infine

$$P_2(X = 2) = P(6 \cap 6) = p^2 = \left(\frac{1}{6}\right)^2 = \frac{1}{36}.$$

Sommando tutte le probabilità otteniamo 1:

$$P_2(X = 0) + P_2(X = 1) + P_2(X = 2) = \frac{25}{36} + \frac{10}{36} + \frac{1}{36} = 1.$$

Le 3 **ipotesi fondamentali** per la distribuzione binomiale sono le seguenti:

1. esiste un numero fisso n di esperimenti, ognuno dei quali dà luogo ad uno dei 2 risultati *mutuamente esclusivi*;
2. i risultati degli n esperimenti sono *indipendenti*;
3. la probabilità di successo p è costante per *ciascun* esperimento.

La probabilità che la variabile casuale binomiale X assuma il valore (intero) x (cioè che si verifichino esattamente x successi) è:

$$P_n(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

dove

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

è detto *coefficiente binomiale* e

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$$

è detto n *fattoriale* (rappresenta il numero di *permutazioni* di n oggetti). Per convenzione, $0! = 1$.

Osservazioni.

1. X può assumere qualsiasi valore da 0 a n.
2. il coefficiente binomiale $\binom{n}{x}$ rappresenta il numero di *combinazioni* di n oggetti presi a gruppi di x, è dunque il numero di modi in cui posso selezionare x oggetti da una gruppo di n, *senza considerarne l'ordine*.

Nel caso della **distribuzione binomiale** inoltre abbiamo che

- **media** $\mu = np$: è il numero *medio* di successi in n esperimenti di Bernoulli.
- **varianza** $\sigma^2 = np(1-p)$.

Esempio. La probabilità che un individuo estratto da una popolazione sia fumatore è $p = 0.29$. Se selezioniamo campioni ripetuti di dimensione $n = 10$, il numero *medio* di fumatori per campione è

$$\mu = np = 10(0.29) = 2.9,$$

mentre la deviazione standard è

$$\sigma = \sqrt{np(1-p)} = \sqrt{10(0.29)(0.71)} = \sqrt{2.059} = 1.4.$$

Esempio. Calcolare la probabilità che un paziente punto con un ago infetto da virus dell'epatite B sviluppi realmente la malattia.

Sia X la variabile casuale che conta il numero di pazienti infetti. Poichè sono risultati mutuamente esclusivi ed esaustivi, X è una variabile di Bernoulli.

Selezioniamo 5 soggetti dalla popolazione di pazienti punti con un ago infetto. Il numero dei pazienti in questo campione che svilupperà la malattia è una variabile casuale binomiale con parametri $n = 5$ e $p = 30\%$ (in base ad un indagine statistica).

La probabilità che, ad esempio, esattamente 2 pazienti sviluppino la malattia è

$$\begin{aligned} P(X = 2) &= \binom{5}{2} 0.30^2 \, (1 - 0.30)^{5-2} = \\ &= 0.309 \approx 31\%. \end{aligned}$$

La probabilità che almeno 3 individui tra i 5 sviluppino la malattia è

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) + P(X = 5) = \\ &= 0.132 + 0.028 + 0.003 = \\ &= 0.163 = 16.3\% \end{aligned}$$

oppure potremmo anche calcolarla come

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \end{aligned}$$

La probabilità che **al massimo** un paziente sviluppi la malattia è

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) = \\ &= 0.168 + 0.360 = \\ &= 0.528 = 52.8\% \end{aligned}$$

oppure potremmo anche calcolarla come

$$\begin{aligned} P(X \leq 1) &= 1 - P(X > 1) = \\ &= 1 - P(X = 2) - P(X = 3) - \\ &\quad - P(X = 4) - P(X = 5) \end{aligned}$$

2.2.4 Distribuzione di Poisson

Se n è *molto grande* e p è *molto piccola* la distribuzione binomiale è approssimata da quella di Poisson

Essa modellizza eventi *discreti* che si verificano **raramente** nel tempo o nello spazio (*distribuzione degli eventi rari*).

Sia X la variabile casuale che conta il numero di volte in cui un evento si verifica in un certo intervallo di tempo. Allora X varia da 0 a $+\infty$.

Definiamo il parametro

$$\begin{aligned} \lambda = &\text{ numero } \textit{medio} \text{ di volte in cui si verifica} \\ &\text{l'evento in un certo intervallo di tempo} \end{aligned}$$

detto *parametro di Poisson.*

L'**espressione matematica** della distribuzione di Poisson è la seguente:

$$P(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

dove $e = 2.71828\ldots$ è la base dei logaritmi naturali.

Le 3 **ipotesi fondamentali** per la distribuzione di Poisson sono le seguenti:

1. la probabilità che un singolo evento di verifichi in un determinato intervallo di tempo è proporzionale alla lunghezza dell'intervallo;
2. teoricamente, in un singolo intervallo di tempo è possibile che l'evento si verifichi un numero *infinito* di volte;
3. gli eventi si verificano *indipendentemente* nello stesso intervallo di tempo e tra intervalli consecutivi.

Nella distribuzione di Poisson si ha che

$$\text{media} = \text{varianza} = np = \lambda$$

Esempio. Calcoliamo la probabilità che x individui siano coinvolti in un incidente d'auto su una popolazione di 10000 individui nell'arco di un anno sapendo che $p = 0.00024$.

Avremo

$$\lambda = np = 10000(0.00024) = 2.4.$$

Allora

$$P(X = 0) = \frac{e^{-2.4}(2.4)^0}{0!} = 0.0907$$
$$P(X = 1) = \frac{e^{-2.4}(2.4)^1}{1!} = 0.2177$$

La probabilità che *almeno* 2 individui siano coinvolti è allora

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = \\ &= 1 - [P(X = 0) + P(X = 1)] = \\ &= 1 - 0.0907 - 0.2177 = \\ &= 0.6916 = 69.19\% \end{aligned}$$

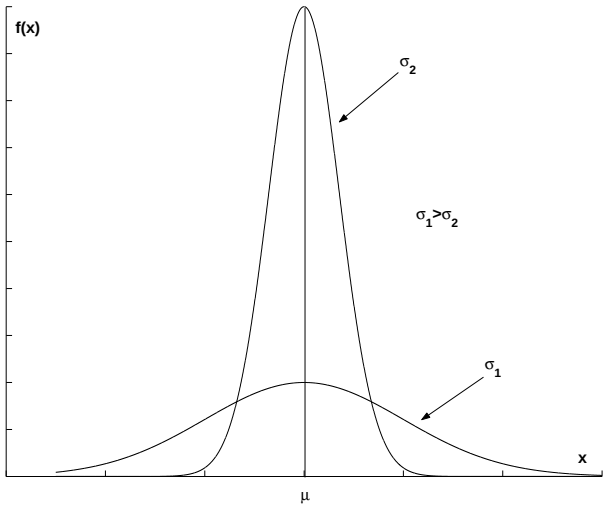
2.2.5 Distribuzione normale

La distribuzione continua di probabilità più comune è la **distribuzione normale**, la cui forma è quella di una binomiale con $p = \frac{1}{2}$ e n che tende all'infinito (Figura 2.2).

La sua densità di probabilità è data dalla funzione

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

dove μ è la *media* di X e σ^2 è la varianza di X.



In questa distribuzione (simmetrica rispetto alla media μ)

$$\text{media=moda=mediana}$$

Poichè una distribuzione normale può avere un numero infinito di valori al variare dei 2 parametri μ e σ, è impossibile tabulare le aree associate ad ogni singola curva. Pertanto è tabulata una sola curva, quella in cui $\sigma = 1$ e $\mu = 0$. Si chiama distribuzione **normale standardizzata** e la indichiamo con Z:

$$Z = \frac{X - \mu}{\sigma},$$

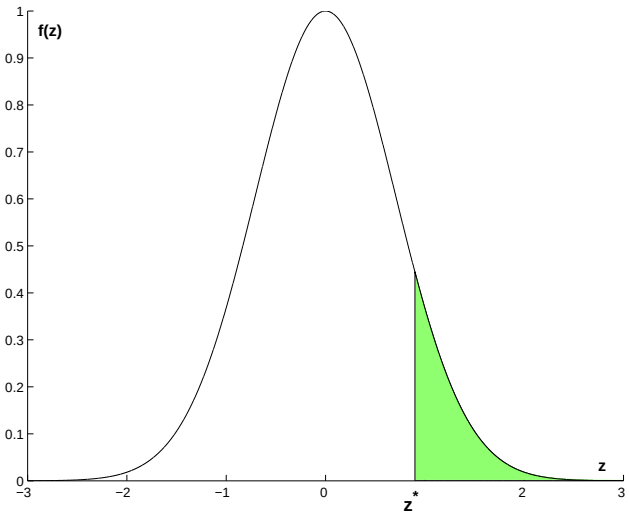
dove μ e σ sono media e deviazione standard della distribuzione normale X. La sua densità di probabilità è allora

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

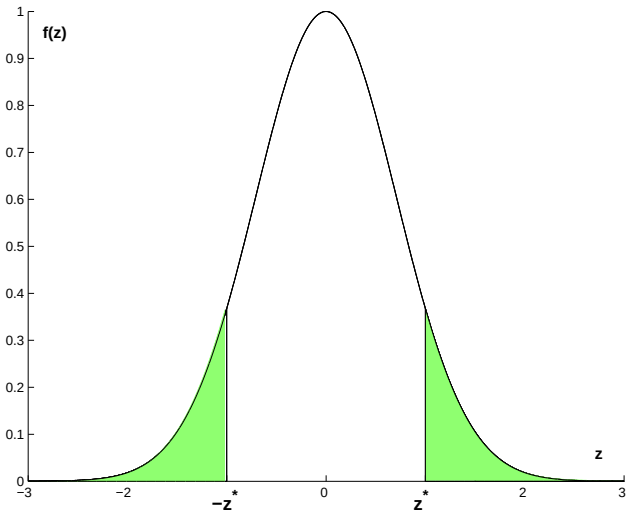
Avremo:

$$P(X \leq x^*) = P\left(Z \leq \frac{x^* - \mu}{\sigma}\right)$$
$$P(x_1 < X \leq x_2) = P\left(\frac{x_1 - \mu}{\sigma} < Z \leq \frac{x_2 - \mu}{\sigma}\right)$$
$$P(X > x^*) = P\left(Z > \frac{x^* - \mu}{\sigma}\right).$$

La Tabella A.3 riporta le aree in un lato della distribuzione, cioè l'area sottesa dalla curva $f(z)$ a destra di $z = z^*$, al variare di z^*:



La curva è simmetrica rispetto a $z = 0$, quindi l'area a destra di $z = z^*$ è uguale all'area a sinistra di $z = -z^*$:



Esempi.

1. Le altezze degli uomini di un certo paese sono distribuite normalmente con media $\mu = 173.6$ cm e varianza $\sigma^2 = 40.96$ cm^2. Si vuole calcolare la probabilità che un uomo scelto a caso

   (a) sia più alto di 187 cm

   (b) sia alto meno di 166.9 cm

   (c) abbia altezza compresa fra 170 e 180 cm.

   Abbiamo (usando la Tabella A.3):

   (a) $P(X > 187) = P\left(Z > \frac{187-173.6}{6.4}\right) = P(Z > 2.09) = 0.018.$

   (b) $P(X \leq 166.9) = P\left(Z \leq \frac{166.9-173.6}{6.4}\right) = P(Z \leq -1.05) = P(Z > 1.05) = 0.147.$

   (c) $P(170 < X \leq 180) =$
   $P\left(\frac{170-173.6}{6.4} < Z \leq \frac{180-173.6}{6.4}\right) =$
   $= P(-0.56 < Z \leq 1) =$
   $= 1 - P(Z > 0.56) - P(Z > 1) =$
   $= 1 - 0.288 - 0.159 = 0.553.$

2. Nella curva normale standardizzata

   (a) quale valore di z lascia una probabilità uguale a 0.10 nella coda di destra?

   (b) quale valore di z lascia una probabilità uguale a 0.05 nella coda di sinistra?

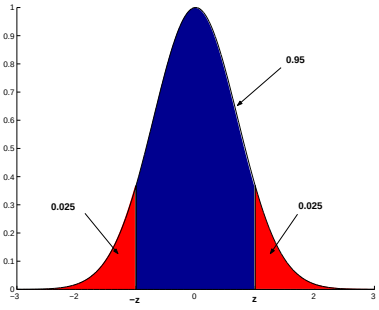
   (c) quali valori di z, simmetrici rispetto all'origine, comprendono una probabilità pari a 0.95?

Si ha (dalla Tabella A.3):

(a) $z = 1.28$

(b) $z = 1.64$ e quindi $-z = -1.64$.

(c) Abbiamo:



Dunque $z = \pm 1.96$.

3. Le altezze degli uomini di un certo paese sono distribuite normalmente con media $\mu = 173.6$ cm e varianza $\sigma^2 = 40.96$ cm^2. Per quale altezza si ha che

   (a) la probabilità di trovare un'altezza maggiore è uguale a 0.1?

   (b) la probabilità di trovare un'altezza minore è uguale a 0.01?

   Risolviamo il problema.

   (a) Cerchiamo sulla Tabella A.3 il valore di z che lascia alla sua destra una probabilità uguale a 0.1: $z = 1.28$. Dalla relazione

   $$\frac{x - 173.6}{6.4} = 1.28$$

   ricaviamo

   $$x = 1.28(6.4) + 173.6 = 181.79\,\text{cm}$$

   (b) Cerchiamo sulla Tabella A.3 il valore di z che lascia alla sua destra una probabilità uguale a 0.01; per simmetria il valore $-z$ lascerà alla sua sinistra una probabilità uguale a 0.01. Troviamo $z = 2.34$ e quindi $-z = -2.34$. Abbiamo così

   $$\frac{x - 173.6}{6.4} = -2.34 \Rightarrow x = -2.34(6.4) + 173.6 = 158.63\,\text{cm}.$$

Capitolo 3

Campionamento

Nello studio delle distribuzioni teoriche di probabilità si suppone di conoscere i principali parametri della popolazione (ad esempio la media). Nelle applicazioni, i valori di questi parametri non sono noti. Occorre perciò descrivere una popolazione utilizzando le informazioni contenute in un campione di osservazioni. Il processo attraverso il quale si traggono conclusioni su un'intera popolazione in base ad una campione si chiama *inferenza statistica*.

Il **problema** che vogliamo risolvere è il seguente: stimare la media μ di X variabile casuale quantitativa (es. pressione arteriosa sistolica di maschi di 30-40 anni che svolgono una certa attività lavorativa).

Possiamo utilizzare la media $\overline{x}$ di un campione estratto dalla popolazione come *stima* per la media μ della popolazione. Perchè $\overline{x}$ sia una buona approssimazione di μ occorre che il campione sia rappresentativo della popolazione in esame e che la dimensione del campione sia sufficientemente grande. Si dice che $\overline{x}$ è uno *stimatore* del parametro μ. Anzi, si può dimostrare che esso è lo *stimatore di massima verosimiglianza* se la popolazione da cui è estratto il campione è distribuita normalmente.

Estraiamo campioni *casuali* di n valori di X, che avranno medie $\overline{x}_1, \overline{x}_2, \overline{x}_3 \ldots$. Allora si genera una nuova variabile casuale $\overline{X}$: se ciascuna di queste medie campionarie è considerata come una singola osservazione, la distribuzione di probabilità di queste medie si chiama *distribuzione della media campionaria* di campioni di dimensione n.

Nelle applicazioni non si selezionano campioni ripetuti di dimensione n da una popolazione, ma la conoscenza della distribuzione della media campionaria consente di fare inferenze in base ad un singolo campione di dimensione n.

La variabilità di $\overline{X}$ dipende da

1. σ (più la pressione arteriosa varia rispetto alla media nella popolazione e maggiore è l'aumento della variabilità delle medie dei campioni di dimensione n);

2. n (più grandi sono i campioni casuali e maggiormente vicini tra loro sono i valori di $\overline{x}$).

I 3 risultati di base per la distribuzione di $\overline{X}$

Supponiamo che la distribuzione di probabilità di una popolazione, o di una variabile casuale quantitativa, abbia media μ e deviazione standard σ. Allora:

1. la media della distribuzione della media campionaria coincide con la media μ della popolazione;

2. la varianza della media campionaria è uguale alla varianza della popolazione *divisa* per la dimensione n del campione $\frac{\sigma^2}{n}$
   La quantità $\frac{\sigma}{\sqrt{n}}$ viene chiamata *errore standard*.

3. anche se la distribuzione X *NON* è normale, la distribuzione di $\overline{X}$ si avvicina sempre più alla normale con media μ e varianza $\frac{\sigma^2}{n}$ *al crescere di* n (questo è l'enunciato del cosiddetto **Teorema del limite centrale**).

NOTA BENE

- Il punto 2. dice che c'è minore dispersione rispetto alla media tra le medie campionarie che tra le singole osservazioni. Inoltre al crescere di n diminuisce la variabilità tra le medie campionarie.

- Il punto 3. dice che, se n è sufficientemente grande, la distribuzione della media campionaria è approssimativamente normale. Più la popolazione originaria si allontana dalla normale, maggiore sarà il valore di n necessario ad assicurare che la distribuzione della media campionaria sia normale con media μ e deviazione standard $\frac{\sigma}{\sqrt{n}}$, che si chiama *errore standard*.

- La variabile
$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
è normalmente distribuita con media 0 e deviaz. standard 1.

Esempio. Consideriamo la distribuzione dei livelli di colesterolo in individui di età compresa fra i 20 ei 74 anni. La media della popolazione è $\mu = 211$ mg/100 ml e la deviazione standard è $\sigma = 46$ mg/100 ml. Selezioniamo campioni ripetuti di dimensione $n = 25$ dalla popolazione.

1. Quale proporzione dei campioni avrà un valor medio $\overline{x}$ superiore a 230 mg/100 ml?

Dal Teorema del limite centrale sappiamo che la distribuzione della media campionaria per campioni di dimensione $n = 25$ è normale con media $\mu = 211$ mg/100 ml e errore standard $\frac{\sigma}{\sqrt{n}} = \frac{46}{5} = 9.2$ mg/100 ml. La variabile

$$Z = \frac{\overline{X} - 211}{9.2}$$

è una normale standardizzata. Se $\overline{x} = 230$ allora

$$z = \frac{230 - 211}{9.2} = 2.07.$$

Dalla Tabella A.3, l'area a destra di $z = 2.07$ è 0.019. Quindi circa l'1.9% dei campioni di dimensione 25 avrà una media maggiore o uguale a 230 mg/100 ml.

2. Quale valore medio campionario $\overline{x}$ delimita il 10% inferiore della distribuzione campionaria?

   Dalla Tabella A.3, $z = 1.28$ delimita un'area pari a 0.1 nella coda di destra; allora, per simmetria, si ha che $z = -1.28$ delimita un'area pari a 0.1 nella coda di sinistra. Da

   $$-1.28 = \frac{\overline{x} - 211}{9.2}$$

   otteniamo

   $$\overline{x} = 211 + (-1.28)(9.2) = 199.2.$$

   Quindi il 10% dei campioni di dimensione 25 ha medie minori o uguali a 199.2 mg/100 ml.

3. Quale intervallo, simmetrico rispetto alla media μ, comprende il 95% delle medie campionarie $\overline{x}$ per campioni di dimensione 25?

   Dalla Tabella A.3, il valore $z = 1.96$ stacca nella coda destra un'area pari a 0.025. Per simmetria allora

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

   Da

   $$-1.96 \leq \frac{\overline{x} - 211}{9.2} \leq 1.96$$

   ricaviamo l'intervallo per $\overline{x}$:

   $$\begin{aligned} &211 - 1.96(9.2) \leq \overline{x} \leq 211 + 1.96(9.2) \Rightarrow \\ &\Rightarrow 193.0 \leq \overline{x} \leq 229.0 \end{aligned}$$

   Dunque circa il 95% delle medie campionarie di dimensione 25 è compreso fra 193.0 e 229.0 mg/100 ml. Se selezioniamo un campione di dimensione 25 con media minore di 193 o superiore a 229 possiamo dedurre che esso è stato estratto da un'altra popolazione oppure che si è verificato un *evento raro*.

4. Quale deve essere la dimensione n del campione affinchè il 95% delle medie campionarie $\overline{x}$ sia compreso nell'intervallo $[\mu - 5\,\text{mg}/100\text{ ml}, \mu + 5\,\text{mg}/100\text{ ml}]$?

   Dobbiamo trovare n per cui

   $$P(\mu - 5 \leq \overline{x} \leq \mu + 5) = 0.95$$

   $\Leftrightarrow$

   $$P\left(\frac{-5}{\frac{\sigma}{\sqrt{n}}} \leq Z \leq \frac{5}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

   Dalla Tavola A.3 sappiamo che il 95% dell'area sottesa dalla curva normale standardizzata è compreso fra $z = -1.96$ e $z = 1.96$. Allora

   $$1.96 = \frac{5}{\frac{46}{\sqrt{n}}}$$

   $\Rightarrow$

   $$n = \left(\frac{46}{5} \cdot 1.96\right)^2 = 325.2.$$

   La dimensione cercata è quindi $n = 326$.

5. Quale valore di $\overline{x}$ è il limite superiore per il 95% dei livelli medi di colesterolo di campioni di dimensione 25?

   Dalla Tabella A.3 il valore $z = 1.64$ delimita nella coda di destra un'area pari a 0.05. Quindi

   $$P(Z \leq 1.64) = 0.95$$

   da cui

   $$\begin{aligned} &P\left(\frac{\overline{X} - 211}{9.2} \leq 1.64\right) = \\ &= P(\overline{X} \leq (1.64)(9.2) + 211) = \\ &= P(\overline{X} \leq 226.08) = 0.95. \end{aligned}$$

   Il valore cercato è circa $\overline{x} = 226.1$, quindi circa il 95% dei campioni di dimensione 25 ha medie minori o uguali a 226.1 mg/100 ml.

   Se volessimo il limite inferiore per il 95% dei livelli medi di colesterolo, ci interesserebbero i valori $z \geq -1.64$. Allora otterremmo

   $$\overline{x} = (-1.64)(9.2) + 211 = 195.92.$$

Capitolo 4

Inferenza sulle medie

Supponiamo di non conoscere le caratteristiche della popolazione (media, varianza...). Allora estraiamo un campione casuale da essa. A questo punto, utilizzando le nostre conoscenze di teoria campionaria, desideriamo fare sulla popolazione tutte le inferenze possibili, sulla base della osservazione sul singolo campione casuale estratto.

4.1 Intervalli di confidenza

Vediamo un esempio di **problematica**: Quanto è maggiore l'efficacia di un nuovo farmaco rispetto ad un trattamento precedente, per coloro che sono affetti da una certa patologia?

I **metodi** che possiamo utilizzar sono:

1. *Stima puntuale*: calcolo un singolo numero per stimare il parametro in esame (es. la media). Essa però non fornisce alcuna informazione circa l'accuratezza della stima (es. non sappiamo quanto $\overline{x}$ è vicino a μ).

2. *Stima intervallare*: fornisce un intervallo di possibili valori entro cui si ritiene sia compreso il parametro in esame con un certo *grado di confidenza*. E' questo il concetto di intervallo di confidenza.

Il grado di confidenza più utilizzato è quello del 95%.

NOTA BENE

Dire che l'intervallo di confidenza contiene il valore del parametro sconosciuto della popolazione con una probabilità (=grado di confidenza) del 95% **NON** SIGNIFICA dire che il valore ignoto della popolazione ha una probabilità del 95% di rientrare nell'intervallo (infatti, il valore del parametro della popolazione non è una variabile casuale) **BENSI'** SIGNIFICA dire che,

selezionando 100 campioni casuali dalla popolazione ed utilizzando questi campioni per calcolare 100 diversi intervalli di confidenza, circa 95 intervalli conterranno il parametro *reale* della popolazione e 5 no.

Stimiamo la **media** μ della popolazione.

σ **nota**

Se X è una var. casuale normale con media μ e deviazione standard σ, allora per qualunque n

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

è una var. *normale* standardizzata.

Se X non segue la distribuzione normale, allora Z è una variabile normale standardizzata *solo* se n è abbastanza grande.

4.1.1 Intervallo di confidenza bilaterale

Dalla Tabella A.3 sappiamo che il 95% delle osservazioni è compreso fra -1.96 e 1.96:

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \longrightarrow$$
$$P\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Dunque

$$\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$

è l'intervallo di confidenza *bilaterale* al 95% per la media μ della popolazione, *nota* la deviazione standard σ.

Si considerano anche altri intervalli di confidenza, ad esempio al 99%. Dalla Tabella A.3 il 99% delle osservazioni in una distribuzione normale standardizzata è compreso fra -2.58 e 2.58. Allora l'intervallo di confidenza per μ al 99% è

$$\left(\overline{X} - 2.58\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.58\frac{\sigma}{\sqrt{n}}\right).$$

Osserviamo che questo intervallo è più ampio dell'intervallo di confidenza al 95%. Infatti meno ampio è l'intervallo, meno confidenti siamo che la media μ vi cada all'interno. Se vogliamo restringere l'intervallo senza ridurre il grado di confidenza abbiamo bisogno di maggiori informazioni su μ; dobbiamo quindi selezionare un campione di dimensione n maggiore.

Al crescere di n l'errore standard sulla media campionaria $\frac{\sigma}{\sqrt{n}}$ diminuisce e questo implica un intervallo di confidenza di minore ampiezza.

Esempio. Consideriamo la distribuzione dei livelli di colesterolo della popolazione maschile di ipertesi fumatori. E' una distribuzione normale con media μ sconosciuta e deviazione standard $\sigma = 46$ mg/100 ml. Supponiamo di selezionare un campione casuale di dimensione $n = 12$. La media di tale campione è $\overline{x} = 217$ mg/100 ml. L'intervallo di confidenza al 95% per la media μ è

$$\left(217 - 1.96\frac{46}{\sqrt{12}}, 217 + 1.96\frac{46}{\sqrt{12}}\right) = (191, 243).$$

L'ampiezza dell'intervallo è $243 - 191 = 52$ mg/100 ml.

Siamo confidenti al 95% che questi limiti comprendano la media μ, cioè il reale livello medio di colesterolo degli ipertesi fumatori. **NON** diciamo che c'è una probabilità pari a 0.95 che μ sia compresa fra 191 e 243, poichè il valore di μ è fisso e può essere o meno compreso fra 191 e 243.

L'intervallo di confidenza al 99% invece sarà

$$\left(217 - 2.58\frac{46}{\sqrt{12}}, 217 + 2.58\frac{46}{\sqrt{12}}\right) = (183, 251).$$

L'ampiezza di tale intervallo è $251 - 183 = 68$ mg/100 ml.

Ci chiediamo ora quanto dovrebbe essere grande la dimensione n del campione per ridurre l'ampiezza dell'intervallo di confidenza al 99% a 20 mg/100 ml.

L'intervallo di confidenza al 99% è

$$\left(\overline{X} - 2.58\frac{46}{\sqrt{n}}, \overline{X} + 2.58\frac{46}{\sqrt{n}}\right).$$

La sua ampiezza è uguale a $2 \cdot 2.58\frac{46}{\sqrt{n}}$. Allora deve essere

$$2 \cdot 2.58\frac{46}{\sqrt{n}} = 20$$

da cui

$$\sqrt{n} = 2.58\frac{46}{10} = 11.868 \rightarrow n = 140.8.$$

Dobbiamo dunque selezionare un campione di dimensione $n = 141$ individui.

4.1.2 Intervallo di confidenza unilaterale

Supponiamo di essere interessati solo al livello *superiore* (analogamente si può ragionare per il livello inferiore) per la media μ della popolazione. Dalla Tabella A.3 rileviamo che il 95% delle osservazioni giace al di sopra di -1.645:

$$P(Z \geq -1.645) = 0.95 \longrightarrow$$
$$P\left(\mu \leq \overline{X} + 1.645\,\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Dunque

$$\overline{X} + 1.645\,\frac{\sigma}{\sqrt{n}}$$

è il limite *superiore* dell'intervallo di confidenza *unilaterale* al 95%.

Esempio. Il livello di emoglobina nei bambini al di sotto dei 6 anni esposti ad elevati livelli di piombo è distribuita normalmente con media μ sconosciuta e deviazione standard $\sigma = 0.85$ g/100 ml. Sappiamo che i bambini intossicati da piombo hanno un livello di emoglobina generalmente molto più basso rispetto ai bambini sani. Siamo quindi interessati al livello *superiore* per μ. Selezioniamo un campione di $n = 74$ bambini esposti ad elevati livelli di piombo. La media è $\overline{x} = 10.6$ g/100 ml. In base a questo campione l'intervallo di confidenza unilaterale al 95% per μ è

$$\mu \leq 10.6 + 1.645\frac{0.85}{\sqrt{74}} = 10.8.$$

Siamo confidenti al 95% che 10.8 g/100 ml sia superiore al reale livello medio di emoglobina nei bambini intossicati dal piombo. Supponiamo di conoscere il livello medio μ_s di emoglobina nei bambini sani. Se $\mu_s \leq 10.8$ il campione di bambini analizzato è un campione di bambini sani. Se $\mu_s > 10.8$ il campione analizzato è di bambini intossicati.

σ ignota

Sostituiamo la deviazione standard σ con la deviazione standard campionaria s:

$$t_{n-1} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

t_{n-1} segue una distribuzione campionaria *prossima* alla distribuzione normale standardizzata per *n grande*, perchè allora s approssima bene σ. Quando n è piccolo, s può differire considerevolmente da σ e ciò fa sì che t_{n-1} abbia una variabilità casuale sostanzialmente più grande di Z.

t_{n-1} segue la cosiddetta *distribuzione t di Student con* $n-1$ *gradi di libertà* (vedi Tabella A.4).

Osservazioni.

1. t è simmetrica con code più spesse rispetto alla normale.

2. I *gradi di libertà* misurano la quantità di informazione disponibile nei dati per approssimare σ con s.

3. Pochi gradi di libertà implicano una maggior dispersione.

4. Molti gradi di libertà determinano una vicinanza alla normale, perchè per n grande s diventa una stima sempre più affidabile di σ (per $n \geq 90$ si può sostituire la t di Student con la normale).

5. La distribuzione t di Student è strettamente valida **solo se** la distribuzione di X è *normale*.

6. t è *robusta* nel senso che è approssimativamente valida anche per marcate deviazioni dalla normalità.

Esempio. Consideriamo un campione di $n = 10$ bambini selezionato fra la popolazione di neonati cui viene somministrato un medicinale contenente alluminio. Non conosciamo la media μ e la deviazione standard σ dei livelli di alluminio plasmatico di questa popolazione. Sappiamo che la media del campione estratto è $\overline{x} = 37.2\,\mu$g/l e la deviazione standard campionaria è $s = 7.13\ \mu$g/l. Calcoliamo l'intervallo di confidenza al 95% per la media μ utilizzando la distribuzione t di Student. Dalla Tabella A.4, essendo il numero di gradi di libertà $10 - 1 = 9$, abbiamo che il 95% delle osservazioni cade nell'intervallo $(-2.262, 2.262)$. Pertanto l'intervallo di confidenza al 95% per la media μ è

$$\begin{aligned}
&\left(\overline{x} - 2.262\frac{s}{\sqrt{n}}, \overline{x} + 2.262\frac{s}{\sqrt{n}}\right) = \\
&= \left(37.2 - 2.262\frac{7.13}{\sqrt{10}}, 37.2 + 2.262\frac{7.13}{\sqrt{10}}\right) = \\
&= (32.1, 42.3).
\end{aligned}$$

Siamo confidenti al 95% che questo intervallo contenga il livello medio reale di alluminio plasmatico.

Si può anche calcolare l'intervallo di confidenza al 99%. In tal caso, dalla Tabella A.4 con 9 g.d.l., si ha che i valori -3.250 e 3.250 comprendono il 99% delle osservazioni.

Quindi l'intervallo di confidenza al 99% per μ è

$$\begin{aligned}
&\left(\overline{x} - 3.25\frac{s}{\sqrt{n}}, \overline{x} + 3.25\frac{s}{\sqrt{n}}\right) = \\
&= \left(37.2 - 3.25\frac{7.13}{\sqrt{10}}, 37.2 + 3.25\frac{7.13}{\sqrt{10}}\right) = \\
&= (29.87, 44.53).
\end{aligned}$$

4.2 Test d'ipotesi (test di significatività)

Vediamo un esempio di **problematica**:Un nuovo farmaco può portare dei miglioramenti ad una certa patologia rispetto ai farmaci esistenti?

Concentriamo ancora l'attenzione sul problema di stimare la **media** μ della popolazione.

Formuliamo l'*ipotesi nulla* H_0: la media della popolazione μ è uguale ad un valore postulato μ_0.

Il test d'ipotesi nell'inferenza statistica consiste nel trarre una delle 2 seguenti conclusioni:

1. si rifiuta l'ipotesi nulla H_0. Allora μ_0 **NON** è la media della popolazione;

2. *non* si rifiuta H_0. Allora μ_0 può essere considerata la media della popolazione.

Si giunge ad una di queste 2 conclusioni analizzando i risultati di un campione di dimensione n e confrontando la media campionaria $\overline{x}$ con μ_0.

La **domanda** che ci si pone è la seguente: se la media della popolazione è μ_0, qual è la probabilità che un campione abbia una media campionaria $\overline{x}$ che si scosta da μ_0 per un ammontare pari o maggiore a quello della $\overline{x}$ osservata?

Queste sono le possibili **risposte**:

- se questa probabilità è "sufficientemente piccola" vi è ragione di credere che la media campionaria osservata $\overline{x}$ *non* sia plausibile. Pertanto l'ipotesi nulla H_0 deve essere **rifiutata**. Questo risultato del test è detto *statisticamente significativo*;

- se questa probabilità non è "sufficientemente piccola" allora la media campionaria osservata $\overline{x}$ è un risultato plausibile e l'ipotesi nulla H_0 **non** viene **rifiutata**.

La probabilità "sufficientemente piccola" si denota in genere con α e determina il **livello di significatività** del test. Di solito si utilizza $\alpha = 0.05$ oppure $\alpha = 0.01$.

La probabilità che un campione abbia una media campionaria che si scosta da μ_0 per un ammontare pari o maggiore a quello della $\overline{x}$ osservata si indica con p e si chiama **valore p del test**.

- se $p \leq \alpha$ **rifiutiamo** H_0

- se $p > \alpha$ **non rifiutiamo** H_0

La probabilità p è data dalle aree delle code della distribuzione delle medie campionarie.

Si calcola

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

se σ è nota oppure se σ non è nota ma n è sufficientemente grande ($n \geq 90$) ed in tal caso $\sigma \simeq s$ (s deviazione standard campionaria). Si usano le tavole della distribuzione normale (**test z**).

Se σ non è nota e n non è sufficientemente grande e la popolazione da cui abbiamo estratto il campione è normale, allora si sostituisce a σ il valore s della deviazione standard campionaria e si usa la variabile

$$t_{n-1} = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

che segue una distribuzione t di Student con $n-1$ gradi di libertà. Si usano le tavole della t di Student (**test t**).

Per fare il confronto fra p e α occorre specificare se lo scarto da μ_0 è in una direzione (test d'ipotesi **unilaterale**) oppure in 2 direzioni (test d'ipotesi **bilaterale**).

Possiamo specificare la scelta di un test bilaterale anche associando all'ipotesi nulla $H_0 : \mu = \mu_0$ la cosiddetta **ipotesi alternativa** $H_A : \mu \neq \mu_0$.

Nel caso di test unilaterale in cui siamo interessati a scostamenti verso *destra* rispetto a μ_0, possiamo associare l'ipotesi alternativa $H_A : \mu > \mu_0$.

Nel caso di test unilaterale in cui siamo interessati a scostamenti verso *sinistra* rispetto a μ_0, possiamo associare l'ipotesi alternativa $H_A : \mu < \mu_0$.

Riassumendo, un test di ipotesi sulla media richiede la specificazione dei seguenti punti:

1. dichiarazione dell'ipotesi nulla $H_0 : \mu = \mu_0$
2. scelta del livello di significatività α
3. scelta fra il test bilaterale oppure unilaterale

E' generalmente molto improbabile che un'ipotesi nulla sia esattamente vera. Perchè allora testarla piuttosto che respingerla immediatamente?

1. *Per testare un'ipotesi semplificatrice.* A volte l'ipotesi nulla definisce un modello semplice per una situazione reale che è molto più complessa di quella indicata dal modello.
2. *Per testare un'ipotesi nulla che può essere approssimativamente vera.* Se testiamo un nuovo farmaco rispetto ad un placebo, può accadere che il farmaco sia o praticamente inattivo o molto efficace. L'ipotesi nulla che il farmaco sia completamente inattivo è allora un'approssimazione vicina a uno stato dei fatti possibile.
3. *Per testare la direzione della differenza da una valore critico.* Se l'ipotesi che $\mu = \mu_0$ viene contraddetta in modo significativo, sarà una buona prova a favore o di $\mu > \mu_0$ o di $\mu < \mu_0$.

Esempio. Consideriamo la popolazione rappresentata dai tempi di sopravvivenza di pazienti affetti da tumore trattati con un nuovo farmaco. Si sa che la deviazione standard è $\sigma = 43.3$ mesi e che il tempo medio di sopravvivenza dei pazienti non trattati con il nuovo farmaco è 38.3 mesi.

Specificazioni

1. $H_0 : \mu = 38.3$

2. livello di significatività $\alpha = 0.05$

3. test bilaterale: interessano scostamenti dalla media in ambedue le direzioni ($H_A : \mu \neq 38.3$)

Osservazione

Consideriamo un campione di $n = 100$ pazienti e calcoliamo la media campionaria. Si trova $\overline{x} = 46.9$ mesi.

Analisi

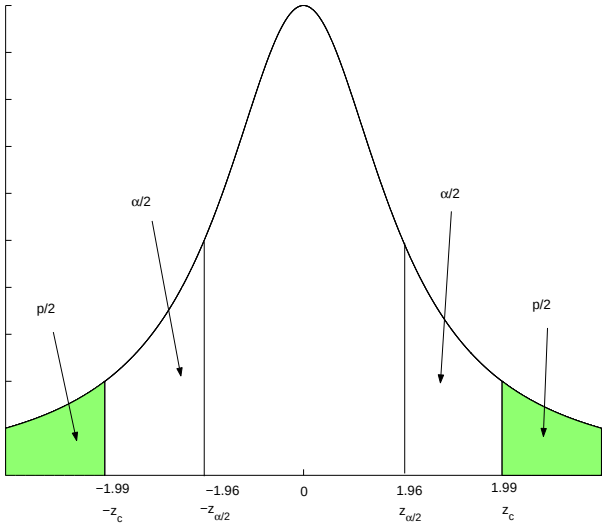
La dimensione n del campione è sufficientemente grande per garantire che la distribuzione delle medie campionarie sia ben approssimata dalla normale, anche se la distribuzione dei tempi di sopravvivenza non è normale.

Calcoliamo

$$z = \frac{46.9 - 38.3}{\frac{43.3}{\sqrt{100}}} = \frac{8.6}{4.33} = 1.99.$$

Dalla Tabella A.3, si trova che $z = 1.99$ stacca nella coda destra un'area $\frac{p}{2} = 0.023$; per simmetria $z = -1.99$ stacca a sinistra un'area $\frac{p}{2} = 0.023$. Quindi $p = 2 \cdot 0.023 = 0.046$.

Conclusione



$p < \alpha$, quindi *rifiutiamo* $H_0 : \mu = 38.3$. Il test è statisticamente significativo. In base al campione osservato, possiamo concludere che la media μ della popolazione è diversa da $\mu_0 = 38.3$. Possiamo anche dire che il valore campionario osservato $\overline{x} = 46.9$ *non* è compatibile con il valore definito dall'ipotesi nulla H_0.

La fluttuazione di campionamento non è una spiegazione verosimile della discrepanza fra il valore definito dall'ipotesi nulla ed i valori osservati nel campione.

Valori critici del test statistico

Chiamiamo z_c il valore calcolato di z in base al campione osservato. Nel nostro esempio $z_c = 1.99$.

Osserviamo che per qualunque z_c esterno all'intervallo (-1.96,1.96), la conclusione del test è di rifiutare l'ipotesi nulla H_0.

Invece, per qualunque z_c interno all'intervallo (-1.96,1.96) la conclusione del test è di non rifiutare l'ipotesi nulla H_0.

I valori -1.96 e 1.96 si chiamamo **valori critici** del test statistico.

Limiti dell'intervallo di confidenza

Per lo stesso campione calcoliamo i limiti dell'intervallo di confidenza al 95% per la media μ dei tempi di sopravvivenza per i pazienti trattati con un nuovo farmaco:

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 46.9 \pm 1.96\frac{43.3}{\sqrt{100}} = 46.9 \pm 8.49.$$

L'intervallo (38.41,55.39) contiene il tempo medio di sopravvivenza per gli

individui trattati con il nuovo farmaco, con un livello di confidenza del 95%.

$$\mu_0 = 38.3 \notin (38.41, 55.39)$$

I limiti di confidenza al 95% non comprendono il tempo medio di sopravvivenza per gli individui non trattati con il nuovo farmaco, in accordo con quanto trovato nel test d'ipotesi ad un livello di significatività $\alpha = 0.05$.

Esempio. Consideriamo la distribuzione dei livelli di colesterolo degli ipertesi fumatori. Assumiamo che la deviazione standard sia $\sigma = 46$ mg/100 ml. Conosciamo la media $\mu = 211$ mg/100 ml dei livelli di colesterolo della popolazione generale di età compresa fra i 20 e i 74 anni.

Specificazioni

1. $H_0 : \mu = 211$ mg/100 ml
2. livello di significatività $\alpha = 0.05$
3. test bilaterale, poiché il livello medio di colesterolo degli ipertesi fumatori può essere maggiore o minore di μ_0 ($H_A : \mu \neq 211$)

Osservazione

Consideriamo un campione di $n = 12$ ipertesi fumatori e calcoliamo la media campionaria: $\overline{x} = 217$ mg/100 ml.

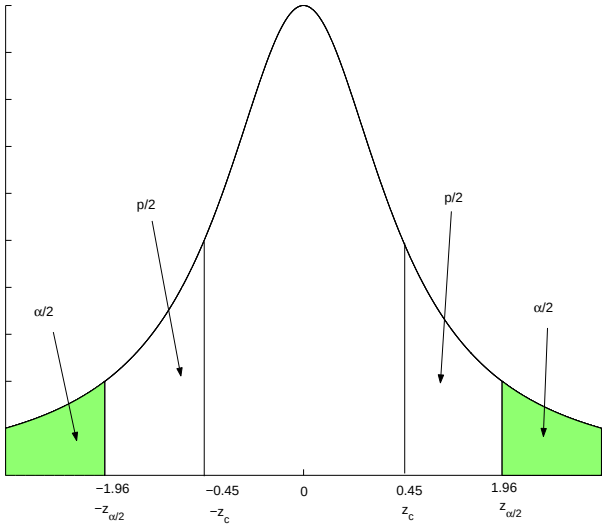
Analisi

La deviazione stan. σ è nota, il livello di colesterolo segue la distribuzione normale. Calcoliamo quindi il test z:

$$z = \frac{217 - 211}{\frac{46}{\sqrt{12}}} = 0.45$$

Dalla Tabella A.3, si nota che $z_c = 0.45$ stacca nella coda destra un'area $\frac{p}{2} = 0.326$. Per simmetria anche l'area a sinistra di $-z_c = -0.45$ sarà $\frac{p}{2} = 0.326$. Quindi $p = 2 \cdot 0.326 = 0.652$.

Conclusione



$p > \alpha$, quindi *non* rifiutiamo H_0. In base al campione osservato, non abbiamo sufficiente evidenza per concludere che il livello medio di colesterolo degli ipertesi fumatori sia diverso da 211 mg/100 ml. La media campionaria osservata $\overline{x} = 217$ mg/100 ml è compatibile con il valore definito dall'ipotesi nulla H_0.

La fluttuazione di campionamento è una spiegazione verosimile della discrepanza fra il valore specificato da H_0 ed i valori campionari osservati.

Qualunque valore di z_c compreso fra -1.96 e 1.96 produrrebbe un valore $p > 0.05$. In tutti questi casi l'ipotesi nulla H_0 non sarebbe rifiutata.

Limiti dell'intervallo di confidenza

Per lo stesso campione, calcoliamo i limiti dell'intervallo di confidenza al 95% per la media μ dei livelli di colesterolo degli ipertesi fumatori:

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 217 \pm 1.96\frac{46}{\sqrt{12}} = 217 \pm 13.28$$

Con una confidenza del 95%, confidiamo che l'intervallo (203.72,230.28) contenga il livello medio di colesterolo degli ipertesi fumatori.

$$211 \in (203.72, 230.28)$$

I limiti di confidenza al 95% comprendono il livello medio di colesterolo della popolazione generale, in assonanza con il risultato trovato nel test d'ipotesi al 5%.

Esempio. Consideriamo la popolazione dei livelli di alluminio plasmatico nei neonati che assumono farmaci a base di alluminio (antiacidi). La media e la deviazione standard di questa popolazione non sono note. Sappiamo

che il livello medio di alluminio plasmatico nei neonati che non assumono farmaci contenenti alluminio è 4.13 μg/l.

Specificazioni

1. $H_0 : \mu = 4.13\,\mu$g/l
2. livello di significatività $\alpha = 0.05$
3. test bilaterale, perchè siamo interessati a scostamenti dalla media 4.13 in ambedue le direzioni ($H_A : \mu \neq 4.13$)

Osservazione

Consideriamo un campione casuale di $n = 10$ neonati che assumono farmaci con alluminio. La media campionaria è $\overline{x} = 37.20\,\mu$g/l. La deviazione standard campionaria è $s = 7.13\,\mu$g/l.

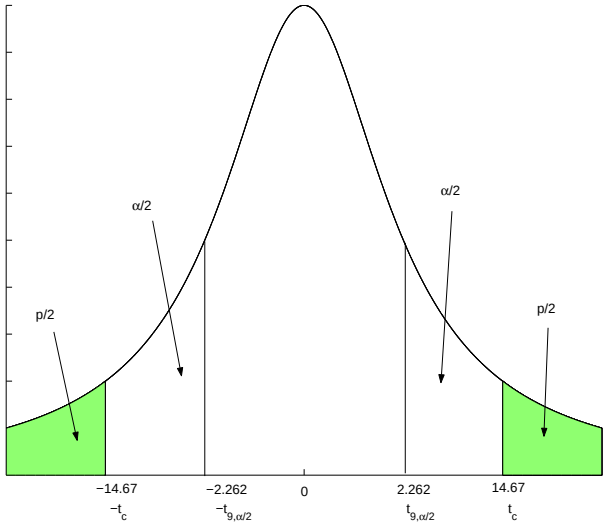
Analisi

Poiché non conosciamo la deviazione standard σ della popolazione, utilizziamo un test t. Calcoliamo

$$t_9 = \frac{37.20 - 4.13}{\frac{7.13}{\sqrt{10}}} = 14.67.$$

Dalla Tabella A.4, per una distribuzione t con 9 gradi di libertà l'area a destra di 14.67 è minore di 0.025 e l'area a sinistra di -14.67 è minore di 0.025.

Conclusione



$p < \alpha$, quindi rifiutiamo l'ipotesi nulla. Il test è statisticamente significativo. Questo campione di neonati fornisce sufficiente evidenza che il livello medio di alluminio dei neonati che assumono farmaci sia diverso da quello dei neonati che non assumono farmaci.

Limiti dell'intervallo di confidenza

I limiti dell'intervallo di confidenza al 95 % per μ sono

$$\overline{x} \pm 2.262\frac{s}{\sqrt{n}} = 37.20 \pm 2.262\frac{7.13}{\sqrt{10}} = \\ = 37.20 \pm 2.25$$

Dunque l'intervallo è (34.95,39.45).

Notiamo che

$$\mu_0 = 4.13 \notin (34.95, 39.45)$$

in accordo con quanto trovato eseguendo il test d'ipotesi ad un livello di significatività $\alpha = 0.05$.

Esempio. Consideriamo la distribuzione dei livelli di emoglobina dei bambini al di sotto dei 6 anni esposti ad elevati livelli di piombo. La media μ è sconosciuta. Si sa che $\sigma = 0.85$ g/100 ml e che il livello medio di emoglobina nei bambini non esposti ad elevati livelli di piombo è 12.29 g/100 ml. Riteniamo che i livelli di emoglobina dei bambini esposti siano mediamente inferiori a quelli dei bambini non esposti.

Specificazioni

1. $H_0 : \mu = 12.29$ g/100 ml
2. livello di significatività $\alpha = 0.05$
3. test unilaterale, relativo ai valori minori di $\mu_0 = 12.29$ ($H_A : \mu < 12.29$)

Osservazione

Consideriamo un campione casuale di $n = 74$ bambini esposti ad elevati livelli di piombo e troviamo $\overline{x} = 10.6$ g/100 ml.

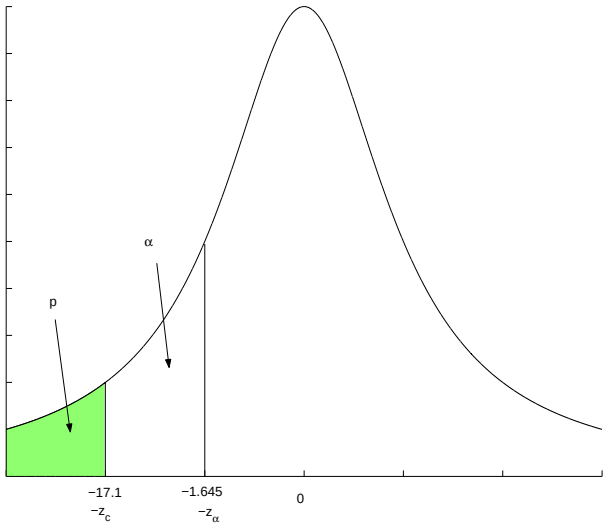
Analisi

Calcoliamo

$$z = \frac{10.6 - 12.29}{\frac{0.85}{\sqrt{74}}} = -17.10$$

Dalla Tabella A.3, l'area a sinistra di $z = -17.10$ è minore di 0.05.

Conclusione



$p < \alpha$, quindi rifiutiamo H_0. Questo campione fornisce sufficiente evidenza che il livello medio di emoglobina dei bambini esposti ad elevati valori di piombo sia minore di 12.29 g/100 ml. Qualunque altro valore $z_c \leq -1.645$ ci avrebbe portato al rifiuto dell'ipotesi nulla H_0 (-1.645 è il valore critico).

Limite superiore dell'intervallo di confidenza

Troviamo che il limite superiore dell'intervallo di confidenza al 95% è

$$\overline{x} + 1.645\frac{\sigma}{\sqrt{n}} = 10.6 + 1.645\frac{0.85}{\sqrt{74}} = 10.8$$

Confidiamo al 95% che 10.8 sia maggiore del livello di emoglobina medio dei bambini esposti al piombo.

$$10.8 < \mu_0 = 12.29$$
$$\mu_0 \notin (-\infty, 10.8)$$

in accordo con quanto trovato nel test d'ipotesi al 5%.

Esempio. Consideriamo la popolazione dei livelli di acido urico nei pazienti affetti da diabete. Sappiamo che la deviazione standard è $\sigma = 1.0$ mg/ml. Sappiamo che il livello medio nelle persone non diabetiche è 5.4 mg/ml. Riteniamo che il livello di acido urico nei diabetici sia mediamente più alto di quello dei non diabetici.

Specificazioni

1. $H_0 : \mu = 5.4$ mg/ml

2. livello di significatività $\alpha = 0.05$

3. test unilaterale, relativo ai valori maggiori di $\mu_0 = 5.4$ mg/ml ($H_A : \mu > 5.4$)

Osservazione

Consideriamo un campione di $n = 25$ diabetici, la cui media campionaria è $\overline{x} = 5.9$ mg/ml.

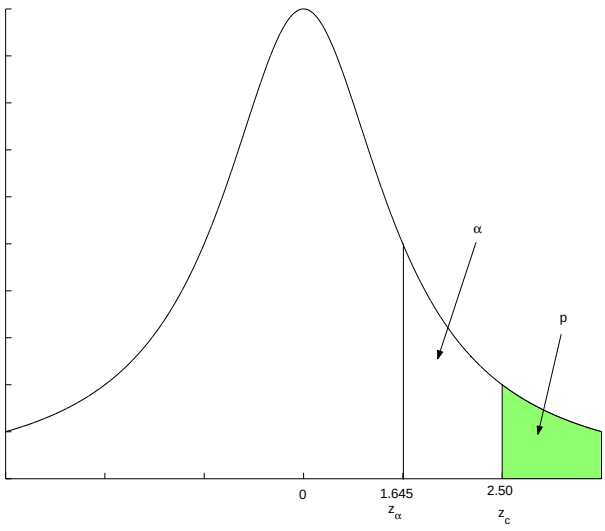
Analisi

Calcoliamo

$$z = \frac{5.9 - 5.4}{\frac{1.0}{\sqrt{25}}} = \frac{0.5}{0.2} = 2.5.$$

Dalla Tabella A.3, il valore $z = 2.5$ stacca nella coda destra un'area $p = 0.006$.

Conclusione



$p < \alpha$, quindi rifiutiamo H_0. Il campione fornisce evidenza ragionevole per asserire che il livello medio di acido urico nei diabetici sia superiore a quello dei non diabetici.

Limite inferiore dell'intervallo di confidenza

Calcoliamo il limite inferiore dell'intervallo di confidenza al 95% per μ:

$$\begin{aligned}\overline{x} - 1.645\frac{\sigma}{\sqrt{n}} &= 5.9 - 1.645\frac{1.0}{\sqrt{25}} = \\ &= 5.9 - 0.33 = \\ &= 5.57\end{aligned}$$

Confidiamo allora al 95% che 5.57 sia minore del livello medio di acido urico dei diabetici

$$\begin{aligned}5.57 > \mu_0 = 5.4 \\ \mu_0 \notin (5.57, +\infty)\end{aligned}$$

4.3 Tipi di errore e potenza

Nel trarre conclusioni con un test di ipotesi possiamo commettere due tipi di errore:

1. Errore del I tipo o errore α

   Si verifica quando rifiutiamo H_0, mentre H_0 è vera.

   $$\alpha = P(\text{rifiutare } H_0 | H_0 \text{ è vera})$$

2. Errore del II tipo o errore β

   Si verifica quando non rifiutiamo H_0, mentre H_0 è falsa.

   $$\beta = P(\text{non rifiutare } H_0 | H_0 \text{ è falsa})$$

Se β è la probabilità di commettere un errore del II tipo, $1 - \beta$ è la *potenza* del test di ipotesi.

La potenza è la probabilità di rifiutare H_0 quando H_0 è falsa:

$$\text{potenza} = P(\text{rifiutare } H_0 | H_0 \text{ è falsa})$$

4.4 Confronto fra 2 medie

Supponiamo di avere 2 campioni di osservazioni da 2 popolazioni sottostanti (es. gruppi di soggetti sottoposti a trattamento e di soggetti di controllo) con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2.

Formuliamo l'ipotesi nulla:

$$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow \mu_1 = \mu_2.$$

Posta *vera* H_0, si calcola la probabilità p di ottenere differenze tra le medie campionarie pari o maggiori di quelle osservate.

Se p è sufficientemente piccola allora è ragionevole dubitare della validità di H_0 e quindi rifiutiamo H_0.

E' necessario anche in questo caso specificare il *livello di significatività* e indicare se l'interesse è nelle differenze in una o entrambe le direzioni relativamente ad H_0 (test a 1 o 2 code)

4.4.1 Campioni appaiati

Ciascuna osservazione in un campione si associa con una ed una sola osservazione dell'altro campione. Si possono presentare i seguenti casi:

1. *autoappaiamento*: il soggetto serve come controllo di se stesso.

   Esempi.

   (a) Sperimentazioni cliniche in cui ciascun soggetto riceve 2 farmaci o 2 procedure in 2 momenti differenti (prima-dopo);

   (b) trattamento applicato ad una gamba, braccio, occhio, orecchio e diverso trattamento applicato all'altra gamba, braccio, occhio, orecchio.

2. *appaiamento naturale*

   Esempi.

   (a) 2 topini dello stesso sesso sono selezionati da una nidiata e un membro della coppia è assegnato ad un trattamento, mentre l'altro membro della coppia è assegnato ad un trattamento diverso;

   (b) in campo umano molte ricerche cliniche selezionano, per ciascun paziente affetto da una malattia, un fratello di controllo dello stesso sesso, il più vicino possibile per età al paziente e privo della malattia (appaiamento per nascita)

   (c) uno studio che considera come casi i ragazzi di una scuola affetti da una certa malattia e, come controlli, ragazzi, senza malattia, della stessa scuola;

3. *appaiamento artificiale*: è creato dal ricercatore. Consiste nell' appaiare soggetti per caratteristiche importanti in modo che i membri di un paio siano il più possibile simili fra loro riguardo a queste caratteristiche. Caratteristiche importanti si intendono quelle che sono associate al risultato sotto studio.

   Esempio. Eseguiamo uno studio sulla prematurità legata al peso alla nascita. Età della madre,razza,peso,pressione sanguigna, numero di gravidanze, abitudine al fumo influiscono sul peso alla nascita. Il ricercatore potrebbe scegliere di appaiare donne per molte o forse tutte

queste caratteristiche. Allora assegna a caso un membro di ciascuno di tali paia al nuovo farmaco sotto studio e l'altro al controllo.

Si presentano 2 difficoltà:

(a) conoscenza a priori delle caratteristiche rilevanti ai fini dello studio;

(b) quando le caratteristiche sono note e sono molte è estremamente difficile ottenere un appaiamento rispetto a tutti i fattori considerati.

Come si procede

Supponiamo di avere 2 campioni di dimensione n:

$$x_{11}, x_{12}, \ldots, x_{1n}$$
$$x_{21}, x_{22}, \ldots, x_{2n}$$

estratti a caso da due distribuzioni normali X_1 e X_2 con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2 rispettivamente.

Concentriamo l'attenzione sulla nuova variabile casuale *differenza* $D = X_1 - X_2$. Calcoliamo le singole differenze per ogni coppia:

$$\begin{aligned} d_1 &= x_{11} - x_{21} \\ d_2 &= x_{12} - x_{22} \\ d_3 &= x_{13} - x_{23} \\ &\ldots \\ d_n &= x_{1n} - x_{2n} \end{aligned}$$

La *media* delle differenze è

$$\overline{d} = \frac{d_1 + d_2 + \ldots + d_n}{n}.$$

La *deviazione standard* delle differenze è

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}.$$

Sia $\delta = \mu_1 - \mu_2$. L'ipotesi nulla da testare è dunque

$$H_0 : \delta = 0.$$

In generale NON si conosce σ_d, la deviazione standard della popolazione delle differenze. Dunque si esegue un test statistico t di Student con $n-1$ gradi di libertà:

$$t_{n-1} = \frac{\overline{d} - \delta}{\frac{s_d}{\sqrt{n}}} = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}}.$$

Esempio. Un gruppo di soggetti ipertesi riceve un farmaco di prova in un momento e un placebo in un altro momento. Le medie μ_1 (ipertesi con placebo) e μ_2 (ipertesi con farmaco) non sono note. La deviazione standard della popolazione delle differenze σ_d non è nota.

Specificazioni

1. $H_0 : \delta = 0$, con $\delta = \mu_1 - \mu_2$
2. livello di significatività $\alpha = 0.05$
3. test bilaterale perchè siamo interessati alle differenze in entrambe le direzioni

Osservazione

Consideriamo un campione di 11 pazienti ipertesi. Confrontiamo l'effetto del placebo e del medicinale sulla pressione sistolica:

$$\begin{aligned} \text{placebo} &: 211, 210, 210, 203, 196, \ldots, 163 \\ \text{medicinale} &: 181, 172, 196, 191, 167, \ldots, 156 \\ d_i &: 30, 38, 14, 12, 29, \ldots, 7 \end{aligned}$$

Calcoliamo

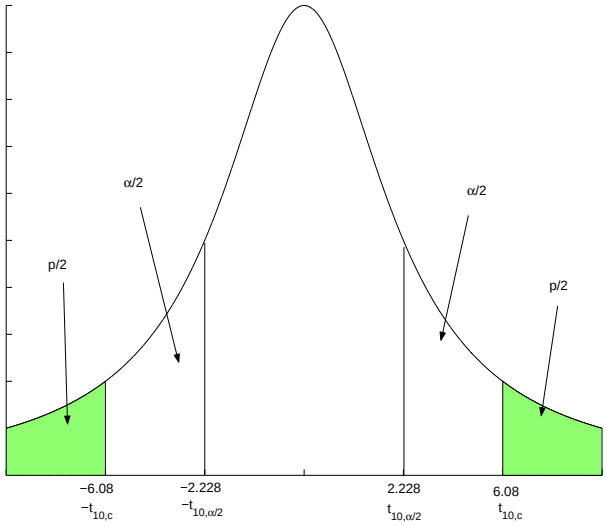
$$\overline{d} = \frac{\sum_{i=1}^{11} d_i}{11} = 24.0$$
$$s_d = \sqrt{\frac{\sum_{i=1}^{11} (d_i - \overline{d})^2}{11 - 1}} = \sqrt{171.4} = 13.09$$

Analisi

Non conosciamo σ_d, quindi utilizziamo un test t. Calcoliamo

$$t_{10} = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{24.0}{\frac{13.09}{\sqrt{11}}} = \frac{24.0}{3.95} = 6.08.$$

Conclusione



$p < \alpha$, quindi il test è statisticamente significativo, cioè indica che la fluttuazione di campionamento è una spiegazione altamente inverosimile della differenza osservata fra placebo e medicinale. Rifiutiamo H_0.

Limiti dell'intervallo di confidenza

I limiti dell'intervallo di confidenza bilaterale al 95% per $\delta = \mu_1 - \mu_2$ sono

$$\overline{d} \pm t_{n-1,0.025} \frac{s_d}{\sqrt{n}} = 24.01 \pm (2.228)(3.95) =$$
$$= 24.01 \pm 8.80$$

Quindi l'intervallo cercato è

$$(15.2, 32.8)$$

Siamo dunque confidenti al 95% che l'intervallo (15.2,32.8) contenga la reale differenza $\delta = \mu_1 - \mu_2$.

$$\delta = 0 \notin (15.2, 32.8)$$

in accordo con quanto trovato nel test d'ipotesi.

Esempio. Consideriamo dei soggetti ipertesi che ricevono un farmaco di prova in un momento ed un placebo in un altro momento. La pressione media μ_1 dei soggetti con placebo e la pressione media μ_2 dei soggetti con farmaco non sono note. Non conosciamo neppure σ_d. Sappiamo che il farmaco *abbassa* la pressione.

Specificazioni

1. $H_0 : \delta = 0$, con $\delta = \mu_1 - \mu_2$

2. livello di significatività $\alpha = 0.05$

3. test unilaterale: siamo interessati a differenze in una direzione, in particolare $\delta > 0$ ($H_A : \delta > 0$).

Osservazione

Consideriamo un campione di 11 pazienti ipertesi. Confrontiamo l'effetto del placebo e del medicinale sulla pressione.

$$\begin{aligned} \text{placebo} &: 211, 210, \ldots, 163 \\ \text{medicinale} &: 181, 172, \ldots, 156 \\ d_i &: 30, 38, \ldots, 7 \end{aligned}$$

Calcoliamo

$$\overline{d} = \frac{\sum_{i=1}^{11} d_i}{11} = 24.0$$

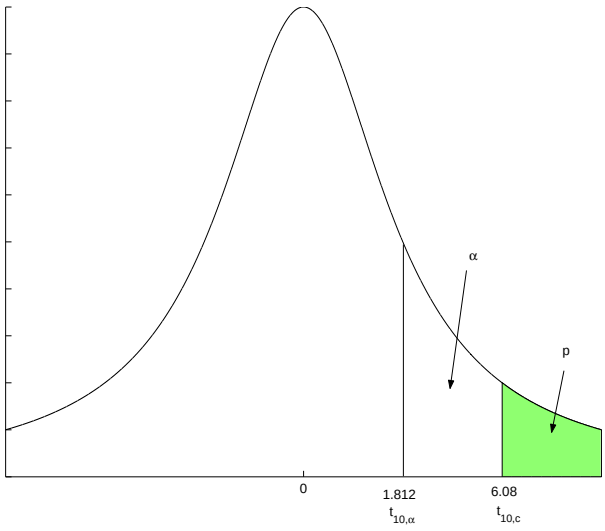
$$s_d = \sqrt{\frac{\sum_{i=1}^{11}(d_i - \overline{d})^2}{11-1}} = \sqrt{171.4} = 13.09$$

Analisi

Calcoliamo

$$t_{10} = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{24.0}{\frac{13.09}{\sqrt{11}}} = 6.08.$$

Conclusione



$p < \alpha$, quindi rifiutiamo H_0: il test è statisticamente significativo.

La pressione media dei soggetti con placebo è più alta della pressione media dei soggetti con il farmaco.

Intervallo di confidenza unilaterale

Possiamo essere interessati a calcolare il limite *inferiore* per l'intervallo di confidenza per la reale differenza $\delta = \mu_1 - \mu_2$ fra le medie.

Per una distribuzione t con 10 gradi di libertà, il 95% delle osservazioni cade a sinistra di $t_{10,0.05} = 1.812$. Quindi

$$P(t \leq 1.812) = P\left(\frac{\overline{d} - \delta}{\frac{s_d}{\sqrt{n}}} \leq 1.812\right) = 0.95.$$

Ma

$$\frac{\overline{d} - \delta}{\frac{s_d}{\sqrt{n}}} \leq 1.812 \Leftrightarrow \overline{d} - \delta \leq 1.812 \cdot \frac{s_d}{\sqrt{n}} \Leftrightarrow$$

$$\delta \geq 24.0 - 1.812 \cdot \frac{13.09}{\sqrt{11}} = 24.0 - 7.1 = 16.9.$$

Siamo dunque confidenti al 95% che 16.9 sia *minore o uguale* a $\delta = \mu_1 - \mu_2$.

$$\delta = 0 \notin [16.9, +\infty)$$

in accordo con quanto trovato nel test d'ipotesi al 5%.

4.4.2 Campioni indipendenti

In molti casi non si conoscono i fattori rilevanti per l'appaiamento; questi addirittura possono non esistere. Inoltre l'appaiamento può essere amministrativamente difficile e provocare uno spreco di tempo.

In alternativa possiamo avere 2 campioni *indipendenti* di osservazioni e quindi un insieme di osservazioni relative al trattamento ottenute indipendentemente da quelli di controllo.

Con campioni indipendenti NON è necessario che i numeri delle osservazioni del gruppo sottoposto a trattamento e di quello di controllo siano gli stessi.

Come si procede

Siano dati due campioni

$$x_{11}, x_{12}, \ldots, x_{1n_1}$$
$$x_{21}, x_{22}, \ldots, x_{2n_2}$$

estratti da 2 popolazioni normali indipendenti con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2. Siano $\overline{x}_1$ e $\overline{x}_2$ le due medie campionarie.

Si dimostra che *quando operiamo con campioni di 2 popolazioni normali indipendenti, la differenza delle medie campionarie* $\overline{X}_1$ *e* $\overline{X}_2$ *è approssimativamente normale con media* $\mu_1 - \mu_2$ *ed errore standard*

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Sia $\delta = \mu_1 - \mu_2$. L'ipotesi nulla è

$$H_0 : \delta = 0.$$

Consideriamo i seguenti casi:

1. varianze **note e diverse**: $\sigma_1^2 \neq \sigma_2^2$.

   Calcoliamo il test z:
   $$z = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

2. varianze **note e uguali**: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

   Calcoliamo il test z:
   $$z = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

3. varianze **sconosciute e uguali**: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

   Dal primo campione abbiamo la varianza campionaria
   $$s_1^2 = \frac{\sum_{j=1}^{n_1}(x_{1j} - \overline{x}_1)^2}{n_1 - 1},$$
   dal secondo campione
   $$s_2^2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \overline{x}_2)^2}{n_2 - 1}.$$

   Una stima comune potrebbe essere la semplice *media* di s_1^2 e s_2^2. Si dimostra che ciò è inappropriato: dato che le grandezze campionarie possono differire sostanzialmente nei 2 gruppi, una varianza campionaria potrebbe essere una stima di σ^2 molto più affidabile dell'altra. Pertanto sembra appropriata una media *ponderata* di s_1^2 e s_2^2 avente come pesi quantità che

dipendono dall'affidabilità di ciascuna varianza campionaria. Matematicamente si può verificare che i pesi ottimali sono i *gradi di libertà* di ciascuna varianza campionaria. La stima risultante combinata di σ^2 si chiama *varianza pooled* della varianza comune:

$$s_p^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{(n_1-1)+(n_2-1)} = \\ = \frac{\sum_{j=1}^{n_1}(x_{1j}-\overline{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\overline{x}_2)^2}{n_1+n_2-2}.$$

Eseguiamo allora in questo caso un test statistico t di Student con $n_1 + n_2 - 2$ gradi di libertà:

$$t_{n_1+n_2-2} = \frac{(\overline{x}_1-\overline{x}_2)-0}{\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

E' ragionevole supporre che le varianze delle 2 popolazioni siano uguali?

Il fondamento logico sta nel fatto che, in molte situazioni, l'applicazione di un nuovo trattamento ad un gruppo, mentre l'altro è lasciato come controllo, tende al massimo ad influenzare il valore medio e a lasciare immutata la variabilità. L'esperienza ha posto questo in evidenza.

Si potrebbe anche testare l'ipotesi di uguaglianza fra le 2 varianze, ma questo tipo di test è estremamente sensibile alla assunzione di *normalità* per le 2 popolazioni originarie.

Esempio. Consideriamo le distribuzioni dei livelli di ferrosierico della popolazione dei bambini sani e della popolazione dei bambini malati di fibrosi cistica. Denotiamo con μ_1 il livello medio di ferro nei bambini sani e con μ_2 il livello medio di ferro nei bambini malati. Le deviazione standard σ_1 e σ_2 non sono note. Supponiamo che siano *uguali*: $\sigma_1 = \sigma_2$. Vogliamo stabilire se i bambini con fibrosi cistica hanno un livello normale di ferro.

Specificazioni

1. $H_0 : \mu_1 = \mu_2$
2. livello di significatività $\alpha = 0.05$
3. test bilaterale: siamo interessati alle differenze fra le medie in entrambe le direzioni

Osservazione

Selezioniamo un campione casuale da ciascuna popolazione. Il campione di $n_1 = 9$ bambini sani ha un livello medio di ferro $\overline{x}_1 = 18.9\,\mu$mol/l ed una deviazione standard $s_1 = 5.9\,\mu$mol/l. Il campione di $n_2 = 13$ bambini con fibrosi cistica ha un livello medio di ferro $\overline{x}_2 = 11.9\,\mu$mol/l ed una deviazione standard $s_2 = 6.3\,\mu$mol/l.

E' possibile che la differenza osservata nelle medie campionarie sia il risultato della variabilità dovuta al caso oppure dobbiamo concludere che la differenza sia dovuta ad una reale differenza fra le medie delle popolazioni?

Analisi

Applichiamo il test t per 2 campioni indipendenti con varianze *uguali*.
Calcoliamo la stima pooled della varianza:

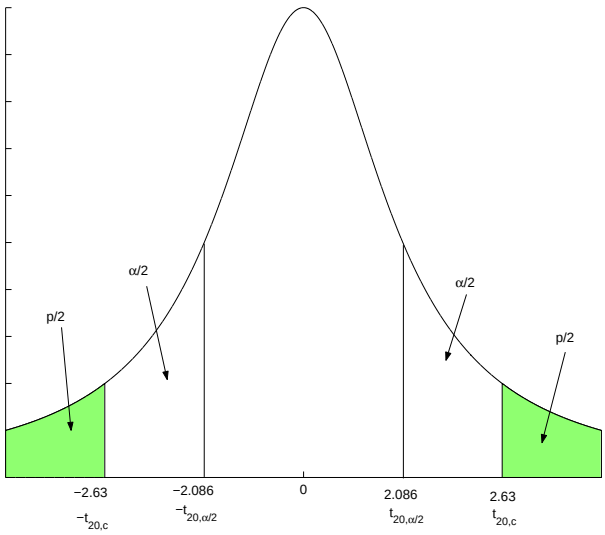
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \\ = \frac{(9-1)(5.9)^2 + (13-1)(6.3)^2}{9+13-2} = 37.74.$$

Calcoliamo

$$t_{9+13-2} = t_{20} = \frac{18.9-11.9}{\sqrt{37.74(\frac{1}{9}+\frac{1}{13})}} = 2.63.$$

Dalla Tabella A.4, troviamo che l'area a destra di $t_{20} = 2.63$ è compresa fra 0.005 e 0.01.

Conclusione



$p < \alpha$, quindi rifiutiamo H_0.

La differenza fra il livello medio di ferro dei bambini sani e quello dei bambini malati è statisticamente significativa.

Intervallo di confidenza bilaterale

Per una distribuzione t con 20 gradi di libertà, il 95% delle osservazioni cade in (-2.086,2.086). Quindi

$$P\left(-2.086 \leq \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}) \leq 2.086\right) = 0.95.$$

La disuguaglianza

$$-2.086 \leq \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \leq 2.086$$

porta a

$$(\overline{x}_1 - \overline{x}_2) - 2.086\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})} \leq \mu_1 - \mu_2 \leq$$
$$\leq (\overline{x}_1 - \overline{x}_2) + 2.086\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}.$$

Quindi i limiti dell'intervallo di confidenza sono

$$(\overline{x}_1 - \overline{x}_2) \pm 2.086\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})} =$$
$$= (18.9 - 11.9) \pm 2.086\sqrt{37.74(\frac{1}{9} + \frac{1}{13})}.$$

Siamo pertanto confidenti al 95% che l'intervallo (1.44,12.55) contenga la reale differenza fra i livelli medi $\mu_1 - \mu_2$.

$$\mu_1 - \mu_2 = 0 \notin (1.44, 12.55)$$

in accordo con quanto trovato nel test d'ipotesi al 5%.

4.5 Analisi della varianza ad 1 criterio di classificazione

Esso è

- un metodo efficace per analizzare l'effetto prodotto dalle classificazioni di vario genere dei dati sul valore medio di una variabile;
- una generalizzazione del test t per un numero *qualsiasi* di campioni indipendenti.

Vediamo alcuni **esempi** di classificazione a 1 criterio dei dati in più gruppi:

1. riduzione della glicemia registrata in gruppi di conigli a cui si somministrano diverse dosi di insulina;
2. valore di un certo test funzionale respiratorio registrato in uomini dello stesso gruppo di età di categorie professionali diverse;
3. volumi di liquido prelevato da uno sperimentatore, che usa diverse pipette per misurare una quantità standard, raggruppando le misure ripetute con la stessa pipetta.

In ognuno degli esempi si potrebbe porre la stessa *domanda*:

Cosa si può dire circa la *variabilità* della glicemia da un gruppo di dosi all'altro, sulla variabilità del test di funzionalità respiratoria da una categoria professionale all'altra, sulla variabilità del volume da una pipetta all'altra?

Prendiamo ora in considerazione k popolazioni diverse, supponendo che siano

- indipendenti;
- normalmente distribuite.

Siano $\mu_1, \mu_2, \ldots, \mu_k$ la medie delle k popolazioni.

Noi vogliamo *testare l'ipotesi nulla*

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k.$$

Supponiamo che le varianze delle k popolazioni siano uguali:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2.$$

Operando con numerose differenti popolazioni, possiamo calcolare 2 misure di variazione:

1. la variazione dei valori individuali rispetto alla media della loro popolazione. Questa è la *varianza entro gruppi*:

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

2. la variazione delle medie delle popolazioni rispetto alla media generale. Questa è la *varianza tra gruppi*:

$$s_B^2 = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_k(\overline{x}_k - \overline{x})^2}{k - 1}$$

dove $n_1, n_2, \ldots, n_k$ sono le dimensioni dei k campioni, $\overline{x}_1$,$\overline{x}_2$,...,$\overline{x}_k$ sono le medie dei k campioni estratti dalle k popolazioni, $\overline{x}$ è la *media globale*:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k}.$$

Se la variabilità *all'interno* delle k popolazioni è piccola rispetto alla variabilità *tra le loro rispettive medie*, significa che le medie delle popolazioni sono realmente differenti.

Un test appropriato per l'ipotesi nulla è perciò basato sul rapporto di varianza

$$F = \frac{s_B^2}{s_W^2}.$$

F è una nuova distribuzione di probabilità, detta *distribuzione di Snedecor*.

Osservazioni.

1. Si dimostra che s_B^2 e s_W^2 si comportano come due stime indipendenti della varianza con rispettivamente $k-1$ e $n_1 + \cdots + n_k - k$ gradi di libertà. Sotto l'ipotesi nulla, dunque, F è vicino a 1.
2. Deviazioni dall'ipotesi nulla tendono a dare valori maggiori dell'unità. Quindi un test di significatività dell'ipotesi nulla deve considerare significativi solo quei valori di F che sono sufficientemente grandi. Pertanto è richiesto un *test unilaterale*.
3. Se $k = 2$ allora il test F si riduce al test t per 2 campioni indipendenti.

Esempio. Consideriamo i dati relativi al volume espiratorio forzato in un secondo in pazienti con patologia coronarica provenienti da 3 diversi centri medici.

Specificazioni

1. $H_0 : \mu_1 = \mu_2 = \mu_3$
2. livello di significatività $\alpha = 0.05$
3. il test in questo caso è sempre unidirezionale (testiamo $s_B^2 > s_W^2$).

Osservazioni

Consideriamo i 3 campioni di Fig. 4.1 di dimensione $n_1 = 21$, $n_2 = 16$, $n_3 = 23$, con medie $\overline{x}_1 = 2.63$ l, $\overline{x}_2 = 3.03$ l, $\overline{x}_3 = 2.88$ l e deviazione standard $s_1 = 0.496$ l, $s_2 = 0.523$ l, $s_3 = 0.498$ l. Calcoliamo la stima della varianza *entro* gruppi:

$$s_W^2 = \frac{(21-1)(0.496)^2 + (16-1)(0.523)^2 + (23-1)(0.498)^2}{21 + 16 + 23 - 3} == 0.254.$$

Calcoliamo la media globale

$$\overline{x} = \frac{21(2.63) + 16(3.03) + 23(2.88)}{21 + 16 + 23} = 2.83.$$

Calcoliamo la stima della varianza tra gruppi:

$$s_B^2 = \frac{21(2.63 - 2.83)^2 + 16(3.03 - 2.83)^2 + 23(2.88 - 2.83)^2}{2} = 0.769.$$

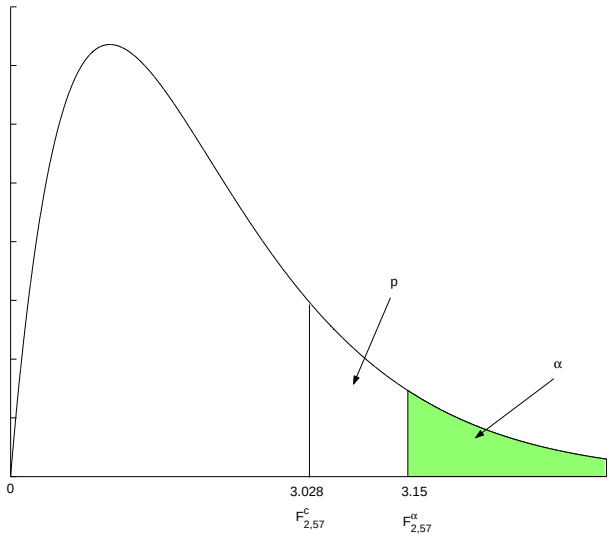
Analisi

Calcoliamo

$$F_{2,57} = \frac{s_B^2}{s_W^2} = \frac{0.769}{0.254} = 3.028.$$

Dalla Tavola A.5 troviamo che $F_{2,57;0.05} = 3.15$.

Conclusione



$p > \alpha$, qundi non rifiutiamo H_0: i campioni a nostra disposizione non ci possono far concludere che esiste una differenza tra le reali medie delle 3 popolazioni.

Capitolo 5

Metodi non parametrici

Per tutti i test statistici finora studiati abbiamo supposto che le popolazioni da cui erano selezionati i dati fossero distribuite secondo una curva normale o approssimativamente normale.

Se invece i dati non rispettano le assunzioni necessarie per l'applicazione delle tecniche tradizionali allora devono essere utilizzati quelli che sono comunemente chiamati **metodi non parametrici** di inferenza statistica.

Le tecniche non parametriche si basano su un minor numero di assunzioni sulla natura delle distribuzioni originarie.

I *test* di ipotesi *non parametrici* seguono la stessa procedura generale dei test parametrici già visti.

1. Prima di tutto facciamo delle *supposizioni* sulle popolazioni originarie attraverso la formulazione della *ipotesi nulla*.

2. Calcoliamo quindi il valore del test statistico utilizzando i dati di un campione casuale di osservazioni.

3. Infine, a seconda del risultato statistico, rifiutiamo o meno l'ipotesi nulla.

5.1 Test di Wilcoxon dei ranghi con segno: campioni appaiati

Viene utilizzato per confrontare campioni di osservazioni quando le popolazioni da cui sono estratti **non** sono **indipendenti**. Esso è quindi simile al test t per i dati appaiati. Come il test t, esso non esamina i due gruppi singolarmente ma si concentra sulla *differenza* tra i valori di ciascuna coppia ed il *segno* di ciascuna differenza.

Tuttavia, esso *non* richiede che la popolazione delle differenze sia normalmente distribuita.

Il test di Wilcoxon dei ranghi con segno è utilizzato per testare l'*ipotesi nulla* che, nella popolazione originaria delle differenze tra le coppie, la **differenza mediana** sia uguale a 0.

Studiamo come si effettua il test di Wilcoxon dei ranghi con segno su un esempio (Tabella 5.1).

Esempio. Supponiamo di voler esaminare l'uso dell'amiloride nella terapia di pazienti con fibrosi cistica. Si ritiene che il farmaco possa favorire la ventilazione polmonare e quindi ritardare la perdita di funzionalità polmonare associata alla malattia. La *capacità vitale forzata* è il volume d'aria che una persona può espellere in 6 secondi; vogliamo confrontare la riduzione della capacità vitale forzata che si verifica in un periodo di 25 settimane di trattamento con il farmaco, con quanto si verifica durante lo stesso periodo di trattamento con placebo.

Come eseguire il test di Wilcoxon dei ranghi con segno

1. Selezioniamo un campione casuale di n coppie di osservazioni.

2. Calcoliamo la differenza di ciascuna coppia di osservazioni.

3. Ignorando i segni delle differenze calcolate, ordiniamo i loro valori assoluti dal più piccolo al più grande. Una differenza uguale a 0 *non* è ordinata e si esclude pertanto dall'analisi, così che la dimensione del campione è ridotta di un'unità.

4. Alle differenze uguali è assegnato un *rango medio*; se le due differenze più piccole assumono entrambe il valore 11, ad esempio, ciascuna osservazione riceverà un rango pari a $(1 + 2)/2 = 1.5$.

5. Infine, assegniamo a ciascun rango un segno positivo o negativo a seconda del segno della differenza.

6. Calcoliamo ora la somma dei ranghi positivi e dei ranghi negativi. Ignorando i segni, indichiamo con T la somma più piccola. Sotto l'ipotesi nulla che la mediana della popolazione originaria delle differenze è uguale a 0, ci aspettiamo che un campione abbia approssimativamente un numero uguale di ranghi positivi e ranghi negativi. Inoltre, la grandezza della somma dei ranghi positivi deve essere confrontabile con la somma dei ranghi negativi.

7. Testiamo allora l'ipotesi nulla considerando il test statistico

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

dove

$$\mu_T = \frac{n(n+1)}{4}$$

è la somma media dei ranghi (infatti ricordiamo che la somma dei primi n numeri naturali è data da $\frac{n(n+1)}{2}$) e

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

è la deviazione standard di T.

Quando la dimensione n del campione è grande, la variabile casuale

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

segue una distribuzione approssimativamente normale con media 0 e deviazione standard 1.

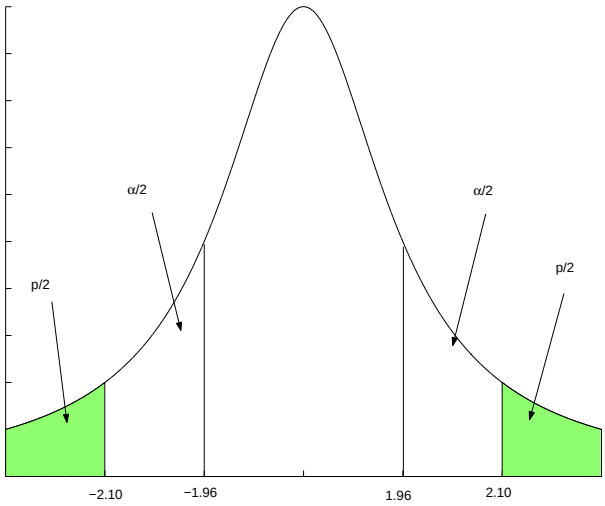
Nel nostro esempio

$$\mu_T = \frac{n(n+1)}{4} = \frac{14(14+1)}{4} = 52.5,$$
$$\sigma_T = \sqrt{\frac{14(14+1)[2(14)+1]}{24}} = 15.93$$

Otteniamo allora che

$$z_T = \frac{T - \mu_T}{\sigma_T} = \frac{19 - 52.5}{15.93} = -2.10$$

Conclusione



L'area sotto la curva normale standardizzata a sinistra di $z = -2.10$ e a destra di $z = 2.10$ è $2(0.018) = 0.036$. Poichè il valore p del test è minore del livello di significatività $\alpha = 0.05$, **rifiutiamo** l'ipotesi nulla e concludiamo che la differenza mediana non è uguale a 0.

La maggior parte delle differenze è positiva. Ciò suggerisce che la riduzione della capacità vitale forzata è *maggiore* durante il trattamento con il placebo che durante il trattamento con il farmaco. L'uso del farmaco, quindi, riduce la perdita di funzionalità polmonare.

Osservazione. Se invece n (dimensione del campione) è *piccolo*, non possiamo supporre che il test statistico z_T segua una distribuzione normale standardizzata. In questo caso, sono disponibili tabelle che ci permettono di valutare se rifiutare o meno l'ipotesi nulla (vedi Tabella A.6).

5.2 Vantaggi e svantaggi dei metodi non parametrici

Vediamo alcuni **vantaggi**:

1. Non richiedono che tutte le popolazioni originarie siano *normalmente* distribuite. Al massimo, le popolazioni devono avere la stessa forma di base.

2. Considerando i ranghi anzichè i valori reali delle osservazioni, possono essere eseguiti rapidamente per piccoli campioni.

3. L'utilizzo dei ranghi li rende *meno sensibili* ad errori di misurazione e permette l'utilizzo di misurazioni *ordinali* piuttosto che continue.

Vediamo alcuni **svantaggi**:

1. Se le ipotesi di un test parametrico sono soddisfatte, il test non parametrico è *meno potente* della corrispondente tecnica parametrica. Se l'ipotesi nulla è falsa, il test non parametrico richiede un campione più ampio per fornire sufficiente evidenza per rifiutarla.

2. Le ipotesi testate con tecniche non parametriche sono *meno specifiche* di quelle testate con metodi parametrici. Infatti, basandosi sui ranghi, essi non utilizzano tutte le informazioni note di una distribuzione.

3. Se molte osservazioni sono uguali, σ_T è una *sovrastima* della deviazione standard di T.

Capitolo 6

Inferenza sulle proporzioni

Applichiamo qui le tecniche dell'inferenza statistica alle *frequenze*. Nello studio delle frequenze siamo di solito interessati alla *proporzione* (frequenza relativa) più che al numero di volte che si verifica un evento (frequenza assoluta).

6.1 Approssimazione normale alla binomiale

Esempio (Esempio 1)**.** Supponiamo di selezionare dalla popolazione di adulti degli Stati Uniti un campione casuale di 30 individui. La probabilità che un individuo sia fumatore è $p = 0.29$. Ci chiediamo qual è la probabilità che al massimo 6 fra i 30 selezionati siano fumatori.

Applicando il principio della somma delle probabilità, avremo, secondo quanto già visto nello studio della distribuzione binomiale:

$$\begin{aligned} P_{30}(X \leq 6) =& P(X = 0) + P(X = 1) + P(X = 2) + \\ &+ \ldots + P(X = 6) = \ldots = 0.19. \end{aligned}$$

Quando la dimensione del campione è grande, l'uso della distribuzione binomiale diventa difficoltoso dal punto di vista del calcolo.

Possiamo allora calcolare le probabilità associate ai risultati di una variabile casuale binomiale X utilizzando una *approssimazione* della distribuzione binomiale basata sulla distribuzione normale.

All'aumentare della dimensione n del campione, la forma di una distribuzione binomiale si avvicina a quella di una normale con la stessa media np e la stessa varianza $np(1-p)$ della binomiale.

Un criterio molto usato afferma che n è "sufficientemente grande" per approssimare una binomiale con una normale quando

$$np \geq 5 \quad \text{e} \quad n(1-p) \geq 5.$$

In questo caso

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

segue approssimativamente la distribuzione normale standardizzata.

Esempio (continuazione Esempio 1)**.** Utilizzando l'approssimazione normale, vogliamo trovare la proporzione di campioni di dimensione $n = 30$ in cui ci sono al massimo 6 fumatori.

Osserviamo innanzitutto che

$$np = 30(0.29) = 8.7 > 5 \text{ e } n(1-p) = 30(0.71) = 21.3 > 5$$

e quindi l'approssimazione è valida.

Calcoliamo

$$z = \frac{6 - 30(0.29)}{\sqrt{30(0.29)(0.71)}} = -1.09.$$

Dalla Tavola A.3, si trova che l'area sottesa dalla curva normale standardizzata a sinistra di -1.09 è pari a 0.138. Quindi 0.138 è la probabilità che al massimo 6 individui siamo fumatori. Questo valore è un'*approssimazione* di $P(X \leq 6) = 0.19$.

6.1.1 Correzione per la continuità

E' stato dimostrato che si può ottenere una migliore approssimazione alla distribuzione binomiale utilizzando il rapporto

$$Z = \frac{X - np + 0.5}{\sqrt{np(1-p)}}, \quad \text{se } X < np,$$

e

$$Z = \frac{X - np - 0.5}{\sqrt{np(1-p)}}, \quad \text{se } X > np.$$

Il termine 0.5 al numeratore del rapporto si chiama *correzione per la continuità.*

Esempio (continuazione Esempio 1)**.** Applichiamo la correzione per continuità per trovare la proporzione di campioni di dimensione $n = 30$ in cui al massimo 6 individui sono fumatori.

Si ha che $X < np$ perchè $6 < 30(0.29) = 8.7$.

Quindi calcoliamo

$$z = \frac{6 - (30)(0.29) + 0.5}{\sqrt{30(0.29)(0.71)}} = -0.89.$$

Dalla Tavola A.3, l'area sottesa a sinistra di $z = -0.89$ è pari a 0.187.

0.187 è una *approssimazione* migliore di $P(X \leq 6) = 0.19$.

6.2 Distribuzione campionaria di una proporzione

Supponiamo di voler stimare la proporzione p di volte in cui si verifica un determinato evento in una determinata popolazione, sulla base di un campione casuale estratto dalla popolazione stessa.

Se la dimensione del campione è n e il numero di volte in cui si verifica l'evento è x, possiamo stimare la proporzione p della popolazione con

$$\hat{p} = \frac{x}{n} \quad \text{(frequenza relativa)}.$$

La proporzione del campione $\hat{p}$ è lo stimatore di massima verosimiglianza di p, cioè è il valore del parametro p che più verosimilmente ha prodotto il campione.

Indicato con 1 il *successo* (uscita dell'evento che stiamo studiando) e con 0 l'*insuccesso*, la media per p è uguale alla proporzione di 1 nella popolazione, cioè p. La deviazione standard è $\sqrt{p(1-p)}$.

Selezioniamo un campione di dimensione n ed indichiamo la proporzione di 1 nel campione con $\hat{p}_1$. Selezioniamo un secondo campione di dimensione n ed indichiamo con $\hat{p}_2$ la proporzione di 1 in questo secondo campione.

Se continuiamo a selezionare all'infinito campioni di dimensione n, otteniamo una *distribuzione campionaria delle proporzioni*.

La distribuzione campionaria delle proporzioni ha le seguenti proprietà (dal Teorema del limite centrale):

1. la media della distribuzione campionaria è la media p della popolazione;

2. la deviazione standard della distribuzione campionaria delle proporzioni è $\sqrt{\frac{p(1-p)}{n}}$. Questa quantità si chiama *errore standard* della proporzione campionaria $\hat{p}$;

3. la forma della distribuzione campionaria è approssimativamente normale se n è sufficientemente grande.

Poichè la distribuzione di $\hat{p}$ è approssimativamente normale con media p e deviazione standard $\sqrt{\frac{p(1-p)}{n}}$, sappiamo che

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

è distribuita normalmente con media 0 e deviazione standard 1.

Esempio. Consideriamo la sopravvivenza a 5 anni in pazienti cui è stato diagnosticato un tumore al polmone. La proporzione media di pazienti sopravvissuti è $p = 0.10$, la deviazione standard è $\sqrt{0.10(1-0.10)} = 0.30$.

Selezioniamo un campione casuale di dimensione $n = 50$ e ricaviamo da esso la proporzione campionaria $\hat{p} = 0.20$.

Se selezioniamo da questa popolazione ripetuti campioni di dimensione $n = 50$, quale frazione avrà una proporzione campionaria maggiore o uguale a 0.20?

Verifichiamo che

$$np = 50(0.10) = 5 \geq 5$$
$$n(1-p) = 50(0.90) = 45 \geq 5.$$

Dal Teorema del limite centrale, sappiamo che la distribuzione campionaria delle proporzioni $\hat{p}$ è approssimativamente normale con media $p = 0.10$ ed errore standard $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.10(1-0.10)}{50}} = 0.0424$.

Cerchiamo l'area sottesa dalla curva normale a destra di $\hat{p} = 0.20$.

Introduciamo allora la variabile standardizzata

$$Z = \frac{\hat{p} - 0.10}{\sqrt{\frac{0.10(1-0.10)}{50}}}.$$

Sappiamo che

$$P(\hat{p} \geq 0.20) = P\left(Z \geq \frac{0.20 - 0.10}{\sqrt{\frac{0.10(1-0.10)}{50}}}\right) = P(Z \geq 2.36).$$

Dalla Tabella A.3, l'area sottesa dalla curva normale standardizzata a destra di 2.36 è 0.009.

Concludiamo che solo circa lo 0.9% dei campioni avrà una proporzione campionaria maggiore o uguale a 0.20.

6.3 Intervalli di confidenza per proporzioni

Sappiamo che la variabile casuale

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

segue la distribuzione normale standardizzata se n è sufficientemente grande. Quindi

$$P\left(-1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96\right) = 0.95$$

da cui otteniamo

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95.$$

Pertanto

$$\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

sono i limiti dell'intervallo di confidenza al 95% per la proporzione p della popolazione.

Poichè *non* conosciamo p, lo stimiamo utilizzando la proporzione campionaria $\hat{p}$. Pertanto

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

sono i limiti dell'intervallo di confidenza, approssimato per p, al 95%.

Analogamente, possiamo ricavare un intervallo di confidenza *unilaterale*:

$$P\left(p \leq \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$
$$\Rightarrow \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

è il *limite superiore* dell'intervallo di confidenza al 95% per p.

$$P\left(p \geq \hat{p} - 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$
$$\Rightarrow \hat{p} - 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

è il *limite inferiore* dell'intervallo di confidenza al 95% per p.

Esempio. Consideriamo la sopravvivenza a 5 anni dei pazienti di età inferiore a 40 anni cui è stato diagnosticato un tumore al polmone.

La distribuzione della proporzione di sopravvivenza ha una media p non nota.

In un campione di $n = 52$ pazienti solo 6 sopravvivono 5 anni. Allora una stima puntuale di p è

$$\hat{p} = \frac{x}{n} = \frac{6}{52} = 0.115.$$

Poichè

$$n\hat{p} = 52(0.115) = 5.98 > 5$$
$$n(1-\hat{p}) = 52(0.885) = 46.02 > 5$$

la dimensione del campione è sufficientemente grande per giustificare l'uso dell'approssimazione normale.

Gli estremi dell'intervallo di confidenza al 95% sono

$$0.115 \pm 1.96\sqrt{\frac{0.115(1-0.115)}{52}}$$

e quindi l'intervallo di confidenza approssimato è

$$(0.028, 0.202).$$

Siamo confidenti al 95% che questo intervallo contenga la *reale* proporzione di pazienti di età inferiore a 40 anni che sopravvivono 5 anni.

6.4 Test d'ipotesi per proporzioni

Esempio. La distribuzione della sopravvivenza a 5 anni dei pazienti sotto i 40 anni cui è stato diagnosticato un tumore al polmone ha una proporzione p non nota.

Sappiamo che la proporzione di pazienti di età superiore a 60 anni che sopravvivono 5 anni è pari a 0.082.

E' possibile che la proporzione di sopravvivenza a 5 anni per i pazienti al di sotto dei 40 anni sia 0.082?

Specificazioni

1. $H_0 : p = 0.082$
2. livello di significatività $\alpha = 0.01$
3. test bilaterale: siamo interessati alle deviazioni in entrambe le direzioni

Osservazione

Consideriamo il campione di $n = 52$ pazienti al di sotto dei 40 anni. La proporzione di sopravvivenza osservata è

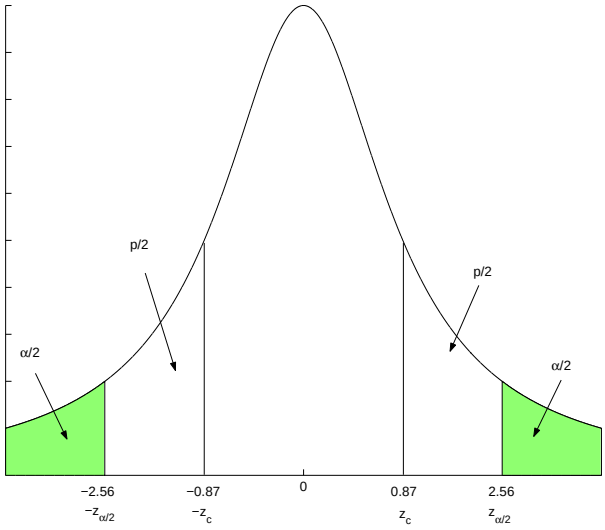
$$\hat{p} = \frac{6}{52} = 0.115.$$

Analisi

Calcoliamo il test z:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.115 - 0.082}{\sqrt{\frac{0.082(1-0.082)}{52}}} = 0.87.$$

Conclusione



$p > \alpha$, quindi non rifiutiamo H_0. Questo campione *non* fornisce evidenza di una differenza fra le proporzioni di sopravvivenza a 5 anni nei due gruppi di età.

Capitolo 7

Tabelle di contingenza

L'approssimazione normale alla distribuzione binomiale può essere utilizzata per eseguire test d'ipotesi per il confronto di 2 proporzioni nel caso di campioni indipendenti. Si possono ottenere gli stessi risultati con una diversa procedura, il *test* χ^2, che permette di confrontare 2 o più proporzioni.

Quando si lavora con dati *nominali* raggruppati in categorie, spesso le frequenze sono organizzate in un formato tabulare, noto come *tabella di contingenza*. Nel caso più semplice sono coinvolte 2 variabili dicotomiche. Le righe della tabella rappresentano i risultati di una delle 2 variabili e le colonne i risultati dell'altra variabile.

Lo scopo è quello di voler stabilire se esiste una qualche relazione tra le 2 variabili in questione.

L'**ipotesi nulla** H_0 è la seguente: *non* esiste alcuna relazione fra le 2 variabili.

7.1 Tabelle 2×2

Cominciamo con l'esaminare il caso di una tabella 2x2.

Si scrive la *tabella delle frequenze* **osservate**:

	Variabile 2		
Variabile 1	A	B	Totale
A	a	b	a+b
B	c	d	c+d
Totale	a+c	b+d	n

dove n è il numero di osservazioni eseguite.

Si scrive poi la *tabella delle frequenze* **attese**:

	Variabile 2		
Variabile 1	A	B	Tot.
A	$\frac{(a+b)(a+c)}{n}$	$\frac{(a+b)(b+d)}{n}$	a+b
B	$\frac{(c+d)(a+c)}{n}$	$\frac{(c+d)(b+d)}{n}$	c+d
Tot.	a+c	b+d	n

7.1.1 Come calcolare le frequenze attese

Sotto l'ipotesi nulla H_0, le proporzioni relative alla prima variabile e quelle relative alla seconda variabile sono uguali; quindi possiamo ignorare questa distinzione e trattare tutti gli n soggetti testati come un unico campione omogeneo. In questo campione la proporzione totale di soggetti relativi al caso A della prima variabile è

$$\frac{a+b}{n}$$

Per ottenere la frequenza attesa relativa al caso in cui entrambe le variabili presentano il dato A, devo allora moltiplicare il totale della prima colonna, cioè a+c, per la proporzione trovata prima:

$$(a+c) \cdot \frac{a+b}{n}$$

Ripeto questo procedimento per ogni cella della tabella di contingenza.

7.1.2 Come eseguire il test d'ipotesi

Siano O le frequenze *osservate* ed E le frequenze *attese*.

Approssimativamente, più grande è lo scarto $O - E$, tanto più valide sono le indicazioni che rifiutano l'ipotesi nulla. E' perciò ragionevole basare un test di ipotesi su questi scarti.

Il test appropriato da usare è il cosiddetto *test del chi quadro*:

$$\chi_1^2 = \sum_{i=1}^{2\cdot 2} \frac{(O_i - E_i)^2}{E_i} \tag{7.1}$$

dove $2 \cdot 2 = 4$ è il numero di celle della tabella e l'indice 1 di χ_1^2 è il numero dei *gradi di libertà* nel caso di un tabella 2x2.

NOTA BENE

Per garantire che la dimensione del campione sia abbastanza grande da rendere valida l'approssimazione del chi quadro (continua) con la sommatoria (7.1) (discreta), nessuna cella deve avere frequenza attesa minore di 1 e al massimo il 20% delle celle deve avere una frequenza attesa minore di 5.

Proprietà del chi quadro

1. non è una distribuzione simmetrica;

2. non può essere negativa;

3. può assumere valori da 0 ad infinito ed è asimmetrica a destra;

4. l'area totale sotto la curva è 1;

5. c'è una diversa distribuzione chi quadro per ogni possibile valore di gradi di libertà: se abbiamo pochi gradi di libertà allora c'è molta asimmetria; se abbiamo molti gradi di libertà allora vi è una minore asimmetria.

Esempio (Esempio 1)**.** Consideriamo la tabella 2×2 che illustra i risultati di uno studio sull'efficacia dei caschi protettivi per bicicletta nella prevenzione dei traumi cranici:

	Casco protettivo		
Trauma cranico	SI'	NO	Totale
SI'	17	218	235
NO	130	428	558
Totale	147	646	793

I dati si riferiscono ad un campione di 793 soggetti coinvolti in incidenti con la bicicletta nell'arco di 1 anno.

Vogliamo sapere se l'uso del casco protettivo modifica la proporzione dei traumi cranici in caso di incidente.

Specifichiamo il livello di significatività $\alpha = 0.05$.

Testiamo allora la seguente *ipotesi nulla* H_0:

la proporzione di soggetti che hanno riportato traumi cranici tra coloro che indossavano il casco è *uguale* alla proporzione di soggetti che hanno riportato traumi cranici tra coloro che non indossavano il casco.

Scriviamo ora la tabella delle frequenze attese, a partire da quella delle frequenze osservate:

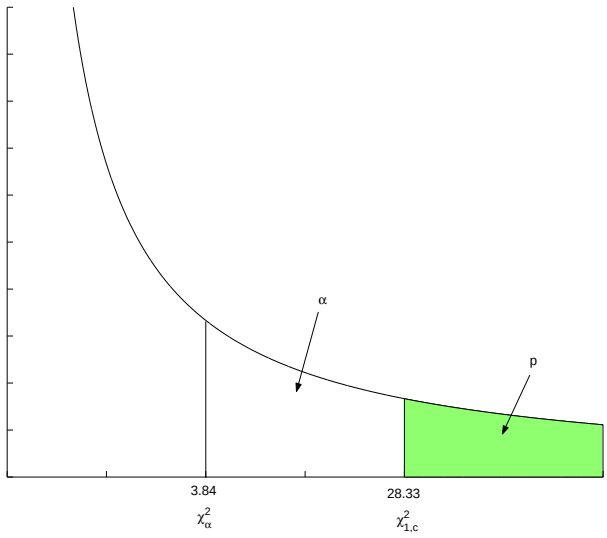
Freq.oss.	**Casco protettivo**		
Trauma cranico	SI'	NO	Totale
SI'	17	218	235
NO	130	428	558
Totale	147	646	793

Freq.attese	**Casco**		
Trauma	SI'	NO	Totale
SI'	$\frac{235 \cdot 147}{793} = 43.6$	$\frac{235 \cdot 646}{793} = 191.4$	235
NO	$\frac{558 \cdot 147}{793} = 103.4$	$\frac{558 \cdot 646}{793} = 454.6$	558
Totale	147	646	793

Le frequenze attese sono tutte maggiori di 5, quindi possiamo applicare il test χ^2. Abbiamo:

$$\begin{aligned}\chi^2_1 &= \frac{(17-43.6)^2}{43.6} + \frac{(130-103.4)^2}{103.4} + \\ &+ \frac{(218-191.4)^2}{191.4} + \frac{(428-454.6)^2}{454.6} = \\ &= 16.23 + 6.84 + 3.70 + 1.56 = \\ &= 28.33.\end{aligned}$$

Dalla Tavola A.8 troviamo che $\chi^2_\alpha = 3.84$ per 1 grado di libertà.



Si ha $p < \alpha$. Pertanto rifiutiamo H_0: la proporzione di soggetti che hanno riportato traumi cranici tra coloro che indossavano il casco è **diversa** da quella di soggetti che hanno riportato traumi cranici tra coloro che non indossavano il casco.

NOTA BENE

Osserviamo che il test è bilaterale anche se consideriamo una sola coda della distribuzione. Infatti è possibile ottenere grandi valori di $(O_i - E_i)^2$

quando la frequenza osservata è maggiore oppure minore della frequenza attesa.

Osservazioni.

1. Nel caso di tabelle 2x2, poichè il numero di gradi di libertà è molto basso, affinchè l'approssimazione (7.1) sia abbastanza valida è necessario applicare il cosiddetto *fattore di correzione di Yates*:

$$\chi^2_1 = \sum_{i=1}^{4} \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

   Secondo alcuni, però, questa correzione rende il test troppo conservativo ed induce a non rifiutare un'ipotesi nulla quando essa è falsa.

2. Nel caso di **tabelle rxc** con r righe e c colonne, la formula del chi quadro si generalizza facilmente alla seguente:

$$\chi^2_{(r-1)\cdot(c-1)} = \sum_{i=1}^{r\cdot c} \frac{(O_i - E_i)^2}{E_i},$$

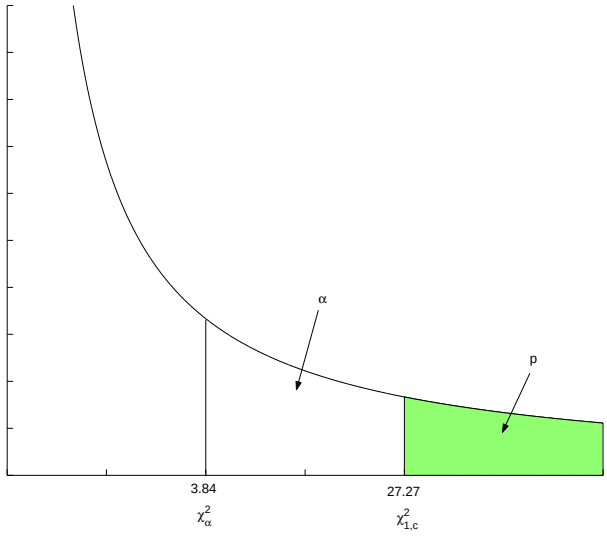
   dove l'indice $(r-1)\cdot(c-1)$ in $\chi^2_{(r-1)\cdot(c-1)}$ rappresenta il numero dei *gradi di libertà*.

   In questo modo il test χ^2 può essere utilizzato per effettuare il *confronto di 3 o più proporzioni*.

Esempio (continuazione Esempio 1)**.** Consideriamo il test del χ^2 sull'efficacia dell'uso dei caschi protettivi negli incidenti di bicicletta, con la correzione di Yates.

Abbiamo:

$$\begin{aligned}\chi^2_1 &= \frac{(|17-43.6|-0.5)^2}{43.6} + \frac{(|130-103.4|-0.5)^2}{103.4} + \\ &+ \frac{(|218-191.4|-0.5)^2}{191.4} + \frac{(|428-454.6|-0.5)^2}{454.6} = \\ &= 15.62 + 6.59 + 3.56 + 1.50 = \\ &= 27.27.\end{aligned}$$



Anche in questo caso $p < \alpha$: rifiutiamo dunque H_0.

Esempio. Consideriamo i dati relativi ad uno studio che esamina la accuratezza dei certificati di morte. I risultati di 575 autopsie sono stati confrontati con le cause di morte riportate nei certificati in 2 ospedali. Il primo è un ospedale pubblico (A), il secondo è un ospedale universitario (B).

I dati sono riportati in forma di tabella di contingenza 2×3:

	Certificato morte			
Osp.	accurato	carente	inesatto	Totale
A	157	18	54	229
B	268	44	34	346
Totale	425	62	88	575

Vogliamo stabilire se i risultati dello studio suggeriscono pratiche differenti nella compilazione dei certificati nei 2 ospedali.

Testiamo l'ipotesi nulla H_0:

le proporzioni delle diverse categorie di certificati nell'ospedale A sono uguali alle corrispondenti proporzioni nell'ospedale B.

Fissiamo il livello di significatività $\alpha = 0.05$.

Troviamo, con il solito calcolo, le frequenze attese a partire dalle frequenze osservate:

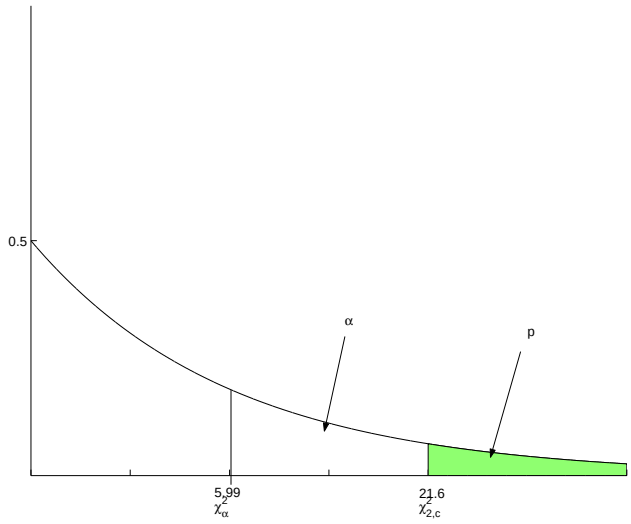
freq. oss.	**Certif. morte**			
Osp.	accurato	carente	inesatto	Totale
A	157	18	54	229
B	268	44	34	346
Totale	425	62	88	575

freq. att.	**Certif. morte**			
Osp.	accurato	carente	inesatto	Totale
A	$\frac{229 \cdot 425}{575} = 169.3$	24.7	35.0	229
B	255.7	37.3	53.0	346
Totale	425	62	88	575

Calcoliamo il test χ^2:

$$\chi^2_2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(157 - 169.3)^2}{169.3} + \frac{(18 - 24.7)^2}{24.7} + \\ + \frac{(54 - 35)^2}{35} + \frac{(268 - 255.7)^2}{255.7} + \frac{(44 - 37.3)^2}{37.3} + \\ + \frac{(34 - 53)^2}{53} = 21.62$$

Dalla Tabella A.8 troviamo che $\chi^2_{2,\alpha} = 5.99$.



Quindi $p < \alpha$: rifiutiamo H_0. Concludiamo che le proporzioni dei certificati di morte nell'ospedale A per le 3 categorie **non** sono uguali alle corrispondenti proporzioni nell'ospedale B.

7.2 Caso di 2 campioni appaiati

Esempio. Vogliamo testare se l'incidenza del diabete è la stessa fra gli individui che hanno subito un infarto e quelli che non hanno patologie cardiache.

144 soggetti con infarto vengono appaiati per età e sesso a 144 soggetti non affetti da patologie cardiache.

Ai membri di ciascuna coppia viene chiesto se gli era stato mai diagnosticato il diabete. I risultati sono riportati nella seguente tabella 2×2:

	Infarto		
Diabete	SI'	NO	Totale
SI'	46	25	71
NO	98	119	217
Totale	144	144	288

Poichè il test del χ^2 non considera l'appaiamento dei dati, in questa situazione non è appropriato. Classifichiamo allora i dati tenendo conto dell'appaiamento, sotto forma di tabella a doppia entrata:

	Infarto NO		
Infarto SI'	diabete SI'	diabete NO	Totale
diabete SI'	9	37	46
diabete NO	16	82	98
Totale	25	119	144

Ci sono 144 coppie, 144 è il totale complessivo della nuova tabella. Riportiamo nei totali marginali i corrispondenti elementi della tabella 2×2. Riportiamo nel corpo centrale della nuova tabella le risposte concordanti e quelle discordanti.

Dei 46 soggetti con infarto e diabetici, 9 sono stati appaiati a non affetti da infarto con diabete e 37 a non affetti da infarto e non diabetici. Dei 98 soggetti con infarto e non diabetici, 16 sono appaiati a diabetici senza infarto e 82 a non diabetici senza infarto.

Formuliamo l'ipotesi nulla H_0:

il numero di coppie in cui il soggetto con infarto è diabetico ed il soggetto appaiato non affetto da infarto non lo è, ***è uguale*** *al numero di coppie in cui il soggetto non affetto da infarto è diabetico ed il soggetto appaiato affetto da infarto non lo è,*

o, più brevemente:

non esiste alcuna associazione tra diabete ed infarto.

Fissiamo un livello di significatività $\alpha = 0.05$.

Le *coppie concordanti* (due diabetici o 2 non diabetici appaiati) non forniscono alcuna informazione per testare H_0. Pertanto ci concentriamo solo sulle *coppie discordanti* (soggetto diabetico appaiato ad un soggetto non diabetico).

Sia r il numero di coppie in cui il soggetto con infarto è diabetico ed il soggetto senza infarto non è diabetico; sia s il numero di coppie in cui il soggetto non affetto da infarto è diabetico ed il soggetto con infarto non è diabetico.

Se la differenza $|r - s|$ è grande, rifiutiamo l'ipotesi nulla H_0 di assenza di associazione.

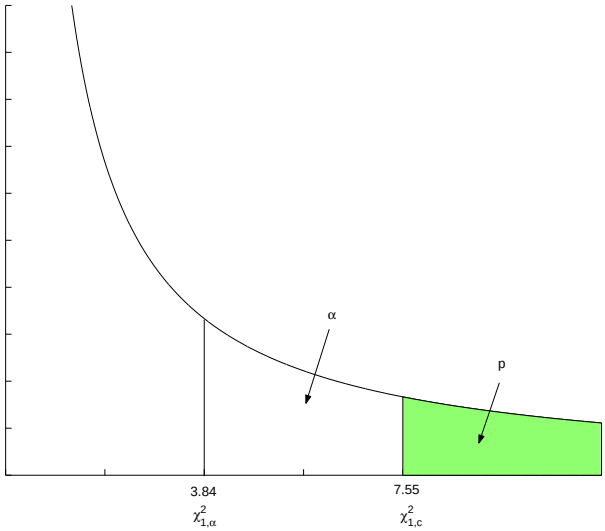
Nel nostro esempio $r = 37$ e $s = 16$.

Usiamo il *test di McNemar*:

$$\chi^2_1 = \frac{(|r - s| - 1)^2}{r + s} = \frac{(|37 - 16| - 1)^2}{37 + 16} = 7.55,$$

che segue una approssimata distribuzione χ^2 con 1 grado di libertà.

Dalla Tabella A.8 ricaviamo che $\chi^2_{1,\alpha} = 3.84$.



Dunque vediamo che $p < \alpha$. Rifiutiamo allora H_0 e concludiamo che l'incidenza del diabete nei soggetti con infarto è *diversa* da quella nei soggetti sani appaiati per età e per sesso.

Capitolo 8

Correlazione

E' il metodo appropriato per descrivere la relazione tra 2 variabili casuali normali mutuamente dipendenti, **ammesso che** la relazione tra esse sia **lineare**.

Siano X e Y le due variabili casuali mutuamente dipendenti e (x_i, y_i) il campione di punti selezionato dalle popolazioni originarie.

Esempio (Esempio 1)**.** Vogliamo esaminare la relazione tra la percentuale di bambini che sono stati vaccinati contro difterite, pertosse e tetano (DPT) in un dato Paese ed il tasso di mortalità al di sotto dei 5 anni.

Nella tabella di Fig. 8.1 sono riportati i dati di un campione casuale di 20 Paesi.

Se X rappresenta la percentuale di bambini vaccinati e Y il tasso di mortalità abbiamo una coppia di risultati (x_i, y_i), $i = 1, \ldots 20$, per ogni Paese. Possiamo rappresentare ciascun Paese con un punto del piano cartesiano.

Per esempio, la Bolivia sarà rappresentata dal punto di coordinate (0.40,0.165).

Esaminando il grafico dell'insieme di punti si osserva che il tasso di mortalità tende a diminuire all'aumentare della percentuale di bambini vaccinati.

8.1 Covarianza e coefficiente di correlazione di Pearson

Sia ρ la correlazione tra X e Y nelle popolazioni originarie.

Essa quantizza la forza della relazione lineare tra i risultati x_i e y_i:

$$\rho = \text{media} \left(\frac{X - \mu_X}{\sigma_X} \frac{Y - \mu_Y}{\sigma_Y} \right)$$

Sia r lo *stimatore* della correlazione delle popolazioni:

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_X}\right)\left(\frac{y_i-\overline{y}}{s_Y}\right) = \\ &= \frac{\mathrm{cov}(X,Y)}{s_X\, s_Y} = \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \end{aligned}$$

dove $\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$ è detta *covarianza* e s_X e s_Y sono le deviazioni standard campionarie di x_i e y_i rispettivamente.

r è detto *coefficiente di correlazione di Pearson*

Si dimostra che

$$-1 \leq r \leq 1$$

Quando $r = -1$ o $r = 1$ si ha *perfetta* correlazione lineare; ciò vale per tutte le rette tranne per quelle orizzontali o verticali.

Quando la relazione tra X e Y devia dalla perfetta linearità, r si allontana da -1 o da 1 e si avvicina a 0.

La grandezza di r è determinata dal grado di approssimazione con cui i punti tendono a disporsi lungo una linea retta.

Se $r = 0$ allora non esiste una relazione lineare tra X e Y (ma può esistere una relazione di altra natura matematica).

In Fig. 8.2 sono riportati alcuni esempi.

L'interpretazione di r dipende fondamentalmente dalle caratteristiche della ricerca e dall'estensione delle conoscenze che si hanno sulla materia oggetto di studio. L'esperienza precedente nel campo specifico serve comunemente come base di confronto per determinare se un particolare coefficiente di correlazione è degno di nota.

Abbiamo la seguente classificazione in base ai valori di r:

- $0 \leq r \leq 0.25$: poca o nessuna associazione lineare;
- $0.25 < r \leq 0.50$: discreto grado di associazione lineare;
- $0.50 < r \leq 0.75$: grado di associazione lineare tra moderato e buono;
- $r > 0.75$: grado di associazione lineare tra molto buono ed eccellente.

Esempio (continuazione Esempio 1)**.** Per i dati sui bimbi vaccinati contro DPT, la media campionaria dei bimbi vaccinati è

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 0.763.$$

La media campionaria del tasso di mortalità sotto i 5 anni è

$$\overline{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 0.0622.$$

Inoltre

$$\sum_{i=1}^{20}(x_i - \overline{x})(y_i - \overline{y}) = -0.2168$$
$$\sum_{i=1}^{20}(x_i - \overline{x})^2 = 0.7656$$
$$\sum_{i=1}^{20}(y_i - \overline{y})^2 = 0.08932$$

Quindi

$$r = \frac{-0.2168}{\sqrt{0.7656}\sqrt{0.08932}} = -0.829.$$

In base a questo campione sembra esserci una *forte* relazione lineare tra la percentuale di bambini vaccinati contro DPT ed il corrispondente tasso di mortalità al di sotto dei 5 anni.

Poichè $r < 0$, il tasso di mortalità diminuisce al crescere della percentuale di vaccinazioni.

NOTA BENE

Un efficace programma di vaccinazione potrebbe essere la principale causa della diminuzione del tasso di mortalità, ma potrebbe anche essere uno degli aspetti di un efficace sistema di assistenza sanitaria che, a sua volta, è la causa della diminuzione del tasso di mortalità.

8.2 Inferenza su ρ

Per determinare se esiste una correlazione tra le variabili casuali X e Y possiamo testare l'ipotesi nulla che NON esista correlazione nelle popolazioni originarie:

$$H_0 : \rho = 0$$

Fissiamo il livello di significatività α.

Posto che H_0 sia vera, ci chiediamo quale sia la probabilità di trovare un coefficiente di correlazione campionario di valore pari o maggiore di quello osservato.

Eseguiamo il test statistico

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}.$$

Se assumiamo che le coppie di osservazioni (x_i, y_i) siano state selezionate casualmente e che X e Y siano normalmente distribuite, sotto l'ipotesi nulla si dimostra che t_{n-2} segue una distribuzione t di Student con $n-2$ gradi di libertà.

Esempio (continuazione Esempio 1)**.** Supponiamo di voler sapere quanto è forte la relazione lineare tra le percentuali di vaccinati X ed il tasso di mortalità sotto i 5 anni Y.

Specificazione

1. $H_0 : \rho = 0$

2. livello di significatività $\alpha = 0.05$

3. test bilaterale: siamo interessati agli scarti da $\rho = 0$ in entrambe le direzioni.

Osservazione

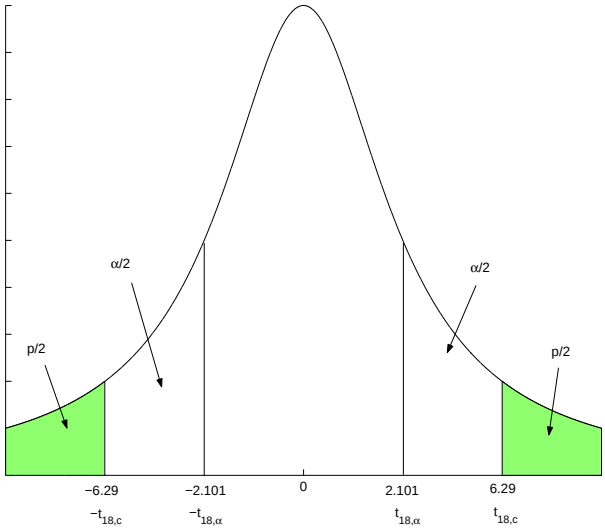
Dal campione di 20 Paesi abbiamo calcolato prima $r = -0.829$.

Analisi

Calcoliamo

$$t_{20-2} = t_{18} = -0.829\sqrt{\frac{20-2}{1-(-0.829)^2}} = -6.29.$$

Dalla Tabella A.4 troviamo che $t_{18,\alpha} = 2.101$.



Conclusione

$p < \alpha$, dunque rifiutiamo H_0: c'è evidenza che la correlazione reale nella popolazione sia diversa da 0. Il tasso di mortalità diminuisce linearmente all'aumentare della percentuale di bambini vaccinati.

8.3 Cautele interpretative

1. Il coefficiente di correlazione quantizza solo la relazione lineare tra X e Y; se X e Y hanno una relazione non lineare esso non fornisce una misura valida di questa relazione.

2. Il coefficiente di correlazione campionario r è molto sensibile a coppie di osservazioni molto al di fuori del campo di variabilità degli altri punti. Quindi se nei dati sono presenti uno o più valori estremi, r può fornire risultati fuorvianti.

3. La correlazione stimata non deve mai essere estrapolata al di fuori dell'intervallo in cui cadono i dati; la relazione fra X e Y infatti può essere diversa al di fuori di questo intervallo.

4. Un'alta correlazione tra due variabili non implica una relazione causa-effetto. Una variabile può influenzare o essere causa delle variazioni dell'altra, ma è anche possibile che qualche altra variabile o un'intera moltitudine di altre variabili possa influenzare le due che sono tra loro correlate.

   Pertanto X e Y possono essere correlate, ma semplicemente perchè ciascuna di esse è fortemente influenzata da una terza variabile Z.

Spesso si incontrano correlazioni che sembrano essere prive di senso o spurie tra 2 variabili che logicamente appaiono essere non correlate.

Esempio. E' stata calcolata un'alta correlazione positiva tra il numero di nidi di cicogna ed il numero di nascite nell'Europa nord occidentale. Le osservazioni sono state rilevate al variare del tempo (1800-1900). L'interpretazione più sensata che si può dare è che gli aumenti di popolazione comportano un aumento nel numero delle costruzioni che quindi forniscono più spazio per la nidificazione delle cicogne.

Capitolo 9

Regressione lineare

E' una tecnica utilizzata per studiare la natura della relazione tra 2 variabili continue.

La differenza rispetto alla correlazione consiste nel fatto che la regressione consente di esaminare la variazione di una variabile Y, detta *variabile di risposta*, che corrisponde ad una determinata variazione dell'altra X, detta *variabile esplicativa*.

L' obiettivo finale è quello di predire o stimare il valore della variabile di risposta associato ad un determinato risultato della variabile esplicativa.

Prima di eseguire qualsiasi analisi, è bene che il ricercatore esegua il grafico dei suoi dati e studi il tipo di relazione. Il *diagramma a punti* costituisce il metodo più conveniente di rappresentazione grafica: esso consiste in un grafico la cui ascissa rappresenta una variabile e l'ordinata l'altra variabile. Ciascuna delle n osservazioni, essendo composta da un valore x_i e y_i, è rappresentata da un punto sul grafico di coordinate (x_i, y_i) (Fig. 9.1).

La *regressione lineare* riguarda l'interpolazione di una linea retta tra punti come quelli della Fig. 9.2.

Esempio. In bambini di entrambi i sessi, la circonferenza cranica (Y) sembra aumentare in modo lineare dai 2 ai 18 anni di età (X), come si vede in Fig. 9.3 e Fig. 9.4.

Esempio. Y rappresenta le misurazioni in centimetri della circonferenza cranica dei neonati con peso alla nascita inferiore a 1500 grammi, X rappresenta l'età gestazionale misurata in settimane.

Sappiamo che la circonferenza cranica media dei neonati con basso peso alla nascita è $\mu_Y = 27$ cm e la deviazione standard è $\sigma_Y = 2.5$ cm. La distribuzione delle misurazioni è approssimativamente normale, quindi circa il 95% dei neonati ha circonferenza cranica contenuta nell'intervallo

$$(27 - 1.96(2.5), 27 + 1.96(2.5)) = (22.1, 31.9).$$

Supponiamo di sapere che

- le circonferenze craniche aumentano al crescere dell'età gestazionale con legge di tipo lineare;
- per ogni età gestazionale x la distribuzione delle misurazioni di circonferenza cranica $Y|x$ è approssimativamente normale.

Ad esempio, le circonferenze craniche dei neonati la cui età gestazionale è 26 settimane sono distribuite normalmente con media $\mu_{Y|26} = 24$ cm e deviazione standard $\sigma_{Y|26} = 1.6$ cm.

Analogamente $\mu_{Y|29} = 26.5$ cm e $\sigma_{Y|29} = 1.6$ cm.

Infine per i neonati nati dopo 32 settimane $\mu_{Y|32} = 29$ cm e $\sigma_{Y|32} = 1.6$ cm.

Per ciascun valore di età gestazionale x, la deviazione standard $\sigma_{Y|x}$ è costante e minore di σ_Y. Infatti è possibile dimostrare che

$$\sigma^2_{Y|x} = (1 - \rho^2)\,\sigma^2_Y,$$

dove ρ è la *correlazione* tra X e Y nelle popolazioni originarie. Se X e Y *non* hanno alcuna relazione lineare allora $\rho = 0$ e

$$\sigma^2_{Y|x} = \sigma^2_Y.$$

Per i dati sulla circonferenza cranica e l'età gestazionale abbiamo

$$(1.6)^2 = (1 - \rho^2)\,(2.5)^2,$$

da cui

$$\rho = \sqrt{1 - \frac{(1.6)^2}{(2.5)^2}} = \pm 0.768.$$

Esiste dunque una correlazione piuttosto forte tra circonferenza cranica ed età gestazionale nella popolazione originaria con basso peso alla nascita; utilizzando questo metodo, però, non possiamo stabilire se la correlazione è positiva o negativa.

Poichè $\sigma_{Y|x} < \sigma_Y$, considerare un singolo valore di età gestazionale ci consente di essere più precisi nelle nostre descrizioni.

Ad esempio, possiamo dire che circa il 95% dei valori della circonferenza cranica della popolazione dei neonati la cui età gestazionale è 26 settimane, è compreso nell'intervallo

$$(24 - 1.96(1.6), 24 + 1.96(1.6)) = (20.9, 27.1).$$

Inoltre, circa il 95% dei neonati la cui età gestaz. è 29 settimane ha circonferenze craniche comprese nell'intervallo

$$(26.5 - 1.96(1.6), 26.5 + 1.96(1.6)) = (23.4, 29.6),$$

mentre il 95% dei neonati la cui età gestaz. è 32 settimane ha misurazioni comprese nell'intervallo

$$(29 - 1.96(1.6), 29 + 1.96(1.6)) = (25.9, 32.1).$$

In sintesi, i rispettivi intervalli sono i seguenti:

Età gestaz.	**Intervallo con 95% delle osservazioni**
26	(20.9,27.1)
29	(23.4,29.6)
32	(25.9,32.1)

Ognuno di questi intervalli è calcolato in modo da includere il 95% dei valori delle circonferenze craniche della popolazione di neonati di una determinata età gestazionale. Nessuno di essi è ampio quanto (22.1,31.9), l'intervallo calcolato per l'intera popolazione di neonati con basso peso alla nascita. Inoltre, gli intervalli tendono verso destra all'aumentare dell'età gestazionale.

9.1 Retta di regressione della popolazione

Se riportiamo su un diagramma a 2 dimensioni i 3 punti di coordinate ($x = 26, \mu_{Y|26} = 24$), ($x = 29, \mu_{Y|29} = 26.5$) e ($x = 32, \mu_{Y|32} = 29$), troviamo che essi giacciono su una retta, cioè la relazione fra X e $\mu_{Y|x}$ è lineare (vedi Fig. 9.5).

La retta di equazione

$$\mu_{Y|x} = \alpha + \beta x,$$

dove $\mu_{Y|x}$ è la circonferenza cranica *media* dei neonati la cui età gestazionale è x settimane e x è l'età gestazionale misurata in settimane, si dice *retta di regressione della popolazione.*

α e β sono costanti chiamate *coefficienti della retta.*

α è l'intercetta con l'asse delle ordinate, cioè il valore di $\mu_{Y|x}$ quando $x = 0$.

β è la pendenza o coefficiente angolare. Essa è la variazione in $\mu_{Y|x}$ che corrisponde alla variazione di un'unità in x.

Anche se la relazione tra $\mu_{Y|x}$ e x è esattamente lineare, la relazione fra le singole misure di circonferenza e l'età gestazionale non lo è (vedi Fig. 9.1).

Le misure di circonferenza cranica dei neonati di una certa età gestazionale x sono distribuite approssimativamente in modo normale con media $\mu_{Y|x}$ e deviazione standard $\sigma_{Y|x}$. La dispersione rispetto alla media è il risultato della naturale variabilità fra i neonati.

Cerchiamo quindi una retta di equazione

$$\hat{y} = a + bx$$

che rappresenti una *stima* per la retta di regressione della popolazione

$$\mu_{Y|x} = \alpha + \beta x,$$

che noi, nelle applicazioni, non conosciamo.

Costruiremo la retta $\hat{y} = a + bx$ mediante un campione di osservazioni.

Facciamo le seguenti **ipotesi** nell'inferenza sulla retta di regressione:

1. Per un determinato valore di X, che si considera misurato senza errore, la distribuzione dei valori Y è *normale* con media $\mu_{Y|x}$ e deviazione standard $\sigma_{Y|x}$ (vedi Fig. 9.6).

2. La relazione tra $\mu_{Y|x}$ e x è descritta dalla retta

$$\mu_{Y|x} = \alpha + \beta x.$$

3. Per ogni determinato valore di X, $\sigma_{Y|x}$ non cambia (*omoschedasticità*) (vedi Fig. 9.6).

4. I risultati y_i sono *indipendenti*.

9.2 Retta di regressione campionaria

Consideriamo il diagramma a punti delle circonferenze craniche in funzione dell'età gestazionale per un campione di neonati con basso peso alla nascita: la variabile esplicativa è sull'asse orizzontale,quella di risposta sull'asse verticale (vedi Fig. 9.1).

Come secondo esempio consideriamo il diagramma a punti delle pressioni sistoliche in funzione dell'età per un campione di 33 donne (vedi Fig. 9.2).

Nei 2 diagrammi, i singoli punti variano molto, ma il profilo generale suggerisce che al crescere della variabile esplicativa X la variabile di risposta Y tenda a crescere con andamento lineare.

Stimiamo i coefficienti della retta di regressione utilizzando un singolo campione di misurazioni. Abbiamo infiniti modi per far passare una retta attraverso la nuvola di punti dei dati. E' necessario un criterio per stabilire

quale delle rette descrive meglio la relazione tra la circonferenza cranica e l'età gestazionale, oppure tra la pressione sistolica e l'età.

Il criterio che conduce alla retta di migliore approssimazione per l'insieme di punti è una tecnica matematica nota come **metodo dei minimi quadrati**. La retta che stiamo cercando è decritta dall'equazione $\hat{y} = a+bx$.

Il metodo dei minimi quadrati consiste nel *determinare i 2 coefficienti della retta a e b in modo tale che risulti minima la somma dei quadrati delle distanze verticali fra le osservazioni (x_i, y_i) e la retta $\hat{y} = a + bx$* (vedi Fig. 9.2).

Indichiamo con e_i la distanza verticale tra il punto (x_i, y_i) e la retta $\hat{y} = a + bx$:

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i.$$

e_i è anche detto **residuo**.

Se tutti i residui fossero uguali a 0, allora tutti i punti (x_i, y_i), $i = 1, \ldots, n$, si troverebbero sulla retta di regressione. Questo è impossibile nei casi reali.

Consideriamo allora la somma dei quadrati dei residui, detta *devianza* o *devianza residua*:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Il metodo dei minimi quadrati determina i coefficienti a e b che rendono minima la devianza.

I valori calcolati a e b sono le stime per i coefficienti α e β della retta di regressione della popolazione.

Operando con il metodo dei minimi quadrati si trovano:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathrm{cov}(X, Y)}{s_X^2} : \textbf{pendenza}$$

$$a = \overline{y} - b\overline{x} : \textbf{intercetta}$$

dove $\overline{x}$ e $\overline{y}$ sono le medie campionarie e s_X^2 è la deviazione standard campionaria di X.

Otteniamo così l'equazione della retta di regressione:

$$\hat{y} = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x})$$

Osserviamo che la retta di regressione passa sempre per il punto di coordinate $(\overline{x}, \overline{y})$ che si chiama **baricentro** dell'insieme di punti.

Dopo aver trovato a e b possiamo sostituire diversi valori di x nell'equazione della retta e trovare i corrispondenti valori di y (valori *predetti*) sulla retta.

Esempio (Esempio 1)**.** In Fig. 9.7 è tracciata la retta di regressione ai minimi quadrati per i dati relativi alla circonferenza cranica e all'età gestazionale. L'equazione della retta è

$$\hat{y} = 3.9143 + 0.7801x.$$

Questa retta ha una devianza che è minore di quella di qualsiasi altra retta che può essere tracciata attraverso la nuvola di punti.

L'intercetta sull'asse verticale è 3.9143. Questo valore è, in teoria, il valore predetto della circonferenza cranica ad un'età gestazionale di 0 settimane. In questo esempio un'età di 0 settimane non ha alcun significato reale. L'intervallo in cui variano i dati è lontano da $x = 0$ settimane, quindi $\hat{y}(0)$ è un' *estrapolazione* rispetto a tale intervallo.

La pendenza della retta è 0.7801. Questo significa che per ogni settimana gestazionale la circonferenza cranica aumenta mediamente di 0.7801 cm.

9.3 Inferenza sulla retta di regressione

Si devono pensare ripetuti campioni di n coppie di osservazioni estratte dalla popolazione originaria. Si calcola la retta di regressione di ogni campione. Nella popolazione originaria esiste una particolare relazione di regressione lineare dei valori di Y sui valori di X, che è naturalmente sconosciuta.

Quali conclusioni si possono trarre circa la relazione sconosciuta nella popolazione, in base a quella determinata dalla retta di regressione interpolata tra i dati campionari?

Sarà necessario considerare la fluttuazione di campionamento delle quantità calcolate. Pertanto l'inferenza attinente la retta di regressione richiede la determinazione degli errori standard della pendenza e dell'intercetta.

Ricordiamo che la retta di regressione dei minimi quadrati è

$$\hat{y} = a + bx$$

mentre la retta di regressione della popolazione originaria è

$$\mu_{Y|x} = \alpha + \beta x.$$

a è una stima dell'intercetta α, b è una stima della pendenza β.

Si dimostra che gli errori standard (es) relativi all'intercetta a e alla pendenza b sono dati da

$$es(b) = \frac{\sigma_{Y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$es(a) = \sigma_{Y|x}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

es(a) e es(b) dipendono dalla deviazione standard dei valori di Y per un dato x ($\sigma_{Y|x}$).

Il problema è che non conosciamo $\sigma_{Y|x}$. Pertanto stimiamo $\sigma_{Y|x}$ attraverso la seguente quantità:

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

detta *deviazione standard della regressione.*

Allora utilizzeremo nei calcoli:

$$\hat{es}(b) = \frac{s_{Y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$\hat{es}(a) = s_{Y|x}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

come stime di $es(a)$ e di $es(b)$.

Esempio (continuazione Esempio 1)**.** Per il campione relativo alla circonferenza cranica in funzione dell'età gestazionale abbiamo

$$s_{Y|x} = 1.5904$$
$$\hat{es}(b) = 0.0631$$
$$\hat{es}(a) = 1.8291.$$

9.3.1 Test d'ipotesi sulla pendenza

La pendenza è di solito il coefficiente più importante nella retta di regressione in quanto essa fornisce l'informazione di base sulla relazione tra X e Y, precisamente la variazione media di Y per unità di variazione di X.

L' ipotesi nulla è la seguente:

$$H_0 : \beta = \beta_0$$

Il test da usare è il seguente:

$$t_{n-2} = \frac{b - \beta_0}{\hat{es}(b)}$$

che segue una distribuzione t di Student con $n-2$ gradi di libertà.

Spesso testiamo $\beta_0 = 0$ cioè che nella popolazione originaria la media di Y non varia al variare di X.

Esempio (continuazione Esempio 1)**.** Abbiamo visto che la retta di regressione per il campione relativo alla circonferenza cranica in funzione dell'età gestazionale è

$$\hat{y} = 3.9143 + 0.7801x.$$

Testiamo l'ipotesi nulla

$$H_0 : \beta = 0$$

con un livello di significatività $\alpha = 0.05$. Il campione ha dimensione $n = 100$. Eseguiamo un test bilaterale:

$$t_{100-2} = t_{98} = \frac{b}{\hat{es}(b)} = \frac{0.7801}{0.0631} = 12.36.$$

Sulla Tavola A.4 troviamo che $t_{98,\frac{\alpha}{2}} = 1.98$. Quindi riscontriamo che $p < \alpha$. Allora rifiutiamo H_0 e quindi concludiamo che nella popolazione originaria di neonati con basso peso alla nascita esiste una relazione lineare statisticamente significativa tra circonferenza cranica ed età gestazionale.

9.3.2 Intervallo di confidenza per la pendenza

Esempio (continuazione Esempio 1)**.** Per t_{98} circa il 95% delle osservazioni cade nell'intervallo (-1.98,1.98).

Quindi l'intervallo di confidenza al 95% per β è

$$\begin{aligned}
&(b - 1.98\,\hat{es}(b), b + 1.98\,\hat{es}(b)) = \\
&= (0.7801 - 1.98(0.0631), 0.7801 + 1.98(0.0631)) = \\
&= (0.6564, 0.9038)
\end{aligned}$$

Siamo confidenti al 95% che questo intervallo comprenda la pendenza β della retta di regressione.

Osserviamo che $\beta_0 = 0 \notin (0.6564, 0.9038)$, in accordo con quanto trovato nel test d'ipotesi al 5%.

9.3.3 Test d'ipotesi sulla intercetta

L' ipotesi nulla è la seguente:

$$H_0 : \alpha = \alpha_0$$

Il test da usare è il seguente:

$$t_{n-2} = \frac{a - \alpha_0}{\hat{es}(a)}$$

che segue una distribuzione t di Student con $n-2$ gradi di libertà.

Nell'eseguire le inferenze sulla intercetta, si dovrebbe per prima cosa notare se l'intercetta cade entro l'intervallo dei punti osservati. Quando i punti sono molto distanti da $x = 0$, la determinazione dell'intercetta implica una considerevole estrapolazione della retta. Tale estrapolazione è quanto meno rischiosa e spesso del tutto non attendibile.

9.4 Come valutare il modello lineare

Si può valutare il modello lineare seguendo sostanzialmente 3 strade:

1. attraverso il calcolo del *coefficiente di determinazione* R^2;
2. attraverso il *grafico dei residui* in funzione dei valori predetti della variabile di risposta Y;
3. attraverso opportune trasformazioni di una delle 2 variabili.

9.4.1 Il coefficiente di determinazione R^2

R^2 è definito come

$$R^2 = r^2$$

dove r è il coefficiente di correlazione di Pearson.

Si ha che $0 \leq R^2 \leq 1$.

Esso rappresenta la proporzione di variabilità tra i valori osservati di Y che è spiegata dalla regressione lineare di Y su X.

Esempio (continuazione Esempio 1)**.** Nel nostro esempio troviamo che

$$R^2 = 0.6095.$$

Quindi il 60.95% della variazione fra i valori osservati della circonferenza cranica è dovuto alla sua relazione lineare con l'età gestazionale. Il restante $(100 - 60.95)\% = 39.05\%$ della variazione non rimane spiegato.

9.4.2 Il grafico dei residui

Esempio (continuazione Esempio 1)**.** Il primo bambino nel campione di neonati con basso peso alla nascita ha un'età gestaz. $x_1 = 29$ settimane ed una circonferenza cranica $y_1 = 27$ cm. Il valore della circonferenza cranica sulla retta di regressione è

$$\hat{y}_1 = 3.9143 + 0.7801(29) = 26.536\,\text{cm}.$$

Il residuo di questa prima osservazione è

$$e_1 = y_1 - \hat{y}_1 = 27 - 26.536 = 0.464,$$

quindi il punto di coordinate $(26.536, 0.464)$ sarà incluso nel grafico dei residui. In Fig. 9.8(a) è riportato il diagramma dei punti $(\hat{y}_i, e_i)$,$i = 1, \ldots, n$.

Il grafico dei residui presenta 3 obiettivi:

1. aiuta ad individuare le osservazioni atipiche del campione;

   Nella Fig. 9.8(a) il residuo maggiore è associato ad un bambino la cui età gestaz. è 31 settimane e la cui circonferenza cranica è 35 cm (vedi Fig. 9.7).

   Il valore di circonferenza cranica sulla retta di regressione è

   $$\hat{y} = 3.914 + 0.7801(31) = 28.10\,\text{cm}.$$

   Il metodo dei minimi quadrati è molto sensibile ai valori atipici dei dati. Quando si ritiene che il valore atipico sia dovuto ad un errore di misura, la rimozione di questo punto migliora l'adattamento della retta di regressione ai dati. Occorre naturalmente essere molto cauti evitando di eliminare punti insoliti che sono in realtà validi o addirittura i più interessanti della serie di dati.

2. suggerisce un errore nell'assunzione di omoschedasticità.

   Se il range di ampiezza dei residui aumenta o diminuisce all'aumentare di Y allora $\sigma_{Y|x}$ non è costante per i valori di X (vedi Fig. 9.8(b)). In questo caso la regressione lineare semplice non è la tecnica corretta per rappresentare la variazione di Y in funzione di X;

3. può suggerire che la reale relazione tra X e Y non è lineare, se i residui non presentano una dispersione casuale ma seguono un andamento preciso.

   In questo caso può essere utile una trasformazione di X o di Y.

9.5 Trasformazioni

Trasformare una variabile significa misurarla su una scala diversa.

In molte situazioni una relazione curvilinea può essere trasformata in una lineare. In questo caso possiamo usare la regressione lineare sui dati trasformati.

In Fig. 9.9(a) è rappresentato il diagramma di dispersione del tasso di natalità in funzione del prodotto nazionale lordo (PNL) in 127 Paesi.

Il tasso di natalità diminuisce al crescere del PNL, ma con legge NON lineare.

Se vogliamo descrivere la relazione tra tasso di natalità e PNL utilizzando la regressione lineare, dobbiamo fare qualche trasformazione (vedi Fig. 9.9(b)).

Quando la relazione fra X e Y non è lineare, consideriamo le trasformazioni (preferibilmente in X):

$$x' = x^p$$

oppure

$$y' = y^p$$

con

$$p = \ldots, -3, -2, -1, -\frac{1}{2}, \frac{1}{2}, 1, 2, 3, \ldots$$

oppure

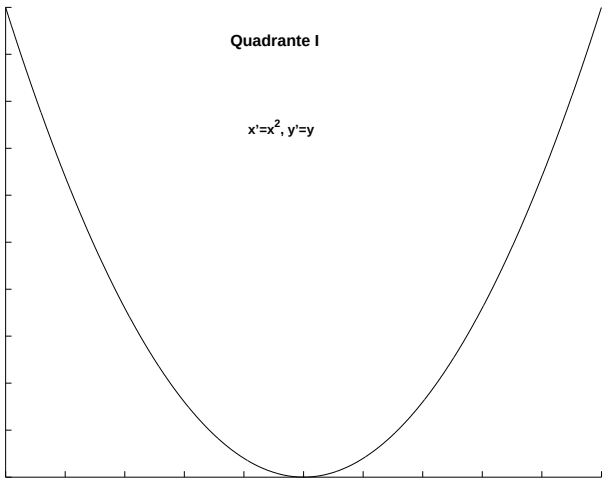
$$x' = \log x$$

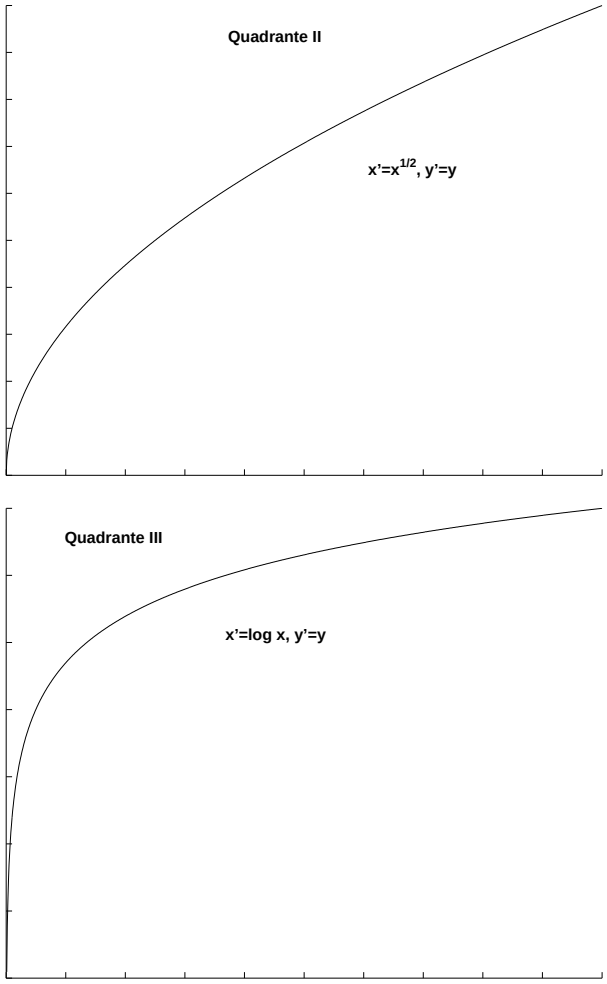
oppure

$$y' = \log y$$

dove log indica il *logaritmo naturale*.

Il *cerchio delle potenze* o *scala delle potenze* aiuta a scegliere il tipo di trasformazione da usare in base all'andamento del diagramma di dispersione (vedi Fig. 9.10).





9.5.1 Trasformazione logistica

La Fig. 9.11 indica una relazione di tipo sigmoidale.

Tali curve sono frequenti in farmacologia, dove

- l'ascissa x rappresenta una serie di dosi (o il logaritmo delle dosi) somministrata a gruppi di animali in laboratorio;
- l'ordinata y rappresenta la percentuale di animali che rispondono, in ciascun gruppo, alla dose somministrata.

A dosi molto basse, non risponde alcun animale. Si raggiunge un livello di dose a cui gli animali cominciano a rispondere e la percentuale dei rispondenti aumenta finchè a tutte le dosi alte tutti gli animali rispondono.

Applicando ai dati la trasformazione detta *trasformazione logistica*

$$\begin{cases} x' = x \\ y' = \log \frac{y}{100-y} \end{cases}$$

si ottiene nel piano (x', y') una relazione lineare (vedi Fig. 9.11).

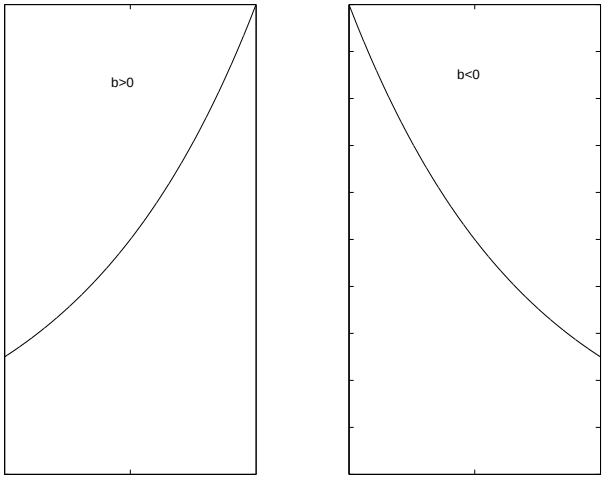
Quindi la relazione fra la dose e la risposta può essere determinata con una regressione lineare fra la risposta logistica (y') e la dose $(x' = x)$.

9.5.2 Regressione non lineare

Quando sappiamo che l'andamento dei dati non è lineare e conosciamo la legge funzionale, possiamo applicare ancora il metodo dei minimi quadrati che diventa più complesso e richiede una grande mole di calcoli. In alcuni casi è possibile rendere lineare il modello mediante una trasformazione.

Supponiamo ad esempio che i dati seguano una legge di tipo **esponenziale**:

$$y = c\,a^{bx}, \quad a > 1 \text{ noto.}$$



Nella relazione esponenziale

$$y = c\,a^{bx} \tag{9.1}$$

calcoliamo il logaritmo (naturale o in base 10) dei 2 membri e applichiamo le proprietà dei logaritmi:

$$\log y = \log(c\,a^{bx}) = \log c + bx \log a.$$

Operiamo ora le seguenti trasformazioni:

$$\begin{cases} x' = x \\ y' = \log y \end{cases} \qquad \begin{cases} a' = \log c \\ b' = b \log a. \end{cases} \tag{9.2}$$

Otteniamo così la relazione

$$y' = a' + b'x'$$

che è di tipo lineare.

Applichiamo il metodo dei minimi quadrati con il modello lineare ai dati trasformati secondo la (9.2) e troviamo a' e b'.

Le relazioni (9.2) ci permettono inoltre di ricavare c e b che, sostituiti nella (9.1), ci danno la curva esponenziale che approssima la nuvola dei dati originali, rendendo minima la devianza.

Vediamo altri casi riconducibili a quello lineare:

1. $$y = \frac{1}{bx + a}$$
   Da qui otteniamo che $\frac{1}{y} = bx + a$ e operando le seguenti trasformazioni
   $$\begin{cases} x' = x \\ y' = \frac{1}{y} \end{cases} \qquad \begin{cases} a' = a \\ b' = b \end{cases}$$
   ricaviamo il modello linearizzato
   $$y' = a' + b'x'. \tag{9.3}$$

2. $$y = \frac{x}{b + ax}$$
   Da qui otteniamo $\frac{1}{y} = \frac{b+ax}{x} = b\frac{1}{x} + a$ e operando le seguenti trasformazioni
   $$\begin{cases} x' = \frac{1}{x} \\ y' = \frac{1}{y} \end{cases} \qquad \begin{cases} a' = a \\ b' = b \end{cases}$$
   otteniamo il modello linearizzato (9.3).

3. $$y = b \log x + a$$
   Operando le seguenti trasformazioni
   $$\begin{cases} x' = \log x \\ y' = y \end{cases} \qquad \begin{cases} a' = a \\ b' = b \end{cases}$$
   otteniamo il modello linearizzato (9.3).

4. $$y = ax^b$$
   Da qui otteniamo $\log y = \log a + b \log x$ e operando le seguenti trasformazioni
   $$\begin{cases} x' = \log x \\ y' = \log y \end{cases} \qquad \begin{cases} a' = \log a \\ b' = b \end{cases}$$
   otteniamo il modello linearizzato (9.3).

Bibliografia

1. P.Armitage, G. Berry, Statistica medica, McGraw-Hill, 1996
2. T. Colton, Statistica in medicina, Piccin Editore, 1979
3. M. Pagano, K. Gauvreau, Biostatistica, seconda edizione, Idelson-Gnocchi, 2003

Appendice 1: Tabelle delle distribuzioni

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

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
0,1	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
0,2	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
0,3	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
0,4	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
0,5	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
0,6	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
0,7	0,242	0,239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
0,8	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
0,9	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
1,0	0,159	0,156	0,154	0,152	0,149	0,147	0,145	0,142	0,140	0,138
1,1	0,136	0,133	0,131	0,129	0,127	0,125	0,123	0,121	0,119	0,117
1,2	0,115	0,113	0,111	0,109	0,107	0,106	0,104	0,102	0,100	0,099
1,3	0,097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0,084	0,082
1,4	0,081	0,079	0,078	0,076	0,075	0,074	0,072	0,071	0,069	0,068
1,5	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
1,6	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0,047	0,046	0,046
1,7	0,045	0,044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
1,8	0,036	0,035	0,034	0,034	0,033	0,032	0,031	0,031	0,030	0,029
1,9	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
2,0	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0,019	0,018
2,1	0,018	0,017	0,017	0,017	0,016	0,016	0,015	0,015	0,015	0,014
2,2	0,014	0,014	0,013	0,013	0,013	0,012	0,012	0,012	0,011	0,011
2,3	0,011	0,010	0,010	0,010	0,010	0,009	0,009	0,009	0,009	0,008
2,4	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,007	0,007	0,006
2,5	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
2,6	0,005	0,005	0,004	0,004	0,004	0,004	0,004	0,004	0,004	0,004
2,7	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
2,8	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
2,9	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	0,001	0,001
3,0	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
3,1	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
3,2	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
3,3	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3,4	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

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	30,37	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

	Dimensione del campione										
T_0	2	3	4	5	6	7	8	9	10	11	12
1	0,5000	0,2500	0,1250	0,0625	0,0313	0,0157	0,0079	0,0040	0,0020	0,0010	0,0005
2		0,3750	0,1875	0,0938	0,0469	0,0235	0,0118	0,0059	0,0030	0,0015	0,0008
3		0,6250	0,3125	0,1563	0,0782	0,0391	0,0196	0,0098	0,0049	0,0025	0,0013
4			0,4375	0,2188	0,1094	0,0547	0,0274	0,0137	0,0069	0,0035	0,0018
5			0,5625	0,3125	0,1563	0,0782	0,0391	0,0196	0,0098	0,0049	0,0025
6				0,4063	0,2188	0,1094	0,0547	0,0274	0,0137	0,0069	0,0035
7				0,5000	0,2813	0,1485	0,0743	0,0372	0,0186	0,0093	0,0047
8					0,3438	0,1875	0,0977	0,0489	0,0245	0,0123	0,0062
9					0,4219	0,2344	0,1250	0,0645	0,0323	0,0162	0,0081
10					0,5000	0,2891	0,1563	0,0821	0,0420	0,0210	0,0105
11						0,3438	0,1915	0,1016	0,0528	0,0269	0,0135
12						0,4063	0,2305	0,1250	0,0655	0,0337	0,0171
13						0,4688	0,2735	0,1504	0,0801	0,0416	0,0213
14						0,5313	0,3204	0,1797	0,0967	0,0508	0,0262
15							0,3711	0,2129	0,1163	0,0616	0,0320
16							0,4219	0,2481	0,1377	0,0738	0,0386
17							0,4727	0,2852	0,1612	0,0875	0,0462
18							0,5274	0,3262	0,1875	0,1031	0,0550
19								0,3672	0,2159	0,1202	0,0647
20								0,4102	0,2461	0,1392	0,0757
21								0,4551	0,2784	0,1602	0,0882
22								0,5000	0,3125	0,1827	0,1019
23									0,3477	0,2066	0,1167
24									0,3848	0,2325	0,1331
25									0,4229	0,2598	0,1507
26									0,4610	0,2886	0,1697
27									0,5000	0,3189	0,1902
28										0,3501	0,2120
29										0,3824	0,2349
30										0,4156	0,2593
31										0,4493	0,2847
32										0,4830	0,3111
33										0,5171	0,3387
34											0,3667
35											0,3956
36											0,4251
37											0,4549
38											0,4849
39											0,5152

Tabella A.8 Percentili della distribuzione chi-quadrato

	Area nella coda superiore				
gl	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	Suicidio	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

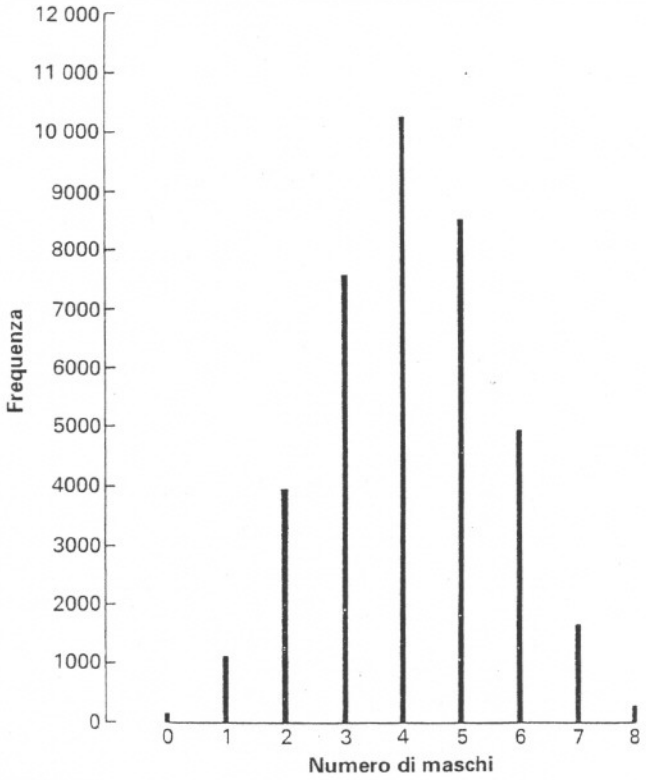

12 000
11 000
10 000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequenza
0
1
2
3
4
5
6
7
8
Numero di maschi

Fig. 1.1

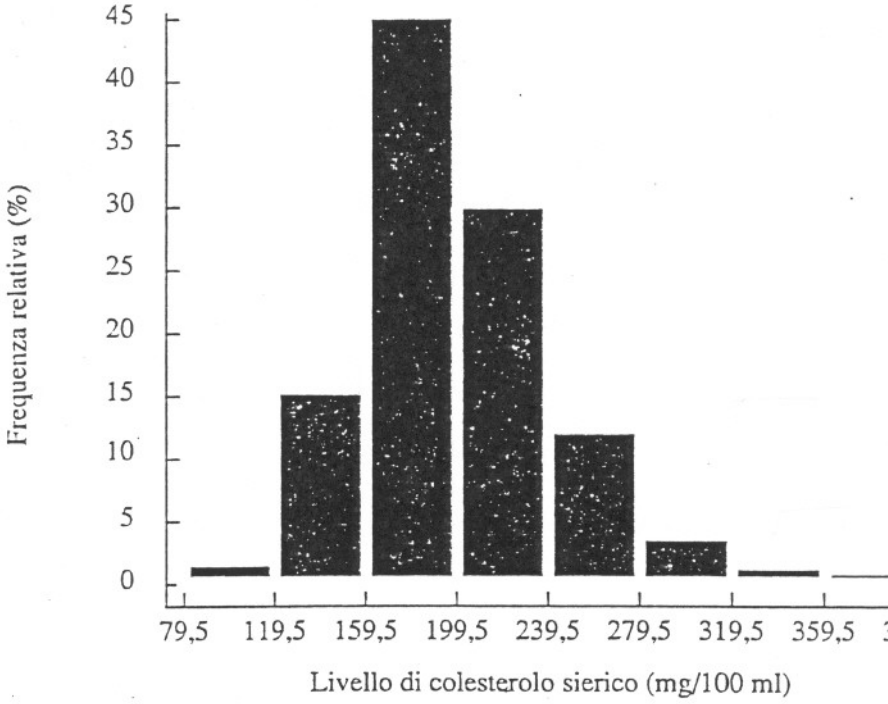


Figura 1.2 Istogramma: frequenze relative dei livelli di colesterolo sierico in 1.067 soggetti della popolazione maschile degli Stati Uniti di età compresa tra 25 e 34 anni, 1976-1980

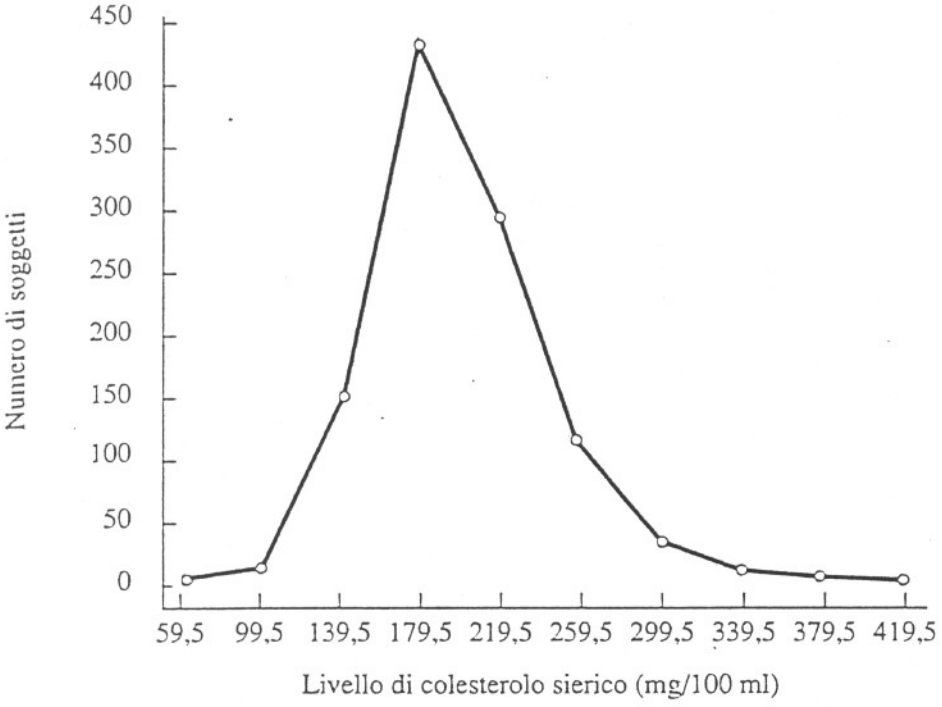


Figura 1.3 Poligono di frequenza: frequenze assolute dei livelli di colesterolo sierico in 1.067 soggetti della popolazione maschile degli Stati Uniti di età compresa tra 25 e 34 anni, 1976-1980

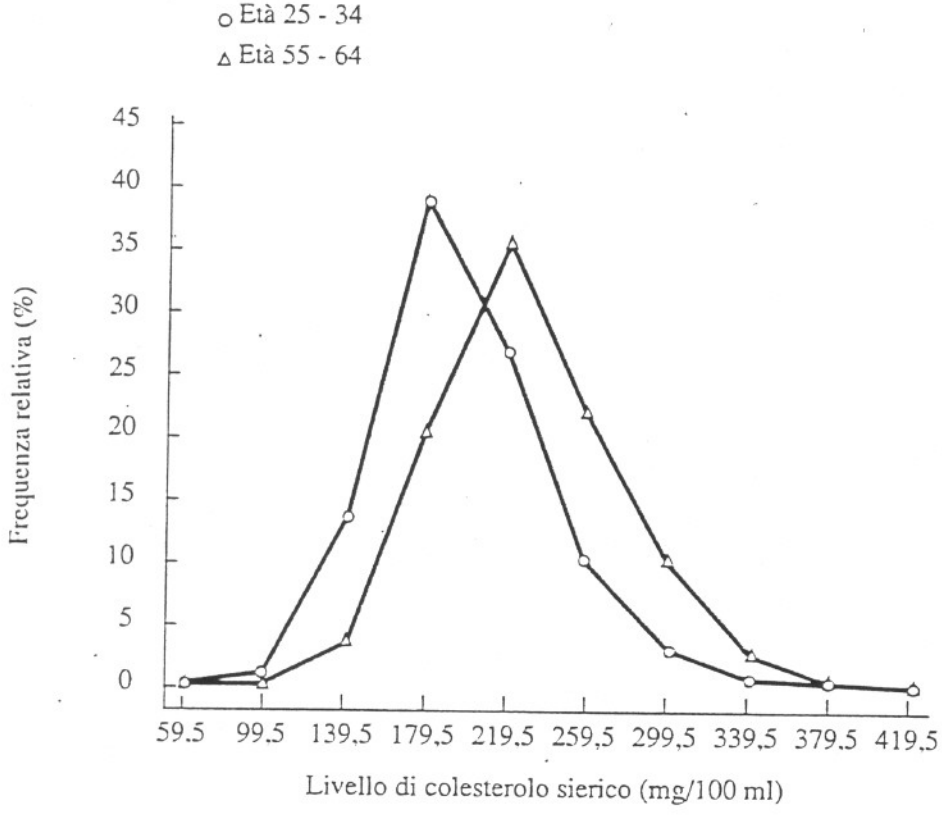


Figura 1.4 Poligono di frequenza: frequenze relative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

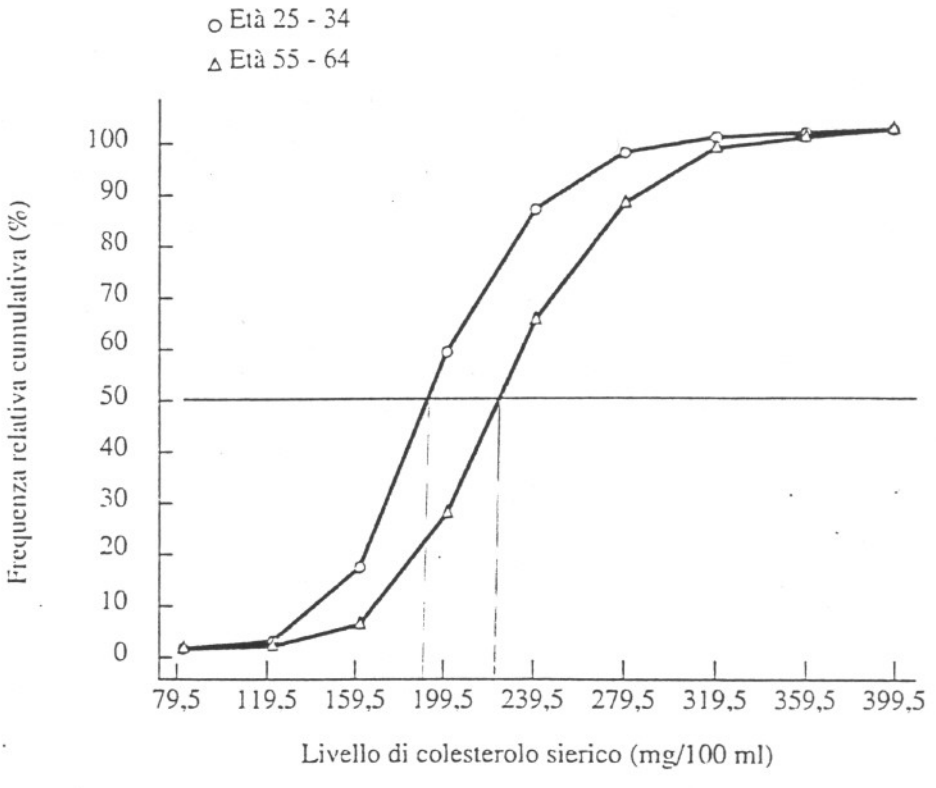


Figura 1.5 Poligono di frequenza cumulativa: frequenze relative cumulative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

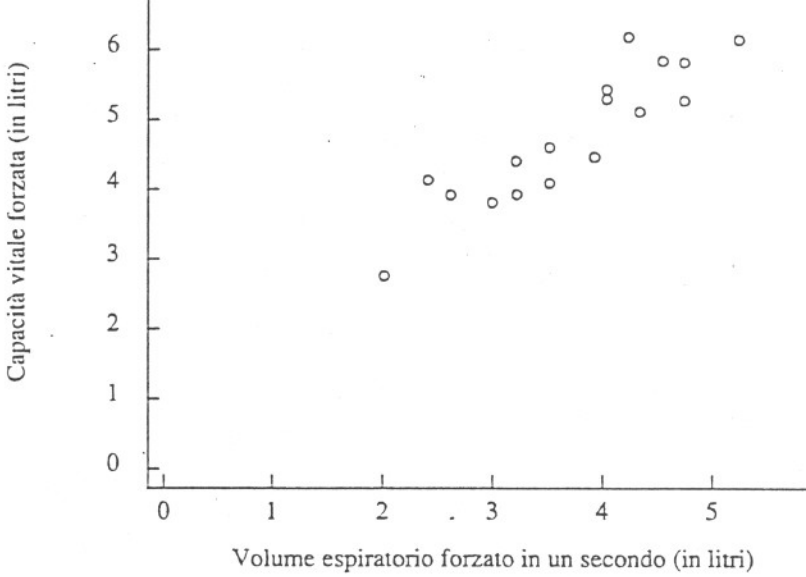


Figura 1.6 Diagramma a punti: capacità vitale forzata in funzione del volume espiratorio forzato in un secondo in 19 soggetti asmatici

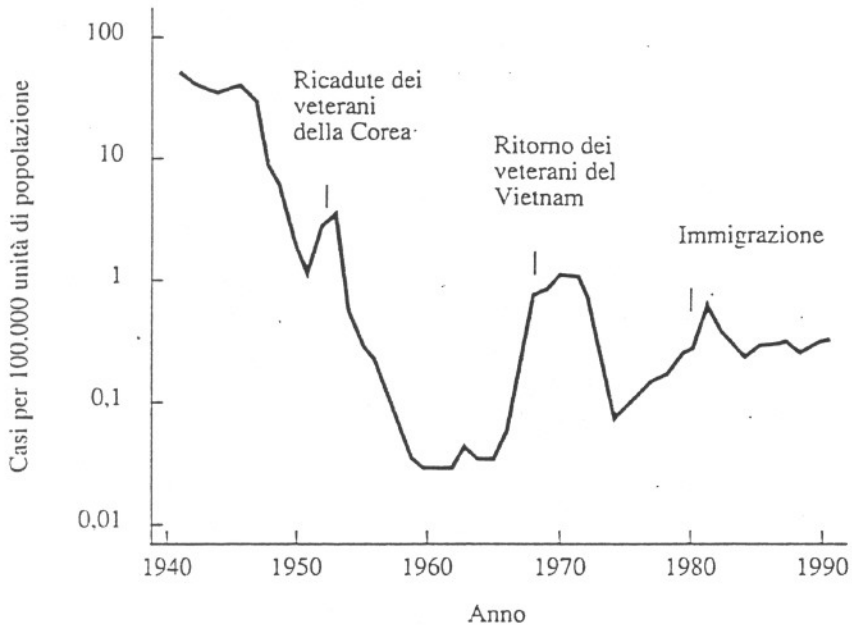
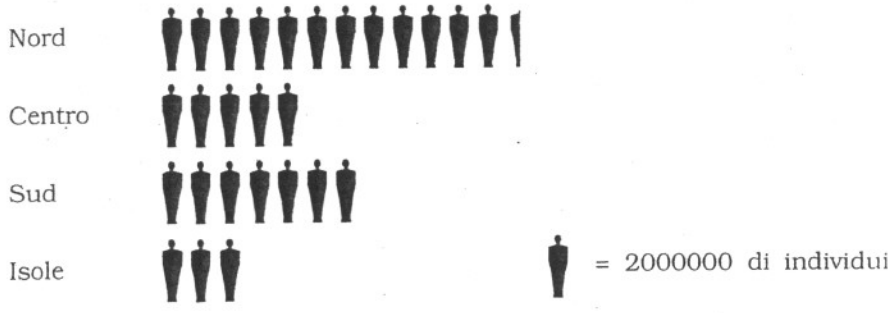
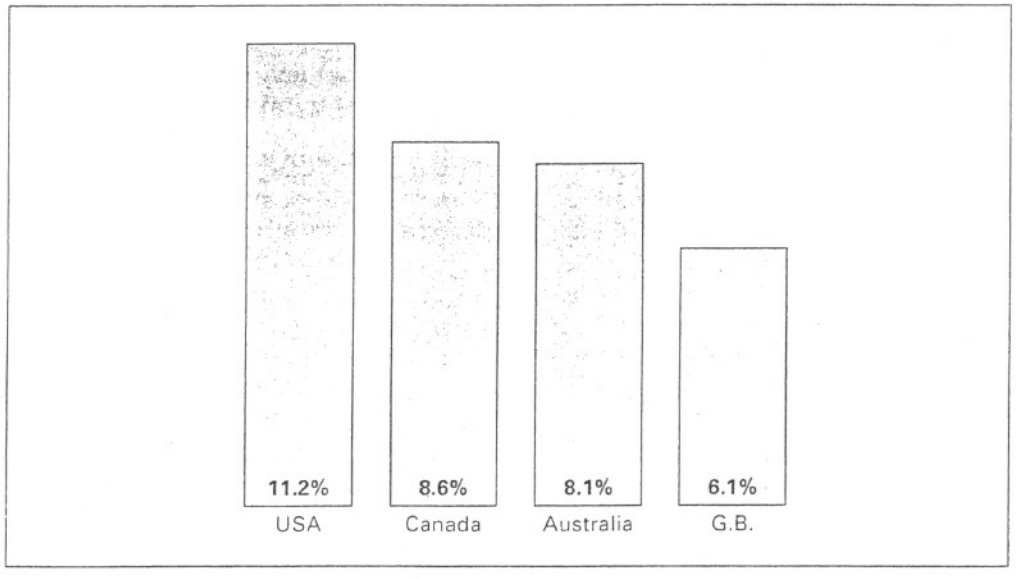


Figura 1.7 Diagramma lineare tassi di malaria registrati per anno Stati Uniti, 1940-1989

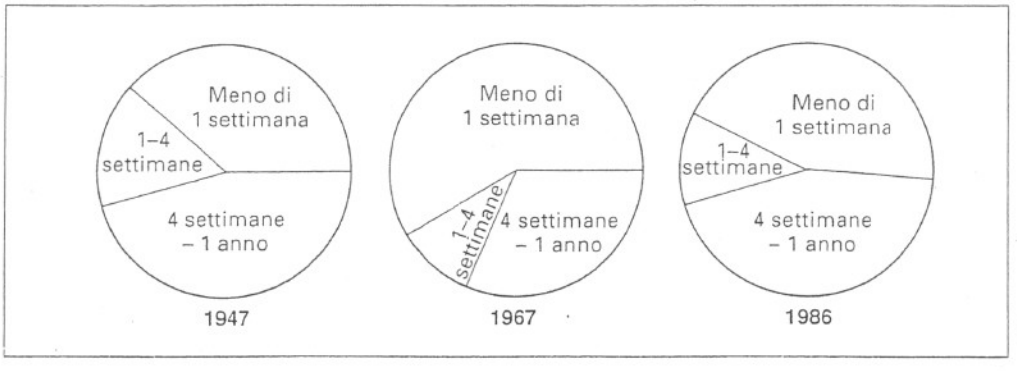
Secondo il *Compendio Statistico* ISTAT 1988 la popolazione residente nel territorio italiano (al 1° Gennaio 1988) risulta così distribuita per gruppi di regioni (il *Nord* comprende Piemonte, Valle d'Aosta, Lombardia, Liguria, Trentino-Alto Adige, Friuli-Venezia Giulia, Emilia-Romagna; il *Centro* comprende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

Nord	25 518 866
Centro	10 952 361
Sud	14 135 320
Isole	6 792 561
Totale	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

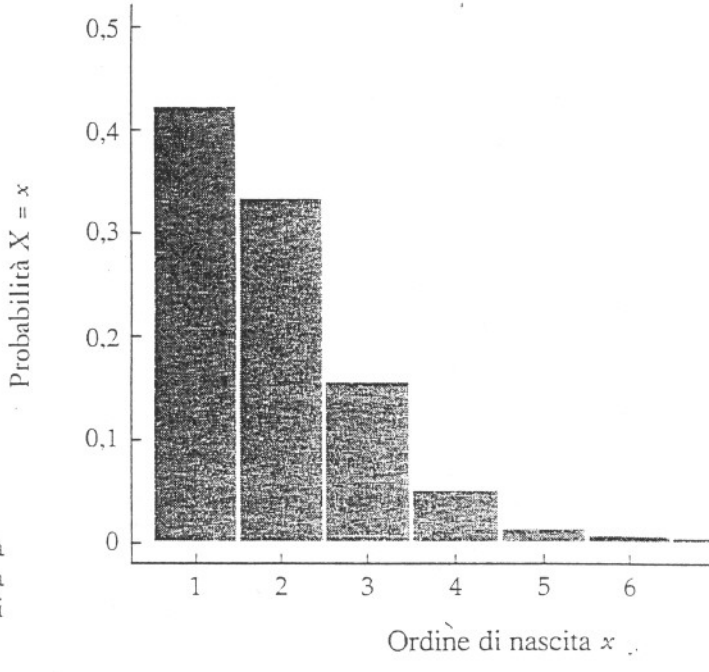


Fig 2.1 Distribuzione di probabilità di una variabile casuale che rappresenta l'ordine di nascita di neonati negli Stati Uniti

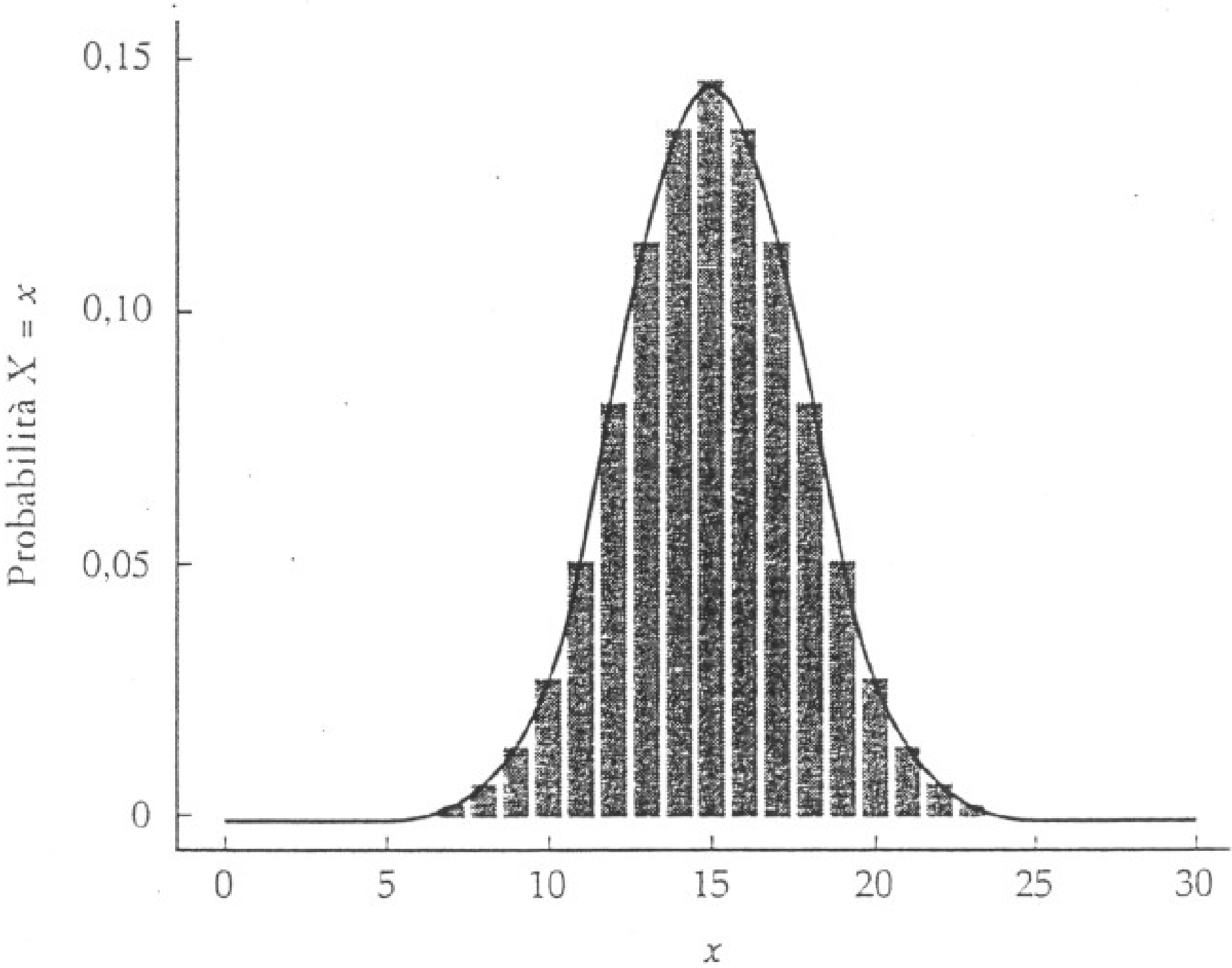


Fig. 2.2

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	$\bar{x}_2$ = 3,03 litri	$\bar{x}_3$ = 2,88 litri
s_1 = 0,496 litri	s_2 = 0,523 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	
					86	−19

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

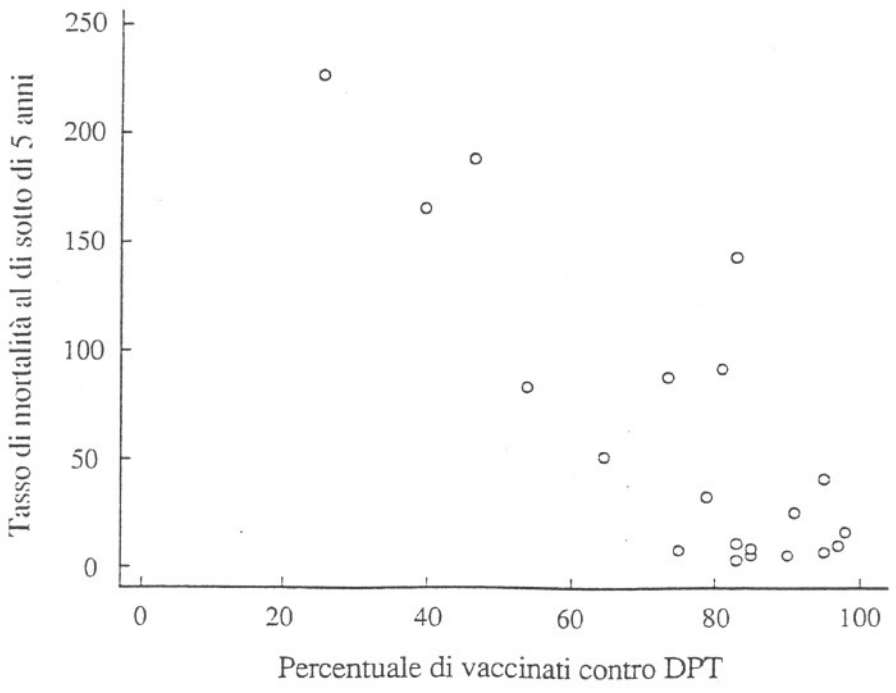
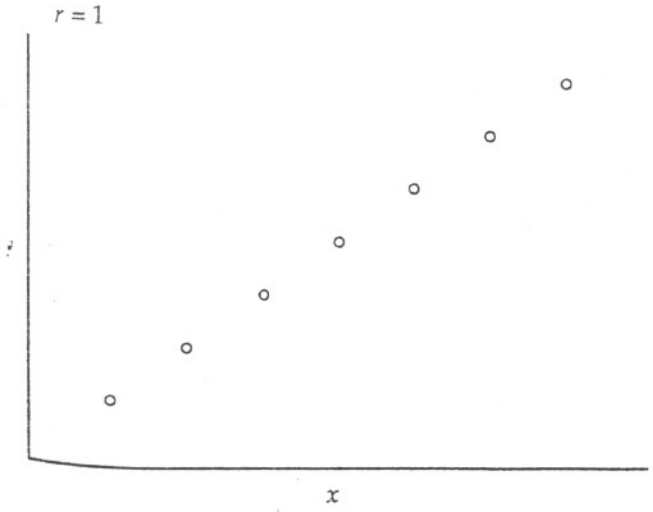
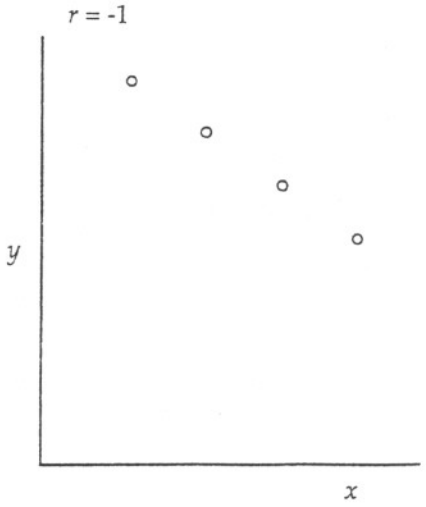


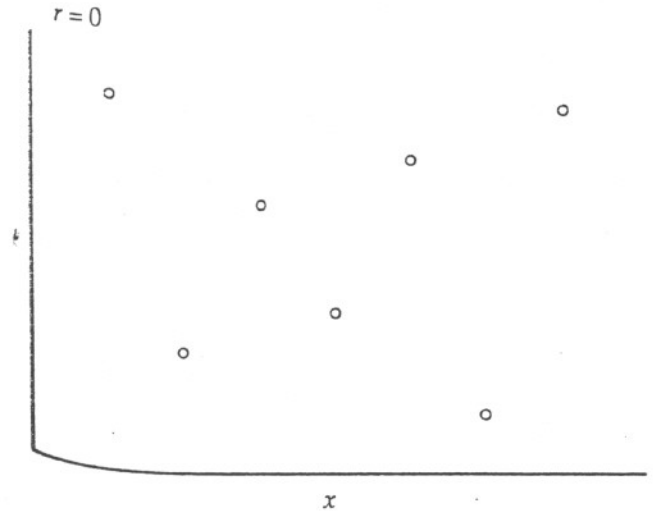
Fig. 8.1



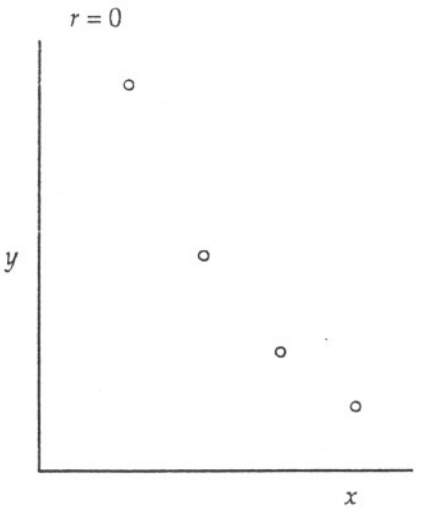
(a)



(b)



(c)



(d)

Diagrammi a punti che illustrano possibili relazioni tra X e Y

Fig. 8.2

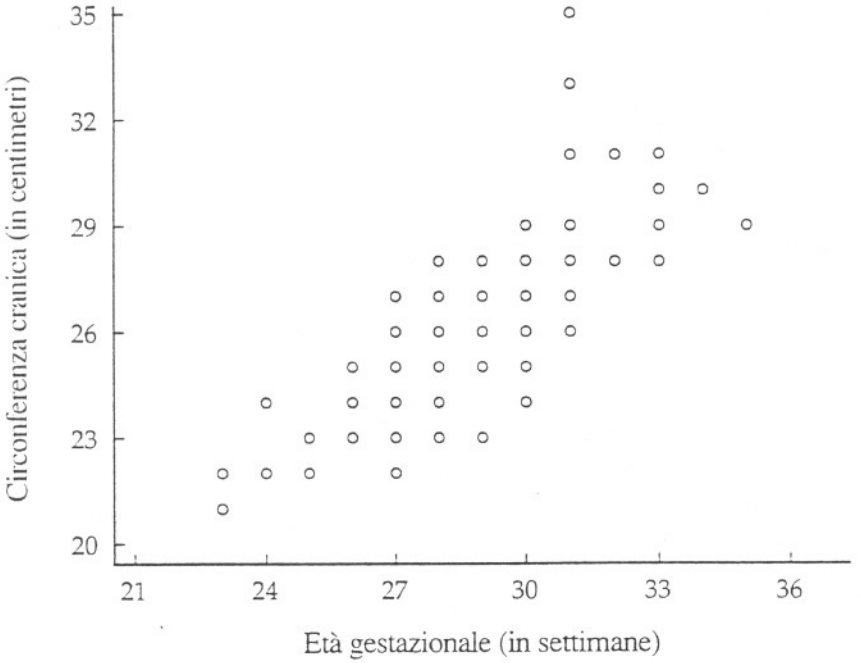
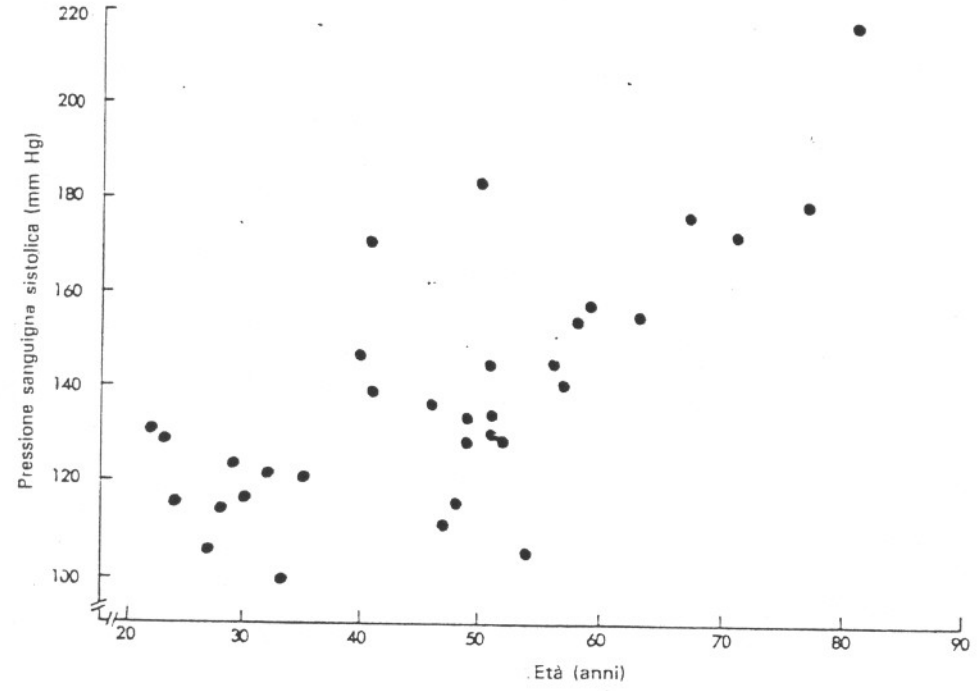

Circonferenza cranica (in centimetri)
35
32
29
26
23
20
21
24
27
30
33
36
Età gestazionale (in settimane)

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



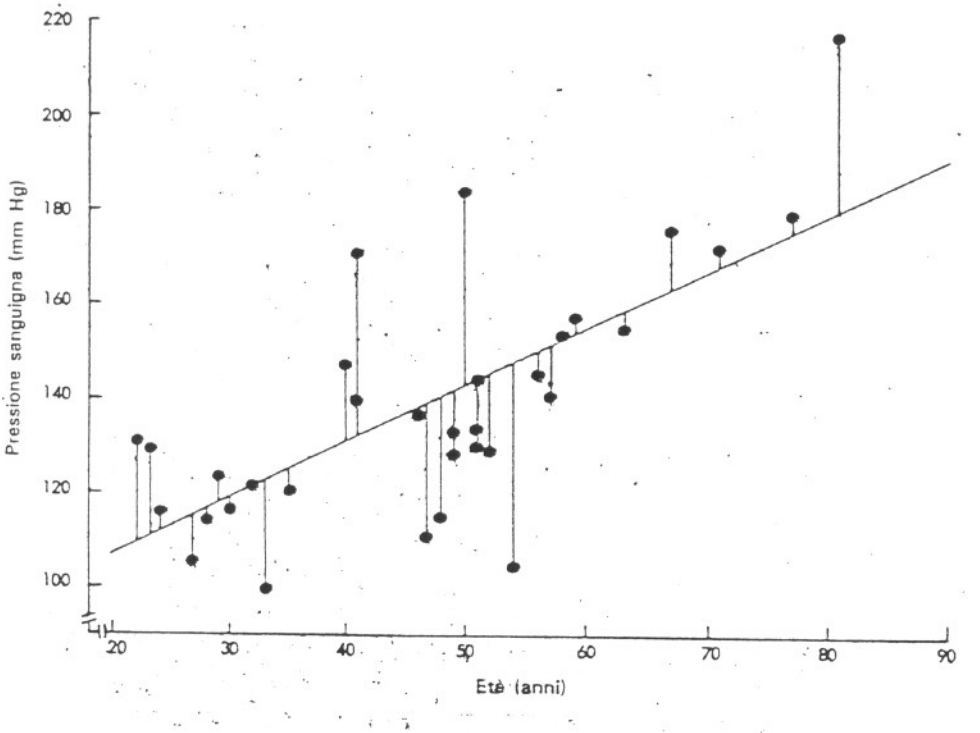


Fig. 9.2

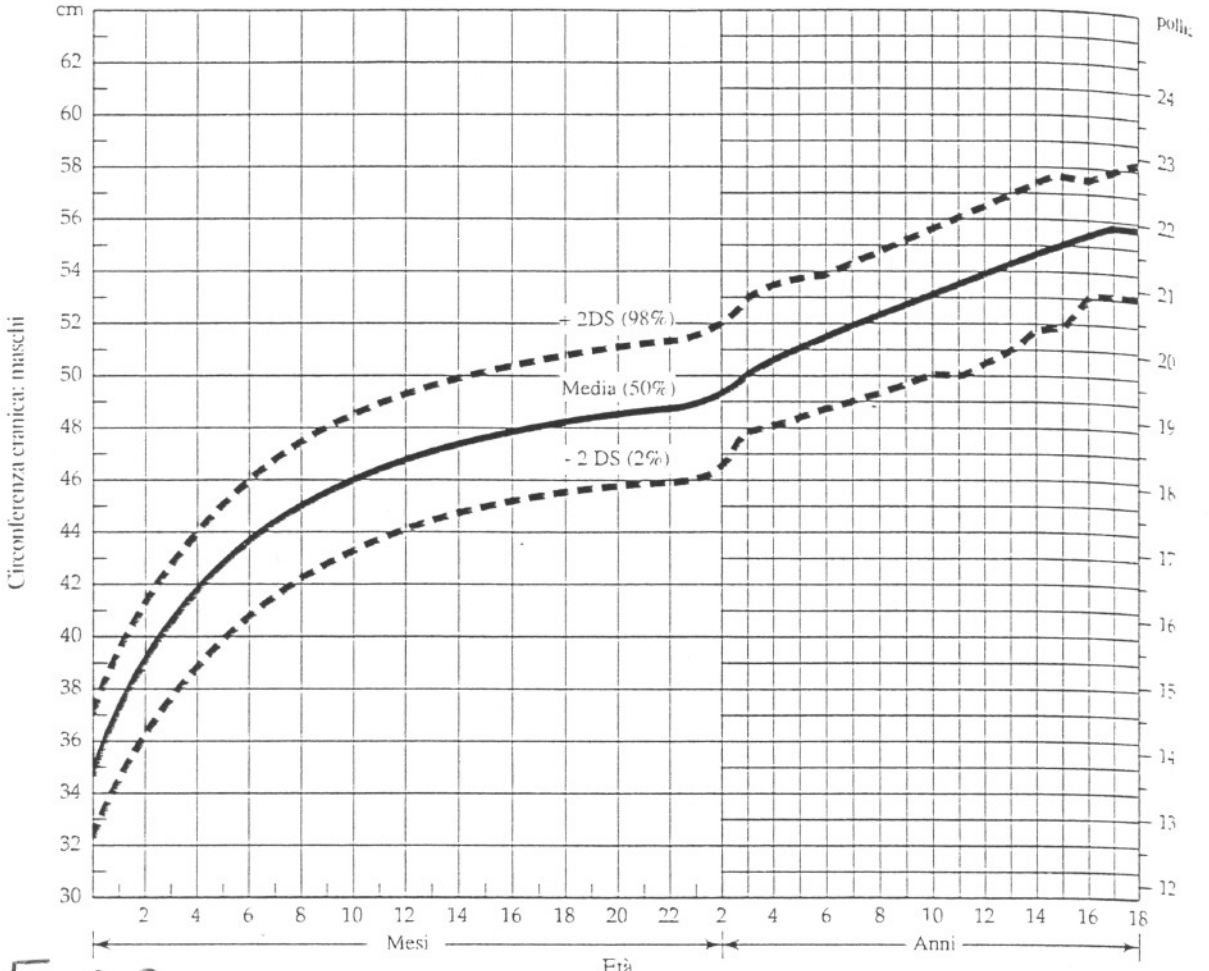


Fig. 9.3 Circonferenza cranica in funzione dell'età nei maschi (Da: G. Nellhaus, *Pediatrics*, 41:106, 1968. Copyright 1968 American Academy of Pediatrics)

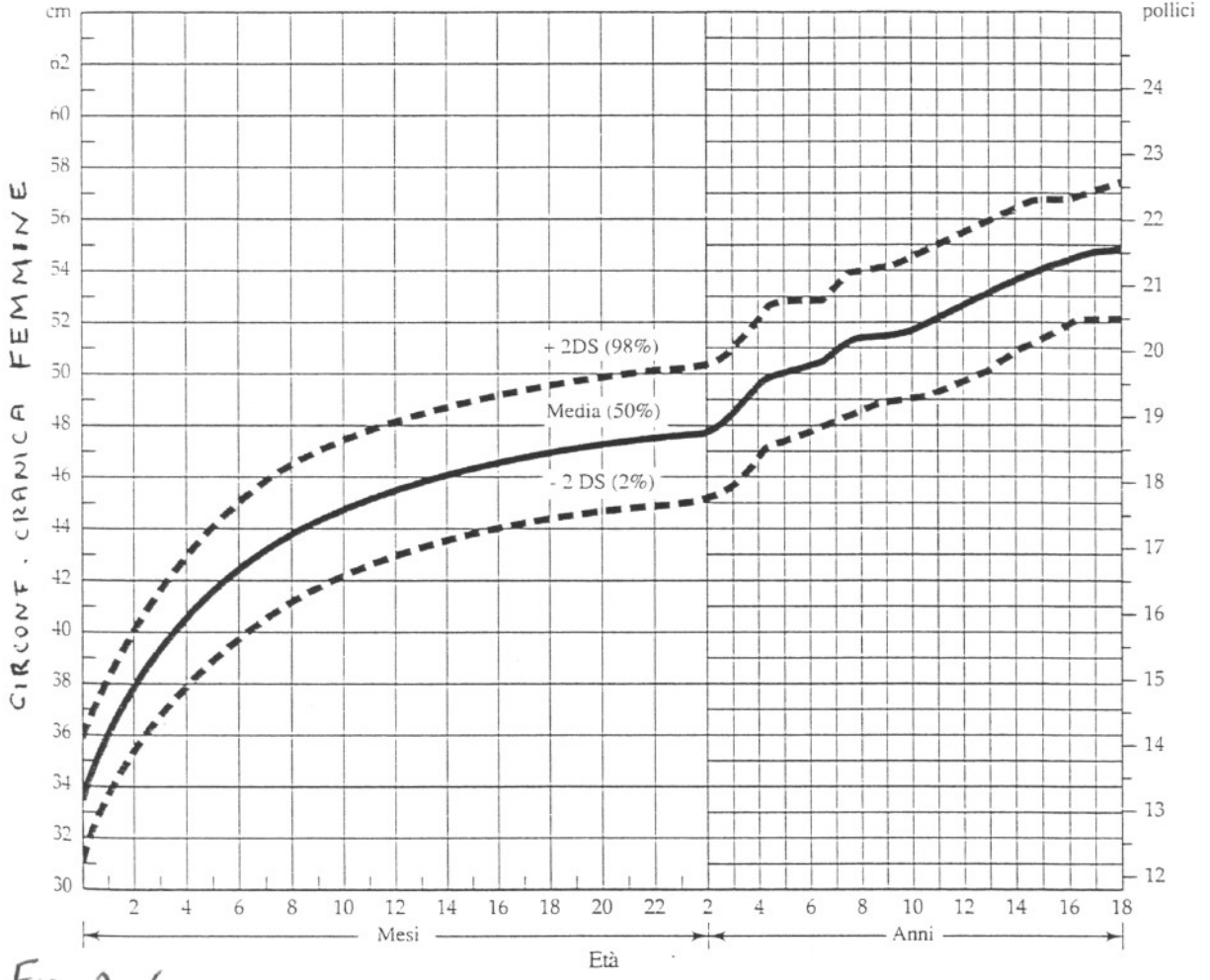


Fig. 9.4 Circonferenza cranica in funzione dell'età nelle femmine (Da: G. Nellhaus, *Pediatrics*, 41:106, 1968.

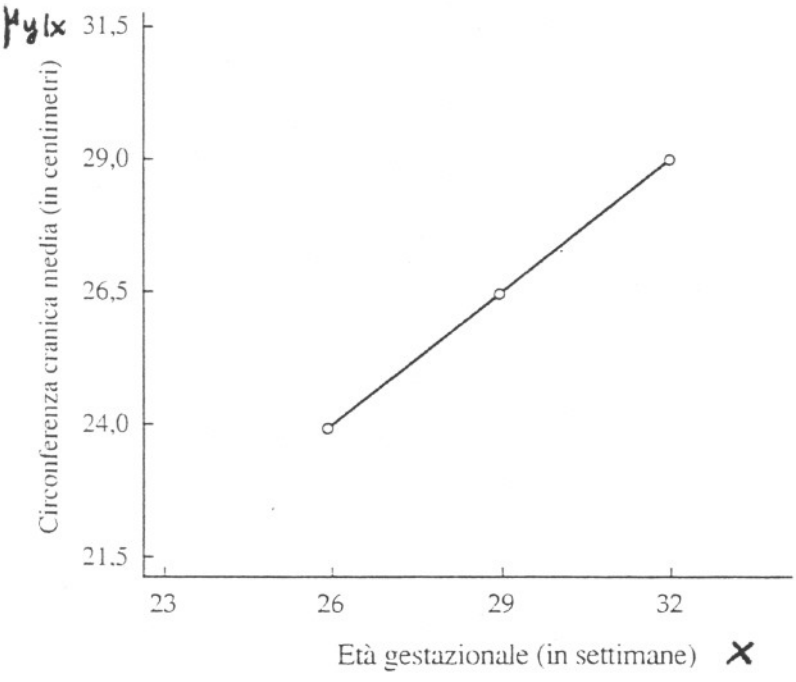

μy|x
31,5
29,0
26,5
24,0
21,5
Circonferenza cranica media (in centimetri)
23
26
29
32
Età gestazionale (in settimane)
X

Fig. 9.5

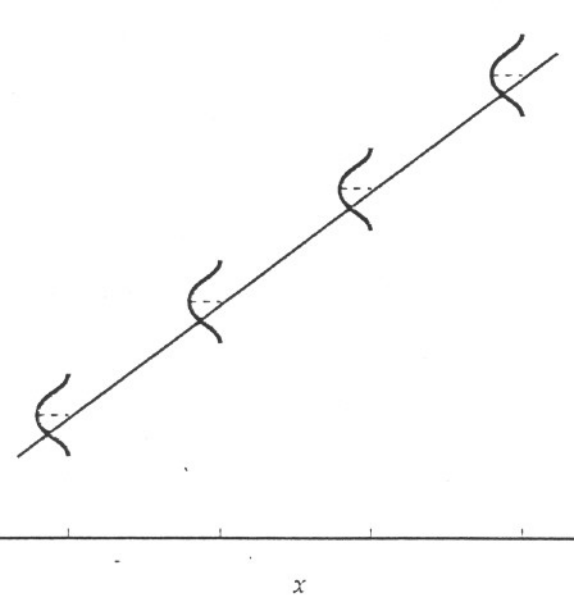


Fig. 9.6

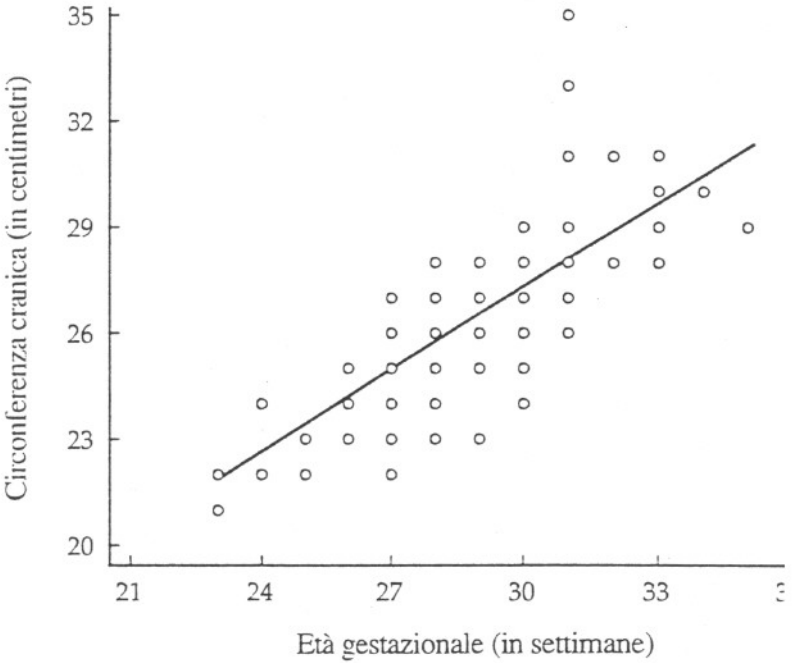
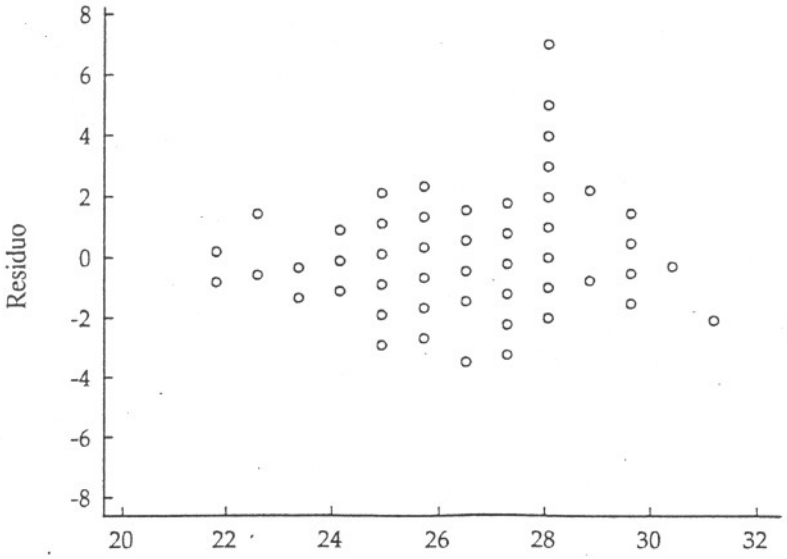
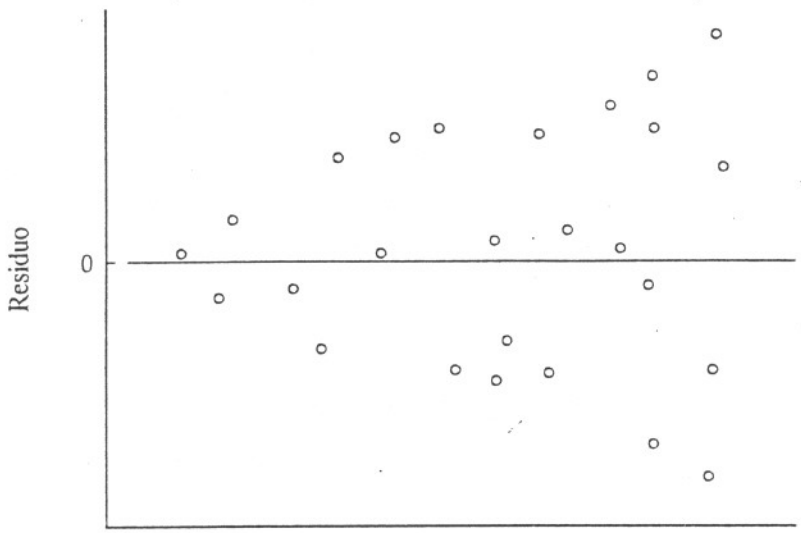

Circonferenza cranica (in centimetri)
35
32
29
26
23
20
21
24
27
30
33
Età gestazionale (in settimane)

Fig. 9.7

RESIDUI IN FUNZIONE DEI VALORI SULLA RETTA AI MINIMI QUADRATI

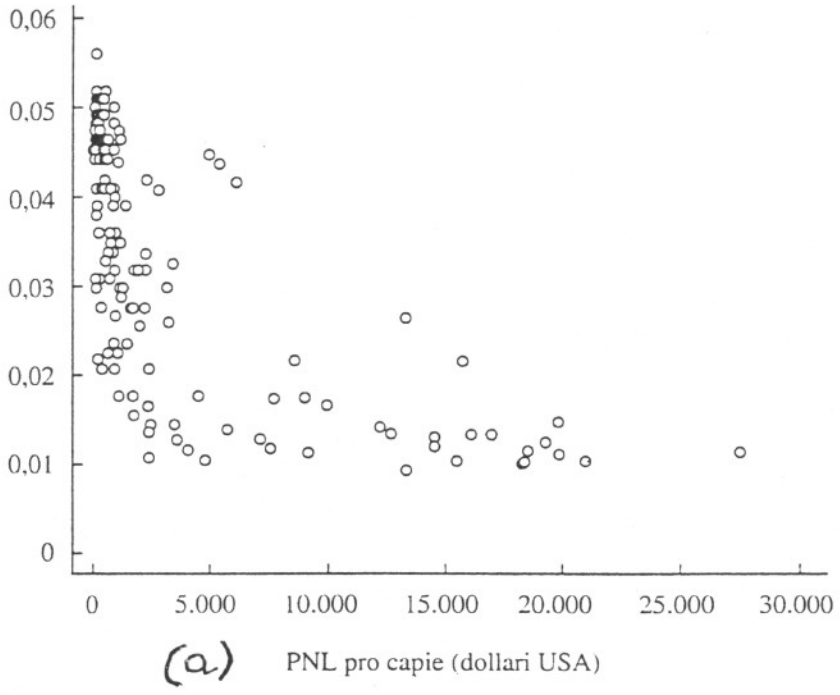


(a) Valore adattato della circonferenza cranica



(b) Valore adattato di y

Fig. 9.8


0,06
0,05
0,04
0,03
0,02
0,01
0
0
5.000
10.000
15.000
20.000
25.000
30.000
(a)
PNL pro capie (dollari USA)

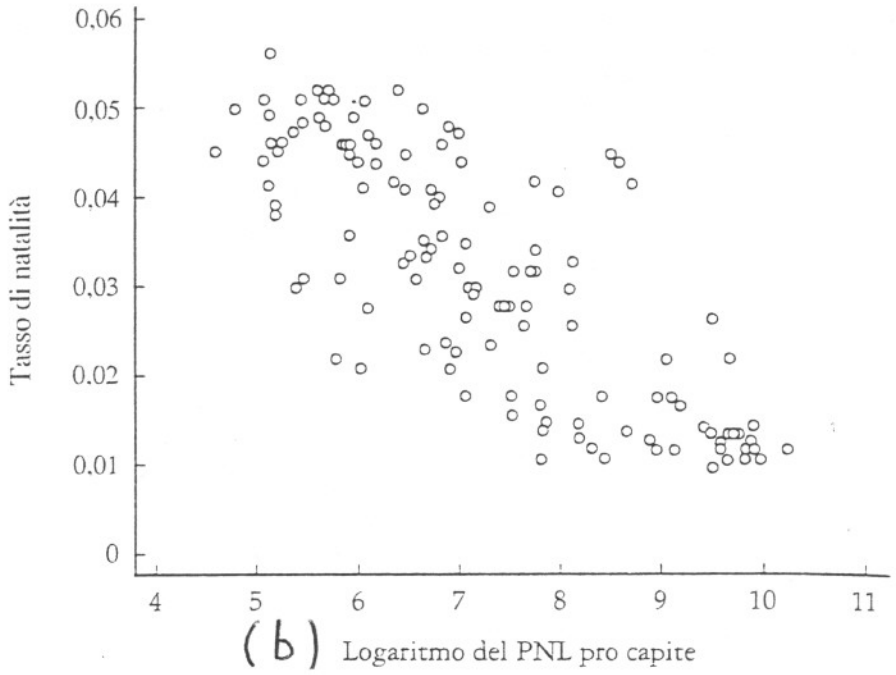

Tasso di natalità
0,06
0,05
0,04
0,03
0,02
0,01
0
4
5
6
7
8
9
10
11
(b)
Logaritmo del PNL pro capite

Fig. 9.9

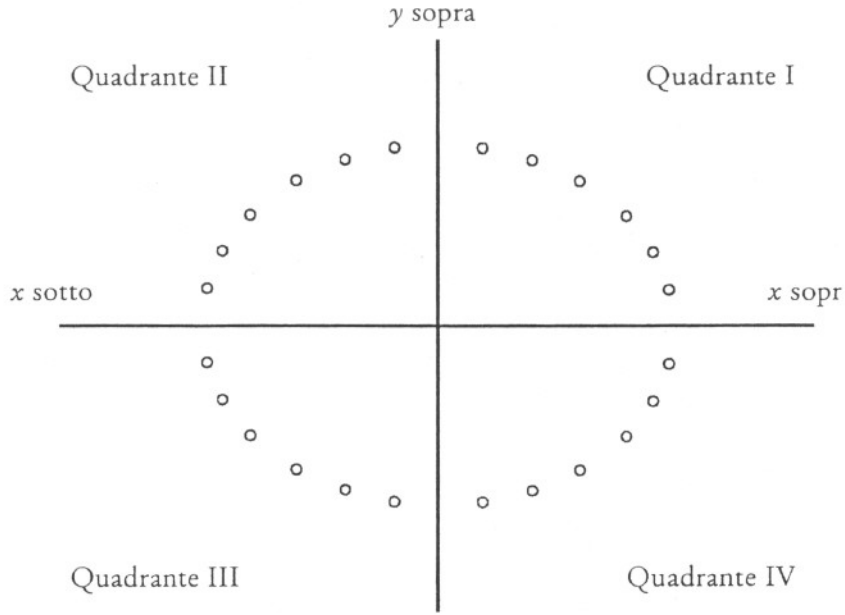
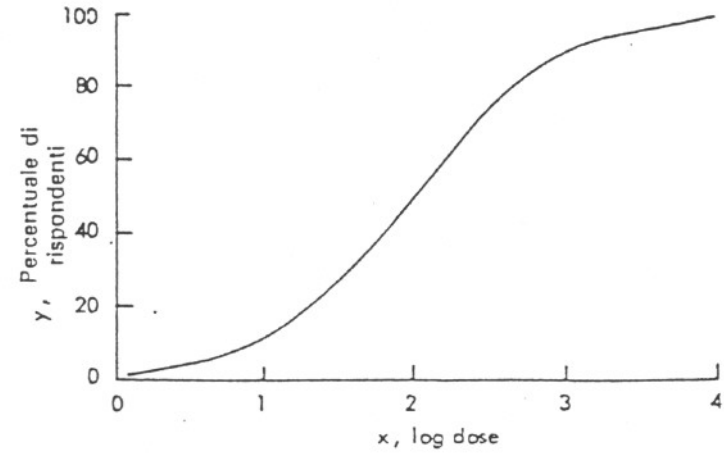


Figura 9.10 Il cerchio delle potenze


100
80
60
40
20
0
y, Percentuale di rispondenti
0
1
2
3
4
x, log dose

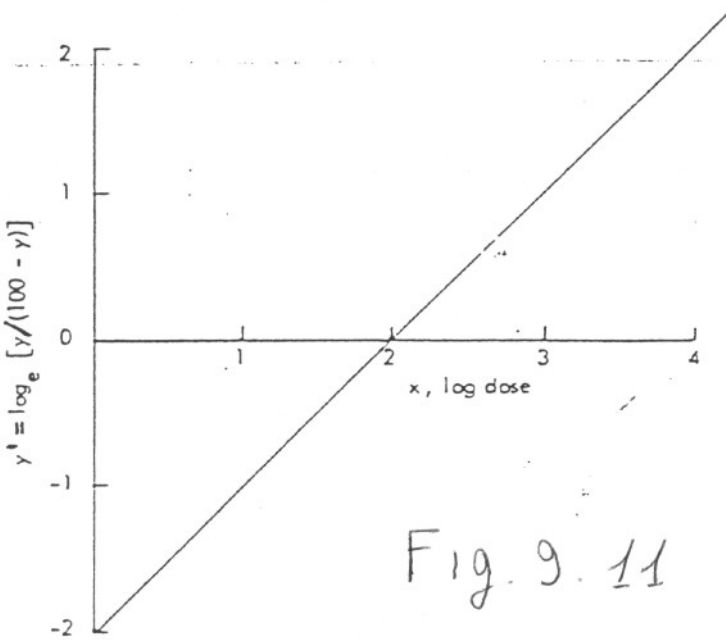

2
1
0
-1
-2
y' = log_e [y/(100 - y)]
1
2
3
4
x, log dose

Fig. 9.11