

## 1 INTRODUZIONE AL CALCOLO DELLE PROBABILITA'

### Plausibilità

Compatibilità tra ipotesi particolare e quadro di conoscenze generali (probabilità a priori)

### Non contraddittorietà

Capacità di una ipotesi di formulare previsioni corrette (verosimiglianza della ipotesi)

### Probabilità

Rapporto tra eventi favorevoli ed eventi possibili, tutti comunque equiprobabili. NON è una definizione perché utilizza il concetto di equiprobabilità, vuoto di significato finché la probabilità non è stata definita.

Limite a cui tende la proporzione di volte in cui si realizza l'evento, al tendere all'infinito delle occasioni (Definizione frequentista)

E' la misura della propensione soggettiva a credere alla veridicità dell'asserzione. (Definizione soggettivista)

### Assiomi<sup>1</sup> di Kolmogorov

$$0 \leq Pr(A) \leq 1 \quad [1]$$

$$Pr(S) = 1 \quad [2] \quad S \text{ spazio campionario}$$

$$Pr(A \cup B) = Pr(A) + Pr(B) \quad [3]$$

$$Pr(A \cap B) = Pr(A) Pr(B|A) \quad [4]$$

La [3], valida se A e B si escludono, riguarda la proprietà additiva della probabilità e afferma che, dati due eventi mutuamente esclusivi, la probabilità dell'evento composto, che consiste nel realizzarsi **dell'uno o dell'altro**, è pari alla somma delle probabilità degli eventi singoli.

La [4] riguarda infine la proprietà moltiplicativa delle probabilità. Essa afferma che, dati due eventi, la probabilità dell'evento composto, che consiste nel realizzarsi **congiunto** di due eventi, è pari al prodotto della probabilità dell'uno per la probabilità dell'altro, condizionata al primo.

---

<sup>1</sup> Assioma: principio evidente in sé, e che perciò non ha bisogno di essere dimostrato

### Odds favorevole ad un evento

E' il rapporto tra la probabilità dell'evento e il suo complemento a uno, ovvero la probabilità che l'evento non si realizzi. Tale quantità è pari al reciproco della 'quota' che un book-maker paga nel caso in cui l'evento sul quale si è scommesso si sia realizzato.

L'odds rappresenta quanto si è disposti a rischiare, per unità di vincita, scommettendo su di un esito particolare. Esso esprime, su di una scala da zero a infinito, il grado di sicurezza con cui prevediamo detto esito.

$$Odds = \frac{Pr}{Pr-1}$$

Così se l'odds vale 1 se a un determinato esito si assegna probabilità pari a 1/2, vale 0.5 se gli si assegna probabilità 1/3 e così via.

L'Odds sarà tanto maggiore di 1 quanto più sicuro è considerato l'esito su cui si scommette e tanto minore di 1 quanto meno probabile esso è considerato.

### Esercizi

**1.1** Calcolare la probabilità, lanciando due volte lo stesso dado, di ottenere per due volte un sei.

I due eventi sono indipendenti; allora si ha:

$$Pr(2\text{volte } 6) = Pr(6 \text{ la prima volta}) Pr(6 \text{ la seconda volta})$$

$$Pr(2 \text{ volte } 6) = \frac{1}{6} \frac{1}{6} = \frac{1}{36}$$

**1.2** Calcolare la probabilità, lanciando due volte lo stesso dado, di ottenere almeno una volta un sei.

Soluzione a

$$Pr(\text{almeno un } 6) = Pr(6 \wedge n6) + Pr(n6 \wedge 6) + Pr(6 \wedge 6)$$

$$Pr(\text{almeno un } 6) = \frac{1}{6} \frac{5}{6} + \frac{5}{6} \frac{1}{6} + \frac{1}{6} \frac{1}{6}$$

$$Pr(\text{almeno un } 6) = \frac{11}{36}$$

Soluzione b

La probabilità di ottenere almeno un sei è pari al complemento a 1 della probabilità di non ottenere nemmeno un sei su due lanci.

$$Pr(n6 \wedge n6) = Pr(n6) Pr(n6)$$

$$Pr(n6 \cap n6) = \frac{5}{6} \frac{5}{6} = \frac{25}{36}$$

$$Pr(\text{almeno un } 6) = [1 - Pr(n6 \cap n6)] = 1 - \frac{25}{36} = \frac{11}{36}$$

**1.3** Con ciascuna somministrazione di un farmaco vi è la probabilità del 10% che si sviluppi una reazione

$$Pr(\text{non reaz. con 5 somm.}) = 0.9^5 = 0.59$$

tossica. Un paziente richiede 5 somministrazioni consecutive del farmaco. Qual è la probabilità di rilevare almeno una reazione tossica?

$$Pr(\text{non reazione}) = 1 - Pr(\text{reazione}) = 1 - 0.1 = 0.9$$

$$Pr(\text{reaz. con 5 somm.}) = 1 - Pr(\text{non reaz. con 5 somm.}) = 0.41$$

**1.4** Qual è la probabilità per un 25enne di compiere 65 anni sapendo che alla nascita la probabilità di compiere 25 e 65 anni è rispettivamente 0.95364 e 0.6544

$$Pr(25 \cap 65) = Pr(25) Pr(65|25)$$

$$Pr(65|25) = \frac{Pr(65 \cap 25)}{Pr(25)} = \frac{0.6544}{0.95364} = 0.6862$$

**1.5** In una famiglia con due figli sappiamo che uno di questi è maschio. Qual è la probabilità che l'altro figlio sia di sesso femminile?

Definiamo l'universo campionario: [m;f] [m;m] [f;m]

$$Pr(f|m) = \frac{2}{3}$$

**1.6\*** Date n persone calcolare la probabilità che almeno due siano nate nello stesso giorno. (considera gli anni costituiti da 365 giorni)

$$Pr(\textit{nessuna}) = \frac{364}{365} \frac{363}{365} \dots \dots \dots$$

$$Pr(\textit{nessuna}) = \frac{365}{365} \frac{364}{365} \frac{363}{365} \dots \dots \dots$$

$$Pr(\textit{nessuna}) = \frac{365!}{(365-n)!} \frac{1}{365^n}$$

$$Pr(\textit{nessuna}) = \binom{365}{n} n! \frac{1}{365^n}$$

$$Pr(\textit{nessuna})_{n=25} = 0.413$$

$$Pr(\textit{almeno una})_{n=25} = 0.587$$

$$Pr(\textit{nessuna})_{n=40} = 0.1087$$

$$Pr(\textit{almeno una})_{n=40} = 0.8913$$

[\(Vedi grafico\)](#)

**1.7** Una soggetto decide di sottoporsi ad una serie di accertamenti consistenti in venti test diagnostici diversi, ognuno dei quali, su trenta ripetizioni fornisce un valore falsamente patologico.

Si determini la probabilità che l'individuo, peraltro perfettamente sano, presenti almeno un risultato, tra i venti test a cui si è sottoposto, di valore patologico. [R.0.492]

## 2 APPLICAZIONI DEL CALCOLO DELLE PROBABILITA' ALLA DIAGNOSI CLINICA

### **Sensibilità di un test diagnostico**

La capacità del test di risultare positivo quando vi si sottopone un ammalato. Dal punto di vista formale possiamo dire che la sensibilità di un test è la probabilità di ottenere un valore positivo condizionata allo stato di malato del soggetto esaminato.

$$SENS = Pr(+|M)$$

### **Specificità di un test diagnostico**

La capacità di un test di risultare negativo quando vi si sottopone un sano. Formalmente la specificità è la probabilità di un esito negativo del test, condizionata allo stato di non malato del soggetto esaminato.

$$SPEC = Pr(-|S)$$

### **Valore predittivo di un test positivo**

E' il grado di certezza con cui, di fronte a un esito positivo, si può affermare la presenza della patologia.

$$VPP = Pr(M|+) = Pr(M) \frac{Pr(+|M)}{Pr(+)} = \frac{Pr(M)Pr(+|M)}{Pr(M)Pr(+|M) + Pr(S)Pr(+|S)}$$

$$VPP = \frac{Pr(M) SENS}{Pr(M) SENS + Pr(S) (1-SPEC)}$$

Al di là dei formalismi matematici è importante sottolineare che il **VPP** è funzione non solo dei parametri di validità del test (sensibilità e specificità), ma anche dalla probabilità a priori della patologia che si intende identificare.

### **Valore predittivo di un test negativo**

E' il grado di certezza con cui, di fronte a un esito negativo, si può escludere la presenza della patologia.

$$VPN = Pr(S|-) = Pr(S) \frac{Pr(-|S)}{Pr(-)} = \frac{Pr(S)Pr(-|S)}{Pr(S)Pr(-|S) + Pr(M)Pr(-|M)}$$

$$VPN = \frac{Pr(S) SPEC}{Pr(S) SPEC + Pr(M) (1-SENS)}$$

Anche in questo caso è facile rendersi conto che il **VPN** dipende anche dalla probabilità a priori della patologia che si intende identificare.

**Variazione di VPP in funzione di Pr(M)**

Sensibilità 0.94

Specificità 0.92

[\(vedi grafico\)](#)

**Variazione di VPN in funzione di Pr(M)**

Sensibilità 0.94

Specificità 0.92

[\(vedi grafico\)](#)

### **Esempio Clinico** ( da Marubini e coll. Introduzione alla Statistica Medica-NIS)

Consideriamo l'impostazione di uno screening atto ad individuare precocemente i neonati affetti da sindrome di Crigler-Najjar<sup>2</sup>.

La distribuzione dei valori di bilirubina in neonati sani e affetti dalla patologia in esame è rappresentata nella [figura](#).

E' evidente come i livelli di bilirubina dei sani e dei malati siano in parte sovrapposti. Ciò significa, in altri termini, che vi sono dei sani che presentano dei livelli di bilirubina maggiori o uguali a quelli dei malati. E' altresì evidente che la probabilità di un malato di avere un livello di bilirubina inferiore a 3 mg/dl è estremamente bassa, tanto che un test basato su di un tale valore di soglia avrebbe una sensibilità praticamente del 100%. Tuttavia, a causa della sovrapposizione delle due distribuzioni la specificità sarebbe molto bassa.

Al contrario se si scegliesse come valore soglia 7 mg/dl , si avrebbe un test di specificità pari al 100% ma con una ridotta sensibilità.

Come si vede allora la specificità e sensibilità di un test sono quantità variabili, tra l'altro, con il valore soglia; e per la loro non indipendenza uno spostamento di soglia che comporti un aumento di sensibilità non può che comportare una riduzione di specificità e viceversa.

Alcune regole pratiche:

#### **1 Probabilità a priori della malattia molto bassa**

Per quanto elevata sia la specificità un test positivo sarà probabilmente un falso positivo. In tale contesto, inoltre, la specificità non può essere molto alta altrimenti la bassa sensibilità indotta renderebbe possibile *'il passaggio dei pochi casi attraverso le maglie del setaccio'*.

L'esito positivo del test di screening non rappresenta un diretto indirizzo diagnostico, ma solamente l'indicazione della necessità di un approfondimento diagnostico. Se la sensibilità è molto elevata, l'esito negativo del test potrebbe essere sufficiente a escludere la presenza della patologia indagata.

(riesamina i grafici dell'andamento di VPP e VPN)

#### **2 Probabilità a priori della malattia molto elevata**

Un test atto a confermare la presenza della malattia dovrà avere una specificità praticamente assoluta, anche se di bassa sensibilità. Un risultato negativo darà luogo ad ulteriori accertamenti, mentre un risultato positivo confermerà la diagnosi.

---

<sup>2</sup> I neonati affetti da questa sindrome presentano livelli di bilirubinemia sierica molto elevati anche a distanza dalla nascita.

### Esercizi

**2.1** Calcolare con i dati di tabella: SENS, SPEC, Pr(M), Pr(+), VPP, VPN

		Sano	Malato	Totale
TEST	Positivo	50	50	100
	Negativo	600	2	602
		650	52	702

**2.2** Calcolare con i dati di tabella: SENS, SPEC, Pr(M), Pr(S), Pr(+|S), Pr(-|M), VPP, VPN

		Sano	Malato	Totale
TEST	Positivo	8	94	102
	Negativo	92	6	98
		100	100	200

**2.3** Un test con sensibilità 0.8 e specificità 0.7 viene introdotto per valutare la presenza di una malattia che si presume sia attribuibile allo 0.1% della popolazione .

Determina: Pr(+|S), Pr(-|M), Pr(M|+), Pr(S|-)

**2.4** Risolvere il problema precedente facendo ora riferimento ad un test con sensibilità 0.7 e specificità 0.8.

**2.5** Risolvere il problema 2.3 facendo riferimento ad una patologia attribuibile al 10% della popolazione.

**2.6** Risolvere il problema 2.5 facendo riferimento ad un test con sensibilità 0.7 e specificità 0.8.

### CONCETTO DI VERO POSITIVO (TP) E VERO NEGATIVO (TN)

I soggetti TP sono i soggetti malati e con test positivo. La probabilità di imbattersi in un TP è perciò pari alla probabilità del verificarsi degli eventi combinati **Test+** e **essere malato**.

Ricordando l'assioma [4] di Kolmogorov e sostituendo ad A e B gli eventi sopra considerati possiamo scrivere:

$$Pr(TP) = Pr(+ \cap M) = Pr(M) Pr(+|M) = Pr(M) SENS$$

La probabilità di imbattersi in un TP è pari alla probabilità di imbattersi in un malato moltiplicata per la Sensibilità. La probabilità di un evento TP non dipende solo dalle caratteristiche del test usato ma è notevolmente condizionata anche dalla rarità della malattia.

Se desideriamo perciò ridurre al minimo la probabilità di imbatterci in falsi positivi (FP)<sup>3</sup> ci troviamo nella necessità di usare test a elevata sensibilità, ma soprattutto dobbiamo orientare lo screening su una sottopopolazione ad elevato rischio di malattia.

I soggetti TN sono i soggetti sani e con test negativo. Come nel caso dei TP possiamo scrivere:

$$Pr(TN) = Pr(- \cap S) = Pr(S) Pr(-|S) = Pr(S) SPEC$$

Se desideriamo ridurre al minimo la probabilità di ottenere dei falsi negativi (FN)<sup>4</sup> dobbiamo usare test ad elevata specificità.

Possiamo ora, tenuto presente le definizioni di TP e TN, dare una veste formale diversa sia alla sensibilità sia alla specificità.

$$SENS = \frac{Pr(TP)}{Pr(M)} = \frac{n.TP}{n.malati}$$

$$SPEC = \frac{Pr(TN)}{Pr(S)} = \frac{n.TN}{n.sani}$$

#### Sensibilità

Percentuale di soggetti con la malattia che risultano malati in base al test

#### Specificità

Percentuale di soggetti senza malattia che risultano sani in base al test

---

<sup>3</sup>Un soggetto FP è un soggetto che, pur essendo sano, ha un test positivo.

<sup>4</sup>Un soggetto FN è un soggetto che, pur essendo malato, presenta un test negativo.

**ESEMPIO CLINICO** (da LILIENFELD Fondamenti di Epidemiologia Piccin)

La [tabella](#) riportata mostra i livelli di glicemia a due ore dal pasto in un gruppo di 70 veri diabetici e di 510 veri non diabetici. La percentuale di dei diabetici identificati dal test (**Sensibilità**) e la percentuale di dei non diabetici identificati dal test (**specificità**) sono indicate per vari livelli di glicemia.

Se si decidesse di usare il livello di 110 mg/dl quale valore soglia per il diabete, si identificherebbero il 92.9 % dei veri diabetici e il 48.4% dei veri non diabetici. Se si stabilisse una soglia più bassa, ad esempio 80 mg/dl, il test assumerebbe una Sensibilità pari al 100%. Peraltro ciñ avverrebbe a scapito di una errata identificazione di numerosi soggetti normali come diabetici in conseguenza della bassa specificità.

Se si stabilisse una soglia più elevata, ad esempio, 200 mg/dl, il test sarebbe in grado di identificare il 100% dei veri non diabetici (specificità = 100%) , d'altro canto però, molti diabetici veri (62.9%) verrebbero erroneamente classificati come non diabetici in conseguenza della bassa sensibilità.

Nello stabilire il livello di un test atto a identificare i soggetti affetti da una malattia, si devono valutare attentamente i costi<sup>5</sup> relativi alla determinazione sia di falsi negativi che di falsi positivi.

Si pensi ad esempio al costo conseguente alla catalogazione HIV-negativo di un donatore in realtà HIV-positivo.

Effetto dei differenti valori di soglia glicemica sulla determinazione di falsi positivi e falsi negativi ([vedi grafici](#))

---

<sup>5</sup>Il termine costo viene qui inteso in un'ottica sociale piuttosto che meramente economica

### 3 MISURE DI OCCORRENZA DI MALATTIA

#### Tasso di incidenza

Misura la velocità con cui un processo morboso tende a svilupparsi in un determinato contesto.

Dal punto di vista formale il tasso di incidenza è il rapporto tra i **nuovi** casi di malattia insorti e la somma dei periodi di osservazione di tutti gli individui della popolazione in esame.

$$INCIDENZA = \frac{N. \text{ nuovi casi}}{\sum t_i}$$

Il denominatore rappresenta il cosiddetto prodotto 'persone-tempo' e individua l'entità dell'osservazione temporale che ha definito un determinato tasso di incidenza.

E' importante sottolineare che uno stesso prodotto 'persone-tempo' può derivare da metodi osservazionali diversi. Ad esempio 100 persone-anno può derivare dall'osservazione per un anno di 100 soggetti, ma può pure derivare dall'osservazione di 50 individui per due anni!

#### Incidenza Cumulata

E' la proporzione di individui, tratti da una popolazione chiusa, che si ammala in un determinato periodo. Rappresenta il rischio medio di ammalarsi *in un determinato periodo*.

$$IC = \frac{P_0}{P_1}$$

$P_0$       numero di nuovi ammalati

$P_1$       numero di sani arruolati nello studio

#### Prevalenza

Misura la distribuzione del fenomeno nella popolazione, fornendo una '*fotografia*' della situazione.

Formalmente è rappresentato dal rapporto tra malati e tutti i soggetti presenti .

$$PREVALENZA = \frac{N. \text{ Malati}}{N. \text{ Totale Soggetti}}$$

#### Esercizio

**3.1** Supponiamo che nel gennaio 1976, 1000 soggetti adulti residenti in una cittadina abbiano accettato di sottoporsi ad una visita per ipotiroidismo. La malattia fu accertata in 8 soggetti; di questi, 5 erano già in terapia, negli altri 3 la malattia fu diagnosticata per la prima volta.

Le stesse persone furono esaminate nuovamente nel gennaio 1978. Vennero riconosciuti 6 nuovi casi di ipotiroidismo; 2 di questi avevano manifestato i sintomi della malattia alcuni mesi prima ed erano stati

sottoposti a terapia dal loro medico di base. Si apprese in seguito che delle 8 persone cui era stata posta diagnosi di ipotiroidismo nel 1976, una aveva interrotto la terapia ed era deceduta per mixedema.

Tranne questo caso tutti gli individui esaminati nel 1976 si sottoposero alla seconda visita.

a) Quale fu la prevalenza di ipotiroidismo, in terapia o meno, nel gruppo esaminato nel gennaio 1976?

E nel gennaio 1978?

b) Quale fu il tasso di incidenza relativo a questo gruppo?

c) Se dei 1000 soggetti esaminati che componevano la collettività all'inizio dell'indagine, solo 900 si fossero sottoposti alla visita nel 1978, come avreste modificato le risposte precedenti?

## 4 LA MISURA

### Misura

È il rapporto tra una grandezza e un'altra ad essa omogenea assunta come unità.

### Accuratezza

L'accuratezza di una procedura di misura è quella proprietà per cui la procedura tende a fornire misure coincidenti con la grandezza effettiva.

Essa si determina replicando, in condizioni costanti, la misurazione di una grandezza fissa e nota. Si indica come misura di *'inaccuratezza'* (**bias**) la differenza fra la quantità nota misurabile e la media delle  $n$  misure effettuate ( $n$  deve essere molto elevato).

### Precisione

La precisione di una procedura di misura è la proprietà di fornire risposte fra loro molto prossime, ovvero poco o per nulla disperse intorno al loro valore medio.

Si indica come misura di *'imprecisione'* la deviazione standard delle misure effettuate (vedi in seguito)

Nella [figura](#) è schematizzata la situazione di quattro procedure di misura con varia precisione e accuratezza.

### Esercizio

**4.1** Cinque provette ciascuna con un contenuto di colesterolo pari 150 mg/dl vengono sottoposte a due test [A,B] per la determinazione del livello di colesterolo.

I risultati ottenuti con le dieci misurazioni sono:

**Test A**            150    145    155    165    135

**Test B**            154    152    155    151    153

Determinare qualitativamente, riportando i risultati su di un grafico, quale dei due test è più preciso e/o più accurato.

## 5 LA RAPPRESENTAZIONE GRAFICA DEI DATI

L'utilizzazione di un [diagramma a torta](#) (pie chart) è uno dei modi più intuitivi per rappresentare variabili nominali.

Anche un [diagramma a barre](#) (bar chart) si presta molto bene a rappresentare delle grandezze nominali. Le variabili continue, qualora siano suddivise in classi discrete, possono rappresentarsi tramite [istogrammi](#)

Se le varie classi non hanno la medesima ampiezza, occorre ricordare che non è l'altezza dei rettangoli che deve essere proporzionale alle frequenze, bensì l'area; risulta ovvio allora che quando una classe ha ampiezza doppia rispetto ad un'altra, la frequenza corrispondente deve essere dimezzata prima di essere riportata sul [grafico](#).

La distribuzione di frequenza può essere ben rappresentato da un [diagramma a punti](#) (dot diagram). Il [diagramma Stem-Leaf](#) (ramo-foglia) non è altro che un istogramma la cui realizzazione è basata sulla utilizzazione diretta della grafia dei dati numerici.

Il numero dei rami corrisponde al numero delle classi, mentre il numero delle foglie rappresenta la frequenza della classe di competenza del ramo.

### Esercizi

**5.1** Rappresenta, su di un diagramma a torta, la distribuzione dei gruppi sanguigni ipotizzando che abbiano le seguenti frequenze:

B	14%	O	41%
A	39%	AB	6%

**5.2** Rappresenta il diagramma a punti relativo alla seguente distribuzione di frequenza:

f	3	6	5	2	2
x	0	1	2	3	4

**5.3** Rappresenta l'istogramma di frequenza relativa alla seguente distribuzione:

classe	limiti	frequenza
1	10-14	4
2	15-19	12
3	20-24	10
4	25-29	5
5	30-34	1

**5.4** Risolvere l'esercizio precedente considerando le classi 4 e 5 raggruppate.

**5.5** Costruire il diagramma Stem-Leaf relativo ai seguenti dati:

25	29	30	35	37
40	45	62	18	19
30	31	22	15	18

**ESERCITAZIONE FINALE** (primo periodo)

E' stato approntato un nuovo test per la diagnosi di artrite reumatoide. Il test è stato valutato in quattro comunità ospedaliere con i seguenti risultati ([vedi tabella](#)):

- a) calcola la prevalenza di artrite reumatoide in ciascuna delle comunità;
- b) calcola per ciascuna comunità la sensibilità e specificità del test nonché il VPP;
- c) riporta, su di un unico diagramma a barre, la sensibilità, specificità, VPP e prevalenza di malattia per ciascuna comunità.
- d) come si modificherà presumibilmente il VPP quando il test verrà esteso alla comunità non ospedaliera?

## 6 MISURE DI POSIZIONE O DI TENDENZA CENTRALE

### Media (aritmetica)

E' la somma delle osservazioni divise per il loro numero

$$M = \frac{Y_1 + Y_2 + \dots + Y_N}{N} = \frac{\sum_{i=1}^N Y_i}{N}$$

### Media aritmetica 'pesata'

E' una media aritmetica eseguita su dati raggruppati in classi.

$$M = \frac{\sum_{i=1}^I Y_i f_i}{\sum_{i=1}^I f_i}$$

dove  $i = 1, 2, \dots, I$  indica il numero d'ordine delle classi;  $y_i$  è il valore centrale della  $i$ -esima classe;  $f_i$  è la frequenza della  $i$ -esima classe.

La media così calcolata prende il nome di media pesata in quanto, per il calcolo, ciascun valore della variabile (il valore centrale di ciascuna classe) è moltiplicato (pesato) per la corrispondente frequenza osservata.

Due sono le proprietà fondamentali della media aritmetica:

- 1) la somma algebrica degli scarti dalla media aritmetica è sempre uguale a zero;
- 2) la somma dei quadrati degli scarti dalla media aritmetica è minima.

### Mediana

Se si dispongono i dati in ordine crescente o decrescente l'osservazione che occupa la posizione centrale dei dati osservati è la mediana.

Se il numero dei dati osservati  $N$  è pari non esiste un'osservazione centrale e la mediana si definisce per, convenzione, come la media aritmetica delle due osservazioni centrali.

La mediana, come abbiamo visto, divide la distribuzione ordinata dei valori in due frazioni uguali. Con lo stesso principio possiamo dividere l'insieme dei dati in quattro parti uguali. I valori che costituiscono i punti di suddivisione della distribuzione sono chiamati **primo quartile, secondo quartile (mediana) e terzo quartile**.

### Media geometrica

$$Mg = \sqrt[N]{Y_1 Y_2 \dots Y_N} = \sqrt[N]{\prod_{i=1}^N Y_i}$$

E' importante sottolineare che il logaritmo della media geometrica corrisponde alla media aritmetica del logaritmo dei dati; quindi la media geometrica coincide con l'antilogaritmo (esponenziale) della media aritmetica dei logaritmi.

$$\log(Mg) = \frac{\sum_{i=1}^N \log(Y_i)}{N}$$

La media geometrica è usata abbastanza frequentemente per le grandezze biologiche misurate in scala di diluizione 'al raddoppio', quali i titoli anticorpali nel siero, i titoli di particelle infettanti, i dosaggi farmacologici e le prove di tossicità.

### Moda

La moda corrisponde, quando esiste, alla osservazione più frequente

### ESERCIZI

**6.1** Calcolare la media e la mediana delle seguenti osservazioni: 34, 36, 30, 35, 100

**6.2** Con riferimento all'esercizio precedente, quale tra media e la mediana rappresenta meglio la tendenza dei dati?. Perché?

**6.3** Calcola la media dei [dati tabellati](#) raggruppati in classi

**6.4** Calcola la media geometrica dei seguenti quattro titoli di sieri anti-H: 1/200 1/400 1/1600 1/800

(Ricorda:  $\log(Mg) = \text{Media Logaritmi}(y) \rightarrow$  da cui.....)

## 7 MISURE DI DISPERSIONE

### Range (campo di variazione)

$$RANGE = (Y_{MAX} - Y_{MIN})$$

### Box and whiskers plot ( schema a scatola e baffi)

Rappresenta un modo sintetico per riassumere i dati in termini di range, mediana, quartili etc..

Per costruire il diagramma è sufficiente riportare la mediana (+) e i quartili (I) (che formano i due lati della scatola), il 5 e 95 centile (baffi).



### Varianza

Data una popolazione costituita da N osservazioni la varianza ( $\sigma^2$ ) è definita dalla seguente relazione:

$$F^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}$$

Questo è facilmente estensibile al caso di dati raggruppati in classi<sup>6</sup>:

$$F^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2 f_i}{\sum_{i=1}^N f_i}$$

---

<sup>6</sup> Una formula alternativa per il calcolo di  $\sigma^2$  è la seguente:

$$F^2 = \left( \frac{\sum_{i=1}^N y_i^2 f_i}{\sum_{i=1}^N f_i} - \bar{y}^2 \right)$$

## Deviazione Standard

La deviazione standard ( $\sigma$ ) è la radice quadrata positiva della varianza, e viene espressa nella stessa unità di misura utilizzata nella raccolta dei dati.

La [tabella](#) esemplifica i vari passaggi da effettuare per giungere alla determinazione della  $\sigma$  nel caso di dati raggruppati in classe.

### ESERCIZI

**7.1** Calcola la deviazione standard dei seguenti dati:

232 456 56 34 565 123 100 33 56

**7.2** Calcola la deviazione standard dei seguenti dati:

x	4.9	5.0	5.1	5.2	5.3	5.4	5.5	5.6
f	54	47	43	29	21	20	14	9

**7.3** Costruisci il diagramma box-whisker e descrivi il tipo di distribuzione dei seguenti dati:

## 8 MISURE DI POSIZIONE

Le misure di posizione sono usate per descrivere il posizionamento di un'osservazione relativamente al resto dei dati.

### Percentile

Date  $N$  osservazioni ordinate (dal valore minimo al valore massimo) l' $n$ -esimo percentile ( $P_n$ ) è quell'osservazione che risulta così allocata:

$$\text{se } \left(\frac{n}{100}\right) \cdot N = \text{numero intero} \quad \text{allora} \quad P_n = \left(\frac{n}{100}\right) \cdot N + 0.5$$

$$\text{se } \left(\frac{n}{100}\right) \cdot N \neq \text{numero intero} \quad \text{allora} \quad P_n = \text{Parte intera} \left( \left(\frac{n}{100}\right) \cdot N + 1 \right)$$

### Standard Score

LO standard score ( $z$ ) esprime la distanza dell'osservazione, espressa in numero di  $\sigma$ , rispetto al valore medio:

$$z = \frac{y_i - \bar{y}}{s}$$

## ESERCIZI

**8.1** Calcola il 10 e il 60 percentile dei seguenti dati:

24	46	57	64	65	82	89	90	90	111
117	128	143	148	152	166	171	186	191	197
209	223	230	247	249	254	258	264	269	270
273	284	294	304	304	332	341	393	395	487
510	516	518	518	534	608	642	697	955	1160

**8.2** Trova lo standard score ( $z$ ) delle seguenti osservazioni appartenenti ad un insieme di dati con media 30 e  $\sigma=1$

**8.3** Con riferimento all'esercizio precedente trova il valore corrispondente ai seguenti standard score

-2      1      2.5      1.5      0

## 9 LA CURVA NORMALE

Per molte variabili casuali la funzione di densità di probabilità è una curva di tipo campanulare chiamata curva normale o gaussiana del tipo di quella indicata in [figura](#) ( $\mu=0$ ;  $\sigma=1$ )

La distribuzione normale è importante, tra l'altro, perché approssima molto bene la reale distribuzione di molte misure mediche<sup>7</sup> nella popolazione generale (uricemia, colesterolemia, pressione sanguigna, peso, statura etc...)

L'espressione analitica della funzione di probabilità di una distribuzione normale è la seguente:

$$p(x) = \frac{1}{s \sqrt{2\pi}} \exp\left[-\frac{(x - m)^2}{2s^2}\right]$$

con  $\sigma$  = deviazione standard                       $\mu$  = valore medio

L'area sottesa dalla curva rappresenta la popolazione. Il valore medio determina la posizione della distribuzione lungo l'asse delle ascisse, mentre l'entità della  $\sigma$  definisce 'l'ampiezza' della curva. A  $\sigma$  elevate corrispondono curve piatte (dati molto dispersi) a  $\sigma$  piccole corrispondono curve strette, espressioni di dati molto addensati rispetto al loro valore medio. ([figura](#))

### Curva Normale Standard

E' una curva normale con **media zero** e **deviazione standard unitaria**. L'espressione analitica della funzione di probabilità si semplifica:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

Le aree sottese da una curva normale standard sono tabulate e come vedremo possono essere utilizzate anche per applicazioni che riguardino distribuzioni con media e deviazione standard diverse.

### Standardizzazione di una curva normale generalizzata

Una qualsiasi curva normale può essere ricondotta ad una curva normale standard procedendo alla seguente trasformazione di variabile<sup>8</sup>:

$$z = \frac{x - m}{s}$$

---

<sup>7</sup> alcune misure mediche, che non si distribuiscono normalmente, possono ricondursi facilmente ad una distribuzione gaussiana mediante una semplice trasformazione logaritmica (vedi ad esempio le transaminasi)

<sup>8</sup> z prende il nome di deviatore normale

E facile dimostrare che:

$$\text{Media}(z) = 0 \quad \text{Dev. St.}(z) = 1$$

### ESERCIZI

**9.1** Nella curva normale standardizzata quale valore di  $z$  lascia:

il 10% nella coda di destra

il 5% nella coda di sinistra

**9.2** Quali livelli di  $z$  (simmetrici rispetto a zero) comprendono:

il 38.3%      68.3%      95.4%      100% della popolazione

**9.3** Si assuma che tra i non diabetici il livello di glicemia a digiuno sia distribuito in modo approssimativamente normale con media 105 mg/dl e una  $\sigma$  di 9 mg/dl.

a) quale percentuale di non diabetici ha livelli di glicemia compresi tra 90 e 125 mg/dl

b) quale livello di glicemia lascia una coda di destra pari al 10% dei non diabetici

c) quali livelli (simmetrici rispetto alla media) comprendono il 95% della popolazione di non diabetici

**9.4** Una popolazione ha un'altezza distribuita in modo normale con media 173 cm e  $\sigma = 7$  cm

Calcola la percentuale di persone che:

a) sono alte più di 180 cm; b) sono alte da 160 a 180 cm; c) sono alte meno di 160 cm

**9.5** Si supponga che  $\text{Log}(\text{ALT})$  sia distribuito, nella popolazione sana, con media 1.25 e  $\sigma$  0.12, mentre, nei soggetti affetti da epatite, con media 1.55 e  $\sigma$  0.13. Dobbiamo impostare uno screening che, sulla base della valutazione del livello di ALT, permetta di selezionare i donatori di una banca del sangue ospedaliera esenti da eventuale epatite anitterica.

a) si determini una soglia di livello di ALT tale che la procedura di screening porti ad accettare il 95% degli individui sani ;

b) determinare la percentuale di malati che sarebbero catalogati come falsi negativi dal livello di soglia calcolato al punto a.

**9.6** In un ampio gruppo di pazienti coronaropatici si trovò che il loro livello di colesterolo serico si distribuiva approssimativamente in modo normale. Si trovò inoltre che il 10% del gruppo aveva livelli di colesterolo al di sotto di 182.3 mg/dl, mentre il 5% aveva valori superiori a 359 mg/dl. Quali sono la media e la deviazione standard della distribuzione?

**9.7** Si supponga che nei maschi normali il livello di acido urico sia distribuito in modo approssimativamente normale con media 5.4 mg/dl e  $\sigma$  pari a 1 mg/dl.

a) quale è la probabilità che un maschio sano selezionato in modo random abbia un livello di acido urico serico al di fuori dell'intervallo 4.0-7.0 mg/dl?

b) quale è la probabilità che tra quattro maschi sani selezionati a caso ve ne sia almeno uno il cui livello di acido urico serico sia al di fuori dell'intervallo 4.0-7.0 mg/dl?

c) quanti maschi sani devono essere selezionati in modo tale che vi sia una probabilità almeno 0.95 che almeno uno abbia i livelli di acido urico serico al di fuori dell'intervallo 4.0-7.0 mg/dl?

**9.8** The length of life  $x$  (in months) of a hair dryer is approximately normally distributed with mean 96 and standard deviation 18.

a) the manufacturer decides to guarantee the product for 5 years. What percentage of the product will fail to satisfy the guarantee?

b) the manufacturer decides to replace only 1% of all hair dryers. What should the length (in months) of the guarantee be?

**9.9** Let  $x$  be the number of minutes after 11 o'clock a bus leaves the bus station. Assume that the distribution of times is approximately normal with mean 15 and standard deviation 4.

a) if a person gets to the bus station at 11:10, what is the probability the person has missed the bus?

b) if a person is willing to risk a 20% chance of not making the bus, what is the maximum number of minutes after 11 o'clock that the person can reach the station?

c) what time should the person reach the station in order to have a 50-50 chance of catching the bus?

## 10 CAMPIONAMENTO

Alcune definizioni:

- Universo:** insieme esaustivo di unità che è di interesse del ricercatore
- Campione:** sottoinsieme dell'universo estratto seguendo una procedura probabilistica
- Parametro:** una caratteristica dell'universo (ad esempio la percentuale dei fumatori)
- Stimatore:** corrispettivo campionario del parametro soggetto quindi a fluttuazione casuale di campionamento
- Inferenza:** è un processo logico deduttivo per effetto del quale da misurazioni ottenute su di un campione siamo in grado estendere le nostre conoscenze all'universo

### CAMPIONAMENTO DA POPOLAZIONE NORMALE

Sia assegnata una popolazione infinita A distribuita in modo normale con media  $\mu$  e deviazione standard  $\sigma$ . Immaginiamo ora di estrarre da A, con reinserimento<sup>9</sup>, infiniti campioni di numerosità n.

*Si può dimostrare che la media dei campioni così definiti (media campionaria) si distribuisce in modo normale con media pari alla media dell'universo generatore e deviazione standard pari a:*

$$\mathbf{F}_{\bar{x}} = \frac{\mathbf{F}}{\sqrt{n}} \quad \text{errore standard della media } ES(\bar{x})$$

(figura)

Il 95% delle medie degli infiniti campioni avrà una media che sarà compresa tra il valore medio dell'universo  $\mu \pm 1.96$  l'errore standard prima definito:

$$Pr\left(\mu - 1.96 \frac{\mathbf{F}}{\sqrt{n}} < \bar{x} < \mu + 1.96 \frac{\mathbf{F}}{\sqrt{n}}\right) = 95\%$$

Nota la media di un campione, di numerosità n, estratto da una popolazione infinita con deviazione standard  $\sigma$  si può allora affermare che la media dell'universo, nel 95% dei casi, sarà compresa tra il valore medio campionario  $\pm 1.96$  l'errore standard. Tale intervallo viene definito intervallo di confidenza, al 95%, della media .

$$Pr\left(\bar{x} - 1.96 \frac{\mathbf{F}}{\sqrt{n}} < \mu < \bar{x} + \frac{\mathbf{F}}{\sqrt{n}}\right) = 95\%$$

---

<sup>9</sup> se il campionamento viene effettuato da una popolazione infinita il reinserimento è, in realtà, del tutto ininfluente. Se invece il campionamento viene realizzato, senza reinserimento, da un popolazione finita di numerosità N, la deviazione standard campionaria vale:

$$\mathbf{F}_{\bar{x}} = \mathbf{F} \sqrt{\frac{N-n}{n(N-1)}}$$

E' facile vedere che, per N tendente all'infinito, la deviazione standard campionaria, coincide con il valore determinato campionando con reintroduzione

### CAMPIONAMENTO DA POPOLAZIONE QUALSIASI

Le considerazioni che prima sono state svolte prendendo come riferimento una popolazione normale sono sorprendentemente estensibili, quando i campioni hanno una numerosità sufficientemente elevata, ad una distribuzione qualsiasi. (figura)

Vale cioè il teorema del limite centrale o teorema fondamentale della statistica:

Aumentando la dimensione  $n$  del campione, la distribuzione campionaria delle medie, estratta da una popolazione *qualsiasi*<sup>10</sup>, si approssima alla distribuzione *normale* con media coincidente alla media della popolazione e deviazione standard campionaria pari al rapporto tra la deviazione standard della popolazione e la radice quadrata della numerosità campionaria.

### CAMPIONAMENTO DA POPOLAZIONI CON DEVIAZIONE STANDARD SCONOSCIUTA

In precedenza abbiamo sempre ritenuto di conoscere il valore della deviazione standard della popolazione che veniva utilizzata come base di campionamento (universo). Questa situazione, in realtà si presenterà molto di rado. Normalmente ci troveremo di fronte ad un campione estratto da un universo di cui saranno sconosciuti sia  $\mu$  sia  $\sigma$ .

Il campione tuttavia presenterà una certa dispersione ed è logico ritenere che questa informazione possa essere utilizzata per prevedere la dispersione dell'universo  $\sigma$ .

Si può dimostrare che la miglior stima della varianza dell'universo, deducibile dall'informazione campionaria è:

$$\text{Miglior stima di } \mathbf{F}^2 = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Le relazioni che definivano gli intervalli di confidenza della media saranno ancora valide, nel caso di deviazione standard dell'universo sconosciuta, a condizione di sostituire a  $\sigma$  il valore  $s$ <sup>11</sup> prima definito.

Nel caso di campioni sufficientemente numerosi possiamo scrivere allora:

$$Pr\left(\bar{x} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{s}{\sqrt{n}}\right) = 95\%$$

---

<sup>10</sup> un requisito è che la popolazione abbia varianza finita

<sup>11</sup>  $s$  prende il nome di deviazione standard campionaria ed è concettualmente molto diversa dalla deviazione standard del campione.

Nella formula precedente il valore **1.96** rappresenta la deviatore normale che comprende il 95% della popolazione. Si suppone perciò implicitamente che la variabile **t** (t di Student) di seguito definita abbia una distribuzione di tipo normale.

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

In realtà t approssima la normale solo per n abbastanza elevati. Quando n è piccolo l'intervallo di confidenza deve essere determinato con la seguente relazione:

$$Pr\left(\bar{x} - t_{0.025} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \frac{s}{\sqrt{n}}\right) = 95\%$$

## ESERCIZI

**10.1** Si faccia riferimento alla popolazione di maschi sani definita al problema 9.7:

- tra i campioni ripetuti di dimensione 25 estratti da questa popolazione, quale proporzione ha media pari o maggiore di 5.9 mg/dl;
- quale valore lascia alla sua destra il 5% della distribuzione delle medie campionarie di dimensione 25?
- quale deve essere la dimensione del campione perché il 5% delle medie dei campioni di tali dimensioni superi la media della popolazione di 0.2 mg/dl?

**10.2** Si consideri una popolazione di pazienti con sopravvivenza media 38.3 mesi e deviazione standard 43.3 mesi.

(Si noti che in questa popolazione il tempo di sopravvivenza non è distribuito in modo normale dato che, tra l'altro, la media meno  $\sigma$  fornisce un valore negativo. La dimensione campionaria di 100 o più unità, qui presa in considerazione, è certamente grande abbastanza da assicurare che la distribuzione di campionamento delle medie sia normale.)

- tra i campioni di grandezza 100 estratti da questa popolazione quale proporzione avrà sopravvivenza media superiore a 46.9 mesi?
- quali limiti comprendono il 95% delle medie dei campioni di dimensione 100 estratti da questa popolazione?
- quale dimensione del campione è necessaria perché il 95% delle medie campionarie sia compreso nell'intervallo  $\pm 6$  mesi dalla media della popolazione?

### 11 DIFFERENZA FRA DUE MEDIE (campioni indipendenti)

Spesso ci troveremo ad esaminare due campioni e ci verrà chiesto di valutare se entrambi siano stati estratti dal medesimo universo.

Se fosse vero che entrambi i campioni derivano dal medesimo universo allora dovremo utilizzare le due deviazioni standard campionarie per stimare la deviazione standard della popolazione di partenza. Un'ipotesi verosimile potrebbe essere quella di stimare la  $\sigma$  dell'universo tramite una media, in qualche modo pesata, delle due deviazioni campionarie. In effetti si può dimostrare che la miglior stima di  $\sigma^2$  è la  $s^2$  **pooled** così definita:

$$s^2_{pooled} = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Possiamo allora impostare il seguente test:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

se il valore di t stacca un'area molto piccola della distribuzione possiamo ritenere che molto difficilmente i due campioni provengono dallo stesso universo. In caso contrario affermiamo che non abbiamo ragione di ritenere che i due campioni provengano da universi differenti.

#### Schematizzazione del processo inferenziale

1) Si imposta un test statistico in grado di rifiutare o non rifiutare l'ipotesi nulla così definita:

$$H_0: \mu_1 = \mu_2$$

2) Lo sperimentatore stabilisce il livello di significatività del test ( $p < 0.05$ ). Con un livello di significatività del 5% si è disposti a correre il rischio del 5% di rifiutare erroneamente l'ipotesi nulla quando essa è vera (vediamo cioè delle differenze che in realtà non esistono)

3) Si calcola la t di Student

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

4) Il valore di t stacca un'area superiore o inferiore al 5%

$A(t) < 5\%$	Abbiamo elementi sufficienti per rifiutare l'ipotesi nulla	<b>(RIFIUTO)</b>
$A(t) > 5\%$	Non abbiamo elementi sufficienti per rifiutare l'ipotesi nulla	<b>(NON RIFIUTO)</b>

### CONFRONTO FRA CAMPIONI NON INDIPENDENTI

La caratteristica peculiare dei campioni non indipendenti è che ciascuna osservazione in un campione si accoppia con una e una sola osservazione dell'altro campione. Se, ed è il caso più probabile, la deviazione standard delle differenze non è nota il test utilizza la deviazione standard campionaria delle differenze ( $s_d$ ) e si basa sulla distribuzione t con n-1 gradi di libertà. Il rapporto critico sarà allora:

$$t_{n-1} = \frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$  = media delle differenze

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

Si noti che i gradi di libertà sono pari al numero di coppie considerate meno uno.

### ESERCIZI

(da T.Colton Statistica in medicina Piccin)

**11.1** Si supponga che la uricemia sia distribuita, nei maschi sani, in modo normale con media pari a 5.4 mg/dl e  $\sigma = 1$  mg/dl

a) qual è la probabilità che tra quattro maschi sani, selezionati a caso, il livello medio di uricemia sia al di fuori dell'intervallo 4-7 mg/dl?

b) quanti maschi sani devono essere selezionati perché vi sia una probabilità maggiore o uguale al 95% che il loro livello di acido urico serico medio sia entro l'intervallo 4.9-5.9 mg/dl?

**11.2** Si supponga che la colesterolemia sia distribuita, nei maschi sani, in modo normale con media pari a 240 mg/dl con  $\sigma = 50$  mg/dl

a) qual è la probabilità che un campione casuale di 36 elementi evidenzi un livello di colesterolemia media inferiore o uguale a 215 mg/dl?

**11.3** In un esperimento ciascun topo, di un campione casuale di 25 unità, deve essere iniettato con un farmaco ad un livello di dose di 0.004 mg per grammo di peso corporeo. Per questo ceppo di topi è noto che il peso è approssimativamente distribuito in maniera normale con una media di 19g e una  $\sigma$  di 4g

a) se il ricercatore possiede un totale di 2 mg del farmaco, qual è la probabilità che questo non sia sufficiente per trattare tutti i topi?

b) quanto farmaco dovrebbe possedere il ricercatore al fine di correre al più un rischio dell' 1% di non trattare tutti gli animali?

**11.4** In uno studio condotto su 49 uomini sani è emerso che i limiti di confidenza al 95% della capacità vitale media erano compresi tra 4.62 e 4.94 litri.

Se la capacità vitale nella popolazione sottostante è distribuita in modo normale, quale range (simmetrico rispetto al valore medio) comprende il 90% degli uomini sani?

**11.5** La tabella sotto riportata registra i risultati di uno studio condotto, su dati appaiati, allo scopo di valutare l'efficacia del trattamento dell'ipertensione tramite l'uso di idroclorotiazide.

paziente	placebo	ict
FB	211	181
IF	210	172
PG	210	196
HF	203	191
RR	196	167
LP	190	161
BK	191	178
IF	177	160
MK	173	149
MT	170	119
JM	163	156

a) Con un livello di significatività del 5%, vi sono elementi per affermare che il trattamento è efficace?

b) Sempre ad un livello di significatività del 5%, determinare i limiti di confidenza della differenza fra i trattamenti

(da W. Case General Statistics Wiley)

**11.6** An alkaline battery for an AM-FM stereo cassette radio was designed to last 30 hours, on average, for FM play. There were consumers complaints that batteries were lasting less than 30 hours. The manufacturer randomly sampled 38 batteries. The mean life was 29.3 hours with standard deviation of 2.95 hours. Is there sufficient evidence, at the 5% level of significance, to indicate the mean battery life is less than 30 hours?

**11.7** The Speedy Oil Change Company advertised a 15-minute wait for an oil change. A sample of 23 oil changes showed a mean time of 16.5 minutes with standard deviation of 4.2 minutes. At the 5% level of significance, is there evidence that the mean time for an oil change is different from 15 minutes?

**11.8** A medical doctor sensed that smokers in the 40-45 age group with a rare disease had smoked on average more than 20 years. A sample of 10 patients gave the following years of smoking

22.0 21.3 19.6 19.6 21.4 24.0 25.9 19.7 25.5 25.1

Using a 1% significance level, is there sufficient evidence to justify the doctor's belief?

## 12 IL CONTROLLO DI QUALITA'

Il controllo di qualità è uno strumento che consente di valutare se una determinata procedura fornisce delle prestazioni compatibili con un prefissato standard.

Le caratteristiche di produzione variano per effetto di due gruppi di cause:

- a) gruppo limitato di cause ( differenze tra macchine, tra addetti, tra materie prime.....) perfettamente individuabili e controllabili in modo da ottenere le caratteristiche produttive desiderate;
- b) una miriade di cause non individuabili né controllabili ciascuna delle quali esercita un'influenza impercettibile sul processo produttivo.

Posto che il primo gruppo di cause sia mantenuto costante, la variabilità del processo dipende in misura casuale dall'insieme dei fattori appartenenti al secondo gruppo.

### Carte di Controllo

Le carte di controllo rappresentano uno degli strumenti più importanti per il controllo statistico di qualità. La carta di controllo è corredata da tre rette parallele all'asse delle ascisse che esprimono, quella centrale o **media**, il valore medio della statistica o la sua stima ottenuta in base ai primi  $m$  campioni, e quelle esterne, dette di **controllo**, gli estremi della banda di ampiezza pari a sei volte la deviazione standard  $\sigma$ , o la sua stima  $s$ .

Le carte di controllo possono essere di due tipi:

#### 1) carte di controllo per attributi

Si usano quando la qualità può essere espressa dalla identificazione di una o più caratteristiche qualitative.

#### 2) carte di controllo per variabili

Si usano quando la qualità può essere espressa da caratteristiche quantitative.

Le principali carte di controllo per variabili sono:

- a) carta di controllo per la **media**
- b) carta di controllo per **il range**

### Carta di controllo per la media

Per la costruzione di questa carta di controllo viene considerata una successione di  $m$  campioni tutti di ampiezza  $n$ . La miglior stima della deviazione standard della popolazione vale:

$$s = \sqrt{\frac{\sum_{i=1}^{n \cdot m} (x_i - \bar{x})^2}{nm-1}}$$

Lo stimatore della media  $\mu$  della popolazione vale:

$$\bar{x} = \frac{\sum_{i=1}^{n \cdot m} x_i}{n \cdot m} = \frac{\sum_{i=1}^m \bar{x}_i}{m}$$

La carta di controllo della media sarà allora, ricordando le proprietà della distribuzione di campionamento, caratterizzata dalle seguenti rette:

$$y = \bar{x} \quad \text{retta centrale}$$

$$y = \bar{x} \pm 3 \frac{s}{\sqrt{n}} \quad \text{rette di controllo}$$

### Carta di controllo per il range (Range chart)

Viene usata quando occorre verificare la variabilità di un processo. La range chart viene realizzata in modo del tutto analogo alla carta di controllo della media.

- 1) Per ogni campione si determina il range  $R_i$
- 2) si determinano il valore centrale e i limiti superiore e inferiore della carta tramite il valore medio del range.

$$\bar{R} = \frac{\sum_{i=1}^m R_i}{m}$$

$$UCL = \bar{R} D_4 \quad LCL = \bar{R} D_3$$

I valori di  $D_4$  e  $D_3$  sono tabellati in funzione della numerosità  $n$  dei campioni.<sup>12</sup>

([figura 1](#) [figura 2](#))

---

<sup>12</sup> Il range è uno stimatore distorto di  $\sigma$ . Il valore atteso di  $R$  e la deviazione standard del range campionario possono essere espressi in funzione della  $\sigma$ .

$$E(R) = d_2 \sigma \quad \sigma_R = d_3 \sigma$$

dove  $d_2$  e  $d_3$  sono costanti tabellate in funzione di  $n$  (numerosità campionaria).  
La miglior stima di  $\sigma$  è:

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

La linea centrale sarà definita da  $d_2 s$ , mentre i limiti superiori e inferiori saranno definiti dalle seguenti relazioni:

$$UCL = d_2 s + 3d_3 s$$

$$LCL = d_2 s - 3d_3 s$$

Il valore atteso del range può anche essere stimato dal suo valore medio. In tal caso le linee caratteristiche saranno identificate dalle seguenti relazioni:

Linea centrale:  $\bar{R}$

$$UCL: \quad \bar{R} + 3 \frac{d_3}{d_2} \bar{R} = \bar{R} \left( 1 + 3 \frac{d_3}{d_2} \right) = \bar{R} D_4$$

$$LCL: \quad \bar{R} - 3 \frac{d_3}{d_2} \bar{R} = \bar{R} \left( 1 - 3 \frac{d_3}{d_2} \right) = \bar{R} D_3$$

### 13 COME ANALIZZARE PROPORZIONI

Consideriamo una popolazione sufficientemente ampia costituita dal 40% da maschi e dal 60% da femmine.

La probabilità di estrarre a caso un soggetto di sesso femminile sarà  $p = 0.6$

La variabile considerata (sesso) è una variabile dicotomica ossia può assumere soltanto due valori (maschio/femmina)

Associamo alla condizione femmina il valore  $X=1$  e alla condizione maschio il valore  $X=0$ .

Con questa schematizzazione la percentuale di popolazione femminile può essere interpretata come il valore medio della variabile  $X$  prima definita. Infatti:

$$\mu = \frac{\sum X}{N} = \frac{1+1+1+1+\dots\dots\dots+0+0+0\dots\dots}{N} = 0.6 = p$$

Visto allora che possiamo maneggiare la  $p$  come se si trattasse di una media di una variabile fittizia associata ( $X$ ), vediamo di definirne, se possibile, in modo analogo la deviazione standard.

Per definizione la deviazione standard di una variabile  $X$  presente in una popolazione di numerosità  $N$  vale:

$$\mathbf{F} = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

Se la variabile  $X$  è dicotomica la deviazione standard può essere facilmente espressa in funzione di  $p$

$$\mathbf{F} = \sqrt{p(1-p)}$$

Poiché  $\sigma^{13}$  dipende solamente da  $p$ , essa non contiene alcuna informazione aggiuntiva. Tuttavia sarà utile per definire il valore dell'errore standard associato **alla stima di  $p$**  basata su campioni estratti casualmente da popolazioni come quella definita in precedenza.

---

<sup>13</sup> la deviazione standard raggiunge il valore massimo (0.5) quando  $p$  vale 0.5, mentre per  $p=0$  o per  $p=1$  la deviazione standard assume valore nullo

## STIMA DI PROPORZIONI OTTENUTE DA CAMPIONI

Supponiamo di estrarre, dalla popolazione prima definita e con reinserimento, infiniti campioni tutti di numerosità  $n$ .

In ogni campione vi sarà una diversa proporzione di femmine ( $p_1$   $p_2$   $p_3$  ..... )

La distribuzione di campionamento sopra considerata ha tre proprietà fondamentali:

- 1) la media delle proporzioni campionarie tende alla proporzione della popolazione generale  $p$
- 2) l'errore standard della media vale<sup>14</sup>

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

- 3) la forma della distribuzione è approssimativamente normale, posto che  $n$  sia sufficientemente grande<sup>15</sup>

Per il computo degli intervalli di confidenza e per l'impostazione dei test statistici possiamo, anche nel caso di variabili dicotomiche, utilizzare le formule che erano state ricavate in precedenza con riferimento a variabili continue.

Intervallo di confidenza di una proporzione

$$Pr \left( p_c - 1.96 \sqrt{\frac{p_c(1-p_c)}{n}} < p < p_c + 1.96 \sqrt{\frac{p_c(1-p_c)}{n}} \right) = 95\%$$

Confronto fra due proporzioni

$$p_c = \frac{p_{1c}n_1 + p_{2c}n_2}{n_1 + n_2} \quad z = \frac{p_{1c} - p_{2c}}{\sqrt{p_c(1-p_c) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

<sup>14</sup> l'errore standard [ES] dipende dalla vera proporzione  $p$  che è, in genere, l'incognita del problema. Come nel caso di una variabile continua si può dimostrare che la miglior stima di ES, ottenibile dall'esame di un solo campione, vale:

$$ES = s = \sqrt{\frac{p_c(1-p_c)}{n}} \quad p_c = \text{proporzione campionaria}$$

<sup>15</sup> per sufficientemente grande intendiamo che deve essere  $np \geq 5$  e  $n(1-p) \geq 5$   
 Nel caso che una o entrambe le condizioni non fossero rispettate, si renderebbe necessario risolvere il problema in modo esatto utilizzando la distribuzione binomiale anziché quella normale.

## ESERCIZI

# 13.1 La tabella sotto riportata illustra i risultati di una sperimentazione clinica controllata condotta su pazienti dializzati allo scopo di valutare l'efficacia di basse dosi di aspirina nei confronti dello sviluppo di trombi.

I pazienti sono stati randomizzati e lo studio è stato condotto in doppio cieco. [Glantz]

Gruppo campionario	Numero di pazienti	
	Sviluppo trombi	Non sviluppo trombi
Placebo	18	7
Aspirina	6	13

Commenta i risultati

- 1) l'aspirina si è dimostrata efficace?
- 2) perché i pazienti sono stati randomizzati?
- 3) cosa si intende per studio condotto in doppio cieco?
- 4) ritieni etico trattare un gruppo di pazienti con placebo?

# 13.2 La tabella sotto riportata illustra i risultati di una sperimentazione clinica atta a valutare la relazione tra tipo di anestesia e mortalità negli interventi chirurgici a cuore aperto. [Glantz]

Anestetico	Vivi	Morti	Tot
Alotano	53	8	61
Morfina	57	10	67
Tot	110	18	128

Commenta i risultati

- 1) vi è evidenza di una associazione tra tipo di anestesia e mortalità?
- 2) sarebbe corretto impostare questo studio in doppio cieco?

**# 13.3** La tabella sotto riportata si riferisce ai risultati di uno studio condotto allo scopo di valutare l'efficacia del propranololo nei pazienti infartuati. I due gruppi di pazienti sono quelli trattati con propranololo e un gruppo di controllo che non riceve alcun farmaco. [Colton]

La risposta dicotomica consisteva nell'essere ciascun paziente ancora vivo al ventottesimo giorno dopo la sua immissione allo studio, o nell'essere egli venuto a mancare in un certo momento compreso entro questo periodo di 28 giorni

Gruppo campionario	Vivi a 28 gg	Deceduti Tot	
propranololo	38	7	45
controllo	29	17	46
Tot	67	24	91

Commenta i risultati

- 1) ritieni efficace l'uso del propranololo?
- 2) avresti impostato questo studio in doppio cieco?
- 3) ammesso che lo studio in oggetto abbia dimostrato una efficacia del propranololo a 28 giorni, potresti reimpostare uno studio analogo al precedente ma che abbia come scopo la valutazione dell'efficacia del propranololo a 60 giorni?

## CONFRONTO FRA PIU' DI DUE PROPORZIONI [Armitage]

Siano dati  $k$  gruppi di osservazioni e nel gruppo  $i$ -esimo siano stati osservati  $n_i$  soggetti dei quali  $r_i$  presentano una certa caratteristica (sono cioè positivi). Si indica con  $p_i$  la frequenza relativa di positivi  $r_i/n_i$ . I dati possono essere rappresentati come segue:

Gruppo	1	2	3	$k$	totale
Positivi	$r_1$	$r_2$	$r_3$	$r_k$	$R$
Negativi	$n_1 - r_1$	$n_2 - r_2$	$n_3 - r_3$	$n_k - r_k$	$N - R$
Totale	$n_1$	$n_2$	$n_3$	$n_4$	$N$
Freq. di positivi	$p_1$	$p_2$	$p_3$	$p_4$	$P = R/N$

Le frequenze assolute formano una tavola di contingenza  $2 \times k$ ; vi sono infatti 2 righe e  $k$  colonne, escludendo i totali marginali. In corrispondenza di ogni frequenza osservata  $O$  si calcola la frequenza attesa  $E$  mediante la formula

$$E = \frac{\text{Totale Riga} \times \text{Totale Colonna}}{N}$$

Si calcola quindi la quantità  $(O - E)^2 / E$  e infine

$$X^2 = \sum \frac{(O - E)^2}{E} = \frac{\sum n_i (p_i - P)^2}{P(1 - P)}$$

sommando sulle  $2k$  celle della tabella.

In base all'ipotesi nulla che tutti i campioni siano scelti a caso da una popolazione con la stessa frequenza relativa di positivi,  $X^2$  è distribuita approssimativamente come  $X^2_{(k-1)}$  [chi-quadro con  $k-1$  gradi di libertà]. L'approssimazione migliora all'aumentare delle frequenze attese, e il metodo può essere applicato con una certa tranquillità se le frequenze attese sono superiori a 5.

### ESERCIZI

# 13.4 Risolvere, col metodo del chi-quadro gli esercizi 13.1, 13.2, 13.3

# 13.5 La tabella sotto riportata mostra l'associazione fra la quantità di acqua bevuta e il tasso con cui gli abitanti si ammalarono di gastroenterite.

Acqua bevuta [bicchieri/giorno]	Ammalati	Non ammalati
Meno di 1	39	121
Da 1 a 4	265	258
5 o più	265	146

# 13.6 La tabella sotto rappresentata riporta i risultati di uno studio clinico randomizzato condotto su tre gruppi di ragazze di età compresa tra i 3 e i 16 anni, che erano state colpite in passato da infezioni ricorrenti del tratto urinario.

I tre farmaci hanno efficacia diversa?

Antibiotico	Ricaduta	Non ricaduta
Ampicillina	20	7
Trimetoprim-Sulfametoxazolo	24	19
Cephalexin	14	1

# 13.7 Le stime di sensibilità di due test per diagnosticare la stessa malattia sono risultati del 75% per T1 (su un campione di 40 malati) e dell'87% per T2 (su un campione di 60 malati). Si può concludere che T2 è più sensibile di T1?

# 13.8 Un dado ha solamente i numeri 1, 2, 3 ripetuti 3 volte. Lanciandolo 90 volte ottengo il numero 1 20 volte, il numero 2 30 volte e il numero 3 40 volte.

Vi è evidenza che il dado sia sbilanciato?

## 14 MISURE DI ASSOCIAZIONE

### **Rischio** [Risk]

Esprime la probabilità di un soggetto di poter sperimentare un determinato evento (malattia). Si può determinare tramite il rapporto tra i soggetti che sperimentano un evento (N) e coloro che ne sono stati a rischio (P)

$$R = \frac{N}{P}$$

Il concetto di rischio non può essere individualizzato, ma deve essere sempre riferito al gruppo di persone oggetto di studio. Se una persona muore, noi non potremo mai sapere se si tratti di una persona ad alto rischio di morte, oppure di una persona a basso rischio, ma particolarmente sfortunata.

### **Tasso** [Rate]

Il concetto di tasso è legato al concetto di rischio, ma è matematicamente distinto. Si può determinare tramite il rapporto tra il numero di eventi (N) e il numero di persone-tempo.

### **Rischio Relativo**

E' il rapporto tra i rischi calcolati su due gruppi di soggetti di cui uno solo esposto al fattore in esame.

### **Rischio Attribuibile (di popolazione)**

Proporzione della malattia nella popolazione che è attribuibile alla caratteristica. Il RA esprime la diminuzione percentuale nell'incidenza di una malattia se l'intera popolazione cessasse di essere esposta al sospetto agente etiologico.

### Metodi di calcolo

L'odds ratio (OR) può essere espresso come rapporto crociato (CP).

Quando la malattia è rara, indipendentemente dal tipo di studio, OR è una buona stima di RR.

	MALATI	NON MALATI	
ESPOSTI	a	b	n1
NON ESPOSTI	c	d	n2
	m1	m2	N

$$RR \cong ODDS\ RATIO = \frac{a\ d}{c\ b} \qquad RA = \frac{a-b}{1-b}$$

$$Z_{tab} = \frac{(ad - bc) \sqrt{(N-1)}}{\sqrt{n_1 n_2 m_1 m_2}} \qquad ES\ \log(OR) = \frac{\log(OR)}{Z_{tab}}$$

$$I.C.(\log(OR)) = \log(OR) \mp 1.96\ ES\ \log(OR) \qquad [Miettinen]$$

$$I.C.(\log(OR)) = \log(OR) \pm 1.96\ \sqrt{VAR(OR)} \qquad [Woolf]$$

$$VAR(OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Intervalli di confidenza di una proporzione secondo Miettinen [Armitage pag.119]

n = numerosità del campione      p = r/n

$$\mathbf{B}_L = \frac{r}{r + (n-r+1) F_{0.025, 2n-2r+2, 2r}}$$

$$\mathbf{B}_U = \frac{r+1}{r+1 + (n-r) (F_{0.025, 2r+2, 2n-2r})^{-1}}$$

Cognome.....Nome.....

1) Ogni volta che un individuo riceve prodotti ricavati da un pool di sangue, vi è una probabilità del 1.5% che sviluppi epatite virale. Un individuo riceve prodotti provenienti da pool di sangue per 5 volte. Qual è la sua probabilità di sviluppare un'epatite serica?

- 7.5%                       92.%                       7.28%

2) Associate con 3 malattie A,B,C, vi sono le rispettive probabilità del 20%, 35%, 55% di essere ricoverati in ospedale. Se un individuo ha tutte e tre le malattie (A,B e C) e queste esercitano il loro effetto indipendentemente, qual è la probabilità per questo soggetto di essere ricoverato in ospedale?

- 55%                       110%                       76.6%

3) Un test con sensibilità 0.85 e specificità 0.8 viene introdotto per valutare la presenza di una malattia che si presume sia attribuibile allo 0.1% della popolazione.

Determina:

- |           |                          |        |                          |        |                          |       |
|-----------|--------------------------|--------|--------------------------|--------|--------------------------|-------|
| Pr(+   S) | <input type="checkbox"/> | 20%    | <input type="checkbox"/> | 80%    | <input type="checkbox"/> | 15%   |
| Pr(-   M) | <input type="checkbox"/> | 85%    | <input type="checkbox"/> | 15%    | <input type="checkbox"/> | 20%   |
| Pr(M   +) | <input type="checkbox"/> | 1.42%  | <input type="checkbox"/> | 12.5%  | <input type="checkbox"/> | 0.42  |
| Pr(S   -) | <input type="checkbox"/> | 99.98% | <input type="checkbox"/> | 9.998% | <input type="checkbox"/> | 12.5% |

4) Un test con sensibilità 0.8 e specificità 0.7 viene sottoposto ad una popolazione di 4000 sani e 800 malati.

Determina il numero di veri positivi (TP) e di veri negativi (TN)

- |    |                          |      |                          |      |                          |     |
|----|--------------------------|------|--------------------------|------|--------------------------|-----|
| TP | <input type="checkbox"/> | 3840 | <input type="checkbox"/> | 640  | <input type="checkbox"/> | 560 |
| TN | <input type="checkbox"/> | 2800 | <input type="checkbox"/> | 3200 | <input type="checkbox"/> | 640 |

5) Un processo epidemico che si è sviluppato in un tempo relativamente recente sarà caratterizzato da:

- elevata incidenza e prevalenza ~~a~~ elevata incidenza e bassa prevalenza  
 bassa incidenza e prevalenza  bassa incidenza e elevata prevalenza

6) Una grave epidemia (A) è stata debellata recentemente con l'introduzione di un nuovo principio attivo.

La malattia A sarà ora caratterizzata da:

- elevata incidenza e prevalenza ~~a~~ elevata incidenza e bassa prevalenza  
 bassa incidenza e prevalenza  bassa incidenza e elevata prevalenza

Cognome.....Nome.....

7\*) In un vasto gruppo di pazienti inviati alla visita specialistica perché sospetti di essere affetti da sindrome di Cushing alla fine si trovò che ogni tre pazienti che effettivamente avevano la malattia ve ne era uno che non l'aveva. Inoltre il 65% di coloro che alla fine erano diagnosticati come portatori della malattia, mostravano osteoporosi all'esame iniziale. In contrapposizione, soltanto il 3%, tra coloro che alla fine furono diagnosticati come non affetti da malattia, avevano mostrato osteoporosi all'esame iniziale.

[ AIUTO       $Pr(C)=3/4$        $Pr(NC)=1/4$        $Pr(O|C)=65%$        $Pr(O|NC)=3%$   
 $Pr(NO|C)=...$        $Pr(NO|NC)=.....]$

a) qual è la probabilità che un paziente che si presenti con osteoporosi all'esame iniziale alla fine sia diagnosticato affetto da sindrome di Cushing [si chiede la  $Pr(C|O)=VPP.....]$

- 49.55%                                       48.75%                                       65%

b) se l'osteoporosi viene usata come procedura di screening per la sindrome di Cushing quali sono le frequenze di falsi positivi [ $Pr(O|NC)$ ]

- 25%     3%     65%

e di falsi negativi [ $Pr(NO|C)$ ]

- 25%     3%     35%

c) assumendo che i pazienti si presentino in ordine casuale quale proporzione si troverà:

c1) con osteoporosi iniziale e malattia

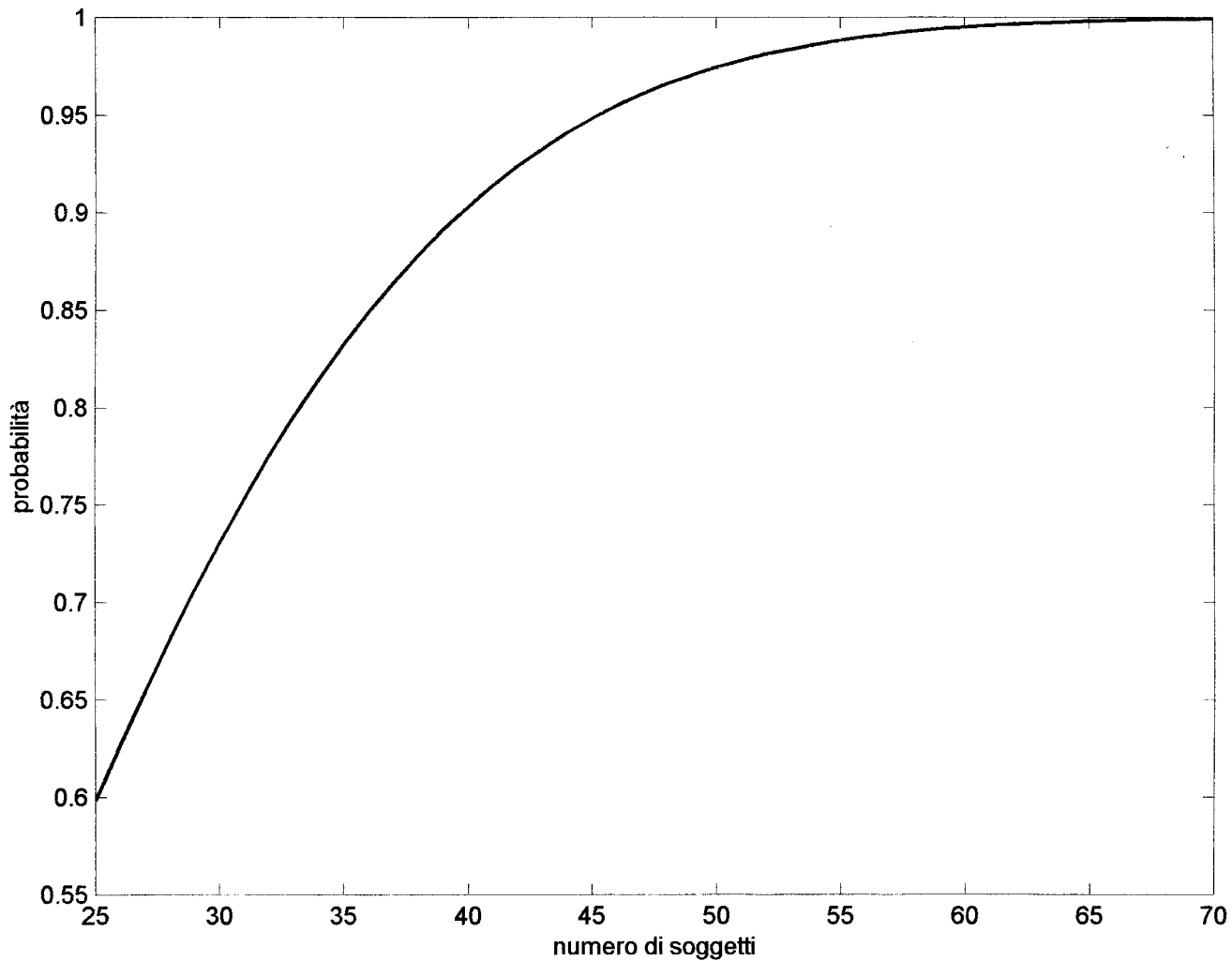
- 35%     16.25%     48.75%

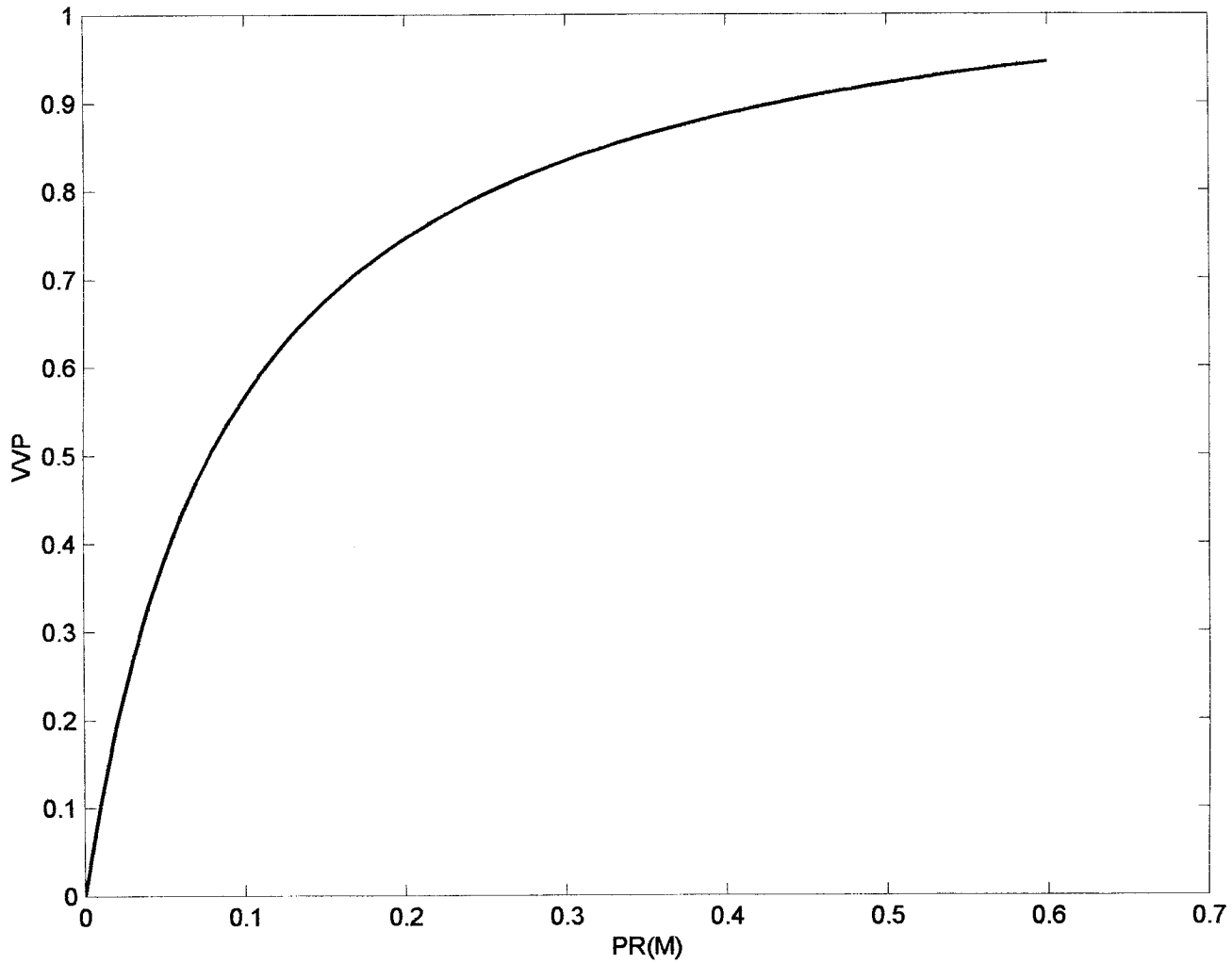
c2) con osteoporosi iniziale e priva di malattia

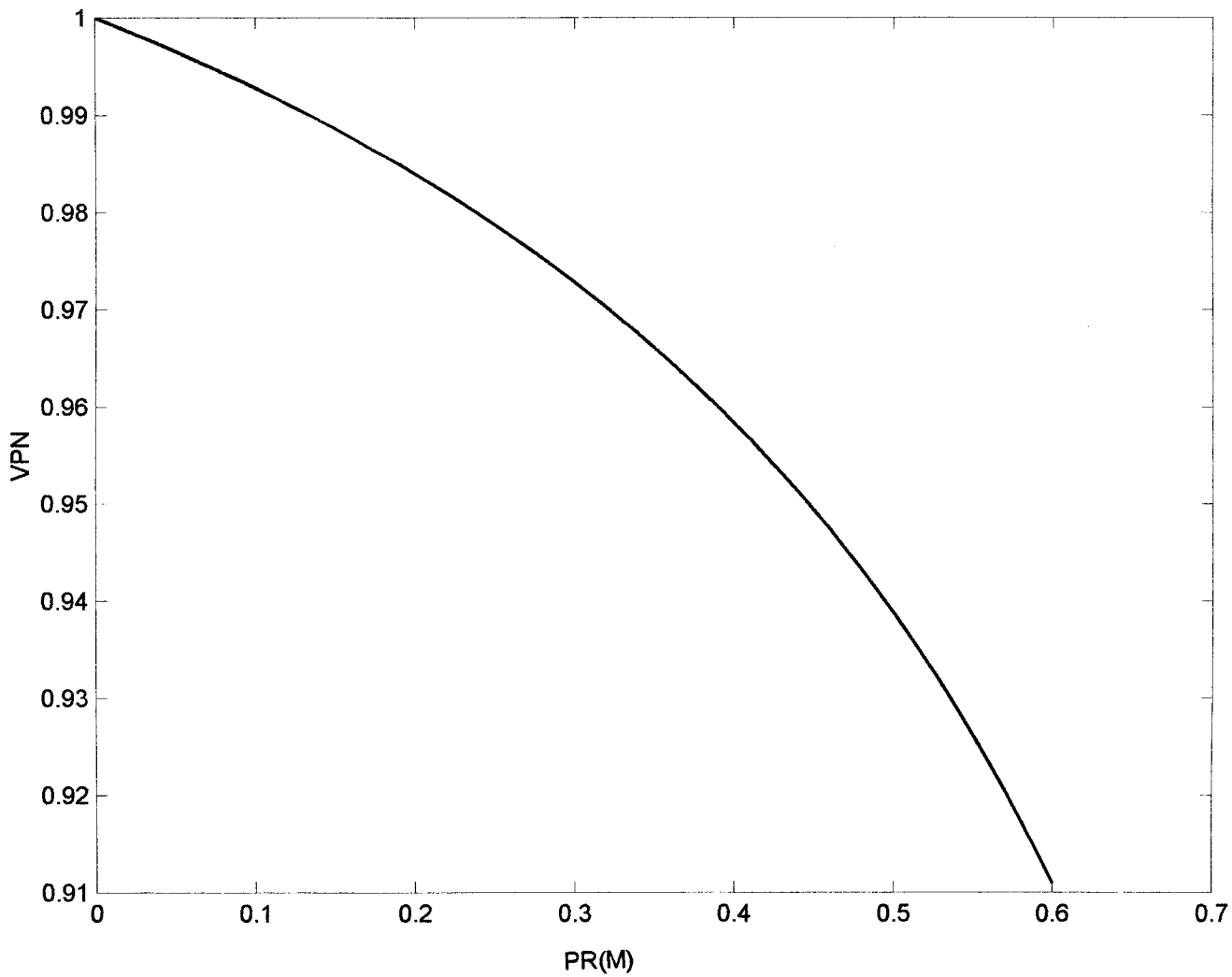
- 3%     49.55%     0.75%

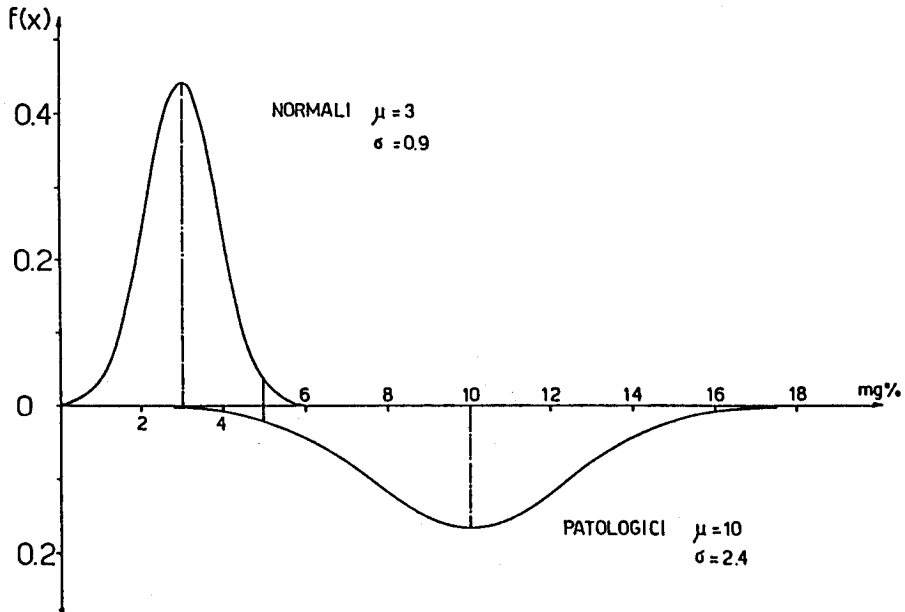
c3) senza osteoporosi iniziale e priva di malattia

- 25%     24.75%     0.75%





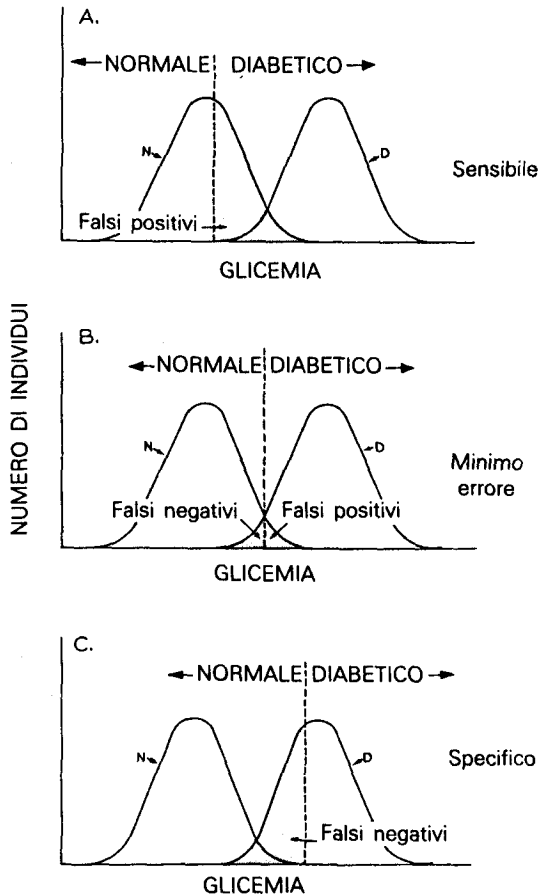




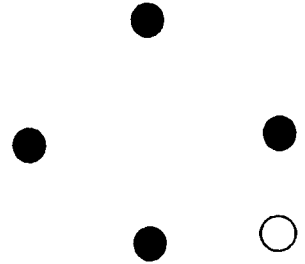
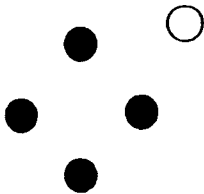
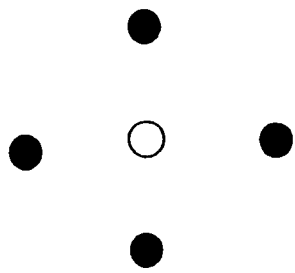
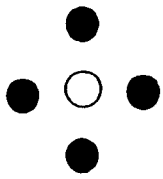
Sensibilità e specificità dei valori di glicemia a due ore dai pasti in un gruppo di 70 veri diabetici e di 510 veri non diabetici per differenti livelli della soglia di glicemia.

Livelli di glicemia (mg/100 ml)	Sensibilità (percentuale di diabetici identificata)	Specificità (percentuale di non diabetici identificata)
80	100.0	1.2
90	98.6	7.3
100	97.1	25.3
110	92.9	48.4
120	88.6	68.2
130	81.4	82.4
140	74.3	91.2
150	64.3	96.1
160	55.7	98.6
170	52.9	99.6
180	50.0	99.3
190	44.3	99.8
200	37.1	100.0

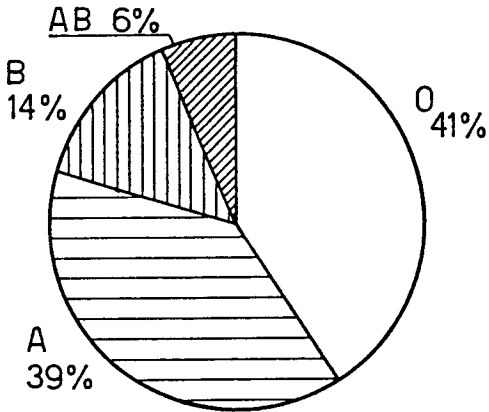
Fonte: United States Public Health Service

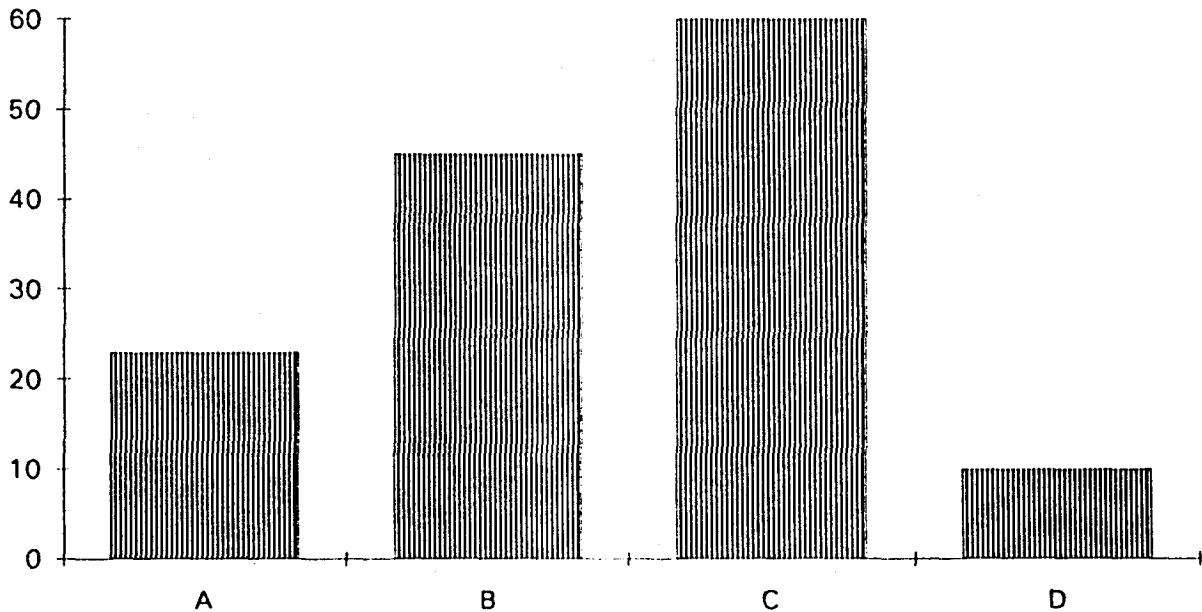


Effetto di differenti valori soglia di glicemia su falsi positivi e falsi negativi: (A) una soglia bassa significa che il test è più sensibile; (B) una soglia intermedia comporta un'errore globale minimo e (C) una soglia alta conduce ad un test più specifico.

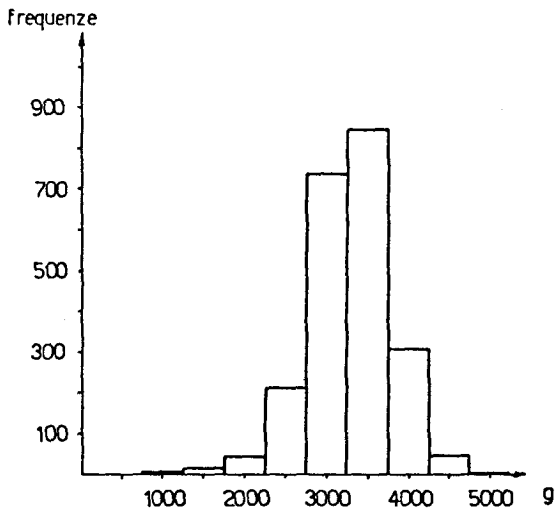


# GRUPPO SANGUIGNO





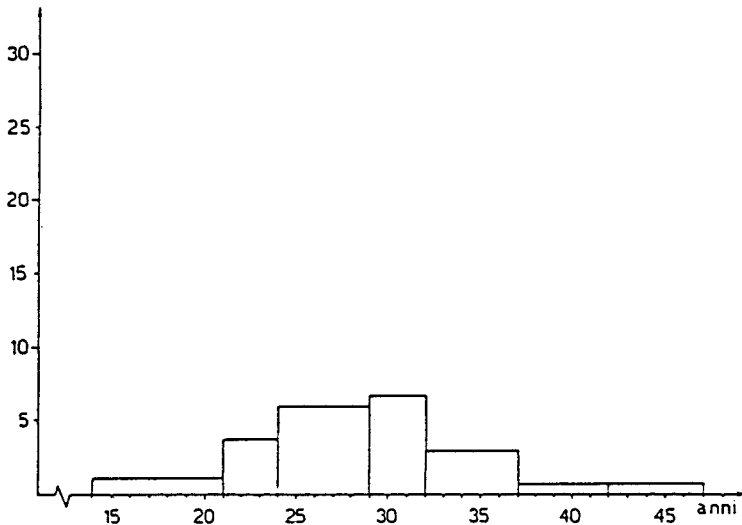
## PESO ALLA NASCITA



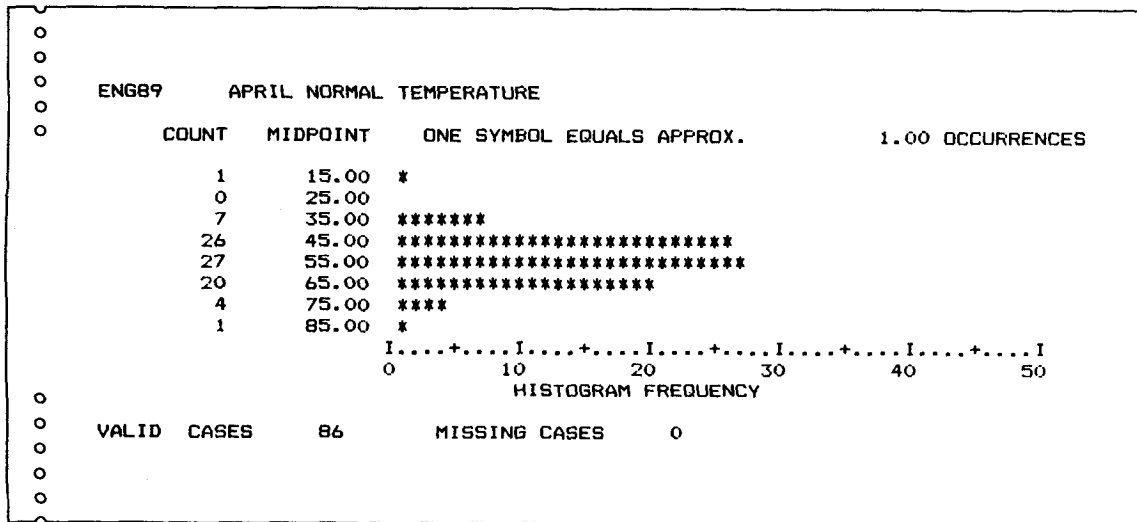
## Istogramma a canne d'organo (classi di diversa ampiezza)

### ETA' DELLA MADRE AL PARTO

frequenze



Computer-generated histogram for April normal temperature data, obtained using SPSS.





Comunità A:

		<u>Artrite Reumatoide</u>		
		<u>Presente</u>	<u>Assente</u>	<u>Totale</u>
Risultato del testo	Positivo	887	888	1775
	Negative	99	7989	8088
	Totale	986	8877	9863

Comunità B

		<u>Artrite Reumatoide</u>		
		<u>Present</u>	<u>Assente</u>	<u>Totale</u>
Risultato del testo	Positivo	1485	385	1870
	Negative	165	3465	3630
	Totale	1650	3850	5500

Coimunità C:

		<u>Artrite Reumatoide</u>		
		<u>Presente</u>	<u>Assente</u>	<u>Totale</u>
Risultato del testo	Positivo	3634	269	3903
	Negative	404	2423	2827
	Totale	4038	2692	6730

Comunità D:

		<u>Artrite Reumatoide</u>		
		<u>Presente</u>	<u>Assente</u>	<u>Totale</u>
Risultato del testo	Positivo	2866	56	2922
	Negative	318	506	824
	Totale	3184	562	3746

**Distribuzione di frequenza con classi di ampiezza costante (= 500 g): peso alla nascita**

---

<i>Peso alla nascita</i>	<i>Valore centrale della classe</i>	<i>Frequenze</i>
750 ┆ 1250	1000	8
1250 ┆ 1750	1500	17
1750 ┆ 2250	2000	46
2250 ┆ 2750	2500	214
2750 ┆ 3250	3000	737
3250 ┆ 3750	3500	846
3750 ┆ 4250	4000	308
4250 ┆ 4750	4500	47
4750 ┆ 5250	5000	4
	mancante	2
<i>Totale</i>		2229

---

## Calcolo della deviazione standard su dati raggruppati in classi

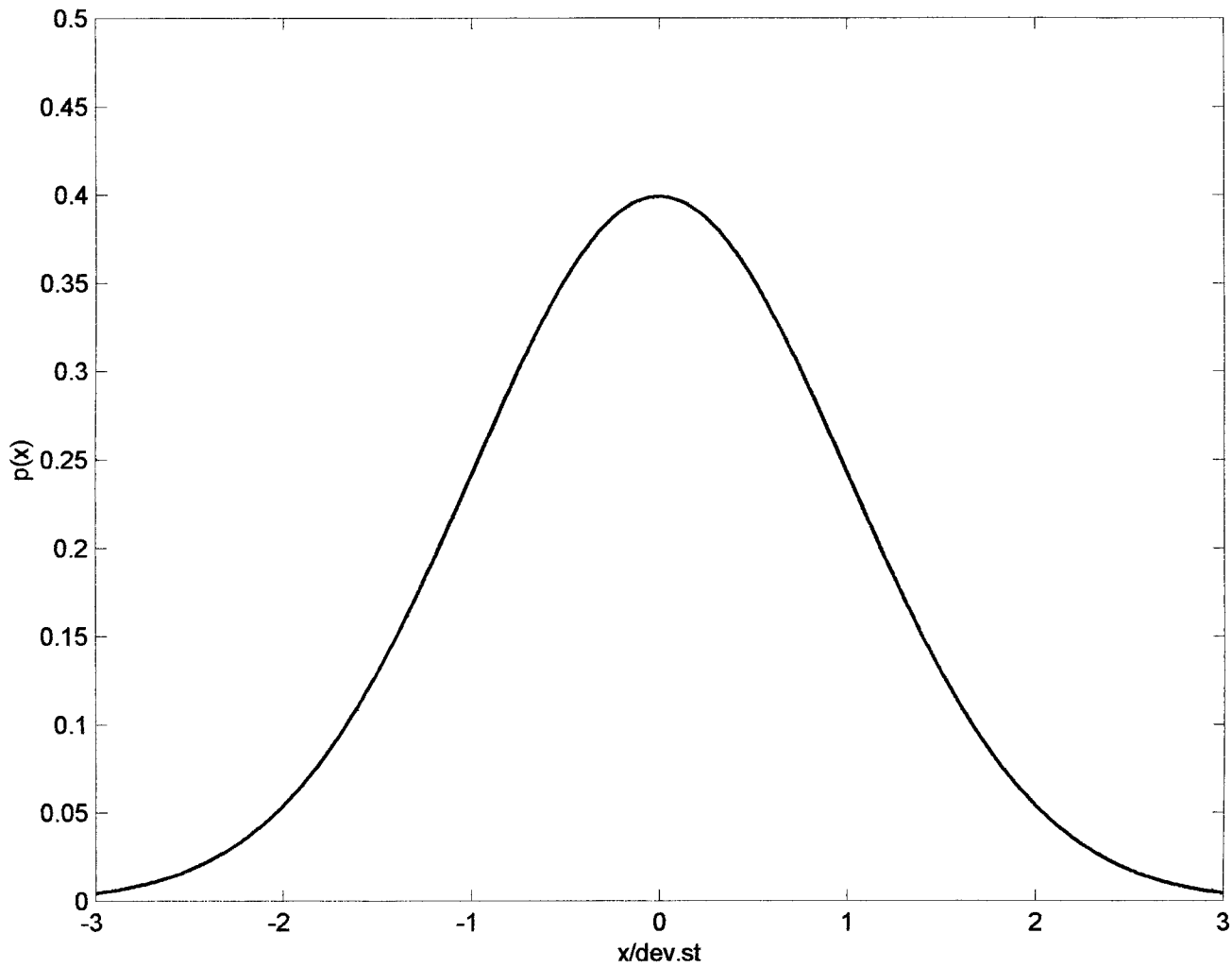
Valore centrale della classe	Frequenze	Scarti dalla media					
y	f	y · f	(y - M)	(y - M) <sup>2</sup>	(y - M) <sup>2</sup> · f	y <sup>2</sup>	y <sup>2</sup> · f
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2500	1	2500	-762.5	581406	581406	6250000	6250000
2750	12	33000	-512.5	262656	3151875	7562500	90750000
3000	13	39000	-262.5	68906	895781	9000000	117000000
3250	24	78000	-12.5	156	3750	10562500	253500000
3500	19	66500	237.5	56406	1071719	12250000	232750000
3750	9	33750	487.5	237656	2138906	14062500	126562496
4000	1	4000	737.5	543906	543906	16000000	16000000
4250	1	4250	987.5	975156	975156	18062500	18062500
<b>Totali</b>	<b>80</b>	<b>261000</b>			<b>9362500</b>		<b>860874996</b>

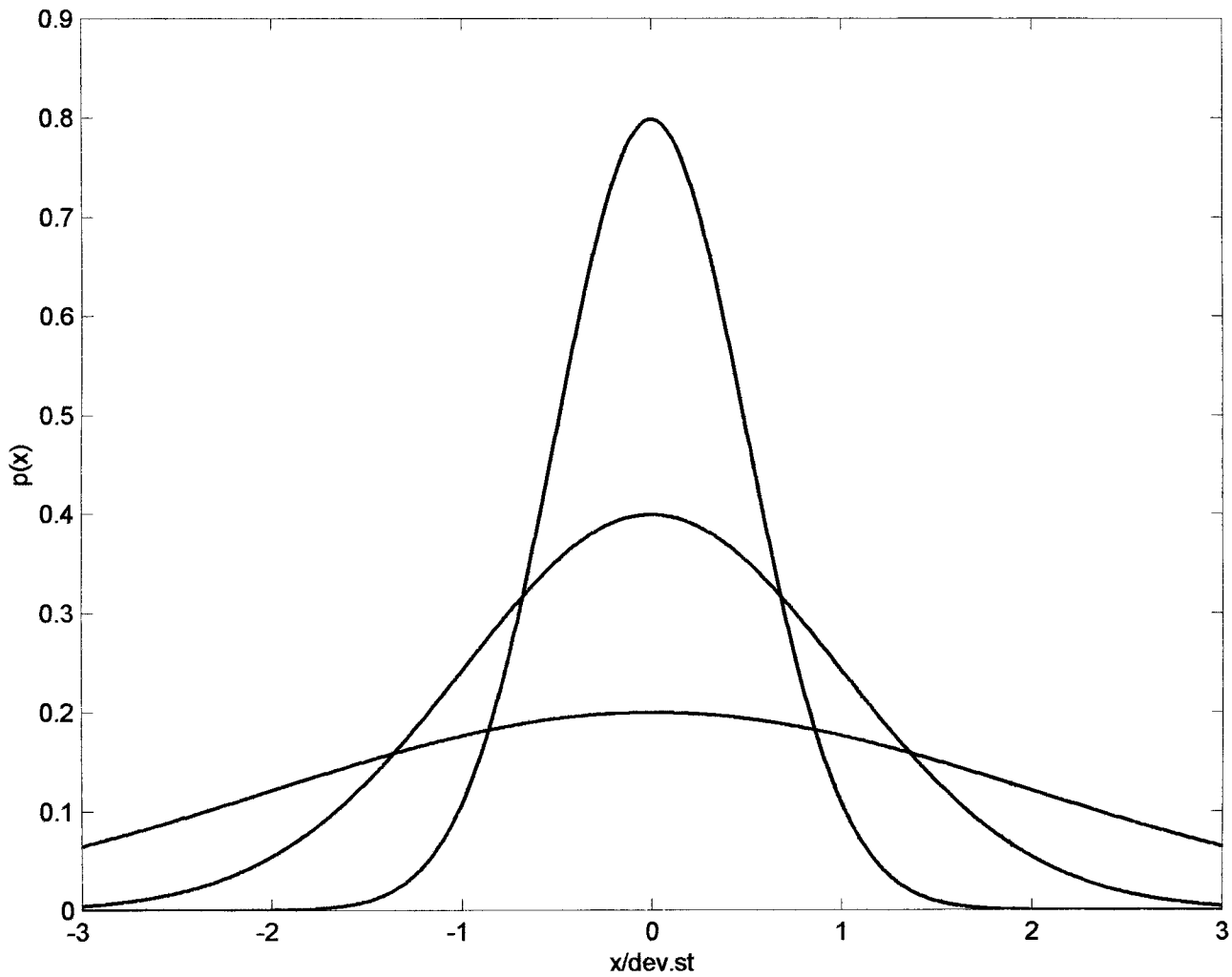
$$D.S. = \sqrt{\frac{\sum_{i=1}^n (y_i - M)^2 \cdot f_i}{\sum_{i=1}^n f_i}} = \sqrt{\frac{9362500}{80}} = \sqrt{117031.25} = 342.0983$$

oppure:

$$D.S. = \sqrt{\frac{\sum_{i=1}^n (y_i^2 \cdot f_i) - \frac{(\sum_{i=1}^n y_i \cdot f_i)^2}{\sum_{i=1}^n f_i}}{\sum_{i=1}^n f_i}} = \sqrt{\frac{\sum_{i=1}^n y_i^2 \cdot f_i}{\sum_{i=1}^n f_i} - M^2}$$

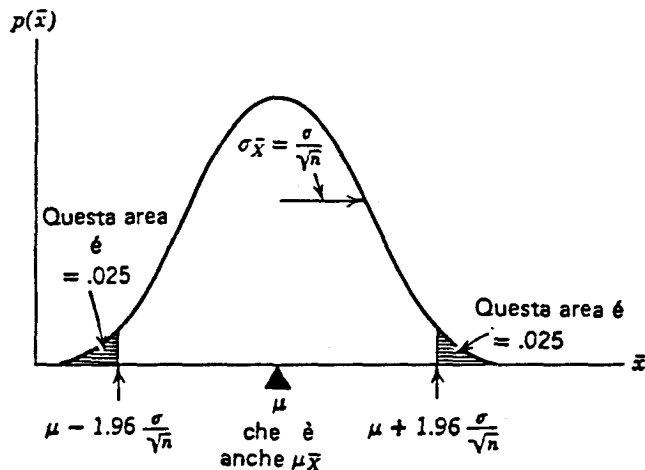
$$= \sqrt{\frac{860874996 - \frac{261000^2}{80}}{80}} = \sqrt{117031.2} = 342.082$$



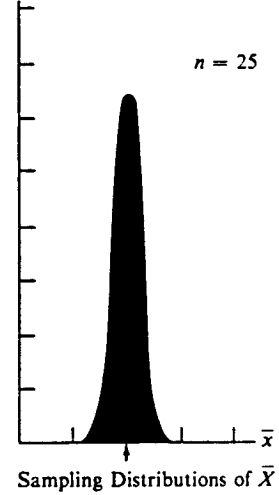
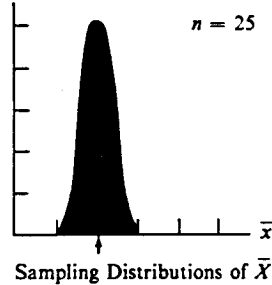
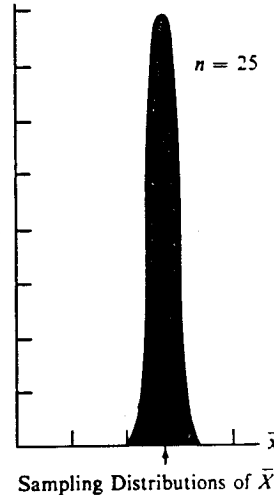
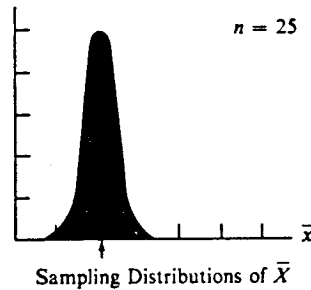
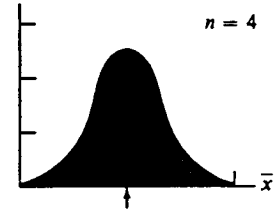
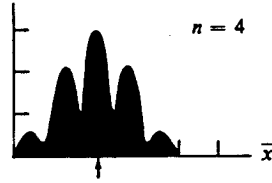
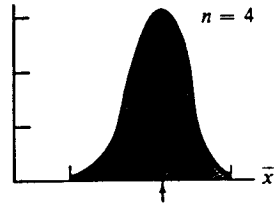
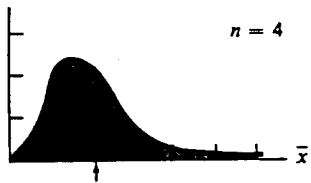
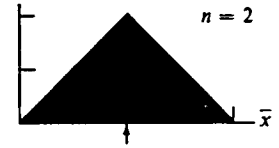
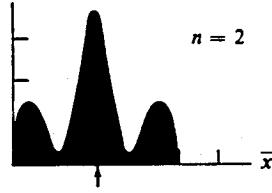
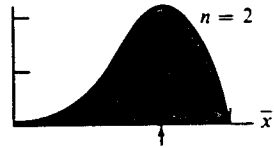
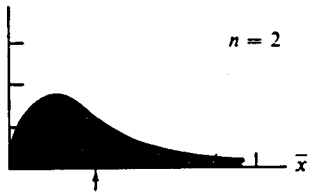
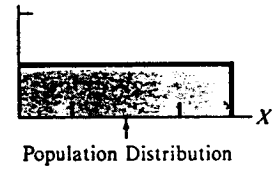
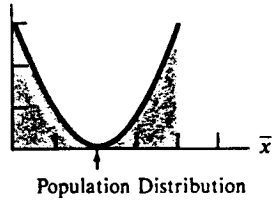
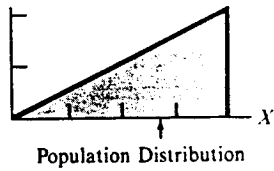
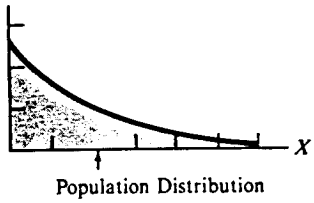


$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

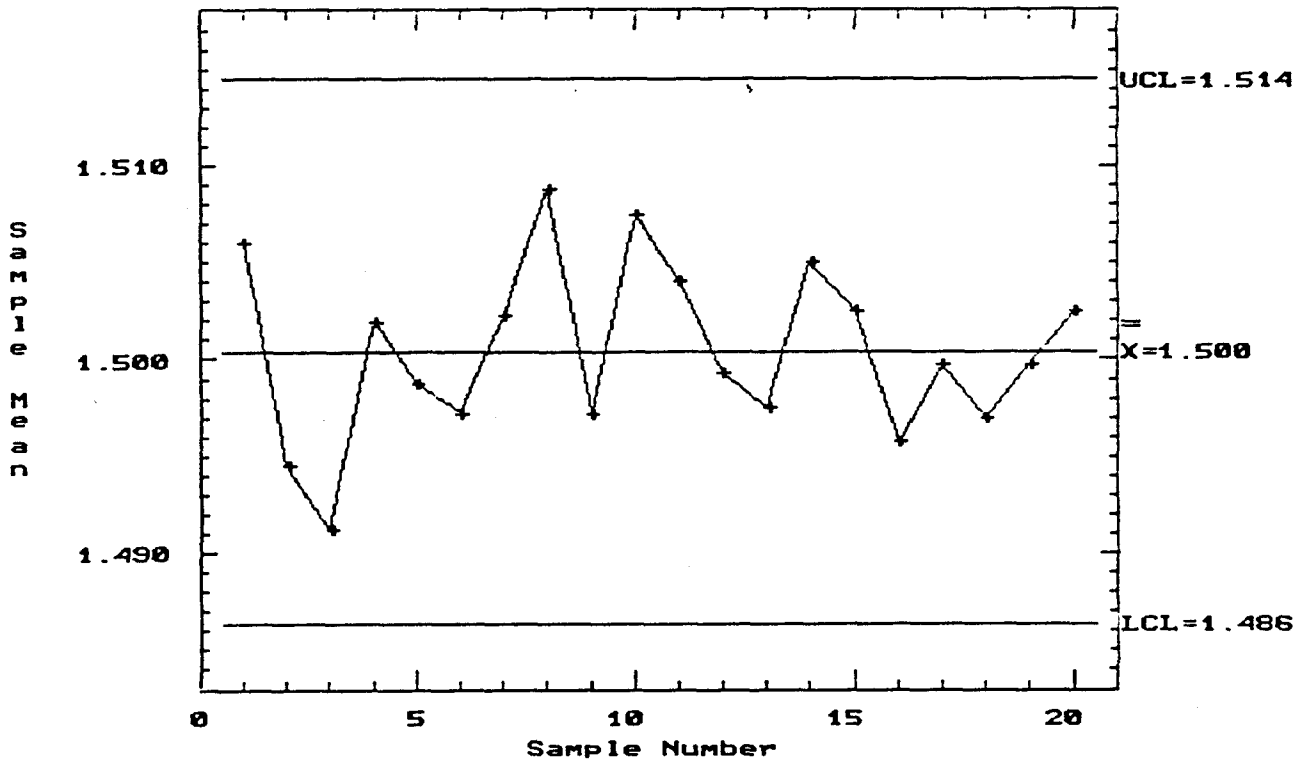
errore standard della media  $ES(\bar{x})$



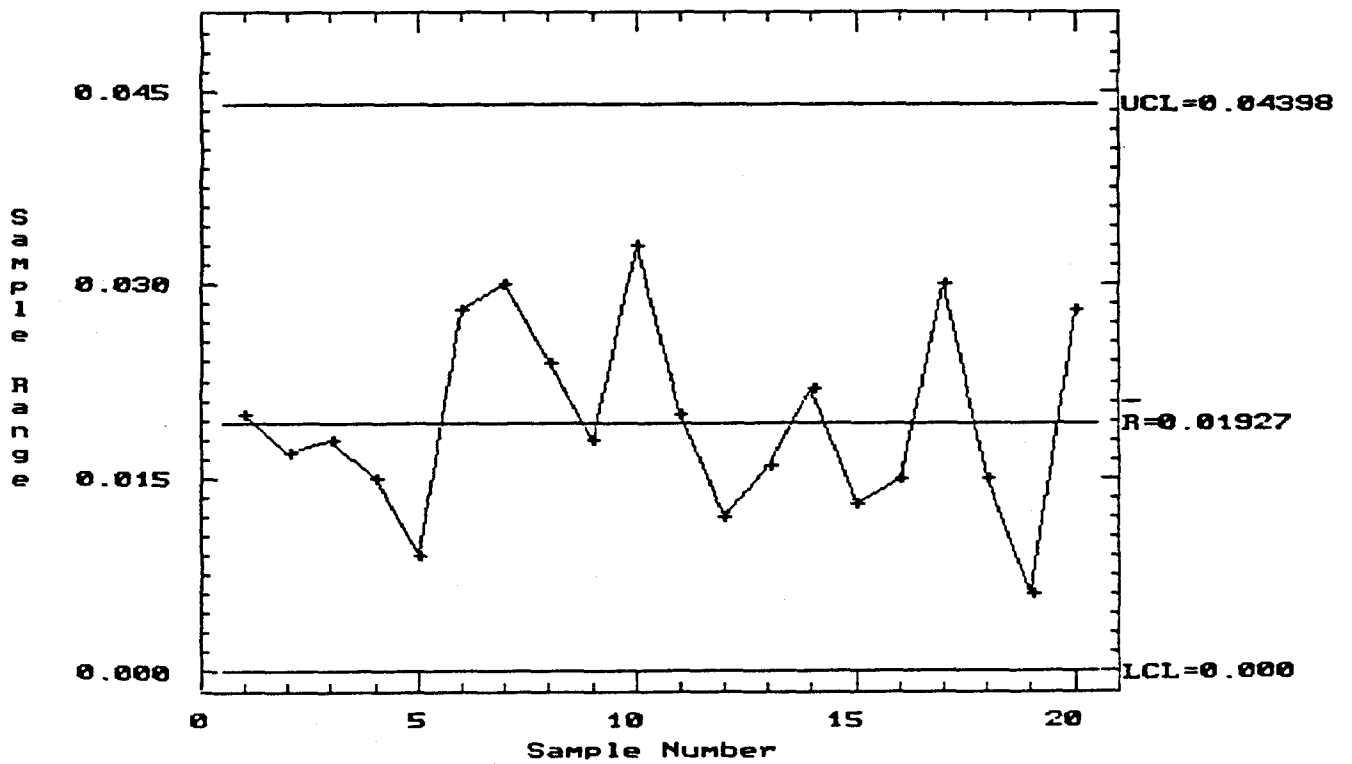
Distribuzione di una media campionaria  $\bar{X} \sim N [\mu, (\sigma^2/n)]$ . (Si noti che  $\mu$  è una costante ignota; ma supponiamo che qualunque sia il suo valore la variabile  $\bar{X}$  si distribuisce attorno ad essa come nel grafico).



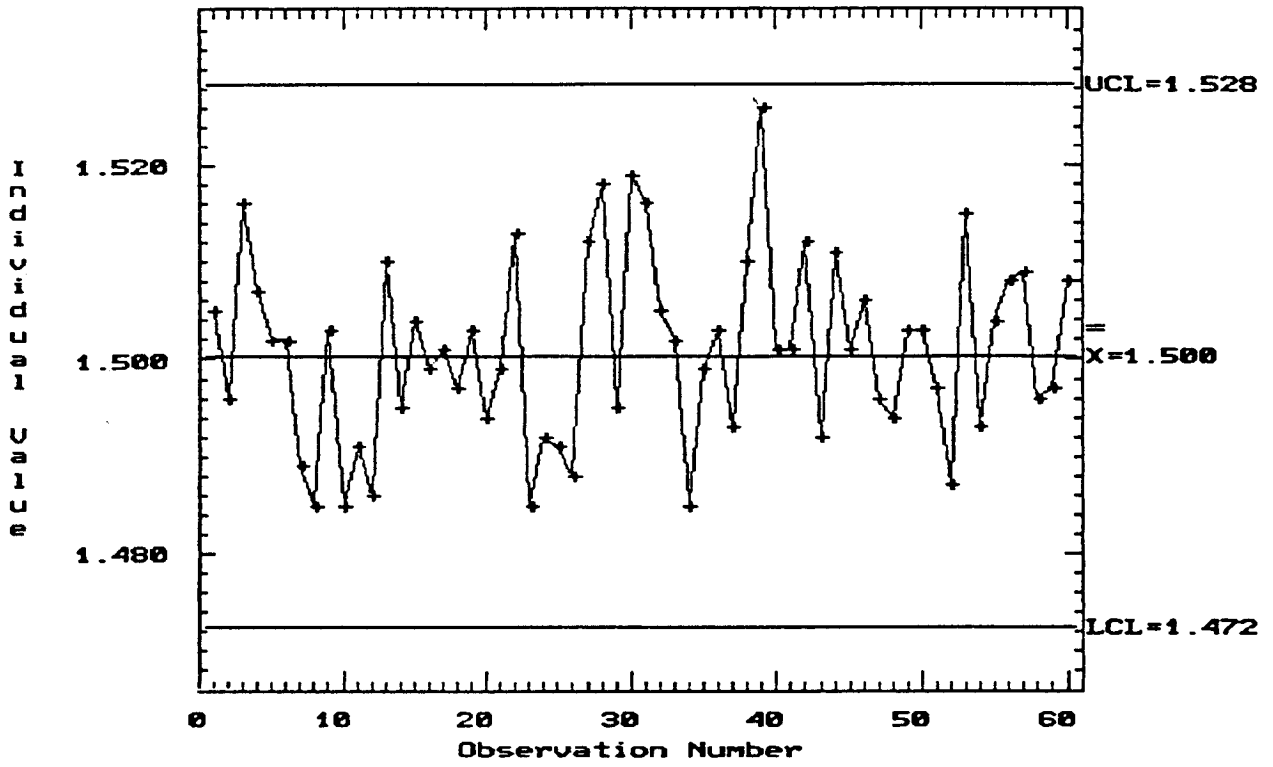
X-bar Chart for C6



R Chart for C6



I Chart for C6



S Chart for C6

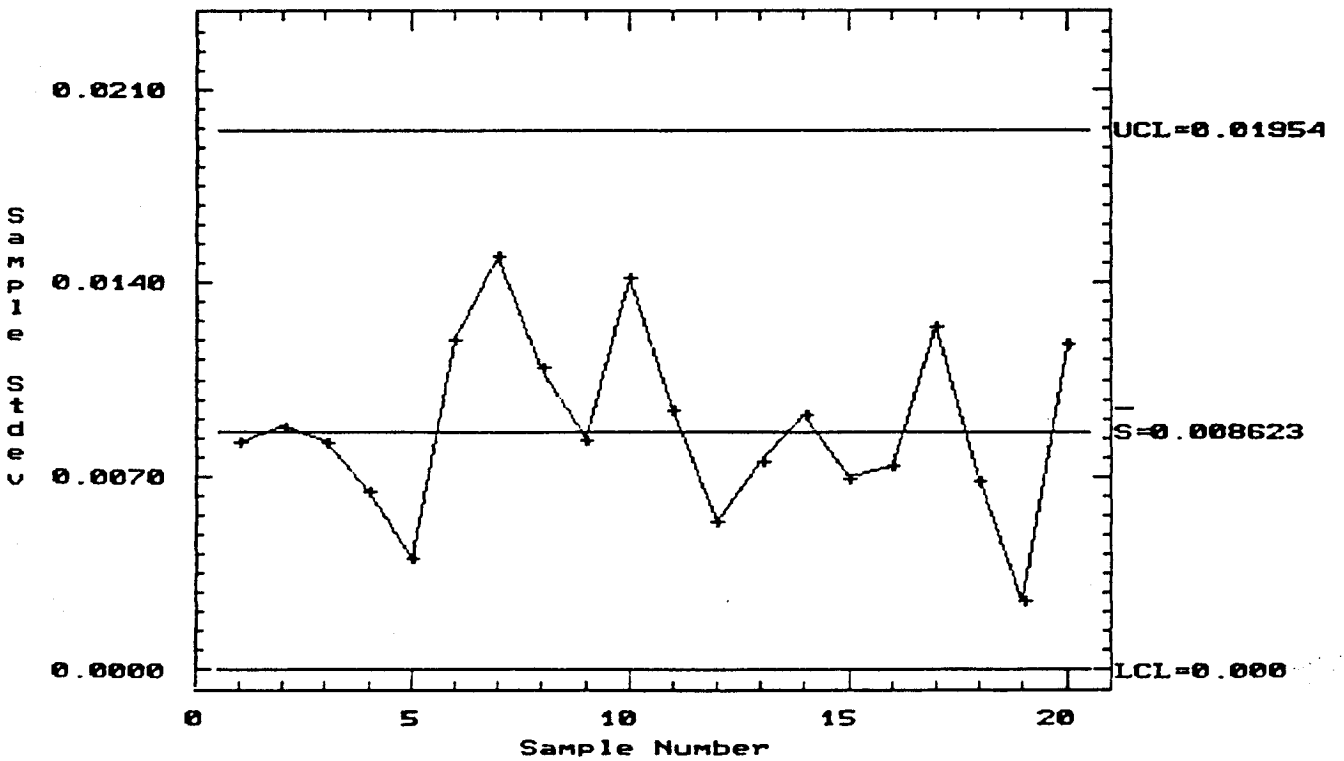


Tabella A1. Aree in una coda della curva normale standardizzata

Questa tabella riporta l'area tratteggiata

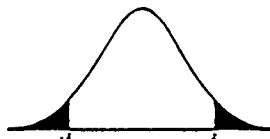


$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	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	0.048	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.037	
1.8	0.036	0.035	0.034	0.034	0.033	0.032	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									

Adattata da Croxton [25].

Tabella A2. Aree nelle due code della curva normale standardizzata

Questa tabella riporta le aree tratteggiate

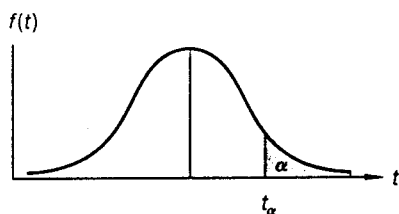


<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0.1	0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0.2	0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0.3	0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0.4	0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0.5	0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0.6	0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0.7	0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0.8	0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0.9	0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1.0	0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1.1	0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1.2	0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1.3	0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1.4	0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1.5	0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1.6	0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1.7	0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1.8	0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1.9	0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2.0	0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2.1	0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2.2	0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2.3	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.17	0.17
2.4	0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2.5	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2.6	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2.7	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2.8	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2.9	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3.0	0.003									

**Tabella A3.** Percentili della distribuzione  $t$  (questa tabella dà i valori di  $t$  che, per diversi g.l., staccano una specificata proporzione in una o nelle due code della distribuzione  $t$ )

g.l.	Area nelle due code				
	0.10	0.05	0.02	0.01	0.001
	Area in una coda				
	0.05	0.025	0.01	0.005	0.0005
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
$\infty$	1.645	1.960	2.326	2.576	3.291

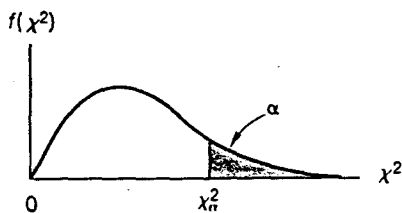
TABLE 6 Critical Values for Student's  $t$



$\nu$	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Source: This table is reproduced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.), *The Biometrika Tables for Statisticians*, Vol. 1, 3d ed., Biometrika, 1966.

TABLE 7 Critical Values of  $\chi^2$



DEGREES OF FREEDOM	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$
1	.0000393	.0001571	.0009821	.0039321	.0157908
2	.0100251	.0201007	.0506356	.102587	.210720
3	.0717212	.114832	.215795	.351846	.584375
4	.206990	.297110	.484419	.710721	1.063623
5	.411740	.554300	.831211	1.145476	1.61031
6	.675727	.872085	1.237347	1.63539	2.20413
7	.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

Source: From C. M. Thompson, "Tables of the Percentage Points of the  $\chi^2$ -Distribution," *Biometrika*, 1941, 32, 188-189. Reproduced by permission of the *Biometrika* Trustees.

DEGREES OF FREEDOM	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169