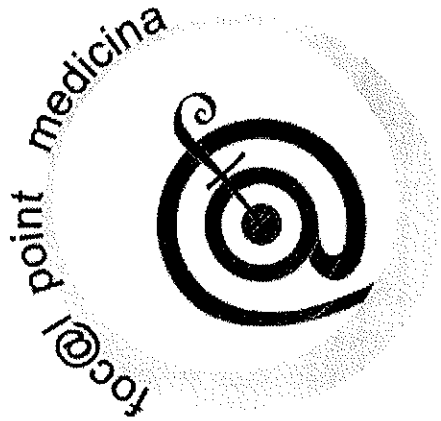
UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

STATISTICA MEDICA

(Prof. CARLA ROSSI)

IV Lezione



Intervalli di confidenza

Applicazioni

Stime di intervallo

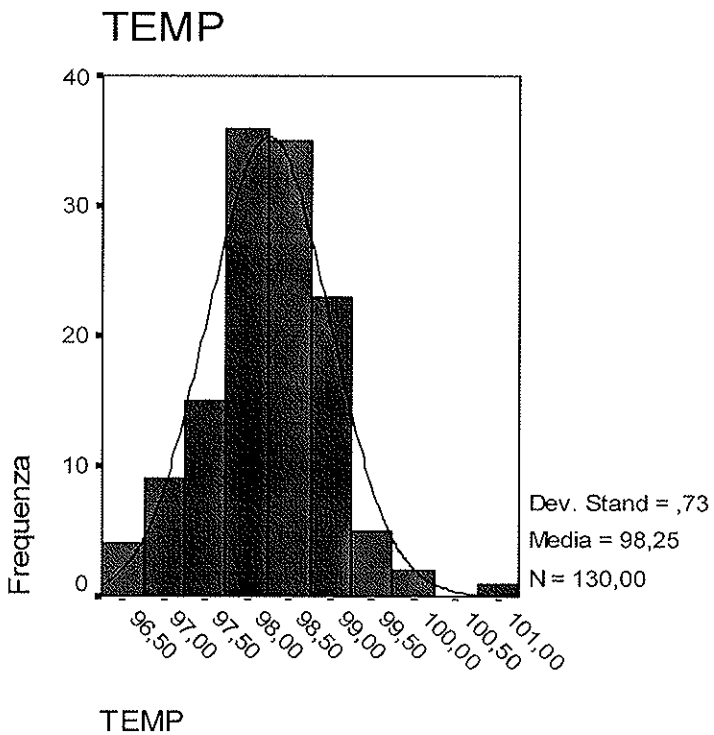
- La distribuzione dello stimatore, permette di stabilire con che "probabilità" il suo valore apparterrà ad un preciso intervallo.
- Si determina un intervallo, attorno alla stima puntuale, che contiene il valore incognito del parametro con un prefissato valore di probabilità.
- Tale valore prende il nome di livello di confidenza e l'intervallo si denota "intervallo di confidenza".

Temperatura corporea e battiti

- Numerosità campionaria: n=130
- Metà maschi e metà femmine
- Numerosità sufficienti per utilizzare l'approssimazione normale

Temp	Sex	Battiti
96,30	1	70
96,70	1	71
96,90	1	74
97,00	1	80
97,10	1	73
97,10	1	75
97,10	1	82
99,10	2	74
99,20	2	77
99,20	2	66
99,30	2	68
99,40	2	77
99,90	2	79
100,00	2	78
100,80	2	77

Approssimazione normale



Standardizzazione

- Se la deviazione standard empirica delle misure è s=0,73, allora la deviazione standard della media è:

$$s(\overline{X}) = \frac{s}{\sqrt{n}} = 0{,}06$$

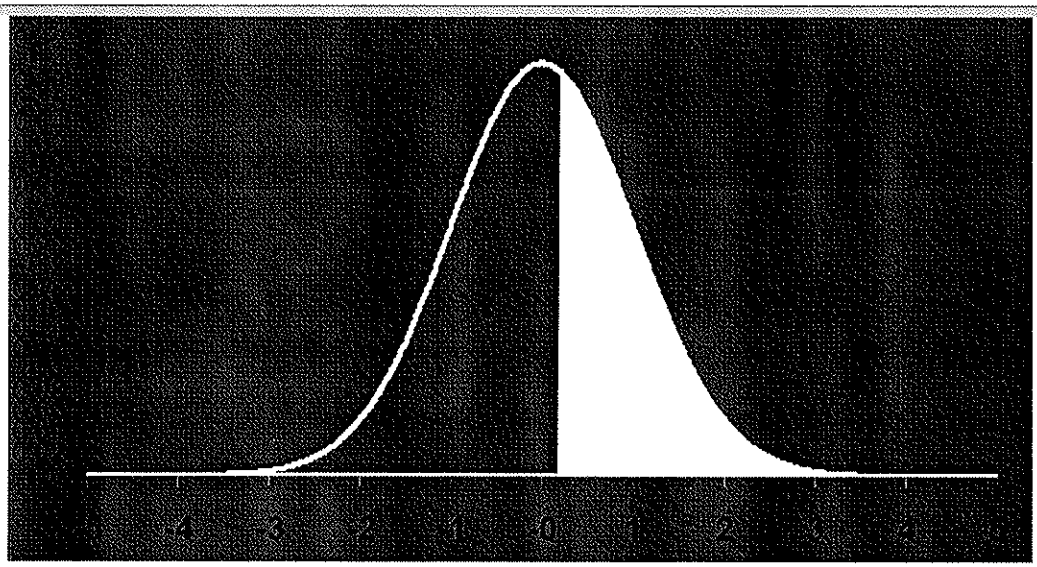
- E' tradizione chiamarla errore standard (errore standard stimato) oppure "s.e" oppure "e.s.e." (estimated standard error).

Calcolo dell'intervallo

- L'intervallo di confidenza al 95% si ricava considerando che, per la variabile standardizzata:
- esattamente il 95% delle osservazioni è tra - 1,96 e + 1,96
- Al variare del campione allora osserveremo circa il 95% delle volte un valore di Z= $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ compreso tra - 1,96 e + 1,96.

Uso delle Tavole

Nelle tavole numeriche delle distribuzioni statistiche sono riportate alcune informazioni sulle funzioni di ripartizione ovvero alcune informazioni sulle probabilità di intervalli particolari o di semirette, come nel caso della normale (da una certa ascissa al $+\infty$).



7

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									
0.2	0.421									
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.5							0.005			

8

Intervallo di confidenza al 95%

- Possiamo scrivere:

$$P(\mu - 1{,}96\sigma/\sqrt{n} \le \bar{X} \le \mu + 1{,}96\sigma/\sqrt{n}) = 0{,}95$$

- Esplicitiamo mettendo al centro la media incognita:

$$P(\bar{X} - 1{,}96\sigma/\sqrt{n} \le \mu \le \bar{X} + 1{,}96\sigma/\sqrt{n}) = 0{,}95$$

- Se non conosciamo σ, lo sostituiamo con s che è la sua stima puntuale.
- L'intervallo è centrato su $\bar{X}$ e ha semiampiezza 1,96s.

Temperatura corporea e battiti

deviazione standard della media=0,06

semiampiezza dell'intervallo attorno alla media=0,06(1,96)=0,13 al livello 95%

N. soggetti	Media	I.C. 95%			
130	98,25	98,12	-	98,38	
	N	Media		I.C. 95%	
Maschi	65	98,1	97,93	-	98,28
Femmine	65	98,39	98,21	-	98,58

Peso alla nascita

Peso alla nascita				
	media	deviazione standard	n	errore standard
madri fumatrici	114,11	18,1	484	0,82
madri non fumatrici	123,05	17,4	742	0,64

Intervalli di confidenza al 95%		
	estremo inferiore	estremo superiore
madri fumatrici	114,11-1,96(0,82)	114,11+1,96(0,82)
madri non fumatrici	123,05-1,96(0,64)	123,05+1,96(0,64)
madri fumatrici	112,5	115,72
madri non fumatrici	121,80	124,3

Interpretazione

- I due intervalli non hanno punti comuni.
- La conclusione è che con probabilità 95% i neonati da madri non fumatrici hanno peso medio superiore ai neonati da madri fumatrici.
- Più precisamente, ripetendo l'esperimento, con probabilità 95% il peso medio dei neonati di madri non fumatrici apparterrà all'intervallo determinato in corrispondenza e analogamente per le fumatrici.

Calcium and Blood Pressure

- Does increasing calcium intake reduce blood pressure? Observational studies suggest that there is a link, and that it is strongest in African-American men.
- Twenty-one African-American men participated in an experiment to test this hypothesis. Ten of the men took a calcium supplement for 12 weeks while the remaining 11 men received a placebo. Researchers measured the blood pressure of each subject before and after the 12-week period. The experiment was double-blind.

Dati (appaiati)

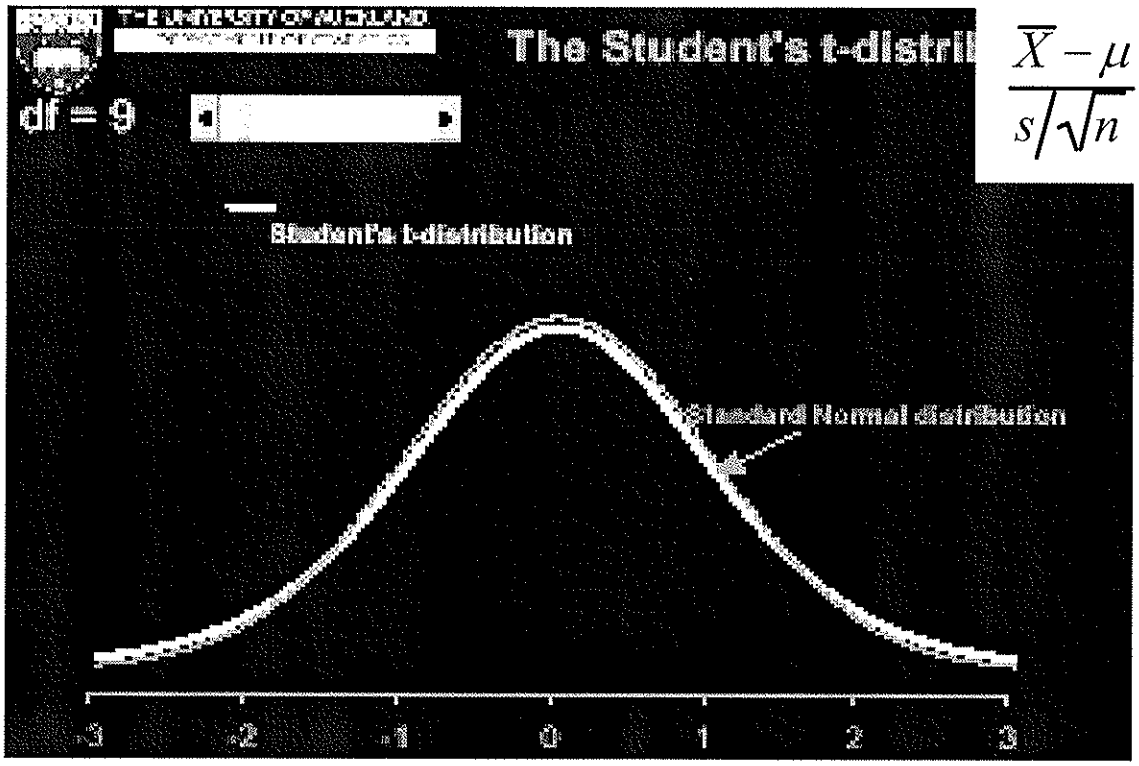
Consideriamo la misura di efficacia data dalla differenza (prima e dopo il trattamento) e determiniamo la media per i due gruppi di soggetti e l'intervallo di confidenza al 95%.

Treatment	Begin	End	Decrease
Calcium	107	100	7
Calcium	110	114	-4
Calcium	123	105	18
Calcium	129	112	17
Calcium	112	115	-3
Calcium	111	116	-5
Calcium	107	106	1
Calcium	112	102	10
Calcium	136	125	11
Calcium	102	104	-2
Placebo	123	124	-1
Placebo	109	97	12
Placebo	112	113	-1
Placebo	102	105	-3
Placebo	98	95	3
Placebo	114	119	-5
Placebo	119	114	5
Placebo	112	114	2
Placebo	110	121	-11
Placebo	117	118	-1
Placebo	130	133	-3

Il modello adatto

- In questo caso i dati sono troppo pochi per utilizzare l'approssimazione normale.
- La distribuzione che si utilizza è il modello t di Student, che è una famiglia di distribuzioni dipendenti dal numero di gradi di libertà dello stimatore della deviazione standard.
- E' possibile determinare il moltiplicatore della deviazione standard per determinare la semiampiezza dell'intervallo centrato sulla media utilizzando tabelle numeriche della t standardizzata.

Il gruppo placebo



$$\frac{\overline{X}-\mu}{s/\sqrt{n}}$$

Uso delle tavole (t standardizzata)

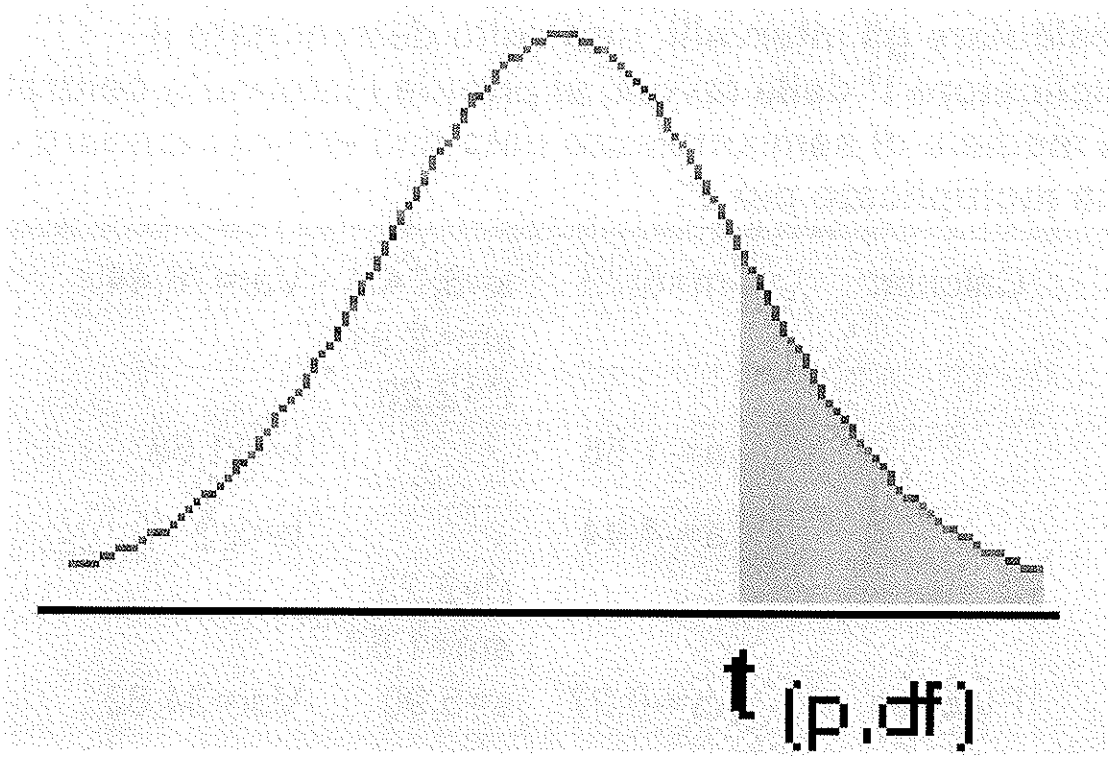


Tavola della distribuzione t di Student

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	**2.26216**	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	**2.22814**	2.76377	3.16927	4.5869

Calcolo degli intervalli

Per calcolare gli intervalli, una volta determinato il moltiplicatore t* dalla tavola, si prende l'intervallo centrato sulla media e di semiampiezza data dall'errore standard della media moltiplicato per t*.

Calcium	m=5	s=8,74	Var=76,44
		sm=2,91	
		t*=2,26	
	a=-1,58		b=11,58
	a=-4,44		b=3,9
		t*=2,23	
		sm=1,87	
Placebo	m=-0,27	s=5,90	Var=34,82

Interpretazione

- Gli intervalli determinati hanno ampia sovrapposizione, non è possibile escludere che le differenze osservate nei risultati dei due trattamenti siano del tutto casuali.
- Occorrono comunque approfondimenti.

Alternativamente...

- Consideriamo le stime ottenute per le probabilità di successo nei due gruppi e determiniamo gli intervalli di confidenza utilizzando l'approssimazione normale.

	p	1-p	s	e.s	CV della media
calcio	0,6	0,4	2,4	0,76	126
placebo	0,3	0,7	2,31	0,70	232

	a	b	a	b
calcio	-0,89	2,09	0	1
placebo	-1,07	1,67	0	1
	intervallo non ammissibile		intervallo ammissibile	

Interpretazione

- In questo modo gli intervalli ammissibili determinati hanno totale sovrapposizione, le differenze osservate nei risultati dei due trattamenti appaiono del tutto casuali.
- Il risultato ottenuto dipende principalmente dalle numerosità campionarie troppo esigue che lasciano grande variabilità prevista, come appare dai coefficienti di variazione delle medie.

Se si aumentano le numerosità campionarie?

- Supponiamo di ottenere le stesse stime per le due proporzioni ma con campioni di 100 soggetti per ogni trattamento. Come variano gli intervalli?

numerosità campionarie n=100 per entrambi i trattamenti

	p	1-p	s	e.s	CV della media
calcio	0,6	0,4	2,4	0,24	40
placebo	0,3	0,7	2,31	0,23	77

	a	b	
calcio	0,13	1,07	intervallo non ammissibile
placebo	-0,2	0,75	intervallo non ammissibile

Un esperimento di Darwin

- Mi è spesso capitato di pensare che sarebbe stato consigliabile appurare se pianticelle ottenute da semi provenienti da impollinazione incrociata dei fiori fossero in qualche modo superiori a quelle derivanti da auto-impollinazione.

continuazione

- *Ma, dato che generalmente non si conoscono per gli animali casi di danno che diventino evidenti in una singola generazione ottenuta da accoppiamenti tra parenti prossimi come fratello e sorella, ho ritenuto che la stessa regola potesse applicarsi alle piante e che sarebbe stato necessario, a costo di impiegare molto tempo, considerare diverse generazioni successive di piante ottenute con auto-impollinazione e impollinazione incrociata per ottenere qualche risultato....*

Piano sperimentale

- *Pertanto decisi di iniziare una lunga serie di esperimenti con varie piante che continuarono per 11 anni...*
- *Gli esperimenti si svolgevano nel modo seguente. Una singola pianta nel caso producesse abbastanza fiori, o due o tre piante, erano poste sotto una rete abbastanza grande da coprire completamente le piante senza toccarle.*

Gruppo sperimentale e gruppo controllo

- *Questo punto è molto importante perché se i fiori toccano la rete possono essere fertilizzati in modo incrociato da insetti…. Sulle piante protette venivano marcati più fiori e fertilizzati con il loro stesso polline. Un uguale numero delle stesse piante, marcate in modo diverso, veniva allo stesso tempo fertilizzato in modo incrociato con polline di altre piante…..*

Le unità statistiche

- *In seguito i semi provenienti da autoimpollinazione e da impollinazione incrociata venivano posti in recipienti separati….*
- *Nel confrontare i due insiemi di piante ottenute…l'altezza di ogni pianta veniva misurata accuratamente dalle due parti. Spesso più di una volta….*

Descrizione dei risultati

- *..Le 15 piante ottenute da impollinazione incrociata mostravano un'altezza media di 20.19 pollici mentre le 15 ottenute da auto-impollinazione avevano un'altezza media di 17.57 pollici.... Mr. Galton rappresentò graficamente i risultati....*

I dati

Codice di identificazione della pianta madre.	Altezza in pollici delle piante figlie ottenute da impollinazione incrociata	Altezza in pollici delle piante figlie ottenute da auto-impollinazione
1	23.5	17.375
1	12	20.375
1	21	20
2	22	20
2	19.125	18.375
2	21.5	18.625
3	22.125	18.625
3	20.375	15.25
3	18.25	16.5
3	21.625	18
3	23.25	16.25
4	21	18
4	22.125	12.75
4	23	15.5
4	12	18

Calcolo degli intervalli

20,19	17,54	medie
3,62	2,04	dev.st
0,93	0,53	dev.st.medie

moltiplicatore t*=2,14

a=20,19-1,99=18,02
b=20,19+1,99=22,18

a=17,54-1,13=16,41
b=17,54+1,13=18,67

Interpretazione

- Gli intervalli sono poco sovrapposti, occorrono comunque ulteriori approfondimenti o ulteriori dati per arrivare a un'interpretazione affidabile.

Problemi e metodi di stima

Dati scambiabili

- L'impostazione più usuale si basa su modelli di osservazione che producono ***dati scambiabili***, ovvero *su misure o osservazioni che si possa supporre siano avvenute sempre nella stessa situazione sperimentale o osservazionale e indipendentemente una dall'altra.*

Dati scambiabili

- Si dicono ***scambiabili*** *perché, nella situazione considerata, l'ordine con cui si presentano le osservazioni non fornisce alcuna informazione per il problema di interesse: ogni permutazione dell'n-pla dei dati osservati è equivalente ai fini della soluzione del problema inferenziale.*

Schema del campionamento ripetuto

- I diversi metodi di soluzione del problema di stima possono essere valutati attraverso le proprietà che possiedono quando si pensi di ripetere la procedura di osservazione allo stesso modo e indipendentemente un numero rilevante di volte.
- Tale schema teorico è lo ***schema del campionamento ripetuto***.

Il metodo di stima dei momenti

- I problemi di stima dei parametri vengono risolti uguagliando i momenti teorici, espressi in funzione del parametro incognito, agli analoghi momenti statistici (o empirici), calcolati numericamente sulla base dei dati disponibili.
- Per stimare si calcola sul campione quella funzione statistica che piu' "direttamente" corrisponde al significato del parametro nella distribuzione, con l'idea che la media del campione ($\overline{X}$) corrisponda al valore atteso della distribuzione.

Stima della varianza

- La varianza è il valore atteso della deviazione di un'osservazione dal valore atteso al quadrato e si stima calcolando la media delle deviazioni osservate dalla media campionaria al quadrato, tenendo conto dei gradi di libertà:
-

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Stima di una proporzione

- Se si vuole stimare una proporzione o percentuale incognita θ, si utilizza il risultato dato dalla frequenza osservata f_n su un campione rappresentativo di numerosità n e si pone la stima θ^* di θ pari a :

$$\theta^* = f_n.$$

Stimatori e Stime

- La regola (formula) che definisce il metodo con cui effettuare la stima si dice **Stimatore.**
- Il risultato numerico dell'applicazione della regola ai dati campionari osservati si dice **Stima.**
- Gli stimatori sono variabili statistiche e vengono denotati con lettere maiuscole.
- Le stime sono valori numerici e vengono denotati con le lettere minuscole.

Il metodo di stima dei momenti

- Nel caso della frequenza lo stimatore risulta allora dato da:

$$\Theta(X)=f_n(X)=X/n$$

dove X è il numero di "successi" e la stima relativa da

$$\theta(x)=f_n(x)=x/n.$$

- Con lo stesso approccio possono risolversi altri problemi di stima. Ma come ***valutare le proprietà degli stimatori ottenuti?***

Proprietà

- Essendo delle variabili aleatorie, gli stimatori sono caratterizzati dalla loro distribuzione al variare del campione (distribuzione campionaria).
- Per valutare la bontà di uno stimatore occorrerà studiarne la distribuzione.
- In particolare si studieranno gli indici di posizione e di dispersione

Le proprietà degli stimatori

- Prendiamo in considerazione la media e lo scarto standard nel caso della stima di una proporzione.

- $\Theta(X)=f_n(X)=X/n$

si può calcolarne la media:

$E(\Theta(X))=E(X)/n=n\,\Theta\,/n=\Theta$

E la deviazione standard:

$\sigma(\Theta(X))=\sqrt{[\sigma^2(\Theta(X)]}=\sqrt{[\Theta(1-\Theta)/n]}=\sigma\,/\sqrt{n}$

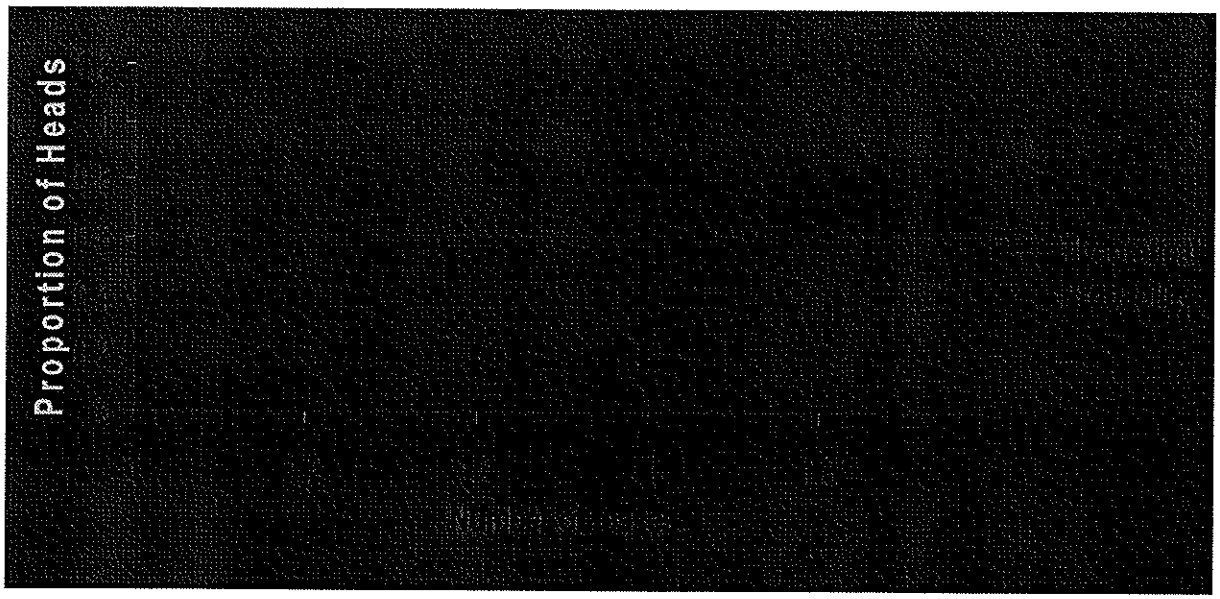
Le proprietà degli stimatori

1. *la regola di stima utilizzata (stimatore) non introduce mediamente distorsioni sistematiche nella stima (non tende a sottostimare, né a sovrastimare il parametro incognito);*
2. *se si aumenta la numerosità campionaria, l'informazione fornita dallo stimatore è più precisa in quanto diminuisce la variabilità (la dispersione) della sua distribuzione.*

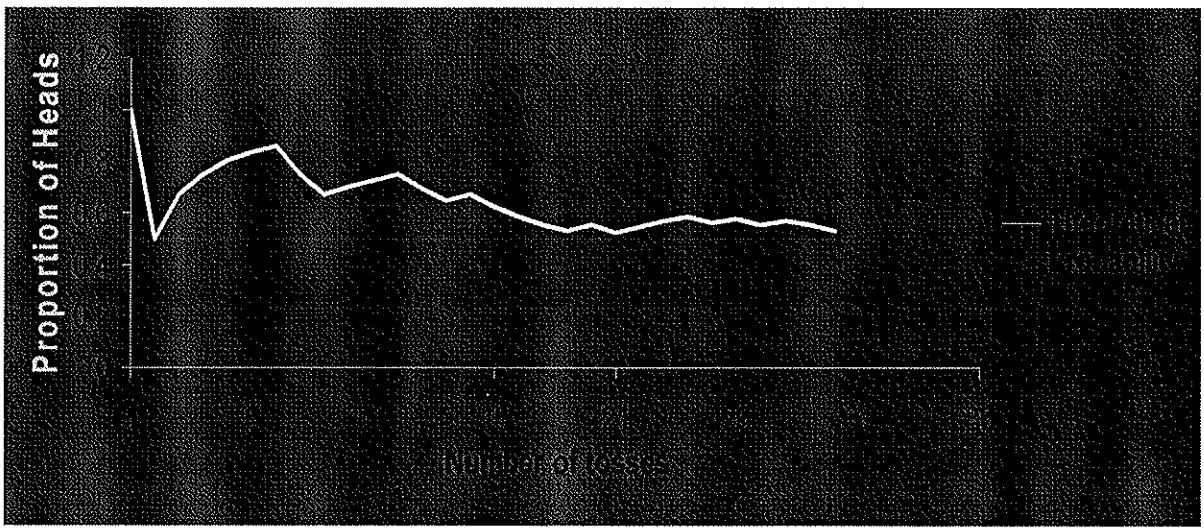
Simulazione: stima di una proporzione

- Supponiamo di voler stimare la probabilità di "Testa" nel lancio di una moneta (sappiamo che la probabilità vera è p=0,5).
- Lanciamo tante volete una moneta e registriamo la frequenza relativa di T al crescere del numero di prove.
- Grafichiamo gli andamenti e verifichiamo che al crescere del numero di prove la frequenza relativa si allontana sempre meno dal valore 0,5.

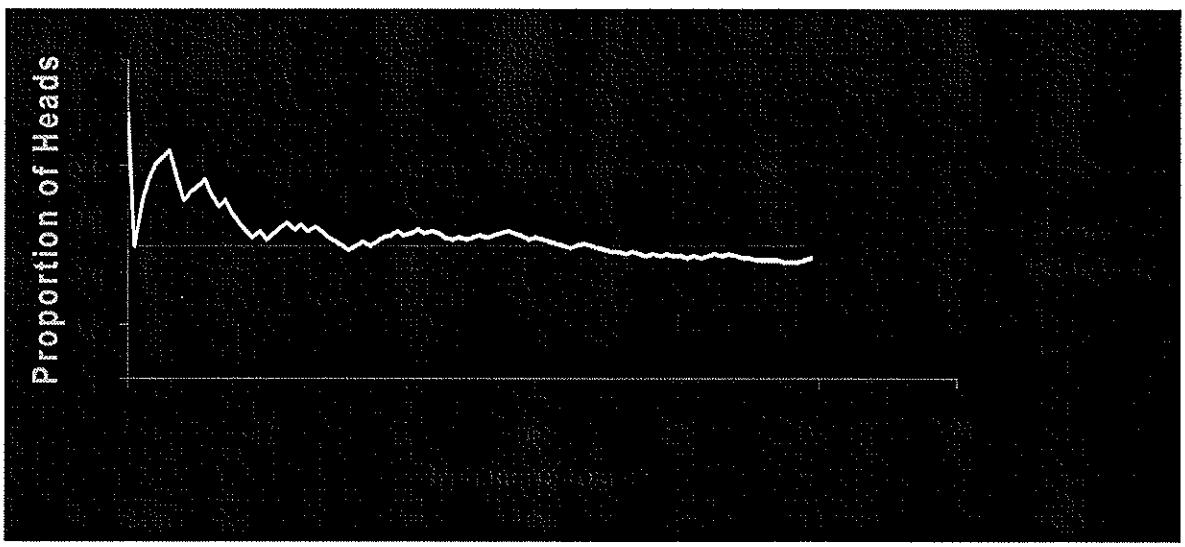
Verifica: Testa e Croce



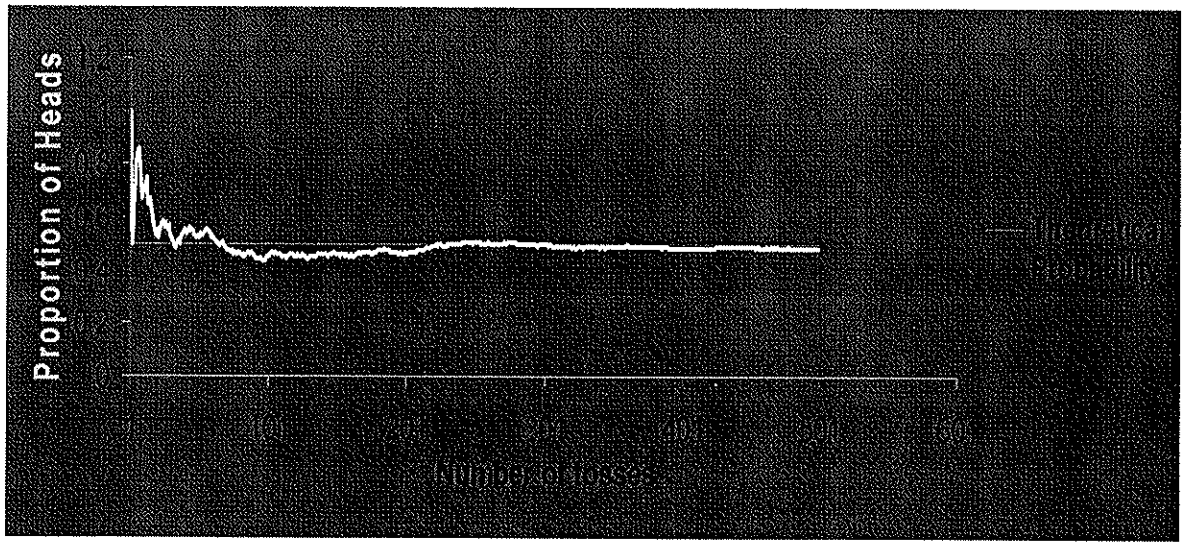
30 lanci



100 lanci



500 lanci



Le proprietà degli stimatori

- In genere si richiede ad uno stimatore di soddisfare le due proprietà: la ***proprietà di correttezza o di non distorsione***:

1. *per qualsiasi valore di n, il valore centrale (media) della distribuzione dello stimatore coincide con il parametro da stimare, ovvero la distribuzione dello stimatore è centrata attorno al parametro da stimare;*

e la ***proprietà di consistenza***:

2. *lo stimatore $\Theta(X)$ tende al crescere di n a concentrarsi intorno a θ.*

Stima puntuale

- Consideriamo un campione casuale semplice di n misure ***con media e varianza incognite.***
- Utilizziamo il ***metodo dei momenti*** e *uguagliamo le incognite ai corrispondenti momenti statistici calcolati sul campione,* otteniamo:
- Per la media: la media empirica.
- Per la varianza: la varianza empirica (basata sui gradi di libertà).

Stime di intervallo

- La stima del parametro data dal singolo valore dello stimatore in corrispondenza ai dati osservati si dice: *stima puntuale del parametro.*
- Tutte le informazioni rilevanti devono essere fornite al momento di comunicare il risultato di un processo di stima e il risultato costituito dalla sola stima puntuale del parametro non è completo.

Stime di intervallo

- Un modo completo di fornire le informazioni rilevanti sul processo è di fornire: il modello statistico utilizzato, il risultato della stima puntuale, la corrispondente stima dello scarto standard dello stimatore e la numerosità campionaria.

Stime di intervallo

- Generalmente le informazioni sul modello e sulla numerosità campionaria vengono riportate nel momento in cui si descrive l'esperimento o l'osservazione, mentre i risultati relativi al processo di stima vero e proprio, stima e scarto standard dello stimatore, vengono riportati nella forma:

stima±scarto standard.

Il caso binomiale

- Se, sulla base di un modello "binomiale" e di 100 unità statistiche, si è ottenuto un valore stimato di θ: $\theta^*=0,4$ (lo scarto standard della stima è 0,024), si riporta il risultato nel modo seguente:
  - $0,4 \pm 0,024$.

Stime di intervallo

- Questo modo di fornire l'informazione prevede, pertanto, *un risultato finale in forma di intervallo di cui la stima puntuale costituisce il punto medio.*
- L'approccio al problema di stima è del tipo ***stima di intervallo***.

Statistica Inferenziale

La Statistica Inferenziale

- Quando si esaminano dei dati, ci si può chiedere se questi dati, anche se interessanti, rappresentano solo se stessi o se contengono un'informazione rilevante anche per altri fenomeni, spesso più interessanti perché più generali.

Dal particolare al generale

- Prima di un'elezione si fa un sondaggio di opinione.
- Benché in Italia ci siano almeno 40 milioni di cittadini con diritto di voto, un sondaggio viene di solito eseguito interrogando un migliaio di persone.
- Le risposte di questi individui ci dicono qualcosa sulle tendenze in tutta la popolazione?

Dal particolare al generale

- Sperimentando un nuovo farmaco, si individuano 100 ammalati che si dividono in 2 gruppi di 50.
- Un gruppo riceve la nuova medicina, l'altro un trattamento già in uso.
- Nel primo gruppo, guarisce il 60% dei pazienti, nell'altro gruppo solo il 40%.
- Basta questo risultato per convincersi che la nuova medicina avrà un effetto benefico anche su pazienti futuri?

Modellizzazione

- La statistica inferenziale per affrontare questi problemi costruisce un modello per le osservazioni, concentrando, se possibile, il dubbio che si vuole risolvere (la tendenza di voto di tutta la popolazione, l'effetto "vero" del farmaco) nel valore di un singolo parametro che rimane indefinito.

La Statistica Inferenziale

- I **modelli** permettono di generalizzare, in modo induttivo, i risultati dall'insieme dei dati osservati (**campione**) alla **popolazione** di riferimento.
- Tali modelli permettono anche, in molti casi, di valutare, **probabilisticamente**, eventuali margini di **incertezza** nella generalizzazione dei risultati.

Sondaggio pre-elettorale

- Supponiamo, per semplicità, che il voto abbia due esiti possibili e che nella popolazione intera una percentuale P_A intenda votare per l'alternativa A e, di conseguenza, il resto per l'altra alternativa. Interessa dunque sapere qualcosa sul valore di P_A.
- Scegliamo un campione adatto per il nostro sondaggio.

Rilevazioni campionarie

- Quando la rilevazione dei dati è parziale, si tratta cioè di indagine campionaria, particolare attenzione va posta alla metodologia di scelta delle unità da inserire nel campione, cioè agli ***schemi di campionamento***, al fine di ottimizzarne la rappresentatività e di poter valutare l'estendibilità dei risultati alla popolazione di origine (inferenza).

Campione casuale semplice

- In questo tipo di campione le differenze tra campione e popolazione, e tra diversi campioni analogamente scelti, sono dovute solo al caso, sono cioè frutto di variabilità accidentale e non dipendono da cause sistematiche.
- L'estrazione del campione è casuale semplice quando tutti gli elementi della popolazione hanno la stessa probabilità di essere estratti.

Il campione del sondaggio

- Scegliamo un campione di poche persone rispetto a tutta la popolazione, ma se scegliamo a caso, è ragionevole ritenere che ogni soggetto campionato abbia probabilita' P_A di appartenere al gruppo che intende votare *A* e che le nostre osservazioni siano indipendenti.
- Denotiamo con X il numero di persone intervistate (da un totale di n) che rispondono *A*. X è la quantità che possiamo osservare.

Come si modella X?

- X rappresenta il numero di risposte positive su n, dove ogni risposta è indipendente dalle altre e la probabilità di risposta affermativa (voto per A) è P_A.
- Si tratta di osservazioni modellabili con una distribuzione binomiale con parametri n (numero delle prove) e P_A (probabilità di successo in ogni singola prova).

Il modello

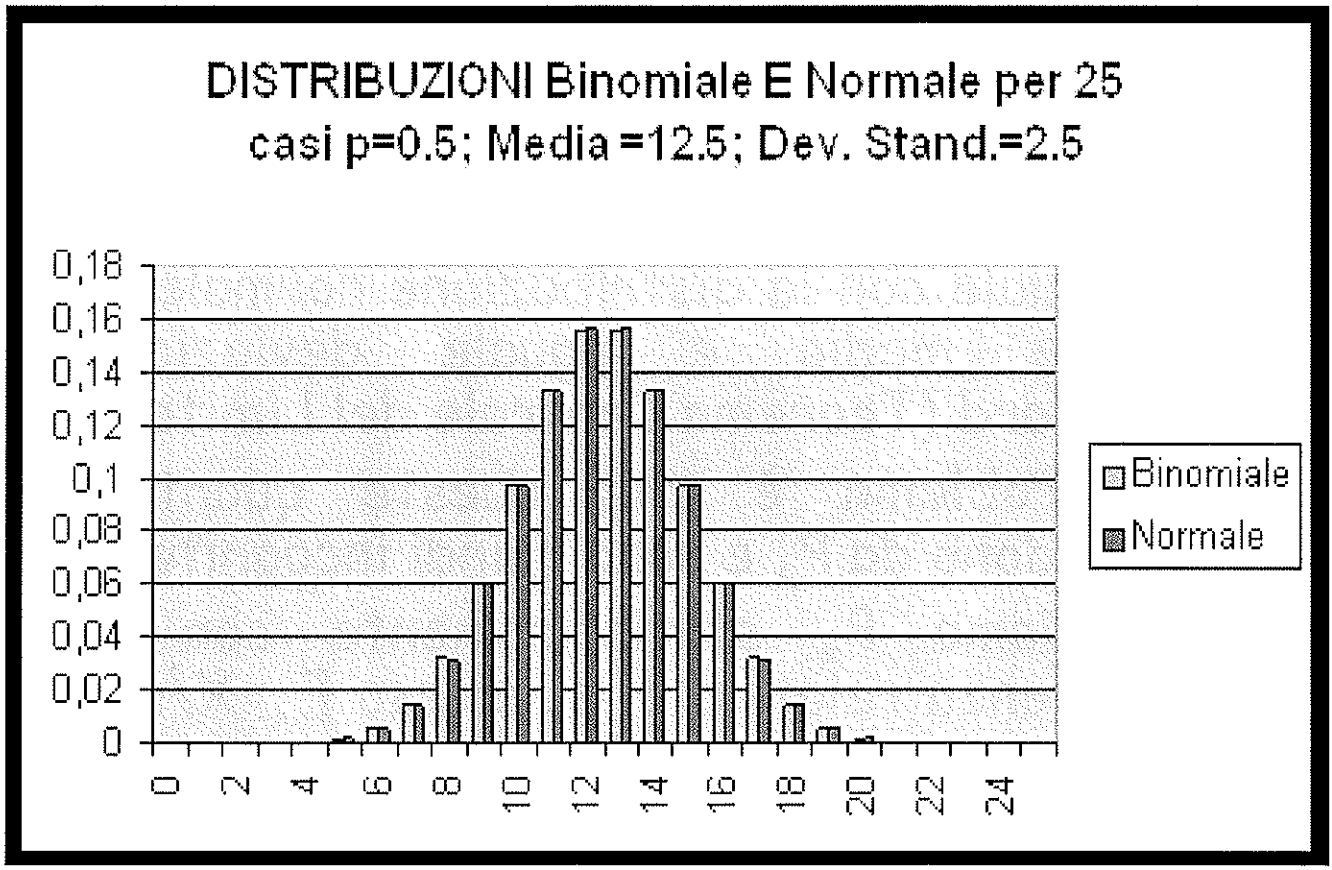
- Possiamo scrivere:
  - $P(X = k) = \binom{n}{k} P_A{}^k (1- P_A)^{n-k}$
- Per questa distribuzione binomiale si ha:
  - $E(X)=n P_A$
  - $\sigma(X)=\sqrt{[n P_A(1- P_A)]}$

Approssimazione normale

- Di solito n è abbastanza grande per poter approssimare la distribuzione binomiale con la distribuzione normale di uguale media e uguale deviazione standard (Teorema Centrale del Limite).
- Il problema è che queste quantità dipendono da P_A e sono quindi incognite.
- Valutiamo P_A statisticamente con la frequenza relativa di successo $F_A=X/n$ e di conseguenza gli altri parametri incogniti sostituendo nelle formule.

Verifichiamo

- Una distribuzione normale e binomiale forniscono probabilità simili per un dato campo di valori, man mano che il numero di prove binomiali aumenta. L'approssimazione è migliore per valori di p prossimi a 0,5.
- Il grafico mostra la differenza tra una curva normale e una distribuzione binomiale con p=0,5, n=25 prove, media μ= np = 12,5, deviazione standard $\sigma = \sqrt{npq} = 2{,}5$ e una distribuzione normale con media e deviazione standard uguali.



Calcoliamo (stimiamo)

- Supponiamo di aver intervistato 10000 soggetti e che 5300 abbiano risposto di voler votare per *A*.
- Allora valutiamo:

$$F_A = X/n = 53\%$$

Chiameremo la proporzione osservata nel nostro sondaggio una stima della vera proporzione nella popolazione.

Questa stima e' stata ottenuta usando lo stimatore (formula) X/n, del quale conosciamo la distribuzione.

Calcolo dei parametri della distribuzione

- Considerando i parametri della distribuzione binomiale abbiamo:

$$E(X/n)=n F_A /n= F_A =0,53$$

$$\sigma(X/n)=\sqrt{[n F_A(1- F_A)]}/n$$

$$\sigma(X/n)=\sqrt{[F_A(1- F_A)/n]}= \sqrt{[(0,53)(0,47)]}/100$$

$$\sigma(X/n)=0,005$$

Approssimiamo e prevediamo

- La distribuzione normale da utilizzare sarà allora quella con media μ=0,53 e deviazione standard σ=0,005.
- Dalle proprietà della distribuzione normale sappiamo che, se prendiamo altri campioni della stessa dimensione, cioè ripetiamo l'esperimento nelle stesse condizioni, dobbiamo aspettarci che praticamente tutti i valori siano entro 3 D.S. della media.

Generalizziamo (induzione)

- Quindi tutti i risultati di analoghi sondaggi dovrebbero variare tra 0,53-0,015 e 0,53+0,015 (0,515 e 0,545).
- L'intervallo è abbastanza piccolo.
- L'osservazione del nostro campione si può generalizzare a tutta la popolazione, anche se rimane un intervallo di possibilità.
- Però siamo abbastanza certi che vinca A.
- Questo e' il tipo di conclusioni inferenziali che e' possibile fare in presenza di variabilità...

Il metodo generale: l'inferenza

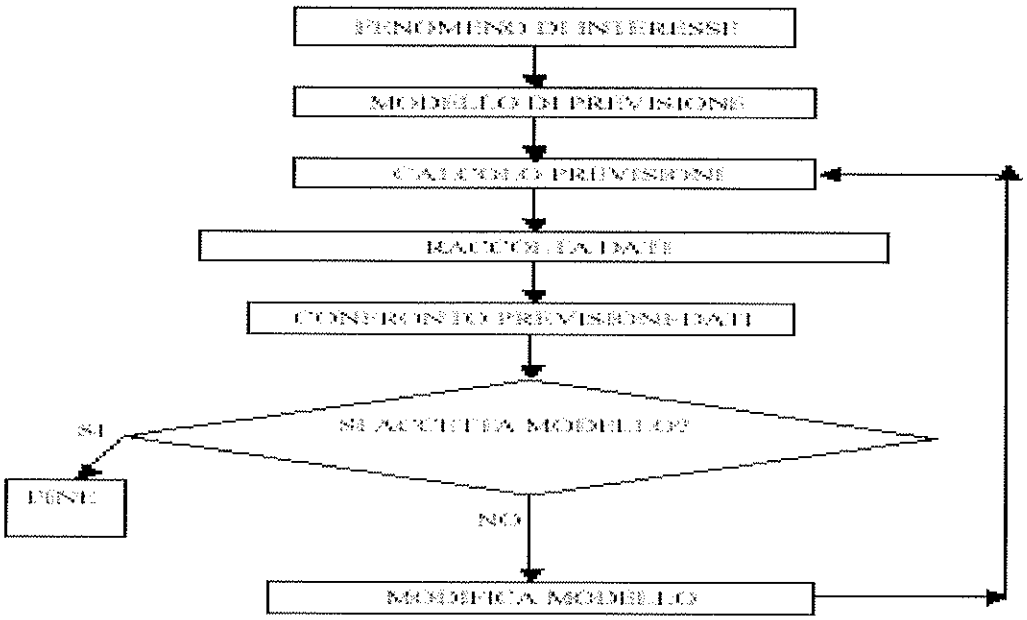
- Si utilizza l'informazione contenuta nelle osservazioni per produrre una "stima" del parametro d'interesse, accompagnata da una misura di "precisione" della stima (basata sulla deviazione standard), con l'eventuale aggiunta di considerazioni sulla possibilità che il parametro abbia certi valori critici per le conclusioni (verifica di ipotesi).
- Nel caso precedente: "vincerà A? (P_A >50%?).

Statistica inferenziale

La ***statistica inferenziale*** consiste nella *modellizzazione del processo di produzione dei dati* (fatti osservabili), e *nell'identificazione del relativo modello* attraverso la *stima* dei coefficienti incogniti, ovvero dei *parametri del modello*, mediante *qualche criterio "ragionevole"*.

Si utilizzano quindi i risultati per assumere decisioni o trarre conclusioni.

Processo inferenziale



Problema inferenziale

- Un ***problema inferenziale*** prevede ***una fase "diretta"*** in cui viene costruito il modello per il processo di produzione dei dati e ***una fase "inversa"*** in cui, dal confronto tra le previsioni prodotte dal modello e i valori effettivamente osservati, sulla base del criterio scelto vengono prodotte le stime dei parametri incogniti o verificata l'adeguatezza del modello.

Problema inferenziale

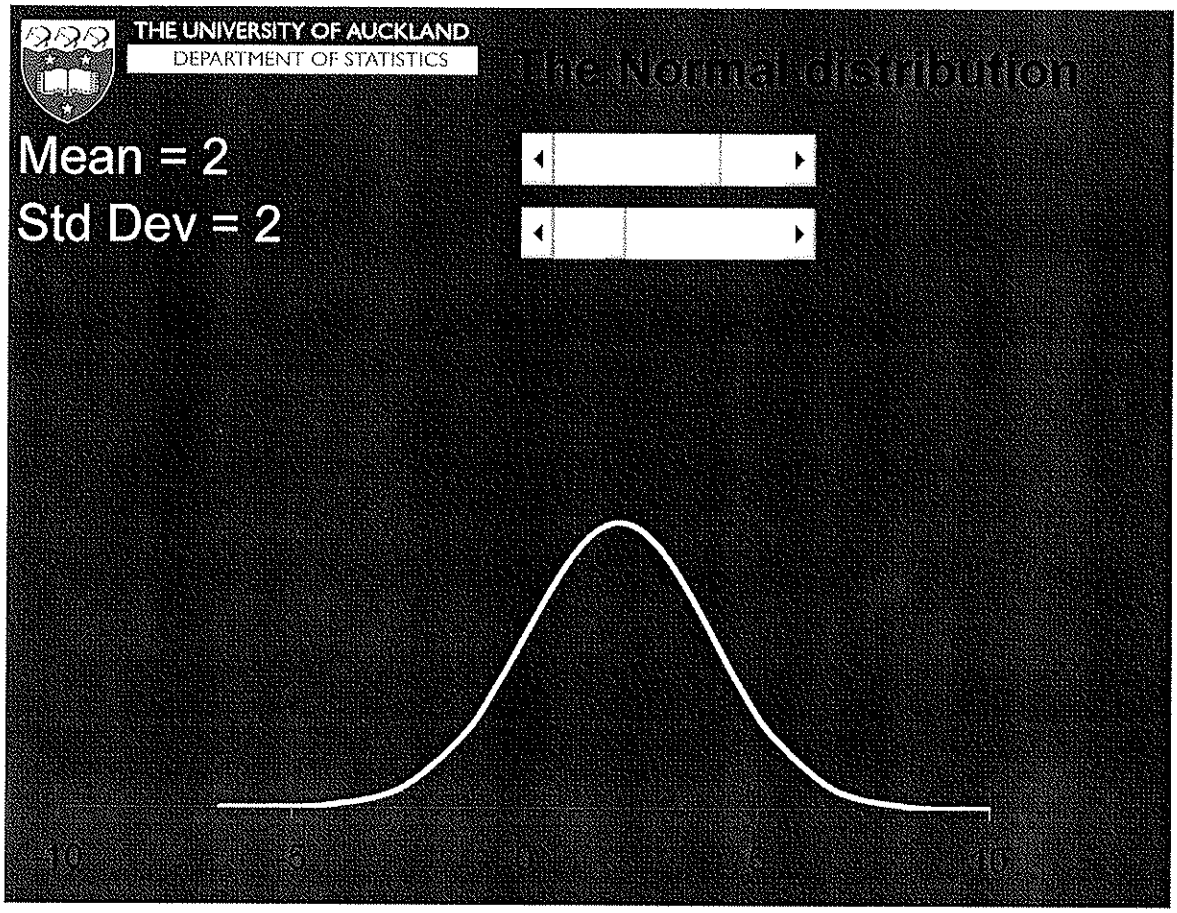
- Se il problema prevede solo la *valutazione dei parametri* del modello, stiamo risolvendo un ***problema di stima***, se è prevista anche la possibilità di *"cambiare" il modello*, stiamo risolvendo un problema di ***verifica di ipotesi***.

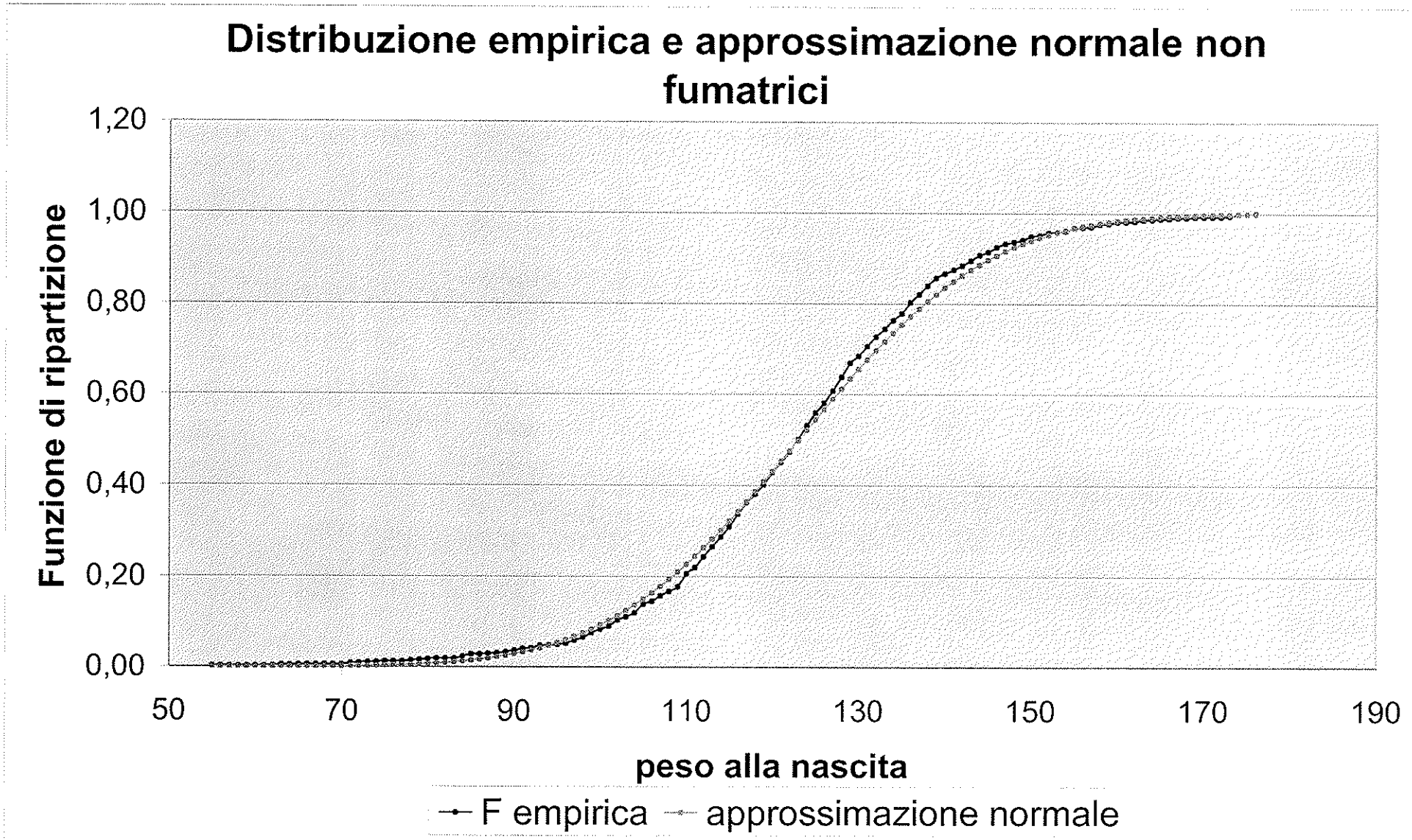
Parole chiave

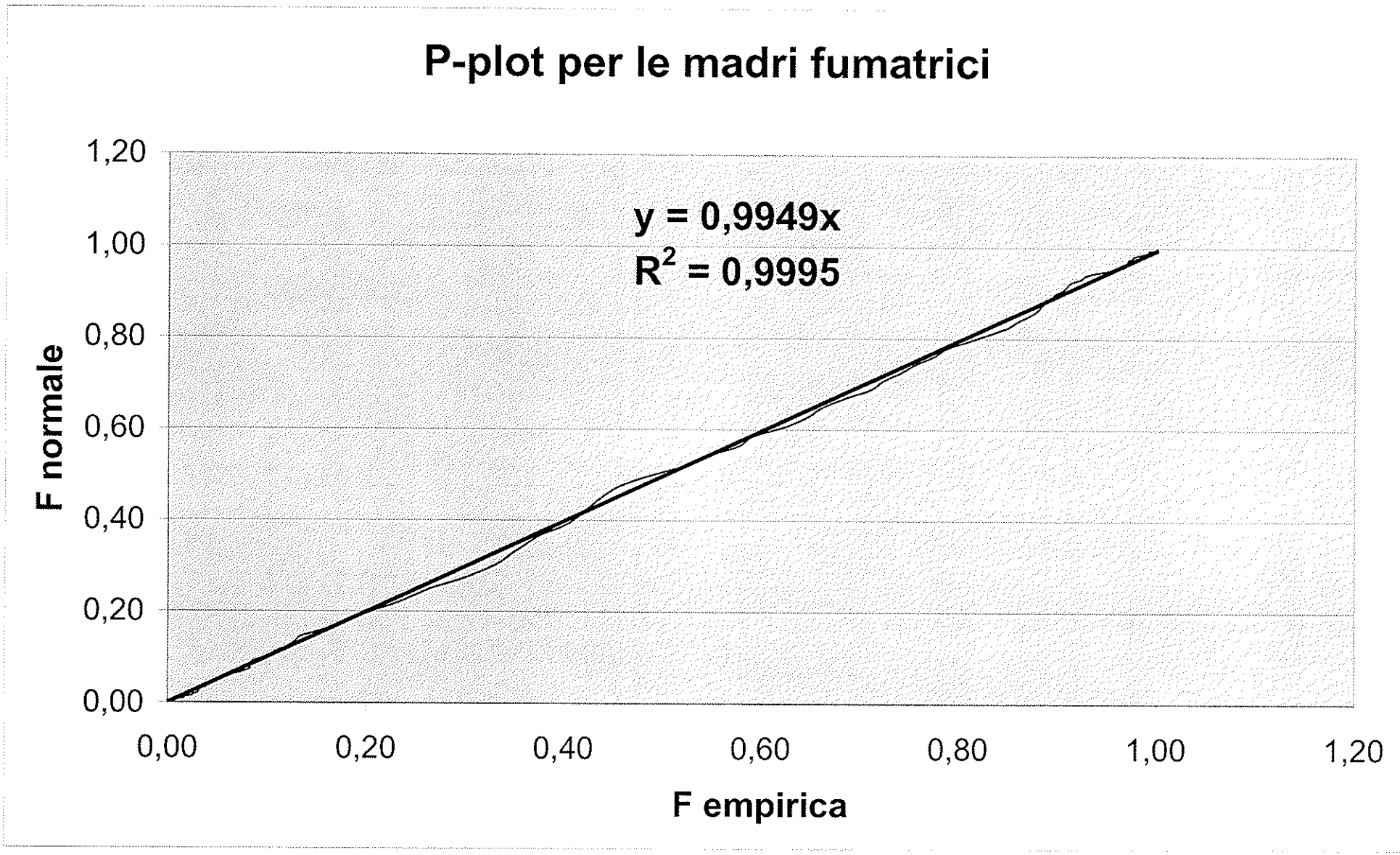
- Popolazione
- Campione
- Modello (di previsione)
- Parametro incognito (denotato con θ)
- Induzione
- Problema inferenziale di stima
- Problema inferenziale di verifica di ipotesi

5	74	11,18	5,00	95%	13,88	60,12	87,88	1
5	69,4	12,58	5,63	95%	15,62	53,78	85,02	1
5	70,8	3,70	1,66	95%	4,60	66,20	75,40	0
5	79,6	14,99	6,71	95%	18,62	60,98	98,22	1
5	70,6	10,31	4,61	95%	12,80	57,80	83,40	1
5	73,4	7,40	3,31	95%	9,19	64,21	82,59	1
5	79,4	7,67	3,43	95%	9,52	69,88	88,92	1
5	76	10,75	4,81	95%	13,34	62,66	89,34	1
5	77	6,36	2,85	95%	7,90	69,10	84,90	1
5	80	12,75	5,70	95%	15,83	64,17	95,83	1
10	72,7	9,93	3,14	95%	7,11	65,59	79,81	1
10	75,3	11,19	3,54	95%	8,00	67,30	83,30	1
10	78,7	10,53	3,33	95%	7,53	71,17	86,23	1
10	70,4	13,98	4,42	95%	10,00	60,40	80,40	1
10	82,4	8,13	2,57	95%	5,81	76,59	88,21	0
10	70,9	7,80	2,47	95%	5,58	65,32	76,48	1
10	75,3	12,50	3,95	95%	8,94	66,36	84,24	1
10	80,1	9,36	2,96	95%	6,70	73,40	86,80	1
10	75,1	11,65	3,68	95%	8,33	66,77	83,43	1
10	83,3	9,62	3,04	95%	6,88	76,42	90,18	0
20	76,9	11,57	2,59	95%	5,42	71,48	82,32	1
20	78,4	13,01	2,91	95%	6,09	72,31	84,49	1
20	75,5	11,69	2,61	95%	5,47	70,03	80,97	1
20	72,35	10,45	2,34	95%	4,89	67,46	77,24	1
20	82,1	14,45	3,23	95%	6,76	75,34	88,86	1
20	77,8	8,83	1,97	95%	4,13	73,67	81,93	1
20	75,75	10,51	2,35	95%	4,92	70,83	80,67	1
20	78,5	10,67	2,39	95%	5,00	73,50	83,50	1
20	78,25	11,60	2,59	95%	5,43	72,82	83,68	1
20	77,45	11,39	2,55	95%	5,33	72,12	82,78	1

sample size	sample mean	sample std.	standard error	confidence level	margin error	lower confidence limit	upper confidence limit	contain true mean


THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS
The Normal distribution
Mean = 2
Std Dev = 2
-10
-5
0
5
10


Distribuzione empirica e approssimazione normale non fumatrici
Funzione di ripartizione
1,20
1,00
0,80
0,60
0,40
0,20
0,00
50
70
90
110
130
150
170
190
peso alla nascita
F empirica
approssimazione normale

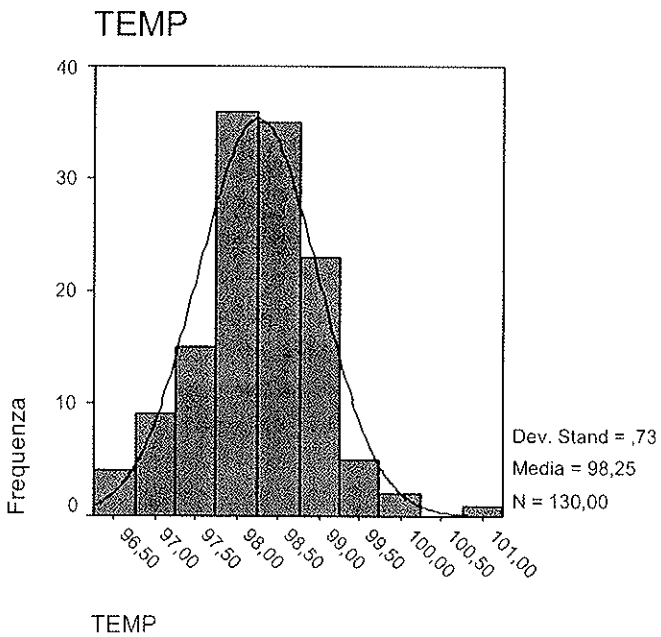

P-plot per le madri fumatrici
y = 0,9949x
R2 = 0,9995
F normale
F empirica
1,20
1,00
0,80
0,60
0,40
0,20
0,00
0,00
0,20
0,40
0,60
0,80
1,00
1,20

Esempio: test sulle medie (grandi campioni)

Descrizione dei dati.
Consideriamo il seguente insieme di dati riferiti a 130 soggetti sani di cui 65 maschi e 65 femmine, estratti da uno studio pubblicato sul *Journal of the American Medical Association.*.
Per ogni soggetto, sono state misurate la temperatura corporea (in °F) e la frequenza cardiaca (battiti/min). La matrice dei dati è riportata in appendice.
La media delle temperature corporee del campione totale risulta uguale a 98.25 °F.
La deviazione standard è uguale a 0.73.
La varianza è uguale a 0.54. Tali valori sono riportati nella tabella seguente accanto agli anologhi indici per la variabile "frequenza cardiaca".

	Temp °F	Battiti
Media	98,25	73,76
Varianza	0,54	49,87
DS	0,73	7,06
Mediana	98,30	74,00
Min	96,30	57,00
Max	100,80	89,00
CV	0,01	0,10

Esaminando l'istogramma dei dati si può notare che i valori mostrano una distribuzione statistica approssimabile con una distribuzione normale con la stessa media e la stessa deviazione standard (curva continua).



Per la distribuzione normale nell'intervallo $\mu \pm 2\sigma$ è compreso il 95.4% dei casi: per il nostro campione tale intervallo risulta: 98.2 ± 1.4 (96.8 – 99.6).

Stima di intervallo della media.

La media delle temperature corporee del campione analizzato è 98.2 °F, mentre la temperatura corporea della popolazione sana è tradizionalmente fissata a 98.6 °F.

La differenza tra i due valori potrebbe essere dovuta a una serie di motivi, per esempio alle fluttuazioni fisiologiche giornaliere della temperatura corporea (circa 0.9 °F/die); all'utilizzazione di termometri imprecisi per rilevare le temperature campionarie; alla possibilità che il campione considerato non sia rappresentativo della popolazione generale oppure che il valore tradizionalmente considerato come valore medio della temperatura corporea della popolazione (98.6 °F) non sia aggiornato, quindi non corretto. Effettivamente, se trasformiamo in °C i 2 valori considerati, otteniamo per 98.6 °F una temperatura di 37°C (comunemente considerato un indice di alterazione per la popolazione media); per 98.25 °F una temperatura di 36.8 °C (comunemente considerato indice di normalità per la popolazione media). Per valutare la significatività della differenza riscontrata possiamo calcolare l'intervallo di confidenza per la media della popolazione sulla base del nostro campione di 130 soggetti.

Fissando un livello di confidenza, determiniamo, sulla base dei dati rilevati, un intervallo entro il quale ci aspetteremmo di trovare il valore medio della popolazione.

Comunemente tale livello di confidenza si indica con α= 1-p dove p è la probabilità che l'intervallo non contenga il valore del parametro della popolazione.

Quindi, nota la distribuzione campionaria, una volta fissato α, si ottengono i valori che delimitano l'intervallo di confidenza cercato.

Al crescere di n (numerosità del campione) aumenta la precisione delle stime perché diminuisce lo scarto quadratico medio, proporzionale a $1/\sqrt{(n)}$, quindi gli estremi dell'intervallo di confidenza si restringono rendendo più accurata la stima d'intervallo del parametro di interesse.

Nel nostro caso, non conoscendo la varianza della popolazione, l'intervallo di confidenza della media per $\alpha = 0.95$ è dato da:

$98.25 - 1.96\ [0.7 / \sqrt{(130)}\ ] \leq \mu \leq 98.25 + 1.96\ [0.7 / \sqrt{(130)}]$

Per $\alpha = 0.95$ otteniamo, quindi, un intervallo compreso tra 98.13 e 98.37, intervallo che esclude il valore tradizionale 98.6.

Per $\alpha = 0.999$ l'intervallo di confidenza è: $98.25–3.291\ [0.7/11.4] \leq \mu \leq 98.25 + 3.291[0.7/11.4]$, cioè un intervallo compreso tra 98.05 e 98.45 che esclude anch'esso il valore medio della popolazione 98.6.

Verifica di ipotesi sulla media della popolazione totale.

Per formalizzare la verifica di ipotesi sulla media della popolazione possiamo applicare il test t di Student per grandi campioni (approssimazione normale).

L'ipotesi nulla H_0 è: il campione è estratto da una popolazione con temperatura media μ=98.6.

L'ipotesi alternativa H_1 è: il campione è estratto da una popolazione con temperatura media $\mu \neq 98.6$.

La temperatura media del nostro campione (n = 130) è 98.2, la varianza corretta del campione è 0.53 e la deviazione standard è 0.73.

Vogliamo verificare, al livello del 5% (p=0.05), se le osservazioni campionarie si accordano con l'ipotesi nulla, quindi, che il campione possa considerarsi estratto da una popolazione la cui temperatura corporea media è di 98.6 °F.

La zona di accettazione del test C* coincide con l'intervallo di confidenza a livello α=1-p=0,95, mentre la zona critica C è la zona complementare. Pertanto, in base alla stima di intervallo già determinata, possiamo rifiutare l'ipotesi nulla al livello di significatività fissato.

Osserviamo anche che possiamo rifiutare l'ipotesi nulla anche al livello di significatività 0,001.

Verifica di ipotesi sulla differenza delle medie.

Potremmo anche investigare la significatività della differenza fra la temperatura media del campione maschile e del campione femminile, rispettivamente 98.1 °F e 98.4 °F. Tali valori sono riportati, insieme alle altre statistiche anche per la variabile "frequenza cardiaca", nella tabella seguente.

maschi	Temp	Battiti	femmine	Temp	Battiti
Media	98,10	73,37	Media	98,39	74,15
DS	0,70	5,88	DS	0,74	8,11
Mediana	98,10	73,00	Mediana	98,40	76,00
Min	96,30	58,00	Min	96,40	57,00
Max	99,50	86,00	Max	100,80	89,00
CV	0,7%	8,0%	CV	0,8%	10,9%

Possiamo applicare il test t per grandi campioni (approssimazione normale), dopo aver verificato che i due campioni hanno varianza uguale e stimato tale varianza comune.

Verifica di ipotesi sull'uguaglianza delle varianze.

Per il test sulle varianze poniamo:

Ipotesi nulla H_0: le differenze sono dovute a fluttuazioni casuali: $\sigma_1^2 = \sigma_2^2$ $(\sigma_1^2/\sigma_2^2 = 1)$.

Ipotesi alternativa H_1: le differenze sono sistematiche: $\sigma_1^2 \neq \sigma_2^2$ $(\sigma_1^2/\sigma_2^2 \neq 1)$.

Per i nostri campioni otteniamo:

$s_1^2 = 0,49$

$s_2^2 = 0,55$

Nel calcolo del rapporto si pone sempre al numeratore il valore più elevato, per cui la statistica test risulta:

$F_{64}^{64} = 1,12$

Valore non significativo al livello 5%: possiamo considerare uguali le due varianze.

La varianza combinata si stima utilizzando la formula seguente:

$$S^2_{n+m-2} = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

che applicata ai nostri dati fornisce il valore:

$$s^2_{65+65-2} = \frac{64(0,49) + 64(0,55)}{128} = 0,52$$

da cui si ricava il valore della deviazione standard delle misure: s=0,72 e la deviazione standard di ogni media:

$$s_{12} = 0,72/\sqrt{(65)} = 0,09$$

La deviazione standard della somma e della differenza delle medie si ottiene dal seguente calcolo:

$$S(X \pm Y) = \sqrt{(\frac{s_X^2}{n} + \frac{s_Y^2}{m})}$$

che applicata ai nostri dati fornisce il valore:

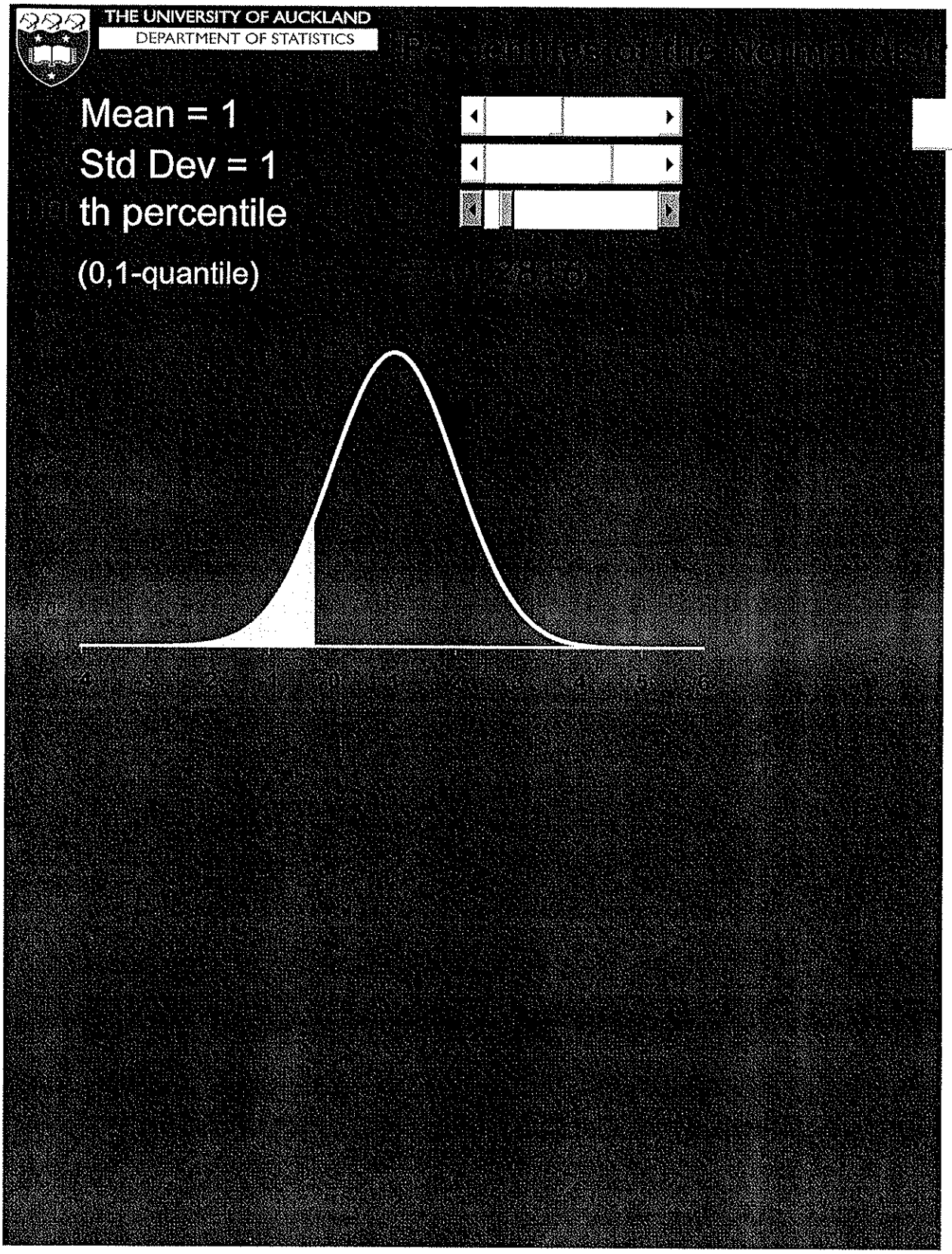
$$S(M - F) = 0{,}09\sqrt{2} = 0{,}13$$

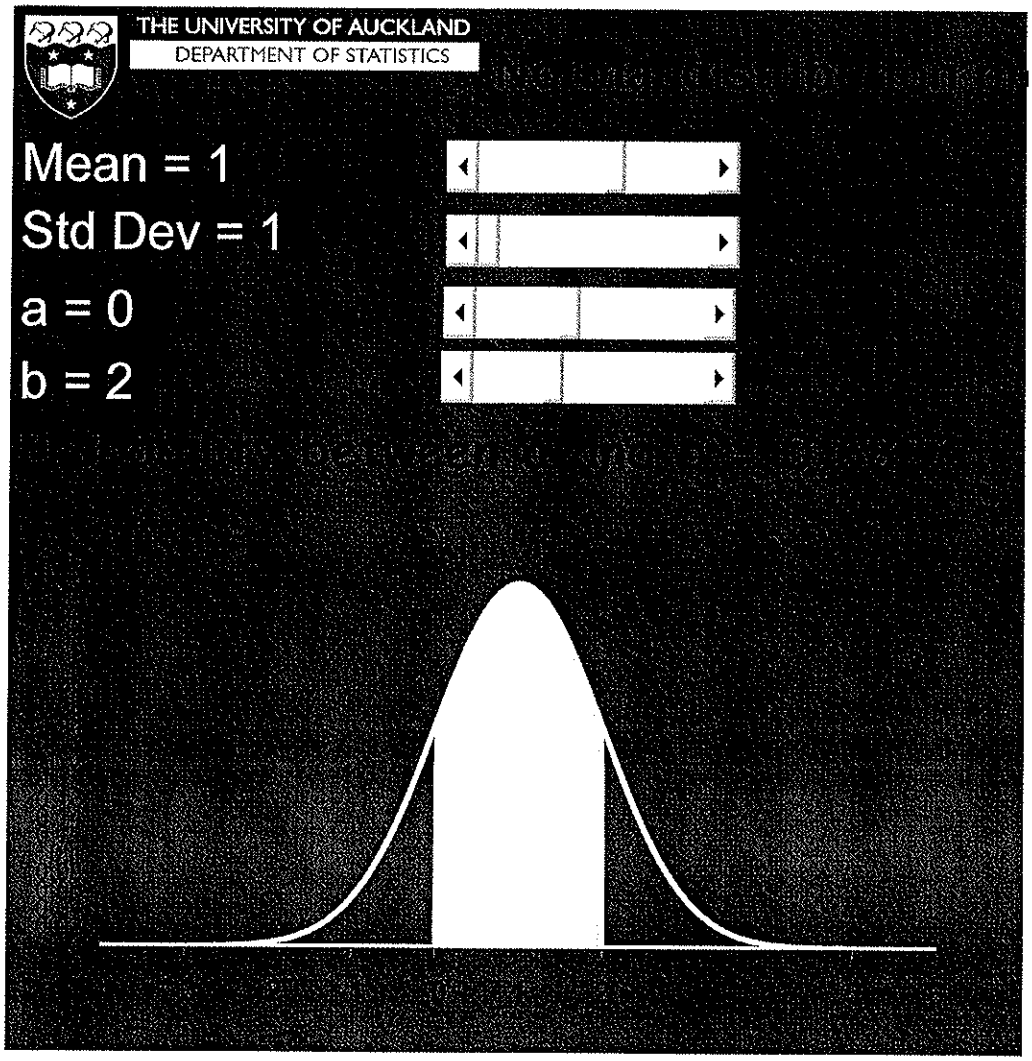
Il valore della statistica test è $t=(m_1-m_2)/s(m_1-m_2)=2{,}23$, che risulta significativo al livello 5%, dato che fornisce un livello di significatività osservato pari a $p=0.024$.

Si suggerisce di ripetere per esercizio le analisi per la variabile "frequenza cardiaca".

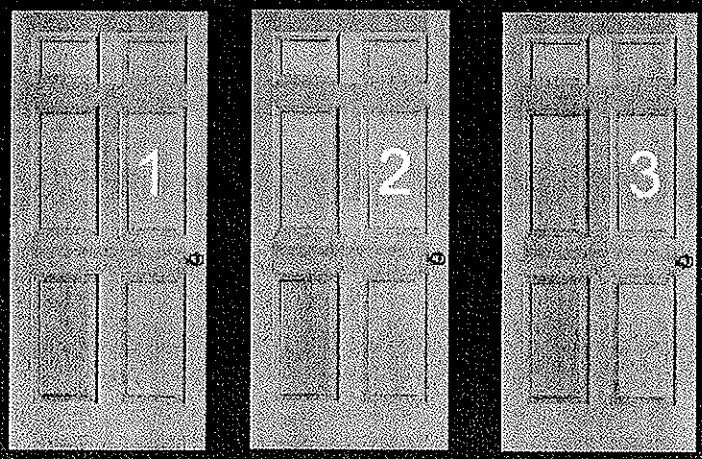
Appendice: Matrice dei dati

Temp °F	Sesso	Battiti	Temp °F	Sesso	Battiti
96,3	Maschi	70	96,4	Femmine	69
96,7	Maschi	71	96,7	Femmine	62
96,9	Maschi	74	96,8	Femmine	75
97,0	Maschi	80	97,2	Femmine	66
97,1	Maschi	73	97,2	Femmine	68
97,1	Maschi	75	97,4	Femmine	57
97,1	Maschi	82	97,6	Femmine	61
97,2	Maschi	64	97,7	Femmine	84
97,3	Maschi	69	97,7	Femmine	61
97,4	Maschi	70	97,8	Femmine	77
97,4	Maschi	68	97,8	Femmine	62
97,4	Maschi	72	97,8	Femmine	71
97,4	Maschi	78	97,9	Femmine	68
97,5	Maschi	70	97,9	Femmine	69
97,5	Maschi	75	97,9	Femmine	79
97,6	Maschi	74	98,0	Femmine	76
97,6	Maschi	69	98,0	Femmine	87
97,6	Maschi	73	98,0	Femmine	78
97,7	Maschi	77	98,0	Femmine	73
97,8	Maschi	58	98,0	Femmine	89
97,8	Maschi	73	98,1	Femmine	81
97,8	Maschi	65	98,2	Femmine	73
97,8	Maschi	74	98,2	Femmine	64
97,9	Maschi	76	98,2	Femmine	65
97,9	Maschi	72	98,2	Femmine	73
98,0	Maschi	78	98,2	Femmine	69
98,0	Maschi	71	98,2	Femmine	57
98,0	Maschi	74	98,3	Femmine	79
98,0	Maschi	67	98,3	Femmine	78
98,0	Maschi	64	98,3	Femmine	80
98,0	Maschi	78	98,4	Femmine	79
98,1	Maschi	73	98,4	Femmine	81
98,1	Maschi	67	98,4	Femmine	73
98,2	Maschi	66	98,4	Femmine	74
98,2	Maschi	64	98,4	Femmine	84
98,2	Maschi	71	98,5	Femmine	83
98,2	Maschi	72	98,6	Femmine	82
98,3	Maschi	86	98,6	Femmine	85
98,3	Maschi	72	98,6	Femmine	86
98,4	Maschi	68	98,6	Femmine	77
98,4	Maschi	70	98,7	Femmine	72
98,4	Maschi	82	98,7	Femmine	79
98,4	Maschi	84	98,7	Femmine	59
98,5	Maschi	68	98,7	Femmine	64
98,5	Maschi	71	98,7	Femmine	65
98,6	Maschi	77	98,7	Femmine	82
98,6	Maschi	78	98,8	Femmine	64
98,6	Maschi	83	98,8	Femmine	70
98,6	Maschi	66	98,8	Femmine	83
98,6	Maschi	70	98,8	Femmine	89
98,6	Maschi	82	98,8	Femmine	69
98,7	Maschi	73	98,8	Femmine	73
98,7	Maschi	78	98,8	Femmine	84
98,8	Maschi	78	98,9	Femmine	76
98,8	Maschi	81	99,0	Femmine	79
98,8	Maschi	78	99,0	Femmine	81
98,9	Maschi	80	99,1	Femmine	80
99,0	Maschi	75	99,1	Femmine	74
99,0	Maschi	79	99,2	Femmine	77
99,0	Maschi	81	99,2	Femmine	66
99,1	Maschi	71	99,3	Femmine	68
99,2	Maschi	83	99,4	Femmine	77
99,3	Maschi	63	99,9	Femmine	79
99,4	Maschi	70	100,0	Femmine	78
99,5	Maschi	75	100,8	Femmine	77


THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS
Mean = 1
Std Dev = 1
th percentile
(0,1-quantile)


THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS
Mean = 1
Std Dev = 1
a = 0
b = 2

THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS



Go Back

Should you change doors?

Many people think "it doesn't matter if you
the chances of winning are the same", bu

At the start of the game:
P(Money is behind your door) = **1/3**
P(Money not behind your door) = **2/3**

Note: we assume that the game show ho
random when the money is behind **neith**

When the money ***is*** behind one of these
game show host lets you know it's behind
change to.

So now, if you **change** doors, you have a
choosing the door with the money.

Many peopl
to understand. For lots more detailed info
problem, see

Should you change doors?

Back to Menu

Many people think "it doesn't matter if you change doors or not, the chances of winning are the same", but in fact...

At the start of the game:
P(Money is behind your door) = **1/3**
P(Money not behind your door) = **2/3**

Note: we assume that the game show host shows you a door at **random** when the money is behind **neither**.

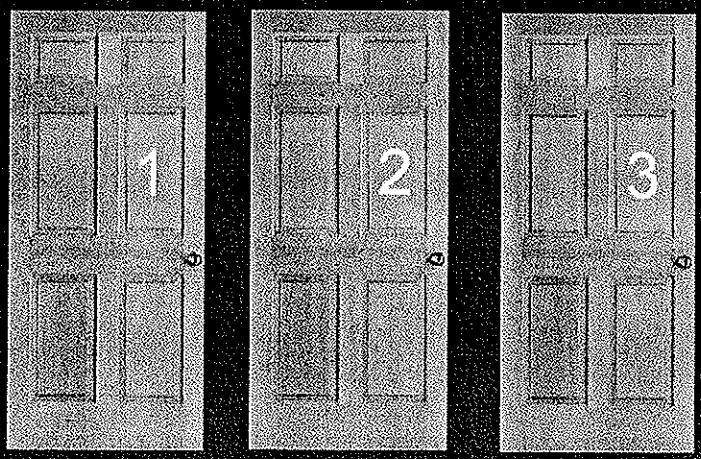
When the money ***is*** behind one of these other two doors, the game show host lets you know it's behind the door you can change to.

So now, if you **change** doors, you have a probability of **2/3** of choosing the door with the money.

Many people find this problem hard to understand. For lots more detailed information on this problem, see

THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

Choose a door:



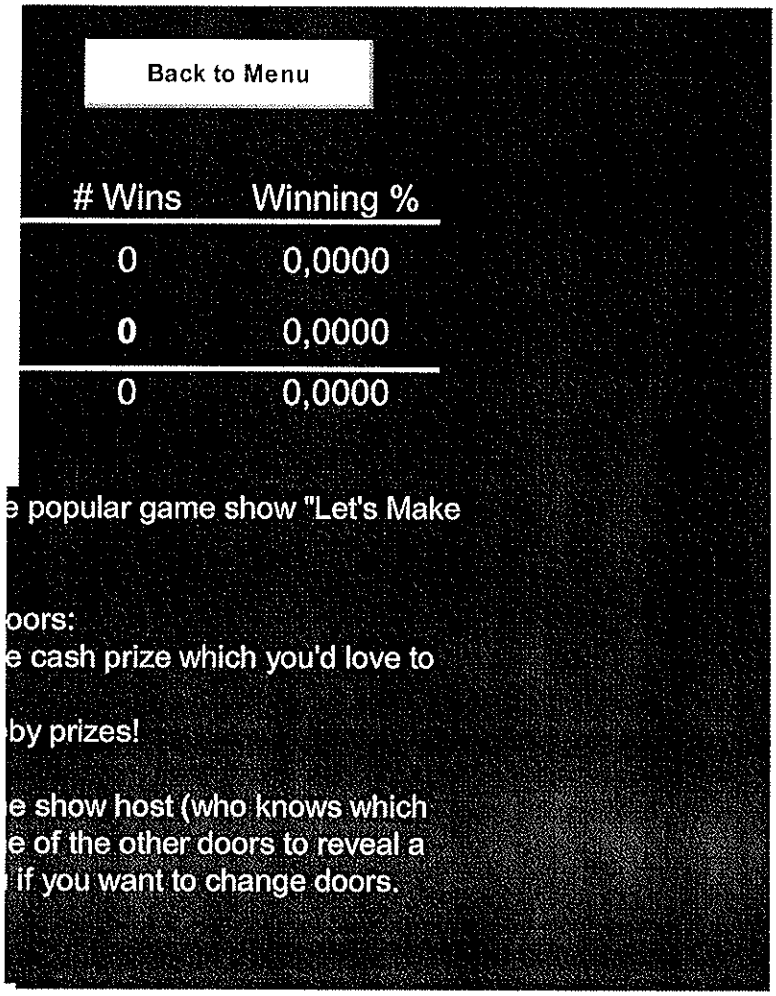
	# Games
Changed Doors	0
Didn't Change	0
Total	0

Here's your chance to take part in the po
a Deal".

You get to choose one of the three door
* Behind one of these doors is a large ca
take home.
* Behind the other two doors are booby

Once you've chosen a door, the game s
door the prize is behind) will open one o
booby prize. Then the host asks you if y

Should you change?


Back to Menu
Wins
Winning %
0
0,0000
0
0,0000
0
0,0000
e popular game show "Let's Make
oors:
e cash prize which you'd love to
by prizes!
e show host (who knows which
e of the other doors to reveal a
if you want to change doors.

I TEST: LA VERIFICA DELLE IPOTESI

La scelta del modello statistico

Il fumo della madre e la salute del neonato

- Una delle raccomandazioni mediche che appaiono sui pacchetti di sigarette negli Stati Uniti dice che il fumo in gravidanza può provocare danni al feto, nascita prematura, e peso basso alla nascita.

- Su che basi possiamo decidere che il peso alla nascita per madri fumatrici e non fumatrici è proprio diverso?

La verifica d'ipotesi

- Spesso è necessario decidere, sulla base dei dati osservati, della veridicità di asserzioni o ipotesi, che si assume siano formulate precedentemente all'esperimento.
- Mentre nei problemi di stima si cerca di determinare un valore (stima puntuale) o un insieme di valori (stima per intervallo) che approssimino il "vero" valore del parametro di interesse, in questo caso si vuole stabilire se i dati osservati suffraghino l'ipotesi formulata a priori.

L'ipotesi nulla e la falsificazione

- Le impostazioni classiche del problema si basano sulla logica della falsificazione, che ha il corrispettivo "matematico" nelle dimostrazioni per assurdo.
- Si formula quindi un'ipotesi, detta ***ipotesi nulla,*** che è in genere ***un'ipotesi di "casualità",*** che si ***"vorrebbe rifiutare",*** verificando che i dati sono in contrasto con tale ipotesi.

Sperimentazioni e verifica di ipotesi

- Quando un ricercatore sperimenta un nuovo farmaco su un gruppo di pazienti è perché crede che questo sia migliore delle cure precedenti.
- La pratica statistica però vuole che si elegga a "ipotesi privilegiata" H_0 l'ipotesi che le due cure siano uguali, cioè che le differenze eventualmente osservate siano solo dovute al caso.

Causalità e casualità

- La critica più probabile contro un esperimento dove i pazienti che usano il nuovo farmaco hanno avuto esito migliore di quelli che hanno usato la vecchia terapia, una volta accertato che l'esperimento è stato ben condotto, che i pazienti nei due gruppi erano a priori paragonabili e che il risultato è stato valutato in modo onesto, è quella di dire che la differenza nell'esperimento non è reale, ma dovuta al caso e non alla superiorità del nuovo trattamento.

Il livello di significatività

- Sarà quindi necessario valutare la probabilità che si verifichi un simile tipo di errore.
- Quando un test ci porta alla conclusione che "la differenza tra i trattamenti e' statisticamente significativa al livello 5%", stiamo proprio dicendo "i dati sono poco compatibili con l'ipotesi di nessuna differenza tra i trattamenti, per cui è opportuno concludere che c'è differenza tra i trattamenti e la differenza è così grande che non può ragionevolmente essere stata causata dal caso (la probabilità che questo sia successo è appunto inferiore a 5%, per costruzione).

Le due possibili alternative

- ***test d'ipotesi:*** si formulano due ipotesi, dette rispettivamente *ipotesi nulla* ed *ipotesi alternativa*, *l'una delle quali verrà rifiutata in favore dell'accettazione dell'altra sulla base di una regola di decisione*,
- ***test di significatività:*** prevedono la formulazione della sola *ipotesi nulla*, *che verrà rifiutata o mantenuta in vita (concetto diverso dall'accettazione) sulla base dei dati osservati.*

Test parametrici

- Abbastanza spesso la situazione vuole che si confrontino due gruppi di osservazioni, provenienti dallo stesso modello a parte, eventualmente, il valore di un parametro.
- Situazioni tipo sono che un gruppo è un campione dalla distribuzione normale $N(m_1,s^2)$, mentre l'altro gruppo è un campione dalla distribuzione normale $N(m_2,s^2)$ (medie potenzialmente differenti ma stessa varianza) ed è interessante il confronto tra m1 e m2.

Alcuni casi di interesse

- In questo caso la verifica si riconduce al confronto tra le stime di parametri e i test si chiamano: test parametrici.
- Per esempio un gruppo è un campione dalla distribuzione normale $N(m_1,s_1^2)$, mentre l'altro gruppo è un campione dalla distribuzione normale $N(m_2,s_2^2)$ ed è interessante il confronto tra s_1^2 e s_2^2;
- si ha un'osservazione da una distribuzione $Bin(n, p_1)$ e un'altra da $Bin(m,p_2)$ ed è d'interesse il confronto tra p_1 ep_2.
-

Un esperimento di Darwin

- Mi è spesso capitato di pensare che sarebbe stato consigliabile appurare se pianticelle ottenute da semi provenienti da impollinazione incrociata dei fiori fossero in qualche modo superiori a quelle derivanti da auto-impollinazione.....

I dati

Codice di identificazione della pianta madre.	Altezza in pollici delle piante figlie ottenute da impollinazione incrociata	Altezza in pollici delle piante figlie ottenute da auto-impollinazione
1	23.5	17.375
1	12	20.375
1	21	20
2	22	20
2	19.125	18.375
2	21.5	18.625
3	22.125	18.625
3	20.375	15.25
3	18.25	16.5
3	21.625	18
3	23.25	16.25
4	21	18
4	22.125	12.75
4	23	15.5
4	12	18

L'ipotesi nulla

- L'ipotesi nulla è in generale l'ipotesi di casualità:

- H_0: le differenze di altezza osservate nei due insiemi di piante sono dovuti a fluttuazioni statistiche.

L'ipotesi alternativa

- L'ipotesi alternativa è quella che si vorrebbe "dimostrare"

- H_1: le differenze di altezza osservate nei due insiemi di piante sono dovuti a alla "superiorità" dei semi ottenuti da impollinazione incrociata.

I test d'ipotesi

- Lo statistico, sulla base del campione estratto deve decidere se accettare l'ipotesi nulla H_0, ritenendola vera, oppure se rifiutarla ritenendola invece falsa; in questa seconda eventualità l'accettazione dell'ipotesi alternativa H_1 è obbligata, in quanto la falsità di H_0 implica secondo tale impostazione, la veridicità di H_1.

I test d'ipotesi

- In altre parole, si rifiuta H_0 e si accetta quindi H_1 quando il risultato sperimentale (cioè il campione osservato) si ritiene meno "verosimile" sotto (o condizionatamente a) H_0 piuttosto che sotto H_1.

Statistica test

- Solitamente nel condurre un test si preferisce lavorare, anziché con il campione complessivo, con una funzione di questo che prende il nome di statistica test ed in base alla quale verrà individuata una regola di decisione.

Statistica test

- ***Definizione:*** Si dice ***statistica test*** ogni funzione delle osservazioni campionarie da utilizzare in un test:

$$T(\mathbf{X}) = T(X_1, X_2, \ldots, X_n)$$

Statistiche test sono tutti i parametri che caratterizzano un modello statistico e che sono stimati dal campione: ***media, deviazione standard, probabilità di successo in una prova....***

Zea Mais

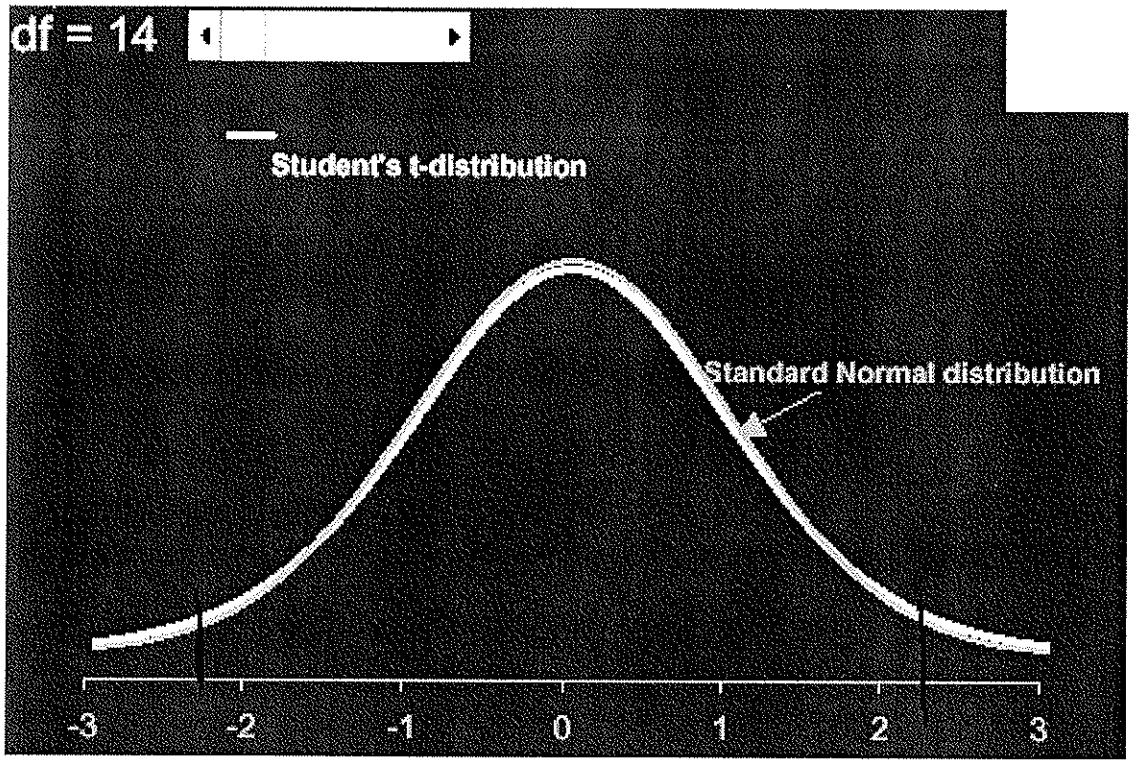
- In questo caso la statistica test è certamente la media di cui conosciamo la legge di variabilità.
- Dalle tavole della distribuzione t di Student con 14 gradi di libertà possiamo ricavare le informazioni utili per determinare una regola di decisione che, in base al valore ottenuto dal calcolo della statistica test, ci permetta di decidere se tale valore contrasta o no con l'ipotesi nulla di causalità.

Dagli intervalli al test (5%)

20,19	17,54	medie	
3,62	2,04	dev.st	
0,93	0,53	dev.st.medie	

a=20,19-1,99=18,02
b=20,19+1,99=22,18
a=17,54-1,13=16,41
b=17,54+1,13=18,67

La zona di rifiuto (5%, 2,14)



Che cosa si può osservare?

- la media del primo gruppo di piante è più piccola del limite inferiore dell'intervallo di confidenza al 95% determinato per il secondo gruppo;
- la media del secondo gruppo è superiore al limite superiore dell'intervallo determinato per il primo gruppo.
- In entrambi i casi i valori ottenuti sono poco probabili, meno del 5%, sotto l'ipotesi di scostamenti casuali.

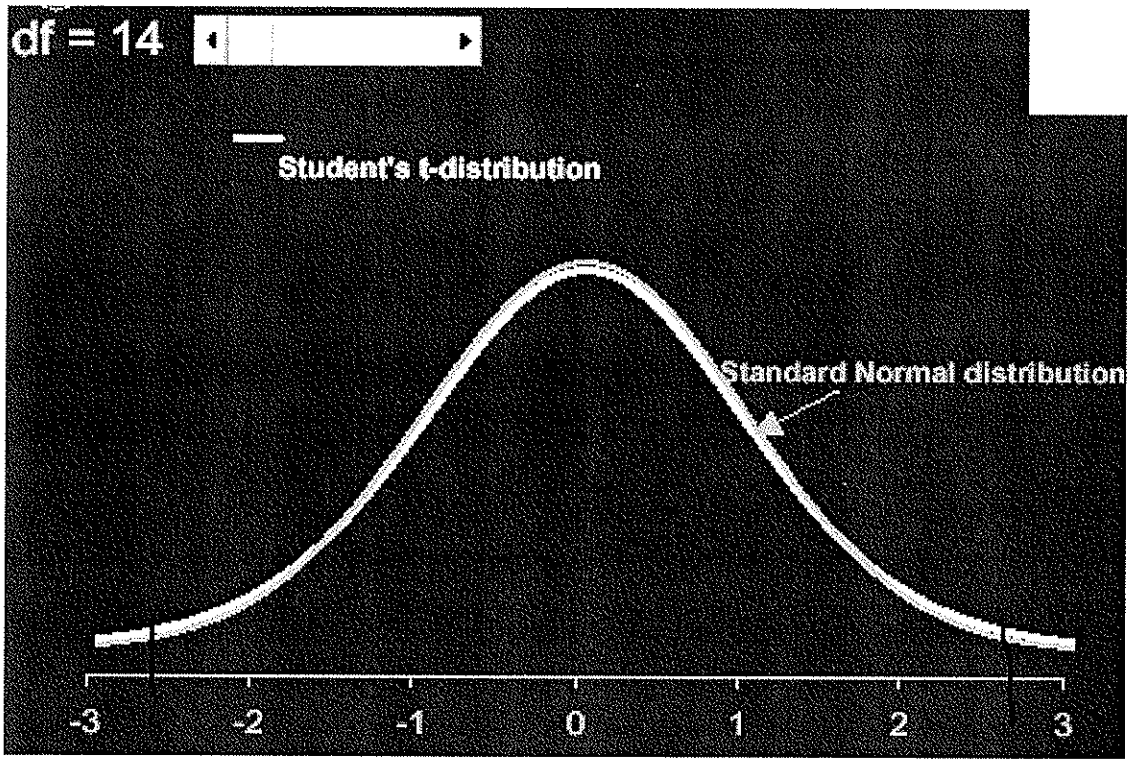
Decisione

- Possiamo decidere che l'ipotesi di casualità deve essere rifiutata e sappiamo che corriamo il rischio di sbagliare con una probabilità inferiore al 5%.
- Si dice che il risultato ottenuto mostra una significatività al livello 5%. Cioè che l'ipotesi H_0 può essere rifiutata con un livello di significatività del 5%.

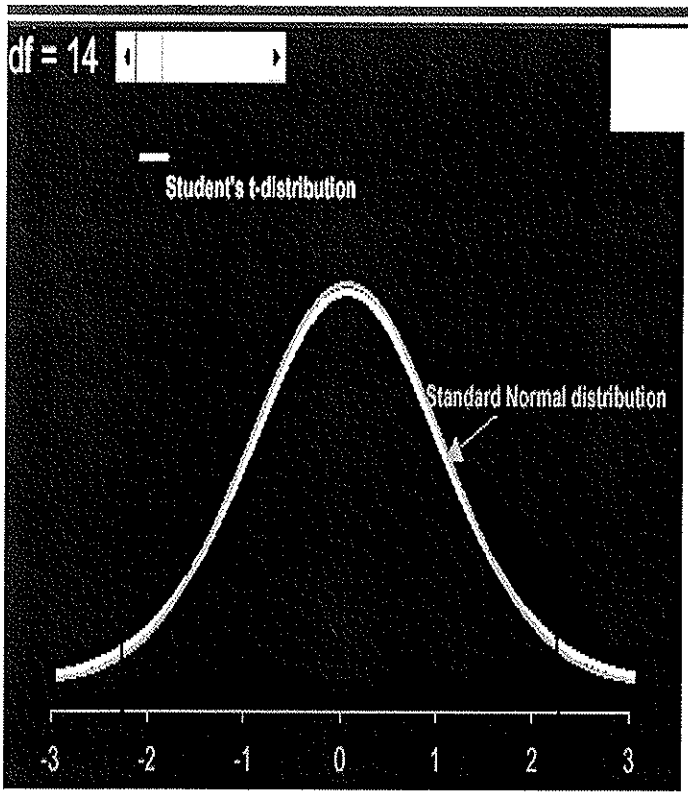
Dagli intervalli al test (2%)

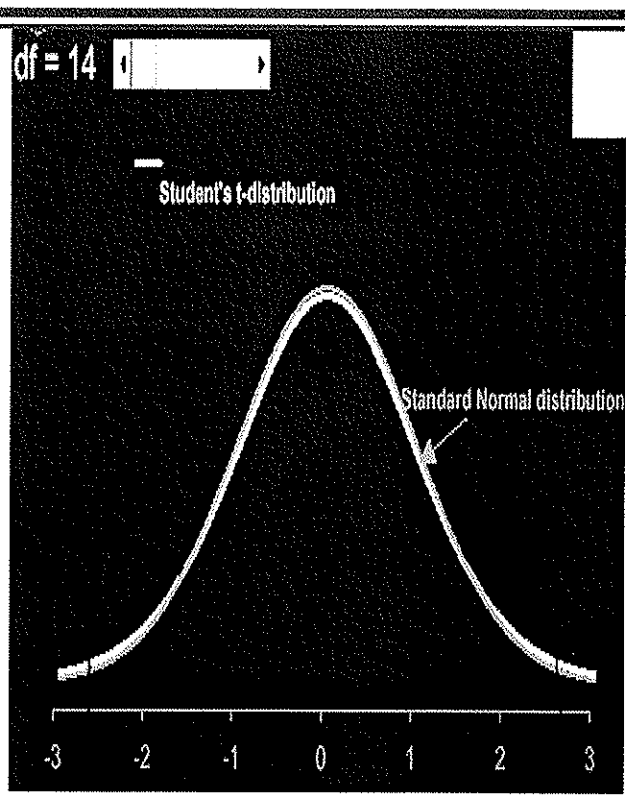
a=20,19-2,44=17,75
b=20,19+2,44=22,63
a=17,54-1,39=16,15
b=17,54+1,39=18,93

La zona di rifiuto (2%, 2,62)



Confrontiamo le regole e i livelli





La zona critica al livello α

- La zona (di rifiuto) della distribuzione che cade fuori dell'intervallo di confidenza (al livello 1-α) si chiama zona critica e si indica con C.
- Tale zona rappresenta quella regione dello spazio in cui variano i valori della statistica test, che corrisponde a una probabilità di appartenenza sotto H_0 fissata, che indichiamo con α (complementare del livello di confidenza).
- Se osserviamo un valore in tale regione rifiutiamo H_0 con un livello di significatività α.

Errori di I e II specie

- Nel condurre un test con tale procedura si potrebbe incorrere in due tipi di errore, e cioè nell'***errore di prima specie***, *rifiutando H_0 quando in realtà questa è vera*, e nell'***errore di seconda specie*** *accettando H_0 quando invece questa è falsa.* Si ha quindi la seguente definizione:
- ***Definizione.*** si dicono:
  - ***Probabilità di errore di prima specie*** =
$$P(T \in C \mid H_0) = \alpha$$
  - ***Probabilità di errore di seconda specie*** =
$$P(T \notin C \mid H_1) = \beta.$$

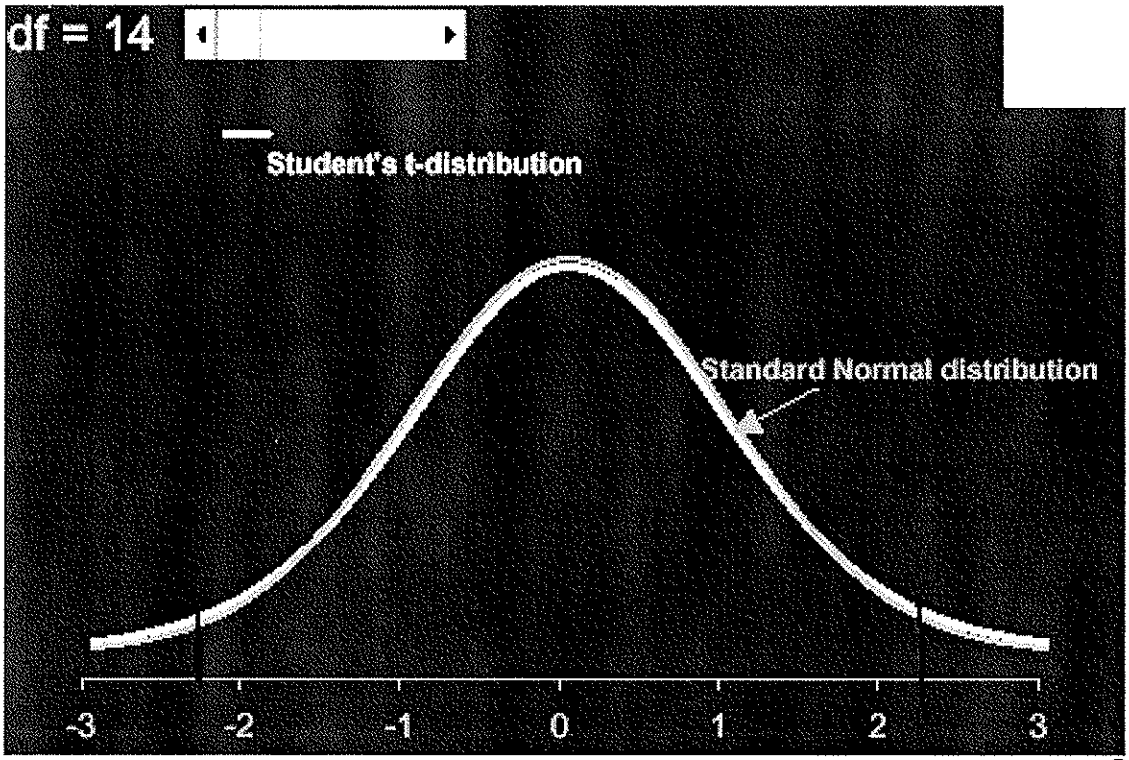
Errori di I e II specie

- Sarebbe auspicabile che tali probabilità fossero entrambe pari a 0 ma ciò, se non in casi banali, e impossibile; si dovrà tentare allora, di mantenerle simultaneamente ad un livello ragionevolmente basso. La strategia che di solito si segue, è quella di fissare l'estremo superiore rispetto della probabilità dell'errore di prima specie ad un livello α basso (0.05 o 0.01).

Test bilaterale e unilaterale

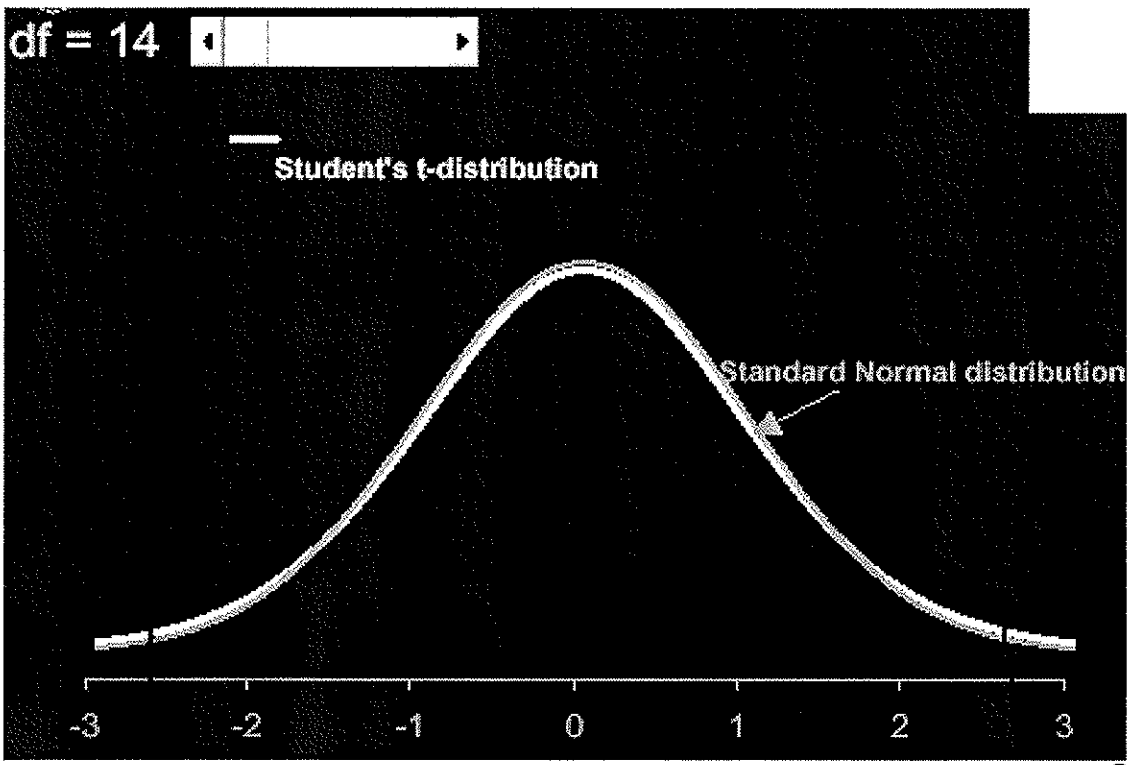
- Sarebbe più preciso formulare le due ipotesi come segue:
- H_0: le differenze di altezza osservate nei due insiemi di piante sono dovuti a fluttuazioni statistiche, ovvero l'altezza media delle due popolazioni è la stessa: $\theta_A=\theta_I$
- H_1: le differenze di altezza osservate nei due insiemi di piante sono dovuti a alla "superiorità" dei semi ottenuti da impollinazione incrociata: $\theta_A<\theta_I$

La zona critica (2,5%, 2,14)



31

La zona critica (1%, 2,62)



32

Confrontiamo

- Prendiamo il modello di riferimento sotto H_0, cioè la distribuzione delle piante ottenute per autoimpollinazione, e confrontiamo il valore della media dell'altro gruppo con tale modello standardizzando, cioè verifichiamo dove si situa il valore della statistica test:

$$T_{14} = \frac{20,59 - 17,54}{0,53} = 5,75$$

- rispetto alla distribuzione t_{14} standardizzata.

Livello di significatività osservato

- Controllando la tavola otteniamo che la probabilità di valori superiori a quello osservato è inferiore a 0,0005, che corrisponde all'ascissa 4,14.
- Il valore ottenuto è quindi significativo a qualunque livello si voglia fissare almeno fino a 0,0005, e anche più piccolo.
- La probabilità della coda a destra del valore osservato è praticamente zero.
- Tale probabilità si chiama: livello di significatività osservato e si indica con p.

p e i test di significatività

- Per effettuare un test di significatività, le informazioni fornite dal campione vengono sintetizzate in una *statistica test* $D(X)$ che possa essere interpretata come misura della distanza tra le osservazioni campionarie e l'ipotesi H_0. In altre parole D viene definita in maniera che suoi valori grandi costituiscano un'indicazione contro H_0, mentre suoi valori piccoli costituiscano un'indicazione a favore di H_0.

I test di significatività

- La scelta di tale statistica test deve effettuarsi prima di osservare i dati campionari, ed è basata sul tipo di deviazione dall'ipotesi nulla, che si vuole individuare.

  Dopo aver estratto il campione si può calcolare il valore osservato $d = D(x)$ della statistica $D(X)$ sulla base del quale si calcola il *livello di significatività osservato p*.

Livello di significatività osservato

Quindi il valore p è la probabilità che la statistica D assuma, nel caso in cui l'ipotesi H_0, sia vera, un valore "grande" almeno come quello osservato d.

In questo modo si capisce come il livello di significatività osservato p(d) possa essere interpretato come una misura di evidenza sperimentale a favore di H_0.

Livello di significatività osservato

Infatti se $p(d)$ è molto piccolo allora vuol dire che se H_0 fosse vera sarebbe ben difficile ottenere un valore di D maggiore o uguale di quello osservato, e quindi una distanza tra l'ipotesi e i dati campionari maggiore o uguale di quella fornita da d;

ciò porta ad affermare che H_0 è falsa e quindi a rifiutarla.

Livello di significatività osservato

Viceversa, un valore $p(d)$ elevato, essendo un'indicazione a favore di H_0, non dà luogo a valutazioni di tipo conclusivo, in quanto rivela soltanto una mancanza di evidenza contro H_0.

Livello di significatività osservato

- Solitamente si considera il valore $p(d) = 0,05$, come quello che divide i livelli di significatività osservati "piccoli" da quelli "grandi": se $p(d) < 0,05$, allora si dice che l'ipotesi è contraddetta dai dati al livello del 5%, mentre se $p(d) > 0,05$, allora l'ipotesi si dice consistente con i dati, ancora al livello del 5%. E' chiaro però che questa non è nulla di più di una convenzione e che livelli di significatività pari a 0,051 e a 0,049 indicano una pari evidenza contro l'ipotesi H_0.

Ipotesi su una media

Test normale

- Sia $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ un campione casuale estratto da una popolazione normale di media μ ignota e varianza σ^2 nota, e si voglia testare l'ipotesi nulla H_0: $\mu=\mu_0$ contro una delle ipotesi alternative H_1: $\mu\neq\mu_0$, H_1: $\mu < \mu_0$, o H_1: $\mu > \mu_0$.

Ipotesi su una media

- Per la scelta della statistica test appare naturale sfruttare il fatto che, se $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$ è un campione casuale proveniente da una popolazione $N(\mu,\sigma^2)$, allora la variabile media aritmetica del campione segue ancora una distribuzione normale di media μ e di varianza σ^2/n, cioè $N(\mu,\sigma^2/n)$.

Test Normale

- Si può allora assumere come statistica test la funzione:

$$T(\boldsymbol{X}) = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

che sotto l'ipotesi nulla H_0 è una $N(0,1)$, cioè una normale standardizzata, di cui esistono le tavole di probabilità.

Sulla base dei valori critici ricavabili da tali tavole, si determinano le regioni di accettazione e di rifiuto.

Verifica d'ipotesi per grandi campioni

- Il test normale può essere utilizzato anche nel caso in cui si voglia verificare un'ipotesi sulla media μ di una distribuzione arbitraria con varianza incognita. Si dimostra infatti che, dato il campione $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$ proveniente da una popolazione arbitraria, se s^2 è uno stimatore consistente di σ^2.

Verifica d'ipotesi per grandi campioni

- per n sufficientemente grande (per convenzione $n > 30$) la statistica test

$$T(\boldsymbol{X}) = \frac{\overline{X} - \mu_0}{\sqrt{\hat{\sigma}^2 / n}}$$

converge, sotto H_0, ad una $N(0,1)$ (Teorema centrale). Sotto queste condizioni il problema può essere risolto secondo i criteri di un test normale.

Confrontiamo le medie dei pesi alla nascita

Prendiamo come popolazione di riferimento quella delle madri non fumatrici e confrontiamo la media dei pesi dei neonati delle madri fumatrici con il modello di riferimento a partire dagli intervalli di confidenza.

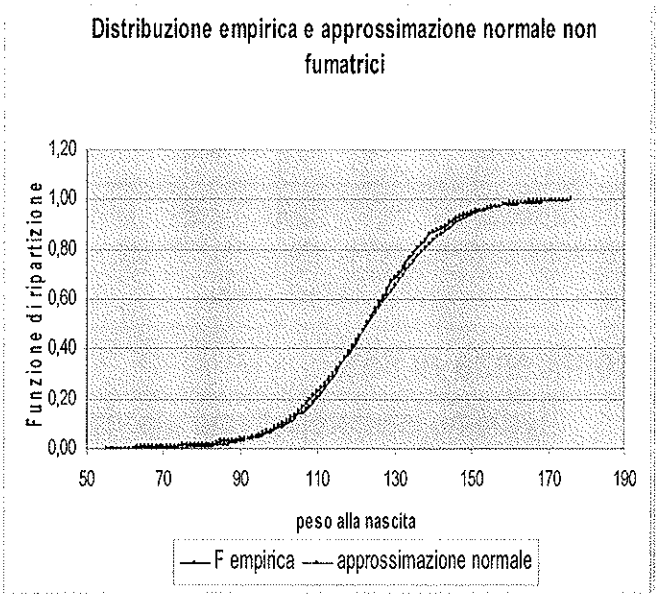
	Peso alla nascita			
	media	deviazione standard	n	errore standard
madri fumatrici	114,11	18,1	484	0,82
madri non fumatrici	123,05	17,4	742	0,64

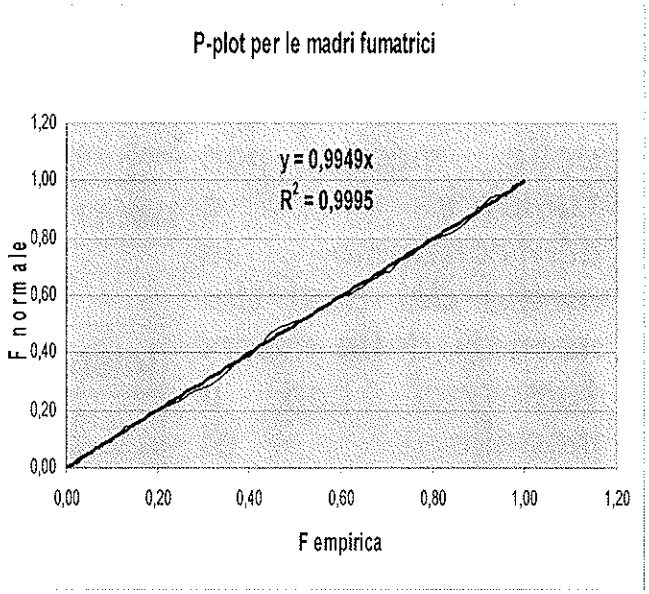
Intervalli di confidenza al 95%		
	estremo inferiore	estremo superiore
madri fumatrici	114,11-1,96(0,82)	114,11+1,96(0,82)
madri non fumatrici	123,05-1,96(0,64)	123,05+1,96(0,64)
madri fumatrici	112,5	115,72
madri non fumatrici	121,80	124,3

Confronto con il modello normale

- Data la numerosità delle osservazioni in entrambi i gruppi possiamo utilizzare per la distribuzione delle medie campionarie l'approssimazione normale.
- Anche le misure, comunque, sono ben approssimate dalla distribuzione normale.
- Possiamo verificarlo confrontando i grafici delle funzioni di ripartizione e i p-plot.

Confronti grafici





Come si vede in entrambi i gruppi la distribuzione statistica del peso alla nascita è ben approssimata dal modello normale con media e varianza stimate dai due campioni.

L'ipotesi nulla

- L'ipotesi nulla H_0 si può formulare:

- **Il peso medio alla nascita per i neonati di madri fumatrici è uguale a quelle dei neonati di madri non fumatrici.**

Statistica test

- La statistica test è la media dei pesi relativi alle madri fumatrici standardizzata rispetto alla popolazione presa come riferimento (madri non fumatrici).

$$Z(X) = \frac{\overline{X} - \mu_0}{\sqrt{\hat{\sigma}^2 / n}} = \frac{114{,}11 - 123{,}05}{0{,}82} = -10{,}9$$

- Il valore di p osservato è praticamente 0.
- Il risultato è significativo a qualunque livello standard.

Verifica d'ipotesi su una frequenza relativa

- Si consideri il caso in cui si voglia verificare un'ipotesi formulata circa la proporzione p di individui di una popolazione caratterizzati dalla presenza di un certo attributo.
- ipotesi nulla H_0: $p=p_0$
- ipotesi alternativa H_1: $p \neq p_0$

Il campione

- $X = (X_1, X_2, \ldots X_n)$
- X_i rappresenta il verificarsi di un certo evento nella generica *i-ma* prova e assume valore 1 (corrispondente ad un successo) con probabilità p e valore 0 (corrispondente ad un insuccesso, ovvero un fallimento) con probabilità $1-p$,

La statistica test

- La statistica test è il numero di successi (frequenza assoluta):

$$T = \sum_{i=1}^{n} X_i$$

- che, sotto l'ipotesi nulla, si distribuisce come una binomiale di parametri (n, p_0) la cui media è $m=np_0$ e la varianza è $\sigma^2=np_0(1-p_0)$.

Grandi campioni

- Se la numerosità n del campione è grande, ci si riconduce al test normale, in quanto, per il teorema centrale del limite, una binomiale di parametri (n, p) è approssimabile, per n grande, una normale con la stessa media e la stessa varianza, cioè $N(np, np(1-p))$:

$$Z = \frac{T/n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{f_n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Esempio

- Se su 120 prove si sono avuti 40 successi e l'ipotesi nulla si pone:
- ipotesi nulla H_0: $p=0,38$
- *La statistica test vale* $T=40$ *e* $f_{120}=1/3$
- *Allora si ottiene il valore non significativo al 5%:*

$$Z = \frac{1/3 - 0,38}{\sqrt{0,38(1-0,38)/120}} = \frac{0,047}{0,044} = 1,07$$

Dati appaiati: Zea Mais

- "Student" utilizzò i dati di Darwin sulle piantine di zea mais considerandoli appaiati e utilizzando le differenze delle colonne come misure di una nuova variabile statistica "differenza" da confrontare con lo zero.
- La matrice dei dati è la seguente:

Dati appaiati: Zea Mais

unit	diff (inches)
1	49
2	-67
3	8
4	16
5	6
6	23
7	28
8	41
9	14
10	29
11	56
12	24
13	75
14	60
15	-48
media	20,93
dev. stand. dati	37,74
gradi di liberta'	14
dev. stand. media	9,75
moltiplicatore test a 1 coda 5%	1,75
zona di rifiuto >a	17,06
moltiplicatore test a 1 coda 1%	2,6
zona di rifiuto >a	25,35

Il test per le differenze

- Se le differenze provengono da campioni di una stessa popolazione la loro media teorica è nulla e si distribuiscono secondo un t di student con n-1 gradi di libertà. Per n maggiore di 30 si utilizza generalmente l'approssimazione normale
- Il test consiste nel confrontare la media osservata con lo zero.

Le ipotesi

- Ipotesi nulla:
- H_0: le differenze sono dovute a fluttuazioni casuali: $\theta=0$.
- Ipotesi alternativa:
- H_1: le differenze sono dovute al fatto che l'impollinazione incrociata produce piante superiori: $\theta>0$.

La distribuzione delle differenze

- Statistica test T(X)=Media osservata
- T(x)=20,93
- Deviazione standard dei dati: 37,74
- Deviazione standard della media: 9,75
- Zona critica del test unilaterale al 5%: T(X)>1,76(9,75)=17,16
- Zona critica del test unilaterale al 1%: T(X)>2,62(9,75)=25,54
- Il valore ottenuto è significativo al livello 5%, ma non al livello 1%.
- Rifiutiamo l'ipotesi nulla al livello 5%, ma non al livello 1%.

Test normale (dati temperatura)

- Numerosità campionaria: n=130
- Metà maschi e metà femmine
- Vogliamo verificare se c'è differenza in media tra le temperature dei maschi e delle femmine

Temp	Sex	Battiti
96,30	1	70
96,70	1	71
96,90	1	74
97,00	1	80
97,10	1	73
97,10	1	75
97,10	1	82
99,10	2	74
99,20	2	77
99,20	2	66
99,30	2	68
99,40	2	77
99,90	2	79
100,00	2	78
100,80	2	77

Le ipotesi

- Ipotesi nulla:
- H_0: le differenze sono dovute a fluttuazioni casuali: $\theta_1=\theta_2(\theta_1-\theta_2=0)$.
- Ipotesi alternativa:
- H_1: le differenze sono sistematiche: $\theta_1\neq\theta_2$ $(\theta_1-\theta_2\neq0)$.
- Per effettuare il test usiamo come statistica la differenza delle medie di cui occorre calcolare la deviazione standard.

Varianze temperature

- Per calcolare la deviazione standard della differenza occorre prima verificare se le varianze dei due campioni da confrontare si possono considerare uguali.
- In tal caso:
- Ipotesi nulla:
- H_0: le differenze sono dovute a fluttuazioni casuali: $\sigma_1^2 = \sigma_2^2 (\sigma_1^2 / \sigma_2^2 = 1)$.
- Ipotesi alternativa:
- H_1: le differenze sono sistematiche: $\sigma_1^2 \neq \sigma_2^2$ $(\sigma_1^2 / \sigma_2^2 \neq 1)$.

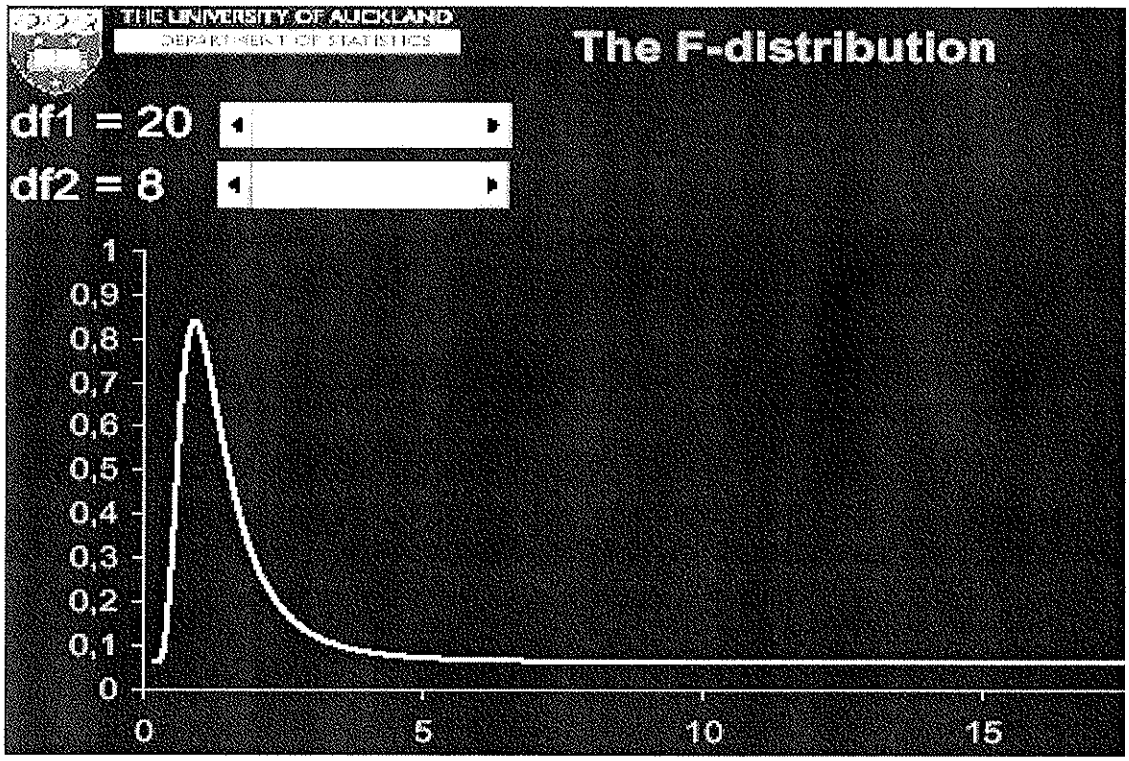
La statistica F

- $s_1^2 = 0,49$
- $s_2^2 = 0,55$
- Nel calcolo del rapporto si pone sempre al numeratore il valore più elevato

- $F_{64}^{64} = 1,12$

- Valore non significativo: possiamo considerare uguali le due varianze.

La distribuzione F



La varianza

- In tal caso possiamo stimare un'unica varianza calcolando la media ponderata delle due varianze:

$$S^2_{n+m-2} = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

$$s^2_{65+65-2} = \frac{64(0{,}49) + 64(0{,}55)}{128} = 0{,}52$$

Il test

- La deviazione standard delle misure risulta:
- s=0,72
- e la deviazione standard di ogni media:
- S_{12}=0,72/√(65)=0,09

La deviazione standard della somma e della differenza

- La deviazione standard della somma e della differenza delle medie si ottiene dal seguente calcolo:

$$S(X \pm Y) = \sqrt{(\frac{s_X^2}{n} + \frac{s_Y^2}{m})}$$

- che nel nostro caso risulta:

$$S(M - F) = 0{,}09\sqrt{2} = 0{,}13$$

Il risultato del test

- La statistica test vale:
- t=(m1-m2)/s(m1-m2)=2,23
- che risulta significativo al livello 5% utilizzando la tavola della normale dato che si tratta di un grande campione.
- Il valore osservato della significatività è p=0,024.

Confrontiamo le medie dei pesi alla nascita

Applichiamo il test sulla differenza delle medie per campioni indipendenti e, data la gaussianità delle misure di partenza e le numerosità campionarie calcoliamo direttamente la statistica test per varianze diverse e utilizziamo le tavole della normale N(0,1). Si usa il test unilaterale e il valore è significativo a ogni livello prefissato (p=0)

	Peso alla nascita			
	media	deviazione standard	n	errore standard
madri fumatrici	114,11	18,1	484	0,82
madri non fumatrici	123,05	17,4	742	0,64

$$T(X-Y)=(\overline{X}-\overline{Y})/\sqrt{(\frac{s_X^2}{n}+\frac{s_Y^2}{m})}$$

$$T(X-Y)=(123{,}05-114{,}11)/\sqrt{(\frac{302{,}76}{742}+\frac{327{,}61}{484})}$$

$$T(X-Y)=8{,}6$$

Test sulle medie (dati non appaiati)

- Does increasing calcium intake reduce blood pressure? Observational studies suggest that there is a link, and that it is strongest in African-American men. Twenty-one African-American men participated in an experiment to test this hypothesis. Ten of the men took a calcium supplement for 12 weeks while the remaining 11 men received a placebo. Researchers measured the blood pressure of each subject before and after the 12-week period. The experiment was double-blind.

La matrice dei dati

Treatment	Begin	End	Decrease
Calcium	107	100	7
Calcium	110	114	-4
Calcium	123	105	18
Calcium	129	112	17
Calcium	112	115	-3
Calcium	111	116	-5
Calcium	107	106	1
Calcium	112	102	10
Calcium	136	125	11
Calcium	102	104	-2
Placebo	123	124	-1
Placebo	109	97	12
Placebo	112	113	-1
Placebo	102	105	-3
Placebo	98	95	3
Placebo	114	119	-5
Placebo	119	114	5
Placebo	112	114	2
Placebo	110	121	-11
Placebo	117	118	-1
Placebo	130	133	-3

Test sull'uguaglianza delle varianze

- La varianza delle differenze dei trattati è 76,44; quella dei non trattati (placebo) è 34,82. Il valore della statistica test è
- $F_{10}^{9}=2,2$
- Il valore critico al 5% è 3,02, dunque le varianze non sono significativamente diverse e si può fare la stima combinata.
- La varianza comune risulta 54,53 e la deviazione standard della differenza media 3,25.

Il test

- A pooled t-test is appropriate for comparing the change in blood pressure between the treatment and placebo groups. Normal probability plots of blood pressure change for the two groups do not show problematic departures from normality, and the sample standard deviations do not rule out equal population standard deviations. The pooled t-statistic (standardized difference of the means) is 1,63, with a p-value of 0,059 (unilateral test).

Differenza di proporzioni

Esito	Maschi	Femmine	Totale
Allontanamento	35 (33,3%)	18 (22%)	53 (28,4%)
Dimissioni	44 (42%)	46 (56,1%)	90 (48,1%)
Trasferimento	5 (4,7%)	7 (8,5%)	12 (6,4%)
Decesso	-	1 (1,2%)	1 (0,5%)
Non Rilevati	21 (20%)	10 (12,2%)	31 (16,6%)
Totale	105 (100%)	82 (100%)	187 (100%)

Le ipotesi

- Supponiamo di voler verificare la seguente ipotesi:
- L'esito allontanamento è più frequente tra i maschi che tra le femmine.
- Possiamo formulare la seguente ipotesi nulla:
- H_0: $\theta_m = \theta_f$
- E l'ipotesi alternativa:
- H_1: $\theta_m > \theta_f$

La statistica test

- t=(f1-f2)/s(f1-f2)
- Con:
- f1=0,33
- f2=0,22
- s(f1)=√[0,33(1-0,33)/105]=0,046
- s(f2)=√[0,22(1-0,22)/82]=0,046
- s(f1-f2)= √[s^2(f1)+s^2 (f2)]=0,046 √2=0,065
- t=1,69
- Il valore critico al 5% per il test unilaterale con approssimazione normale è 1,65. Il risultato è significativo al livello fissato.

Test di ipotesi per tabelle di contingenza

- Si può impostare un test di ipotesi per verificare la dipendenza tra la classificazione di riga e quella di colonna in una tabella di contingenza.
- L'ipotesi nulla sarà sempre:
- H_0: i due caratteri considerati (sulle righe e sulle colonne) sono indipendenti.
- Diversi indici possono essere usati come statistiche test, ma solo dell'indice χ^2 **si conosce la distribuzione teorica** nell'ipotesi nulla.

Statistica test: l'indice χ^2

- $D(X) = \chi^2$

$$\chi^2 = \frac{\sum_i (n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

- Tale indice misura la distanza euclidea pesata tra le celle della tabella osservata e quelle corrispondenti della tabella attesa.
- Il test è sempre unilaterale.

La variabilità dell'indice

- Il modello statistico che descrive la variabilità casuale (sotto H_0) dell'indice χ^2 è la distribuzione χ^2 con lo stesso numero di gradi di libertà della tabella, nel nostro caso (r-1)(c-1)=4.
- Le ascisse critiche sono 9,49 (al 5%) e 13,28 (al 1%).
- Il valore ottenuto χ^2=63,282 appartiene alle zone critiche, p osservato è praticamente uguale a 0, rifiutiamo l'ipotesi nulla.

Il daltonismo rosso-verde

Tabelle osservate

Dalton r-v	maschio	femmina	Tot
positivo	420	68	488
negativo	4900	4600	9500
Tot	5320	4668	9988

Dalton r-v	sordo	non sordo	Tot
positivo	45	7500	7545
negativo	450	90000	90450
Tot	495	97500	97995

Il daltonismo rosso-verde

Tabelle attese

Dalton r-v	maschio	femmina	Tot
positivo	259,93	228,07	488
negativo	5060,07	4439,93	9500
Tot	5320	4668	9988

Dalton r-v	sordo	non sordo	Tot
positivo	38,11	7506,89	7545
negativo	456,89	89993,11	90450
Tot	495	97500	97995

Risultati dei test

- Il valore del χ^2 è rispettivamente:
- χ^2=221,76
- χ^2=1,36
- In entrambi i casi il numero di gradi di libertà è 1
- Le ascisse critiche sono:**3,84 (al 5%) e 6,63 (al 1%).**
- Il valore p è nel primo caso maggiore del 10%, nel secondo praticamente zero.
- **Il daltonismo risulta significativamente legato al sesso, ma non alla sordità.**

Un altro esempio

Effettuiamo il test per lo studio del rischio associato alla presenza di un anticorpo nel sangue in relazione alla gravità di attacco cardiaco. Il numero di gradi di libertà è 2.

risulta χ^2 =10,54 e p<0,01

Rifiutiamo l'ipotesi di indipendenza.

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	85	40	125
++	125	95	220
+	150	145	295
Marginale	360	280	640

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	70,31	54,69	125
++	123,75	96,25	220
+	165,94	129,06	295
Marginale	360	280	640

Verifica sui fattori di rischio

MASCHI

Stenosi Aortica	Fumatore		totale
	SI	NO	
SI	37	25	62
NO	24	20	44
Totale	61	45	106

$$OR_M = \frac{(37)(20)}{(25)(24)} = 1{,}23$$

ODDS RATIO

- Supponiamo che i nostri dati - rappresentati da un campione di *n* soggetti - siano organizzati nella forma di una tabella di contingenza 2x2.

	esposti	non esposti	Totale
malattia	*a*	*b*	*a + b*
non malattia	*c*	*d*	*c + d*
totale	*a + c*	*b + d*	*n*

ODDS RATIO

- possiamo esprimere l'odds ratio come:

$$OR = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

- Questo è il rapporto del prodotto crociato dei valori nella tabella 2x2

Test sull'odds ratio

- Per effettuare il test sull'odds ratio occorre passare al logaritmo, la cui distribuzione si può approssimare con una normale, e usare come statistica test:
- T(a,b,c,d)=ln(OR)/s.e.(ln(OR))
- che è N(0,1), con:

$$s.e.(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

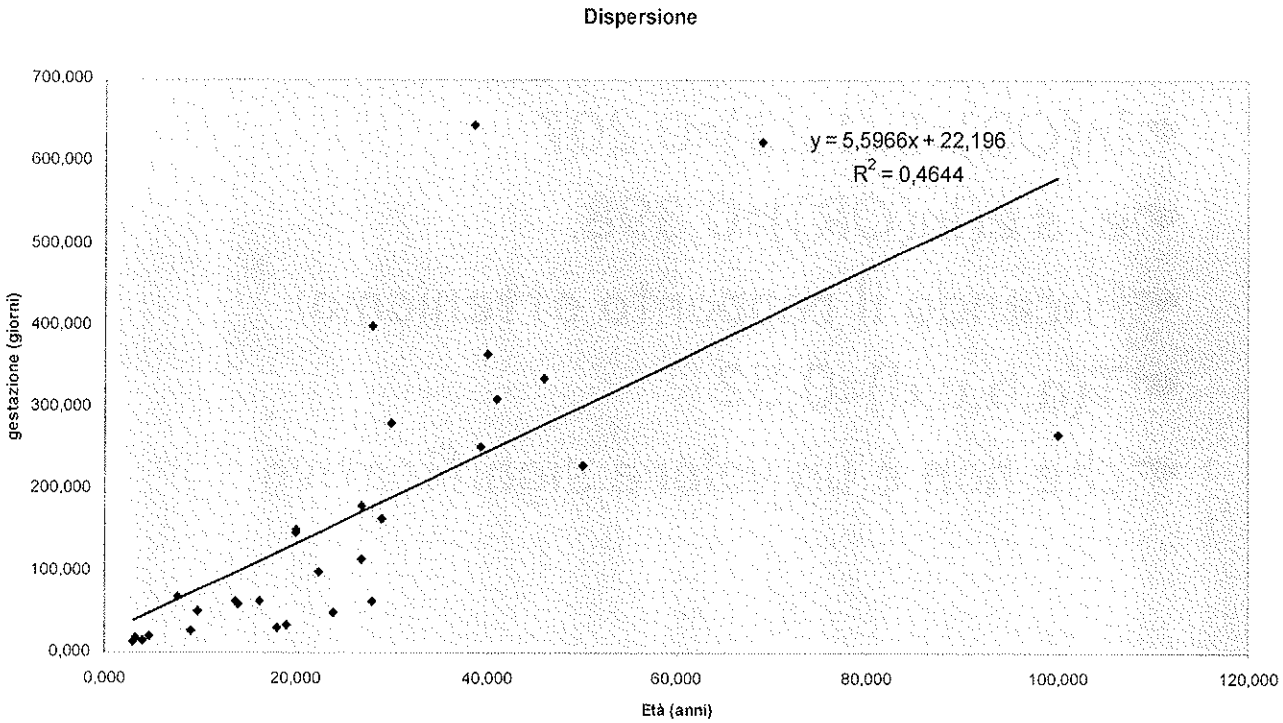
Esempio

- Nel nostro esempio si ha:
- ln(OR)=ln(1,23)=0,207
- e s.e(ln(OR))=0,398.
- Facendo il rapporto otteniamo:
- T(a,b,c,d)=0,52
- che non è un valore significativo.

Regressione lineare (esempio)

Species	maximum life span (years)	Gestation time (days)
African elephant	38,600	645,000
Arctic Fox	14,000	60,000
Asian elephant	69,000	624,000
Baboon	27,000	180,000
Big brown bat	19,000	35,000
Cat	28,000	63,000
Chimpanzee	50,000	230,000
Cow	30,000	281,000
Donkey	40,000	365,000
Giraffe	28,000	400,000
Goat	20,000	148,000
Golden hamster	3,900	16,000
Gorilla	39,300	252,000
Gray seal	41,000	310,000
Gray wolf	16,200	63,000
Ground squirrel	9,000	28,000
Guinea pig	7,600	68,000
Horse	46,000	336,000
Jaguar	22,400	100,000
Little brown bat	24,000	50,000
Man	100,000	267,000
Mouse	3,200	19,000
Pig	27,000	115,000
Rabbit	18,000	31,000
Raccoon	13,700	63,000
Rat	4,700	21,000
Red fox	9,800	52,000
Rhesus monkey	29,000	164,000
Sheep	20,000	151,000
Water opossum	3,000	14,000

Regressione lineare (esempio)



Il test sul coefficiente di regressione b

- E' possibile verificare se il coefficiente angolare di una retta di regressione è significativamente diverso da zero.
- Si ipotizza che la sua distribuzione sia normale e l'ipotesi nulla prevede che la sua media sia nulla.
- Occorre però stimare la sua deviazione standard per usare la tavola normale standardizzata

La varianza di b*

- Si dimostra che:

$$\sigma^2(b^*) = \frac{\sum(x_i - \bar{x})^2 \sigma^2(Y)}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sigma^2(Y)}{\sum(x_i - \bar{x})^2}$$

- e:

$$\sigma^2(Y) = \frac{\sum(y_i - a^* - b^* x_i)^2}{n-2} = \sum \frac{residui^2}{n-2}$$

- Perché, essendo basata su due parametri già stimati, ha n-2 gradi di libertà.

Tabella dei residui per i calcoli

Species	maximum life span (years)	gestation time (days)	Residui
African elephant	38,600	645,000	406,64
Arctic Fox	14,000	60,000	-40,6
Asian elephant	69,000	624,000	215,4
Baboon	27,000	180,000	6,6
Big brown bat	19,000	35,000	-93,6
Cat	28,000	63,000	-116
Chimpanzee	50,000	230,000	-72,2
Cow	30,000	281,000	90,8
Donkey	40,000	365,000	118,8
Giraffe	28,000	400,000	221
Goat	20,000	148,000	13,8
Golden hamster	3,900	16,000	-28,04
Gorilla	39,300	252,000	9,72
Gray seal	41,000	310,000	58,2
Gray wolf	16,200	63,000	-49,92
Ground squirrel	9,000	28,000	-44,6
Guinea pig	7,600	68,000	3,24
Horse	46,000	336,000	56,2
Jaguar	22,400	100,000	-47,64
Little brown bat	24,000	50,000	-106,6
Man	100,000	267,000	-315,2
Mouse	3,200	19,000	-21,12
Pig	27,000	115,000	-58,4
Rabbit	18,000	31,000	-92
Raccoon	13,700	63,000	-35,92
Rat	4,700	21,000	-27,52
Red fox	9,800	52,000	-25,08
Rhesus monkey	29,000	164,000	-20,6
Sheep	20,000	151,000	16,8
Water opossum	3,000	14,000	-25

Esempio

- otteniamo:
- b*=5,6, n-2=28
- σ(Y)=127,26

$$\sum (x_i - \bar{x})^2 = 12551{,}6$$

$$\sigma(b^*) = 0{,}01 \qquad b^* / \sigma\ (b^*) = 560$$

- Significativo a qualunque livello (p=0).

Tabelle di contingenza e test χ^2

Nell'analisi fin qui fatte (distribuzione di frequenza e analisi di regressione) abbiamo sempre preso in considerazione dati a carattere **Quantitativo**, o semmai qualora erano presenti variabili **Qualitative**, queste venivano ricodificate. In questo tipo di studio, invece, veranno utilizzati dati **Qualitativi** raggruppati in categorie determinate dalle classificazioni effettuate nella popolazione in esame. Supponiamo di avere un campione di N unità statistiche classificate rispetto a due criteri o variabili (aleatorie), una avente r categorie e l'altra c. Lo scopo è quello di determinare se questi due criteri di classificazione sono **indipendenti**, ossia se la distribuzione della popolazione campionaria è la medesima in entrambi i criteri.

Per permettere che ciò sia vero si costruisce, innanzitutto, la cosidetta **Tabella di Contingenza** ovvero una tabella a doppia entrata in cui le osservazioni relative a due variabili qualitative vengono rappresentate simultaneamente allo scopo di determinare l'esistenza di eventuali relazioni tra di esse. I valori presenti nelle celle della tabella sono ottenuti suddividendo l'intero spazio campionario delle N unità statistiche rispetto alle due variabili di interesse, quindi l'elemento generico n_{ik} rappresenta il numero di individui che verificano **simultaneamente** i criteri i e k. Nel caso generale ($r \times c$), essa si presenta in codesta configurazione:

k / i	1	2	3	.	.	c	$n_{i.}$
1	n_{11}	n_{12}	n_{13}	.	.	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	.	.	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	.	.	n_{3c}	$n_{3.}$
4	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
r	n_{r1}	n_{r2}	n_{r3}	.	.		$n_{r.}$
$n_{.k}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	.	.	$n_{.c}$	$n_{..}$

Vediamo cosa rappresentano tutti i simboli presenti all'interno di tale tabella.

Le due variabili si posizionano una in riga e l'altra in colonna suddivise rispettivamente in r e c categorie. Al variare dei pedici i (i = 1, 2, 3,....,r) e k (k = 1, 2, 3,.....,c) vengono rappresentate le frequenze congiunte assolute osservate nik della i-esima categoria della prima variabile e la k-sima categoria della seconda variabile.

Con $ni.$ viene indicato il numero totale di osservazioni nella i-esima categoria della variabile riga mentre con $n.k$ il numero totale di osservazioni della k-sima categoria della variabile colonna, sono noti come **totali marginali** e non sono altro che, rispettivamente, la somma di tutte le frequenze presenti sulla riga i e la somma di quelle presenti nella colonna k.

Con $n..$ viene indicato il numero totale delle osservazioni nel campione (=N) e corrisponde alle somme di tutte le frquenze di riga e di colonna.

Nel più semplice dei casi vengono coinvolte due variabili dicotomiche (due modalità: si/no or 0/1) le quali danno origine ad una tabella di contingenza 2×2.

Vediamo, in Excel, come sia possibile rappresentare le tabelle di contingenza.

Le normali versioni di Excel non hanno strumento alcuno che permetta, a meno che non si voglia ricostruirla manualmente, di determinare tale tabella; ma esiste una "aggiunta" che, qualora caricata, è in grado di effettuare analisi statistiche molto complesse tra le quali quanto da noi voluto.

Per rendere disponibile tale strumento, dobbiamo aprire il file **Phstat.zip** (disponibile alla pagina web http:/www.apogeonline.com/libri/00805/allegati/studenti/PHStat) e dopo aver letto le istruzioni contenute in **Phstat.doc** istalliamo tale componente aggiuntiva mediante il file **Setup.exe.** Tale aggiunta contiene delle macro che potrebbero presentare dei virus, quindi si consiglia di renderle attive solo in presenza di un programma antivirus tipo Norton o simili. Una volta effettuata l'istallazione nel modo corretto (che peraltro presenta non poche difficoltà), nella barra del Menù comparirà, tra **Dati** e **Finestra**, l'opzione **PHStat** che sarà selezionata allor quando dovremmo costruire le nostre tabelle di contingenza e le analisi successive.

Vediamo,ora, il caso di una tabella 2×2.

Nel file **diabete.xls** sono presenti informazioni su 144 (=N) indiani indigeni della tribù dei Navajos. Questo campione viene classificato rispetto alla presenza/assenza (si/no) di episodi di infarto al miocardio e in individui che sono diabetici o meno. Come vedete sono presenti dati qualitativi dicotomici. Vogliamo determinare la relativa tabella a doppia entrata; selezioniamo:

PHStat, Two-Way Tables &Charts...(tabelle e grafici a due vie - per due variabili) la finestra di dialogo che si presenta richiede Row Variable cell Range: in cui inseriamo la colonna degli infartati (le cui frequenze compariranno nelle righe della tabella) e Column Variable cell Range: in cui poniamo la colonna dei diabetici, viene richiesto, inoltre, se entrambi le colonne contengono i nominativi dei campi e noi lo selezioniamo (se precedentemente inglobati nei ranges richiesti). E' possibile inoltre opzionare il titolo della tabella che verrà costruita, e la presenza o meno di un vertical bar chart rappresentativo dei dati presenti nel file (il quale comparirà eventualmente nel foglio Side by Side Chart). Una volta dato Ok, PHStat crea due o tre nuovi fogli di lavoro, a seconda se includuiamo il bar chart o meno, nei quali sono presenti una copia del file dati (DataCopy) e la tabella di contingenza (nel foglio Two Way Table).

Riportiamo qui nel seguito la tabella di contingenza ottenuta mediante Excel:

Tabella a doppia entrata

	diabetici		
infartati	no	si	Totale complessivo
No	82	16	98
Si	37	9	46
Totale complessivo	119	25	**144**

Come si può notare dei 46 Navajos che hanno avuto esperienze di infarto 9 hanno "matchato" con la presenza di diabete e 37 con la sua assenza, dei 98 non infartati 16 sono stati appaiati coi diabetici e 82 con i non.

I totali marginali rispecchiano le somme tra righe ecolonne e il nostro $n_{..}$ coincide con la totalità campionaria N.

Da questa tabella possiamo ricavare informazioni sulla proporzione di indiani Navajos che hanno avuto infarti:

$$\frac{46}{144} \times 100 \cong 32\%$$

e per differenza (68%) la proporzione dei non infartati; e sulla proporzione di diabetici:

$$\frac{25}{144} \times 100 \cong 17{,}4\%$$

e sempre per differenza (82,6%) dei non diabetici.

Torniamo allo scopo di tale analisi. Ricordiamo che volevamo dimostrare l'indipendenza delle due variabili che costituiscono i criteri di classificazione della totalità campionaria (in questo caso: infartati e diabetici).

Lo strumento che permette di determinare ciò prende il nome di test del χ^2

Il test chi-quadro serve a stimare la possibilità di errare concludendo dalle osservazioni statistiche che le due variabili siano statisticamente dipendenti. Esso confronta le frequenze osservate in ogni cella della tabella di contingenza con le frequenze che ci aspetteremmo (**frequenze attese**) se non ci fosse associazione tra i contenuti che definiscono le righe e le colonne.

Quindi il test parte da una assunzione di base (**ipotesi nulla**) nella quale si ipotizza che le due proporzioni (infartati e diabetici) di individui (determinati dalle due variabili) siano uguali (quindi dipendenti). Per dimostrare l'indipendenza dovremmo rifiutare (in gergo "rigettare") l'ipotesi di ugualianza e quindi l'ipotesi nulla con un livello di significatività α (che in genere = 0,05).

Assumono particolare importanza i gradi di libertà (degree of freedom=df) che rappresentano numericamente la quantità di informazione disponibile nei dati che può essere usata per la stima della significatività del test. Il loro calcolo avviene mediante: **df=(*r*-1)*(c-1)** e nel caso di una tabella due per due, come il nostro, df=1.

Vediamo come Excel calcola il test del χ^2:

Selezioniamo **PHStat, c-Sample Tests, Chi-Square Test** la finestra di dialogo richiede il livello di significatività e noi digitiamo 0.05 (!!occhio!!: il punto e non la virgola), il numero di righe (=2) e di colonne (=2); è opzionabile il titolo dell'output che, ad esempio, possiamo chiamare test chi quadro per l'indipendenza.

questo punto, dopo O.K., Excel crea un nuovo foglio che denomina Hypothesis in cui devono essere inseriti i valori della tabella di contingenza ottenuti (in **Observed Frequencies**), al che immediatamente determina la tabella delle frequenze attese (**Expected Frequencies,** si noti che le marginali totali coincidono con i rispettivi della tabella di contingenza delle frequenze osservate), il valore critico (=3,841) per i df in questo caso per df=1 (presente in qualsiasi testo di statistica nella tabella valori critici di χ^2 al variare dei df), il valore del test chi-quadro (=0,228; sarà sempre ≥0) , il p-value che, per poter rifiutare l'ipotesi nulla di dipendenza, dovrà necessariamente essere < α. Nel nostro caso il p-value=0,63 che è > di α=0,05 e infatti Excel ci dice che non possiamo rifiutare l'ipotesi nulla di dipendenza. Questo è quanto appare del nostro esempio una volta svolte le operazioni descritte:

Test Chi-quadro per l'indipendenza

Observed Frequencies:

	Column variable		
Row variable	**C1**	**C2**	**Total**
R1	82	16	98
R2	37	9	46
Total	119	25	144

Expected Frequencies:

	Column variable		
Row variable	**C1**	**C2**	**Total**
R1	80,99	17,01	98
R2	38,01	7,99	46
Total	119	25	144

Level of Significance	0,05
Number of Rows	2
Number of Columns	2
Degrees of Freedom	1
Critical Value	**3,841455**
Chi-Square Test Statistic	**0,228875**
***p*-Value**	**0,63236**
Do not reject the null hypothesis	

Andrebbe corretto con il test McNemar ma non mi sembra il caso, potrebbe essere un esempio in cui il chi –quadro fallisce.

Oltre ad essere relativamente semplice da stimare il test del chi-quadro si presta alla generalizzazione nel confronto tra tre o più proporzioni di popolazione.

Estendiamo, quindi, tale analisi al coso in cui la tabella di contingenza relativa non sarà più una 2×2 bensì una $r \times c$.

Apriamo il file **certificati_morte.xls** (non proprio una cosa allegra!) in cui sono presenti dati quantitativi relativi ad uno studio che investiga sull'accuratezza dei certificati di morte redatti, in due ospedali differenti, su 575 decessi.

I risultati delle autopsie, effettuate su tali decessi, vengono comparate con le cause di morte presenti nei cerificati.

I due ospedali vengono codificati con A e B, mentre lo stato dei certificati di morte può essere: **esatto** ovvero le cause di morte accertate nelle autopsie corrispondono a quelle presenti nei certificati; **impreciso** ovvero mancanti di informazioni o contenenti errori, non particolarmente gravi da richiedere una nuova stesura degli stessi; **errato** ovvero tali da giustifiacare una nuova stesura delle cause di morte.

Quindi la relativa tabella di contingenza sarà una 2×3.

Selezioniamo **PHStat, Two-Way tables &Charts...** nelle cella di riga richiesta inseriamo la colonna ospedale e in quella di colonna lo stato e, una volta opzionato il nome dell'output, dando o.k. appare:

Tabella di contingenza 2x3

	Stato			
Ospedale	errato	Esatto	impreciso	Totale complessivo
A	54	157	18	229
B	34	268	44	346
Totale complessivo	88	425	62	**575**

(si osservi che il nome delle categorie appare sempre in ordine alfabetico e non come vengono inseriti).

Quindi dei 575 certificati considerati: 425 sono corretti, 62 mancano di precisione ma non hanno bisogno di esere ristilati e 88 invece devono essere rifatti.

Ciò che vorremmo determinare è se i risultati osservati suggeriscono una differenza nelle pratiche di compilazione dei certificati di morte nei due ospedali in esame.

Testiamo, quindi, la non associazione tra ospedale e lo stato dei certificati di morte o in altre parole (ipotesi nulla) che le proporzioni, all'interno di ogni stato, siano identiche.

Procediamo con il test χ^2:

selezionamo **PHStat → c-Sample Test → Chi-Square Test** e nella finestra di dialogo successiva 0.05 nel livello di significatività, 2 in numero di righe e 3 nelle colonne, titolo dell'output e O.K.

Verrà creato il foglio Hypothesis, e una volta inseriti i valori all'interno della tabella delle frequenze osservate comparirà:

Test del Chi –Quadro

Observed Frequencies:

	Column variable			
Row variable	C1	C2	C3	Total
R1	54	157	18	229
R2	34	268	44	346
Total	88	425	62	575

Expected Frequencies:

	Column variable			
Row variable	C1	C2	C3	Total
R1	35,05	169,26	24,69	229
R2	52,95	255,74	37,31	346
Total	88	425	62	575

Level of Significance	0,05
Number of Rows	2
Number of Columns	3
Degrees of Freedom	2
Critical Value	**5,991476**
Chi-Square Test Statistic	**21,55006**
***p*-Value**	**2,12E-05**
Reject the null hypothesis	

Allora come si può notare df=2, il valore critico del chi-quadro con tali gradi di libertà e a un livello di significatività del 5% è 5,99 il test del chi-quadro è 21,55, infine $p<\alpha$ quindi possiamo rifiutare l'ipotesi nulla e concludere che le proporzioni dei certificati di morte non sono identiche per i tre diversi stati o equivalentemente che non c'è associazione tra ospedale e stato. Abbiamo dimostrato l'indipendenza tra le variabili ospedale e stato.