

Corso Integrato di **laboratorio di informatica,** **statistica e analisi dei dati**

Dr C. Meneghini

Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84
meneghini@fis.uniroma3.it
st. - 83 -
tel.: 06 5733 7217

<http://www.fis.uniroma3.it/~meneghini>

Prof. M. A. Ricci

Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84
riccim@fis.uniroma3.it
st. - 145 -
tel.: 06 5733 7226

<http://scienze.fis.uniroma3.it>

C. Meneghini - Laboratorio di Informatica

1

Questionario

Nome e Cognome (leggibile)	Esperienza		Tipo di utilizzo
	Uso del PC	Sistemi operativi	
1			Lavoro, Studio, gioco, altro

Livello di esperienza nell'uso del calcolatore:

Buona,
media,
nulla

Conoscenza di sistemi operativi:

Windows (98,2000,XP, Vista...)
Linux
altro

Esperienza di applicativi... calcolo

Excel, origin, kaleidagraph, etc...

Utilizzo del PC:

gioco, lavoro, comunicazione, etc...

C. Meneghini - Laboratorio di Informatica

4

Introduzione

Organizzazione del corso

Piano delle lezioni

&

... stimolare interesse e curiosità per gli argomenti

C. Meneghini - Laboratorio di Informatica

2

Laboratorio di informatica, statistica e analisi dei dati sperimentali.

Scopo del corso

Il metodo scientifico è lo studio **sistematico, controllato, empirico** e **critico** di ipotesi formulate sulle **relazioni** tra vari **fenomeni**

Il metodo scientifico si basa su osservazioni sperimentali e richiede la **misura**, l'**analisi** e l'**interpretazione** di **dati sperimentali**.

Per un'interpretazione corretta dei dati sperimentali è essenziale l'**analisi rigorosa** del processo di **misura** e dei **dati sperimentali** basata su **metodi statistici** (oggettivi).

Il PC (elaboratore elettronico) è uno strumento essenziale per l'**elaborazione** e la **presentazione** efficiente dei dati e dei risultati sperimentali.

Organizzazione del corso: **II semestre**

Teoria: Mar. 9-11 / Giov. 11-13 / Ven. 15-16 Aula 1 (Biologia-V* Marconi)

24 h. Informatica e statistica (prof. C. Meneghini)

32 h. statistica ed analisi dati (prof. M. A. Ricci)

~ 56 h

Esercitazioni al calcolatore:

~ 30 h

Aula di informatica Mart. (A) Giov.(B) 14-17.30

a partire da Mar. 31-Mar.2009 Aula 1 (Biologia-Vle Marconi)

Studio guidato:

~ 20 h

Valutazione

3 **esoneri**: AAA: Idoneo AAB: orale. Altrimenti: deve sostenere l'esame
prova pratica: A: idoneo; B: orale; C: non idoneo

IMPORTANTE possono sostenere gli esoneri solo gli studenti in corso o che seguano regolarmente il corso.

IMPORTANTE sessioni d'esame straordinarie solo in casi straordinari e solo per chi, essendosi presentato ad una sessione ordinaria, non abbia passato l'esame

Modulo: **Laboratorio di informatica**

Uso dell'elaboratore elettronico (PC) per il calcolo e applicazioni scientifiche.

In particolare:

- elaborazione di dati sperimentali,
- analisi statistica di dati sperimentali

Il programma del corso **non riguarda** l'uso elementare del computer. I seguenti argomenti saranno considerati noti, sia pur ad un livello elementare:

- uso di software per elaboratori di testo (word processor);
- uso software di presentazioni elettroniche (es.: powerpoint);
- navigazione in internet;
- uso della posta elettronica;
- etc...

Nota: I programmi dei corsi ECDL coprono solo parzialmente gli argomenti trattati nel corso. Il possesso di una **patente ECDL** non può sostituire il corso (neanche in parte).

- Modulo di informatica - teoria

- Il PC che conosciamo: S.O. Windows, informazioni sull'uso e la configurazione. Nozioni elementari di **sicurezza**. Definizione e classificazione dei **Malware**. Strumenti di amministrazione. Norme di comportamento in rete.
- Gli Algoritmi: un algoritmo è un procedimento che consente di ottenere un risultato atteso eseguendo, in un determinato ordine, un insieme di passi (azioni/operazioni/etc...) semplici. Elementi costitutivi, struttura degli algoritmi, nozioni di progettazione di un algoritmo.
- Codifica delle informazioni: Definizione e unità di misura dell'informazione. Codice binario, codifica di numeri, caratteri, suoni e immagini. **Errori nel calcolo numerico**: overflow, underflow, troncamento.
- Elaboratori: cenni storici, macchine analogiche e digitali. Macchina di von Neumann e funzionamento del PC. Hardware e software.
- Calcolo: Logica e matematica, le porte logiche. Logica e calcolo. Calcolo numerico

- Modulo di informatica - esercitazioni -

- Nozioni di base (ma non troppo) sull'utilizzo del PC in ambiente Windows.

Applicativi:

- Gnuplot: Un semplice programma per il calcolo, la grafica di funzioni e la presentazione dei dati.
- Uso del foglio elettronico (Excel) per calcolo ed elaborazione dati.
- Analisi statistica di dati

Perchè programmare?

Programmare = istruire il computer affinché faccia che vi serve in modo efficiente.

Una volta istruito a dovere un computer diventa prezioso per migliorare la produttività. Progettare e implementare un algoritmo ci obbliga a pensare in modo schematico e rigoroso.

- 1) Il computer è una macchina passiva, senza iniziativa e non fa nessuno sforzo per capire quello che intendete ... 
- 2) I computers hanno circuiti speciali che sentono la fretta, l'emozione e il panico. Probabilmente il sistema operativo (windows in particolare) ha istruzioni che gli permettono di andare in crash nel momento meno opportuno. 
- 3) i computers riconoscono l'operatore e funzionano solo quando è qualcun'altro a scrivere! 
- 4) Il computer può fare calcoli molto complessi ma, in alcuni casi, sbagliare una semplice divisione ($3/2 \neq 2$)! Probabilmente lo fa per dispetto! 
- 5) Un PC, nonostante i soldi spesi, la potenza, tutte le estensioni che gli avete comprato, non è in grado di fare proprio quell'operazione semplicissima che a mano ci mettereste un attimo (es. digitare (r) o (c))... sempre per dispetto?!

Il computer non fa nessuno sforzo per capire quello che intendete dire.

Scrivere un programma costringe a:

- pensare in modo sistematico e rigoroso
- applicare il *metodo scientifico*

Il **metodo scientifico** è lo studio **sistematico**, **controllato**, **empirico** e **critico** di **ipotesi** formulate sulle **relazioni** supposte tra vari **fenomeni** (Frédéric Kerlinger)

Posso istruire un PC a capire quello che penso?

www.20q.net

Materiale didattico e testi consigliati

Nonostante sia difficile trovare un testo che riassume e segua il programma del corso, i **lucidi** e gli **appunti** delle lezioni rappresentano un **sunto schematico** degli argomenti trattati a lezione e durante le esercitazioni. Pertanto sono una guida utile per seguire gli argomenti trattati ma **non possono sostituire la presenza alle lezioni ed esercitazioni**. Sono inoltre ben lunghi dal trattare in modo esaustivo gli argomenti, possono presentare errori, sviste, inesattezze. Pertanto è bene **integrarli con i testi consigliati o altri equivalenti**.

A Informazioni, appunti e lucidi delle lezioni (modulo di statistica):

<http://www.fis.uniroma3.it/~meneghini>

A M. Middleton: *Analisi statistica con EXCEL* – APOGEO:

uso del foglio elettronico per l'analisi statistica e la rappresentazione di dati
www.apogeeonline.com/libri/88-503-2475-8/scheda

B A. Camusi: *Metodi statistici per la sperimentazione biologica* Zanichelli

Statistica descrittiva, probabilità, test statistici, distribuzioni, confronti, regressione lineare etc...

C Stanton A. Glantz: *Statistica per discipline biomediche* McGraw-Hill

Analisi statistica avanzata. Utile soprattutto per il futuro!

Per chi non ha alcuna dimestichezza con il PC

0 ECDL – la guida alla patente europea del computer (W2K) McGraw-Hill
(o altri testi equivalenti, guide in internet etc...)

La didattica è fatta di interazioni
docente/studente &
studente/studente

Scopo di un corso
universitario è:

~~testare le proprie conoscenze~~

superare i propri limiti di conoscenza

domande/discussioni/ricieste di chiarimenti/dubbi:

chiariscono il grado di conoscenza della classe e migliorano il profitto del gruppo, soprattutto se espressi pubblicamente... quindi **fate domande!**

Dr C. Meneghini
Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84
st. - 83 -
tel.: 06 5733 7217
meneghini@fis.uniroma3.it

Controllate che la vostra casella di
posta abbia spazio per ricevere la
risposta!!!!

to: meneghini@fis.uniroma3.it
from: aaa@bbb.ccc.ddd
subject: informazioni, corso laboratorio di informatica, Biologia

Salve/ buongiorno/ Gentile Prof./ etc...

sono Pinco Pallino, riguardo al corso (2006/07) avrei bisogno
di... / vorrei sapere se.... / mi sa dire quando / etc... ?

Grazie / arrivederci / cordialmente / distinti saluti/ a presto / etc....

Pinco Pallino

Informazione, Statistica e democrazia

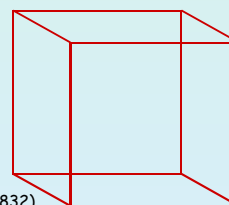
Statistica: da **statistico** = *che concerne il governo e la vita dello
stato, che riguarda lo studio di fenomeni interessanti la collettività*

La statistica ufficiale in uno stato democratico è un
servizio pubblico essenziale per garantire i diritti di
egualianza (e di scelta libera) a tutti i cittadini.

La diffusione dell'informazione statistica è strettamente
associata alla diffusione delle democrazie (dopo 1830)

Domande?

L'illusione della certezza: processo mentale ed emotivo
che ci permette di risolvere uno stato di incertezza



cubo di Necker (1832)

La propensione a ignorare l'incertezza è innata

L'illusione della certezza guida la scelta soggettiva, l'alfabetizzazione
statistica è requisito essenziale per una libertà di scelta consapevole,
ovvero per una **valutazione del rischio**

Perchè la statistica? Prendere una decisione richiede informazione

G. Barbieri: *Statistica Ufficiale, patrimonio di tutti* (2004)



Giovanni Buridano ~1300

La fiducia nel dato statistico è ... bassa

LA STATISTICA (Trilussa)

Sai ched'è la statistica? E' 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego, da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perchè c'è un antro che se ne magna due.

Scarsa fiducia nelle informazioni statistiche

La decisione richiede informazione.

Più l'informazione è completa più la decisione è semplice:

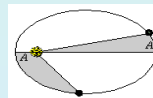
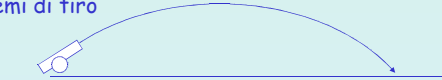
Fenomeni deterministici: informazione esatta e completa

Fenomeni non deterministici: informazione incompleta e incerta

previsioni del tempo	diffusione di un virus
orario dell'autobus	Andamento dei prezzi
traiettoria di una pallina	Mercato del lavoro
temperatura di ebollizione dell'acqua	Stato delle risorse energetiche
voto di laurea	evoluzione di una popolazione
	test del DNA

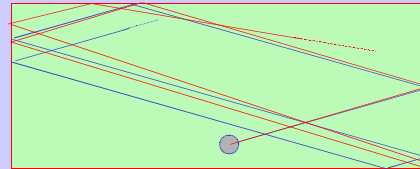
Quanto sono deterministici i fenomeni?

Problemi di tiro



leggi di Keplero:
orbite planetarie

Il biliardo



Illusione della certezza

Il desiderio di sicurezza porta spesso a "mentire"

- ✓ governi su **encefalopatia spongiforme bovina**
- ✓ medici su **pericoli sottostimati di analisi cliniche**
- ✓ case farmaceutiche su **effetti collaterali farmaci**
- ✓ ditte sigarette su **pericolo da fumo**

...a questo mondo non c'è niente di certo, a parte la morte e... le tasse (Benjamin Franklin)

Il test del DNA è sicuro! (errore umano?!)

- ✓ c'è una concordanza casuale su 1000000
- ✓ su 10^6 persone una sarà concordante

Sul totale della popolazione italiana ($5.7 \cdot 10^7$ abitanti) ci saranno 60 coincidenze

Statistica: analisi quantitativa dei fenomeni collettivi che hanno attitudine a variare, allo scopo di descriverli e di individuare le leggi o i modelli che permettono di spiegarli e di prevederli.

(dizionario Garzanti)

1 mole gas = 1022 atomi

(i) $x_i(t) = x_i^{(0)} + v_{x_i}^{(0)} t + a_{x_i} t^2 / 2$

$$x_i(t) = x_i^{(0)} + v_{x_i}^{(0)} t + a_{x_i} t^2 / 2$$

$$y_i(t) = y_i^{(0)} + v_{y_i}^{(0)} t + a_{y_i} t^2 / 2 \quad i = 1, 2, 3, \dots, 10$$

$$z_i(t) = z_i^{(0)} + v_{z_i}^{(0)} t + a_{z_i} t^2 / 2$$

310 equazioni, 610 parametri iniziali

Legge dei gas perfetti

$$PV = RT$$

Non so cosa faccia la singola particella ma conosco l'energia cinetica media per particella e quindi la velocità media (T). So che aumentando la temperatura aumenta la pressione (quantità media di moto trasferita alle pareti) etc...

Fenomeni deterministici

$$F = ma$$

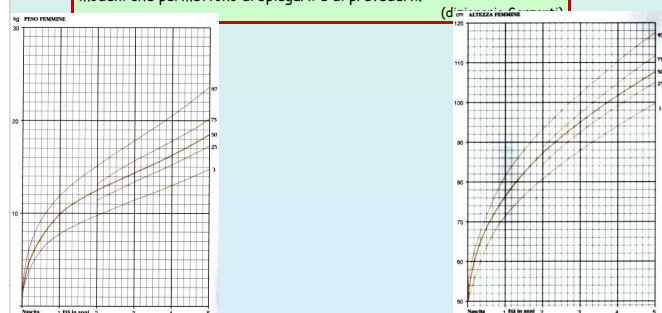
moto uniformemente accelerato

$$x(t) = x_0 + v_0 t + (1/2) a t^2$$

Noti a (accelerazione), x_0 (posizione iniziale) e v_0 (velocità iniziale) posso prevedere la posizione del corpo all'istante t (futuro).

La previsione sarà tanto più precisa quanto meglio conosco a , x_0 e v_0 .

Statistica: analisi quantitativa dei fenomeni collettivi che hanno attitudine a variare, allo scopo di descriverli e di individuare le leggi o i modelli che permettono di spiegarli e di prevederli.

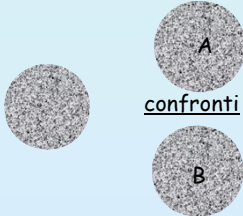


25%-75% percentile: perfetto stato di normalità
 75%-90% percentile: livello "borderline" o normo-superiori o normo-inferiori
 90%-97% e 10%-3% percentile: patologia lieve o comunque ad alto rischio
 97% e 3% percentile: patologia conclamata, in riferimento alla statura, un'altezza inferiore al 3° percentile (nanismo) non è necessariamente patologica; per questo motivo si richiede una visita specialistica approfondita (valutazione dell'età ossea, della velocità di crescita, esami ematochimici, ecc.); analogo discorso per stature superiori al 97° percentile.

Statistica: analisi quantitativa dei fenomeni collettivi che hanno attitudine a variare, allo scopo di descriverli e di individuare le leggi o i modelli che permettono di spiegarli e di prevederli.

(dizionario Garzanti)

Coltura batterica:



Statistica descrittiva:

- Tasso di crescita,
- densità di batteri
- mortalità,
- vita media, ...

Analisi statistica:

quanto sono affidabili/significative le analogie e le differenze tra risultati di diversi esperimenti?

C. Meneghini - Laboratorio di Inform.

25

LA STATISTICA

1) cerca di tenere sotto controllo l'incertezza che deriva dal trattare fenomeni collettivi (molti individui/molti eventi)

2) Semplificare la complessità mediante modelli basati sul calcolo delle probabilità. Conservare l'informazione riducendo la molteplicità dei fenomeni

3) Calcolo quantitativo del rischio

C. Meneghini - Laboratorio di Inform.

28

La **statistica descrittiva** studia i criteri di rilevazione, *classificazione* e di *sintesi* delle informazioni relative a una popolazione oggetto di studio. La statistica descrittiva raccoglie le informazioni sulla popolazione o su una parte di essa (**campione**) in distribuzioni e sintetizza i risultati attraverso **famiglie** di indici: **valori medi**, **variabilità**, **forma**, rapporti statistici, relazioni statistiche etc...

La **statistica inferenziale** è l'insieme dei procedimenti di generalizzazione che, avvalendosi dei metodi del calcolo delle probabilità, consentono di estendere all'intera popolazione le informazioni fornite da un campione.

La **test statistici** è una **regola di decisione**: permette di verificare determinate ipotesi sui parametri. Consentono di quantificare il rischio di sbagliare affermando la validità o meno di determinate ipotesi

C. Meneghini - Laboratorio di Inform.

26

Definizioni di Probabilità

a) Frequenza relativa di un evento all'interno di una classe di riferimento, calcolata per un numero molto grande (infinito) di osservazioni.

Es.: Su 100 000 prove ottengo 95750 risultati positivi, quale è la probabilità di un risultato positivo per il 100 001 esperimento?

b) Propensione è legata alle caratteristiche fisiche.

Es.: la probabilità che esca testa lanciando una moneta è 0.5, la probabilità che esca 6 lanciando un dado è 1/6, etc...

c) Probabilità soggettiva di un evento è il grado di fiducia che un individuo attribuisce al verificarsi dell'evento.

Lancio una moneta e ottengo 10 volte di seguito testa... quale è la probabilità che io stia barando?

Schiaccio il pedale del freno, quale è la pb. che la macchina rallenti?

Tipico nel passare dai modelli alla realtà: In una simulazione ottengo una sopravvivenza pari al 95 % per un dato comportamento. Quale è la probabilità che la popolazione reale sopravviva?...

Fenomeni collettivi:

moto delle particelle in un gas, previsioni meteorologiche, andamento dei prezzi, evoluzione di un ecosistema, etc...

LA STATISTICA

riduce la fonte di incertezza creando modelli (basati sul calcolo delle probabilità) in grado di conservare l'informazione quantitativa riducendo la molteplicità dei fenomeni. Riduce l'incertezza e permette di quantificare il rischio

Rischio

Il rischio indica la probabilità del verificarsi di un evento

INCERTEZZA

Un risultato è incerto quando non è possibile, in base ai dati a disposizione, assegnare un valore di probabilità ai diversi esiti possibili

C. Meneghini - Laboratorio di Inform.

27

In uno stagno ci sono NT pesci.

Ne pesco No, li segno e li rimetto in libertà.

Pesco di nuovo N1 pesci e vedo che ci sono N' pesci già segnati.

Posso stimare il numero totale di pesci?

Frequenza relativa: N_r/N_t

$\frac{N_o}{N_T}$ = probabilità di pescare un pesce segnato (propensione)

$\frac{N}{N_1}$ = frequenza osservata: la migliore stima che ho a disposizione per stimare la probabilità di pescare un pesce segnato

$$\frac{N_o}{N_T} \sim \frac{N}{N_1} \quad N_T \sim \frac{N_1}{N} N_o$$

NT una stima del numero di pesci nel lago!

Posso migliorare questa stima?
Come?

30

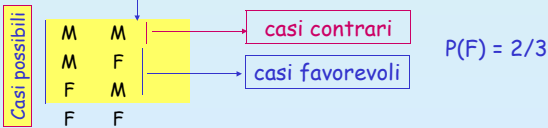
b) Propensione (Numero di casi favorevoli)/(numero di casi possibili)

A proposito della probabilità di eventi complessi il giudizio del senso comune può essere totalmente sbagliato!

L'elaborazione dei dati statistici è semplificata (software dedicati)

L'interpretazione dei risultati statistici è complessa e può essere fonte di errori grossolani e spesso gravi

In una popolazione la probabilità di maschi e la probabilità di femmine è eguale (1/2). Una famiglia ha esattamente due figli, se almeno uno dei due è maschio, quale è la probabilità che l'altro sia femmina?



C. Meneghini - Laboratorio di Inform.

31

Alcune proprietà delle probabilità

Siano A e B due eventi indipendenti

$$P(A \text{ oppure } B) = P(A) + P(B)$$

$$P(A \text{ e } B) = P(A) * P(B)$$

$$P(\text{non } A) = 1 - P(A)$$

Es.: lancio un dado, la probabilità di ottenere 1 oppure 2 è:

$$P(1 \text{ oppure } 2) = 1/6 + 1/6 = 1/3$$

Es.: lancio due monete, la probabilità di ottenere due croci è:

$$P(X \text{ e } X) = 1/2 * 1/2 = 1/4$$

Es.: lancio due monete, la probabilità di non ottenere due croci è:

$$P(\text{not}(X \text{ e } X)) = 1 - 1/4 = 3/4$$

C. Meneghini - Laboratorio di Inform.

34

In una popolazione tutte le famiglie hanno esattamente due figli. La probabilità di maschi e la probabilità di femmine è eguale (1/2). Incontro un papà con un figlio maschio. Quale è la probabilità che abbia una sorella?

I	II
M	M
M	M
F	M
M	F

$P(F) = 2/4 = 50\%$

Non è molto diverso dal lancio di due monete, se ottengo testa su una, quale è la pb. che l'altra sia croce?

Quale è la probabilità che in un gruppo di N studenti almeno due di essi abbiano il compleanno lo stesso giorno (trascuriamo gli anni bisestili)?

Data una persona, la pb. che una seconda persona (N=2) non abbia il compleanno lo stesso giorno è:

$$N = 2 \quad Pb(\text{non stesso compleanno}) = 364/365$$

$$N = 3 \quad Pb(\text{non stesso compleanno}) = 363/365 * 364/365$$

$$Pb(N) = \frac{364}{365} \frac{363}{365} \frac{362}{365} \dots \frac{365 + 1 - N}{365} = \frac{365!}{365^N (365 - N)!}$$

$$Pb(\text{stesso}) = 1 - P(N)$$

$$23 \text{ persone} \dots Pb(\text{stesso}) \sim 50\%$$

C. Meneghini - Laboratorio di Inform.

35

Un'urna contiene 2 palline bianche e 2 palline nere. Ne estraggo una a caso, quale è la probabilità che sia bianca?

Un'urna contiene 2 palline bianche e 2 palline nere. Ne estraggo una a caso e, senza guardarla ne estraggo una seconda. Se la seconda è bianca, quale è la probabilità che la prima sia anch'essa bianca?

BB	NN
BN	BN
NB	BN
BN	NB
NB	NB
NN	BB

C. Meneghini - Laboratorio di Inform.

33

Esami clinici

	positivo	negativo
malato	sensibilità (veri positivi)	falsi negativi
sano	falsi positivi	specificità (veri negativi)

Efficienza:
Sensibilità + specificità

$$Pb(\text{falsi negativi}) = 1 - pb(\text{veri positivi})$$

$$Pb(\text{falsi positivi}) = 1 - pb(\text{veri negativi})$$

Test di gravidanza:

falso positivo: test positivo ma non è incinta!

falso negativo: test negativo ma è incinta! ⚠

Test HIV:

falso positivo: test positivo ma non è malato!

falso negativo: test negativo ma è malato! ⚠

falsi negativi = maggiore danno

5

Un po' di logica

A e B sono due variabili logiche, possono assumere solo valore Vero (V) o Falso (F)

L'affermazione:

Se A è vero allora B è vero

$A \Rightarrow B$

A vero è una condizione necessaria ma non sufficiente per B vero

A	B	$A \Rightarrow B$
V	V	V
V	F	F
F	V	V
F	F	V

Implica che:

Se B è Falso allora A è Falso

$\bar{B} \Rightarrow \bar{A}$

Se B è Vero allora A può essere vero o falso !

Se A è Falso allora B può essere vero o falso !

Se il paziente è malato (A) allora il test è positivo (B)



Se il test è positivo (B) allora il paziente è malato (A)

Laboratorio di Inform.

37

Occorre saper calcolare

Quando è vantaggioso fare una mammografia?

Tra i 40 e 50 anni la probabilità che una donna sviluppi il cancro al seno è del 0.8% (8‰)

Se una donna ha un cancro al seno la probabilità che il mammogramma sia positivo è del 90%

Se non ha il cancro al seno c'è comunque una probabilità del 7% che il mammogramma risulti positivo lo stesso

Una donna, con età tra 40 e 50 anni, fa un mammogramma ed il risultato è positivo

Quanto è probabile che abbia il cancro al seno?

	positivo	negativo	Pop.
malato	sensibilità 90%	10%	8%
sano	7%	specificità 93%	92%

	pos.	neg.	Pop.
M	$P(P M)$	$P(N M)$	$P(M)$
S	$P(P S)$	$P(N S)$	$P(S)$

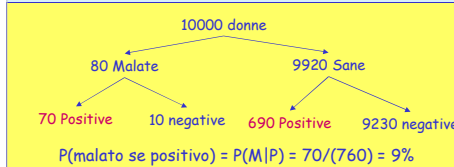
$$P(\text{Malato}) = P(M) = 0.008$$

$$P(\text{positivo se malato}) = P(P|M) = 0.9$$

$$P(\text{positivo se sano}) = P(P|S) = 0.07$$

$$P(\text{positivo se malato}) = P(P|M) = 90\%$$

$$P(M|P) = 9\%$$



40

Esami clinici

	positivo	negativo
malato	sensibilità (veri positivi)	falsi negativi
sano	falsi positivi	specificità (veri negativi)

Sensibilità 100%



malato $\Rightarrow$ Positivo

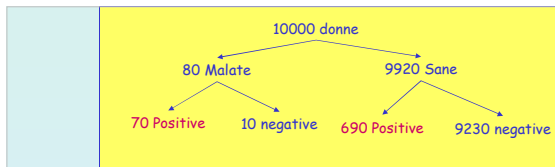
Stato	risultato
Malato	Positivo
Malato	Negativo
Sano	Positivo
Sano	Negativo

Se il paziente è malato (A) allora il test è positivo (B)

Mi serve altra informazione per conoscere lo stato del paziente

C. Meneghini

38



$$P(\text{malato se positivo}) = 70/760 = 0.09 = 9\%$$

$$P(\text{sano se positivo}) = 690/760 = 1.9\% = 91\%$$

$$P(\text{malato se negativo}) = 10/9240 = 0.001 = 0.1\%$$

$$P(\text{sano se negativo}) = 9230/9240 = 0.999 = 99.9\%$$

Teorema di Bayes

$$p(\text{malattia}) = p(m) = 0.008$$

$$p(\text{non_malattia}) = p(s) = 1 - 0.008 = 0.992$$

$$p(\text{positivo}|\text{malattia}) = p(p|m) = 0.90$$

$$p(\text{positivo}|\text{non_malattia}) = p(p|s) = 0.07$$

$$p(m|p) = \frac{p(p|m) p(m)}{p(p|m) p(m) + p(p|s) p(s)}$$

$$= \frac{0.90 \cdot 0.008}{0.90 \cdot 0.008 + 0.07 \cdot 0.992} = \frac{0.0072}{0.0766} = 0.0939$$

$$p(\text{malattia}|\text{positivo}) = 9\%$$

	positivo	neg.	Pop.
malato	$P(P M)$ 90%	10%	$P(M)$ 8%
sano	$P(P S)$ 7%	93%	$P(S)$ 92%

Meneghini - Laboratorio di Inform.

41

Occorre saper calcolare

Quando è vantaggioso fare una mammografia?

Tra i 40 e 50 anni la probabilità che una donna sviluppi il cancro al seno è del 0.8% (8‰)

Se una donna ha un cancro al seno la probabilità che il mammogramma sia positivo è del 90%

Se non ha il cancro al seno c'è comunque una probabilità del 7% che il mammogramma risulti positivo lo stesso

Una donna, con età tra 40 e 50 anni, fa un mammogramma ed il risultato è positivo

Quanto è probabile che abbia il cancro al seno?

	positivo	negativo	Pop.
malato	sensibilità 90%	10%	8%
sano	7%	specificità 93%	92%

	pos.	neg.	Pop.
M	$P(P M)$	$P(N M)$	$P(M)$
S	$P(P S)$	$P(N S)$	$P(S)$

$$P(\text{positivo se malato}) = P(P|M) = 0.9 = 90\%$$

$B \Leftarrow A$

$$P(\text{malato se positivo}) = P(M|P) = 0.9 = 90\%$$

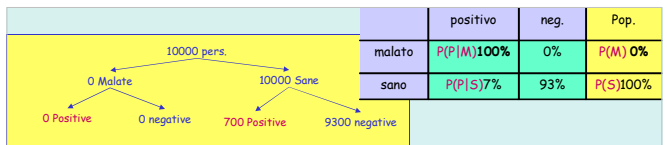
$$P(\text{negativo se sano}) = P(N|S) = 0.93 = 93\%$$

$B \Leftarrow A$

$$P(\text{sano se negativo}) = P(N|S) = 0.93 = 93\%$$

C. Meneghini - Laboratorio di Inform.

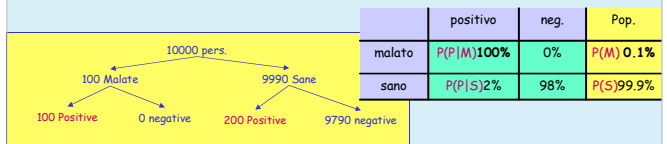
39



$$P(M) = 0$$

$$P(P|M) = 100\% = \text{sensibilità assoluta}$$

$$P(\text{malato se positivo}) = 0$$



$$P(M) = 0.1\%$$

$$P(P|M) = 100\% = \text{sensibilità assoluta}$$

$$P(\text{malato se positivo}) = 50\%$$

Nota: per la classe "persone a rischio" la $P(M)$ è molto maggiore!

$$p(m|p) = \frac{p(p|m) p(m)}{p(p|m) p(m) + p(p|s) p(s)}$$

Es. Si sa che lo 0.5% dei soggetti di una città è AIDS. Si sa che i tests diagnostici danno diagnosi corretta nell'80% dei sani e nel 98% dei malati. Qual è la probabilità di esser sano posto che il test sia stato positivo?

Esercitiamoci

14-01-07, 20:47

aleph
Utente Junior

Re: Un test sul test

Se ho capito bene, su 10.000 persone devo fare 10.000 test, i cui risultati sono 9.998 negativi e 2 positivi. Allora è bene spiegare ai due eventuali che sarebbe il caso di ripetere il test. Almeno per loro due. Comunque mi sembra che la riflessione di questo medico sia particolarmente acuta. Dimostra ancora una volta quanto sia importante farsi una cultura matematica (almeno di base) per sopravvivere. No?

L'elaborazione dei dati statistici è semplificata (software dedicati)

L'interpretazione dei risultati statistici è complessa e può essere fonte di errori grossolani e spesso gravi

Risultati di Test di ammissione a corsi universitari

	M	F
Ammessi	1198	557
Respinti	1493	1278

	M	F
Ammessi	44%	30%
Respinti	56%	70%

Conclusione:

Le donne sono discriminate nei test di ammissione!

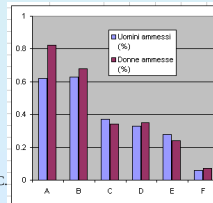
L'elaborazione dei dati statistici è semplificata (software dedicati)

L'interpretazione dei risultati statistici è complessa e può essere fonte di errori grossolani e spesso gravi

Risultati dei Test di ammissione: dettagli per dipartimento

	Dip. A		Dip. B		Dip. C		Dip. D		Dip. E		Dip. F	
	M	F	M	F	M	F	M	F	M	F	M	F
Amm.	512	89	353	17	120	202	138	131	53	94	22	24
Resp.	313	19	207	8	205	391	279	244	138	299	351	317

% ammessi elevata



% ammessi bassa

Risultati di un trattamento terapeutico

	Dottore A		Dottore B	
	M	F	M	F
efficace	2000 (40%)	1000 (33%)	2000 (67%)	3000 (60%)
inefficace	3000 (60%)	2000 (67%)	1000 (33%)	2000 (40%)

Trattamento più efficace sugli uomini

Trattamento meno efficace sulle donne

M Dottore A

	F	M	F	totale
	Dottore A+B			
efficace	4000 (50%)	4000 (50%)		
inefficace	4000 (50%)	4000 (50%)		

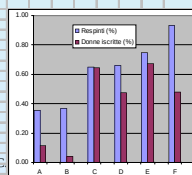
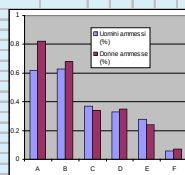
Il trattamento ha la stessa efficacia per uomini e donne

Ipotesi:

1. le iscrizioni ai vari dipartimenti sono casuali?
2. La frazione di ammessi per dipartimento è bilanciata?

Le donne tendono a far domanda nei dipartimenti dove è maggiore la probabilità di essere respinti.

	A		B		C		D		E		F	
	M	F	M	F	M	F	M	F	M	F	M	F
Ammessi	512	89	353	17	120	202	138	131	53	94	22	24
Respinti	313	19	207	8	205	391	279	244	138	299	351	317
Totale (M/F)	825	108	560	25	325	593	417	375	191	393	373	341
Totale	933	585	568	33	530	984	666	619	329	692	726	685
% ammessi	0.62	0.82	0.63	0.68	0.37	0.34	0.33	0.35	0.28	0.24	0.06	0.07
% respinti	0.36	0.17	0.37	0.32	0.63	0.66	0.67	0.65	0.76	0.75	0.94	0.93
% donne	0.12	0.04	0.04	0.03	0.06	0.05	0.07	0.06	0.07	0.07	0.09	0.08



Il grafico mostra qualitativamente una correlazione positiva tra frazione di respinti e numero di donne iscritte.

Domande?

1. Il 60% degli studenti di un liceo sono maschi e il 40% femmine. L' 80% delle ragazze fuma, mentre l'80% dei ragazzi non fuma. Quanto vale la probabilità che uno studente scelto a caso risulti maschio e fumatore? Incontrando lungo il corridoio uno studente, qual'è la probabilità che sia un fumatore? Sapendo invece che uno studente fuma, quanto vale la probabilità che sia un ragazzo? (Risolvere il problema senza considerare un ipotetico numero totale di studenti)
2. Una ditta cerca dell'acqua nel sottosuolo mediante misure di conducibilità del terreno. L'esperienza passata ha mostrato che, nei terreni in cui non c'è acqua, la probabilità che la conducibilità superi un certo valore critico è dell'1%, mentre è praticamente 1 nel caso di acqua. Le misure indicano un valore di conducibilità ben al di sopra di quello critico. Inoltre i contadini della zona ritengono molto probabile che in quel terreno si possa trovare l'acqua. Quantificando il "molto probabile" in 90% valutare la probabilità che scavando ci sia effettivamente dell'acqua.
3. *Agli impiegati piace il lifting* è il titolo di un tra filetto pubblicato sul Corriere della Sera del 12/11/94: "Plastiche facciali, lifting, seni o semplici nasi nuovi. la chirurgia estetica conquista sempre nuovi clienti. ... secondo un sondaggio della scuola di medicina estetica, infatti, solo '8% degli operati appartiene al mondo dello spettacolo. Al primo posto si collocano a sorpresa, gli impiegati (18%), seguiti da artigiani e commercianti (16%), professionisti e casalinghe (15%), dirigenti (12%). Perfino gli operai (11%) superano in numero gli artisti". Perché il giornalista si dice sorpreso dei risultati del sondaggio? (quanto vale la probabilità che una persona sia un artista, nell'ipotesi che la persona si sia sottoposta ad un intervento di chirurgia estetica?)
4. Si hanno due scatole di cui una contiene 8 palline bianche e 2 nere e l'altra 2 bianche e 8 nere. Si estrae una pallina da una scatola scelta a caso e, senza guardarla, la si ripone nell'altra scatola. Successivamente si estrae una pallina da quest'ultima scatola. Calcolare la probabilità che la pallina sia bianca facendo i conti dettagliati.