Statistica descrittiva

Saverio Vicario

Testi consigliati

- WayneW. Daniel, Biostatistica: concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria, EdiSESs.r.l. – Napoli.
- StantonA. Glantz, Statistica per discipline biomediche, 5°Edizione, McGrawHill.
- M. Pagano, K.Gauvreau, Biostatistica, II Edizione, IdelsonGnocchi.
- K. Seefeld, E. Linder. Statistics Using R with Biological Examples. Dispensa disponibile sul sito

La ricerca della verità: come fare?

- Definisci un domanda/problema
- Produci dati/ osservazioni intorno alla tua domanda
- Definisci un ipotesi
- Definisci una predizione basata sulla tua ipotesi
- Produci un osservazioni controllata basata sulla tua ipotesi per testare la predizione
- Analizza e interpreta i dati in modo da confermare o meno l'ipotesi. Questo materiale aiutera' generare nuove domande che renderanno piu specifica l'ipotesi o produrranno una nuovo e diversa ipotesi da testare
- Ritesta l'ipotesi sulla base di nuove predizioni (spesso fatto da altri)

Sommario della lezione

- Statistica descrittiva:
  - Aiuta a esplorare i dati per generare ipotesi
- Teoria delle probabilità:
  - definisce una misura con cui rigettare o corroborare un ipotesi

Definizioni degli elementi che compongono un'osservazione

- Unità statistica

- Caratteristica o attributo dell'unita statistica
- Modalità o valori dell'attributo
- Variabile aleatoria descrive la variazione dei valori dell'attributo nelle varie unita statistiche

- Il campione come insieme di unità statistiche
- La popolazione come insieme da cui e' tratto il campione

Definizioni degli elementi che compongono un'osservazione

- Una unità statistica è un singolo soggetto preso in considerazione nell'analisi.

Es.: se studiamo l'altezza media degli studenti di questa aula, ogni studente è una unità statistica.

Definizioni degli elementi che compongono un'osservazione

- Una caratteristica, detta anche feature, è un attributo o aspetto di una unità statistica.

Es.: se studiamo l'altezza media degli studenti di questa aula, l'altezza è una caratteristica

Definizioni degli elementi che compongono un'osservazione

- Variabile aleatoria o casuale

In generale una caratteristica (feature) è rappresentata da una variabile aleatoria che può assumere valori (modalità) diversi su unità statistiche differenti.

Es .pressione diastolica del sangue, la frequenza cardiaca, la statura di maschi adulti, il livello di espressione di un gene in un tessuto,… .

Variabile descritta con una lettera maiuscola tipicamente X

Mentre singolo valore la corrispondente minuscola x con pedice i che identifica il valore esatto x_i

$X=\{x_1,x_2, \ldots, x_i, \ldots, x_n\}$ sono gli n valori della variabile X

Definizioni degli elementi che compongono un'osservazione

A seconda delle modalità o valori possibili dell'attributo le variabili possono essere definite come

- Qualitative:
  - Ordinate (es. seme grande piccolo, medio, o ricercatore, prof. Associato, prof. Ordinario)
  - Non Ordinate (es. gruppo sanguigno)
- Quantitative:
  - Continue (es. lunghezza in cm di un seme, peso)
  - Discrete (es. numero di semi in una spiga)

Definizioni degli elementi che compongono un'osservazione

- Se esiste una sola variabile per unita statistica si parla di statistica univariata
  - Le domande sono:
    - Descrizione della distribuzione
    - Paragone della variabile tra due campioni
    - Paragone tra un campione e una popolazione teorica
- Diverse variabili per campione
  - Le domande sono:
    - Descrizione della distribuzione multidimensionale
    - Relazione tra variabili(es. predizione di un carattere sulla base di uno o più altri caratteri

Statistica descrittiva

- Come descrivere una distribuzione di valori osservata in una variabile aleatoria.
- X={36, 50, 85, 14, 97, 77, 98, 73, 8, 83)
- Tre approcci:
  - Classi di frequenza (direttamente applicabile a variabili con modalità qualitativa o quantitativa discreta e con approssimazioni a variabili continue)
  - Quantili (applicabile a varabili quantitative discrete o continue)
  - Momenti (applicabile solo a variabili continue e con approssimazione a variabili discrete)
- Ogni approccio definisce:
  - Indice di centralità ed eventuale indice di dispersione
  - Rappresentazione grafica

Classi di frequenza

Tipica rappresentazione tabellare per variabili qualitative o per variabili quantitative discrete. Nella tabella sono riportate:

- le **modalitàdella variabile**
- **le frequenze associate a ciascuna modalità**
- **Caso qualitativo**

```
>semi=sample(c('liscio', 'rugoso','spinoso'),10,replace=TRUE)
>semi
 [1] spinoso liscio  liscio  spinoso rugoso  rugoso  spinoso spinoso rugoso
[10] liscio
Levels: liscio rugoso spinoso
>table(semi)
semi
 liscio  rugoso spinoso
      3       3       4   #Frequenze assolute
>table(semi)/sum(table(semi))
semi
 liscio  rugoso spinoso
    0.3     0.3     0.4   #Frequenze relative
```

Classi di frequenza

```
>table(semi)
semi
 liscio  rugoso spinoso
      3       3       4   #Frequenze assolute
>table(semi)/sum(table(semi))
semi
 liscio  rugoso spinoso
    0.3     0.3     0.4   #Frequenze relative
>100*table(semi)/sum(table(semi))
Q
 liscio  rugoso spinoso
     30      30      40   #Percentuali
>cumsum(table(semi))
 liscio  rugoso spinoso
      3       6      10   #Frequenze cumalative
>cumsum(table(semi)/sum(table(semi)))
 liscio  rugoso spinoso
    0.3     0.6     1.0   # Percentuali cumalative
```

Classi di frequenza

- Per dati quantitativi discreti
- Prediamo caso InsectSpray (72 conteggi di numeri di insetti in piante trattate con 12 diversi pesticidi)

```
>data(InsectSprays)
>Q<-InsectSprays[,1]
> es<-data.frame(ni=table(Q)[],fi=table(Q)/sum(table(Q)),
  Ni=cumsum(table(Q)),Fi=cumsum(table(Q)/
  sum(table(Q))) )
```

Classi di frequenza: Indici e grafici

- Indice centrale= moda (la classe più frequente)

Moda= >names(which.max(table(semi)))

Indici di dispersione= campo di variazione ={min.max} o tabelle

Campo di variazione= >diff(range(Q))

>table(Q)

- Rappresentazione grafica:
- Grafici a barre (variabili qualitative) e Istogrammi (variabili quantitative)

>barplot(table(semi))

>hist(Q, 26)

>?hist

Classi di frequenza:caso continuo

- Dati di esempio

>data(iris) # lunghezze di petali e sepali in 3 specie di Iris

- Si definiscono classi dividendo il campo di variazione della variabile (i.e. min,max) in un numero k arbitrario di segmenti di ugual lunghezza (l=(max-min)/k)e si tabulano i risultati

>q<-hist(iris[,1])

>q

>range(iris[,1])

- Densità di probabilità
- $d_i = n_i / (l*n)$
- La densità di probabilità e'una quantità che non dipende dal'ampiezza delle classi come la frequenza

Classi di frequenza:caso continuo

- La densità di probabilità e'una quantità che non dipende dal'ampiezza delle classi come la frequenza

```
>q<-hist(iris[,1])
>q16<-hist(iris[,1],16)
>plot(q$mids,q$c, type='l')
>lines(q16$mids,q16$c, col=2)

>plot(q$mids,q$density, type='l')
>lines(q16$mids,q16$density, col=2)
```

- Si noterà che la rappresentazione del dato dipende dal numero di classi. Il numero di classi usato nella funzione hist garantisce una buona visione di insieme ma non consente di guardare i dettagli della distribuzione
- Dunque, la moda e' difficile da stabilire per variabili continue [Per il curioso (scaricare libreria ash e leggere referenze incluse per avere una versione migliorata di istogramma)]
- Ritroveremo la densità di probabilità quanto andremo a studiare le distribuzione teoriche

I quantili

- Si applica a tutti i dati quantitativi
- Se ordiniamo le nostre osservazione quantitative in ordine (con primo valore 0) crescente il numero d'ordine diviso il numero totale di oggetti rappresenta il quantile
- Il quantile varia tra 0 e 1 e non dipende dal valore esatto ma solo dalla posizione d'ordine

```
>quantile(c(0,10,20,30,40))
>quantile(c(0,12,20,38,61))
```

Indice di centralità mediana o quantile 50%

In caso di di valori pari la mediana e' uguale a

mediana=$(x_k+x_l)/2$ in cui $k<n/2<l$

I quantili

- Indice di dispersione: distanza interquartile
- Quartili= quantili 0,0.25, 0.5, 0.75, 1
- Interquartile = quantile 0.75 – quantile 0.25

- Rappresentazione grafica: box-plot

```
>quantile(iris[,1], c(0,0.25,0.5,0.75,1))
>boxplot(iris[,1])
```

Utile per paragonare distribuzioni

```
>boxplot(iris[,1:2])
>boxplot(list(Q,iris[,1]), notch=TRUE, varwidth=TRUE)
```

Quantili: paragone tra distribuzioni

- Il qqplot paragona i valori dei quantili di una distribuzione con l'altra

```
>qqplot(Q,iris[,1])
```

- Il primo elemento di uno e ' paragonato al primo elemento dell'altro scalandole nel campo di variazione di ogni variabile. Le due variabili non devono avere lo stesso numero di punti.
- qqnorm paragone una distribuzione X con la distribuzione teorica gaussiana o normale che vedremo prossima lezione

```
>hist(rnorm(10000))  #istogramma da una distribuzione normale
>qqnorm(Q)
```

Momenti e momenti centrali

- M_k=(1/n)*somma(x^k_i) Momento k-esimo
- Mc_k=(1/n)*somma((x_i-M_1)^k) Momento centrale k-esimo
- M_1=media indice di centralità
- M_{c2} = varianza indice di dispersione
- M_{c3}=simmetria -> simmetria standardizzata (M_{c3}/(M_{c2})$^{3/2}$)
- M_{c4}=curtosi -> curtosi standardizzata (M_{c4}/(M_{c2})^2)-3

Curtosi o quanto la distribuzioni e' larga o stretta

- Per una distribuzione di n elementi i primi n momenti centrali (piu' la media) definiscono la distribuzione in maniera univoca

```
>momentoC<-function(x,k){mean((x-mean(x))^k)}.
```

Proprietà della media

- Media si calcola con la funzione mean()
- Lo scarto quadratico medio o standard deviation e' la radice quadrata della varianza. E' nella stessa unità di misura della media
- La media $\bar{z}$: e' quel valore che minimizza lo scarto quadratico medio
- La media z di un campione Z nato dall'unione di un campione X con n elementi e Y con m elementi può essere calcolata dalla media di X e Y e dal numero di elementi in ognuna assumendo k=n+m

$$\bar{z} = \frac{n\bar{x} + m\bar{y}}{k}$$

Varianza

- Lo scarto quadratico medio di X e' indicato con s_x mentre la varianza e indicata con s^2_x

```
>sd(Q)^2
>var(Q)
```

- Grandi valori di s^2_x indicano la presenza di dati molto lontani dalla media; piccoli valori di s^2_x indicano dati concentrati intorno alla media. Se $s^2_x=0$ allora tutti i dati sono uguali e coincidono con la media.
- Esercizio:
- Produci un istogramma di Q e iris[,1]
- Quali sono le tue aspettative su media, varianza, simmetria e curtosi
- Paragona con i risultati calcolati con la funzione momentoC

Proprietà dei momenti per trasformazione lineari delle variabili

- Dato un campione $x_1,\ldots,x_n$ *di X con media* m_x *e varianza* s^2_X *e due numeri a e b, definito la nuovo variabile*
- $y_i=ax_i+b$
- *si ha:*
- $M_y=am_x+b$ e $s^2_y=a^2s^2_X$

- *Esercizio:*
  - *Dimostrare questa proprietà di media e varianza sapendo che*
  - M_y = *1/n * somma*$((y_i)^2)$=*1/n * somma*$((ax_i+b)^2)$

Campione standardizzato

- E' un campione con $M_x=0$ e $s^2_x=1$
- Qualunque campione può essere standardizzato sottraendo la media e dividendo per la deviazione standard
- Più formalmente applicando la trasformazione lineare
- $y_i=ax_i+b$ *con* $a=1/s_x$ e $b=-M_x/s_x$

```
a=1/sd(Q)
> b=-mean(Q)/sd(Q)
> y=a*Q+b
> mean(y)
[1] -5.692927e-17    # molto prossimo a zero
> var(y)
[1] 1
```

Accenni al multivariato

- Come rappresentazione multivariata basata sui momenti si usa la covarianza per la relazione tra due variabili e la matrice di varianza/covarianza per n variabili

- Cov=1/n somma$(x-M_x)(y-M_y)$
- Covarianza varia –Inf a +Inf
- Valori vicini a zero per variabili che non covariano
- Valori negativi per variabili che una cresce al diminuire dell'altra
- Valori positivi per opposto

```
>cov(iris[,1:4])
>plot(iris[,1:4])
```

Differenze tra moda, mediana, media

- Assumiamo un campione disomogeneo con 10% dei dati che provengono da un errore sistematico che sposta la lettura di 12

```
>B=10
>X=round(c(rnorm(1000), rnorm(100,mean=12))*B)
>hist(X)
>c(media=mean(X), mediana=median(X),
  moda=names(which.max(table(X))))
```

Con B=10 la moda funziona meglio, quando il numero dei valori possibili aumenta (B=100) la moda diventa sempre più difficile da calcolare mentre la mediana rimane robusta.

La media invece non tenta di correggere

Teoria delle probabilità

Definizioni di probabilità

- Probabilità(definizione classica) Se un evento si può verificare in *N modi mutuamente esclusivi ed ugualmente possibili*, e se *m di questi possiede la caratteristica E, la probabilitàdi verificarsi di E è: P(E) =m/N*
- *Limiti:*
  - *Non e' chiaro in tutti le applicazioni statistiche come si definiscono queste unità equiprobabili con cui si misura il P(E)*
- *Esempio:*
  - Se si lancia un dado a 6 facce non truccato, la probabilità che si osservi 2 è 1/6 ed è uguale per le rimanenti facce.

Definizioni di probabilità

- Probabilità (definizione frequentista) Se si ripete un esperimento un gran numero di volte *n* e se *un certo evento con caratteristica E si verifica m volte, la frequenza relativa di successo di E, m/n, sarà approssimativamente uguale alla probabilità di E:*
- $P(E) \approx m/m$
- *Limiti:*
  - *E' difficile immaginare infinita serie di esperimenti*
- *Esempio*
  - Supponiamo di effettuare *n=100 lanci di una moneta non truccata e di osservare m=47 volte Testa. Allora la probabilità di osservare Testa è approssimativamente 47/100.*

Definizioni di probabilità

- Probabilità (definizione soggettivista). Assume che la probabilità E *di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica, 0 se l'evento non si verifica.*
- Probabilità come rischio o incertezza
- Sempre applicabile ma soggettivo

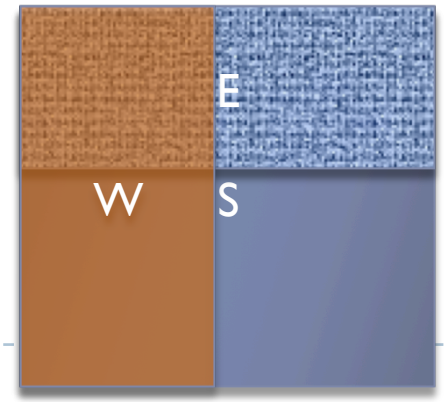
Definizioni

- Si dice *evento certo o spazio degli eventi, e si denota con S, l'insieme di tutti i risultati di un esperimento.*
- *Gli elementi di S sono chiamati risultati sperimentali.*
- *Si dice evento E un qualsiasi sottoinsieme di S e si denota con* $E \subset S$
- *Un evento si dice elementare quando è costituito da un solo elemento di S.*

Es. con rappresentazione con diagramma di Venn

Se S uguale ai numeri interi

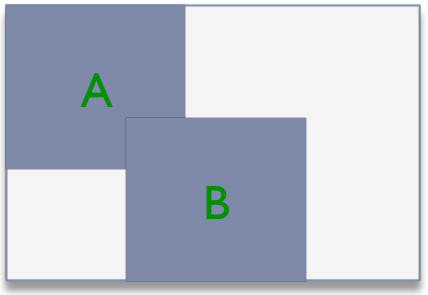
- *E uguale ai numeri negativi*
- *W uguale ai numeri pari*



Probabilità come insiemistica

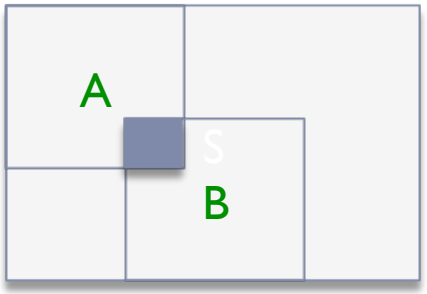
$A \cup B$

L'area degli eventi se A o B o entrambi si verificano

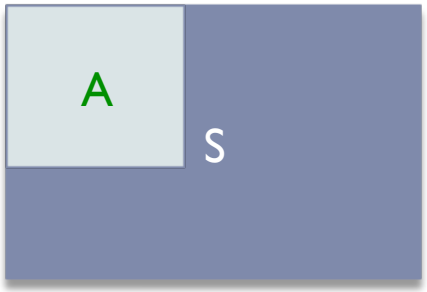


$A \cap B$

L'area degli eventi se A e B si verificano
Se l'intersezione e' vuota A e B sono mutuamente esclusivi



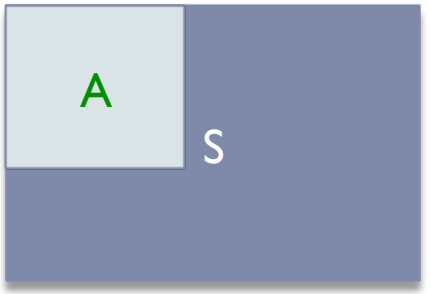
$\overline{A}$ L'area degli eventi se A non si verifica



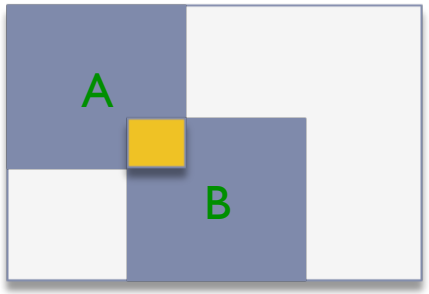
- P(A e/o B)=A U B / S

Proprietà

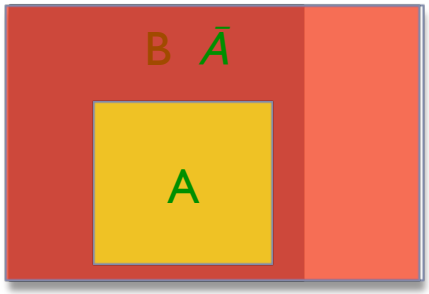
- $P(\emptyset) = 0$
- Per ogni evento *A risulta:*
  - $P(A) = 1 - P(\bar{A})$ e $0 \leq P(A) \leq 1$ assumendo $P(S)=1$



- $P(AUB) = P(A) + P(B) - P(A \cap B)$



- Gli eventi *A* e *B sono tali che* $A \subset B$.
  - *1)* $P(B) = P(A) + P(B \cap \bar{A})$
  - *2)* $P(B) \leq P(A)$.



Esempio

- Ricerca sull'uso della cocaina da parte degli uomini e delle donne (Erickson& Murray'89).

Num. volte in cui è stata usata la cocaina	Maschio (M)	Femmina (F)	Totale
1-19 volte (A)	32	7	39
20-99 volte (B)	18	20	38
>100 volte (C)	25	9	34
Totale	75	36	111

```
>z=data.frame(M=c(A=32,B=18,C=25),F=c(7,20,9))
> mosaicplot(z)
> dev.print('file)
>colSums(z)
>rowSums(z)
```

Esempio

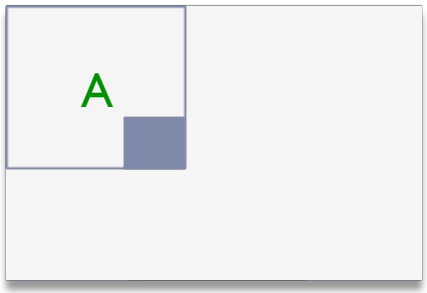
- **Qual è la probabilità che la persona scelta sia un maschio?**
- Soluzione:
- assumiamo che M e F siano categorie mutuamente esclusive e che ogni persona abbia la stessa probabilità di essere scelta. Allora la probabilità di essere M è il numero di soggetti con la caratteristica M diviso il numero totale di soggetti: *P(M)= Numero di maschi / Numero totale di soggetti = 75 / 111 = 0.6757 P(M) prende il nome di* ***probabilità marginale***.

Esempio

- **Qual è la probabilità che una persona scelta a caso dai 111 soggetti sia maschio (M) *e abbia usato la cocaina un numero di volte maggiore o uguale a 100 (C) nella sua vita?***
- **Soluzione:** l'evento che ci interessa è $M \cap C$ che si verifica quando gli eventi Me Csi verificano congiuntamente. $M \cap C$ ècostituito da tutti i soggetti maschi che hanno usato cocaina un numero di volte maggiore o uguale a 100. Quindi: $P(M \cap C) = 25 / 111 = 0.2252$.
- $P(M \cap C)$ prende il nome di **probabilità congiunta**.

Probabilità condizionata

- Consideriamo due eventi *A* e *B* e *supponiamo di sapere che l'evento A si è verificato. Il verificarsi di A modifica la probabilità di B? La probabilità di un evento B dato che l'evento A si è verificato prende il nome di probabilità condizionata* e si *denota con: P(B | A).*



Ma l'area dell'intersezione non va più divisa per l'area di S ma per l'area di A visto che A e' avvenuto ed il resto di S e' ormai impossibile

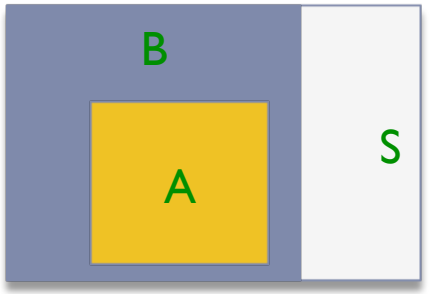
Esempio

- **Supponiamo di scegliere a caso un soggetto tra i 111 soggetti e che esso sia maschio (M). Qual èla *Pche questo maschio abbia usato cocaina un numero di volte maggiore o uguale a 100 durante la sua vita (C) ?***
- ***Soluzione:*** avendo estratto un maschio, siamo interessati al consumo di cocaina trai maschi e non tra tutti i 111 soggetti. Le femmine sono escluse. Quindi l'evento C | M è l'insieme dei maschi che ha usato cocaina un numero di volte maggiore o uguale a 100. Quindi: P(C | M)= 25 / 75 = 0.33.

- La probabilità condizionata ci consente di calcolare in modi diversi la probabilità congiunta. Siano A e B due eventi. Allora:
- *Sapendo che:* $P(A \mid B) = P(A \cap B)/P(B)$
- $P(A \cap B) = P(A \mid B)\ P(B)$.
- *Oppure:* $P(A \cap B) = P(B \mid A)\ P(A)$.

Proprietà

- Se $A \subset B$ *allora* $P(B \mid A) = 1$



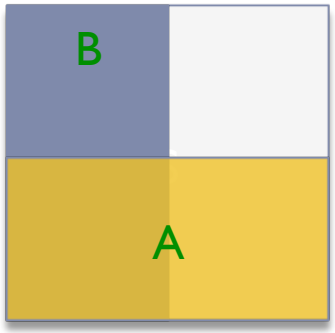
- Se $A \subset B$ *allora* $P(A \mid B) \geq P(A)$

Esempio

- Scegliamo un soggetto tra i 111 dell'esempio 1. Qual è la probabilità che questo soggetto sia maschio (M) oppure abbia usato cocaina un numero di volte maggiore o uguale a 100 nella sua vita (C)?
- Soluzione:
- gli eventi M e C non sono mutuamente esclusivi e quindi: $P(M \cup C) = P(M) + P(C) - P(M \cap C) = 75/111 + 34/111 - 25/111 = 84/111 = 0.7568$

Proprietà

- Due eventi A e B si dicono indipendenti se:
- $P(A \cap B) = P(A)\,P(B)$



- Come conseguenza si ha:
- $P(A \mid B) = P(A)$.
- *E inoltre:* $P(B \mid A) = P(B)$.

Esempio

- In un campione di 100 studenti (60 *F* e *40 M) 40 portano gli occhiali (24 O F* e *16 OM).*
- *La probabilità P(O)che uno studente porti gli occhiali* è:
- *P(O)= 40/100 = 0.4*
- *Calcoliamo: P(O | M) = P(O ∩M) / P(M) = (16/100) / (40/100) = 0.4*
- *Quindi l'informazione aggiuntiva che lo studente è un ragazzo non muta la probabilitàche uno studente porti gli occhiali*

Nota Bene

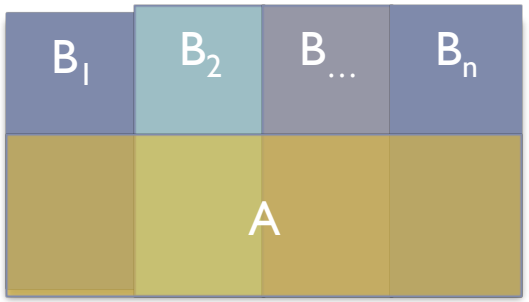
- Se *A* e *B sono eventi indipendenti allora:*
  - $P(A \mid B) = P(A)$

.

- *Ma se A e B sono eventi mutuamente esclusivi allora:*
  - $P(A \mid B) = 0$ *in quanto* $P(A \cap B) = 0$.

Teorema delle probabilità totali

- *Assumendo che:*
- $1 = B_1 \cup B_2 \cup \ldots \cup B_n$
- $P(A \cap B_1) = P(A|B_1)\, P(B_1)$



- $P(A) = P(A|B_1)\, P(B_1) + P(A|B_2)\, P(B_2) + \ldots + P(A|B_n)\, P(B_n)$

Riprendiamo l'esempio

- Ricerca sull'uso della cocaina da parte degli uomini e delle donne (Erickson& Murray'89).

Num. volte in cui è stata usata la cocaina	Maschio (M)	Femmina (F)	Totale
1-19 volte (A)	32	**7**	**39**
20-99 volte (B)	**18**	**20**	**38**
100 volte (C	**25**	**9**	**34**
Totale	75	36	111

```
>z=data.frame(M=c(A=32,B=18,C=25),F=c(7,20,9))
> mosaicplot(z)
> dev.print('file)
>colSums(z)
>rowSums(z)
```

Esempio

- $P(M) = P(M \cap A) + P(M \cap B) + P(M \cap C)$.
- $P(M) = 32/111 + 18/111 + 25/111 = 75/111$.
- *Per il teorema delle probabilità totali si ha:*
- $P(M) = P(M \mid A)\, P(A) + P(M \mid B)\, P(B) + P(M \mid C)\, P(C)$. $P(M) = (32/39)\,(39/111) + (18/38)\,(38/111) + (25/34)\,(34/111)$. $P(M) = 32/111 + 18/111 + 25/111 = 75/111$.

Teorema di Bayes

- Nota: siano *A* e *B due eventi.*
- *E' noto che:*
- $P(A \cap B) = P(A \mid B)\, P(B)$ e $P(A \cap B) = P(B \mid A)\, P(A)$.
- *Quindi:* $P(B \mid A)\, P(A) = P(A \mid B)\, P(B)$,
- *da cui:* $P(B \mid A) = P(A \mid B) P(B) / P(A)$

Nulla di particolarmente emozionante ma se A=Dati e B=Ipotesi

- $P(I \mid D) = P(D \mid I) P(I) / P(D)$
- $P(D \mid I)$ *si può calcolare (verosimiglianza dell'ipotesi o Likelihood)*
- $P(I \mid D)$ *e' quello che si vorrebbe sapere*

Ma come calcolare P(D) e P(I)?

Probabilità dei dati e delle ipotesi

- *Per il teorema probabilità totali posso dire che*
- $P(D) = P(D|I_1)\,P(I_1) + P(D|I_2)\,P(I_2) + \ldots + P(D|I_n)\,P(I_n)$

- $P(I_n)$ *probabilità a priori (prima di fare l'esperimento) dell'ipotesi n. Per interpretazione frequentista non e' quantificabile mentre soggettivista si. Rappresenta l'aspettativa di rischio che l'ipotesi* I_n *sia vera.*

La ricerca della verità: come fare?

- Definisci un domanda/problema
- Produci dati/ osservazioni intorno alla tua domanda
- Definisci un ipotesi
- Definisci una predizione basata sulla tua ipotesi
- Produci un osservazioni controllata basata sulla tua ipotesi per testare la predizione
- Analizza e interpreta i dati in modo da confermare o meno l'ipotesi. Questo materiale aiutera' generare nuove domande che renderanno piu specifica l'ipotesi o produrranno una nuovo e diversa ipotesi da testare
- Ritesta l'ipotesi sulla base di nuove predizioni (spesso fatto da altri)

Test diagnostici

- Un *test diagnostico o di screening* è un procedimento che si applica in genere a soggetti sani per determinare la presenza in essi di una patologia o per stabilire una loro eventuale suscettibilità (predisposizione) alla patologia.

Test diagnostico

- Indichiamo con *H(healthy) l'evento che un soggetto è sano* e *con D (diseased) l'evento che un soggetto è malato. Allora:*
- $H \cap D = \emptyset$ e $H \cup D = S$.
- *Inoltre, indichiamo con P l'evento che un soggetto è risultato positivo al test e con N l'evento che un soggetto è risultato negativo al test.*
- $P \cap N = \emptyset$ e $P \cup N = S$.
- *Vogliamo determinare:* $P(D \mid P)$ e $P(H \mid N)$

Test diagnostico

Risultati del test diagnostico

Vero stato del paziente	**Positive**	**Negative**
D	TP	FN
H	FP	TN

TP (TruePositive): soggetti positivi al test che sono malati.
FN (False Negative): soggetti negativi al test che sono malati.
FP (False Positive): soggetti positivi al test che sono sani.
TN (TrueNegative): soggetti negativi al test che sono sani.

Sensibilità e Specificità

- Si definisce *sensibilità di un test diagnostico:*
- *Se = P(P | D).*
- *Quindi:*
- *Sp= TN/(TN + FP)*
- *Si definisce specificità di un test diagnostico:*
- *Sp= P(N | H).*
- *Quindi: Se = TP/(TP + FN)*

Esempio

- La probabilità di recupero in soggetti affetti da cancro della cervice uterina è elevata in caso di individuazione precoce. Il Paptest è una procedura di screening che può individuare il cancro anche in casi asintomatici. Studi condotti negli anni 1972-1978 in 306 laboratori in 44 stati hanno rilevato che:
- *P(N | D)= 0.1625 (probabilità di un FN);*
- *P(P | D)= 0.8375 (Se).*
- *P(P | H)= 0.1864 (probabilità di un FP);*
- *P(N | H)= 0.8136 (Sp).*

Nota Bene

- Questi sono indicatori della bontà di un test diagnostico. Infatti per la loro stima il test si applica a soggetti il cui fenotipo è noto *a priori*.
- *Un test con elevata Se è preferibile quando non si vuole perdere nessun malato (FN≈0).*
- *Un test con elevata Sp è preferibile quando può essere dannoso avere falsi positivi.*

Valore predittivo di un test positivo

- Determiniamo *P(D | P)*
- *Applichiamo il teorema di Bayes:*
- *P(D | P) = P(P | D) P(D) /(P(P | D) P(D) + P(P | H) P(H))*
- *Sapendo che le donne affette da cancro della cervice sono 8.3 su 100.000 negli anni '82-'83:*
- *P(D)= 0.000083 e P(H) = 1–P(D)= 0.999917.*
- *Allora: P(D | P)= 0.000373.*
- *P(D | P) è detto valore predittivo di un test positivo.*

Valore predittivo di un test negativo

- Determiniamo *P(H | N)*
- *Applichiamo il teorema di Bayes:*
- *P(H | N) =P(N | H) P(H/(P(N | D) P(D) + P(N | H) P(H))*
  Nelle stesse ipotesi precedenti si ha: P(H | N)= 0.999983.
- *P(H | N) è detto valore predittivo di un test negativo.*
- *Questo indica che su 1.000.000 di donne risultate negative al Paptest, 999983 erano sane.*

Odd e Rischio relativo

- Rischio relativo Il rischio relativo (*RR*) è *il rapporto tra la probabilità di sviluppare la patologia in soggetti esposti e la probabilità di sviluppare la patologia in soggetti non esposti:*
- *RR =P(D | esposto) /(P(D | non esposto))*
- *Nota che con il termine esposto si vuole indicare una qualsiasi feature che caratterizza un gruppo di soggetti*

Rischio ed eventi rari

- Uno studio condotto negli Stati Uniti su uomini di età≥35 anni ha mostrato che:
- *P(morte per cancro al polmone | fumatore) = 0.002679*
  P(morte per cancro al polmone | non fumatore) = 0.000154

RR= 17.4

- *Quindi anche se la probabilità dell' evento considerato è bassa (evento raro), il rischio relativo mette in evidenza l'effetto del fumo sulla probabilità che un soggetto muoia di cancro al polmone.*

Tavole dei fattori Bayesiani

- Usati per interpretare i risultati dei rapporti delle probabilità di ipotesi alternative

Kass e Raferty 1995 propongono

- 1-3 appena menzionabile
- 3-20 supporto positivo
- 20-150 supporto forte
- >150 supporto decisivo

Relazioni logiche tra un test diagnostico ed un test statistico

- Il test diagnostico e' una procedura di laboratorio deve decidere tra due ipotesi (H o D) e produce un risultato binario (P o N) che viene tradotto in probabilità sulla base di uno studio di riferimento
- Un test statistico frequentista tipicamente prende in considerazione una sola ipotesi e calcola una statistica descrittiva prodotta con una combinazione degli indici descrittivi presentati e calcola la probabilità di osservare quel valore di statistica descrittiva assumendo una distribuzione nota di riferimento (probabilità condizionata della statistica assumendo l'ipotesi o verosimiglianza dell'ipotesi)