Corso Integrato di Statistica Informatica e Analisi dei Dati Sperimentali

Note A.A. 2009-10

C. Meneghini

1 Elementi di calcolo delle probabilitá, teorema di Bayes e applicazioni

1.1 Definizione di probabilitá e operazioni elementari.

Nel linguaggio comune il concetto di probabilitá si associa a situazioni non prevedibili, in pratica esprime il nostro grado di incertezza rispetto a un dato evento o fenomeno. Nella vita comune siamo abituati a trattare in modo istintivo l'incertezza in base all'esperienza e al buon senso. Vediamo ora come definire in modo quantitativo il concetto di probabilità.

Dal punto di vista scientifico la probabilità è una legge che associa un determinato valore alla possibilità che si realizzi un dato evento, in questo modo è possibile quantificare l'incertezza in modo sistematico, riproducibile e verificabile.

Il calcolo delle probabilitá é una branca relativamente giovane della matematica nata principalmente dalla curiositá dei matematici unita a questioni molto pratiche legate, ad esempio, al gioco d'azzardo. Si fa risalire l'origine della teoria del calcolo delle probabilitá alla metá del 1600 quando il cavalier de Méré (1654) chiese consiglio al matematico e filosofo francese Blaise Pascal su come ripartire le sue puntate in denaro in un gioco di dadi. Pascal discusse il problema con il matematico Pierre Fermat e da questa discussione derivano i primi approcci scientifici al calcolo delle probabilitá. Secondo il giocatore la probabilitá di avere almeno un 6 lanciando 4 volte un dado doveva essere eguale alla probabilitá di avere almeno un doppio 6 lanciando 24 volte due dadi. Tuttavia l'esperienza suggeriva il contrario. Il calcolo corretto[1] mostrava che l'esperienza aveva ragione.

La prima trattazione sistematica della teoria delle probabilitá si fa risalire ai lavori di P. S. Laplace. E' interessante notare come negli anni si siano sviluppati diversi approcci alla teoria basati su interpretazioni diverse del concetto di probabilitá. Lo sviluppo della teoria vede prima l'affermarsi dell'approccio frequentista (J. Venn, A. Cournot, R. von Mises), molto pratico, basato su osservazioni sperimentali elementari quali lanci di dadi e monete. Segue di poco lo sviluppo della concezione soggettiva, in cui si valuta la probabilitá unicamente in base alle informazioni a disposizione, questa é una concezione assai importante se pensiamo che nella vita di tutti i giorni dobbiamo continuamente effettuare scelte e prendere decisioni su eventi aleatori basandoci molto sulle informazioni (esperienza, istinto, intuito etc...) e poco sulla matematica. Nella prima metá del 900 si sviluppa una teoria assiomatica rigorosa della probabilitá (N. Kolmogorov) che si basa sui due concetti primitivi di *evento* e *probabilitá* e pochi assiomi, un po' l'equivalente della geometria euclidea. L'evento é definito come qualche cosa che puó accadere. Il concetto di probabilitá ha diverse definizioni a seconda dell'approccio utilizzato: *classica*, *frequentistica* e *soggettiva*.

La definizione *classica*é dovuta a P. Laplace (intorno al 1812) e si basa su un approccio pratico e sperimentale: dato un insieme di possibili eventi (risultati), tutti egualmente possibili e mutualmente esclusivi, la probabilitá dell'evento A é il rapporto tra il numero di casi favorevoli al verificarsi di A (N_A) e il numero di casi possibili(N_T):

$$P(A) = \frac{N_A}{N_T} \tag{1}$$

Questa definizione é relativamente semplice da usare quando sia possibile contare (misurare) i casi favorevoli e quelli possibili. Ad esempio: la probabilitá di ottenere il valore 5 lanciando un dado a 6 facce é: $P(5) = 1/6$ essendo il numero di casi favorevoli $N_5 = 1$ e le possibilità: $N_{tot} = 6$). Questa definizione richiede che gli eventi possibili siano tutti equiprobabili. Se ad esempio vogliamo sapere quale è la probabilità di osservare il carattere "e" scegliendo a caso in un punto di una pagina di un libro questa non è 1/21 (o 1/25 considerando anche le lettere x, y, j, k, w) dal momento che alcune lettere sono piú frequenti di altre (vedi ad esempio http://en.wikipedia.org/wiki/Letter_frequency) ma dipende da diversi fattori tra i quali l'alfabeto usato e dalla lingua del testo. Tuttavia se conosciamo il numero di caratteri di tipo "e" N_e e il numero totale di caratteri N_T nel libro in questione, possiamo valutare la probabilità di selezionare la lettera "e" come: $P(e) = N_e/N_T$.

Questa definizione è quindi utilizzata per variabili nominali o discrete mentre puó essere difficile da applicare e puó dar luogo a paradossi nel caso di variabili continue. Si veda, ad esempio, il paradosso di Bertrand (ad esempio Wikipedia: Bertrand, paradosso).

La critica principale alla definizione classica sta nel fatto che essa riconduce ad un circolo vizioso dal momento che nel definite la probabilità deve usare questa stessa definizione per verificare che gli eventi possibili siano effettivamente equiprobabili. Inoltre questo tipo di definizione è spesso difficilmente applicabile: tornando all'esempio precedente per definire correttamente la $P(e)$

[1] Come vedremo nel primo caso $P(A) = 1 - (5/6)^4 = 0.5178$ dove 5/6 è la probabilità di ottenere un qualunque numero diverso da 6. Nel secondo caso $P(B) = 1-(35/36)^{24} = 0.4914$ dove 35/36 è la probabilità di ottenere una qualunque coppia diversa da un doppio 6.

dobbiamo conoscere il numero esatto di lettere (N_T) e di lettere di tipo "e" (N_e) nel libro. Anche se in questo caso sarebbe cosa fattibile, anche se lunga, non sempre è cosi, anzi, di solito sono molti i casi in cui non si ha accesso all'insieme di tutti i casi possibili (popolazione) ma solo ad un sottoinsieme di essi (campione). In questi casi si deve ricorrere ad una definizione di tipo diverso.

La definizione *frequentista* o empirica é un concetto sperimentale di probabilitá definito a partire da dati empirici. La definizione frequentista riprende la definizione classica ma sostituisce al numero di casi possibili e casi favorevoli il numero di prove sperimentali (N_{exp}) e il numero di risultati favorevoli (N_A) ottenuti in esperimenti empirici identici e indipendenti, definisce quindi la probabilitá di un evento (risultato) come il limite, per un numero di prove che tende a infinito, del rapporto N_A/N_T:

$$P(A) = \lim_{N_T \to \infty} \frac{N_A}{N_T} = \lim_{N_T \to \infty} f_A$$

dove f_A è detta frequenza relativa dell'evento A. Ovviamente anche la definizione frequentista si espone a critiche, la piú ovvia é che dipende dall'esperimento: se lanciamo 50 volte una moneta osservando 20 volte testa la frequenza di teste é $f_T = 0.4$. Se noi o un nostro collega ripete l'esperimento ottenendo 35 volte testa la frequenza di teste é $f_T = 0.7$. Tuttavia se il numero di prove aumenta la frequenza relativa osservata tende a stabilizzarsi e le differenze tra esperimenti tende a ridursi. Ci sono esperimenti famosi ad opera di Buffon e Pearson in cui una moneta venne lanciata 4040 volte (ottenendo $f_T = 0.5069$), 12000 volte (ottenendo $f_T = 0.50158$) e ben 24000 volte ($f_T = 0.5005$). Nella definizione empirica il passaggio al limite non è un processo matematico rigoroso ma un'estrapolazione di osservazioni sperimentali insieme a considerazioni sulla natura del processo in esame quali la simmetria del sistema (moneta a due facce). Ad esempio la probabilità di osservare una particolare lettera dell'alfabeto scegliendo a caso in un testo scritto in una data lingua è definita in modo empirico dall'analisi di un certo numero di testi scritti in quella lingua.

Un'altra critica rilevante alla definizione frequentista è che essa si basa su due assunti principali: che le prove siano identiche e che le prove siano indipendenti, anche questi sono spesso parametri difficilmente controllabili con esattezza.

Nella definizione empirica é evidente l'analogia con la frequenza relativa delle osservazioni in una data classe: le frequenze relative di un certo insieme di eventi forniscono una stima della distribuzione di probabilitá degli eventi stessi: $P(A) \approx f_A$ tanto piú accurata quanto maggiore é il numero delle prove effettuate.

Una terza definizione di probabilità é quella *soggettiva* ovvero come misura del grado di fiducia che un individuo attribuisce, secondo tutte le informazioni in suo possesso, all'avverarsi di un evento A. Il concetto di probabilitá soggettiva, per quanto sembri definito vagamente, é estremamente importante, infatti molte delle nostre scelte e decisioni si basano sul concetto di probabilitá soggettiva. Come esempio pensiamo all'atto di frenare andando in bicicletta: se i freni non rispondono la sensazione è quella di accelerare: in base alla nostra esperienza siamo così sicuri che tirando il freno il mezzo rallenti che se il freno non funziona ci sembra di accelerare. Nell'approccio soggettivo la probabilià di un evento è definita, in modo provocatorio, come la massima somma di denaro che un soggetto razionale è disposto a scommettere a fronte di una vincita lorda unitaria. Ad esempio se si prospetta una vincita di 100 Euro nel caso esca il numero 7 al lotto, la giocata razionale dovrebbe essere $100/90 = 1.11$ Euro, scommettere una somma maggiore darebbe luogo ad uno squilibrio a favore del monopolio di stato[2].

Indipendentemente dalla definizione del concetto di probabilità, l'approccio presentato da Kolmogorov (1933) resta tuttora l'approccio deduttivo più soddisfacente al fine di un'applicazione pratica del concetto di probabilità. Come nella geometria Euclidea la teoria di Kolmogorof è di tipo assiomatico assiomatico ovvero si basa su alcune definizioni e assiomi a partire dai quali è poi possibile dedurre tutti gli altri risultati e definizioni utili per il calcolo effettivo delle probabilità di un dato evento.

Definizione 1 **Prova**: rappresenta la realizzazione di un esperimento aleatorio che esso può essere eseguito o osservato ripetutamente sempre nelle medesime condizioni. Un esperimento aleatorio genera un insieme di risultati possibili detti eventi.

Definizione 2 **Eventi**: sono i possibili esiti di un esperimento che indicheremo con lettere maiuscole A, B, C o A_1, A_2, ... A_M. Per ora limitiamoci a eventi di tipo discreto, vedremo poi come generalizzare le cose per eventi continui. L'insieme di tutti gli eventi costituisce lo spezio degli eventi o universo Ω.

Definizione 3 **Probabilità**: è una legge che associa ad ogni sottoinsieme A dello spazio degli eventi un valore $P(A)$. In analogia con la notazione insiemistica $P(A)$ rappresenta la dimensione dell'insieme A.

Assioma 1 $0 \leq P(A) \leq 1$ la Probabilià di un evento è una grandezza positiva minore di uno.

Assioma 2 per eventi certi $P(\Omega) = 1$ e per eventi impossibili $P(\emptyset) = 0$.

Assioma 3 se due eventi sono incompatibili (o disgiunti) $A \cap B = \emptyset$, ovvero se il verificarsi di uno esclude la possibilità del verificarsi dell'altro, allora: $P(A \cup B) = P(A) + P(B)$

[2] Le quote del Lotto pagano 11.232 volte la posta per l'estratto e 250 volte la posta per l'ambo

Il vantaggio dell'approccio assiomatico consiste nel fatto di poter effettuare calcoli e produrre teoremi che sebbene utilizzino il concetto di probabilità, non dipendono da come questa probabilità è definita, purchè in accordo con gli assiomi fondamentali. Il successo di questo approccio è sta nel fatto di associare gli eventi a insiemi misurabili e utilizzare il concetto di *misura* di un insieme per definire la probabilità dell'evento associato. In queso modo è possibile utilizzare molti degli strumenti matematici della teoria degli insiemi per ottenere in modo intuitivo ed elegante le proprietà legate al concetto di probabilita[3]

Dobbiamo ora introdurre alcune definizioni e notazioni che ci saranno utili per la trattazione matematica utilizzando alcuni degli strumenti già noti dalla teoria degli insiemi.

Eventi compatibili e incompatibili Indichiamo con il simbolo $\cap$ l'evento l'intersezione di due eventi: $C = A \cap B$. C é ancora un evento dello spazio degli eventi che rappresenta il verificarsi sia di A che B. Esempio: estraendo una carta da un mazzo di carte napoletane sia A=*carta di coppe*, e $B = figura$ allora $C = A \cap B$ rappresenta l'evento *figura di coppe*.

Due eventi sono incompatibili o disgiunti se la loro intersezione é l'insieme nullo: $A \cap B = \emptyset$.

Se $A = A \cap B$ allora A è un sottoinsieme di B.

Unione e insieme esaustivi Con il simbolo $\cup$ indichiamo l'unione di due insiemi, quindi $C = A \cup B$ é ancora un'insieme dello spazio degli eventi e rappresenta l'evento in cui si verifica A oppure B. Sempre utilizzando il lancio di un dado come esempio l'espressione $2 \cup 4 \cup 6$ indica l'evento in cui il lancio fornisce un valore 2 oppure 4 oppure 6.

Se $A \cup B = \Omega$ A e B sono detti esaustivi ovvero comprendono tutti i possibili eventi delle spazio.

Eventi complementari Indichiamo con il simbolo $\bar{A}$ l'evento complementare ad A (che si legge non-A). Si ha che $A \cup \bar{A}$ corrisponde a tutto lo spazio degli eventi, ovvero A e $\bar{A}$ sono esaustivi: $A \cup \bar{A} = \Omega$.

Ovviamente A e il suo complementare $\bar{A}$ sono incompatibili, quindi $A \cap \bar{A} = \emptyset$ corrisponde all'insieme vuoto, ovvero non è possibile che si verifichino contemporaneamente A e il suo opposto.

Vediamo alcuni teoremi che si possono dimostrare utilizzando le proprietà viste sopra e che ci saranno utili per il calcolo.

Probabilità dell'evento certo La probabilità associata ad un evento certo è uguale a 1: $P(\Omega) = 1$, quindi $P(A \cup \bar{A}) = 1$.

Probabilità dell'evento impossibile La probabilità associata ad un evento impossibile (insieme vuoto) è uguale a 0, quindi: $P(A \cap \bar{A}) = 0$

Probabilità dell'evento unione Se $P(A)$ e $P(B)$ sono le probabilità associate agli eventi A e B, la probabilità associata all'evento unione (ovvero al verificarsi di A oppure di B) è:

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

Se gli eventi sono incompatibili ($A \cap B = \emptyset$) si ha $P(A \cap B) = 0$ quindi:

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

Utilizzando il concetto di probabilità di un evento C come "misura" dell'insieme C e utilizzando i diagrammi di Venn il risultato è intuitivo. Vediamolo con un esempio: si estragga una carta da un mazzo di carte napoletane e si considerino gli eventi *Coppe* (estrazione di una carta di coppe), *Denari* (estrazione di una carta di denari) e *figura* (estrazione di una figura):

A	N_A/N	P(A)
coppe	$10/40$	0.25
denari	$10/40$	0.25
figura	$12/40$	0.30

La probabilità dell'evento *coppe o denari* è $P(coppe \cup denari) = P(coppe) + P(denari) = 0.5$. Mentre possiamo calcolare la probabilità di *coppe o figura* come $P(coppe \cup figura) = 10/40 + 12/40 - 3/40 = 19/40$ dove l'ultimo termine corregge il fatto che le *figure di coppe* sono state contate due volte, sia in *coppe* che in *figure*.

Probabilità dell'evento complementare Dal momento che $P(A \cap \bar{A}) = 0$ si ha $P(A \cup \bar{A}) = P(A) + P(B)$. Essendo inoltre $P(A \cup \bar{A}) = 1$ si ha: $P(\bar{A}) = 1 - P(A)$.

Intersezione di eventi e probabilità condizionate Se $P(A)$ e $P(B)$ sono le probabilità associate agli eventi A e B, la probabilità associata all'evento intersezione è:

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

Dove $P(A|B)$ rappresenta la probabilità che si verifichi A a condizione che si sia verificato B (probabilità condizionata). Nel caso in cui gli eventi siano indipendenti $P(A|B) = P(A)$ e $P(A \cap B) = P(A)P(B)$. Per visualizzare eventi indipendenti pensiamo al lancio di due dadi e mettiamo lungo l'asse x i risultati possibili del lancio del primo dado e sull'asse y quelli relativi al lancio del secondo: il risultato ottenuto sul secondo dado è indipendente dal risultato ottenuto sul primo: gli eventi sono indipendenti. Lo spazio degli eventi contiene 36 possibili risultati:

[3] link utili:
`ishtar.df.unibo.it/stat/avan/prob/matem.html`
`dipastro.pd.astro.it/ciroi/spfis1/sperI_cap3.pdf`
`www-lar.deis.unibo.it/~lmarconi/RichiamiProbabilit%C3%A0.pdf`

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Possiamo usare la definizione classica (propensione) per calcolare la probabilità di ottenere, ad esempio 1 sul primo dado e 6 sul secondo $P(1 \cap 6) = 1/36$. Usando la definizione precedente $P(1 \cap 6) = P(6|1)P(1) = 1/36$. Dal momento che i due eventi sono indipendenti $P(6|1) = P(1 \cap 6)/P(1) = P(6)$ quindi $P(1 \cap 6) = P(1)P(6) = 1/36$. Qualora i due eventi non siano dipendenti bisogna tenerne conto dal momento che la probabilità che uno dei due si presenti dipende dal successo dell'altro. Ad esempio la probabilità di estrarre dal mazzo di carte una figura di coppe é $P(coppe \cap figura) = 3/40$ ovvero la probabilità di estrarre una carta di coppe (*P(coppe) = 1/4*) per la probabilità che, data una carta di coppe, questa sia effettivamente una figura (*P(figura|coppe) = 3/10*).

La formula precedente si può invertire per calcolare la probabilità condizionata come rapporto tra la probabilità dell'evento intersezione e la probabilità dell'evento condizionante:

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \qquad (2)\\
P(B|A) &= \frac{P(A \cap B)}{P(A)}
\end{aligned}
$$

Riprendiamo l'esempio precedente *P(figura|coppe)* rappresenta la probabilità che, una volta estratta una carta di coppe, questa sia una figura. In tal caso $P(figura \cap coppe) = 3/40$ rappresenta la probabilità di estrarre una figura di coppe, ma la condizione: "una volta estratta una carta di coppe" riduce lo spazio degli eventi possibili ai soli casi in cui si sia estratta una carta di coppe.

La definizione di probabilità condizionata puó essere dimostrata in modo semplice:

$$
\begin{aligned}
P(C) &= P((A \cup \bar{A}) \cap C) = \\
&= P((A \cap C) \cup (\bar{A} \cap C)) \\
&= P(A \cap C) + P(\bar{A} \cap C) \qquad (3)\\
& \qquad (4)
\end{aligned}
$$

e dividendo ambo i membri:

$$\frac{P(A \cap C)}{P(C)} + \frac{P(\bar{A} \cap C)}{P(C)} = 1$$

in questo caso lo spazio degli eventi è costituito da due eventi possibili: la realizzazione di A

se si verifica l'evento C si può verificare, contemporaneamente, A oppure il suo opposto $\bar{A}$ quindi il primo termine rappresenta la probabilitá che si verifichi l'evento A a condizione si sia verificato C: $P(A|C)$. Il secondo termine rappresenta la probabilitá che si verifichi l'evento $\bar{A}$ sempre a condizione si sia verificato C: $P(\bar{A}|C)$. Queste due possibilità esauriscono i possibili eventi ovvero $(A|C) \cup (\bar{A}|C) = \Omega$

Essendo $P(A \cap C) = P(C \cap A)$ si ha $P(A|C)P(C) = P(C|A)P(A)$ da cui discende la regola di Bayes:

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \qquad (5)$$

Questa é una regola utile che ci permette di collegare tra loro le probabilitá condizionate e le probabilitá degli eventi condizionanti. A volte il denominatore della 5 si trova scritto:

$$P(C) = P(C|A)P(A) + P(C|\bar{A})P(\bar{A})$$

Attenzione: due eventi incompatibili non vuol dire che sono indipendenti: se A e B sono eventi **indipendenti** allora $P(A|B) = P(A)$ ovvero A si verifica indipendentemente dal verificarsi di B. Se i due eventi sono incompatibili allora il verificarsi di A esclude la possibilità che si verifichi B, quindi: $P(A|B) = 0$

1.2 Teorema di Bayes

Per illustrare il teorema Bayes usiamo un esempio pratico. Compriamo un certo numero N_T di componenti elettronici N_a prodotte dalla ditta a, N_b prodotte dalla ditta b e N_c prodotte dalla ditta c. Di queste alcune risultano difettose, siano R_a, R_b e R_c i componenti difettosi per ogni partita. La tabella seguente riassume i dati:

ditta	N. pezzi	N. difetti
a	N_a	R_a
b	N_b	R_b
c	N_c	R_c
	N_T	N_R

Utilizzando la definizione frequentista usiamo i dati in tabella per calcolare:

$P(R) = \frac{N_R}{N_T}$: la probabilità che un pezzo sia difettoso,

$P(\alpha) = \frac{N_\alpha}{N_T}$: la probabilità che un pezzo sia prodotto dalla ditta α (α = a, b, c).

$P(R|\alpha) = \frac{R_\alpha}{N_\alpha}$: la probabilità che un pezzo della ditta α sia difettoso. Ovvero la probabilità che sia difettoso un pezzo scelto tra quelli della ditta α (probabilità condizionata).

$P(\alpha|R) = \frac{R_\alpha}{N_R}$: la probabilità che un pezzo difettoso sia stato prodotto dalla ditta α. Ovvero la probabilità che un pezzo difettoso sia stato prodotto dalla ditta α.

$P(R \cap \alpha) = P(R|\alpha)P(\alpha) = \frac{R_\alpha}{N_T}$: la probabilità che un pezzo sia prodotto dalla ditta a e che sia difettoso. Da quanto visto sopra è anche: $P(R \cap \alpha) = P(\alpha|R)P(R) = \frac{R_\alpha}{N_T}$

ditta	P(α)	P(R|α)	P(R$\cap\alpha$)	P(α|R)
a	P(a)	$P(R\|a)$	$P(R\cap a)$	$P(a\|R)$
b	P(b)	$P(R\|b)$	$P(R\cap b)$	$P(b\|R)$
c	P(c)	$P(R\|c)$	$P(R\cap c)$	$P(c\|R)$

Il teorema di Bayes permette di calcolare P(α|R) a partire dalla P(R|α) o viceversa utilizzando la:

$$P(\alpha|R) = \frac{P(R|\alpha)P(\alpha)}{P(R)}$$

o la relazione inversa.

Dal momento che a, b, e c sono eventi esaustivi e indipendenti, ovvero: $a \cup b \cup c = \Omega$. Si ha anche: $R \cap \Omega = R$. Quindi la probabilità che un pezzo sia difettoso: $P(R)$ si può scrivere come la somma:

$$P(R) = P(R \cap a) + P(R \cap b) + P(R \cap c)$$

ovvero è la probabilità che sia un pezzo difettoso e prodotto dalla ditta, oppure dalla ditta b oppure dalla ditta c. Questa può essere riscritta:

$$P(R) = P(R|a)P(a) + P(R|b)P(b) + P(R|c)P(c)$$

In questo modo il teorema di Bayes si può anche scrivere:

$$P(\alpha|R) = \frac{P(R|\alpha)P(\alpha)}{P(R|a)P(a) + P(R|b)P(b) + P(R|c)P(c)}$$

Un altro modo per scrivere il teorema di Bayes è il seguente:

$$P(\alpha|R) = \frac{P(R|\alpha)P(\alpha)}{P(R|a)P(a) + P(R|\bar{a})P(\bar{a})}$$

Applicazione 1: test diagnostici

Utilizziamo il teorema di Bayes per l'interpretazione di test diagnostici o a fine di screening. Il risultato di un test diagnostico è una variabile aleatoria che dipende dallo stato di salute del paziente. Anche se sarebbe auspicabile un test infallibile a volte un individuo sano può risultare positivo (falsi positivi) o un individuo malato può risultare sano (falsi negativi). Un test diagnostico è generalmente progettato per minimizzare i falsi negativi dal momento che sono sicuramente più dannosi (es. diffusione di malattie contagiose). Un test di screening viene caratterizzato dando i valori di **specificità**, ovvero la probabilità che su un paziente sano il test risulti negativo, e la **sensibilità**, ovvero la probabilità che su un paziente malato il test risulti positivo.

	Soggetto	
Test	Sano	Malato
Negativo	$P(T^-\|S)$	$P(T^-\|M)$
Positivo	$P(T^+\|S)$	$P(T^+\|M)$

dove:

$P(T^-|S)$: è la probabilità che su un paziente sano il test risulti negativo (specificità);

$P(T^+|M)$: è la probabilità che su un paziente malato il test risulti positivo (sensibilità);

$P(T^-|M)$: è la probabilità che su un paziente malato il test risulti negativo (falsi negativi);

$P(T^+|S)$: è la probabilità che su un paziente sano il test risulti positivo (falsi positivi).

Si ha che: $P(T^-|S) + P(T^+|S) = 1$, ovvero un soggetto sano può risultare solo positivo o negativo al test e lo stesso vale per un soggetto malato: $P(T^-|M) + P(T^+|M) = 1$.

La $P(T^-|S)P(S) + P(T^-|M)P(M) = P(T^-)$ è la probabilità che un individuo qualunque (sano o malato) sia negativo al test, mentre $P(T^+|S)P(S) + P(T^+|M)P(M) = P(T^+)$ è la probabilità che un individuo qualunque (sano o malato) sia positivo al test.

I valori della tabella sono calcolati in base ad un periodo di sperimentazione del test. Ora una una domanda che ci si può porre è quale sia la probabilità che un soggetto risultato positivo sia effettivamente malato ovvero: $P(M|T^+)$. Per questo possiamo utilizzare il teorema di Bayes:

$$P(M|T^+) = \frac{P(T^+|M)P(M)}{P(T^+)}$$

dove $P(T^+)$ è la probabilità di risultare positivi al test (indipendentemente se sani o malati). Per poter applicare il teorema di Bayes è necessario conoscere la $P(M)$ ovvero la probabilità di essere effettivamente malati. Questa è un'informazione aggiuntiva che viene, ad esempio, dai dati statistici sulla diffusione della patologi (ad esempio dal servizio sanitario nazionale). Nota la $P(M)$ e sapendo che la $P(T^+)$ si calcola: $P(T^+) = P(T^+|S)P(S) + P(T^+|M)P(M) = P(T^+|S)(1-P(S)) + P(T^+|M)P(M)$ Quindi:

$$P(M|T^+) = \frac{P(T^+|M)P(M)}{P(T^+|S)(1-P(S)) + P(T^+|M)P(M)}$$

Per fare un'esempio numerico supponiamo che la patologia X colpisca il 2% della popolazione e che il test abbia una sensibilitá del 98.5% e una specificità del 95.5%. Costruiamo la tabella:

	Soggetto	
Test	Sano	Malato
Negativo	0.955	0.015
Positivo	0.045	0.985

Calcoliamo:

$$P(M|T^+) = \frac{0.985 \cdot 0.02}{0.045 \cdot 0.98 + 0.985 \cdot 0.02} = 0.31$$

Scopriamo quindi la probabilità che il soggetto sia effettivamente malato una volta che il test sia risultato positivo, è relativamente bassa. Vediamo quale è la probabilità che sia sano se il test è negativo:

$$P(S|T^-) = \frac{P(T^-|S)P(S)}{P(T^-)}$$

con $P(T^-) = P(T^-|S)P(S) + P(T^-|S)P(M)$ quindi:

$$P(S|T^-) = \frac{0.955 \cdot 0.98}{0.955 \cdot 0.98 + 0.015 \cdot 0.02} = 0.9997$$

E' quindi molto difficile che un soggetto che risulti negativo sia poi malato.

Applicazione 2: probabilità degli effetti o probabilita delle cause?

Possiamo interpretare l'osservazione di un evento come effetto condizionato da una causa (fenomeno fisico): se N_o è il numero di piantine osservate campionando una regione, questo valore è condizionato da diversi fattori (cause) come la distribuzione media di piantine sul territorio $\bar{x}$ insieme a un certo numero di fattori che influenzano l'osservazione. Da osservazioni sperimentali posso stimare la probabilità di ottenere un certo effetto dovuto ad una data causa:

$$P(\text{Effetto}|\text{causa})$$

Il teorema di Bayes mi permette di invertire l'espressione e dedurre la probabilità di una determinata causa come origine delle osservazioni (inferenza statistica)

$$P(\text{causa}|\text{Effetto})$$

ovvero dedurre la probabilità che le osservazioni sperimentali siano originate da un determinato fenomeno fisico (modello) piuttosto che un altro. Vediamo di chiarire questo punto con un esempio: la probabilità che un seme di una data specie di piante sia sterile (e quindi non nasca la piantina) è $p_s = 0.3$. Un venditore sostiene che la probabilità che i suoi semi siano sterili è sensibilmente più bassa: $p_v = 0.1$. Faccio un test piantando un primo seme e questo non germoglia. Ripeto la prova n volte senza che questi germoglino. Quale è la probabilità che il venditore sia disonesto e i suoi semi siano di tipo standard?

La probabilità che gli n semi siano tutti sterili se il venditore è onesto è:

$$P(S_n|O) = p_o^n$$

mentre se il venditore è disonesto è:

$$P(S_n|D) = p_s^n$$

In tabella calcoliamo queste probabilità in funzione del numero di prove fallite:

n	$P(S_n\|O)$	$P(S_n\|D)$
1	0.100	0.300
2	0.010	0.090
3	0.001	0.027
4	10^{-4}	0.008
5	10^{-5}	0.002

Vediamo che se il venditore fosse onesto la probabilità di trovare n il semi sterili di seguito é piú bassa che non nel caso di un seme standard.

Avendo trovato i primi n semi tutti sterili voglio calcolare quale è la probabilità che il venditore sia disonesto: $P(D|S_n)$ in modo da ridurre il rischio di accusare un innocente! Uso il teorema di Bayes per scrivere:

$$P(S_n|D)P(D) = P(D|S_n)P(S_n)$$

dove $P(S_n)$ è la probabilità che tutti gli n semi siano sterili:

$$P(S_n) = P(S_n|D)P(D) + P(S_n|O)P(O)$$

$P(D)$ è la probabilità (a priori) che il venditore sia disonesto mentre $P(O)$ è la probabilità che sia onesto. Non conoscendo la persona devo fare un ipotesi (probabilità soggettiva), sono fiducioso e stimo $P(D) = 0.1$ quindi $P(O) = 0.9$. Sostituendo:

$$P(D|S_n) = \frac{p_s^n P(D)}{p_s^n P(D) + p_o^n P(O)}$$

da cui:

n	$P(D\|S_n)$
1	0.25
2	0.50
3	0.75
4	0.90
5	0.96

Quindi maggiore è il numero di semi sterili maggiore è la probabilità che sia disonesto.

Applicazione 3: Aggiornamento dell'ipotesi.

Posso peró fare di piú. Nell'esperimento descritto sopra ho piantato un seme, quindi osservato il risultato. In caso di fallimento ne ho piantato un secondo e aspettato il germoglio e cosi' via. Anche se inizialmente eravamo abbastanza fiduciosi nel venditore (P(O)=0.9), man mano che provo i semi e questi non germogliano, la mia fiducia nell'onestá del venditore diminuisce. Come posso fare per tener conto di questa informazione? Vediamo come sfruttare la maggiore informazione che viene dall'esperienza per aggiornare la nostra ipotesi iniziale sull'onestà del venditore.

Quando troviamo il secondo seme sterile dobbiamo tener conto del fatto che abbiamo già osservato un seme sterile e che quindi la probabilità che il venditore sia onesto, prima basata solo su un ipotesi (probabilità soggettiva) è cambiata alla luce dell'osservazione. Possiamo quindi utilizzare non $P(D)$ (stima soggettiva) ma $P(D|S_1)$ ovvero la probabilità che il venditore sia disonesto avendo già osservato un primo seme sterile. Procedendo al passo n-esimo dobbiamo tener conto di quel che è successo nei passi precedenti, quindi:

$$P(D|S_n) = \frac{p_s^n P(D|S_{n-1})}{p_s^n P(D|S_{n-1}) + p_o^n (1 - P(D|S_{n-1}))}$$

dove $P(D|S_{n-1})$ è la probabilità che il venditore sia disonesto, aggiornata tenendo conto delle informazioni ottenute nelle osservazioni e $1 - P(D|S_{n-1}) = P(O|S_{n-1})$

é la probabilità che questo sia onesto, visti gli n-1 risultati precedenti. Ovviamente $P(D|S_o) = P(D)$ ovvero é la probabilitá stimata in base a considerazione del tutto soggettive. In tabella è riportata $P(D|S_n)$ aggiornata iterativamente:

n	$P(D\|S_n)$
1	0.25
2	0.75
3	0.96
4	0.996
5	0.9995

Si vede come l'aggiornamento dell'ipotesi permette di tener conto dell'esperienza nel trarre le conclusioni.

Applicazione 4: Screening e calcolo della prevalenza. Un test di screening viene effettuato per stabilire la frazione di soggetti affetti da una data patologia all'interno di una data popolazione ovvero la probabilitá che un soggetto sia malato $P(M)$. Indichiamo con N^+ il numero di casi risultati positivi e con N il numero di soggetti sottoposti a test. Dal momento che il test non é perfetto (esistono falsi positivi cosí come falsi negativi) la frazione:

$$\frac{N^+}{N}$$

indica la probabilitá che il test sia positivo ovvero $P(T^+)$ piuttosto che la probabilitá che un soggetto sia effettivamente malato, ovvero:

$$P(T^+) = P(T^+ \cap M) + P(T^+ \cap S)$$

Usando le probabiliá condizionate:

$$P(T^+) = P(T^+|M)P(M) + P(T^+|S)P(S)$$

Dal momento che: $P(S) + P(M) = 1$ si ha:

$$P(T^+) = P(T^+|M)P(M) + P(T^+|S)(1 - P(M))$$

da cui:

$$\begin{aligned} P(M) &= \frac{P(T^+) - P(T^+|S)}{P(T^+|M) - P(T^+|S)} = \\ &= \frac{\frac{N^+}{N} - P(T^+|S)}{P(T^+|M) - P(T^+|S)} \end{aligned}$$

L'errore (Err) che si compie utilizzando $P(T^+)$ per stimare la prevalenza di malati, ovvero la differenza tra la frazione di malati nella popolazione e la frazione di soggetti risultati positivi al test, é:

$$\begin{aligned} Err &= P(M) - P(T^+) = \\ &= \frac{P(T^+)\left(P(T^-|M) + P(T^+|S)\right) - P(T^+|S)}{P(T^+|M) - P(T^+|S)} \\ &= \frac{P(T^+)\left(P(T^-|M) + P(T^+|S)\right) - P(T^+|S)}{1 - \left(P(T^-|M) + P(T^+|S)\right)} \end{aligned}$$

Indicando con $f_n = P(T^-|M)$ la probabilitá di falsi negativi, con $f_p = P(T^+|S)$ la probabilitá di falsi positivi e con $f^+ = N^+/N$ la frazione di soggetti positivi osservata abbiamo:

$$P(M) = \frac{f^+ - f_p}{1 - f_p - f_n}$$

mentre l'errore:

$$\begin{aligned} Err &= P(M) - f^+ \\ &= \frac{f^+(f_n + f_p) - f_p}{1 - (f_n + f_p)} \qquad (6) \end{aligned}$$

$$(7)$$

ovvero l'errore relativo (*Err(%)*):

$$\begin{aligned} Err(\%) &= \frac{P(M) - f^+}{f^+} = \\ &= \frac{f^+(f_n + f_p) - f_p}{f^+(1 - f_n - f_p)} \end{aligned}$$

Il denominatore é sempre positivo dal momento che $f_n, f_p \ll 1$. Generalmente l'incidenza di una patologia é bassa ($f^+ \sim$ qualche %), quindi $f^+(f_n + f_p) < f_p$ e l'errore é generalmente negativo: ovvero un test di screening tende a sovrastimare l'incidenza di una patologia.

Supponiamo che il 10% dei soggetti sottoposti ad un test diagnostico per la patologia A siano risultati positivi. Sappiamo inoltre che il test ha una sensibilitá del 99% ($f_n = 0.01$) e una specificitá del 97% ($f_p = 0.03$). L'incidenza della patologia A é quindi:

$$P(M) = \frac{f^+ - f_p}{1 - f_n - f_p} = 0.073$$

con un errore relativo:

$$Err(\%) = -0.27$$

del 27% in meno.

Attenzione, la specificitá del test é un parametro importante da tenere sotto controllo perché é quello che principalmente determina la differenza tra valore vero e osservazione. In particolare se $f_p \sim f^+$ l'errore puó essere vicino al 100% . Quindi un test che puó andar bene nel caso di una popolazione a rischio (elevata incidenza di una patologia) puó essere poco accurato per una popolazione a basso rischio (pochi casi reali, bassa incidenza della patologia).