Introduzione alla probabilità

Luca Mari, versione 17.2.16

Contenuti

I principali concetti introdotti in questo capitolo

La generazione combinatoria di campioni

Ricordiamo il concetto alla base della statistica descrittiva: data una popolazione Z, eventualmente ignota, con insieme supporto $A=\{a_1,\dots,a_N\}$, si suppone possibile ottenere da essa uno o più campioni $X=\langle x_i\rangle$, intesi come successioni di elementi di A. Nell'ipotesi che il campione sia sufficientemente numeroso e sia stato ottenuto senza distorsioni (cioè, come si dice in inglese, sia *unbiased*), si considera il campione stesso un rappresentante dell'intera popolazione, e quindi l'informazione ricavata dal campione, in particolare nella forma delle sue statistiche (moda, mediana, media, ...) viene assunta come stima per la corrispondente informazione che si otterrebbe sull'intera popolazione se questa fosse nota o comunque se fosse possibile acquisire i dati corrispondenti. Ciò mette in evidenza l'importanza dell'operazione sperimentale di *campionamento*, che consente appunto di "estrarre" elementi dalla popolazione per generare campioni.
A partire da un campione $X=\langle x_1,\dots,x_n\rangle$ dato (o eventualmente dall'intera popolazione, nel caso in cui questa sia nota) è dunque spesso utile costruire nuovi campioni attraverso un'attività di "estrazione a caso" di elementi del campione di partenza. Date alcune semplici condizioni relativamente a questa attività di selezione degli elementi di X e dato il numero n degli elementi di X stesso, il problema di base che ci si pone è: quanti campioni diversi si possono generare? E di conseguenza: con quale frequenza relativa ci si può aspettare di ottenere ognuno dei campioni che possono essere generati? A questi problemi risponde una parte della matematica nota come *analisi combinatoria* (o anche "calcolo combinatorio").
Data l'importanza dell'analisi combinatoria in molti problemi che coinvolgono questioni di probabilità – di cui ci occuperemo prossimamente – introduciamo qui al proposito alcuni concetti di base.

Supponiamo dunque di voler generare da X dei campioni Y di k elementi. Al proposito si danno due condizioni alternative possibili:

- nella generazione di ogni campione Y lo stesso elemento di X può essere selezionato ripetutamente; in riferimento all'esempio di un'urna, che contiene una pallina per ogni elemento di X e da cui estrarre k palline per formare il campione Y, la possibilità di ripetizione corrisponde alla reinserzione nell'urna di ogni pallina non appena estratta; in analisi combinatoria questo è chiamato il problema delle *disposizioni con ripetizione*;
- nella generazione di ogni campione Y lo stesso elemento di X non può essere selezionato più di una volta; ancora in riferimento all'esempio dell'urna, questa condizione corrisponde alla non reinserzione nell'urna di ogni pallina estratta; in analisi combinatoria questo è chiamato il problema delle *disposizioni semplici* (o: senza ripetizione).

Cominciamo ad affrontare il problema delle disposizioni con ripetizione. Il campione Y è una k-upla di elementi $\langle y_{1,\dots},y_k\rangle$, con l'unico vincolo che ognuno di essi è scelto tra gli n elementi di X. Dunque ci sono n possibilità di scelta per y_1 (cioè y_1 potrebbe essere o x_1, o x_2, o ..., o x_n), ancora n possibilità di scelta per y_2, grazie alla possibilità di ripetizione, e così via per ognuna delle k scelte. Dunque il numero di disposizioni con ripetizione di k elementi scelti tra n è pari a:

$$n^k$$

Prendiamo in esame un esempio con piccoli numeri: una procedura di campionamento che prevede che da un lotto di $n=5$ oggetti si estraggano $k=2$ oggetti per fare un controllo di qualità sul lotto.
Nell'ipotesi che, dopo che un oggetto è stato scelto e controllato, esso sia reinserito nel lotto, e quindi possa essere scelto una seconda volta, abbiamo 5 scelte possibili per la prima scelta e ancora 5 per la seconda, dunque in tutto $5\times5=5^2$ campioni possibili.

Il problema delle disposizioni con ripetizione ammette un'interpretazione semplice e mnemonica: se si considerano gli elementi di X come le cifre di un sistema di numerazione in base n, ogni campione Y può essere inteso come un numero a k cifre in base n; nel caso più ovvio in cui $n=10$, per esempio con $k=3$ cifre si possono scrivere numeri da 0 a 999, e quindi appunto $10^3=1000$ numeri diversi.
In sintesi, la frequenza relativa con cui ci si può aspettare di ottenere un certo campione di k elementi scelti con ripetizione tra n è $1/n^k$.

Rispetto al problema delle disposizioni semplici consideriamo prima di tutto che esso è definito solo per $k\leqslant n$: data la non ripetizione della selezione, al più si potranno infatti generare campioni con tanti elementi quanti ne contiene il campione di partenza X. Questo problema può essere risolto secondo una logica analoga a quella adottata per il problema precedente. Ci sono n possibilità di scelta per il primo elemento

y_1. A questo punto nella scelta del secondo elemento y_2 l'elemento scelto per y_1 non è più disponibile, a causa della non ripetizione, e quindi rimangono $n-1$ possibilità di scelta. Per y_3 la scelta è tra $n-2$ elementi, e così via fino a y_k, per cui rimangono $n-k+1$ scelte possibili. Dunque il numero di disposizioni semplici di k elementi scelti tra n è pari a:

$$n(n-1)...(n-k+1)$$

Se definiamo "*prodotto fattoriale* di n" o anche, più semplicemente, "n fattoriale":

$$n!=\prod_{i=1}^{n} i=n\cdot(n-1)\cdot...\cdot 2\cdot 1$$

allora il numero di disposizioni semplici di k elementi scelti tra n può essere calcolato come $n!/(n-k)!$.

Riprendiamo l'esempio precedente, supponendo ora che la procedura di campionamento, che prevede sempre che da un lotto di $n=5$ oggetti si estraggano $k=2$ oggetti, non consenta però la reintroduzione degli oggetti selezionati. Abbiamo allora 5 scelte possibili per la prima scelta e 4 per la seconda, dunque in tutto $5\times 4=5!/3!=120/6=20$ campioni possibili.

Il caso particolare in cui $k=n$, cioè in cui, uno dopo l'altro, tutti gli elementi di X vengono scelti, corrisponde alla situazione in cui il campionamento ha lo scopo di generare ogni volta un nuovo ordinamento sugli elementi di X. Questo caso è chiamato delle *permutazioni*. Il numero delle permutazioni di n elementi è dunque il numero di possibili modi con cui n elementi distinti possono essere ordinati, ed è pari a $n(n-1)\cdot...\cdot 2\cdot 1$, cioè a $n!$.

I fogli di calcolo mettono a disposizione una singola funzione per calcolare sia il numero di permutazioni sia il numero di disposizioni semplici, considerate come permutazioni di `k` elementi scelti `n`:

```
=PERMUT(n,k)
```

Dunque il numero di permutazioni di `n` elementi è calcolato dalla formula:

```
=PERMUT(n,n)
```

equivalente al prodotto fattoriale di `n`:

```
=FACT(n)
```

Prendiamo in esame, infine, un ultimo caso, variante del problema delle disposizioni semplici, in cui non si tiene conto dell'ordine con cui gli elementi compaiono nel campione Y, e quindi successioni $\langle y_i \rangle$ distinte solo per l'ordine dei loro elementi non vengono conteggiate separatamente, essendo considerate in effetti lo stesso campione. Questo problema si chiama delle *combinazioni semplici*. Il numero di combinazioni semplici di k elementi scelti tra n è pari al numero delle disposizioni semplici di k elementi scelti tra n, diviso per il numero di combinazioni costituite dagli stessi elementi e distinte solo per l'ordine degli elementi stessi. Ma, come abbiamo appena visto, tale numero è proprio il numero di permutazioni di k elementi, cioè $k!$. Dunque il numero di combinazioni semplici di k elementi scelti tra n è pari a:

$$\frac{n(n-1)...(n-k+1)}{k!}=\frac{n!}{k!(n-k)!}$$

Questa è la definizione del *coefficiente binomiale*:

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

che si legge "n su k" (e in inglese, più espressivamente, "n *choose* k").

Ancora in riferimento al nostro esempio, supponiamo ora che la procedura di campionamento di $k=2$ oggetti da un lotto di $n=5$ oggetti non solo non consenta la reintroduzione degli oggetti selezionati, ma anche non mantenga la distinzione sull'ordine degli oggetti selezionati. Abbiamo allora 5 scelte possibili per la prima scelta e 4 per la seconda, ma queste 20 coppie sono a due a due identiche a meno dell'ordine, e quindi i campioni possibili in questo caso sono $5\times 4/2=5!/(2!\times 3!)=120/(2\times 6)=10$.

I fogli di calcolo mettono a disposizione una funzione per calcolare il coefficiente binomiale `n` su `k`:

```
=COMBIN(n,k)
```

il cui nome ricorda dunque, appropriatamente, che essa calcola il numero di combinazioni semplici di `k` elementi scelti `n`.

E' facile verificare che `COMBIN(n,k)` ha lo stesso valore di:
`=PERMUT(n,k)/PERMUT(k,k)`
e perciò di:
`=PERMUT(n,k)/FACT(k)`

Rivediamo dunque il nostro esempio. Se si adotta una procedura di campionamento che prevede che da un lotto di n oggetti si estraggano k oggetti senza ripetizione (cioè senza reintroduzione dopo la selezione) e non si considera rilevante l'ordine con cui i k oggetti sono stati ottenuti, il numero di possibili campioni ottenibili è proprio n su k. Per piccoli valori di n il valore del coefficiente binomiale può essere calcolato direttamente; nel caso $n=3$:

$$\binom{3}{1}=\frac{3!}{1!(3-1)!}=\frac{6}{2}=3$$

è ovvio che ci sono tre campioni di un unico elemento: $\langle x_1\rangle$, $\langle x_2\rangle$ e $\langle x_3\rangle$;

$$\binom{3}{2}=\frac{3!}{2!(3-2)!}=\frac{6}{2}=3$$

i tre campioni di due elementi sono $\langle x_1, x_2\rangle$, $\langle x_1, x_3\rangle$ e $\langle x_2, x_3\rangle$;

$$\binom{3}{3}=\frac{3!}{3!(3-3)!}=\frac{6}{6}=1$$

(si adotta qui la convenzione che $0!=1$) è ancora ovvio che c'è un unico campione di tre elementi: $\langle x_1, x_2, x_3\rangle$.

Proviamo questi concetti con valori più elevati, per esempio campionando $k=5$ oggetti da un lotto di $n=100$ oggetti. Se si ammette che gli oggetti possano essere selezionati più volte, quanti campioni diversi si possono ottenere? Si tratta di disposizioni con ripetizione; dunque $100^5=10^{10}$ (nel calcolo combinatorio si ottengono facilmente numeri molto elevati...). Se invece non sono ammesse ripetizioni, e dunque si tratta di un problema di disposizioni semplici, con soluzione $100\times 99\times 98\times 97\times 96$, un valore pari a circa 9×10^9, in questo caso dello stesso ordine di grandezza del numero di disposizioni con ripetizione. Infine, se i campioni non sono distinti per ordinamento, un problema di combinazioni semplici, il numero di campioni diversi che si possono ottenere è uguale al coefficiente binomiale 100 su 5 = $100\times 99\times 98\times 97\times 96/5!$, circa 7×10^7.
Come detto, questi semplici strumenti di analisi combinatoria consentono di calcolare il numero di campioni che si possono generare in condizioni date, e quindi, nell'ipotesi di simmetria (o: indifferenza) nella scelta dei singoli elementi, la frequenza relativa della loro occorrenza semplicemente come inverso di tale numero.

L'algebra dei campioni

A complemento delle tecniche appena illustrate, campioni possono essere generati per combinazione di campioni pre-esistenti. Consideriamo una popolazione Z, per semplicità su un insieme supporto finito. Allora:

- XY (indicato anche X, Y e $X\cap Y$) è il *campione prodotto* (o: intersezione) costituito dagli elementi che sono contenuti sia in X sia in Y; naturalmente XY può essere vuoto, e in tal caso lo si indica con $\emptyset$;
- $X+Y$ (indicato anche $X\cup Y$) il *campione somma* (o: unione) costituito dagli elementi che sono contenuti in X o in Y o in entrambi; in questo ultimo caso, cioè se $XY\neq\emptyset$, gli elementi nel campione prodotto sono inclusi nel campione somma una volta sola;
- $\bar{X}$ il *campione complemento* costituito dagli elementi della popolazione (o del super-campione) Z che non sono contenuti in X; dunque per ogni X, $X\bar{X}=\emptyset$ e $X+\bar{X}=Z$;
- generalizzando quest'ultimo caso, un insieme di campioni $\{X_1,\ldots,X_n\}$ è una partizione della popolazione (o del super-campione) Z se (i) per ogni $i\neq j$, $X_iX_j=\emptyset$ (condizione di mutua esclusività); (ii) $\sum_i A_i=Z$ (condizione di esaustività); dunque in particolare, per ogni campione X, $\{X, \bar{X}\}$ è una partizione di Z.

La condizione $X\subseteq Y$ indica poi che tutti gli elementi del campione X sono anche elementi del campione Y.

Dato l'interesse a calcolare la frequenza relativa di un campione rispetto alla popolazione da cui esso è estratto (concretamente, il rapporto tra il numero degli elementi del campione e il numero di elementi della popolazione), introduciamo la notazione:

$$X \mid Z$$

per indicare il campione X della popolazione Z (o in modo equivalente il sotto-campione X del campione Z), che si può leggere per brevità "X dato Z" o anche "X condizionato a Z". Con $Z \mid Z$ si indica perciò che si sta considerando un campione coincidente con la popolazione, mentre l'opzione di iterare l'estrazione di sotto-campioni si manifesta nella possibilità di scrivere, per esempio, $X \mid Y$ e poi $Y \mid Z$, in cui Y costituisce il campione di riferimento per X e, nello stesso tempo, un sotto-campione per Z. Questa flessibilità consente dunque di impiegare la notazione $X \mid Y$ anche per generici campioni, e quindi anche nel caso in cui non vale che $X \subseteq Y$.
Si noti che la notazione per la somma e il prodotto può essere applicata a entrambi i termini di $X \mid Z$. Dunque $X \mid YZ$ è il campione X ottenuto dal campione prodotto di Y con Z, e $X \mid Y+Z$ è il campione X ottenuto dal campione somma di Y con Z.
La notazione $X \mid Z$ è specificamente appropriata per scrivere la frequenza relativa di $X \mid Z$, indicata dunque come:

$$f(X \mid Z)$$

Non avrebbe infatti alcun senso parlare di frequenza relativa di un campione senza il riferimento al super-campione, o alla popolazione, da cui esso è ottenuto. D'altra parte, a volte si considera ammissibile la più semplice notazione:

$$f(X)$$

("la frequenza relativa di X"), evidentemente nell'ipotesi che la popolazione di riferimento, e quindi in particolare il numero dei suoi elementi, sia data. Solo per mantenere la notazione la più semplice possibile, nel seguito ove non ambiguo adotteremo questa seconda forma, ricordando però che in essa deve essere sempre intesa la presenza implicita del termine condizionante e quindi del riferimento a un super-campione o a una popolazione.
Si pone ora un problema generale: come è possibile calcolare le frequenze relative dei campioni "composti", $f(XY)$, $f(X+Y)$, ... a partire dalle frequenze relative dei campioni "componenti", $f(X)$, $f(Y)$, ... ? (in accordo alla notazione più corretta, il problema dovrebbe essere dunque posto: come calcolare $f(XY \mid Z)$, $f(X+Y \mid Z)$, ... a partire da $f(X \mid Z)$, $f(Y \mid Z)$, … ?)

Il calcolo delle frequenze relative dei campioni

Ricordiamo dunque che $f(X \mid Z)$ (la frequenza relativa del campione X dato Z, cioè considerato come sotto-campione di Z) è il rapporto tra il numero di elementi di X che sono contenuti anche in Z (e perciò il numero di elementi di XZ) e il numero di elementi di Z. Per definizione di frequenza relativa, per ogni X vale che:

$$0 \le f(X) \le 1$$

e in particolare $f(\varnothing)=0$ e $f(X \mid X)=1$.
E' immediato verificare che per ogni $X \mid Z$ vale inoltre che:

$$f(X \mid Z)=f(XZ \mid Z)$$

Se si suppongono note le frequenze relative $f(X)$, $f(Y)$, ..., (come detto, quando ciò non genera ambiguità manteniamo implicito il termine condizionante) con regole elementari si può calcolare la frequenza relativa di campioni derivati, come segue.

Se $XY=\varnothing$, cioè i campioni X e Y sono incompatibili, allora $f(X+Y)=f(X)+f(Y)$.

Ne segue, in particolare, che per ogni X:

$$f(X)+f(\bar{X})=1$$

Nel caso generale in cui il campione XY sia non nullo, nel calcolo della frequenza relativa occorre non contare due volte gli elementi di XY, e dunque:

$$f(X+Y)=f(X)+f(Y)-f(XY)$$

e più correttamente:

$$f(X+Y|Z) = f(X|Z) + f(Y|Z) - f(XY|Z) \qquad [\textit{regola della somma}]$$

Per esempio: in un campione di 100 oggetti (Z) di cui occorre controllare la qualità relativamente alla possibile presenza di due difetti, in 6 oggetti risulta presente il primo difetto (X), in 10 il secondo (Y) e in 2 entrambi i difetti (XY); qual è la frequenza relativa del sotto-campione degli oggetti che hanno almeno un difetto ($X+Y$)? La risposta si ottiene applicando la regola della somma: $f(X+Y)=f(X)+f(Y)-f(XY)=6/100+10/100-2/100$. Dato che il riferimento è sempre allo stesso campione di 100 oggetti, si sarebbe potuto ottenere lo stesso risultato operando sulle frequenze assolute e solo alla fine dividendo per il numero complessivo degli oggetti: $(6+10-2)/100$.
La regola per calcolare $f(XY)$ è solo un poco più complessa, e si basa sull'osservazione che XY può essere ottenuto considerando $X|Y$ nel contesto di Y , oppure, simmetricamente, $Y|X$ nel contesto di X , e quindi:

$$f(XY)=f(X|Y)f(Y)=f(Y|X)f(X)$$

e più correttamente:

$$f(XY|Z)=f(X|YZ)f(Y|Z) \qquad \text{[}\textit{regola del prodotto}\text{]}$$

(per esempio: nello stesso campione di 100 oggetti (Z) dell'esempio precedente, si vuole conoscere la frequenza relativa del sotto-campione degli oggetti che hanno entrambi i difetti. A questo scopo, si accerta dapprima che gli oggetti che hanno il secondo difetto (Y) sono 10 ; quindi solo su questi si accerta che gli oggetti che hanno il primo difetto ($X|Y$) sono 2 ; dunque $f(Y)=10/100$ e $f(X|Y)=2/10$, e perciò $f(XY)=(2/10)(10/100)$. D'altra parte, si sarebbe anche potuto accertare dapprima che gli oggetti che hanno il primo difetto (X) sono 6 , accertando quindi solo su questi che gli oggetti che hanno il secondo difetto ($Y|X$) sono 2 ; dunque $f(X)=6/100$ e $f(Y|X)=2/6$, e perciò $f(XY)=(2/6)(6/100)$.

Indipendenza statistica

Dati due campioni X e Y , potrebbe accadere che:

$$f(X|Y)=f(X)$$

cioè che la frequenza relativa $f(X)$ non sia influenzata dal fatto se X sia considerato rispetto a Y o meno. Prendendo in esame nuovamente la regola del prodotto, in tal caso vale dunque che:

$$f(XY)=f(X)f(Y)$$

e più correttamente:

$$f(XY|Z)=f(X|Z)f(Y|Z)$$

Quando ciò accade si dice che i campioni X e Y sono *statisticamente indipendenti*.
Per esempio: nello stesso campione di 100 oggetti (Z) degli esempi precedenti, si accerta che gli oggetti che hanno il primo difetto (X) sono 20 , che gli oggetti che hanno il secondo difetto (Y) sono 30 , e che tra questi 30 gli oggetti che hanno il primo difetto ($X|Y$) sono 6 ; allora $f(X|Z)=20/100$, $f(Y|Z)=30/100$ e $f(X|YZ)=6/30$; poiché $20/100=6/30$, i due campioni X e Y sono statisticamente indipendenti.
E' importante mantenere chiara la distinzione tra indipendenza statistica e incompatibilità di campioni. Per definizione, X e Y sono statisticamente indipendenti se:

$$f(XY)=f(X)f(Y)$$

e sono invece incompatibili se:

$$f(XY)=0$$

Esempi semplici e chiari al proposito si ottengono considerando campioni ottenuti da un usuale mazzo di 52 carte da gioco, senza jolly. Supponiamo per esempio: $X_1=$ gli assi; $X_2=$ le figure; $X_3=$ le carte rosse; dunque $f(X_1)=4/52=1/13$, $f(X_2)=4\times 3/52=3/13$; $f(X_3)=1/2$. Questi campioni sono indipendenti? e sono incompatibili? Con pochi calcoli abbiamo:

$$f(X_1X_2)=0 \qquad \text{(non ci sono assi che siano anche figure)}$$

e dunque X_1 e X_2 sono incompatibili, e d'altra parte:

$$f(X_1)f(X_2)=1/13\times 3/13\neq 0$$

e dunque X_1 e X_2 non sono indipendenti (la dipendenza tra i due è infatti evidente: $f(X_1|X_2)$ è uguale a 0 , non certo a $f(X_1)$).
D'altra parte:

$f(X_1X_3)=2/52=1/26\neq 0$ (ci sono 2 assi rossi)

e dunque X_1 e X_3 non sono mutuamente esclusivi, e d'altra parte:

$f(X_1)f(X_3)=1/13\times 1/2=1/26$

quindi:

$f(X_1X_3)=f(X_1)f(X_3)$

cioè X_1 e X_3 sono indipendenti: estrarre un asso non modifica l'informazione che sia o meno una carta rossa, e ugualmente estrarre una carta rossa non modifica l'informazione che sia o meno un asso.

Dalle frequenze relative alle probabilità

Il tema del significato stesso del concetto di probabilità è complesso e controverso. A noi basta qui considerare che le regole che le frequenze relative soddisfano possono essere interpretate in senso più generale, per esempio per caratterizzare il grado di certezza, o fiducia, o credibilità, … che si attribuisce all'accadimento di un evento, o alla verità di un'ipotesi, X a partire dalle conoscenze disponibili Z.
Secondo questa interpretazione generale, con la notazione:

$P(X|Z)$

si indica la probabilità che si attribuisce a X dato Z, per esempio la probabilità che si attribuisce al fatto che l'evento X accada a partire dall'insieme Z delle conoscenze che sono disponibili. $P(X|Z)$ è chiamata *probabilità condizionata* (in inglese *conditional probability*, per cui a volte in italiano si usa il termine "probabilità condizionale").
E' particolarmente semplice, oltre che utile per evidenziare la connessione con i fenomeni statistici, tornare a considerare Z come insieme supporto (in inglese chiamato, in questo caso, *sample space*) e X come un suo sottoinsieme (da cui la giustificazione di chiamarlo anche "evento", nel senso di "evento di scelta degli elementi di X tra quelli possibili, appunto in Z). Come anche per la statistica, vale dunque che la presenza di quel fenomeno (piuttosto misterioso...) chiamato "casualità" non ha necessariamente un ruolo nel calcolo della probabilità. Naturalmente, una possibile ragione di incertezza è l'incapacità di prevedere un evento a motivo della complessità delle sue cause, che a volte si spiega supponendo che sia presente "il caso". E' questo il tipico esempio del lancio di una moneta: con un'assegnazione di probabilità non si pretende che l'evento non segua le leggi, deterministiche, della meccanica classica, ma si rende conto appunto della complessità delle cause che generano l'effetto, "testa" o "croce". D'altra parte, consideriamo una situazione in cui una persona abbia deliberatamente scelto una faccia della moneta ma ce l'abbia mantenuta nascosta. L'unica incertezza presente in questo caso è soggettiva: la faccia della moneta è stata scelta ed è determinata. Nondimeno, rimane (per noi) sensato attribuire una probabilità alle due proposizioni "la faccia scelta è testa" e "la faccia scelta è croce".
Le regole introdotte sopra, somma e prodotto in particolare, possono essere immediatamente ripresentate per la probabilità:

$P(X+Y|Z)=P(X|Z)+P(Y|Z)-P(XY|Z)$ [regola della somma]

$P(XY|Z)=P(X|YZ)P(Y|Z)$ [regola del prodotto]

$P(XY|Z)=P(X|Z)P(Y|Z)$ [condizione di indipendenza statistica]

I due casi estremi:

$P(X|Z)=0$

e:

$P(X|Z)=1$

corrispondono a una situazione in cui, in base alla conoscenza Z disponibile, si considera X certamente vero e certamente falso (o anche, per esempio, necessario e impossibile) rispettivamente. I casi intermedi, in cui dunque la probabilità è maggiore di 0 ma minore di 1 identificano le situazioni di incertezza. E' in questo senso che si può sostenere che *la probabilità è la logica dell'incertezza*.
L'idea che la probabilità dipenda dallo stato di conoscenza disponibile sull'oggetto, evento, ipotesi, ... in esame, e non sia una "caratteristica inerente", è suggestivamente chiarita dal seguente esempio. Un lotto di n oggetti contiene n_A oggetti di tipo A e $n_B = n–n_A$ oggetti di tipo B. Se si estrae un oggetto "a caso", la probabilità che sia di tipo A è perciò n_A/n, e n_B/n che sia di tipo B. Supponiamo che il primo oggetto estratto, di tipo non conosciuto, non sia reintrodotto nel lotto, e se ne estragga un secondo, diciamo di tipo A. Da questa seconda estrazione possiamo ottenere qualche informazione circa il primo oggetto? In effetti sì, come

è chiaro in particolare nel caso $n_A=n_B=1$: se il secondo oggetto è di tipo A, il primo deve essere di tipo B. Ciò significa che la seconda estrazione modifica la probabilità della prima estrazione benché, evidentemente, questa sia già stata effettuata, prima della seconda. Se la prima estrazione fosse caratterizzata da una probabilità indipendente dal nostro stato di conoscenza, tale probabilità sarebbe stabilita a priori, e non cambierebbe certo a causa di un evento successivo all'estrazione stessa. Ma, come abbiamo visto, non è così: la probabilità di X descrive quantitativamente la certezza attribuita a X sulla base della conoscenza disponibile Z, e dunque è appropriatamente indicata $P(X|Z)$.
Nelle situazioni in cui l'omissione di Z non genera ambiguità, si può tralasciare l'indicazione del termine condizionante e scrivere $P(X)$. In conseguenza le regole indicate sopra possono essere riscritte nella versione semplificata:

$P(X+Y)=P(X)+P(Y)-P(XY)$ [regola della somma in versione semplificata]
$P(XY)=P(X|Y)P(Y)$ [regola del prodotto in versione semplificata]
$P(\emptyset)=0$
$P(Z)=1$

In molti casi, una probabilità viene attribuita generalizzando il concetto di frequenza relativa, sulla base della definizione cosiddetta "classica", dovuta tra gli altri a Laplace. L'idea è semplice: supponendo per esempio che X sia un evento che può o meno accadere, la probabilità di X è il rapporto tra il numero di casi favorevoli all'accadimento di X e il numero dei casi possibili. E' evidentemente questa la definizione che si impiega quando, per esempio, nel lancio di un dado si attribuisce all'evento $X=$ 'numero pari' la probabilità $1/2$: gli eventi favorevoli sono ' 2 ', ' 4 ' e ' 6 ', dunque 3, rispetto al totale di 6 eventi possibili. Nonostante la sua utilità, questa definizione presenta vari problemi. Prima di tutto, a proposito dei casi possibili occorre assumere che essi siano "equi-possibili": se per esempio supponessimo che il dado è truccato, non necessariamente attribuiremmo la probabilità $1/2$ all'evento 'numero pari', nonostante questa definizione. Inoltre, non è chiaro in che senso la "equi-possibilità" sarebbe accertabile nel caso di eventi come, per esempio, i fenomeni atmosferici, a proposito dei quali un'assegnazione probabilistica è evidentemente utile. Infine, sono noti alcuni paradossi a cui la definizione classica è soggetta. Si consideri per esempio un'automobile che ha percorso 1 km in un tempo non precisamente noto, tra 1 e 2 minuti. Prendendo in esame i due intervalli [1 ' , 1 '30 "] e [1 '30 " , 2 '], in assenza di ulteriore informazione li potremo trattare come equipossibili, e quindi attribuiremo a ognuno di essi la probabilità $1/2$. Consideriamo ora non il tempo di percorrenza, ma la velocità media, evidentemente tra 60 e 30 km/h. Analogamente al ragionamento precedente, attribuiremo dunque probabilità $1/2$ alle due possibilità [60 , 45] e [45 , 30] km/h. Il fatto però è che un tempo di percorrenza di 1 '30 " corrisponde a una velocità media di 40, e non 45 km/h, così che allo stesso evento staremmo attribuendo probabilità diverse in funzione del modo con cui lo si considera, una posizione quantomeno opinabile, se non appunto paradossale.
Tornando alla prospettiva più generale di interpretare la probabilità come la logica dell'incertezza, si pone infine il problema di giustificare la validità delle regole della somma e del prodotto. Il fatto che esse siano valide per frequenze relative non implica certo, infatti, che si applichino ugualmente ad assegnazioni di "gradi di certezza". E' questo un tema complesso: possiamo qui solo accennare che (i) è stato dimostrato che tali regole possono essere derivate da ipotesi sufficientemente generali circa le condizioni della rappresentazione della conoscenza, ma che, nello stesso tempo, (ii) è possibile definire logiche dell'incertezza più generali della probabilità (ciò di cui la teoria della probabilità è un caso particolare), in un ambito noto come "teoria dell'evidenza". Non è però un tema di cui si occuperemo qui.
Introduciamo invece alcune regole di base del calcolo della probabilità, che useremo ampiamente nel seguito.

Gli odds

A volte l'informazione a proposito di una frequenza o una probabilità viene presentata nella forma di un rapporto, per esempio quando si parla di una "scommessa 1 a 3". Con ciò si intende che si ammette 1 caso favorevole rispetto a 3 sfavorevoli (e quindi ci si aspetta di essere adeguatamente ricompensati nel caso di vittoria nella scommessa, ma questo è un altro discorso...). Non si tratta della frequenza relativa $1/3$, dato che i casi possibili sono $1+3=4$, e perciò la frequenza relativa corrispondente alla scommessa è $1/4$.
Questo genere di informazione si riporta in termini probabilistici normalizzando rispetto al numero di casi possibili, cioè riscrivendo " 1 a 3 " come " $1/4$ a $3/4$ ": in questo modo, se si indica con X l'evento favorevole, e quindi con $\bar{X}$ l'evento sfavorevole, " 1 a 3 " significa che $P(X)=1/4$ e $P(\bar{X})=3/4$. Lo si può intendere così come rapporto: " 1 a 3 " significa che $P(X)/P(\bar{X})=1/3$.

In generale si chiama *odds* (non esiste un termine italiano corrispondente: dovrebbe essere qualcosa del tipo "rapporto dei casi favorevoli sui casi sfavorevoli") dell'evento X :

$$\text{odds}(X) = \frac{P(X)}{1-P(X)}$$

E' facile a questo punto ricostruire $P(X)$ dato $\text{odds}(X)$ invertendo la formula precedente:

$$P(X) = \frac{\text{odds}(X)}{1+\text{odds}(X)}$$

Per esempio, $\text{odds}(X) = 2/3$ (cioè "scommessa 2 a 3") significa che stiamo attribuendo a X la probabilità $2/5$.

La regola di marginalizzazione

A partire dalla regola del prodotto (scritta nella versione semplificata):

$$P(XY) = P(X|Y)\, P(Y)$$

e anche:

$$P(\bar{X}Y) = P(\bar{X}|Y)\, P(Y)$$

sommando si ottiene:

$$P(XY) + P(\bar{X}Y) = [P(X|Y) + P(\bar{X}|Y)]\, P(Y)$$

Ma poiché, per la regola della somma:

$$P(X|Y) + P(\bar{X}|Y) = 1$$

si ottiene:

$$P(XY) + P(\bar{X}Y) = P(Y)$$

relazione che può essere generalizzata al caso di una partizione $\{X_1, ..., X_n\}$:

$$P(Y) = \Sigma_i P(X_i\, Y) \qquad \text{[}\textit{regola di marginalizzazione}\text{]}$$

che può essere scritta anche:

$$P(Y) = \Sigma_i P(Y|X_i)\, P(X_i)$$

La regola di Bayes

Poiché il prodotto di campioni è commutativo, $XY = YX$, la regola del prodotto porta a:

$$P(X|Y)\, P(Y) = P(Y|X)\, P(X)$$

che può essere scritto:

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)}$$

Sostituendo ora al denominatore in accordo alla regola di marginalizzazione, si ottiene:

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{\sum_i P(X_i\, Y)}$$

o anche:

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{\sum_i P(Y|X_i)\,P(X_i)} \qquad \text{[}\textit{regola (o teorema) di Bayes}\text{]}$$

Un'interpretazione della regola di Bayes

Consideriamo che $X = H$ sia un'ipotesi alla cui probabilità siamo interessati e $Y = D$ un dato disponibile, per esempio a seguito di un esperimento, che supponiamo dipendente dall'ipotesi in questione. Riscriviamo così la regola di Bayes:

$$P(H \mid D) = \frac{P(D \mid H)\,P(H)}{P(D)}$$

In questa formulazione ci sono tre generi di probabilità in considerazione:

- probabilità del tipo P(ipotesi): sono chiamate *probabilità a priori* (o anche *prior*), perché dipendono da informazione / conoscenza che si assume disponibile prima di compiere l'esperimento;

- probabilità del tipo P(dato|ipotesi): descrivono la struttura dell'esperimento, e quindi quanto sia verosimile ottenere il dato nel caso in cui l'ipotesi sia vera; sono perciò chiamate *verosimiglianze* (in inglese *likelihood*) e sono intese come funzioni dell'ipotesi (tanto che a volte le si scrive $L_D(H) = P(D|H)$);
- infine, probabilità del tipo P(ipotesi|dato), a cui siamo interessati: sono chiamate *probabilità a posteriori* (o anche *posterior*).

In questa prospettiva, attraverso la regola di Bayes l'esperimento può essere interpretato come uno strumento di *aggiornamento di probabilità* sulle ipotesi: si parte dalle probabilità a priori P(ipotesi) e, a seguito dei risultati dell'esperimento, le si aggiorna in probabilità a posteriori P(ipotesi|dato).

Il termine al denominatore nella regola di Bayes, $P(D)=\sum_i P(D|H_i)P(H_i)$, non ha relazioni con la specifica ipotesi H a cui si è interessati e ha solo il ruolo di fattore di normalizzazione. Si può allora riscrivere la regola come semplice proporzionalità:

$$P(H \mid D) \propto P(D \mid H) P(H)$$

cioè:

$$\text{posterior} \propto \text{verosimiglianza} \times \text{prior}$$

Un'altra opzione è di scrivere la regola di Bayes anche per l'ipotesi complementare $\overline{H}$, $P(\overline{H}|D)$:

$$P(\overline{H}|D)=\frac{P(D|\overline{H})P(\overline{H})}{\sum_i P(D|H_i)P(H_i)}$$

notando appunto che il denominatore rimane lo stesso. Allora:

$$\frac{P(\overline{H}|D)}{P(D|\overline{H})P(\overline{H})}=\frac{P(H|D)}{P(D|H)P(H)}$$

Riorganizzando i termini si ottiene:

$$\frac{P(H|D)}{P(\overline{H}|D)}=\frac{P(D|H)P(H)}{P(D|\overline{H})P(\overline{H})}$$

in cui $P(H)/P(\overline{H})$ = odds(H) ma anche $P(H|D)/P(\overline{H}|D)$ = odds($H|D$). Se poi chiamiamo $P(D|H)/P(D|\overline{H})$ *rapporto di verosimiglianza* di H, otteniamo un'ulteriore formulazione della regola di Bayes:

odds($H|D$) = rapporto di verosimiglianza di H × odds(H)

Mettiamo alla prova questo modo di usare la regola di Bayes (a volte si parla al proposito di "uso bayesiano della regola di Bayes") con un esempio (adattato da B. Efron, *Bayesians, frequentists, and scientists*, http://statweb.stanford.edu/~ckirby/brad/papers/2005BayesFreqSci.pdf). L'ecografia di una donna incinta ha mostrato che aspetta due gemelli maschi: si vuole sapere con quale probabilità siano monozigotici oppure derivino da due distinte cellule uovo. Si sa inoltre che i gemelli monozigotici, che sono sempre dello stesso sesso, sono 1/3 dei casi di gemelli. Dunque:

D = gemelli dello stesso sesso

H = gemelli monozigotici

$P(H)$ = 1/3

e perciò odds(H) = (1/3) / (2/3) = 1/2

P($D|H$) = 1 [se i gemelli sono monozigotici sono sempre dello stesso sesso]

P($D|\overline{H}$) = 1/2 [se i gemelli non sono monozigotici il loro sesso non è condizionato]

e perciò rapporto di verosimiglianza di H = 1 / (1/2) = 2

E in sintesi:

odds($H|D$) = 2 × 1/2 = 1

cioè gli odds che i gemelli siano monozigotici sapendo che sono dello stesso sesso sono "1 a 1", e quindi $P(H|D)$ = 0.5.

Distribuzioni di probabilità

Analogamente a quanto fatto nel passaggio da frequenze relative a probabilità, anche lo studio delle distribuzioni statistiche, che come abbiamo visto sintetizzano l'informazione portata da campioni nell'ipotesi che l'interesse sia solo per le frequenze, può essere interpretato come il correlato sperimentale di una *teoria delle distribuzioni di probabilità*. L'idea è che una distribuzione di probabilità p, definita su un insieme

supporto A, sia una funzione che può essere espressa in forma analitica (invece che ricavata numericamente dal campione, come nel caso delle distribuzioni statistiche), e che dunque come tale può essere studiata nelle sue caratteristiche. Ciò rende possibile l'importante generalizzazione di ammettere che l'insieme supporto possa essere continuo, l'insieme dei numeri reali o un suo sottoinsieme, e non solo discreto come invece in statistica.
Per semplicità, cominciamo comunque a considerare il caso delle distribuzioni *discrete* (in inglese *probability mass function*, termine spesso usato nella forma abbreviata "pmf"):

$$p : A \rightarrow [0,1]$$

tali dunque che, analogamente a quanto accade per le distribuzioni di frequenze relative, per ogni elemento $x \in A$, $p(x)$ è la probabilità dell'elemento x.
La condizione di normalizzazione, che richiede che la somma delle probabilità degli elementi di una partizione di A sia unitaria, si esprime in questo caso:

$$\sum_{x \in A} p(x) = 1$$

cosa che mostra come da una distribuzione di probabilità *p* si possa calcolare la *misura di probabilità* di un qualsiasi sottoinsieme X di A, $P(X)$, come:

$$P(X) = \sum_{x \in X} p(x)$$

da cui seguono immediatamente le due condizioni agli estremi:

$$P(\varnothing) = 0$$

e:

$$P(A) = 1$$

Si noti che mentre la distribuzione p è definita su A, e dunque assume un valore per ogni elemento di A stesso, la funzione P è definita su un'algebra booleana di A, al limite l'intero insieme 2^A dei sottoinsiemi di A (a volte chiamato "insieme delle parti" di A; tale insieme è indicato come 2^A perché se l'insieme A ha n elementi allora esso ha appunto 2^n elementi):

$$P : 2^A \rightarrow [0,1]$$

Per ogni $x \in A$ vale perciò, banalmente, l'uguaglianza:

$$P(\{x\}) = p(x)$$

Come semplice esempio, consideriamo il caso della distribuzione rettangolare, definita sull'insieme supporto dei numeri interi e tale che sono dati due interi a e b, $a < b$, per cui $p(x)$ ha un valore positivo e costante per gli x tra a e b, ed è nulla altrove. Poiché tra a e b inclusi ci sono $b - a + 1$ numeri interi, la distribuzione è definita da:

$$p(x) = \begin{cases} \dfrac{1}{b-a+1} & \text{se } a \leqslant x \leqslant b \\ 0 & \text{altrove} \end{cases}$$

Nel caso in cui l'insieme supporto A sia almeno ordinato (per semplicità potremmo supporre che A sia un sottoinsieme di numeri interi, possibilmente esteso tra $-\infty$ e $+\infty$), a partire da una distribuzione $p(x)$ si definisce la corrispondente *distribuzione di probabilità cumulata* $F(x)$ come:

$$F(x) = \sum_{y=-\infty}^{x} p(y)$$

che soddisfa le seguenti condizioni:

$$F(-\infty) = 0$$

$$F(+\infty) = 1$$

se $x_1 < x_2$ allora $F(x_1) \leqslant F(x_2)$

La distribuzione cumulata $F(x)$ può essere impiegata, alternativamente alla distribuzione $p(x)$, per calcolare la probabilità di un sottoinsieme X connesso di A, $X = \{x \in A \mid x_1 < x \leqslant x_2\}$:

$$P(X) = F(x_2) - F(x_1)$$

Consideriamo, per esempio, una distribuzione rettangolare di estremi $a=1$ e $b=6$. Poiché in tal caso $1+b-a=6$, essa sarà definita come:

$$p(x)=\begin{cases}\frac{1}{6} & \text{se } 1\leqslant x\leqslant 6\\ 0 & \text{altrove}\end{cases}$$

Se, sempre per esempio, $X=\{2,3,4\}$, allora il valore $P(X)$ può essere calcolato come:

$$P(X)=p(2)+p(3)+p(4)=3(1/6)=0.5$$

oppure come:

$$F(4)-F(1)=4(1/6)-1(1/6)=3(1/6)=0.5$$

Tutto ciò è in pratica identico a quanto abbiamo visto in precedenza a proposito delle distribuzioni di frequenze relative, se non per il fatto che nel caso della probabilità la distribuzione è assunta a priori, a partire da un'espressione analitica, invece che essere ricavato a posteriori dal campione disponibile. Consideriamo ora il caso delle distribuzioni *continue*, cioè con insieme supporto A continuo, coincidente con l'insieme dei numeri reali o un suo sottoinsieme (per esempio, certe distribuzioni potrebbero essere definite solo sui reali non negativi, e in tal caso $A=[0,+\infty)$).

La funzione $p(x)$, chiamata in questo caso *funzione di densità di probabilità* (in inglese *probability density function*, termine spesso usato nella forma abbreviata "pdf"), consente di calcolare la probabilità dell'intervallo $X=[x_1,x_2]$ come:

$$P(X)=\int_{x_1}^{x_2} p(x)\,dx$$

(la probabilità di un singolo elemento x non è dunque $p(x)$ – tale valore è 0 – ma $p(x)\,dx$; per questa ragione, il codominio delle distribuzioni continue non è l'intervallo $[0,1]$, come nel caso discreto, ma $[0,+\infty)$; per la stessa ragione, se x è una variabile con dimensione D – per esempio una lunghezza, dunque misurata in metri – allora la pdf $p(x)$ ha dimensione D^{-1}: in questo modo il termine generico $p(x)\,dx$, e quindi il suo integrale, diventa adimensionale, come deve essere una probabilità).

Dalla *funzione di densità di probabilità cumulata* (in inglese *cumulative probability density function*, termine spesso usato nella forma abbreviata "cdf"):

$$F(x)=\int_{-\infty}^{x} p(y)\,dy$$

si ottiene la pdf:

$$p(x)=\left.\frac{dF(y)}{dy}\right|_{y=x}$$

nonché, sempre se $X=[x_1,x_2]$, analogamente al caso discreto:

$$P(X)=F(x_2)-F(x_1)$$

Per esempio, il corrispondente nel continuo della distribuzione rettangolare è la distribuzione uniforme:

$$p(x)=\begin{cases}\frac{1}{b-a} & \text{se } a\leqslant x\leqslant b\\ 0 & \text{altrove}\end{cases}$$

$$F(x)=\begin{cases}0 & \text{se } x<a\\ \frac{x-a}{b-a} & \text{se } a\leqslant x\leqslant b\\ 1 & \text{se } x>b\end{cases}$$

Dunque qualsiasi funzione che soddisfi le condizioni:

$$p:A\rightarrow[0,+\infty)$$

$$\int_{-\infty}^{+\infty} p(x)\,dx=1$$

può essere interpretata come una pdf (e analogamente per A discreto), e qualsiasi funzione che soddisfi le condizioni:

$F : A \rightarrow [0,1]$
$F(-\infty) = 0$
$F(+\infty) = 1$
se $x_1 < x_2$ allora $F(x_1) \leqslant F(x_2)$

può essere interpretata come una cdf (e analogamente per A discreto).
Una volta fissati i parametri da cui essa dipende (gli estremi a e b nell'esempio precedente), una distribuzione di probabilità, espressa come pdf o cdf, è dunque definita. Come tale, essa può essere impiegata per generare numericamente campioni, simulando così l'operazione sperimentale di campionamento. Si dice in tal caso che i valori che si ottengono sono "estratti", per campionamento numerico, dalla distribuzione di probabilità stessa.

E' ciò che, per esempio, realizza nei fogli di calcolo ogni esecuzione della funzione `RAND()`, che genera un valore che si suppone in questo senso estratto dalla distribuzione uniforme di estremi $a = 0.0$ e $b = 1.0$.

Questo concetto di "estrazione" può essere impiegato anche a proposito di dati di origine sperimentale, investigando se e quanto appropriatamente si possa supporre che un campione dato è estratto da una distribuzione di probabilità data, dunque interpretata come *modello* per il campione. Un po' schematicamente, si possono perciò intendere statistica e teoria della probabilità come discipline complementari e convergenti:

- la statistica parte da campioni e si pone lo scopo di sintetizzare l'informazione contenuta in essi, eventualmente sulla base di ipotesi circa la distribuzione di probabilità alla base dei campioni dati; segue dunque un percorso *bottom-up*;
- la teoria della probabilità parte da distribuzioni definite analiticamente e ne studia le caratteristiche, anche al fine di stabilire se esse possano essere impiegate come modelli per campioni dati; segue dunque un percorso *top-down*.

Momenti

Abbiamo considerato in precedenza come calcolare le principali statistiche, media e varianza (e deviazione standard) in particolare, a partire da distribuzioni di frequenze relative. Analogamente a quanto accade per le distribuzioni statistiche, anche le distribuzioni di probabilità sono caratterizzate da entità come la media e la varianza, in questo caso chiamate *momenti* della distribuzione. Nel caso in cui l'insieme supporto è discreto, le espressioni introdotte in statistica sono identicamente impiegabili per i momenti, a meno di reinterpretare le frequenze relative come probabilità e le categorie come elementi dell'insieme supporto. Dunque la media di una distribuzione discreta $p(x)$, che indichiamo con μ_p, è:

$$\mu_p = \sum_{x=-\infty}^{+\infty} x\, p(x)$$

e la sua varianza, σ_p^2, è:

$$\sigma_p^2 = \sum_{x=-\infty}^{+\infty} (x-\mu_p)^2\, p(x)$$

Tre considerazioni su queste formule.
La prima. Come avevamo considerato, le statistiche campionarie sono in effetti esse stesse campioni (o, detto altrimenti "variabili casuali"), dato che ripetendo l'operazione di campionamento di una stessa popolazione in generale si ottengono campioni con, per esempio, medie ogni volta diverse. Le distribuzioni di probabilità sono invece analoghe a distribuzioni statistiche su popolazioni: una volta che la distribuzione $p(x)$, o $F(x)$, è stabilita, ogni suo momento è dato, e dunque è un parametro della distribuzione stessa.
La seconda considerazione. Quelli indicati sopra per μ_p e σ_p^2 sono due casi particolari della formula generale che definisce i momenti di una distribuzione discreta. Il momento di ordine k rispetto a z, $\mu_p(k,z)$, è:

$$\mu_p(k,z) = \sum_{x=-\infty}^{+\infty} (x-z)^k\, p(x)$$

da cui si vede come la media sia il momento primo rispetto a 0 e la varianza sia il momento secondo rispetto alla media.

La terza considerazione, a proposito di notazione. Per entità corrispondenti si usano in statistica e calcolo delle probabilità simboli analoghi, con lettere romane per le funzioni statistiche e lettere greche per quelle probabilistiche. Così, la media e la deviazione standard statistiche sono indicate con le lettere m e s, mentre i corrispondenti parametri probabilistici con le lettere μ e σ.
Il passaggio ai momenti per distribuzioni continue è ora immediato:

$$\mu_p = \int_{-\infty}^{+\infty} x\, p(x)\, dx$$

e:

$$\sigma_p^2 = \int_{-\infty}^{+\infty} (x-\mu_p)^2\, p(x)\, dx$$

Per una data distribuzione di probabilità i momenti possono essere dunque calcolati mediante le formule appena introdotte. Calcoliamo per esempio media e varianza per la generica distribuzione uniforme di estremi a e b:

$$\mu_p = \int_{-\infty}^{+\infty} x\, p(x)\, dx = \frac{1}{b-a}\int_a^b x = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{a+b}{2}$$

(riscopriamo così il fatto ovvio che in questo caso la media è il punto medio dell'intervallo $[a,b]$) e:

$$\sigma_p^2 = \int_{-\infty}^{+\infty} (x-\mu_p)^2\, p(x)\, dx = \frac{1}{b-a}\int_a^b (x-\mu_p)^2\, dx = \frac{1}{b-a}\left[\int_a^b x^2\, dx - \int_a^b 2x\mu_p\, dx + \int_a^b \mu_p^2\, dx\right] =$$

$$= \frac{1}{b-a}\left[\frac{x^3}{3}\Big|_a^b - 2\mu_p\frac{x^2}{2}\Big|_a^b + \mu_p^2 x\Big|_a^b\right] = \frac{1}{b-a}\left[\frac{b^3-a^3}{3} - \mu_p(b^2-a^2) + \mu_p^2(b-a)\right] =$$

$$\frac{(b-a)^3 + 3b^2a - 3ba^2}{3(b-a)} - \frac{\mu_p(b-a)(b+a)}{(b-a)} + \mu_p^2 = \frac{(b-a)^2}{3} + \frac{ba(b-a)}{(b-a)} - \mu_p^2 =$$

$$\frac{4b^2 - 8ab + 4a^2 + 12ab - 3a^2 - 6ab - 3b^2}{12} = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}$$

(da cui segue che $\sigma_p^2 = \frac{b-a}{2\sqrt{3}}$).
Ciò mostra che due parametri sono sufficienti per caratterizzare una distribuzione di probabilità uniforme: media e varianza, oppure media e ampiezza dell'intervallo $[a,b]$, oppure ancora gli estremi stessi a e b.
Naturalmente, altre distribuzioni di probabilità saranno caratterizzate da altri parametri, e il calcolo per ottenere il valore dei momenti sarà diverso. Rimane comunque il concetto di base: *una distribuzione di probabilità è una funzione non negativa definita sull'insieme supporto A da un'espressione analitica parametrica*.

Variabili casuali e distribuzioni di probabilità

Il concetto di *variabile casuale*, introdotto in precedenza in ambito statistico, è significativamente applicabile anche nella teoria della probabilità: si dice che una variabile casuale X segue una certa distribuzione di probabilità $p(x)$, in simboli $X \sim p$, per specificare che la distribuzione a frequenze relative del campione degli elementi $x_1, x_2, \ldots$ ottenuti da X approssima la distribuzione di probabilità data sempre meglio, al crescere nel numeri degli elementi.

> Considerando ancora una volta l'esempio della funzione dei fogli di calcolo `RAND()`, si può dire che essa implementa una variabile casuale che segue la distribuzione uniforme in [0,1], indicata con $U(0;1)$, sostenendo con ciò che ogni volta che tale funzione viene eseguita si ottiene un valore per la variabile casuale e che la distribuzione a frequenze relative di tali valori converge a $U(0;1)$ al crescere della sua numerosità (o meglio: dovrebbe convergere, se il generatore di numeri pseudo-casuali alla base della funzione è implementato correttamente).

Poiché di una distribuzione di probabilità è nota l'espressione analitica, e sono generalmente calcolabili i suoi momenti, è possibile confrontare le statistiche del campione della variabile casuale con i corrispondenti momenti della distribuzione di probabilità. Anche in questo caso, ci si può aspettare che al crescere del numero di elementi del campione la statistica converga al momento corrispondente, la media campionaria alla media della distribuzione, e così via. La figura che segue mostra un esempio: la linea orizzontale

rappresenta il valore della deviazione standard di $U(0;1)$, pari a $\sqrt{1/12}$, a cui la linea corrispondente alla deviazione standard campionaria, per campioni tra 10 e 1000 elementi, converge.

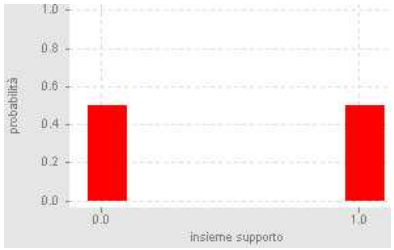
Alcune distribuzioni di probabilità importanti

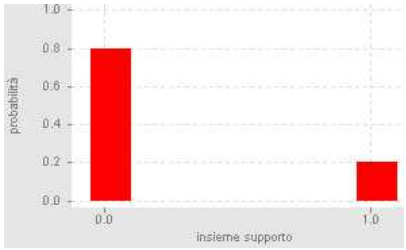
Posto che qualsiasi funzione p che soddisfi le condizioni specificate (se l'insieme supporto A è discreto, $p:A\to[0,1]$ e $\sum_{x\in A} p(x)=1$; se l'insieme supporto A è continuo, $p:A\to[0,+\infty)$ e $\int_{-\infty}^{+\infty} p(x)\,dx=1$)) è interpretabile come una distribuzione di probabilità, nel corso del tempo sono state individuate e studiate alcune distribuzioni particolarmente rilevanti per l'informazione che consentono di formalizzare.
Vediamone qualche esempio, cominciando da distribuzioni definite su supporti discreti.

La *distribuzione bernoulliana* descrive la semplicissima situazione in cui l'insieme A contiene solo due elementi, 1 e 0, interpretati rispettivamente come successo e non successo di un esperimento dato (per esempio un certo oggetto appena prodotto può essere funzionante oppure guasto, e ci si chiede quale sia la probabilità dei due casi). Assumendo che la probabilità dell'evento $x=1$, cioè di successo, sia q, la distribuzione è così definita:

$$p(x)=\begin{cases} q & \text{se } x=1 \\ 1-q & \text{se } x=0 \end{cases}$$

dove dunque la probabilità q è l'unico parametro della distribuzione, così che possiamo scrivere $X\sim \text{bern}(q)$ per indicare che la variabile casuale X segue la distribuzione bernoulliana di probabilità q.
Le due figure mostrano distribuzioni bernoulliane, con $q=0.5$ (sinistra) e $q=0.8$ (destra).





Applicando le definizioni dei momenti delle distribuzioni introdotte sopra è facile verificare che $\mu_p=q$ e $\sigma_p^2=q(1-q)$ (infatti:

$$\mu_p=\sum_{x=-\infty}^{+\infty} x\,p(x)=1\times q+0\times(1-q)=q$$

e:

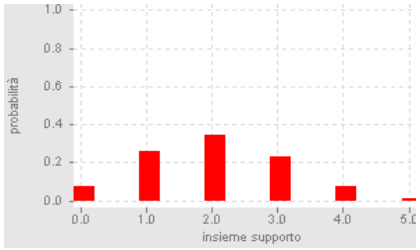
$$\sigma_p^2=\sum_{x=-\infty}^{+\infty}(x-\mu_p)^2\,p(x)=(1-q)^2\times q+(0-q)^2\times(1-q)=q-2q^2+q^3+q^2-q^3=q-q^2=q(1-q)\,).$$

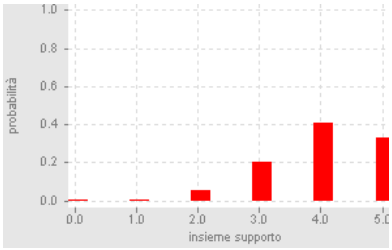
La *distribuzione binomiale* descrive la situazione in cui un esperimento bernoulliano è ripetuto un numero fissato n di volte, ogni volta nelle stesse condizioni, così che ogni ripetizione è descritta dalla stessa distribuzione bernoulliana di parametro q.
Studiamo il caso $n=2$, che prevede 4 eventi possibili: $\langle x_1=1,x_2=1\rangle$ (due successi), $\langle x_1=1,x_2=0\rangle$ e $\langle x_1=0,x_2=1\rangle$ (un successo e un insuccesso), e $\langle x_1=0,x_2=0\rangle$ (due insuccessi) (per esempio ognuno dei

due oggetti appena prodotti da macchine diverse può essere funzionante oppure guasto, e ci si chiede quale sia la probabilità dei quattro casi possibili). Poiché le ripetizioni sono statisticamente indipendenti, le probabilità di tali eventi sono $p(\langle x_1=1, x_2=1\rangle)=q\times q=q^2$, $p(\langle x_1=1, x_2=0\rangle)=p(\langle x_1=0, x_2=1\rangle)=q\times(1-q)$, e $p(\langle x_1=0, x_2=0\rangle)=(1-q)\times(1-q)=(1-q)^2$ (è facile dimostrare, a questo punto, che la somma delle probabilità di questi quattro eventi è sempre 1: $q^2+2q(1-q)+(1-q)^2=q^2+2q-2q^2+1-2q+q^2=1$).
Analogamente, per esempio la probabilità di tre insuccessi in tre ripetizioni è $(1-q)^3$, e la probabilità dell'evento $\langle x_1=1, x_2=1, x_3=0\rangle$ (successo nelle prime due ripetizioni e insuccesso nella terza) è $q\times q\times(1-q)$, che si può scrivere anche nella forma generale $q^x(1-q)^{n-x}$, dove x è il numero di successi su n ripetizioni. Per giungere alla distribuzione binomiale si considera non rilevante l'ordine con cui successi e insuccessi si alternano nella successione, così che per esempio $\langle x_1=1, x_2=1, x_3=0\rangle$ e $\langle x_1=1, x_2=0, x_3=1\rangle$ non sono distinti, e le loro probabilità devono essere perciò sommate. Come abbiamo visto in precedenza, si tratta dunque di contare il numero di combinazioni semplici di x oggetti scelti da n, un valore che è uguale al coefficiente binomiale. Si ottiene perciò:

$$p_{n,q}(x)=\binom{n}{x}q^x(1-q)^{n-x}$$

indicata anche con $\text{binom}(n, q)$, che si mostra ha media nq e varianza $nq(1-q)$.
Le due figure mostrano distribuzioni binomiali, con $n=5$ e $q=0.4$ (sinistra) e $q=0.8$ (destra).





Dunque se $X_1,\ldots,X_n$ sono variabili casuali indipendenti e tutte distribuite in modo bernoulliano con $p(x_i=1)=q$, allora:

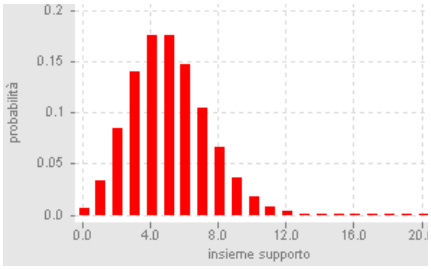
$$\sum_{i=1}^{n} X_i \sim \text{binom}(n, q)$$

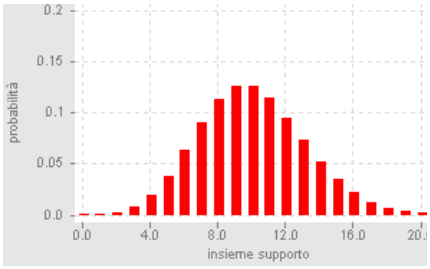
Ciò mostra anche che una distribuzione bernoulliana è un caso particolare di distribuzione binomiale, per $n=1$.

L'insieme supporto A contiene dunque due elementi per distribuzioni bernoulliane e n elementi per distribuzioni binomiali. Una terza distribuzione a insieme supporto discreto è la *distribuzione poissoniana*, che descrive la probabilità che $0, 1, 2,\ldots$ eventi indipendenti accadano in un dato intervallo di tempo (o di spazio) (per esempio nell'arco di una giornata si possono ricevere $0, 1, 2,\ldots$ ordini per la consegna di merce, e ci si chiede quale sia la probabilità di ognuno dei diversi casi possibili), e senza che ci sia a priori un numero massimo di eventi possibili. Tale distribuzione ha un unico parametro, il suo valor medio $\lambda>0$ (che non necessariamente è un numero intero, naturalmente), ed è definita come:

$$p_\lambda(x)=e^{-\lambda}\frac{\lambda^x}{x!}$$

Si dimostra che λ è anche la varianza della distribuzione.
Le due figure mostrano distribuzioni poissoniane, con $\lambda=5$ (sinistra) e $\lambda=10$ (destra).



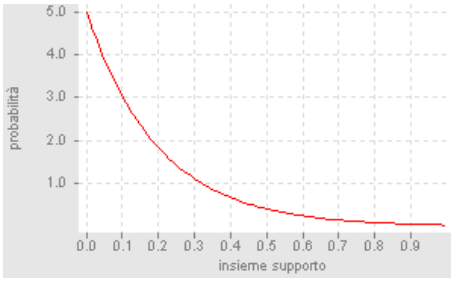


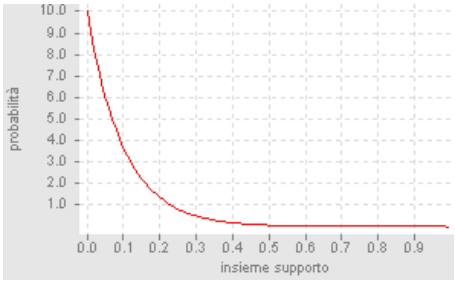
Passiamo ora a considerare distribuzioni definite su supporti continui.

Un'interessante distribuzione è connessa con la poissoniana: se quest'ultima descrive la probabilità di $n=0,1,2,\ldots$ eventi indipendenti in una data unità di tempo, la *distribuzione esponenziale* descrive la probabilità dei diversi intervalli di tempo tra due eventi successivi. Se nell'unità di tempo il numero medio λ di eventi è grande, e quindi è grande la frequenza degli eventi, allora l'intervallo medio $1/\lambda$ inter-eventi sarà piccolo, e viceversa. Ciò mostra che le distribuzioni poissoniana ed esponenziale sono sostanzialmente inverse l'una all'altra. La distribuzione esponenziale dipende da un unico parametro, appunto l'intervallo medio $1/\lambda$ inter-eventi, ed è definita, sull'insieme supporto $[0,+\infty)$, come:

$$p_\lambda(x)=\lambda e^{-\lambda x}$$

Si dimostra che λ è anche la deviazione standard della distribuzione.
Le due figure mostrano distribuzioni esponenziali, con $\lambda=5$ (sinistra) e $\lambda=10$ (destra), dunque corrispondenti a un intervallo medio inter-eventi di 0.2 e 0.1 unità di tempo rispettivamente.

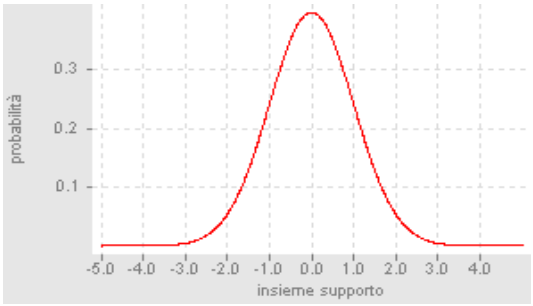


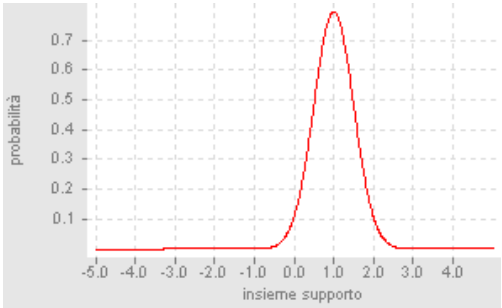


Abbiamo già introdotto la *distribuzione gaussiana* (chiamata anche distribuzione "normale"), che in particolare è quanto risulta dal teorema del limite centrale. Nonostante la sua espressione analitica sia piuttosto complessa, l'idea di base è semplice: una gaussiana è una distribuzione esponenziale quadratica negativa, cioè, a meno di precisazioni, $p(x)=\exp(-x^2)$. L'effettiva definizione consente di specificare media μ_p e deviazione standard σ_p, dunque i due parametri della distribuzione, e naturalmente garantisce la condizione di normalizzazione:

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

indicata in breve $N(\mu,\sigma)$ (la lettera "*N*" deriva da "normale").
Le due figure mostrano distribuzioni gaussiane, con $\mu=0$ e $\sigma=1$ (sinistra) e $\mu=1$ e $\sigma=0.5$ (destra).





Data l'importanza della distribuzione gaussiana, i fogli di calcolo contengono funzioni che ne consentono il calcolo. In particolare, `NORMDIST()` calcola i valori della pdf e della cdf, così che, per esempio, `NORMDIST(2,3,0.5,0)` è il valore della pdf di media `3` e deviazione standard `0.5` nel punto `2`, e `NORMDIST(2,3,0.5,1)` è il valore della cdf con gli stessi parametri e nello stesso punto.
La cdf è una funzione monotona non decrescente; nel caso in cui sia strettamente crescente (cioè nel dominio non ci sono intervalli in cui la pdf ha valore 0; è per esempio il caso della distribuzione gaussiana), essa è invertibile, così che se la cdf è:

$F: A \to [0,1]$

la cdf inversa è:

$F^{-1}: [0,1] \to A$

La cdf inversa può essere impiegata per costruire un *generatore di numeri (pseudo-)casuali* che seguono la distribuzione di probabilità data. In pratica, occorre semplicemente calcolare la cdf inversa su un argomento casuale a distribuzione uniforme, che i fogli di calcolo mettono a disposizione tramite la funzione `RAND()`.
Dunque, data la funzione che calcola la cdf inversa gaussiana, `NORMINV()`, ogni volta che si esegue per esempio `NORMINV(RAND(),3,0.5)` si ottiene un numero casuale estratto da una distribuzione gaussiana di media `3` e deviazione standard `0.5`.