Probabilità e statistica

(appunti di Paolo Gronchi)

An unsophisticated forecaster uses statistics
as a drunken man uses lamp-posts:
For support rather than for illumination.

Andrew Lang

Indice

1. Probabilità

Spesso nella pratica si ha a che fare con circostanze o esperimenti dei quali è impossibile predire con certezza l'esito. Il lancio di una moneta è il tipico esempio. Sono possibili due diversi esiti: o esce testa o esce croce. Ma ce ne sono di più complessi e allo stesso tempo usuali. Esperimenti con esiti casuali sono:

- il risultato di una partita di calcio della prossima giornata di campionato. Esiti possibili: 1, X, 2;
- il sesso di un nascituro al momento del suo concepimento. Esiti possibili: M, F;
- i cinque numeri estratti al lotto nella ruota di Napoli. Esiti possibili: tutte le cinquine ordinate di numeri compresi tra 1 e 90;
- l'altezza di un individuo. Esiti possibili: numeri (interi?) compresi tra un minimo ed un massimo.

Il calcolo delle probabilità si propone di quantificare l'incertezza propria di queste situazioni aleatorie. In altre parole, stabilisce opportune regole per esprimere il grado di fiducia che si attribuisce al verificarsi di un evento (a partire da ipotesi o conoscenze su eventi meno complicati).

Per parlare di probabilità è necessario chiarire l'ambiente nel quale operiamo.

Lo **spazio campione** può essere definito come l'insieme di tutti gli esiti possibili di un esperimento dipendente dal caso. Usualmente è indicato con Ω ed i suoi elementi sono detti **punti campione** o esiti possibili.

Qualche autore chiama *spazio degli eventi* lo spazio campione ed *eventi elementari* i punti campione.

Un **evento** può essere visto come un sottoinsieme dello spazio campione, cioè come l'insieme dei possibili esiti dell'esperimento che indicano il verificarsi dell'evento. Per questo motivo le operazioni booleane definite tra gli insiemi si traducono in operazioni tra eventi.

L'unione di due eventi $A \cup B$ è l'evento accade A o accade B o entrambi.

L'intersezione di due eventi $A \cap B$ è l'evento accadono entrambi A e B.

Il complemento di un evento A^c è l'evento non accade A.

Esempi di spazi campione.
Lancio di un dado. Lo spazio campione è $\Omega = \{1, 2, 3, 4, 5, 6\}$. I punti campione o eventi elementari sono i sei elementi di Ω definibili a parole con il risultato del lancio è il numero n. L'evento $D =$ il risultato del lancio è un numero dispari non è un evento elementare e possiamo scrivere $D = \{1, 3, 5\}$.

Misurazione con cronometro del tempo di caduta di un grave. Lo spazio campione Ω può essere pensato discreto in quanto composto di tutti i numeri interi positivi compresi tra due valori di riferimento (esprimendo il tempo in un'opportuna unità di misura). I punti campione sono i singoli valori temporali. Un evento potrebbe essere il tempo di caduta è superiore ai 15 secondi.

Lancio ripetuto di una moneta (Processo di Bernoulli). Lo spazio campione Ω è l'insieme delle successioni di due simboli o numeri (uno per testa e l'altro per croce). I matematici preferiscono usare i numeri 0 e 1 a simboleggiare il numero di testa uscito all'ennesimo lancio. I punti campione o eventi elementari sono le successioni di 0 e 1. L'evento è uscita testa al quinto lancio non è un evento elementare e non è proponibile descriverlo come sottoinsieme di Ω. Gli eventi è uscita testa all'n-esimo lancio sono detti *eventi generatori* e sono di fondamentale importanza per descrivere eventi più complessi e stabilire quindi la loro probabilità.

Per restare nell'ambito più generale possibile è bene introdurre il concetto di **σ-algebra** di insiemi. Dato un insieme Ω, una famiglia $\Im$ non vuota di sottoinsiemi di Ω si dice una σ-algebra se verifica gli assiomi

(A1) se $A_n \in \Im$ per $n = 1, 2, \ldots$, allora $\bigcup_{n\geq 1} A_n \in \Im$

(A2) se $A \in \Im$ allora $A^c \in \Im$.

È semplice verificare che ogni σ-algebra contiene l'insieme Ω e l'insieme vuoto $\emptyset$. Inoltre è chiusa rispetto alla intersezione numerabile. Nel caso in cui l'insieme Ω contiene solo un numero finito di elementi, allora la (A1) può essere riformulata chiedendo che l'unione di due sottoinsiemi in $\Im$ sia ancora in $\Im$.

A questo punto possiamo introdurre il concetto di probabilità. Seguiremo il cosiddetto metodo assiomatico. Fissato uno spazio campione Ω ed una σ-algebra $\Im$ di parti di Ω, una probabilità è una funzione P che assegna un numero reale $P(A)$ ad ogni evento A appartenente a $\Im$ con le seguenti proprietà:

(P1) $P(A) \geq 0$

(P2) $P(\Omega) = 1$

(P3) se A_1, A_2, ... è una successione di eventi di $\Im$ a due a due disgiunti, allora

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots .$$

Osserviamo che la probabilità risulta così definita soltanto sui sottoinsiemi di Ω che appartengono a $\Im$.

La terna $(\Omega, \Im, P)$ si chiama **spazio di probabilità**.

La proprietà (P3) si chiama *additività completa*; se il numero di eventi che vi compaiono è finito, allora si parla di *additività finita* e la sua necessità è abbastanza evidente.

Le principali proprietà di una probabilità sono le seguenti:

(P4) per ogni evento A è $0 \leq P(A) \leq 1$

(P5) per ogni evento A risulta $P(A^c) = 1 - P(A)$

(P6) per l'evento impossibile $\emptyset$ risulta $P(\emptyset) = 0$

(P7) se A e B sono eventi incompatibili allora $P(A \cup B) = P(A) + P(B)$

(P8) se A è un sottoevento di B allora $P(A) \leq P(B)$

(P9) se A e B sono eventi allora $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(P10) gli eventi A e B sono **indipendenti** se e solo se $P(A \cap B) = P(A)P(B)$

La (P10) è una definizione più che una proprietà. Il concetto di indipendenza tra due eventi è intuitivo e traduce il fatto che due eventi non abbiano effetto l'uno sull'altro. L'esempio classico è il lancio ripetuto di una moneta: i risultati in lanci diversi devono essere indipendenti! Siccome l'indipendenza tra eventi è molto importante nel calcolo della probabilità, è giusto darne una definizione rigorosa. Per non confonderla con il concetto intuitivo di indipendenza a volte si preferisce parlare esplicitamente di **indipendenza stocastica**. La (P10) sarà quindi uno strumento utile per verificare la presunta indipendenza di eventi complessi e per svelare l'indipendenza stocastica di eventi apparentemente correlati.

2. Spazi di probabilità discreti

La teoria della probabilità nasce nel 1654 da una corrispondenza tra Pascal e Fermat su alcuni giochi d'azzardo in uso a quel tempo, giochi che prevedevano soltanto un numero finito di esiti possibili. Cominciare il nostro studio dagli spazi di probabilità discreti (cioè con un numero finito di punti campione) è dovuto comunque ad esigenze didattiche più che a influenze storicistiche.

Dato uno spazio campione finito $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, possiamo scegliere come σ-algebra l'insieme delle parti di Ω, cioè la famiglia costituita da tutti i sottoinsiemi di Ω.

Ogni evento A avrà quindi solo un numero finito di casi favorevoli, cioè di punti campione che implicano il verificarsi di A. Se $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_r}\}$, allora dall'additività di P e

dall'incompatibilità di esiti diversi ricaviamo

$$P(A) = P(\omega_{i_1}) + P(\omega_{i_2}) + \cdots + P(\omega_{i_r})\,.$$

Vediamo pertanto che ogni misura di probabilità P su uno spazio campione finito è determinata dai suoi valori sui punti. Con un po' di abuso delle notazioni, confondendo cioè eventi e sottoinsiemi dello spazio campione, possiamo scrivere la formula

$$P(A) = \sum_{\omega \in A} P(\omega) \tag{1}$$

che permette di esprimere la probabilità di ogni evento in termini delle probabilità degli eventi elementari.

Un esempio semplice e nello stesso tempo fondamentale è quello della **equiprobabilità**, quando cioè si suppone che ognuno degli esiti possibili abbia la stessa probabilità di accadere.

Si deduce che per ogni punto campione ω deve essere

$$P(\omega) = \frac{1}{|\Omega|}\,,$$

dove $|\Omega|$ indica la cardinalità di Ω. Quindi ritroviamo la definizione classica di probabilità

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{casi favorevoli}}{\text{casi possibili}}\,. \tag{2}$$

La semplicità della formula (2) non deve indurre a credere che sia facile calcolare la probabilità di ogni evento. In ultima analisi, questa formula riconduce il calcolo della probabilità su spazi discreti a problemi di conteggio e quindi alla *combinatoria*.

Una situazione tipica per un processo finito è quella del **campionamento**, cioè l'estrazione di un certo numero di unità o campioni da una popolazione fissata. L'esempio classico è l'estrazione di palline da un'urna (gioco del lotto, lancio di n dadi, ecc.). I possibili esiti della prima estrazione sono tanti quanti i campioni presenti nella popolazione. Se analizziamo invece la seconda estrazione ci rendiamo conto che vi sono differenze, ad esempio se pensiamo al gioco del lotto o al lancio di n dadi. Nel primo caso siamo in presenza di *estrazione senza reimbussolamento* (detto anche **campionamento senza rimessa**) nel secondo di *estrazione con reimbussolamento* o **campionamento con rimessa**.

Facciamo un esempio. Sia $\Omega = \{x, y, z\}$ e consideriamo, per brevità, il caso di due estrazioni successive. Indichiamo con Ω_r lo spazio campione con rimessa e con Ω_s lo spazio campione senza rimessa. Abbiamo allora

$$\Omega_r = \{xx, xy, xz, yx, yy, yz, zx, zy, zz\}\ , \quad \Omega_s = \{xy, xz, yx, yz, zx, zy\}\,.$$

A questo punto cerchiamo di calcolare la cardinalità degli spazi Ω_r e Ω_s nel caso generale di k estrazioni da una popolazione di n unità. Lo strumento più consono a questo scopo e la **regola fondamentale del calcolo combinatorio**:
se un oggetto si forma facendo una successione di k scelte tali che ci siano n_1 possibilità per la prima scelta, n_2 possibilità per la seconda scelta, ..., n_k possibilità per la k-esima scelta, allora il numero totale di oggetti che si possono così formare è il prodotto

$$n_1\, n_2\, \ldots n_k\,.$$

Ne segue facilmente che, nel caso di campionamenti di dimensione k (cioè con k estrazioni) da una popolazione di n unità, si ha

$$|\Omega_r| = n^k \quad , \qquad |\Omega_s| = n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!}\,,$$

dove $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$ è l'usuale notazione per il fattoriale di n.

Esercizi.
Menu al ristorante. Un ristorante offre una scelta tra tre antipasti, cinque primi, quattro secondi, tre contorni e tre dolci. Quanti pranzi completi (senza bis) distinti possono essere ordinati? [540]

Insieme delle parti. Quanti sono i sottoinsiemi di un insieme con n elementi? $[2^n]$

Cinquine al lotto. Quante sono le possibili cinquine su una ruota del lotto? [43 949 268]

Il problema dei compleanni. In un'aula ci sono n studenti. Qual è la probabilità che almeno due abbiano lo stesso compleanno? $[1 - \frac{365!}{(365-n)!365^n}]$

Scomposizione di numeri. In quanti modi possiamo scrivere il numero n come somma di k numeri interi positivi ordinati? $[\frac{(n-1)!}{(k-1)!(n-k)!}]$

Estrazioni indipendenti. Considerati i due eventi $A = \{$ i-esimo esito alla j-esima estrazione$\}$ e $B = \{$ h-esimo esito alla k-esima estrazione$\}$, verificare la loro indipendenza o dipendenza stocastica nel caso di campionamenti con o senza rimessa.

3. Probabilità condizionata

Spesso nella vita reale vogliamo valutare la probabilità di un evento avvantaggiandosi della conoscenza parziale dell'esito dell'esperimento.

Un esempio è valutare la probabilità di fare 10 lanciando tre dadi. Semplici calcoli ci mostrano che tale probabilità (nel caso di un dado non truccato) è 1/8. Supponiamo di aver lanciato i dadi e di vedere che un dado si ferma indicando il numero 3. Nel breve attimo che precede l'arresto degli altri dadi possiamo sfruttare questa informazione e rivalutare la probabilità di fare 10. Questa è pari alla probabilità di fare 7 con due dadi e quindi è pari a 1/6. Quindi l'informazione avuta ha modificato la probabilità dell'evento in questione.

Non sempre è così semplice decidere quale cambiamento sia prodotto dall'informazione acquisita. A tal proposito citiamo il cosiddetto *paradosso del carceriere*.

Tre condannati a morte A, B e C vengono informati che due di loro sono stati graziati ed uno solo sarà giustiziato. Comunque conosceranno il loro destino soltanto il giorno successivo, fissato per l'esecuzione. Il condannato A tornando nella sua cella chiede al carceriere di rivelargli il nome del compagno di prigionia che sarà graziato. Il carceriere si rifiuta di dare questa informazione perché altrimenti la sua probabilità di essere giustiziato aumenterebbe, passando da 1/3 a 1/2. Il condannato A ribatte che, siccome almeno uno dei due suoi compagni sarà sicuramente graziato, venire a conoscenza di un nome non può alterare la sua probabilità di essere giustiziato. Chi dei due ha ragione?

Lo strumento ottimale per rispondere a questa domanda è la probabilità condizionata.

Dato uno spazio di probabilità $(\Omega, \Im, P)$ ed un evento H con probabilità non nulla, cerchiamo di valutare la probabilità di un evento A nell'*ipotesi* (o con la **condizione**) che H sia accaduto. Nell'esempio del lancio dei tre dadi H è l'evento un dado indica il numero 3. Ci aspettiamo che in generale la probabilità degli eventi cambi e quindi troviamo un nuovo nome a ciò che vogliamo definire. Chiamiamo P_H la probabilità condizionata da H. Se pensiamo alla probabilità su Ω come all'area di un sottoinsieme, allora siamo immediatamente spinti a riconoscere che la probabilità condizionata di A dato H dipende dall'area di $A \cap H$. Potremmo anche essere più rigorosi, osservando che $P_H(H^c) = 0$ e quindi che $P_H(A) = P_H(A \cap H)$ ed anche che

$P_H(A)P(H) = P(A \cap H)$. Comunque sia si può arrivare a concludere che deve valere la formula

$$P_H(A) = \frac{P(A \cap H)}{P(H)}\,, \qquad P(H) \neq 0\,.$$

La notazione usuale per $P_H(A)$ è $P(A|H)$, dove la stanghetta verticale separa l'evento di cui valutare la probabilità dalla condizione assunta.

La regola per calcolare la **probabilità condizionata** di A dato H è

$$P(A|H) = \frac{P(A \cap H)}{P(H)}\,, \qquad P(H) \neq 0\,. \tag{3}$$

La probabilità condizionata permette di dare definizioni alternative dell'indipendenza stocastica. Le seguenti tre affermazioni sono equivalenti:

(PC1) A e B sono indipendenti, cioè $P(A \cap B) = P(A)P(B)$
(PC2) $P(A|B) = P(A)$
(PC3) $P(B|A) = P(B)$.

In altre parole, l'informazione che un evento si è verificato non altera (chiaramente) la probabilità che si verifichi un evento indipendente.

Dalla (3) ricaviamo la formula $P(A \cap H) = P(A|H)P(H)$ e pertanto, considerando l'evento H^c e la probabilità condizionata dato H^c, otteniamo facilmente la relazione

$$P(A) = P(A \cap H) + P(A \cap H^c) = P(A|H)P(H) + P(A|H^c)P(H^c)\,.$$

Questa può essere generalizzata per arrivare ad una formula molto utile in varie applicazioni concrete. Supponiamo di avere diverse alternative H_i, cioè eventi H_i che verificano

(LA1) $H_i \cap H_j = \emptyset$, per $i \neq j$ (incompatibilità)
(LA2) $\bigcup_i H_i = \Omega$ (esaustività)
(LA3) $P(H_i) \neq 0$ per ogni indice i.

Allora la **legge delle alternative** afferma che

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots = \sum_i P(A|H_i)P(H_i)\,. \tag{4}$$

Supponiamo adesso di aver sottoposto uno studente ad un test a risposta multipla (m risposte possibili di cui solo una corretta). Immaginiamo che lo studente abbia probabilità p di conoscere la risposta esatta e non la conosca con probabilità $1 - p$. Chiaramente possiamo assumere che conoscendo la risposta azzeccherà certamente quella esatta mentre, nel caso che non la conosca, abbia probabilità $1/m$ di indovinare (completamente a caso). Nell'ipotesi che abbia risposto esattamente al test, qual è la probabilità che conosca la risposta?

Questo è un semplice esempio in cui si vuole invertire quello che appare il naturale susseguirsi delle scelte. Spieghiamoci meglio. In questo problema compaiono due scelte casuali: sapere o non sapere la risposta e indovinare o non indovinare la risposta. Nel formulare le ipotesi fatte abbiamo, più o meno esplicitamente, dato una valenza di causa e effetto alle singole scelte. Ci è parso naturale assumere la probabilità di indovinare data la conoscenza dello studente, mentre la domanda chiede esattamente l'opposto, cioè determinare la probabilità della conoscenza data la correttezza della risposta al test.

In termini semplificativi, chiediamo la probabilità di una causa sapendo l'effetto prodotto.

Facciamo un altro esempio. Tre artigiani confezionano in un giorno n_1, n_2 e n_3 borse delle quali, rispettivamente, d_1, d_2 e d_3 difettose. Scelta una borsa a caso, scopriamo che è difettosa. Qual è la probabilità che sia stata confezionata dal primo artigiano?

Anche qui, è naturale definire la probabilità che una delle borse confezionate dal singolo artigiano sia difettosa, mentre chiedere la probabilità che una borsa difettosa sia stata confezionata da un certo artigiano ci appare un ragionamento inverso.

Analizziamo bene quest'ultimo esempio. Il nostro spazio campione è composto dalle n borse, con $n = n_1 + n_2 + n_3$. L'esperimento consiste nello sceglierne una a caso (ipotesi di equiprobabilità) ed abbiamo a che fare con i seguenti eventi:

$A_i = \{$ la borsa è stata confezionata dall'i-esimo artigiano$\}$,

$D = \{$ la borsa è difettosa$\}$.

Sappiamo che $P(A_i) = \frac{n_i}{n}$, per $i = 1, 2, 3$ ed inoltre $P(D|A_i) = \frac{d_i}{n_i}$. Ciò che cerchiamo invece è $P(A_1|D)$. Nel gergo tecnico, $P(A_1)$ è detta *probabilità a priori* mentre $P(A_1|D)$ è detta *probabilità a posteriori*.

Dalla definizione di probabilità condizionata ricaviamo

$$P(A_1|D) = \frac{P(A_1 \cap D)}{P(D)} = \frac{P(D|A_1)P(A_1)}{P(D)}\,.$$

Utilizzando la legge delle alternative arriviamo alla conclusione

$$P(A_1|D) = \frac{P(D|A_1)P(A_1)}{\sum_i P(D|A_i)P(A_i)} = \frac{d_1}{d_1 + d_2 + d_3}\,.$$

Questo è un esempio molto semplice, in cui potevamo arrivare alla soluzione senza scomodare troppe regole e definizioni.

Nel caso generale il ragionamento è del tutto analogo ed il risultato è una formula che va sotto il nome di **legge di Bayes**:

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_j P(A|H_j)P(H_j)}\,, \tag{5}$$

dove le H_i sono alternative e quindi incompatibili ed esaustive (vedi legge delle alternative).

Vediamo a questo punto come si risolve il problema dello studente davanti ad un test a risposta multipla.

Consideriamo i due eventi:

$C = \{$ lo studente conosce la risposta$\}$,

$R = \{$ lo studente risponde esattamente$\}$.

Abbiamo $P(C) = p$, $P(R|C) = 1$ e $P(R|C^c) = 1/m$. Per calcolare $P(C|R)$ applichiamo la legge di Bayes e scopriamo

$$P(C|R) = \frac{P(R|C)P(C)}{P(R|C)P(C) + P(R|C^c)P(C^c)} = \frac{p}{p + (1-p)/m} = \frac{mp}{mp - p + 1}\,.$$

Nel caso di un test con 5 risposte possibili (di cui una sola corretta), se $p = 1/2$, cioè lo studente conosce la metà degli argomenti del test, allora $P(C|R) = 5/6$, che a parole significa che una risposta giusta su sei è in media dovuta al caso. Se $p = 3/4$ allora, in media, soltanto una risposta esatta su 16 è dovuta al caso.

Torniamo adesso al paradosso del carceriere. La vera difficoltà sta nel tradurre correttamente l'enunciato un po' vago del problema in termini probabilistici. Un primo passo utile può essere quello di complicare le cose, anche se pare assurdo. Supponiamo che i condannati fossero 50 (non stiamo ad assegnare un nome ciascuno per ovvi motivi) e che uno solo verrà giustiziato. Il prigioniero A poteva in questo caso chiedere il nome di 49 suoi compagni che avevano ricevuto la grazia. Sarebbe salita ad $1/2$ la sua probabilità di essere giustiziato? Chi di noi nei panni di A avrebbe avuto il coraggio di chiedere lo scambio con quell'unico prigioniero non nominato? Appare più naturale credere che l'informazione ricevuta non alteri le probabilità di A. Eppure, a ben vedere, potremmo pensare che il carceriere cominci ad elencare i graziati in ordine alfabetico (o di numero di matricola). Se salta un unico nome, allora anche noi, nei panni di A avremmo forse un sospiro di sollievo. Quindi istintivamente l'informazione ricevuta potrebbe cambiare le

probabilità precedenti. Questo esempio con molti condannati chiarisce meglio un punto che si rivela fondamentale e che nella formulazione iniziale è del tutto vago. Supponiamo che B e C siano entrambi graziati. Quale nome pronuncerebbe il carceriere?

Per tradurre in termini probabilistici corretti, si possono considerare i seguenti eventi:

$G_A = \{A \text{ sarà giustiziato}\}$,
$G_B = \{B \text{ sarà giustiziato}\}$,
$G_C = \{C \text{ sarà giustiziato}\}$,
$N_B = \{\text{ il carceriere rivelerà il nome di } B\}$,
$N_C = \{\text{ il carceriere rivelerà il nome di } C\}$.

Per ipotesi, ribadita anche dalle parole del carceriere, $P(G_A) = P(G_B) = P(G_C) = 1/3$.

Per quanto concerne invece le probabilità degli ultimi due eventi, nulla si può evincere dal testo del problema nel caso che A venga giustiziato. Supponiamo dunque che $P(N_B|G_A) = p$ e $P(N_C|G_A) = 1 - p$, cioè che, nel caso che B e C siano entrambi graziati, il carceriere riveli il nome di B con probabilità p.

Vogliamo valutare la probabilità condizionata di G_A dati rispettivamente gli eventi N_B e N_C. Dalla (5) segue facilmente:

$$P(G_A|N_B) = \frac{P(N_B|G_A)P(G_A)}{P(N_B|G_A)P(G_A) + P(N_B|G_B)P(G_B) + P(N_B|G_C)P(G_C)} .$$

Osservando che $P(N_B|G_B) = 0$ e $P(N_B|G_C) = 1$ ricaviamo

$$P(G_A|N_B) = \frac{p}{p+1}$$

ed analogamente

$$P(G_A|N_C) = \frac{1-p}{2-p} .$$

Osserviamo che ognuna delle due probabilità è $1/3$ solo nel caso $p = 1/2$. Il ragionamento di A era quindi giusto nell'ipotesi $p = 1/2$. A suo favore potremmo dire che, non conoscendo p, la valutazione migliore è proprio $1/2$. Invece per $p = 0$ oppure per $p = 1$, il ragionamento del carceriere acquista di significato e la probabilità a posteriori di G_A risulta 0 o $1/2$ a seconda della risposta.

Un esempio che sorprende spesso gli studenti è il seguente. Un test antitumorale, come quasi tutti i test diagnostici, non è infallibile e commette due tipi di errori: i cosiddetti *falsi positivi* e *falsi negativi*. I primi sono esiti positivi per pazienti sani mentre i secondi sono esiti negativi per pazienti affetti dalla malattia in esame. La probabilità che un test azzecchi la giusta diagnosi è detta *accuratezza*. La probabilità di errore è in generale diversa tra pazienti sani e pazienti malati. Si chiama *sensibilità* del test la probabilità che fornisca esito positivo in presenza di malattia e *specificità* del test la probabilità che dia esito negativo su un soggetto sano. Supponiamo adesso che un test antitumorale con sensibilità del 98% e specificità del 99% dia esito positivo sul paziente X. Sapendo che la malattia ha un'incidenza dello 0,2% sulla popolazione, che probabilità ha X di essere affatto dalla malattia?

Indichiamo con E l'evento il test ha dato esito positivo e con T l'evento il paziente ha il tumore.

Le nostre informazioni sono: $P(T) = 2/1000$, $P(E|T) = 98/100$ e $P(E|T^c) = 1/100$.

Vogliamo calcolare $P(T|E)$. Dalla (5) otteniamo

$$P(T|E) = \frac{P(E|T)P(T)}{P(E|T)P(T) + P(E|T^c)P(T^c)} = \frac{0,98 \cdot 0,002}{0,98 \cdot 0,002 + 0,01 \cdot 0,998} = 0,1641 .$$

Quindi il paziente risultato positivo al test ha una probabilità pari circa al 16,41% di avere un tumore.

4. Variabili aleatorie e funzioni distribuzioni

Introduciamo un nuovo concetto, quello di *variabile aleatoria*. Come sempre accade in matematica, i concetti vengono introdotti per semplificare ragionamenti usuali in certi campi, anche se inizialmente allo studente appare solo la difficoltà ad incamerare nuove definizioni.

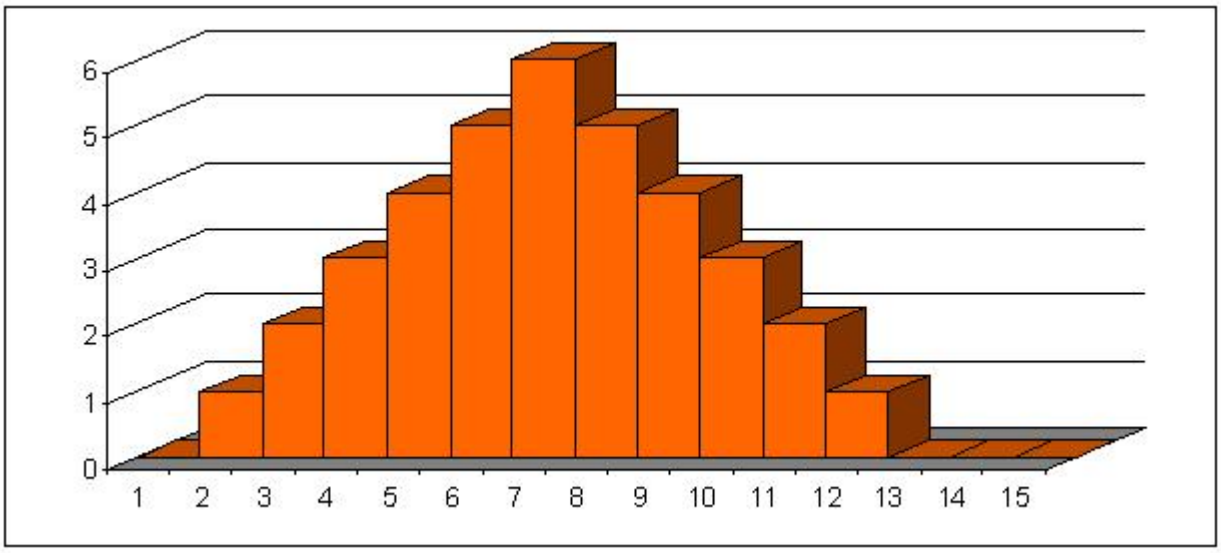
Abbiamo già discusso dell'esempio del lancio di un dado. I sei esiti possibili li abbiamo chiamati eventi, vi abbiamo definito una probabilità e ci siamo più o meno abituati a questa terminologia. Possiamo rileggere questo esempio dando un nome X al numero ottenuto lanciando il dado. Quindi X è un numero, compreso tra 1 e 6, ma non sappiamo quale. Chiamiamo X una variabile aleatoria (intera) e diciamo di conoscerla una volta che abbiamo deciso non solo i suoi valori possibili ma anche la probabilità che essa assuma i singoli valori.

In questo esempio non appare nessuna grande novità. Procediamo per gradi. Analizziamo il lancio di due dadi. Gli esiti possibili sono le coppie di numeri interi compresi tra 1 e 6, in tutto 36 eventi elementari. Abbiamo visto che con un po' di calcoli è semplice determinare la probabilità che la somma dei due numeri usciti sia un certo numero fissato (pensiamo a dadi non truccati, per semplicità). Chiamiamo X tale somma. La variabile aleatoria X non è più equivalente all'esito del lancio, cioè esiti diversi possono produrre lo stesso valore di X. Presentare X vuol dire elencare tutti i valori che può assumere con la relativa probabilità che ciò avvenga. Anche in questo caso la variabile aleatoria si dice intera, perché assume solo valori interi. Un modo di esibire X potrebbe essere la matrice

$$\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{1}{18} & \frac{1}{12} & \frac{1}{9} & \frac{5}{36} & \frac{1}{6} & \frac{5}{36} & \frac{1}{9} & \frac{1}{12} & \frac{1}{18} & \frac{1}{36} \end{pmatrix}$$

in cui ogni colonna riporta un valore possibile e la corrispondente probabilità dell'evento, dando per scontato che valori diversi hanno probabilità nulla.

Un modo più compatto ed efficace è quello di ricorrere ad un grafico. Nel caso della variabile aleatoria X abbiamo il seguente grafico, dove l'unità di misura sulle ordinate è espressa in trentaseiesimi.



Vediamo di generalizzare. Sia $(\Omega, \Im, P)$ uno spazio di probabilità; una **variabile aleatoria** è una funzione $X : \Omega \to \mathbb{R}$. Indichiamo con $\{X \leq k\}$ l'evento definito come l'insieme di tutti i punti campione $\omega \in \Omega$ per i quali $X(\omega) \leq k$. Per la precisione il nome di variabile aleatoria spetta soltanto a quelle funzioni tali che eventi di questo tipo appartengono a $\Im$. Comunque noi ci occuperemo sempre di funzioni che hanno questa proprietà e quindi possiamo pensare ad una variabile aleatoria come ad una qualsiasi funzione a valori reali.

Diremo che una variabile aleatoria X è *intera* o *positiva* quando è tale come funzione. Nel caso del punteggio ottenuto col lancio di due dadi, ad esempio, la funzione è sia positiva che intera.

Abbiamo detto in precedenza che conoscere la variabile aleatoria X significa non solo sapere quali numeri reali sono possibili valori di X, ma anche conoscere la probabilità che ciò avvenga. Supponiamo che X assuma i valori 1, 2, ..., n: conoscere la variabile aleatoria X vuol dire sapere la probabilità degli eventi $\{X = k\}$, per ogni $k = 1, 2, \ldots, n$.

Posto

$$p_k = P(X = k)$$

la successione delle coppie di numeri $\{(k, p_k)\}$ costituisce la **distribuzione di probabilità** di X e possiamo rappresentarla tramite una tabella o matrice oppure tramite un grafico simile a quello visto nel caso del lancio di due dadi. Osserviamo che dovrà valere $\sum_k p_k = 1$.

In seguito saremo interessati anche a variabili aleatorie non intere ma *continue*, che possono assumere cioè tutti i valori reali. Un esempio potrebbe essere dato dal lancio del giavellotto. Anche se le misurazioni sono espresse in centimetri (e quindi possiamo considerarla una variabile aleatoria intera) i risultati possibili sono talmente tanti che conviene utilizzare le notazioni (e tecniche) delle variabili continue. Indichiamo con L la variabile aleatoria che esprime il risultato di un singolo lancio. Come possiamo esprimere la probabilità che L assuma un certo valore? Nell'ipotesi di valori reali, la probabilità di azzeccare esattamente il risultato è evidentemente bassissima, anzi nulla. Cosa significa allora in questo caso conoscere la variabile aleatoria?

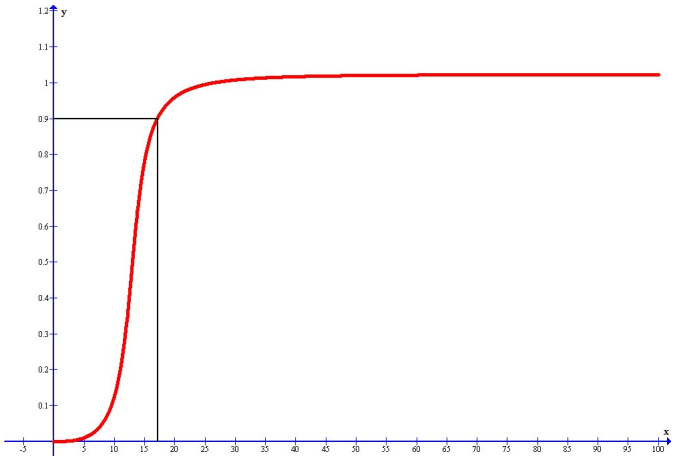
Per le variabili aleatorie continue, e quindi anche per L, gli eventi da prendere in esame non sono quelli del tipo $\{L = x\}$ ma quelli esprimibili come $\{L \leq x\}$, per ogni $x \in \mathbb{R}$.

Al variare di x consideriamo la funzione

$$F(x) = P(L \leq x)\,,$$

che chiamiamo la **funzione di distribuzione di probabilità** di L. Il grafico della funzione F ci fornisce tutte le informazioni che si possono desiderare sulla variabile aleatoria L. Osserviamo che dovrà valere

$$\lim_{x\to+\infty} F(x) = 1 \quad \text{e} \quad \lim_{x\to-\infty} F(x) = 0\,.$$



Supponiamo che il grafico a fianco rappresenti la funzione di distribuzione di probabilità nel lancio del giavellotto da parte di una persona. Potremmo pensare di aver chiesto ad un numero enorme di persone di lanciare il giavellotto ed aver quindi stimato le probabilità in questione in base alle frequenze del campione statisticamente rappresentativo. Il record mondiale del lancio del giavellotto è di 98,48 metri e quindi la probabilità che un lancio sia inferiore ai 100 metri deve necessariamente essere 1 se si basa su lanci già effettuati. In figura sono evidenziati due segmenti correlati alla domanda: quale distanza è irraggiungibile nel 90% dei lanci? Si parte orizzontalmente dallo 0,9 segnato sull'asse delle ordinate e, raggiunto il grafico, si scende fino a leggere circa 17 metri.

Pertanto la funzione di distribuzione di probabilità di una variabile aleatoria continua garantisce lo stesso tipo di informazioni fornite dalla distribuzione di probabilità di una variabile aleatoria intera (o discreta). Ciononostante i due grafici ci appaiono ben diversi. Da un punto di vista matematico il legame tra le due funzioni è molto chiaro: analizziamolo.

Nel definire la funzione di distribuzione di probabilità di una variabile aleatoria continua, ad esempio L, abbiamo evitato di definire la probabilità che un lancio sia esattamente di 84,60 metri (record italiano dal 1989). L'idea che abbiamo enfatizzato è che praticamente nessun lancio percorrerà esattamente quella distanza se prendiamo in considerazione i millimetri o addirittura i millesimi di millimetro. Quindi è la domanda stessa che non ha molto senso. Invece è naturale chiedere la probabilità che un lancio sia poi registrato pari a 84,60 metri da un giudice di gara. In altre parole ci disinteressiamo dell'errore che il giudice commette nell'approssimare il numero. Quindi la domanda potrebbe essere formulata meglio considerando l'evento che L sia compreso tra 84,595 e 84,605 metri. Questo ci porta ad utilizzare la variazione della funzione F più che la funzione stessa, cioè $F(84,605) - F(84,595)$. Infatti l'evento $\{L \leq b\} = \{a \leq L \leq b\} \cup \{L \leq a\}$ e quindi $P(a \leq L \leq b) = P(L \leq b) - P(L \leq a) = F(b) - F(a)$.

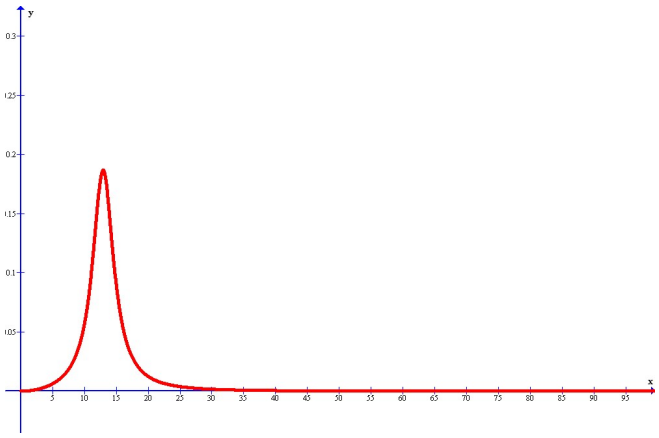
Quando si parla di variazioni di una funzione il concetto di derivata dovrebbe saltare in mente anche agli studenti. Definita dalla formula

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h},$$

si introduce la **densità di probabilità** della variabile aleatoria L. Dunque potremo scrivere

$$F(b) - F(a) = \int_a^b f(x)\, dx.$$

Osserviamo che risulta $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.



Questa funzione assume un significato molto simile a quello visto per le variabili aleatorie intere nel caso dei grafici a barra. Ad esempio, qui a fianco è riportato il grafico della densità di probabilità della variabile L, la cui distribuzione è rappresentata nella pagina precedente. Il picco in corrispondenza dei lanci di 10-15 metri rivela che i dati non sono reali oppure, volendo cercare una giustificazione, che il campione statistico preso in esame per valutare la funzione F non era certo rappresentativo di atleti della disciplina. La densità di probabilità f conserva ancora tutte le informazioni necessarie per rispondere a domande sulla probabilità di eventi espressi in termini della variabile aleatoria L. Ad esempio, la probabilità che un lancio sia compreso tra 10 e 15 metri si legge valutando l'integrale della funzione f sul corrispondente intervallo, cioè calcolando l'area della regione racchiusa dal grafico della f, dall'asse delle ascisse e dalle due rette $x = 10$ e $x = 15$. In formule

$$P(10 \leq L \leq 15) = \int_{10}^{15} f(x)\, dx = F(15) - F(10)\ (= 0,6598)\,,$$

cioè quasi 2 lanci su 3 mandano il giavellotto ad una distanza di soli 10-15 metri dalla pedana.

Torniamo ad un altro tipo di lancio, il lancio di una moneta o di un dado. Analizziamo la variabile aleatoria X che conta il numero di successi (di teste con una moneta o di pari alla roulette o altro ancora) in n lanci. Ogni lancio è indipendente dagli altri e, per mantenere maggiore generalità, immaginiamo che la probabilità di successo in un singolo lancio sia p (e quella di insuccesso sia q, con $p+q=1$). Conoscere X significa capire i suoi valori possibili e le corrispondenti probabilità. I valori possibili sono chiaramente tutti i numeri interi compresi tra 0 e n. Valutiamo adesso la probabilità di avere k successi negli n lanci e conseguentemente $n-k$ insuccessi. Qualsiasi sequenza a noi favorevole ha probabilità $p^k q^{n-k}$ di accadere; il numero di tali sequenze si conta facilmente e risulta pari a $\frac{n!}{k!(n-k)!}$.

Quindi la variabile aleatoria X è rappresentata dalla successione

$$p_k = P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} .$$

Questa distribuzione di probabilità prende il nome di **distribuzione binomiale** o di Bernoulli. Data la generalità del processo, questa distribuzione è molto comune. Quando il numero di lanci n è molto grande, non è agevole calcolare tutti i p_k e certi ragionamenti si semplificano pensando la variabile aleatoria X una variabile aleatoria continua. Praticamente il passaggio che facciamo è quello di sostituire il grafico ad istogrammi della distribuzione con il grafico di una funzione f, cioè di una densità di probabilità. Questo passaggio va definito meglio e ci torneremo quando parleremo della *legge dei grandi numeri*, uno dei capisaldi della statistica moderna, formulata per la prima volta da J. Bernoulli (1654-1705) nella sua opera postuma del 1713, *Ars conjectandi*.

Per il momento ricordiamo soltanto che il fattoriale, definito soltanto sui numeri naturali, può essere esteso ad ogni numero reale positivo tramite la funzione $\Gamma(x)$ data dalla formula

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt .$$

Infatti risulta (facilmente ricavabile dalla regola di integrazione per parti) $\Gamma(x+1) = x\Gamma(x)$ e quindi, verificato che $\Gamma(1) = 1$, vale la relazione $\Gamma(n+1) = n!$, per ogni $n \geq 0$.

Una seconda relazione concernente il fattoriale è la cosiddetta **formula di Stirling** che ne fornisce un'approssimazione asintotica:

$$\lim_{n\to\infty} \frac{n^n e^{-n}\sqrt{2\pi n}}{n!} = 1 .$$

5. Media e varianza di una variabile aleatoria

Continuiamo a considerare il processo di Bernoulli, cioè il lancio di una moneta. Supponiamo di aver assistito a 20 lanci e di aver visto uscire testa una sola volta. Siamo spinti a credere che la moneta sia truccata, cioè che la probabilità di ottenere testa in un singolo lancio non sia $1/2$ come pensavamo. Vedremo in seguito che questo esperimento può essere visto come un test di ipotesi, ma per il momento occupiamoci solo della nostra aspettativa o previsione. Immaginando di avere a che fare con una moneta equa, ci aspettiamo che esca testa circa nella metà dei lanci. Siamo disposti a credere che la casualità provochi un qualche discostamento dal valore preciso (10 in questo caso) ma ci insospettiamo se l'allontanamento è troppo evidente. Come abbiamo calcolato il valore preciso 10? Come possiamo distinguere un discostamento casuale da una truffa?

Proprio per rispondere a queste domande introduciamo nuovi concetti. Come calcolare quel valore 10. Tradotto in termini matematici, il problema è risolto da quella che viene chiamata **media** o **speranza matematica** o **valore atteso** di una variabile aleatoria.

Per darne una definizione precisa conviene distinguere le variabili aleatorie continue da quelle discrete. Se X è una variabile aleatoria (intera o discreta) che assume solo i valori x_1, x_2, …, x_n ed inoltre $p_k = P(X = x_k)$, allora il valore atteso di X è dato da

$$E(X) = \sum_{k=1}^{n} x_k p_k \ . \tag{6}$$

Se invece X è una variabile aleatoria continua con densità di probabilità $f(x)$, allora il valore atteso di X è dato da

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx , \tag{7}$$

se l'integrale improprio è convergente (cosa che noi supporremo sempre verificata).

Formalmente il valore atteso è una media ponderata dei valori assunti da X con pesi pari alla probabilità del singolo valore.

Teorema fondamentale della media. Date due variabili aleatorie X e Y risulta

$$E(X+Y) = E(X) + E(Y)\,. \tag{8}$$

Dimostrazione. Limitiamoci a considerare il caso di variabili aleatorie intere per non complicare inutilmente i ragionamenti. La variabile aleatoria $X+Y$ ha una sua distribuzione di probabilità: indichiamo con q_k la probabilità che $X+Y$ prenda il valore k. Non è semplicissimo vedere come la successione $\{q_k\}$ salti fuori a partire dalle distribuzioni di X e di Y. Conviene introdurre la cosiddetta distribuzione congiunta di X e Y, cioè la successione a due indici (come le matrici)

$$c_{a,b} = P(\{X=a\} \cap \{Y=b\})\,.$$

Adesso possiamo scrivere

$$q_k = P(X+Y=k) = \sum_{a+b=k} c_{a,b}\,,$$

dove l'ultima sommatoria è estesa a tutte le coppie di numeri interi a e b tali che la loro somma sia k. Pertanto, dalla definizione di media di una variabile aleatoria, risulta

$$\begin{aligned} E(X+Y) &= \sum_k k q_k = \sum_k k \sum_{a+b=k} c_{a,b} = \sum_a \sum_b (a+b) c_{a,b} \\ &= \sum_a a \sum_b c_{a,b} + \sum_b b \sum_a c_{a,b} = \sum_a a P(X=a) + \sum_b b P(Y=b) = E(X) + E(Y)\,. \end{aligned}$$

Per giustificare l'aggettivo *fondamentale* dato a questo teorema dobbiamo vederne qualche applicazione.

Sia X il numero di successi su n lanci di moneta. Supponiamo che la probabilità di successo in ogni singolo lancio sia p. Abbiamo visto che X assume i valori tra 0 e n ed inoltre che

$$P(X=k) = p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\,.$$

Pertanto possiamo scrivere $E(X) = \sum_{k=0}^{n} k p_k$. Svolgere questo calcolo non è affatto semplice.

Ragioniamo in modo diverso. Indichiamo con X_k la variabile aleatoria che conta i successi al k-esimo lancio. Chiaramente X_k può assumere soltanto i valori 0 e 1. Sappiamo anche che il valore 1 è assunto con probabilità p. Quindi

$$E(X_k) = 0 \cdot (1-p) + 1 \cdot p = p\,, \text{ per ogni } \ k\,,$$

ossia il valore atteso di X_k è proprio p. Cogliamo l'occasione per sottolineare che il valore atteso non è il valore più probabile! Adesso, osservando che $X = X_1 + X_2 + \cdots + X_n$, il teorema della media ci garantisce che

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = np\,,$$

confermando il risultato immaginabile: su n lanci ci aspettiamo np successi.

Il concetto di media è abbastanza intuitivo ed inoltre è così usuale che sappiamo bene quali indicazioni ci fornisce a proposito di una variabile aleatoria. La nostra esperienza ci suggerisce anche che variabili diverse possono avere la stessa media e ciononostante caratteristiche diverse. Ad esempio, la media del consumo annuo procapite di acqua potabile può essere la stessa in due regioni diverse e non di meno queste stesse regioni possono presentare problemi diversi circa la distribuzione delle risorse idriche tra la popolazione. Ad esempio, la percentuale di abitazioni non raggiunte dalla rete idrica può essere notevolmente diversa.

Quindi la conoscenza della media di una variabile aleatoria non svela ciò che potremmo chiamare la dispersione dei valori intorno alla media stessa. Insomma, stiamo cercando di

rispondere alla domanda precedentemente posta. Come possiamo distinguere uno scostamento dalla media dovuto al caso da uno dovuto ad una truffa, cioè ad una distribuzione di probabilità diversa da quella immaginata?

La misura di dispersione maggiormente utilizzata è la **varianza**. La varianza di X è definita come il valore atteso del quadrato della distanza di X dalla sua media. In formule

$$Var(X) = E\left((X-\mu)^2\right) \ , \quad \text{con} \quad \mu = E(X)\,. \tag{9}$$

La radice quadrata della varianza è la cosiddetta **deviazione standard** o **scarto quadratico medio** della variabile X e si scrive generalmente ricorrendo al simbolo σ

$$\sigma(X) = \sqrt{Var(X)}\,. \tag{10}$$

Le formule per calcolare la varianza in termini della distribuzione di probabilità nel caso discreto e della densità di probabilità nel caso continuo sono semplici conseguenze della definizione. Se X è discreta, $E(X) = \mu$ e $P(X = x_k) = p_k$, allora

$$Var(X) = \sum_k (x_k - \mu)^2 p_k\,. \tag{11}$$

Se X ha densità di probabilità $f(x)$ e $E(X) = \mu$, allora vale

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx\,. \tag{12}$$

Una formula (detta anche Teorema di König) utile in molti casi è la seguente

$$Var(X) = E(X^2) - E(X)^2\,. \tag{13}$$

La dimostrazione è molto semplice

$$Var(X) = E\left((X-\mu)^2\right) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2\,.$$

Tornando al nostro esempio dei 20 lanci di una moneta, possiamo calcolare la varianza del numero di successi X dalle formule appena viste.

$$Var(X) = \sum_{k=0}^{20} (k-10)^2 p_k = \sum_{k=0}^{20} k^2 p_k \ - 10^2\,.$$

Ancora una volta abbiamo qualche difficoltà a svolgere i calcoli.

Potremmo cercare di ripetere il ragionamento fatto per calcolare il valore atteso, cioè utilizzare la scrittura $X = X_1 + X_2 + \cdots + X_{20}$, ma ci manca una formula per la varianza di una somma di variabili aleatorie. Vedremo che questa formula esiste, è semplice, ma vale soltanto se le variabili aleatorie sono indipendenti.

Due variabili aleatorie X e Y si dicono **indipendenti** quando sono indipendenti gli eventi $\{X \le a\}$ e $\{Y \le b\}$, per ogni coppia di numeri reali a e b. In altri termini, la probabilità che X sia minore (o uguale o maggiore) di un certo numero è indipendente dal valore assunto da Y.

Date due variabili aleatorie X e Y, si definisce la **covarianza** di X e Y tramite la formula

$$Cov(X,Y) = E(XY) - E(X)E(Y)\,. \tag{14}$$

Teorema della covarianza. Se X e Y sono variabili aleatorie indipendenti, allora

$$Cov(X,Y) = 0 \ , \quad \text{ovvero } E(XY) = E(X)E(Y)\,.$$

Dimostrazione. Limitiamoci a considerare variabili aleatorie discrete, dato che nel caso continuo dovremmo utilizzare qualche tecnicismo degli integrali. L'evento $\{XY = n\}$ è una unione disgiunta

degli eventi $\{X = a\} \cap \{Y = b\}$ al variare dei numeri a e b tali che $ab = n$. Poiché X e Y sono indipendenti possiamo scrivere

$$P(XY = n) = \sum_{ab=n} P(X = a)P(Y = b)\,.$$

Adesso passando alle medie otteniamo

$$\begin{aligned} E(XY) &= \sum_n nP(XY = n) = \sum_n n \sum_{ab=n} P(X = a)P(Y = b) \\ &= \sum_n \sum_{ab=n} aP(X = a)bP(Y = b) = \sum_a aP(X = a) \sum_b bP(Y = b) \\ &= E(X)E(Y)\,. \end{aligned}$$

Teorema della varianza. Se X e Y sono variabili aleatorie indipendenti, allora

$$Var(X + Y) = Var(X) + Var(Y)\,.$$

Dimostrazione. Questa è una banale conseguenza del risultato precedente. Infatti

$$\begin{aligned} Var(X + Y) &= E\left((X + Y)^2\right) - E(X + Y)^2 = E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\ &= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2Cov(X, Y) = Var(X) + Var(Y)\,. \end{aligned}$$

Altre semplici proprietà della varianza sono

$$Var(X + k) = Var(X) \qquad Var(kX) = k^2 Var(X)\,, \tag{15}$$

per ogni numero k.

Continuiamo la nostra analisi dei 20 lanci di moneta. La variabile aleatoria X definita dal numero di successi ottenuti nei 20 lanci la vediamo come somma delle variabili aleatorie X_k che contano i successi (0 o 1) al k-esimo lancio. Le X_k sono chiaramente indipendenti e quindi il teorema della varianza ci assicura che $Var(X) = 20\,Var(X_1)$. Essendo

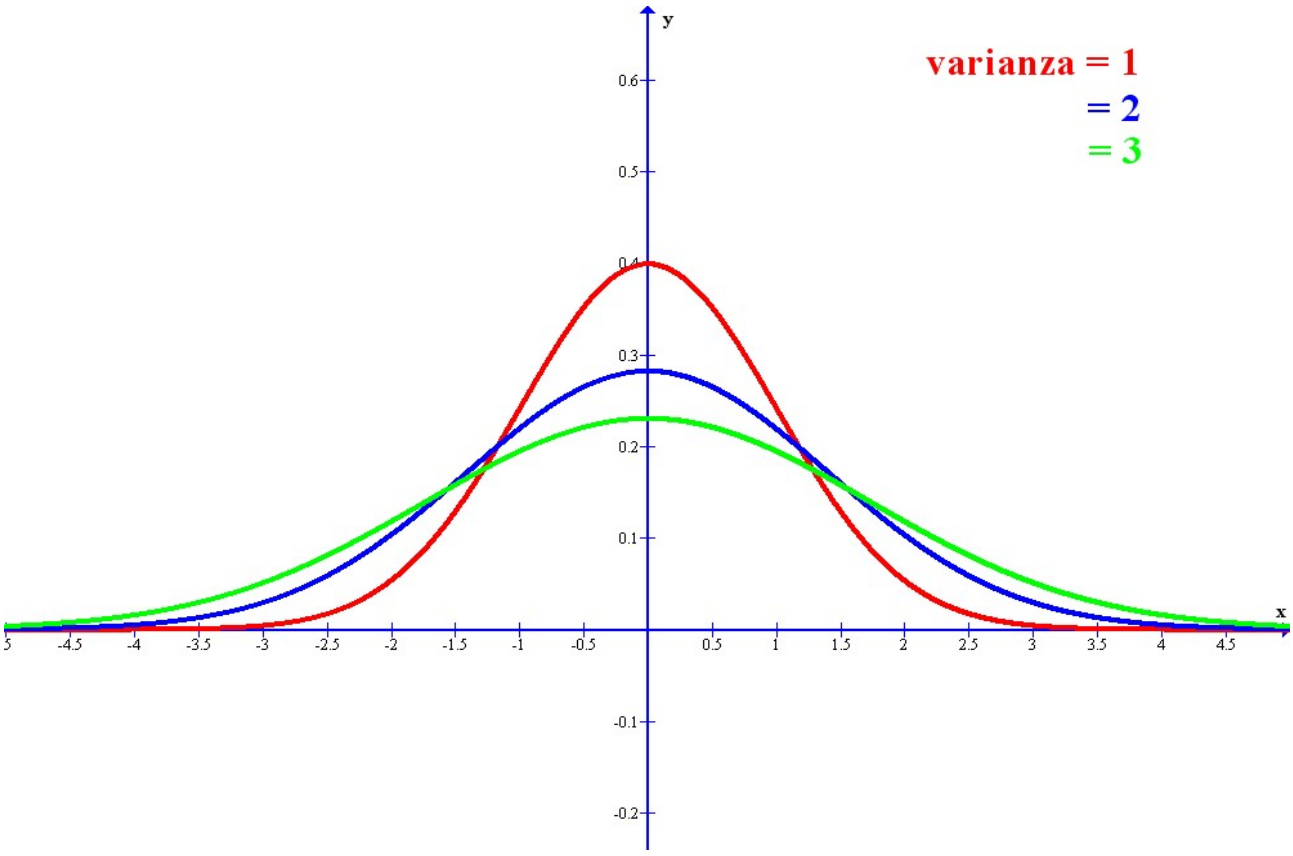
$$Var(X_k) = (0 - p)^2(1 - p) + (1 - p)^2 p = p(1 - p)\,,$$

ricaviamo $Var(X) = 20p(1 - p)$. Nel caso di una moneta equa è $p = 1/2$ e quindi $Var(X) = 5$. Se vogliamo confrontare lo scostamento sperimentale dalla media, l'unità di misura da utilizzare è lo scarto quadratico medio; in questo esempio $\sigma(X) = \sqrt{Var(X)} \simeq 2,23$. Dato che 2,23 è lo scarto quadratico medio, possiamo aspettarci che uno scostamento minore o uguale a 2σ sia del tutto normale, mentre differenze maggiori possono insospettirci. Come rendere rigorosi questi ragionamenti sarà l'argomento del prossimo capitolo.

Introduciamo adesso la più nota funzione di distribuzione di probabilità di una variabile aleatoria: la **distribuzione normale** o **Gaussiana**.

Una variabile aleatoria X ha distribuzione normale con media μ e varianza σ^2 se la sua densità di probabilità è data da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,. \tag{16}$$



La figura sopra riporta i grafici della funzione f per $\mu = 0$ e per tre diversi valori della varianza σ^2. Valori diversi di μ comportano soltanto una traslazione della figura: il valore massimo della funzione è sempre assunto nel valore atteso μ.

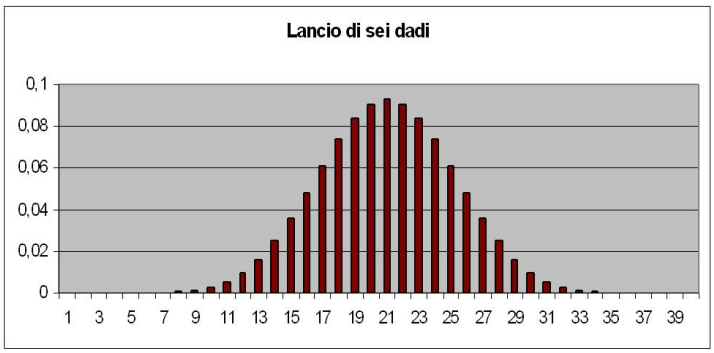
Per prima cosa dovremmo verificare che la funzione data è una densità di probabilità e che realmente la media e la varianza sono quelle volute. Queste affermazioni equivalgono ad espliciti calcoli che fanno intervenire integrali impropri. Alla base di tutto sta l'identità

$$\int_{-\infty}^{+\infty} e^{-x^2}\, dx = \sqrt{\pi}\,,$$

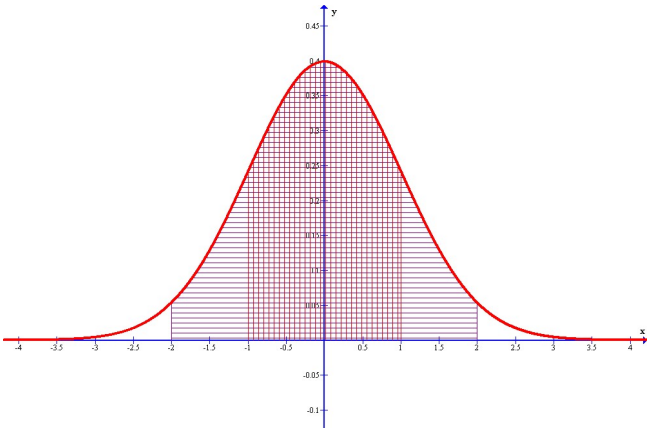
che qualcuno avrà forse incontrato in precedenti corsi di matematica.

La distribuzione normale è usualmente utilizzata per modelizzare l'errore commesso in una qualsiasi misurazione. La media μ rappresenta la misura esatta. Il fatto che la densità sia simmetrica rispetto a μ corrisponde all'osservazione sperimentale che l'errore è per eccesso o per difetto con la medesima probabilità. La varianza, o meglio la sua radice quadrata, cioè la deviazione standard, modula l'errore medio commesso. Al diminuire della varianza, il grafico della f diventa più ripido ed aumenta il valore puntuale in μ.

Nella figura a lato è rappresentata la distribuzione di probabilità della variabile aleatoria somma dei risultati di sei dadi. La somiglianza con la densità della distribuzione normale è notevole. Nel prossimo capitolo vedremo che questa somiglianza è così comune da dirsi appunto *normale*! Proprio questa particolarità rende la distribuzione Gaussiana fondamentale per la statistica.



Vedremo che sarà molto importante conoscere la probabilità di eventi del tipo $\{|X-\mu| \le t\sigma\}$.



Qui a fianco sono evidenziate le regioni le cui aree misurano tali probabilità per $t = 1$ e $t = 2$.

La probabilità che X differisca dalla media per meno di σ è $0,6827$, cioè il 68,27%.

La probabilità che X differisca dalla media per meno di due deviazioni standard è $0,9545$.

La probabilità che X differisca dalla media per meno di tre deviazioni standard è $0,9973$.

Altri numeri utili sono: il 95% dell'area del sottografico si trova a distanza inferiore di 1,96 deviazioni standard dalla media; il 99% si trova a distanza minore di 2,58 deviazioni standard dalla media. Una tabella che riporta i principali valori della distribuzione normale è riportata nell'ultima di queste pagine.

6. Disuguaglianza di Chebyshev, legge dei grandi numeri e teorema centrale

Come abbiamo visto, è naturale aspettarsi che i valori di una variabile aleatoria X si dispongano intorno alla media $\mu(X)$ e che una unità di misura appropriata a X per valutare il discostamento dalla media sia la deviazione standard $\sigma(X)$. Vorremmo a questo punto stimare la probabilità che X ha di differire dalla media per più di k volte $\sigma(X)$.

Lo strumento adatto è la cosiddetta **disuguaglianza di Chebyshev**. Testi diversi riportano scritture diverse del nome Chebyshev. Ciò è dovuto a scelte diverse di traslitterazione dal cirillico e non di attribuzione del risultato. P.L. Chebyshev (1821-1894) dette un enorme contributo allo sviluppo della teoria della probabilità e fu il fondatore della scuola di Pietroburgo, scuola che annoverò tra i suoi aderenti matematici come Liapunov e Markov.

Disuguaglianza di Chebyshev. Sia X una variabile aleatoria e siano $E(X) = \mu$ e $Var(X) = \sigma^2$. Allora, per ogni $t > 0$, risulta

$$P(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}\,. \tag{17}$$

Dimostrazione. Supponiamo che X sia una variabile aleatoria continua con densità di probabilità $f(x)$. Allora

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx \geq \int_{|x-\mu|\geq t\sigma} (x-\mu)^2 f(x)\,dx \\ &\geq \int_{|x-\mu|\geq t\sigma} t^2\sigma^2 f(x)\,dx = t^2\sigma^2 P(|X-\mu|\geq t\sigma)\,, \end{aligned}$$

da cui segue la tesi. Nel caso in cui X sia una variabile aleatoria discreta si può procedere in modo analogo.

Sostituendo t nella (17) con k/σ otteniamo la versione equivalente $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$.

Quest'ultima implica la ben nota **legge dei grandi numeri**, spesso utilizzata a sproposito.

Legge dei grandi numeri. Siano $X_1, X_2, \ldots$ variabili aleatorie indipendenti e con la stessa distribuzione di probabilità. Indichiamo con μ la loro media e con σ^2 la loro varianza. Allora, per ogni $\varepsilon > 0$, risulta

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0\,. \tag{18}$$

Dimostrazione. Consideriamo la variabile aleatoria $S_n = \frac{X_1+X_2+\cdots+X_n}{n}$. Per il teorema fondamentale della media e per il teorema della varianza, abbiamo che $E(S_n) = \mu$ e $Var(S_n) = \frac{\sigma^2}{n}$.

La tesi segue adesso dalla disuguaglianza di Chebyshev.

La legge dei grandi numeri permette di affermare, ad esempio, che la probabilità di successo nel lancio di una moneta è pari alla *frequenza* dei successi in n prove ripetute, cioè al rapporto tra il numero di successi ed il numero di prove effettuate, quando n tende all'infinito. Ciò che non indica è il numero di prove necessarie per avere una buona approssimazione.

Supponiamo che la probabilità di successo in ogni singolo lancio sia p. La frequenza dei successi in n lanci è la variabile aleatoria $S_n = \frac{X_1+X_2+\cdots+X_n}{n}$ dove, come al solito, X_k conta i successi al k-esimo lancio. Abbiamo visto che S_n ha valore atteso p e varianza $\frac{p(1-p)}{n}$. La disuguaglianza di Chebyshev ci dice che, volendo ad esempio essere sicuri al 99% (cioè disposti a sbagliare con probabilità 1/100), possiamo affermare che la differenza tra S_n e p è minore di $10\sqrt{p(1-p)}/\sqrt{n}$. Se vogliamo valutare p con un errore massimo di $0,05$, cioè $1/20$, dovremo prendere n così grande che $10\sqrt{p(1-p)}/\sqrt{n} \leq 1/20$. Tale stima dipende da p, ma se osserviamo che $p(1-p)$ vale al più $1/4$, troviamo che è sufficiente prendere $n \geq 10\,000$.

Questa stima può essere notevolmente migliorata. Ciò non dovrebbe stupire se si tiene conto del tipo di ragionamenti utilizzati nel dimostrare la disuguaglianza di Chebyshev. L'ulteriore passo è rappresentato dal teorema centrale che, nel caso di esperimenti ripetuti, mostra la funzione di distribuzione di probabilità che si ottiene al crescere del numero degli esperimenti: la distribuzione normale.

Teorema centrale. Siano X_1, X_2, ... variabili aleatorie indipendenti e con la stessa distribuzione di probabilità. Indichiamo con μ la loro media e con σ^2 la loro varianza. Allora vale

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{x^2}{2}}\, dx\,. \tag{19}$$

La stima fatta in precedenza può essere migliorata nel seguente modo. La variabile che compare nella (19) è $(S_n - \mu)\sqrt{n}/\sigma$. Volendo trovare n tale che $P(|S_n - p| > \frac{1}{20}) \leq \frac{1}{100}$, è sufficiente scrivere

$$\begin{aligned} P(|S_n - p| > \frac{1}{20}) &= P\left(|S_n - p|\sqrt{n}/\sigma > \frac{\sqrt{n}}{20\sigma}\right) \\ &= P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} > \frac{n}{20\sqrt{p(1-p)}}\right) \to 1 - \frac{1}{\sqrt{2\pi}} \int_{-\frac{n}{10}}^{\frac{n}{10}} e^{\frac{x^2}{2}}\, dx \end{aligned}$$

dove abbiamo utilizzato $E(S_n) = p$, $Var(S_n) = \frac{p(1-p)}{n}$ e $p(1-p) \leq 1/4$. Osservando adesso la tabella in fondo a queste note troviamo che è sufficiente prendere $n/10 \geq 2,58$, cioè $n \geq 26$. A dire il vero non abbiamo affatto tenuto conto del limite che compare nel teorema centrale. Il teorema non afferma che per ogni n vale una certa cosa ma che questa vale al tendere di n all'infinito. Nella pratica si osserva che già dopo pochi lanci di una moneta equa la distribuzione di probabilità si adatta benissimo alla forma a campana della distribuzione normale. L'approssimazione è peggiore per valori di p diversi da $1/2$.

Nell'esempio appena fatto, la probabilità con cui eravamo disposti a sbagliare era fissata all'1%. Questo si chiama il **livello di significatività** del nostro test. Se il nostro test lo avevamo portato avanti per controllare l'equità di una moneta utilizzata da un amico, fissare il livello di significatività vuol dire decidere con quale probabilità siamo disposti ad accusare di truffa l'amico nel caso che la moneta sia equa. Tale numero dipende quindi dall'amicizia, dalla rilevanza del test o quant'altro: comunque sia dipende da fattori che non hanno niente a che vedere con la probabilità o la statistica.

I test del tipo descritto sono così comuni che gli statistici utilizzano una terminologia particolare per due livelli di significatività specifici: 1% e 5%. Un esperimento è detto **significativo** se ci permette di rifiutare il modello congetturato avendo posto il livello di significatività pari al 5%; è detto **molto significativo** se ci permette il rifiuto con il livello di significatività pari all'1%.

Un altro concetto simile al livello di significatività è quello dell'**intervallo di confidenza**. Tutti abbiamo sentito parlare di intervalli di confidenza quando ascoltiamo il susseguirsi di proiezioni dopo una tornata elettorale. Per chiarire il significato facciamo un esempio diverso, dove ancora interviene un processo di Bernoulli.

Supponiamo che due compagnie aeree in competizione abbiano in programma un volo su una stessa tratta allo stesso orario. Entrambe sanno che il numero di passeggeri sulla tratta è costantemente 1000 e che la scelta di ogni passeggero è completamente casuale, indipendente dalle scelte altrui e che privilegia ogni compagnia con probabilità 1/2. Nessuna compagnia ha interesse a rifiutare passeggeri per mancanza di posti ma, allo stesso tempo, utilizzare aerei con un elevato numero di posti aumenta i costi per la compagnia. Per avere la certezza di non dover mai rifiutare passeggeri, l'unica possibilità è prevedere 1000 posti a sedere. Se sono disposte ad accettare una certa probabilità di rifiutare clienti, di quanto può essere ridotto il numero di posti?

Per ogni compagnia il numero di richieste di biglietti per un singolo volo è una variabile aleatoria X la cui distribuzione di probabilità è quella che abbiamo già analizzato in precedenza e che abbiamo immaginato come somma di 1000 variabili X_k, ognuna con valori 0 e 1, con media $p = 1/2$ e varianza $p(1-p) = 1/4$. Ne segue che X ha media 500 e varianza $1000/4$, cioè deviazione standard $15,81$ circa. Per il teorema centrale, la variabile aleatoria $(X - 500)/\sqrt{250}$ ha una distribuzione di probabilità molto simile alla distribuzione normale. Se siamo interessati all'intervallo di confidenza del 95% (cioè siamo disposti a rifiutare passeggeri con la probabilità del 5%) allora dobbiamo cercare nella tabella della $N(0,1)$ quale è il più piccolo valore di z che fornisce un numero maggiore di $0,45$ (al valore della tabella dobbiamo aggiungere l'integrale tra $-\infty$ e 0, che è 1/2); troviamo $1,65$ e quindi possiamo affermare che

$$P\left(\frac{X - 500}{\sqrt{250}} \leq 1,65\right) \geq \frac{95}{100}$$

e quindi

$$P(X \leq 526,08) \geq \frac{95}{100}\,.$$

Dunque è sufficiente che ogni compagnia predisponga aerei con soli 527 posti per non dover rifiutare passeggeri più del 5% delle volte. Se esigiamo una confidenza maggiore, ad esempio il 99%, allora cerchiamo il più piccolo numero che supera $1,49$ nella tabella (trovando $2,33$) e possiamo concludere che sono sufficienti $500 + 2,33\sqrt{250}$ posti a sedere sull'aereo, cioè soli 537 posti.

In questo esempio l'intervallo di confidenza è $[0, 537]$.

La confidenza è la probabilità di non sbagliare una previsione, dove la previsione è espressa dall'affermazione che la variabile X cade in un certo intervallo, detto intervallo di confidenza. Gli intervalli di confidenza possono essere *unilateri* (come nel nostro esempio) o *bilateri* (come avviene nelle proiezioni elettorali).

7. Statistica

Il termine **statistica** fu introdotto da Gerolamo Ghislini nel 1647 per indicare la scienza descrittiva delle qualità ed elementi caratterizzanti degli Stati. Da allora molto è cambiato e non è affatto semplice caratterizzare oggi la statistica.

Una definizione ricorrente afferma che *la statistica si occupa dell'analisi quantitativa dei fenomeni collettivi*, fenomeni cioè composti da un grande numero di unità elementari.

Per cercare di essere più espliciti, potremmo dire che la statistica è la scienza che appronta metodi, fondati sul calcolo delle probabilità, per la raccolta, la sintesi, l'analisi, l'elaborazione e l'interpretazione di dati numerici.

Con lo sviluppo dell'informatica verso la metà del secolo scorso è stato possibile gestire quantità di dati sempre più rilevanti diminuendo altresì il tempo necessario per l'elaborazione. Ciò ha reso accessibili a tutti gli strumenti fondamentali della statistica ed onnipresenti i risultati di elaborazioni più o meno sofisticate di dati.

Questa è la principale ragione della necessità di conoscere i metodi e strumenti basilari della statistica per ogni individuo che voglia consapevolmente partecipare alla realtà attuale.

Il linguaggio della statistica è rappresentativo della sua storia: affonda le sue radici nell'analisi di fabbisogni e caratteristiche di popolazioni, si sovrappone sovente con quello della probabilità ed infine si intreccia con quello delle scienze sperimantali.

Il punto di partenza di un'indagine è una **popolazione** (o *collettivo statistico*) composta di singoli **individui** o **unità statistiche** (ma anche *elemento* o *caso*) ed alcune **caratteristiche** o **variabili** (o *caratteri*) associate agli elementi della popolazione in esame.

Una variabile si realizza in corrispondenza di ogni unità statistica in una **modalità**. Le modalità di un carattere devono essere

- *esaustive*, cioè devono rappresentare tutte le possibilità
- *incompatibili*, quando ad ogni unità è associabile una sola modalità

(come le alternative in probabilità).

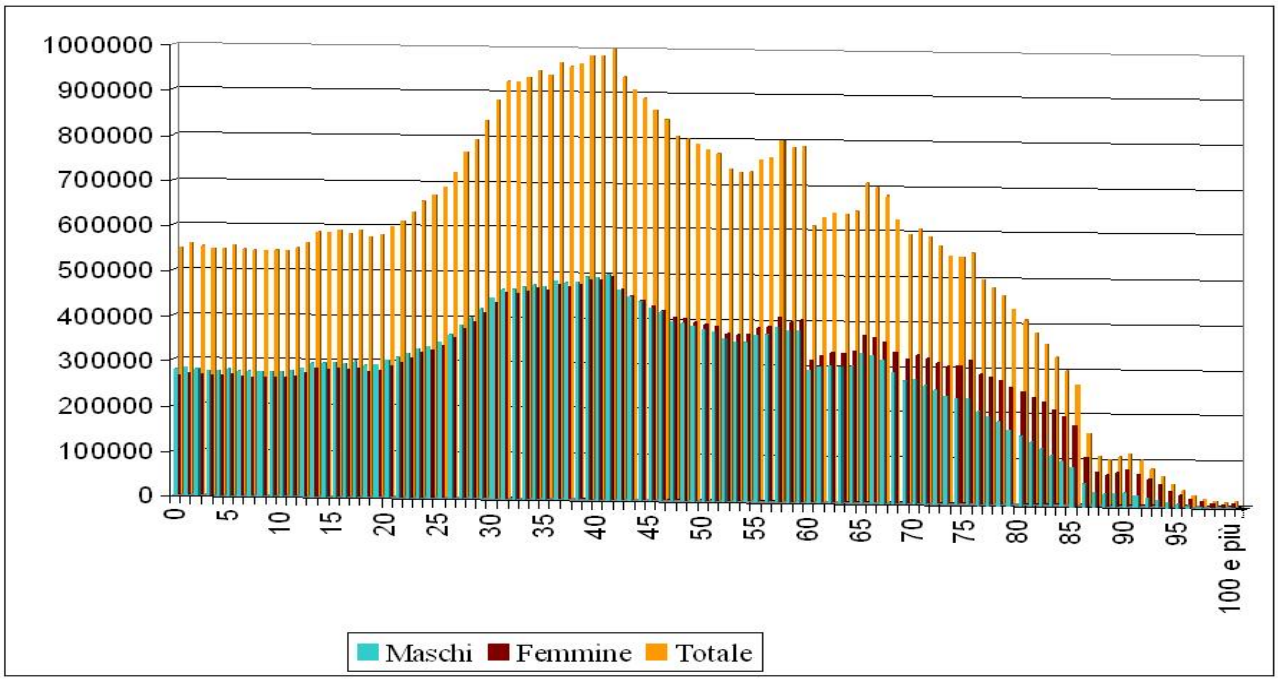
Le variabili si dividono in **qualitative**, espresse da aggettivi o attributi, e **quantitative** se espresse da numeri. Le variabili qualitative si dividono in *ordinali*, quando esiste un ordine naturale e preciso tra le modalità, e *nominali* in caso contrario. Tra le variabili quantitative si distinguono quelle *discrete* e quelle *continue*.

Una volta raccolti i dati relativi ad alcune caratteristiche di una popolazione numerosa, ci sono vari modi di presentarli. Se una variabile qualitativa o quantitativa discreta X assume le modalità $x_1, x_2, \ldots, x_k$, allora indichiamo con $n_1, n_2, \ldots, n_k$ le rispettive **frequenze assolute**, cioè il numero di volte con cui la singola modalità viene osservata nella popolazione.

Ad esempio, i dati forniti dall'ISTAT sull'età della popolazione italiana nel 2006 vengono divulgati attraverso una tabella che riporta le frequenze assolute delle singole età, simile a quella abbozzata a fianco. Il numero 550865 nella stessa riga del numero 2 indica che sono stati rilevati più di mezzo milione di abitanti con età di 2 anni. Osserviamo che l'ISTAT nella colonna delle età dopo il numero 99 pone una sola casella con l'indicazione 100 e più. Probabilmente dovrà presto rivedere questa convenzione, dato che già nel 2006 gli ultracentenari erano 10154.

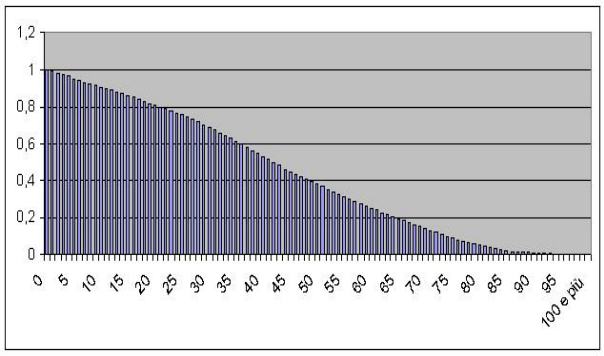
Età	**Maschi**	**Femmine**	**Totale**
0	281997	265162	547159
1	285961	271258	557219
2	282708	268157	550865
3	279183	265925	545108
4	280041	265457	545498
5	284193	268410	552603
⋮	⋮	⋮	⋮
99	1250	5400	6650
100 e più	1729	8425	10154
Totale	28526888	30224823	58751711

Come abbiamo visto per la probabilità degli eventi, un modo efficace per rappresentare i dati è quello di un grafico ad istogrammi. Riportiamo di seguito il grafico relativo all'esempio.



Se avessimo un analogo grafico per la popolazione di un'altra nazione potremmo voler confrontare le due distribuzioni di frequenze, per affermare ad esempio che una è più anziana dell'altra. Chiaramente i singoli valori non forniscono nessuna informazione in proposito: si devono almeno normalizzare. Si considerano quindi le **frequenze relative**, ottenute dividendo la frequenza assoluta per il numero totale di individui della popolazione. Spesso si preferiscono le **frequenze percentuali**, che sono le precedenti moltiplicate per 100. Le frequenze relative sono ovviamente legate al concetto di probabilità. Infatti esprimono la probabilità che un individuo scelto a caso dalla popolazione realizzi quella particolare modalità (in ipotesi di equiprobabilità).

Ancora più significative per un confronto sono le cosiddette **frequenze cumulative**. Queste possono essere definite per caratteristiche qualitative ordinali o quantitative come le frequenze di eventi che non prescrivono la modalità ma ne danno un limite (superiore o inferiore). Nel nostro esempio possiamo considerare il numero di abitanti con età maggiore o uguale a k: la distribuzione delle frequenze cumulative corrispondenti è schematizzata qui a fianco.



Torniamo ad analizzare la tabella della pagina precedente. Possiamo osservare che età e sesso del singolo individuo sono variabili diverse e che quindi quella tabella non riporta soltanto le frequenze delle due variabili (sarebbero state sufficienti l'ultima colonna e l'ultima riga), ma qualcosa di più. Vengono mantenute distinte non solo le singole modalità delle due variabili ma tutte le possibili coppie di modalità. La variabile età ha 101 modalità diverse, cioè 101 possibili valori; la variabile sesso ha solo due modalità. La tabella in questione fornisce le frequenze di tutte le possibili accoppiate (e, s) e quindi 101×2 frequenze. Se avessimo analizzato la variabile età e la variabile regione di residenza, avremmo dovuto riportare 101×20 frequenze. In casi analoghi si parla di **distribuzione doppia di frequenze** o di **distribuzione congiunta** di due variabili statistiche. La rappresentazione completa può essere fornita da una opportuna matrice o da un grafico a istogrammi con più serie come quello della pagina precedente.

Un problema consueto in statistica è quello di cercare di descrivere la distribuzione di frequenze di una variabile su una popolazione ampia partendo dai dati ottenuti su un campione, cioè su un sottoinsieme relativamente piccolo della popolazione totale. Il campionamento, per certi versi simile a quello incontrato in probabilità, può essere effettuato in vari modi (casuale, stratificato, per quote, ecc.). Non entreremo nel merito e supporremo di avere già fissato il campione.

Il problema adesso può essere diviso in due parti. Prima cercare di sintetizzare i dati raccolti in modo da evidenziare quelli più interessanti (caratteristica puramente soggettiva) e poi utilizzare i dati (o i soli indicatori di sintesi scelti) per fare delle previsioni sull'intera popolazione.

Gli ***indicatori sintetici*** o caratteristici più utilizzati sono:

- **campo di variazione** o *range*: quando la variabile è ordinale o quantitativa, cioè quando è possibile ordinare le possibili modalità e quindi parlare di maggiore e minore, il campo di variazione è dato dall'intervallo determinato dalla modalità minima e massima osservata sul campione;
- **moda** o *valore tipico*: è definita come la modalità osservata nel maggior numero dei casi e quindi non dipende dalle altre modalità;
- **media**: anche se solo per variabili quantitative, esistono diverse definizioni di media (aritmetica, geometrica, armonica, quadrata, ecc.); la più utilizzata in statistica è la media aritmetica;
- **mediana**: solo per variabili ordinali o quantitative è definita come la modalità che bipartisce la distribuzione, cioè tale che una metà dei dati osservati sono maggiori o uguali e l'altra metà sono minori o uguali della mediana stessa;
- **percentili** o *quantili* o *centili*: possono essere visti come una generalizzazione della mediana; invece di dividere i dati osservati a metà (una volta ordinati) il k-esimo percentile li divide in un $k\%$ e $(100-k)\%$. Il 25° percentile è detto primo **quartile**, il 50° percentile è la mediana o secondo quartile, il 75° è detto terzo quartile;
- **differenza interquantilica**: definito per le variabili quantitative, è esattamente la differenza tra il terzo ed il primo quartile;
- **varianza**: è definita per caratteristiche quantitative con la stessa formula incontrata per le variabili aleatorie discrete e quindi misura la differenza dalla media;
- **scarto quadratico medio**: è la radice quadrata della varianza.

La relazione tra statistica e probabilità è così evidente che non serve giustificare l'interesse particolare che avremo nello stimare la media e la varianza delle variabili sulla popolazione. La media su un campione si chiama **media campionaria** e si calcola con la formula (6), dove le probabilità p_k vanno sostituite con le frequenze relative, cioè con le frequenze assolute n_k divise per la cardinalità del campione, oppure direttamente con la formula

$$\overline{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \,, \tag{20}$$

dove n indica il numero di elementi nel campione, x_k le singole modalità registrate.

Il teorema centrale garantisce che, all'aumentare della numerosità del campione, la media campionaria tende alla media della variabile in esame con probabilità 1.

La **varianza campionaria** invece si definisce tramite la formula

$$\overline{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{\mu})^2 \,, \tag{21}$$

dove n indica il numero di elementi nel campione, x_k le singole modalità registrate e $\overline{\mu}$ la media campionaria. Osserviamo che non si divide per n, come potremmo aspettarci, ma per $n-1$; chiaramente la differenza è trascurabile quando n è molto grande ma rilevante per piccoli valori di n. Ciò è dovuto al fatto che stiamo stimando contemporaneamente la media e la varianza.

Come conseguenza del teorema centrale abbiamo il seguente principio.

Principio fondamentale della statica. Fatti n rilevamenti indipendenti di una stessa quantità X, per n abbastanza grande ($n \geq 30$) la migliore previsione statistica di ogni ulteriore rilevamento si ottiene considerando X come una variabile aleatoria normale con media $\overline{\mu}$ e deviazione standard $\overline{\sigma}$.

8. Test di ipotesi

Abbiamo già avuto modo di analizzare un test di ipotesi quando abbiamo immaginato di verificare l'equità di una moneta provando a lanciarla un certo numero di volte. Cerchiamo adesso di inquadrare l'esempio in un contesto più generale.

I test riguardano la distribuzione di probabilità di una variabile aleatoria o semplicemente qualche parametro che dipende da questa (quali la media o la varianza).

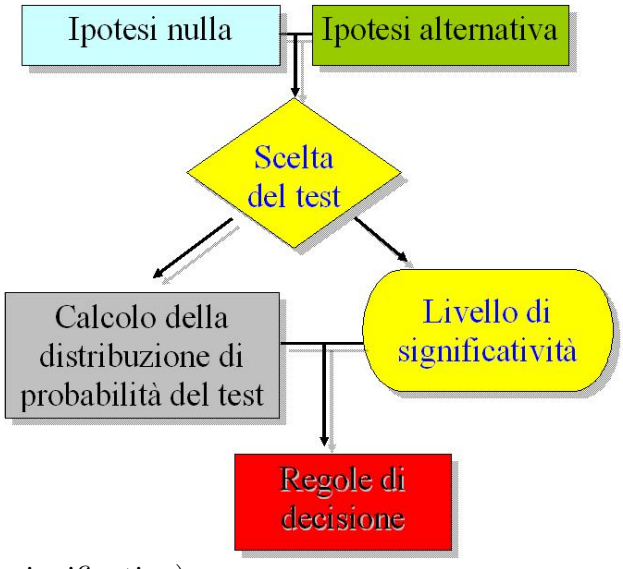
Inizialmente si assume che tale distribuzione (o parametro) ricalchi una previsione teorica o sperimentata precedentemente. Il test statistico mira ad accettare o rifiutare questa assunzione. Nel linguaggio statistico stiamo formulando la cosiddetta **ipotesi nulla**: i dati raccolti dall'indagine possono avere distribuzione (o parametro) diversa da quella assunta, ma le differenze sono imputabili alla casualità del risultato stesso, a fluttuazioni campionarie.

Rifiutare l'ipotesi nulla vuol dire accettare l'**ipotesi alternativa**: le differenze non sono imputabili al caso e quindi la distribuzione di probabilità non è quella assunta.

Nell'esempio della moneta, avevamo assunto come ipotesi nulla l'equità della moneta, da cui l'affermazione che la media dei successi su 20 lanci è 10. Abbiamo anche osservato che pretendere 10 successi su 20 lanci è molto esigente. Ci siamo quindi preoccupati di determinare se il numero di successi ottenuti era *significativamente* diverso da 10.

Per poter prendere una decisione sulla base di un test abbiamo sempre bisogno di fissare un livello di significatività soggettivamente opportuno. Come abbiamo già detto, tale livello esprime la probabilità di rifiutare l'ipotesi nulla nel caso che sia vera.

Nel caso della moneta abbiamo utilizzato la media campionaria (sul campione di 20 lanci) per stimare il valor medio della variabile numero di successi e quindi la probabilità p di successo in ogni singolo lancio. L'analisi teorica dell'esperimento ci ha portato a considerare la distribuzione binomiale con parametro $p = 1/2$ (che poi per semplicità abbiamo approssimato con la distribuzione normale) e conseguentemente ad individuare un intervallo ($[5, 15]$ nel nostro esempio) in modo che la decisione se accettare o rifiutare l'ipotesi nulla con il livello di significatività prescelto è presa a seconda che la media campionaria sia contenuta o no in tale intervallo.



In generale potremo utilizzare una funzione diversa dalla media campionaria che chiameremo **test** o **statistica**; l'aspetto fondamentale è che se ne conosca la distribuzione nel caso in cui l'ipotesi nulla sia vera. E proprio grazie a questa distribuzione dividiamo i valori possibili del test in due regioni: la **regione di rifiuto** e la **regione di accettazione**, con ovvio riferimento alla scelta finale. La regola di decisione del test dipende quindi dal livello di significatività e dalla distribuzione di probabilità della statistica utilizzata.

Nel linguaggio statistico, un esperimento o campione che ci permette di rifiutare l'ipotesi nulla con livello di significatività del 5% (o 1%) è detto **statisticamente significativo** (o molto significativo).

Uno dei test più utilizzati è proprio quello che abbiamo già discusso: l'ipotesi nulla assegna un determinato valore alla media di una variabile aleatoria X ed il test prescelto è la media campionaria $\overline{\mu}$ su successive realizzazioni di X.

Nel caso in cui la distribuzione di probabilità di $\overline{\mu}$ sia teoricamente nota (nell'esempio della moneta quella binomiale) le regioni di rifiuto e accettazione sono più o meno facilmente ottenibili una volta fissato il livello di significatività e l'ipotesi alternativa, che potrebbe essere unidirezionale o bidirezionale. In particolare la regione di accettazione è definita come un intervallo di confidenza (unilatero o bilatero) con livello di confidenza pari a 1 meno il livello di significatività.

Basandoci sul teorema centrale, anche se la distribuzione di probabilità di $\overline{\mu}$ è ignota, quando il campione in esame è sufficientemente ampio (> 100) possiamo approssimarla con una distribuzione normale con media determinata dall'ipotesi nulla (i valori medi di X e di $\overline{\mu}$ sono uguali) e varianza pari alla varianza campionaria.

Se invece il campione è limitato, allora si utilizza come riferimento un'altra distribuzione di probabilità: la **distribuzione di Student di ordine** g. La densità di probabilità di questa distribuzione è la seguente:

$$f(x) = \frac{C(g)}{(1 + \frac{x^2}{g})^{\frac{g+1}{2}}}\,, \tag{22}$$

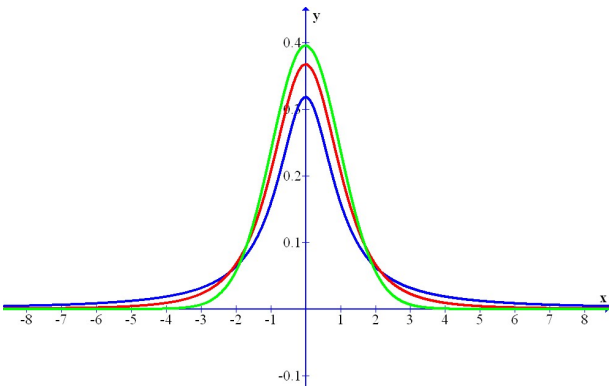
dove g è un parametro detto *ordine* o numero dei **gradi di libertà** della distribuzione e $C(g)$ è la costante giusta affinché l'integrale della f esteso a tutta la retta reale sia 1.

Questa distribuzione, spesso chiamata **t di Student** o semplicemente **t**, prende il nome dallo pseudonimo usato da William Sealy Gosset nell'articolo del 1908 in cui fu introdotta. Gosset era un chimico impiegato nella famosa ditta Guinness di Dublino ed era costretto a pubblicare sotto pseudonimo a causa dell'esclusiva nel contratto con la birreria. Fu il primo a notare che, date n variabili aleatorie con la stessa distribuzione di probabilità, la variabile aleatoria normalizzata

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\overline{\sigma}\sqrt{n}}\,, \tag{23}$$

dove μ è la media e $\overline{\sigma}$ è la deviazione standard campionaria, può avere una distribuzione di probabilità molto diversa dalla normale standard quando n è piccolo. Il teorema centrale può essere utilizzato per dimostrare che al tendere di n all'infinito la normalizzata tende ad uniformarsi alla distribuzione normale standard ma, come abbiamo già notato, non quantifica la vicinanza per specifici valori di n.

Supponendo che le variabili X_k siano tutte normali standard è possibile dimostrare che la variabile normalizzata (23) ha densità di probabilità data dalla (22) con $n - 1$ gradi di libertà, cioè con $g = n - 1$. Una variabile con distribuzione di probabilità t di Student di ordine g ha media nulla e varianza pari a $\frac{g}{g-2}$.



A fianco sono riportati i grafici della densità (22) per g pari a 1, 3 e 30. La forma del grafico è molto simile ad una Gaussiana ma la funzione tende a zero molto più lentamente per x che tende all'infinito. Il valore della funzione in 0 aumenta al variare di g. Per $g = 30$ il grafico è indistinguibile da quello della Gaussiana standard, cioè con varianza 1.

In fondo a questi appunti trovate varie tabelle, tra cui quella con i principali valori della t di Student per diversi gradi di libertà.

Un altro tipo di test abbastanza frequente è quello che riguarda la varianza anziché la media. Ad esempio, supponiamo di avere un certo strumento per effettuare una misurazione. La ditta fornitrice garantisce una specifica precisione. Una misura è inevitabilmente accompagnata da incertezza e quindi dovrebbe essere sempre corredata dall'indicazione dell'intervallo di indeterminazione. Pertanto la misura fornita da ogni strumento è una variabile aleatoria. Indichiamo con X la misura dello strumento in esame. L'*accuratezza* di uno strumento quantifica la differenza tra la media di X e la misura reale, mentre la *precisione* riguarda lo scostamento dalla media (da non confondere con la *sensibilità* e la *portata* di uno strumento che rappresentano il minimo e il massimo valore misurabile). Se lo strumento è accurato, ci aspettiamo che il valore medio di un discreto numero di misure ottenute sia sostanzialmente esatto. Come misura della precisione possiamo prendere lo scarto quadratico medio di X.

Supponiamo adesso di avere una serie di 10 misurazioni

$$X: \quad 216{,}82 \quad 216{,}93 \quad 216{,}95 \quad 216{,}99 \quad 217{,}00 \quad 217{,}02 \quad 217{,}04 \quad 217{,}05 \quad 217{,}08 \quad 217{,}13$$

e di voler sottoporre a verifica l'ipotesi nulla: $Var(X) = 0{,}005$. La media campionaria della serie è 217,001, mentre la varianza campionaria è 0,00752. Quindi la deviazione standard nel campione è 0,08671 e quella ipotizzata è 0,0707.

Dobbiamo decidere se la differenza riscontrata ($0{,}00752 - 0{,}005 = 0{,}00252$) è imputabile alle cosiddette fluttuazioni campionarie, oppure se è rivelatrice di una precisione dichiarata maggiore di quella reale. La decisione, come in ogni test statistico, dipenderà dal livello di significatività prescelto (e questa è la parte facile) e dalla distribuzione di probabilità del test che, nel caso specifico, è la varianza campionaria.

Questo è un caso in cui possiamo dare anche una dimostrazione della scelta della distribuzione. Se l'ipotesi nulla è verificata, allora la variabile $X - \mu$, dove μ rappresenta la media di X, può essere vista come una variabile aleatoria normale con media 217,001 e varianza 0,005. Ciò che dobbiamo determinare è la distribuzione della variabile $(X - \mu)^2$.

Indichiamo con $N(x)$ e $f(x)$ rispettivamente la densità di probabilità di $X - \mu$ e di $(X - \mu)^2$. Per definizione di densità, sfruttando la regolarità di $N(x)$ (uniforme continuità), abbiamo

$$N(x) = \lim_{h \to 0} \frac{P\left(x - h \leq X - \mu \leq x + h\right)}{2h}\,.$$

Analogamente, ipotizzando che anche $f(x)$ sia una funzione regolare, possiamo scrivere

$$f(x) = \lim_{h \to 0} \frac{P\left(x - h \leq (X - \mu)^2 \leq x + h\right)}{2h} = \lim_{h \to 0} \frac{2P\left(\sqrt{x - h} \leq X - \mu \leq \sqrt{x + h}\right)}{2h}\,,$$

dove abbiamo tenuto conto della simmetria di $X - \mu$ rispetto allo 0 (e che $\{a \leq t^2 \leq b\}$ equivale a $\{-\sqrt{b} \leq t \leq -\sqrt{a}\} \cup \{\sqrt{a} \leq t \leq \sqrt{b}\}$). Dunque

$$f(x) = \lim_{h \to 0} \frac{P\left(\sqrt{x - h} \leq X - \mu \leq \sqrt{x + h}\right)}{\sqrt{x + h} - \sqrt{x - h}} \, \frac{\sqrt{x + h} - \sqrt{x - h}}{h} = \frac{N(\sqrt{x})}{\sqrt{x}}\,.$$

Pertanto, dalla formula $N(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}$ ricaviamo

$$f(x) = \frac{e^{-\frac{x}{2\sigma^2}}}{\sqrt{2\pi\sigma^2 x}}\,.$$

Per $\sigma = 1$, la densità di probabilità appena introdotta (e la corrispondente distribuzione) è molto utilizzata nei test statistici ed è comunemente chiamata distribuzione χ^2 di ordine 1 o con 1 grado di libertà.

In generale si definisce la **distribuzione χ^2 di ordine g** o con g **gradi di libertà** tramite la sua densità di probabilità:

$$\chi^2(x) = \frac{x^{\frac{g-2}{2}} e^{-x/2}}{2^{g/2}\Gamma(g/2)}\,. \tag{24}$$

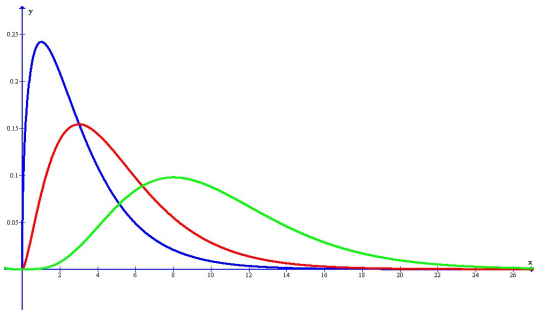
Una variabile aleatoria con distribuzione χ^2 si dice semplicemente una variabile χ^2. La frequente apparizione di variabili χ^2 è in parte giustificata dai seguenti risultati:

(R1) Se X_1, X_2, …, X_n sono n variabili χ^2 indipendenti con g_1, g_2, …, g_n gradi di libertà rispettivamente, allora la loro somma è una variabile χ^2 con $g = g_1 + g_2 + \cdots + g_n$ gradi di libertà;

(R2) Se X_1, X_2, …, X_n sono n normali standard indipendenti, allora la somma dei loro quadrati è una variabile χ^2 con n gradi di libertà;

(R3) Fatti n rilevamenti indipendenti di una stessa variabile normale con media μ e varianza σ^2, la variabile $(n-1)\overline{\sigma}^2/\sigma^2$, dove $\overline{\sigma}$ è la varianza campionaria, è χ^2 con $(n-1)$ gradi di libertà.

A fianco sono rappresentate le densità (24) per g pari a 3, 5 e 10. Il valore massimo della funzione è assunto in $g-2$. La media è g e la varianza è $2g$. Al tendere di g all'infinito la funzione diventa Gaussiana. I percentili di queste distribuzioni si trovano tra le tabelle in fondo a queste pagine.



Torniamo adesso al nostro test. Per il risultato (R3) citato sopra, la statistica più conveniente è il rapporto tra varianza campionaria e varianza ipotizzata moltiplicato per l'ampiezza del campione diminuita di 1, che nel nostro caso specifico diventa

$$9\,\frac{0{,}00752}{0{,}005} = 13{,}536\,.$$

Questo valore va relazionato alla distribuzione χ^2 con 9 gradi di libertà. Nella tabella del χ^2, alla riga corrispondente ai 9 gradi di libertà, il primo numero inferiore al valore ottenuto si trova nella colonna della significatività del 14%. Ciò indica che se il livello di significatività prescelto è inferiore al 14%, allora non dobbiamo rifiutare l'ipotesi che lo strumento abbia una precisione pari a quella dichiarata dalla ditta fornitrice.

In letteratura si trovano molti altri test di ipotesi, basati su altrettante distribuzioni teoriche, che qui non tratteremo. Nel prossimo capitolo incontreremo test statistici di diverso tipo.

9. Indipendenza, correlazione e regressione

Occupiamoci ora di un altro aspetto fondamentale della statistica: l'*analisi dell'associazione* tra due caratteristiche di una popolazione.

Uno dei concetti principali di questa analisi lo abbiamo già incontrato ed è l'indipendenza tra variabili. In statistica si trovano altri concetti di indipendenza e quello che qui analizzeremo va sotto il nome di **indipendenza assoluta**. Due variabili o caratteri di una popolazione sono assolutamente indipendenti quando la conoscenza della modalità con cui si manifesta una delle due variabili non fornisce alcuna informazione sulle possibili modalità della seconda.

Un metodo statistico per verificare l'indipendenza assoluta di due caratteri si basa sulla distribuzione congiunta o distribuzione doppia di frequenze. Come abbiamo già visto, la distribuzione doppia di due variabili può essere rappresentata graficamente oppure con una tabella, detta **tabella a doppia entrata** o **tabella di correlazione**.

Facciamo un esempio. Consideriamo come popolazione gli studenti delle scuole elementari italiane che hanno partecipato al progetto *Censimento a scuola* promosso dall'ISTAT nel 2001. Come caratteri scegliamo il sesso X (due possibili modalità: maschio e femmina) e la zona geografica di appartenenza Y (cinque possibili modalità: nord ovest, nord est, centro, sud e isole). La corrispondente tabella a doppia entrata ha 6 righe e 3 colonne, dato che abbiamo aggiunto alle modalità previste anche una riga ed una colonna finale con i totali. L'ultima

colonna ci indica il numero di individui della popolazione divisi per aree geografiche e quindi rappresenta la distribuzione delle frequenze assolute della variabile Y. Questa, come parte di una distribuzione congiunta, prende il nome di **distribuzione marginale** del carattere Y. Analogamente l'ultima riga rappresenta la distribuzione marginale di X.

	Maschio	Femmina	Totale
Nord Ovest	1355	1350	2705
Nord Est	192	188	380
Centro	484	462	946
Sud	876	814	1690
Isole	743	741	1484
Totale	3650	3555	7205

Tabella a doppia entrata delle frequenze assolute.

Se invece analizziamo una colonna o riga diversa, possiamo ancora vederla come una distribuzione. Ad esempio, la terza riga fornisce le frequenze assolute del carattere X relativamente agli studenti del centro Italia. Questa distribuzione si chiama **distribuzione condizionata** di X alla modalità centro della variabile Y.

Se le variabili X e Y fossero assolutamente indipendenti, allora le distribuzioni condizionate di X alle singole modalità di Y sarebbero tutte uguali e quindi tutte uguali alla distribuzione marginale di X.

Il modo più semplice per verificare l'indipendenza assoluta di due variabili è dunque quello di osservare una tabella a doppia entrata con le frequenze relative (o percentuali) anziché le frequenze assolute. Nel caso della rilevazione precedente otterremmo distribuzioni condizionate che si avvicinano abbastanza alla distribuzione marginale di X, ma non sono esattamente uguali.

	Maschio	Femmina	Totale
Nord Ovest	50,09%	49,91%	100%
Nord Est	50,53%	49,47%	100%
Centro	51,16%	48,84%	100%
Sud	51,83%	48,17%	100%
Isole	50,07%	49,93%	100%
Totale	50,66%	49,34%	100%

Tabella a doppia entrata delle frequenze percentuali.

Quindi le due variabili dell'esempio non sono assolutamente indipendenti.

Siamo nuovamente di fronte ad un risultato contrastante la nostra aspettativa e possiamo pertanto chiederci se la variazione osservata sia imputabile alle ormai note fluttuazioni statistiche oppure riveli una distribuzione del carattere sesso tra gli studenti del campione realmente dipendente dall'area geografica. In termini statistici abbiamo formulato l'ipotesi nulla le due variabili sono indipendenti e l'ipotesi alternativa esiste un legame tra le due variabili e vorremmo programmare un test di ipotesi.

In questo caso il test più utilizzato è il cosidetto **test del Chi-quadrato** introdotto nel 1900 da Karl Pearson (1857-1936). Vediamo come è definito.

Indichiamo con n_{ij} la frequenza assoluta rilevata congiuntamente per la i-esima modalità di X e per la j-esima modalità di Y, in altri termini il numero della tabella scritto nella riga e colonna corrispondenti alle modalità considerate. Indichiamo con n_{i*} la frequenza assoluta della i-esima modalità di X che si trova quindi nella riga del totale. Analogamente con n_{*j} indicheremo le frequenze della distribuzione marginale di Y. Sia infine n la cardinalità del campione (7205 nel nostro caso). Dividiamo il compito in passi successivi.

1°) Si costruisce la **tabella delle frequenze assolute teoriche d'indipendenza** a partire dalle distribuzioni marginali. Indicate con n'_{ij} tali frequenze si osserva che verificano la proporzione $n'_{ij} : n_{*j} = n_{i*} : n$ e risultano quindi definite dalla formula

$$n'_{ij} = \frac{n_{i*}n_{*j}}{n}\,.$$

2°) Si calcolano le **contingenze**, cioè le differenze $(n_{ij} - n'_{ij})$ tra le frequenze osservate e quelle teoriche d'indipendenza per ogni cella della tabella.

3°) Si calcola per ogni cella il quadrato della contingenza diviso per la frequenza teorica d'indipendenza.

4°) Sommando i valori ottenuti per tutte le celle si ottiene il test χ^2 della distribuzione congiunta in esame. Esplicitamente si ha

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}\,.$$

Come distribuzione teorica del test χ^2 possiamo assumere la distribuzione χ^2 di ordine 4; il numero di gradi di libertà da considerare è dato dalla formula

$$(\text{numero di righe} - 1) \times (\text{numero di colonne} - 1)\,.$$

Scelto quindi il livello di significatività desiderato, osservando l'opportuna tabella in fondo a queste pagine, possiamo determinare le regole di decisione. In questo caso è naturale prendere in considerazione intervalli unilateri e pertanto la decisione sarà presa confrontando il χ^2 della distribuzione congiunta in esame con il percentile (determinato dalla significatività richiesta) della distribuzione χ^2 di ordine 4.

Svolgiamo tutti i calcoli nel caso dell'esempio proposto. Scelto il livello di significatività del 5%, la tabella della distribuzione χ^2 di ordine 4 indica come valore limite $9,48773$. Quindi decideremo di imputare al caso le discrepanze tra frequenze osservate e frequenze teoriche se l'indice χ^2 della nostra tabella è inferiore a $9,48773$. Il calcolo esplicito del nostro test prevede i seguenti passaggi:

	Maschio	Femmina	Totale
Nord Ovest	1370,33	1334,67	2705
Nord Est	192,51	187,49	380
Centro	479,24	466,76	946
Sud	856,14	833,86	1690
Isole	751,78	732,22	1484
Totale	3650	3555	7205

Tabella delle frequenze teoriche.

	Maschio	Femmina	Totale
Nord Ovest	-15,33	15,33	0
Nord Est	-0,51	0,51	0
Centro	4,76	-4,76	0
Sud	19,86	-19,86	0
Isole	-8,78	8,78	0
Totale	0	0	0

Tabella delle contingenze.

$$\chi^2 = \frac{(-15,33)^2}{1370,33} + \frac{(-0,51)^2}{192,51} + \frac{(4,76)^2}{479,24} + \cdots + \frac{(-19,86)^2}{833,86} + \frac{(8,78)^2}{732,22} = 1,5879\,.$$

Pertanto l'ipotesi nulla è accettata al livello di significatività del 5%.

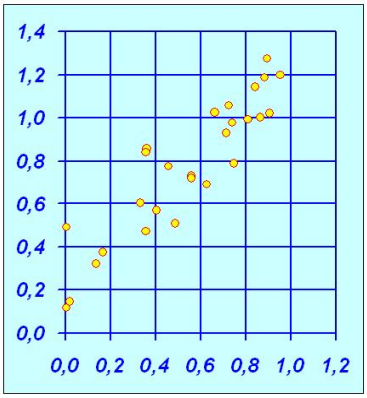
La tabella riportata in queste pagine non è completa e termina al livello di significatività del 15%. Anche per tale livello l'ipotesi sarebbe stata accettata. Il primo percentile al di sotto del valore di χ^2 trovato è l'ottanduesimo.

Un metodo più sbrigativo, anche se meno accurato, per verificare il grado di indipendenza assoluta tra due variabili si basa sull'**indice V di Cramer**. Indicati con r e c rispettivamente il numero di righe e di colonne nella tabella (totali esclusi e quindi il numero di modalità possibili delle due variabili), si definisce

$$\mathbf{V} = \sqrt{\frac{\chi^2}{n \min\{r-1, c-1\}}}\,. \tag{25}$$

L'indice **V** di Cramer è sempre compreso tra 0 e 1; vale 0 quando si ha indipendenza assoluta e vale 1 quando una delle due variabili determina univocamente la modalità dell'altra. Cioè i valori estremi sono assunti per comportamenti opposti della relazione tra due variabili. Se il valore è intermedio allora possiamo ricavarne solo un'indicazione sul comportamento più appropriato e non una presunta probabilità.

Quando l'indice V di Cramer non è nullo o addirittura è vicino a 1, si parla di una più o meno evidente **correlazione** tra le due variabili in oggetto. Altri termini utilizzati sono quelli di **covariazione** e **interrelazione**. Gli esempi a tutti noti sono quelli in cui interviene un rapporto di causa-effetto, ma è bene osservare che l'eventuale interpretazione in senso causale di una correlazione prescinde comunque dai compiti della statistica.



Se le due variabili considerate sono quantitative allora entrano in campo anche altri strumenti della statistica usati frequentemente nelle scienze sperimentali. Il primo per semplicità di realizzazione ed interpretazione è rappresentato dai **grafici di dispersione**, di cui la figura qui a lato è un esempio.

X	Y
0,36	0,86
0,14	0,32
0,73	1,06
0,49	0,51
0,84	1,14
0,91	1,02
0,56	0,73
0,17	0,37
0,66	1,03
0,33	0,61
0,40	0,57
0,95	1,20
0,63	0,69
0,01	0,49
0,46	0,78
0,87	1,00
0,88	1,19
0,89	1,28
0,36	0,47
0,02	0,15
0,36	0,84
0,81	0,99
0,01	0,12
0,74	0,98
0,71	0,93
0,56	0,72
0,75	0,79

I valori assunti dalle variabili X e Y sul singolo individuo o caso del campione in esame vengono interpretati come coordinate in un sistema di riferimento ortogonale del piano. Il grafico di dispersione riporta i punti aventi queste coordinate. L'aspetto negativo di questo tipo di rappresentazione è la perdita dell'informazione sulle frequenze con cui i singoli valori sono rilevati. Tale difetto è statisticamente inesistente se le variabili sono continue. L'aspetto positivo è invece la propensione ad evindenziare leggi matematiche che collegano le due variabili.

Immaginiamo ad esempio di disporre di 27 campioni di un certo materiale e di misurare su ognuno di essi due caratteristiche fisiche che chiameremo X ed Y per semplicità. La tabella a lato riporta le misurazioni ottenute, dove ogni riga contiene le informazioni su un singolo campione.

Il grafico di dispersione corrispondente è quello sopra riportato. È del tutto intuitivo immaginare una relazione forte tra queste due variabili, cioè una legge matematica che per ogni valore osservato di X fornisce un ipotetico valore di Y sufficientemente vicino ai valori eventualmente osservati.

A seconda della funzione matematica chiamata in causa si potranno effettuare calcoli diversi. Per ora limitiamoci al caso in cui la legge desiderata sia lineare. In altri termini immaginiamo di vedere i punti del grafico di dispersione addensati intorno ad una retta particolare. Tra tutte le rette del piano, cerchiamo quella che *meglio* si dispone nella nuvola di punti. Cerchiamo di essere più precisi.

Indichiamo con x_i e y_i i dati ottenuti sull'i-esimo campione e con $y = mx + q$ la generica retta del piano (ovviamente stiamo assumendo che tale retta non sia verticale). Se valesse $y_i = mx_i + q$ per ogni i, allora tutti i punti del grafico di dispersione sarebbero proprio sulla retta. In generale tali equazioni non saranno verificate esattamente, ma solo con un certo scarto o errore.

La retta che meglio approssima i dati raccolti è quella determinata dai valori m e q che minimizzano la funzione

$$S(m,q) = \frac{1}{27}\sum_{i=1}^{27}(y_i - mx_i - q)^2\,,$$

cioè lo scarto quadratico medio. Questo è noto come il **metodo dei minimi quadrati**.

Per trovare gli eventuali punti di minimo della funzione di due variabili S, cerchiamo i punti critici, cioè quei valori di m e q che verificano il sistema

$$\begin{cases} \frac{\partial}{\partial m} S(m,q) = 0 \\ \frac{\partial}{\partial q} S(m,q) = 0 \end{cases} .$$

Otteniamo le equazioni

$$\begin{cases} \sum_{i=1}^{27} x_i(y_i - mx_i - q) = 0 \\ \sum_{i=1}^{27} (y_i - mx_i - q) = 0 \end{cases}$$

da cui ricaviamo facilmente

$$q = \sum_{i=1}^{27} \frac{y_i - mx_i}{27} = \overline{y} - m\overline{x}\,, \tag{26}$$

dove $\overline{y}$ e $\overline{x}$ indicano le medie campionarie di Y e X rispettivamente e da questa, sostituita nella prima equazione del sistema,

$$m = \frac{\sum_{i=1}^{27} x_i y_i \; - \; 27\overline{x}\,\overline{y}}{\sum_{i=1}^{27} x_i^2 \; - \; 27\overline{x}^2} = \frac{Cov(X,Y)}{Var(X)}\,, \tag{27}$$

dove, ricordando le formule (13) e (14), abbiamo indicato con $Cov(X,Y)$ e $Var(X)$ la covarianza e la varianza, calcolate non per le variabili X e Y ma per i dati presi in esame. Queste vengono dette **covarianza empirica** e **varianza empirica**. In particolare, la varianza empirica differisce dalla varianza campionaria per il solo fatto che qui si divide per il numero di dati, mentre nella varianza campionaria si divide per quel numero diminuito di 1.

La retta $y = mx + q$, con m e q che verificano le (26) e (27), è detta **retta di regressione** di Y rispetto a X.

Consideriamo il punto di coordinate $(\overline{x}, \overline{y})$, cioè l'ipotetico baricentro di un sistema di pesi identici disposti sui punti del grafico di dispersione; per la (26) la retta di regressione passa per tale punto.

Scambiando il ruolo di X e Y, cioè prendendo in esame come scarti le differenze $x_i - (y_i - q)/m$ tra le ascisse osservate e quelle teoriche ad ordinata fissata, si ottiene la retta di regressione di X rispetto ad Y. Riscritta nella forma $x = m'y + q'$ ricaviamo formule analoghe alle precedenti per i due coefficienti:

$$q' = \overline{x} - m'\overline{y} \qquad \text{con} \qquad m' = \frac{Cov(X,Y)}{Var(Y)}\,.$$

Osserviamo che la retta di regressione di Y rispetto a X è generalmente diversa da quella di X rispetto a Y. Le due rette risultano coincidenti se e solo se il prodotto dei due coefficienti angolari è 1, cioè se vale ± 1 il numero

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \tag{28}$$

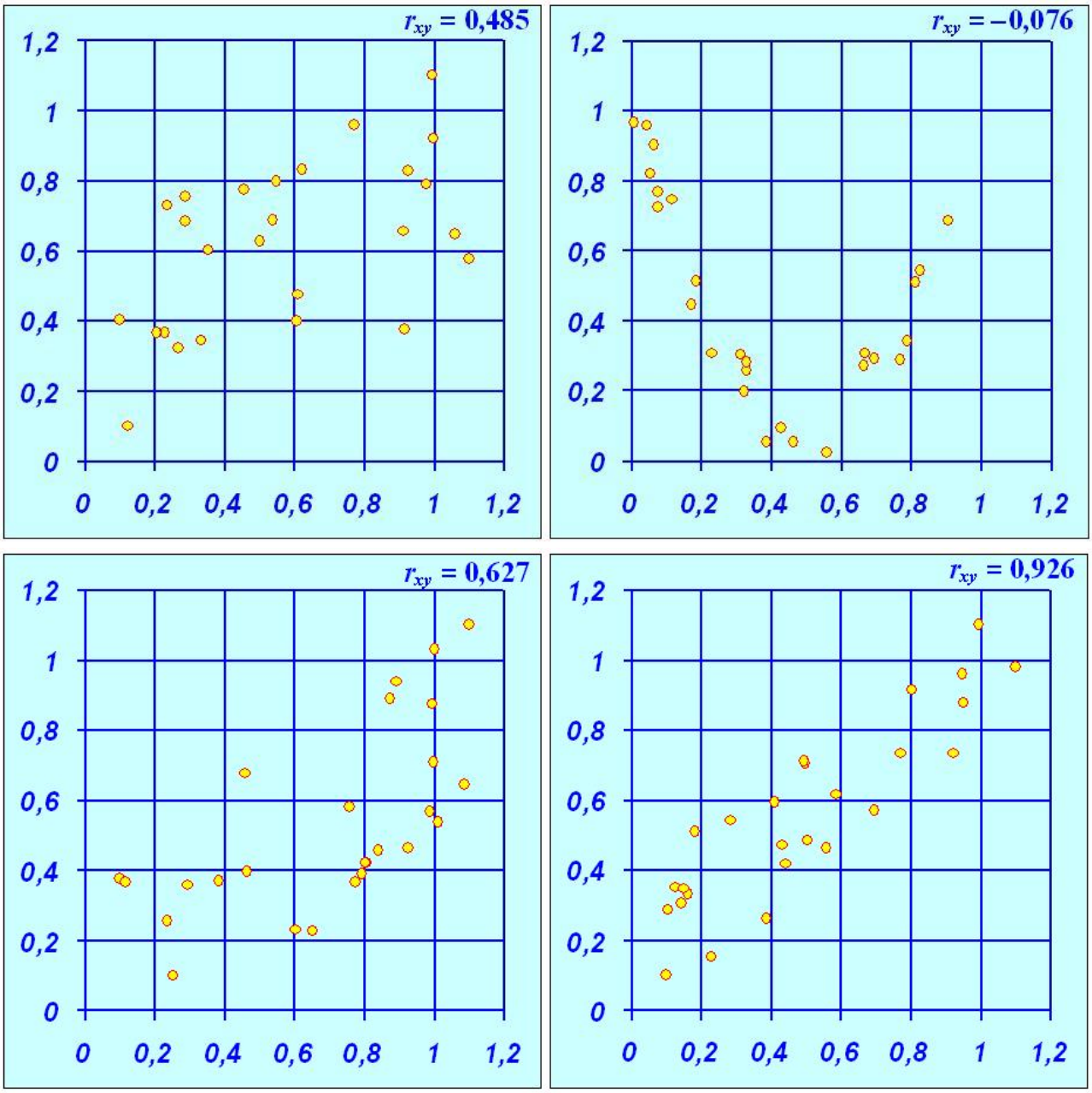
detto **coefficiente di correlazione lineare** o coefficiente di correlazione di Bravais-Pearson.

Per quantificare la bontà dell'approssimazione ottenuta con la retta di regressione potremmo calcolare il valore della funzione $S(m,q)$ nel punto di minimo. Il numero trovato dipende chiaramente dall'unità di misura utilizzata per Y; se per normalizzare dividiamo per $Var(Y)$, allora semplici calcoli conducono al valore $1 - r_{xy}^2$. Questo è un altro modo di introdurre il coefficiente di correlazione lineare ed inoltre chiarisce meglio il suo significato e la sua utilizzazione come misura di interrelazione (lineare) tra due variabili.

Si potrebbe dimostrare che il coefficiente di correlazione lineare è sempre compreso nell'intervallo $[-1, 1]$ ed inoltre

- se il suo valore è 0 allora non vi è certamente dipendenza lineare tra i due caratteri (ma non possiamo parlare di indipendenza assoluta!);
- se il suo valore è positivo e relativamente vicino a 1, allora si è in presenza di una più o meno evidente correlazione diretta;
- se il suo valore è negativo allora si parla di correlazione inversa.

Alcuni esempi di grafici di dispersione con i corrispondenti valori del coefficiente di correlazione lineare sono riportati qui sotto.



Nel grafico in alto a destra si può notare come, a dispetto del coefficiente di correlazione lineare quasi nullo, sia evidente una correlazione tra le due variabili. La funzione che può venire in mente è quadratica, cioè del tipo $y = ax^2 + bx + c$. Quindi, per cercare la parabola che più si avvicina ai punti del grafico, il metodo precedente deve essere modificato. Questo tipo di problema è riportato su molti manuali di statistica e già implementato in diversi software.

Esistono comunque classi di funzioni non lineari per le quali è possibile applicare esattamente lo stesso metodo delle funzioni lineari. Alcune di queste hanno un'importanza tale che è bene

analizzarle separatamente.

A volte accade che il grafico di dispersione relativo a due variabili metta in evidenza una netta correlazione tra di esse con i punti che si addensano intorno ad una curva molto simile a mezza parabola. In questi casi, eventualmente dopo aver effettuato qualche cambiamento nelle unità di misura, si può congetturare una relazione del tipo

$$y = ax^p \,.$$

Sostituendo ad X e Y i corrispondenti logaritmi (in base e o diversa) possiamo scrivere

$$\tilde{y} = \ln y = \ln(ax^p) = \ln a + p \ln x = c + p\tilde{x} \,.$$

I calcoli precedenti ci assicurano che la scelta migliore dei parametri p e c è data da

$$c = \overline{\tilde{x}} - p\overline{\tilde{y}} \qquad \text{con} \qquad p = \frac{Cov(\ln X, \ln Y)}{Var(\ln X)} \,,$$

dove $\overline{\tilde{x}}$ e $\overline{\tilde{y}}$ sono definiti da

$$\overline{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \qquad \overline{\tilde{y}} = \frac{1}{n}\sum_{i=1}^{n} \ln y_i \,,$$

e dunque sono il logaritmo delle corrispondenti medie geometriche.

Prendiamo come esempio i dati contenuti nella seguente tabella.

Pianeta	Mer	Ven	Ter	Mar	Gio	Sat	Ura	Net	Plu
distanza media dal sole	57,9	108,2	149,6	227,9	778,3	1429,4	2871,0	4504,3	5913,5
periodo di rivoluzione	7,6	19,4	31,6	59,4	374,3	929,7	2651,2	5200,4	7816,7

Questi forniscono per ogni pianeta del sistema solare la distanza media dal sole (in milioni di chilometri) ed il periodo di rivoluzione (in milioni di secondi). La relazione esistente tra queste due caratteristiche dei pianeti è collegata alla terza legge di Keplero (1619): *il quadrato del periodo di rivoluzione di un pianeta intorno al sole è direttamente proporzionale al cubo del semiasse maggiore della sua orbita* (che è un'ellisse per la prima legge di Keplero). La differenza tra distanza media dal sole e semiasse maggiore dell'orbita è molto piccola, data la lieve eccentricità delle orbite di ogni pianeta. Pertanto ci aspettiamo che, con buona approssimazione, il periodo di rivoluzione T sia relazionato alla distanza media dal sole D da una funzione del tipo $T = aD^p$.

Mostriamo, con l'aiuto di una tabella, i calcoli per determinare il coefficiente di correlazione lineare e l'equazione della retta di regressione.

	D	T	$x = \ln D$	$y = \ln T$	x^2	y^2	xy
Mer	57,9	7,6	4,0587	2,0281	16,4732	4,1134	8,2317
Ven	108,2	19,4	4,6840	2,9653	21,9397	8,7928	13,8893
Ter	149,6	31,6	5,0080	3,4532	25,0797	11,9243	17,2933
Mar	227,9	59,4	5,4289	4,0843	29,4730	16,6815	22,1733
Gio	778,3	374,3	6,6571	5,9251	44,3171	35,1063	39,4438
Sat	1429,4	929,7	7,2650	6,8349	52,7804	46,7153	49,6553
Ura	2871,0	2651,2	7,9624	7,8828	63,4001	62,1380	62,7659
Net	4504,3	5200,4	8,4128	8,5565	70,7750	73,2135	71,9839
Plu	5913,5	7816,7	8,6850	8,9640	75,4291	80,3536	77,8524
medie	**1782,2333**	**1898,9222**	**6,4624**	**5,6327**	**44,4075**	**37,6710**	**40,3654**

Indicati con x e y i logaritmi delle distanze e dei periodi di ogni pianeta, abbiamo aggiunto tre colonne con i valori x^2, y^2 e xy. Le celle dell'ultima riga contengono la media aritmetica dei nove numeri nella loro colonna; indichiamo con $E(x)$ la media della colonna corrispondente a

x e similmente per le altre colonne. Dalla formula (13), valida anche per la varianza empirica, ricaviamo

$$\begin{aligned} Var(x) &= E(x^2) - E(x)^2 = 44,4075 - (6,4624)^2 = 2,6444\,, \\ Var(y) &= E(y^2) - E(y)^2 = 37,6710 - (5,6327)^2 = 5,9440\,, \\ Cov(x,y) &= E(xy) - E(x)E(y) = 40,3654 - 6,4624 \cdot 5,6327 = 3,9647\,. \end{aligned}$$

Da queste otteniamo immediatamente il coefficiente di correlazione lineare

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}} = \frac{3,9647}{\sqrt{2,6444 \cdot 5,9440}} = 0,99999993\,,$$

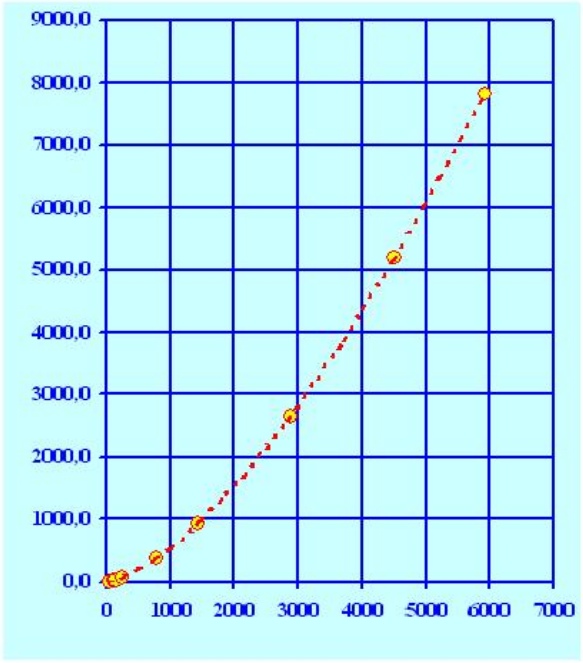
il coefficiente angolare della retta di regressione

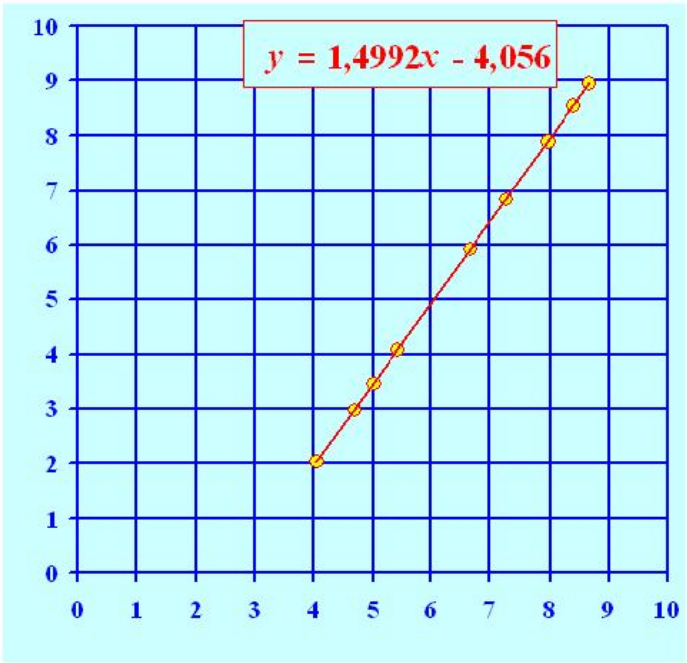
$$m = \frac{Cov(x,y)}{Var(x)} = \frac{3,9647}{2,6444} = 1,4992\,,$$

da cui possiamo dedurre

$$q = E(y) - mE(x) = 5,6327 - 1,4992 \cdot 6,4624 = -4,0560\,.$$

Qui sotto sono rappresentati i grafici di dispersione di T e D a sinistra (con la curva del tipo congetturato $T = D^{1,4992}/57{,}7429$ che meglio approssima i dati) e di $\ln T$ e $\ln D$ a destra.



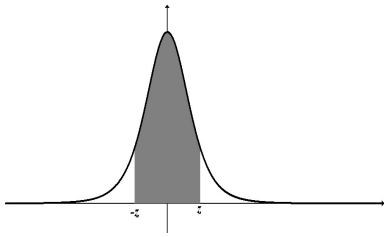


Nel grafico a destra è evidenziata anche la retta di regressione con la relativa equazione. I punti del grafico sono vicinissimi alla retta di regressione ed infatti, come abbiamo appena visto, il coefficiente di correlazione lineare risulta essere praticamente 1.

Distribuzione t di Student

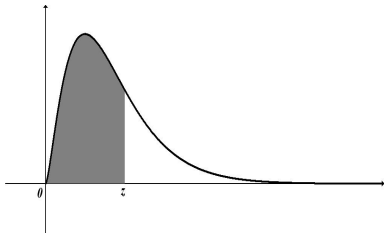
g	0,005	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15
1	127,32111	63,65590	31,82096	21,20505	15,89447	12,70615	10,57888	9,05791	7,91581	7,02636	6,31375	5,72974	5,24218	4,82882	4,47375	4,16530
2	14,08916	9,92499	6,96455	5,64280	4,84873	4,30266	3,89642	3,57824	3,31976	3,10398	2,91999	2,76043	2,62016	2,49541	2,38338	2,28193
3	7,45320	5,84085	4,54071	3,89606	3,48191	3,18245	2,95051	2,76260	2,60543	2,47081	2,35336	2,24939	2,15624	2,07194	1,99502	1,92432
4	5,59754	4,60408	3,74694	3,29763	2,99853	2,77645	2,60076	2,45589	2,33287	2,22610	2,13185	2,04752	1,97123	1,90159	1,83752	1,77819
5	4,77332	4,03212	3,36493	3,00288	2,75651	2,57058	2,42158	2,29739	2,19096	2,09784	2,01505	1,94050	1,87268	1,81043	1,75289	1,69936
6	4,31683	3,70743	3,14267	2,82893	2,61224	2,44691	2,31326	2,20106	2,10431	2,01920	1,94318	1,87444	1,81165	1,75383	1,70020	1,65017
7	4,02935	3,49948	2,99795	2,71457	2,51675	2,36462	2,24088	2,13645	2,04601	1,96615	1,89458	1,82966	1,77021	1,71532	1,66430	1,61659
8	3,83254	3,35538	2,89647	2,63381	2,44899	2,30601	2,18915	2,09016	2,00415	1,92799	1,85955	1,79733	1,74024	1,68744	1,63827	1,59222
9	3,68964	3,24984	2,82143	2,57381	2,39844	2,26216	2,15037	2,05539	1,97265	1,89922	1,83311	1,77291	1,71758	1,66633	1,61854	1,57374
10	3,58137	3,16926	2,76377	2,52749	2,35931	2,22814	2,12023	2,02833	1,94810	1,87677	1,81246	1,75382	1,69984	1,64979	1,60308	1,55924
15	3,28604	2,94673	2,60248	2,39701	2,24854	2,13145	2,03429	1,95094	1,87774	1,81232	1,75305	1,69878	1,64865	1,60200	1,55833	1,51723
20	3,15340	2,84534	2,52798	2,33625	2,19666	2,08596	1,99371	1,91429	1,84433	1,78164	1,72472	1,67249	1,62415	1,57910	1,53685	1,49704
25	3,07820	2,78744	2,48510	2,30113	2,16659	2,05954	1,97009	1,89293	1,82483	1,76371	1,70814	1,65709	1,60979	1,56566	1,52424	1,48517
30	3,02978	2,74998	2,45726	2,27827	2,14697	2,04227	1,95465	1,87894	1,81205	1,75195	1,69726	1,64698	1,60035	1,55683	1,51595	1,47737
35	2,99606	2,72381	2,43772	2,26219	2,13316	2,03011	1,94375	1,86907	1,80302	1,74365	1,68957	1,63983	1,59368	1,55057	1,51008	1,47184
40	2,97117	2,70446	2,42326	2,25027	2,12291	2,02107	1,93566	1,86173	1,79631	1,73747	1,68385	1,63450	1,58871	1,54592	1,50570	1,46772
45	2,95207	2,68959	2,41212	2,24109	2,11500	2,01410	1,92941	1,85606	1,79113	1,73269	1,67943	1,63039	1,58487	1,54232	1,50232	1,46453
50	2,93698	2,67779	2,40327	2,23378	2,10872	2,00856	1,92444	1,85155	1,78700	1,72889	1,67591	1,62711	1,58180	1,53945	1,49962	1,46199
60	2,91457	2,66027	2,39012	2,22292	2,09936	2,00030	1,91702	1,84483	1,78085	1,72322	1,67065	1,62222	1,57723	1,53517	1,49560	1,45820
70	2,89874	2,64790	2,38080	2,21523	2,09273	1,99444	1,91177	1,84005	1,77648	1,71919	1,66692	1,61874	1,57399	1,53212	1,49274	1,45551
80	2,88695	2,63870	2,37387	2,20949	2,08778	1,99007	1,90784	1,83649	1,77321	1,71618	1,66413	1,61614	1,57156	1,52985	1,49060	1,45349
90	2,87790	2,63157	2,36850	2,20504	2,08394	1,98667	1,90480	1,83372	1,77068	1,71384	1,66196	1,61413	1,56967	1,52808	1,48894	1,45192
100	2,87066	2,62589	2,36421	2,20150	2,08088	1,98397	1,90237	1,83152	1,76866	1,71198	1,66023	1,61252	1,56817	1,52667	1,48761	1,45068
1000	2,81329	2,58075	2,33008	2,17319	2,05643	1,96234	1,88293	1,81385	1,75247	1,69704	1,64638	1,59961	1,55610	1,51535	1,47696	1,44064
10000	2,80765	2,57633	2,32672	2,17040	2,05402	1,96020	1,88101	1,81210	1,75086	1,69556	1,64501	1,59833	1,55491	1,51423	1,47591	1,43964
100000	2,80710	2,57587	2,32639	2,17012	2,05378	1,95999	1,88082	1,81193	1,75070	1,69541	1,64487	1,59821	1,55479	1,51412	1,47580	1,43954
1000000	2,80703	2,57583	2,32635	2,17009	2,05375	1,95997	1,88079	1,81191	1,75069	1,69540	1,64486	1,59819	1,55477	1,51410	1,47579	1,43953
10000000	2,80703	2,57583	2,32635	2,17009	2,05375	1,95997	1,88079	1,81191	1,75069	1,69540	1,64485	1,59819	1,55477	1,51410	1,47579	1,43953
∞	2,80706	2,57583	2,32634	2,17009	2,05375	1,95996	1,88079	1,81191	1,75069	1,69540	1,64485	1,59819	1,55477	1,51410	1,47579	1,43953

Nella prima colonna di entrambe le tabella sono indicati i gradi di libertà della corrispondente distribuzione, mentre nella prima riga sono indicati i livelli di significatività. Il numero $2,22814$ della tabella qui sopra nella riga corrispondente a 10 gradi di libertà e nella colonna relativa alla significatività del 5% indica che, per la distribuzione di Student di ordine 10, l'intervallo da $-2,22814$ a $2,22814$ è un intervallo di confidenza al 95%.



L'ultima riga della tabella qui sopra corrisponde ad infiniti gradi di libertà e quindi alla distribuzione normale standard.

Nella tabella relativa alla distribuzione χ^2 gli intervalli di confidenza sono invece unilateri. Il numero $9,83659$ nella sesta riga e decima colonna indica che per la distribuzione χ^2 l'intervallo da 0 a $9,83659$ è un intervallo di confidenza al 92%.



Distribuzione χ^2

g	0,005	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15
1	7,87940	6,63489	5,41190	4,70930	4,21787	3,84146	3,53738	3,28302	3,06490	2,87437	2,70554	2,55422	2,41732	2,29250	2,17796	2,07225
2	10,59653	9,21035	7,82407	7,01310	6,43774	5,99148	5,62682	5,31852	5,05145	4,81589	4,60518	4,41455	4,24053	4,08044	3,93222	3,79424
3	12,83807	11,34488	9,83741	8,94730	8,31115	7,81472	7,40688	7,06031	6,75869	6,49146	6,25139	6,03332	5,83345	5,64888	5,47735	5,31705
4	14,86017	13,27670	11,66784	10,71189	10,02550	9,48773	9,04437	8,66642	8,33653	8,04343	7,77943	7,53904	7,31817	7,11370	6,92325	6,74488
5	16,74965	15,08632	13,38822	12,37461	11,64434	11,07048	10,59622	10,19102	9,83659	9,52107	9,23635	8,97663	8,73758	8,51595	8,30917	8,11520
6	18,54751	16,81187	15,03320	13,96763	13,19781	12,59158	12,08957	11,65992	11,28349	10,94791	10,64464	10,36762	10,11234	9,87537	9,65399	9,44610
7	20,27774	18,47532	16,62243	15,50906	14,70304	14,06713	13,53973	13,08770	12,69118	12,33724	12,01703	11,72423	11,45413	11,20315	10,96848	10,74789
8	21,95486	20,09016	18,16820	17,01052	16,17077	15,50731	14,95634	14,48356	14,06839	13,69746	13,36156	13,05414	12,77032	12,50638	12,25939	12,02707
9	23,58927	21,66605	19,67898	18,47956	17,60826	16,91896	16,34591	15,85371	15,42108	15,03422	14,68366	14,36256	14,06590	13,78983	13,53134	13,28803
10	25,18805	23,20929	21,16075	19,92189	19,02077	18,30703	17,71312	17,20257	16,75348	16,35159	15,98717	15,65318	15,34442	15,05693	14,78759	14,53393
15	32,80149	30,57795	28,25949	26,84796	25,81614	24,99580	24,31079	23,72020	23,19924	22,73192	22,30712	21,91694	21,55550	21,21824	20,90169	20,60300
20	39,99686	37,56627	35,01962	33,46234	32,32056	31,41042	30,64885	29,99097	29,40969	28,88741	28,41197	27,97468	27,56904	27,19009	26,83398	26,49757
25	46,92797	44,31401	41,56603	39,88040	38,64167	37,65249	36,82354	36,10646	35,47211	34,90152	34,38158	33,90287	33,45842	33,04286	32,65199	32,28248
30	53,67187	50,89218	47,96179	46,15995	44,83353	43,77295	42,88314	42,11261	41,43034	40,81614	40,25602	39,73995	39,26045	38,81185	38,38963	37,99024
35	60,27459	57,34199	54,24386	52,33505	50,92810	49,80183	48,85603	48,03638	47,31013	46,65582	46,05877	45,50840	44,99673	44,51777	44,06674	43,63993
40	66,76605	63,69077	60,43607	58,42781	56,94586	55,75849	54,76058	53,89521	53,12798	52,43642	51,80504	51,22271	50,68113	50,17396	49,69616	49,24385
45	73,16604	69,95690	66,55521	64,45350	62,90104	61,65622	60,60941	59,70114	58,89547	58,16891	57,50529	56,89299	56,32333	55,78966	55,28674	54,81047
50	79,48984	76,15380	72,61322	70,42295	68,80386	67,50481	66,41175	65,46291	64,62086	63,86120	63,16711	62,52649	61,93025	61,37154	60,84488	60,34601
60	91,95181	88,37943	84,57990	82,22506	80,48196	79,08195	77,90286	76,87855	75,96889	75,14772	74,39700	73,70368	73,05811	72,45284	71,88204	71,34110
70	104,21477	100,42505	96,38750	93,88126	92,02405	90,53126	89,27306	88,17940	87,20758	86,32985	85,52704	84,78531	84,09438	83,44633	82,83496	82,25535
80	116,32093	112,32879	108,06929	105,42203	103,45877	101,87947	100,54774	99,38945	98,35979	97,42949	96,57820	95,79146	95,05830	94,37044	93,72132	93,10575
90	128,29868	124,11620	119,64845	116,86876	114,80571	113,14523	111,74436	110,52549	109,44151	108,46175	107,56501	106,73594	105,96316	105,23793	104,55340	103,90405
100	140,16971	135,80689	131,14171	128,23668	126,07935	124,34210	122,87586	121,59959	120,46426	119,43777	118,49800	117,62892	116,81869	116,05811	115,34006	114,65881
1000	1118,94751	1106,96899	1093,97726	1085,78684	1079,65250	1074,67937	1070,45861	1066,76693	1063,46855	1060,47474	1057,72396	1055,17143	1052,78432	1050,53686	1048,40905	1046,38493

Distribuzione normale standard

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,17003	0,17364	0,17724	0,18082	0,18439	0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,20540	0,20884	0,21226	0,21566	0,21904	0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,23891	0,24215	0,24537	0,24857	0,25175	0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,27035	0,27337	0,27637	0,27935	0,28230	0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,29955	0,30234	0,30511	0,30785	0,31057	0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,32639	0,32894	0,33147	0,33398	0,33646	0,33891
1,0	0,34134	0,34375	0,34614	0,34849	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41308	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42785	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,44295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327
1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2,0	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158
2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3,0	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49896	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952	0,49953	0,49955	0,49957	0,49958	0,49960	0,49961	0,49962	0,49964	0,49965
3,4	0,49966	0,49968	0,49969	0,49970	0,49971	0,49972	0,49973	0,49974	0,49975	0,49976
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983
3,6	0,49984	0,49985	0,49985	0,49986	0,49986	0,49987	0,49987	0,49988	0,49988	0,49989
3,7	0,49989	0,49990	0,49990	0,49990	0,49991	0,49991	0,49992	0,49992	0,49992	0,49992
3,8	0,49993	0,49993	0,49993	0,49994	0,49994	0,49994	0,49994	0,49995	0,49995	0,49995
3,9	0,49995	0,49995	0,49996	0,49996	0,49996	0,49996	0,49996	0,49996	0,49997	0,49997
4,0	0,49997	0,49997	0,49997	0,49997	0,49997	0,49997	0,49998	0,49998	0,49998	0,49998

La tabella mostra i valori, approssimati alla quinta cifra decimale, della probabilità degli eventi $\{0 \le X \le z\}$ se X ha distribuzione normale standard (media 0 e varianza 1), cioè l'area della regione evidenziata in figura. L'estremo z è la somma dei numeri in grassetto all'inizio delle corrispondenti righe e colonne.

Indicato con $T(z)$ il valore fornito dalla tabella per l'estremo z, sfruttando la simmetria della funzione Gaussiana $G(x)$ possiamo calcolare l'integrale esteso ad un qualsiasi intervallo. Ad esempio avremo:

$$\int_0^{1,24} G(x)\,dx = T(1,24) = 0,39251 = \int_{-1,24}^{0} G(x)\,dx$$

$$\int_{-1,1}^{2,37} G(x)\,dx = T(1,1) + T(2,37) = 0,36433 + 0,49111 = 0,85544$$

$$\int_{1,32}^{1,67} G(x)\,dx = T(1,67) - T(1,32) = 0,45254 - 0,40658 = 0,04596\,.$$