CAPITOLO II

DISTRIBUZIONI E LEGGI DI PROBABILITA'

CAPITOLO II

DISTRIBUZIONI E LEGGI DI PROBABILITÀ'

2.1. ELEMENTI DI CALCOLO COMBINATORIO SEMPLICE

La stima della probabilità di un evento è uno strumento fondamentale della statistica. Nelle sue forme più semplici, si fonda sul calcolo combinatorio. E' evidente ed intuitiva la sua applicazione ai giochi d'azzardo, ai quali effettivamente fu associata alla sua origine. Anche se il risultato di ogni singolo tentativo è imprevedibile, **con un numero elevato di ripetizioni si stabiliscono regolarità** che possono essere previste e calcolate. Dal punto di vista didattico, l'associazione del concetto di probabilità al calcolo combinatorio è un aspetto importante: serve per **collegare una scelta alla probabilità con la quale l'evento** atteso **può avvenire**, nel contesto di tutti gli eventi alternativi possibili. E' la base dell'inferenza statistica, della scelta scientifica in tutti i casi d'incertezza.

I concetti e i metodi del calcolo combinatorio possono essere spiegati in modo semplice, con una serie di esempi, tra loro collegati, per cogliere somiglianze e differenze nelle risposte che forniscono.
In una corsa con 10 concorrenti, che abbiano le medesime possibilità di vittoria, è possibile porsi molti quesiti, tra i quali:

a) quanti differenti ordini d'arrivo sono possibili?
b) quale è la probabilità di indovinare i primi 3 al traguardo, secondo l'ordine?
c) quale la probabilità di indovinare i primi 3, senza considerare il loro ordine?
d) è conveniente scommettere 10 mila lire per guadagnarne 500 mila, se si indovinassero i primi 2 nell'ordine?
e) è conveniente senza stabilire l'ordine?

Per calcolare le probabilità richieste, occorre prestare attenzione alle 4 caratteristiche fondamentali di questi **eventi**:

(1) ***si escludono a vicenda***,
(2) sono ***tutti ugualmente possibili***,
(3) sono ***casuali***,
(4) sono ***indipendenti***.

Il calcolo combinatorio di **raggruppamenti semplici** o **senza ripetizione**, così definiti in quanto ogni elemento compare una volta sola (in altri termini, lo stesso oggetto deve presentarsi in ciascun gruppo

una volta sola), permette di calcolare la probabilità con cui può avvenire ogni evento possibile, che rispetti le 4 condizioni citate.
Se le condizioni fossero differenti, si dovrebbe ricorrere ad altri metodi.
Per esempio, quando lo stesso oggetto può essere presente più volte in uno stesso gruppo (come l'estrazione ripetuta di una carta rimessa ogni volta nel mazzo), si devono utilizzare i raggruppamenti con ripetizione (o calcolo combinatorio con ripetizioni).

Nel **calcolo combinatorio semplice**, i raggruppamenti possibili possono essere distinti in **permutazioni, disposizioni, combinazioni**.

2.1.1 Permutazioni semplici.

Dato un insieme di *n* oggetti differenti a_1, a_2, a_3, ..., a_n, si chiamano permutazioni semplici tutti i sottoinsiemi che si possono formare, collocando gli *n* elementi in tutti gli ordini possibili.
Alcuni esempi di permutazione delle 4 lettere a, b, c, d sono: abcd, abdc, acbd, adcb, cabd, cdba, dbac, cbda, ecc.

Il numero di permutazioni di **n** elementi è

$$P_n = n!$$

dove **n!** (**n** fattoriale) è il prodotto degli **n** elementi: **n!** $= 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$.

Esempio 1: le permutazioni delle 4 lettere (P_4) a, b, c, d, sono 4!

$$\mathbf{P_4} = 4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$$

Esempio 2: le permutazioni di 3 elementi abc sono: abc, acb, bca, bac, cba, cab;
cioè

$$\mathbf{P_3} = 3! = 6$$

Per i calcoli che saranno proposti durante il corso, è utile ricordare che, per definizione,

$$0! = 1$$

e che

$$1! = 1$$

Tabella dei fattoriali di interi (per facilitare i calcoli).

n	*n*!	*n*	*n*!
1	1	26	4.03291×10^{26}
2	2	27	1.08889×10^{28}
3	6	28	3.04888×10^{29}
4	24	29	8.84176×10^{30}
5	120	30	2.65253×10^{32}
6	720	31	8.22284×10^{33}
7	5040	32	2.63131×10^{35}
8	40320	33	8.68332×10^{36}
9	362880	34	2.95233×10^{38}
10	$3.62880 \times 10^{6.}$	35	1.03331×10^{40}
11	3.99168×10^{7}	36	3.71993×10^{41}
12	4.79002×10^{8}	37	1.37638×10^{43}
13	6.22702×10^{9}	38	5.23023×10^{44}
14	8.71783×10^{10}	39	2.03979×10^{46}
15	1.30767×10^{12}	40	8.15915×10^{47}
16	2.09228×10^{13}	41	3.34525×10^{49}
17	3.55687×10^{14}	42	1.40501×10^{51}
18	6.40327×10^{15}	43	6.04153×10^{52}
19	1.21645×10^{17}	44	2.65827×10^{54}
20	2.43290×10^{18}	45	1.19622×10^{56}
21	5.10909×10^{19}	46	5.50262×10^{57}
22	1.12400×10^{21}	47	2.58623×10^{59}
23	2.58520×10^{22}	48	1.24139×10^{61}
24	6.20448×10^{23}	49	6.08282×10^{62}
25	1.55112×10^{25}	50	3.04141×10^{64}

2.1.2 **Disposizioni semplici**.

Dato un insieme di *n* oggetti differenti a_1, a_2, a_3, ..., a_n si chiamano disposizioni semplici i sottoinsiemi di *p* elementi che si diversificano almeno per un elemento o per il loro ordine.

Le disposizioni delle 4 lettere a,b,c,d, raggruppate 3 a 3 sono: abc, acb, bac, dba, bda, abd, ecc.

Il numero di disposizioni semplici di **n** elementi **p** a **p** è

$$D_n^p = \frac{n!}{(n-p)!}$$

Esempio 1: le disposizioni di 4 elementi 3 a 3 sono:

$$D_4^3 = \frac{4!}{(4-3)!} = \frac{24}{1} = 24$$

Derivato dalla semplificazione di questa formula, un altro modo per calcolare le disposizioni semplici di **n** elementi **p** a **p** è

$$D_n^p = n(n-1)(n-2)\ldots(n-p+1)$$

Le disposizioni di 4 elementi 3 a 3 possono quindi essere calcolate anche mediante

$$D_4^3 = 4(4-1)(4-2) = 4 \cdot 3 \cdot 2 = 24$$

Esempio 2: le disposizioni di 7 elementi 3 a 3 sono:

$$D_7^3 = 7(7-1)(7-2) = 7 \cdot 6 \cdot 5 = 210$$

2.1.3 **Combinazioni semplici**

Dato un insieme di *n* oggetti differenti a_1, a_2, a_3, ..., a_n, si chiamano combinazioni semplici di *n* elementi *p* a *p* i sottoinsiemi che si diversificano almeno per un elemento, ma non per il loro ordine.

Le combinazioni semplici delle 4 lettere a,b,c,d, 3 a 3 sono: abc, abd, acd, bcd.

Il numero di combinazioni semplici di **n** elementi **p** a **p** è

$$C_n^p = \frac{n!}{(n-p)!\ p!}$$

Sotto l'aspetto del calcolo e dal punto di vista concettuale, il numero di combinazioni di **n** elementi **p** a **p** corrisponde al rapporto tra il numero di disposizioni di **n** elementi **p** a **p** ed il numero di permutazioni di **p** elementi.

Esempio 1: le combinazioni di 4 elementi 3 a 3 sono

$$C_4^3 = \frac{4!}{(4-3)!3!} = 4$$

Per le applicazioni, è utile ricordare tre **casi particolari**:

a) $$C_n^n = \frac{n!}{n!\ 0!} = 1$$

Il numero di combinazioni di **n** elementi presi **n** ad **n** è **1**: c'è un solo sottoinsieme formato da tutti gli elementi.

$$\text{b)} \qquad \mathbf{C_n^1 = \frac{n!}{1!(n-1)!} = n}$$

Il numero di combinazioni di **n** elementi presi **1** a **1** è uguale a **n**: il numero di sottoinsiemi con **1** solo elemento è **n**.

$$\text{c)} \qquad \mathbf{C_n^0 = \frac{n!}{0!n!} = 1}$$

Il numero di combinazioni di **n** elementi **0** a **0** è **1**: c'è un solo sottoinsieme vuoto.

Come è impostato per il calcolo, **il numero di combinazioni è solo apparentemente frazionario**: risulta sempre **n**, numero intero, che si indica con il simbolo $\binom{p}{n}$ chiamato **coefficiente binomiale** e si legge **p** su **n**.

2.1.4 **Risposte alle domande del paragrafo 2.1**

Si è ora in grado di fornire le risposte ai cinque quesiti introdotti nel paragrafo 2.1

a) In una corsa con 10 concorrenti, i possibili ordini d'arrivo sono le permutazioni di 10 elementi. Il loro numero è

$$P_{10} = 10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3.628.800$$

b) In una corsa di 10 concorrenti, il numero dei possibili gruppi differenti formati dai primi 3 all'arrivo, tenendo conto anche del loro ordine, sono le disposizioni di 10 elementi 3 a 3, cioè

$$D_{10}^3 = \frac{10!}{(10-3)!} = 720$$

La probabilità di indovinare i primi 3 concorrenti secondo l'ordine d'arrivo è 1/720 = 0,001389.

c) In una corsa di 10 concorrenti, i possibili gruppi dei primi 3 concorrenti, senza distinzioni interne di ordine, sono le combinazioni di 10 elementi 3 a 3, cioè

$$C_{10}^3 = \frac{10!}{(10-3)!\ 3!} = 120$$

La probabilità di indovinare i primi 3 concorrenti, senza stabilirne l'ordine, è 1/120 = 0,008333; è 6 (3!) volte più alta di quella in cui si chiede di indovinare anche il loro ordine.

d) Il numero di possibili gruppi formati dai primi 2 concorrenti, stabilendo chi sarà il primo e chi il secondo, in un gruppo di 10 è determinato dalle disposizioni di 10 elementi 2 a 2, cioè

$$D_{10}^{2} = \frac{10!}{(10-2)!} = 90$$

La probabilità di indovinare chi saranno i primi 2 è uguale a 1/90. E un rapporto più sfavorevole del rapporto di 1 a 50 fissato nella scommessa. Per chi scommette non è conveniente vincere 50 volte la posta, quando la probabilità di vincere è 1/90.

e) Il numero di possibili gruppi formati dai primi 2 concorrenti, senza stabilire l'ordine, in un gruppo di 10 è dato dalle combinazioni di 10 elementi 2 a 2, cioè

$$C_{10}^{2} = \frac{10!}{(10-2)!\ 2!} = 45$$

La probabilità di indovinare i primi 2 senza dover stabilire l'ordine uguale a 1/45; è più favorevole del rapporto di 1 a 50 fissato dalla scommessa. Per chi scommette è conveniente, perché l'eventuale guadagno è superiore al rischio. Una scommessa tra i due giocatori è in parità, solamente quando il prodotto tra la probabilità d'indovinare e il moltiplicatore della posta è uguale a 1.

ESERCIZI

1. In un esperimento sulla fertilità del terreno, si vogliono studiare in modo sistematico gli equilibri binari tra i seguenti elementi: Ca, Mg, Na, N, P, K.

A. Quante coppie di elementi occorrerà prendere in considerazione?

(Risposta : $C_{6}^{2} = \frac{6!}{(6-2)!\,2!} = \frac{5 \cdot 6}{2} = 15$)

B. Se si intende valutare tutti gli equilibri ternari, quanti gruppi diversi formati da tre elementi occorrerà formare?

(Risposta : $C_{6}^{3} = \frac{6!}{(6-3)!\,3!} = 20$)

2. Nel timore che, durante una settimana con molto traffico, il tasso d'inquinamento dell'aria in una città fosse in costante aumento, è stata effettuata una serie di rilevazioni.

A. Quale è la probabilità che, senza un reale aumento dell'inquinamento e solo per caso, i valori osservati siano tutti in ordine crescente, se sono state fatte 4 rilevazioni?

(Risposta : $P_4 = 4! = 24$; la probabilità è 1/24 = 0,04166 o 4,166%)

B. E se durante la settimana è stata fatta una rilevazione per giorno, quale è la probabilità che solo per caso siano tutti in ordine decrescente?

(Risposta : $P_7 = 7! = 5.040$; la probabilità è 1/5.050 = 0,000198 o 0,0198%)

3. Per una serie di misure, da un gruppo di animali di dimensioni differenti sono stati estratti alcuni individui.

A. Se si estraggono casualmente 3 individui da un gruppo di 15, quale è la probabilità che essi siano i 3 con dimensioni maggiori?

(Risposta : $C^3_{15} = 15!/ 3!\ 12! = 455$; la probabilità è 1/455 = 0,00219 o 0,219%)

B. Se si estraggono 4 animali da un gruppo di 20, quale è la probabilità che per caso possano essere i 4 con dimensioni minori?

(Risposta: $C^4_{20} = 20! / 4!\ 16! = 4.845$; la probabilità è 1/4.845 = 0,00021 o 0,021%)

2.2. DEFINIZIONI DI PROBABILITA': MATEMATICA, FREQUENTISTA E SOGGETTIVA, CON ELEMENTI DI STATISTICA BAYESIANA

In un gioco fondato su **eventi casuali**, **mutuamente esclusivi**, **ugualmente possibili ed indipendenti**, il risultato di ogni singolo tentativo è imprevedibile. Ma su un elevato numero di tentativi, si stabiliscono regolarità e leggi. A lungo termine, l'esito è prevedibile. Le probabilità dei successi e quelle d'ogni evento alternativo possono essere calcolate con precisione crescente all'aumentare del numero di osservazioni.

La natura del concetto di probabilità è chiarita dal **teorema** di **Bernoulli**. Prende il nome da **Jacques Bernoulli** (1654-1705), matematico svizzero, per quanto egli ha scritto nel suo libro ***Ars Conjectandi,*** pubblicato postumo nel 1713.

Questo teorema è chiamato anche **legge dei grandi numeri** e meno frequentemente **convergenza in probabilità**.

Può essere enunciato in questo modo:

- un evento, che abbia probabilità costanti P, in una serie di prove tende a P, al crescere del numero di tentativi.

Il concetto di **probabilità classica,** fondato su una **probabilità matematica** o **a priori,** è stato il primo ad essere definito. E' la probabilità di ottenere testa o croce con una moneta, di avere un numero da 1 a 6 con un dado o in più lanci, di prevedere i possibili ordini d'arrivo in una gara con diversi concorrenti che rispettino le 4 condizioni citate. Non si richiede nessun dato sperimentale; i risultati sono conosciuti a priori, senza attendere alcuna rilevazione od osservazione, poiché **è sufficiente il solo ragionamento logico per calcolare con precisione le probabilità**. Se la moneta, il dado e la gara non sono truccate, le verifiche sperimentali si allontaneranno dai dati attesi solo per quantità trascurabili, determinate da eventi casuali o da errori di misura.
Queste probabilità non sono determinate solamente da leggi matematiche. Una volta compresi i meccanismi della natura, molte discipline evidenziano regolarità, sovente chiamate leggi, che permettono di stimare in anticipo e con rilevante precisione i risultati di esperimenti od osservazioni. In biologia è il caso sia delle leggi di Mendel, che permettono di calcolare le frequenze genotipiche e fenotipiche attese nell'incrocio tra ibridi, sia della legge di Hardy-Weinberg, utile per definire le frequenze genetiche in una popolazione mendeliana panmittica.
Fermo restando il numero di casi possibili, la probabilità di un evento aumenta quando cresce il numero dei casi favorevoli; oppure quando, fermo restando il numero di casi favorevoli, diminuisce quello dei casi possibili.
La **definizione di probabilità classica** è attribuita sia a **Bernouilli**, sia a **Laplace** (il francese Pierre-Simon, marchese di Laplace, nato nel 1749 e morto nel 1827, le cui opere più importanti sono *Traité de Mécanique Céleste* in cinque volumi e il volume *Théorie Analytique des Probabilité*):
- ***la probabilità di un evento casuale è il rapporto tra il numero di casi favorevoli ed il numero di casi possibili, purché siano tutti equiprobabili***.

La stima di una **probabilità a priori** ha limitazioni gravi nella ricerca sperimentale: per calcolare la probabilità di un evento, è necessario conoscere preventivamente le diverse probabilità di tutti gli eventi possibili. Pertanto, questo approccio non può essere utilizzato sempre. Con la probabilità matematica non è assolutamente possibile rispondere a quesiti che per loro natura richiedono un approccio empirico, che possono essere fondati solo su osservazioni sperimentali ripetute, poiché i diversi risultati ottenibili non sono tutti ugualmente possibili, né casuali. Per esempio, come rispondere alla domanda: “Con un dado truccato, quali sono le probabilità che esca il numero 5?”. Occorre fare esperimenti con lanci ripetuti, per sapere come quel dado è truccato e quindi quali siano le effettive probabilità di ogni numero.

Nello stesso modo, se nella segregazione di un diibrido o nella distribuzione di un carattere ereditario intervengono fenomeni di selezione, con un'intensità ignota, quali saranno le probabilità dei vari fenotipi di essere presenti in una data generazione? Ogni alterazione dei rapporti di equiprobabilità o dei fenomeni di casualità, i soli che possano essere stimati a priori sulla base della logica, richiede l'esperienza di almeno una serie di osservazioni ripetute.

Come stima della **probabilità di un evento sperimentale** può essere utilizzata la sua **frequenza**, quando essa nelle varie ripetizioni si mantiene approssimativamente costante. Una definizione chiara di questo concetto è stata data nel 1920 dal matematico **von Mises (**Richard Martin Edler von Mises, nato in Russia nel 1883 e morto in America nel 1953, dopo aver insegnato matematica applicata a Berlino e a Istanbul**).**
Egli scrive:
- ***la probabilità di un evento casuale è il limite a cui essa tende al crescere del numero delle osservazioni, in una serie di esperienze ripetute nelle stesse condizioni***.

Se ***F*** è la frequenza relativa di un evento in una popolazione, generalmente si può osservare che, all'aumentare del numero di osservazioni (**n),** la frequenza (***f***) del campione tende a diventare sempre più simile a quella reale o della popolazione (F). Questa affermazione non può essere dimostrata né con gli strumenti della matematica, in quanto si riferisce a dati osservazionali, né in modo empirico, poiché nella realtà un esperimento non può essere ripetuto infinite volte. Tuttavia è una **regolarità statistica,** chiamata **legge empirica del caso**, che costituisce la base sperimentale sia di ogni teoria statistica sia del ragionamento matematico.
In questi casi, si parla di **probabilità frequentista** o **frequentistica**, di **probabilità a posteriori**, di **legge empirica del caso** o di **probabilità statistica**.
Essa può essere applicata in tutti i casi in cui le leggi dei fenomeni studiati non sono note a priori, ma possono essere determinate solo a posteriori, sulla base dell'osservazione e delle misure statistiche. Per calcolare la probabilità di trovare un numero prestabilito di individui di una certa specie in una popolazione formata da specie diverse, deve essere nota una legge o regola della sua presenza percentuale, che può essere stabilita solamente con una serie di rilevazioni. Per stimare la probabilità di trovare un individuo oltre una cerca dimensione, è necessario disporre di una distribuzione sperimentale delle misure presenti in quella popolazione.
Solamente in un modo è possibile rispondere a molti quesiti empirici: concepire una serie di osservazioni od esperimenti, in condizioni uniformi o controllate statisticamente, per rilevarne la frequenza.

I due tipi di probabilità presentati, quella classica e quella frequentista, hanno una caratteristica fondamentale in comune: entrambe richiedono che **i vari eventi possano essere ripetuti e verificati in condizioni uniformi o approssimativamente tali**. In altri termini, si richiede che quanto avvenuto nel passato possa ripetersi in futuro. Ma esistono anche fenomeni che non possono assolutamente essere ridotti a queste condizioni generali, perché considerati **eventi unici od irripetibili**. Per esempio, come è possibile rispondere alle domande: "Quale è la probabilità che avvenga una catastrofe o che entro la fine dell'anno scoppi la terza guerra mondiale? Quale è la probabilità che una specie animale o vegetale a rischio effettivamente scompaia? Quale è la probabilità che un lago, osservato per la prima volta, sia effettivamente inquinato?".

E' il caso del medico che deve stabilire se il paziente che sta visitando è ammalato; di una giuria che deve emettere un giudizio di colpevolezza o di assoluzione; di tutti coloro che devono decidere in tante situazioni uniche, diverse ed irripetibili.
Sono situazioni che presuppongono
- il **giudizio di numerosi individui sullo stesso fenomeno** oppure **la stima personale di un solo individuo** sulla base di un suo pregiudizio o della sua esperienza pregressa.

Il valore di probabilità iniziale non è fondato né sulla logica matematica né su una serie di esperimenti.
Nella teoria della probabilità si sono voluti comprendere anche questi **fenomeni non ripetibili**. Da questa scelta, deriva un'altra concezione della **probabilità**: quella **soggettiva o personalistica**.

L'obiezione fondamentale a questa **probabilità logica** è come misurare un grado di aspettativa, quando è noto che individui diversi attribuiscono probabilità differenti allo stesso fenomeno. E' una critica che viene superata dall'approccio soggettivo, secondo il quale
- ***la probabilità è una stima del grado di aspettativa di un evento, secondo l'esperienza personale di un individuo***.

La probabilità nell'impostazione **soggettivista,** detta anche "**bayesiana**", viene intesa come una misura della convinzione circa l'esito di una prova o che accada un certo evento. E' un approccio che ha vaste ed interessanti applicazioni nelle scienze sociali ed in quelle economiche, dove la sola attesa di un fenomeno o la convinzione di una persona influente sono in grado di incidere sui fenomeni reali, come la svalutazione, i prezzi di mercato, i comportamenti sociali. In medicina, è il caso della decisione sulla cura da prescrivere al paziente.
Per la statistica, il problema fondamentale consiste nell'indicare come si debba modificare la probabilità soggettiva di partenza, in dipendenza dei successivi avvenimenti oggettivi, quando non si

dispone di repliche. Per coloro che ritengono che il mondo esterno sia una realtà oggettiva, conoscibile ed indipendente da loro, la conoscenza obiettiva non può derivare da convinzioni personali o da preferenze individuali; pertanto, l'approccio soggettivo non sarebbe attendibile, in quanto non permetterebbe la conoscenza oggettiva del reale.
Questa varietà e contrapposizione di concetti, sinteticamente esposti in modo elementare, sul significato più esteso e comprensivo di probabilità, si riflettono sulla interpretazione delle stime ottenute da dati sperimentali; ma gli aspetti formali del calcolo variano solo marginalmente. Nel contesto delle scienze sperimentali, **esistono casi di applicazione della probabilità soggettiva**; in particolare, **quando si tratta di scegliere una strategia o prendere una decisione.**

Nella ricerca biologica, ecologica ed ambientale, di norma predominano i casi in cui si studiano eventi ripetibili, in condizioni almeno approssimativamente uguali o simili. Pertanto, **quasi esclusivamente si fa ricorso all'impostazione frequentista della probabilità**, trascurando l'impostazione soggettivistica.
Il teorema di **Bayes** (1702-1761) pubblicato nel 1763 (nel volume ***Essay Towards Solving a Problem in the Doctrine of Chance***), due anni dopo la morte dell'autore (il reverendo Thomas Bayes), si fonda su **probabilità soggettive.** Spesso è presentato come alternativo alla inferenza statistica classica, detta anche empirica.
Più recentemente, sono stati sviluppati metodi quantitativi che incorporano l'informazione raccolta **anche con dati campionari**. E' quindi possibile correggere la stima di una probabilità soggettiva originaria, detta **probabilità a priori**, mediante l'informazione fornita da successive rilevazioni campionarie, per ottenere una **probabilità a posteriori**.

Ad esempio, si consideri un evento A la cui probabilità a priori, soggettiva perché espressione di convinzioni e sovente determinata da carenza di informazioni, sia P(A). Dall'esperienza (analisi di un campione) si rileva la probabilità P (intesa, in questo caso, in senso frequentista od empirico) di un certo evento B. Ci si chiede in quale modo la realizzazione dell'evento B modifichi la probabilità a priori di A, cioè qual è il valore di P(A/B) (la probabilità dell'evento A condizionato B).
(Ricordiamo che l'espressione P(A/B) indica la probabilità condizionale di A rispetto a B, cioè la probabilità dell'evento A stimata sotto la condizione che si sia realizzato l'evento B).

Per il **teorema delle probabilità composte, applicato ad eventi non indipendenti**, la probabilità che gli eventi A e B si verifichino contemporaneamente è data da

$$P(A \text{ e } B) = P(A/B) \cdot P(B) \qquad (1a)$$

ed inversamente anche da

$$P(A \text{ e } B) = P(B/A) \cdot P(A) \qquad (1b)$$

Da queste due equazioni si ottiene

$$P(B/A) \cdot P(A) = P(A/B) \cdot P(B) \qquad (2)$$

e, dividendo per P(A), si perviene

$$P(B/A) = \frac{P(A/B) \cdot P(B)}{P(A)} \qquad (3)$$

alla formula generale.

Da essa è possibile dedurre che, per **k eventi reciprocamente incompatibili** e **collettivamente esaustivi**, dove B_1, B_2, ..., B_3 sono gli **eventi mutuamente esclusivi**, si ottiene il **Teorema di Bayes**

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + \ldots + P(A/B_k)P(B_k)} \qquad (4)$$

dove

- $P(B_i)$ è la probabilità a priori che è attribuita alla popolazione B_i prima che siano conosciuti i dati,
- $P(A/B_i)$ rappresenta la probabilità aggiuntiva dopo che è stata misurata la probabilità di A.

Con un esempio, è possibile presentare in modo elementare la successione logica delle stime di probabilità.

1) Si supponga di disporre di un campione A di individui della specie Y e di chiedersi da quale delle tre località indicate (B_1, B_2, B_3) esso possa provenire.

La **stima iniziale o probabilità a priori**, necessaria quando non si possiede alcuna informazione, è fondata sul caso oppure su distribuzioni teoriche. In questa situazione di totale ignoranza iniziale della possibile provenienza degli individui della specie **Y**, la scelta più logica può essere quella di attribuire ad ognuna delle 3 località le medesime probabilità P; è una **distribuzione non informativa** che può essere scritta come:

$$P(B_1) = P(B_2) = P(B_3) = 1/3$$

2) Come secondo passaggio, si supponga ora che un'analisi abbia rivelato che nelle tre diverse località (B_1, B_2, B_3) che possono esserne il luogo d'origine, questa specie sia presente con frequenze percentuali differenti, in comunità delle stesse dimensioni:

- nella prima, che indichiamo con $\mathbf{B_1}$, ha una presenza del **30%**
- nella seconda, indicata con $\mathbf{B_2}$, una presenza del **50%**
- nella terza, indicata con $\mathbf{B_3}$, il **70%**.

Dopo questa informazione che può essere scritta come

$$P(A/B_1) = 0,3 \qquad P(A/B_2) = 0,5 \qquad P(A/B_3) = 0,7$$

il calcolo delle **probabilità a posteriori,** da attribuire ad ogni comunità, può essere attuato mediante il **teorema di Bayes.**

Da esso mediante i rapporti

$$P(B_1 / A) = \frac{0,3(1/3)}{0,3(1/3) + 0,5(1/3) + 0,7(1/3)} = \frac{0,099}{0,4995} = 0,20$$

$$P(B_2 / A) = \frac{0,5(1/3)}{0,3(1/3) + 0,5(1/3) + 0,7(1/3)} = \frac{0,1665}{0,4995} = 0,33$$

$$P(B_3 / A) = \frac{0,7(1/3)}{0,3(1/3) + 0,5(1/3) + 0,7(1/3)} = \frac{0,2331}{0,4995} = 0,47$$

si ricava che il campione **A** di individui della specie **Y** ha

- una probabilità del **20%** di provenire dalla località $\mathbf{B_1}$,
- una probabilità pari al **33%** di essere originario della località $\mathbf{B_2}$,
- una del **47%** di provenire dalla $\mathbf{B_3}$.

Ovviamente **i 3 eventi sono reciprocamente incompatibili e collettivamente esaustivi, per cui la somma delle probabilità è unitaria**, se espressa in rapporti, o pari al 100%, se espressa in percentuale.

3) Con una ulteriore raccolta di dati (sia soggettivi che sperimentali o frequentisti), è possibile calcolare una nuova distribuzione delle probabilità a posteriori, per inglobare le nuove informazioni ed aggiornare la stima.

Se nel conteggio del campione **A**, si osserva che sono presenti **10** individui di cui **8** della specie **Y**, come variano le probabilità che il campione **A** derivi rispettivamente dalla località $\mathbf{B_1}$, dalla $\mathbf{B_2}$ o dalla $\mathbf{B_3}$?

In questa terza fase, il primo passo logico è calcolare la probabilità di trovare la distribuzione osservata nel campione (**8** individui della specie **Y** su **10** complessivi), se fosse stato **estratto casualmente** da ognuna delle 3 comunità:

a - dalla comunità B_1 dove la specie Y ha una proporzione uguale a 0,3
b - dalla comunità B_2 dove la specie Y ha una proporzione uguale a 0,5
c - dalla comunità B_3 dove la specie Y ha una proporzione uguale a 0,7.

Per rispondere a tale quesito, si ricorre alla distribuzione binomiale (che sarà spiegata successivamente), della quale sono riportati i risultati.

a) Nel caso della comunità **B_1** dove la proporzione **p** di individui della specie **Y** è **0,3**

$$P(A / B_1) = \binom{10}{8} \cdot 0{,}3^8 \cdot (1-0{,}3)^2 = 0{,}00145$$

la probabilità **P** che su **10** individui estratti a caso **8** siano della specie **Y** è uguale a **0,00145**.

b) - Nel caso della comunità **B_2** dove la proporzione **p** di individui della specie **Y** è **0,5**

$$P(A / B_2) = \binom{10}{8} \cdot 0{,}5^8 \cdot (1-0{,}5)^2 = 0{,}04394$$

la probabilità **P** che su **10** individui estratti a caso **8** siano della specie **Y** è uguale a **0,04394**.

3 - Nel caso della comunità **B_3** dove la proporzione **p** di individui della specie **Y** è **0,7**

$$P(A / B_3) = \binom{10}{8} \cdot 0{,}7^8 \cdot (1-0{,}7)^2 = 0{,}23347$$

la probabilità **P** che su **10** individui estratti a caso **8** siano della specie **Y** è uguale a **0,23347**.

Le 3 probabilità a posteriori calcolate precedentemente (**0,20; 0,33; 0,47**) diventano le nuove probabilità a priori. Esse devono essere moltiplicate per le probabilità empiriche fornite dalla nuova osservazione:

$$P(B_1 / A) = \frac{0{,}00145 \cdot 0{,}20}{0{,}00145 \cdot 0{,}20 + 0{,}04394 \cdot 0{,}33 + 0{,}23347 \cdot 0{,}47} = 0{,}002$$

$$P(B_2 / A) = \frac{0{,}04394 \cdot 0{,}33}{0{,}00145 \cdot 0{,}20 + 0{,}04394 \cdot 0{,}33 + 0{,}23347 \cdot 0{,}47} = 0{,}116$$

$$P(B_3 / A) = \frac{0{,}23347 \cdot 0{,}47}{0{,}00145 \cdot 0{,}20 + 0{,}04394 \cdot 0{,}33 + 0{,}23347 \cdot 0{,}47} = 0{,}882$$

Sulla base delle ultime informazioni raccolte, la nuova risposta è:

- la probabilità che il campione **A** sia originario della comunità $\mathbf{B_1}$ diventa 2 su mille (**0,002**),
- la probabilità che provenga dalla comunità $\mathbf{B_2}$ diventa 116 per mille (**0,116**),
- la probabilità che sia stato raccolto nella comunità $\mathbf{B_3}$ è 882 su mille (**0,882**).

Non esistono altre possibilità e la loro somma deve essere uguale a 1 (0,002 + 0,116 + 0,882 = 1,0).

L'esempio illustra un aspetto peculiare del procedimento inferenziale bayesiano. Con l'aumento del numero di osservazioni campionarie od empiriche, le probabilità attribuite in modo indifferenziato, o almeno simile (come 1/3 all'inizio) tendono a divergere. Il fenomeno è chiamato **progressiva dominanza dell'informazione sulla distribuzione a priori**.

Un altro aspetto importante è che, partendo **da distribuzioni a priori notevolmente differenti** (per esempio, in partenza la distribuzione poteva essere 0,1 0,2 0,7), **se nessuna di esse è 0 oppure 1 con gli stessi dati sperimentali successivamente raccolti le probabilità tendono a convergere rapidamente.** E' chiamato il **principio della giuria:** se nessuno dei giudici è prevenuto, ritenendo l'accusato sicuramente colpevole o sicuramente innocente, dopo un numero relativamente basso di verifiche tutti giungono alle stessa probabilità, per quanto distanti siano in partenza.

L'analisi bayesiana può essere utilizzata per sommare in modo progressivo l'effetto di tutte le nuove informazioni. Ogni ulteriore indicazione quantitativa permette di aggiornare nuovamente la distribuzione originale precedente; i dati modificano le probabilità stimate inizialmente in condizioni di incertezza o dedotte da una informazione meno precisa. Le ampie possibilità operative di tale teorema derivano dal fatto che le probabilità possono essere tratte indifferentemente da dati oggettivi o da distribuzioni di opinioni, che mutano nel tempo.

Nell'esempio presentato, l'inferenza statistica bayesiana è stata applicata a percentuali o proporzioni. Ma essa può essere estesa a qualsiasi altra caratteristica della popolazione, come la media, la varianza, il coefficiente di regressione, per assegnare probabilità specifiche ad ogni possibile valore di questi parametri.

In questo corso, nei paragrafi successivi saranno illustrate le probabilità matematiche di alcune distribuzioni teoriche.

Il corso è fondato solo sulle probabilità frequentiste o a posteriori.

Per un'inferenza fondata sulle probabilità soggettive, utile a coloro che debbono prendere decisioni in situazioni d'incertezza, si rinvia a testi specifici.

2.3. ALCUNE DISTRIBUZIONI DISCRETE

Le variabili casuali hanno **distribuzioni di probabilità di due tipi: discrete o continue**. Negli esercizi precedenti, con il calcolo combinatorio, si sono stimate distribuzioni di probabilità discrete, che possono essere calcolate per un numero definito di casi. Nelle **variabili casuali discrete,** i valori argomentali sono i numeri naturali: 0, 1, 2, …, n. Servono per calcolare la probabilità di eventi che hanno un numero discreto di ricorrenze.

Una **variabile casuale è continua** quando la sua distribuzione è continua. Con tale variabile continua, si stima la probabilità di estrarre non un singolo valore ma valori ad esso uguali o maggiori (oppure minori). Una distribuzione continua non permette la stima della probabilità di estrarre un particolare valore, ma solo quelli compresi in un dato intervallo. Per esempio, nella distribuzione delle altezze di una popolazione di studenti, non è possibile stimare la probabilità di avere un individuo alto esattamente 176,000 cm ma quella di avere un individuo tra 180 e 190 centimetri.

2.3.1 DISTRIBUZIONE BINOMIALE

La **binomiale è una distribuzione teorica discreta e finita**, per eventi classificati con una **variabile binaria**. E' denominata anche **distribuzione di Bernoulli** o **distribuzione bernoulliana**, in onore del matematico svizzero Jacques Bernoulli (1654-1705), che ha fornito importanti contributi alla teoria della probabilità.

In un collettivo di **n** unità che possono essere ripartite solo in due classi A e B, con frequenze assolute $\mathbf{n_a}$ e $\mathbf{n_b}$, le cui frequenze relative sono **p** e **q** con

$$p = \frac{n_a}{n} \quad e \quad q = \frac{n_b}{n} \qquad \text{tali che} \quad p+q=1$$

la probabilità di avere *i* volte l'evento A (e quindi **n - i** volte l'evento alternativo B) è data da

$$P_i = C_n^i \, p^i \, q^{n-i}$$

ricordando, dalle combinazioni semplici, che

$$C_n^i = \frac{n!}{i!\,(n-i)!}$$

La distribuzione binomiale o bernoulliana fornisce le risposte al problema delle prove ripetute: stima le probabilità che un evento, con probabilità a priori o frequentista *p*, avvenga rispettivamente 0, 1, 2,...i,...n volte, nel corso di n prove identiche ed indipendenti. Le prove

possono essere successive oppure simultanee, purché siano tra loro indipendenti, non si influenzino reciprocamente e quindi le probabilità dei singoli eventi si mantengano costanti.

Le variabili casuali di tipo binario sono numerose: maschio/femmina, successo/insuccesso, malato/sano, inquinato/non inquinato, alto/basso, negativo/positivo. Inoltre tutte le variabili, sia le multinomiali sia quelle continue, possono sempre essere ridotte alla più semplice variabile dicotomica o binaria, seppure con perdita d'informazione. Per esempio, una popolazione classificata in individui di specie diverse (A, B, C, D, E, ...) può sempre essere ricondotta ad una classificazione binaria in specie A e specie non-A; una serie di misure con scala discreta o continua, non importa se di ranghi, d'intervalli o di rapporti, può sempre essere ricondotta ad una classificazione binaria di valori superiori (+) od inferiori (-) ad un limite prefissato.

ESEMPIO 1. E' statisticamente dimostrato, anche se non è stata ancora trovata una spiegazione esauriente, che in tutte le popolazioni umane nascono più maschi che femmine, con un rapporto di 105-106 maschi ogni cento femmine. Possiamo quindi stabilire, a posteriori e sulla base di queste analisi, che la probabilità frequentista della nascita di un maschio è approssimativamente p = 0,52 e che quella di una femmina è, di conseguenza, q = 0,48 (q = 1 - p).

Usando la distribuzione binomiale, possiamo calcolare le specifiche probabilità **P** di avere 0, 1, 2, 3, 4 figli maschi nelle famiglie con 4 figli:

$$P_0 = C_4^0 p^0 q^4 = 1 \cdot 1 \cdot (0{,}48)^4 = 0{,}05$$

$$P_1 = C_4^1 p^1 q^3 = 4 \cdot (0{,}52)^1 \cdot (0{,}48)^3 = 0{,}23$$

$$P_2 = C_4^2 p^2 q^2 = 6 \cdot (0{,}52)^2 \cdot (0{,}48)^2 = 0{,}37$$

$$P_3 = C_4^3 p^3 q^1 = 4 \cdot (0{,}52)^3 \cdot (0{,}48)^1 = 0{,}28$$

$$P_4 = C_4^4 p^4 q^0 = 1 \cdot (0{,}52)^4 \cdot 1 = 0{,}07$$

Se **gli eventi sono casuali ed indipendenti** e **le probabilità** di avere un maschio od una femmina **sono costanti**, in famiglie con 4 figli la probabilità

- P_0 di avere 0 figli maschi è 0,05
- P_1 di avere 1 figlio maschio è 0,23
- P_2 di avere 2 figli maschi è 0,37
- P_3 di avere 3 figli maschi è 0,28
- P_4 di avere 4 figli maschi è 0,07.

Non esistono altri eventi possibili oltre quelli calcolati; di conseguenza, il totale delle probabilità stimate deve necessariamente essere uguale a 1 (0,05 + 0,23 + 0,37 + 0,28 + 0,07 = 1,00); è prassi che gli arrotondamenti siano tali da dare sempre una somma uguale a 1,00.

La **rappresentazione grafica di queste probabilità**

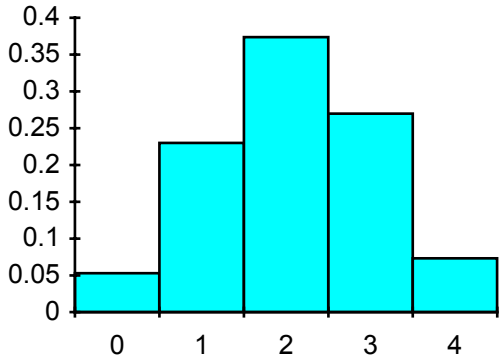


Figura 1. Probabilità del numero di maschi in famiglie con 4 figli.

mediante istogramma mostra con evidenza una distribuzione leggermente asimmetrica. La causa è il differente valore dei due eventi alternativi (p = 0,52; q = 0,48) e del numero basso di eventi (n = 4). Se le probabilità p e q fossero state uguali (ovviamente entrambe 0,5) la distribuzione sarebbe stata simmetrica; con p e q diversi, diventa simmetrica all'aumentare del numero di dati, come sarà di seguito dimostrato empiricamente.

ESEMPIO 2. Applicando la stessa legge, in eventuali famiglie con 10 figli le probabilità **P(i)** di avere **i** figli é

i	**P(i)**
0	0.000649
1	0.007034
2	0.034289
3	0.099056
4	0.187793
5	0.244131
6	0.220396
7	0.136436
8	0.055427
9	0.013344
10	0.001446

La sua rappresentazione grafica con un istogramma è:

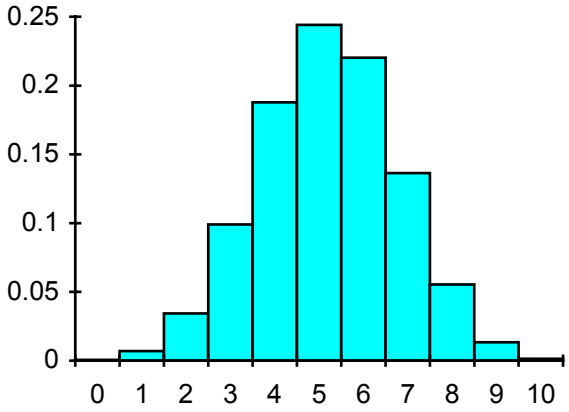


Figura 2. Probabilità del numero di maschi in famiglie con 10 figli

Essa evidenzia meglio della tabella la presenza di una leggera asimmetria.

La distribuzione di probabilità della binomiale dipende da 2 parametri: **p** e **n**.

Se **p** e **q** sono uguali a 0,5 la distribuzione è sempre simmetrica, indipendentemente da **n**. Se **p** è molto più grande o più piccolo di **q**, la distribuzione è asimmetrica, ma l'asimmetria tende a diminuire al crescere di **n**.

Di norma, si ricorre alla distribuzione binomiale per valori di **p** che variano da 0,1 a 0,9. Per valori di **p** esclusi da questo intervallo si ricorre alla distribuzione poissoniana, quando **n** non è grande. Quando **n** è così grande che anche **n·p** è grande, si ricorre comunque alla normale, per qualsiasi valore di **p**.

Quando un campione è di grandi dimensioni, la stima delle probabilità è ottenuta dalla distribuzione normale. A tal fine è intuitivo che in una distribuzione binomiale la media della popolazione se calcolata come **frequenza assoluta** è **n·p**

$$\mu = n \cdot p$$

mentre, se calcolata come **frequenza relativa,** è

$$\mu = p$$

Per esempio, in famiglie di 10 figli il numero medio di maschi è 10 x 0,52 = 5,2 in frequenza assoluta e 0,52 in frequenza relativa.

Senza ricorrere alla dimostrazione, che sarebbe lunga, è ugualmente importante ricordare che la varianza σ^2 della popolazione in **frequenza assoluta** è data dal prodotto di $n \cdot p \cdot q$

$$\sigma^2 = n \cdot p \cdot q$$

mentre in una **frequenza relativa** è

$$\sigma^2 = p \cdot q / n$$

I rapporti tra media e varianza offrono indicazioni importanti, quando dai dati sperimentali sia necessario risalire alla più probabile legge di distribuzione che li hanno determinati.

Nella distribuzione binomiale la varianza è inferiore alla media:

- con una media uguale a $n \cdot p$ e una varianza uguale a $n \cdot p \cdot q$,
- poiché $p + q$ è uguale a **1** ed il valore di **q** è inferiore a 1,
- il valore di $n \cdot p \cdot q$ è inferiore a $n \cdot p$.

ESERCIZIO 1. La distribuzione binomiale è utile anche nei casi in cui le probabilità sono note a priori, come nel lancio di dadi non truccati. Lanciando un dado 5 volte, quale è la probabilità di avere 3 volte il numero 1?

Risposta: ($n = 5$; $i = 3$; $p = 1/6$; $q = 5/6$)

$$P_3 = C_5^3 \, p^3 \, q^2 = \frac{5!}{3!2!} \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = 0,03215$$

ESERCIZIO 2. In un'urna contenente un numero elevatissimo, praticamente infinito, di biglie nel 70% nere e per il rimanente 30% bianche, quale è la probabilità di estrarre 4 biglie tutte nere?

Risposta: ($n = 4$; $i = 4$; $p = 0,7$; $q = 0,3$)

$$P_4 = C_4^4 \, p^4 \, q^0 = \frac{4!}{4!0!} \cdot 0,7^4 \cdot 0,3^0 = 0,2401$$

ESERCIZIO 3. Un esperimento di laboratorio di solito risulta positivo nel 20% dei casi. Con 10 tentativi quale è la probabilità che 9 risultino positivi e 1 solo negativo?

Risposta: ($n = 10$; $i = 9$; $p = 0,2$; $q = 0,8$)

$$P_9 = C_{10}^9 \, p^9 \, q^1 = \frac{10!}{9!1!} \cdot 0,2^9 \cdot 0,8^1 = 0,000004096$$

ESERCIZIO 4. In un lago, la specie A rappresenta il 33% degli individui presenti, la specie B e C entrambi il 25% e la specie D il 17%; estraendo a caso 15 individui, quale è la probabilità che

a) nessuno sia della specie A,

b) tutti siano della specie A

c) almeno 10 siano della specie A

d) meno di 7 siano della specie A

Risposte :

a) la probabilità che nessuno sia della specie A è 0,002461: vedi $P_{(0)}$ nella tabella sottostante;

b) la probabilità che tutti siano della specie A è minore di 1 su 1 milione di casi : vedi $P_{(15)}$

c) la probabilità complessiva che almeno 10 dei 15 individui estratti a caso siano della specie A è data dalla somma delle probabilità calcolate per $P_{(10)}$, $P_{(11)}$, $P_{(12)}$, $P_{(13)}$, $P_{(14)}$ e $P_{(15)}$;

d) la probabilità complessiva che meno di 7 individui siano della specie A è data dalla somma delle singole probabilità da $P_{(0)}$ a $P_{(6)}$ compresi.

i	**P(i)**
0	0.002461
1	0.018183
2	0.062689
3	0.133798
4	0.197702
5	0.214226
6	0.175858
7	0.111364
8	0.054851
9	0.021013
10	0.00621
11	0.00139
12	0.000228
13	0.000026
14	0.000002
15	0.000000

Tabella 2. Distribuzione binomiale con n = 15 e p = 0.33

L'istogramma delle probabilità da $P_{(0)}$ a $P_{(15)}$ mostra come la distribuzione sia già approssimativamente normale, benché la probabilità di ogni singolo evento si discosti da 1/2.

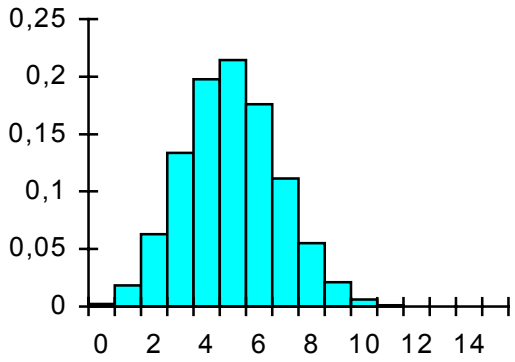


Figura 3. Istogramma della distribuzione binomiale con n = 15 e p = 0.33

ESERCIZIO 5. Per dimostrare come all'aumentare del numero di osservazioni la distribuzione diventi simmetrica e pertanto come essa possa essere bene approssimata da una distribuzione normale, chi è in grado di usare e programmare un calcolatore stimi tutte le probabilità possibili con p = 1/6 e n = 100, ovviamente con **i** che varia a 0 a 100.

Risposta. Le probabilità associate ai diversi tipi di estrazione possono essere espresse anche dai termini dello sviluppo del binomio $(p+q)^n$. La loro rappresentazione grafica, riportata nel grafico sottostante, evidenzia come la distribuzione abbia forma molto simile alla normale, causa dell'alto valore di **n** benché **p** sia lontano da 0,5.

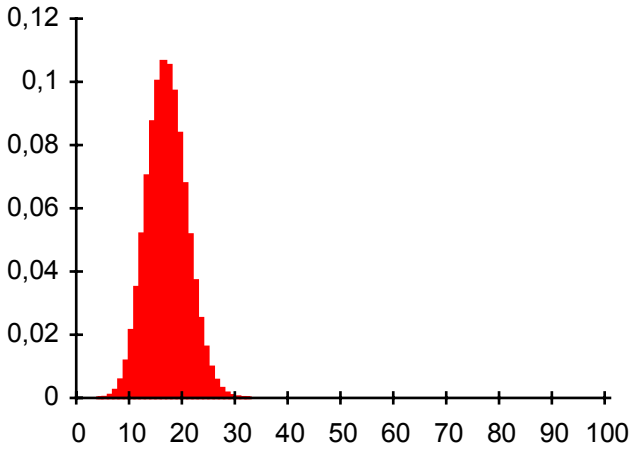


Figura 4. Istogramma della distribuzione binomiale con n = 100 e p = 1/6

2.3.2 DISTRIBUZIONE MULTINOMIALE

La distribuzione multinomiale rappresenta una estensione di quella binomiale; si applica a k eventi indipendenti di probabilità p_1, p_2, ...p_i, ..., p_k (la cui somma è uguale a 1) che possono comparire nel corso di N prove indipendenti, successive o simultanee.

Permette di calcolare la probabilità di ogni evento possibile, quando determinato solo dal caso.

La probabilità che si realizzino congiuntamente tutti gli eventi indicati è determinata dallo sviluppo del multinomio:

$$P_{(n_1 \quad n_2 \quad \dots \quad n_k)} = \frac{N!}{n_1!n_2!\dots.n_k!} p_1^{n_1} p_2^{n_2} \dots.. p_k^{n_k}$$

ESEMPIO. Si supponga che in un lago, con un numero teoricamente infinito di pesci, il 10% (p_1 = 0,10) siano della specie A, il 40% (p_2 = 0,40) siano della specie B, il 20% (p_3 = 0,20) siano di quella C ed il 30% (p_4 = 0,30) della specie D;

a) estraendo 10 pesci, quale è la probabilità di avere 2 A, 3 B, 2 C e 3 D?

b) estraendo 8 pesci quale è la probabilità di avere 4 B e 4 D? (naturalmente con 0 A e 0 C).

Risposte:

a) La probabilità di estrarre 2 individui della specie A, 3 della specie B, 2 della specie C e 3 della specie D, in un lago in cui le quattro specie hanno le frequenze relative della domanda, è calcolata con

$$P_{(2A,3B,2C,3D)} = \frac{10!}{2!3!2!3!} \cdot (0,10)^2 \cdot (0,40)^3 \cdot (0,20)^2 \cdot (0,30)^3$$

$$P_{(2A,3B,2C,3D)} = 25200 \cdot 0,01 \cdot 0,064 \cdot 0,04 \cdot 0,027 = 0,0174$$

e risulta uguale a 0,0174 o 1,74 per cento.

b) La probabilità di estrarre per caso, dallo stesso lago, 4 individui della specie B e 4 della specie D con 8 estrazioni casuali è

$$P_{(0A,4B,0C,4D)} = \frac{8!}{0!4!0!4!} \cdot (0,10)^0 \cdot (0,40)^4 \cdot (0,20)^0 \cdot (0,30)^4$$

$$P_{(0A,4B,0C,4D)} = 70 \cdot 1 \cdot 0,0256 \cdot 1 \cdot 0,0081 = 0,0145$$

uguale a 0,0145 o 1,45 per cento.

Non esiste un limite al numero di fattori che possono essere considerati insieme.

La distribuzione binomiale è utilizzata con frequenza nella statistica non parametrica. Se il numero di dati è ridotto, molti test non parametrici ricorrono ad essa per il calcolo delle probabilità.

Seppure apparentemente più generale, la distribuzione multinomiale ha un uso limitato, circoscritto alla stima della probabilità complessiva di più eventi indipendenti, per ognuno dei quali sia nota la probabilità p_i.

2.3.3 DISTRIBUZIONE POISSONIANA

Quando il numero di dati (**n)** è molto grande e la probabilità (**p)** è molto piccola, la distribuzione binomiale presenta vari inconvenienti pratici, che erano importanti soprattutto prima dell'introduzione del calcolo automatico. Infatti, essa richiede sia l'innalzamento di probabilità (**p**) molto basse a potenze (**i**) elevate, sia il calcolo di fattoriali per numeri (**n**) grandi, che sono operazioni che rendono il calcolo manuale praticamente impossibile.

Per

- **n** che tende all'**infinito,**
- **p** che tende a **0,**
- in modo tale che **n·p** sia **costante**,

il matematico francese Siméon Dennis **Poisson** (1781-1840), già autore di articoli sulla meccanica celeste e sull'elettromagnetismo, nel 1837 entro la sua opera maggiore ***Recherches sur la probabilité des jugements en matière criminelle et en matière civile*** ha dimostrato che la probabilità dell'evento ($\mathbf{P_i}$) è stimata da

$$P_i = \frac{\mu^i}{i!} e^{-\mu}$$

con **e = 2,71828**.

In modo più semplice per i calcoli, la stessa formula può essere scritta come

$$P_i = \frac{\mu^i}{i!} \cdot \frac{1}{2{,}71828^{\mu}}$$

La poissoniana è una **distribuzione** teorica **discreta**, totalmente **definita** da un solo **parametro: la media μ,** quando riferita a una popolazione, che deve **essere costante**. Quando la distribuzione è applicata a un campione, la media **μ** è sostituita da quella **campionaria** $\overline{X}$.

Anche nella distribuzione poissoniana, la media attesa μ è data dal prodotto **n·p**, con (p + q) = 1. Poiché, come in tutte le distribuzioni che possono essere fatte derivare dalla binomiale, **σ^2= n·p·q,** è facile dimostrare come la varianza sia uguale alla media (**σ^2 = μ**). Applicando le tre condizioni appena enunciate,

- **n** che tende all'**infinito**,
- **p** che tende a **0**,
- **p + q = 1**

la varianza della distribuzione di Poisson è

$$\sigma^2 = \lim_{\substack{n\to\infty \\ p\to 0}} npq = \lim_{\substack{n\to\infty \\ p\to 0}} (np)q = \lim_{\substack{n\to\infty \\ p\to 0}} \mu(1-p) = \mu$$

In termini discorsivi,

con un numero **infinito di dati,** se **p** tende a **0** e quindi **q** tende a **1, la varianza è uguale alla media**

n·p·q (σ^2) = n·p (μ).

E' un concetto importante quando si deve individuare la forma reale di una distribuzione campionaria.

La legge di distribuzione poissoniana è detta anche **legge degli eventi rari**, poiché la probabilità (**p**) che l'evento si verifichi per ogni caso e la media (**μ**) degli eventi su tutta la popolazione sono basse. E' chiamata pure **legge dei grandi numeri**, in quanto tale distribuzione è valida quando il numero (**n**) di casi considerati è alto.

Nella pratica della ricerca, la distribuzione poissoniana sostituisce quella binomiale quando **p** < 0,05 e **n** > 100.

La distribuzione poissoniana è utilizzata per eventi che si manifestano sia nello **spazio** che nel **tempo**. E' il caso del numero di insetti presenti in una superficie di piccole dimensioni o del numero di eventi che possono avvenire in un breve intervallo di tempo. In molti testi, il numero medio di eventi è indicato non con **μ** ma con **λ**, in particolare quando si tratta di eventi temporali.

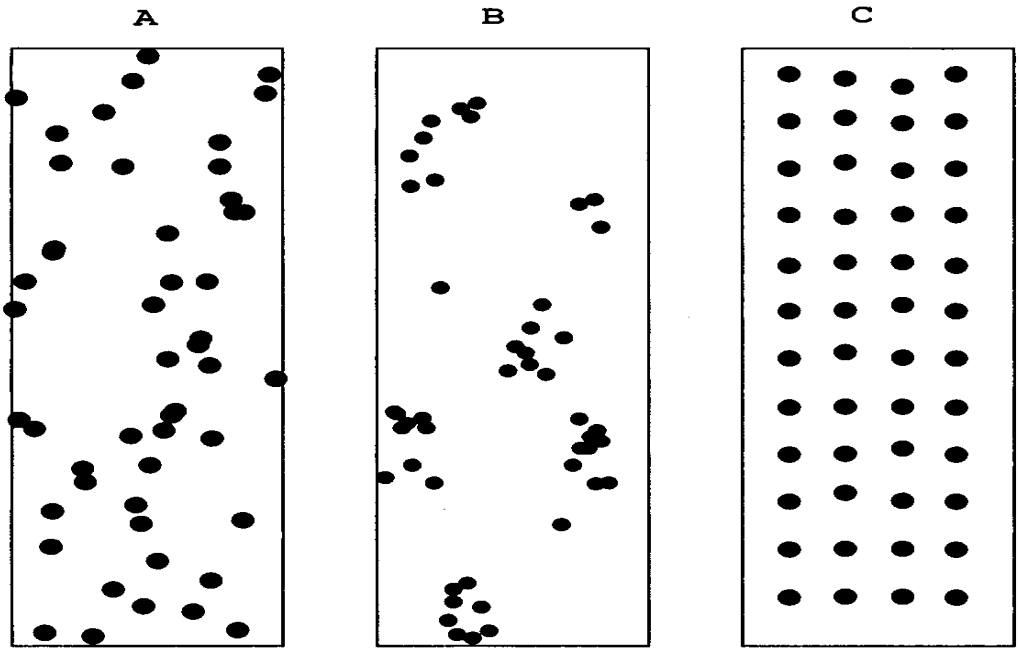


Figura 5. Tipi semplici di distribuzione spaziale degli individui di una popolazione

Considerando lo spazio, più facile da rappresentare e caso più frequente nella ricerca biologica e ambientale, una distribuzione di oggetti segue la **legge poissoniana** quando è **casuale** (in inglese ***random)***.

Le figura precedente illustra i tre tipi fondamentali di distribuzione di una specie o di una comunità su un territorio.

In modo schematico, può avare tre tipi principali di aggregazione:

A - distribuzione **casuale** (***random***),

B - distribuzione **aggregata** (***aggregated***) o **a gruppi,**

C - distribuzione **uniforme** (***uniform***) o **regolare**.

Dalla loro diversa combinazione deriva un numero infinito di possibilità:

i	P_i
0	0.000006
1	0.000074
2	0.000442
3	0.001770
4	0.005309
5	0.012741
6	0.025481
7	0.043682
8	0.065523
9	0.087364
10	0.104837
11	0.114368
12	0.114368
13	0.105570
14	0.090489
15	0.072391
16	0.054293
17	0.038325
18	0.025550
19	0.016137
20	0.009682
21	0.005533
22	0.003018
23	0.001574
24	0.000787
25	0.000378

Tabella 5. Distribuzione di Poisson con $\mu = 12$, per i che varia da 0 a 25

La distribuzione poissoniana ha una forma molto **asimmetrica**, quando la media è piccola.
Quando **μ** < 1, la classe più frequente o più probabile è zero.
E' ancora asimmetrica per valori di **μ** < 3. Ma già con **μ ≥ 5-6** la distribuzione delle probabilità è vicina alla forma simmetrica e può essere bene approssimata dalla distribuzione normale o gaussiana.

Le probabilità di una distribuzione poissoniana con μ = 12 è quasi perfettamente simmetrica, come mostra la tabella precedente e evidenzia ancor meglio il grafico successivo di distribuzione delle probabilità P_i (sull'asse delle ordinate), per i diversi valori di i (sull'asse delle ascisse)

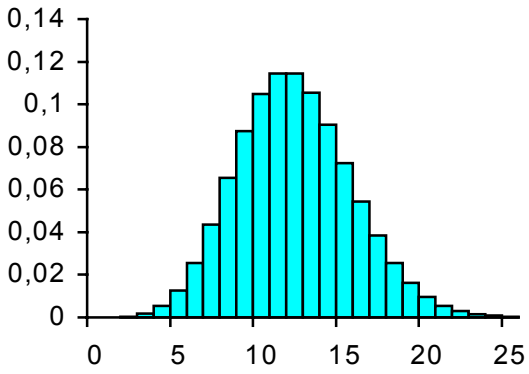


Figura 6. Distribuzione di Poisson con $\mu = 12$.

In termini tecnici, si dice che per avere una distribuzione poissoniana, una **variabile casuale** deve avere **tre requisiti**: **stazionarietà, non-multiplicità, indipendenza**.

Si ha

1 - **omogeneità** o **stazionarietà** quando la media **μ** (o la probabilità **p**) dell'evento è costante, su tutto il periodo considerato o l'area in osservazione; in altri termini, quando la **probabilità** di ogni evento in un intervallo di tempo (**t, t + h**) o in uno spazio infinitesimo è costante, pari a **λh** per ogni **t**;

2 - individualità o **non-multiplicità** quando gli eventi avvengono singolarmente, non a coppie o a gruppi; in tali condizioni, la probabilità che due o più eventi avvengano nello stesso intervallo infinitesimo di tempo o di spazio non è **λh** volte minore di quello di un solo evento;

3 - **indipendenza** quando il presentarsi di un evento, in una particolare unità di tempo o di spazio, non influenza la probabilità che l'evento si ripresenti in un altro istante o luogo.

Tutti questi concetti, evidenziati nelle figure, sono sinteticamente compresi nell'espressione che **i fenomeni non devono essere né contagiosi**, **né regolari**.

ESEMPIO 1 (PER VALUTARE SE UNA DISTRIBUZIONE E' POISSONIANA, tratto con modifiche dal testo di Charles J. **Krebs** del 1999, **Ecological Methodology**, 2nd ed. Benjamin/Cummings, Addison Wesley Longman, Menlo Park, California, p. X + 620).

Una superficie, come una delle tre figure precedenti, è stata suddivisa in 25 quadrati delle stesse dimensioni e in ognuno è stato contato il numero di organismi della specie X con il seguente risultato

3	4	1	1	3	0	0	1	2	3	4	5	0	1	3	5	5	2	6	3	1	1	1	0	1

E' una distribuzione casuale?. In altri termini, il numero di organismi in ogni quadrato segue la legge di Poisson?

Risposta. Come prima parte dell'analisi, dalla serie dei 25 dati

- costruire la **distribuzione di frequenza**,
- calcolare la **media**,
- calcolare la **varianza**.

Osservando che il valore minimo è 0 e il valore massimo è 6,

- la **distribuzione di frequenza** è

Classe	X_i	0	1	2	3	4	5	6
Frequenza assoluta	n_i	4	8	2	5	2	3	1

Da essa si ricavano

- la **media** dei 25 valori

$$\overline{X} = \frac{\sum_{i=1}^{k} X_i \cdot n_i}{\sum_{i=1}^{k} n_i} = \frac{(0x4)+(1x8)+(2x2)+(3x5)+(4x2)+(5x3)+(6x1)}{4+8+2+5+2+3+1} = \frac{56}{25} = 2{,}24$$

che risulta $\overline{X}$ = 2,24 e

- successivamente **la devianza**

$$SQ = \sum_{i=1}^{k} n_i \left(X_i - \overline{X}\right)^2$$

$$SQ = 4(0-2{,}24)^2 + 8(1-2{,}24)^2 + 2(2-2{,}24)^2 + 5(3-2{,}24)^2 + 2(4-2{,}24)^2 +$$
$$+ 3(5-2{,}24)^2 + 1(6-2{,}24)^2 = 78{,}56$$

che risulta SQ = 78,56

e infine **la varianza**

$$s^2 = \frac{SQ}{n-1} = \frac{78{,}56}{24} = 3{,}273$$

che risulta $s^2 = 3{,}273$.

Risulta immediatamente che la varianza (3,273) è maggiore della media (2,24).

In tale situazione ($s^2 > \overline{X}$), la prima indicazione è che la distribuzione non è perfettamente poissoniana, quindi non è casuale, ma è tendenzialmente aggregata: gli individui tendono a distribuirsi sul terreno a gruppi.
Se la varianza fosse stata inferiore alla media, la distribuzione sarebbe stata ugualmente non casuale, ma perché troppo regolare: la popolazione sarebbe stata distribuita in modo tendenzialmente uniforme.

Tuttavia il campione è piccolo. Ne deriva che la differenza rilevata tra media e varianza può essere dovuta non a un fenomeno reale, ma alle variazioni casuali, sempre sono sempre molto importanti quando il numero di osservazioni è limitato, come in questo caso.

Per un valutazione più approfondita, dalla sola conoscenza della media è utile ricavare tutta la distribuzione teorica del numero di individui per quadrato.
Mediante la formula della distribuzione poissoniana

$$P_i = \frac{\overline{X}^i}{i!} \cdot e^{-\overline{X}} = \frac{2{,}24^i}{i!} \cdot \frac{1}{2{,}71828^{2{,}24}}$$

con $\overline{X} = 2{,}24$ e i che varia da 0 a 6
si ricava la seguente serie di probabilità:

Evento	i	0	1	2	3	4	5	6	≥ 7
Probabilità	P_i	0,1065	0,2385	0,2671	0,1994	0,1117	0,0500	0,0187	0,0081

In modo più dettagliato, da $i = 0$ a $i = 6$ le singole probabilità sono calcolate direttamente con

$$P_0 = \frac{2{,}24^0}{0!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{1}{1} \cdot \frac{1}{9{,}393} = 0{,}1065$$

$$P_1 = \frac{2{,}24^1}{1!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{2{,}24}{1} \cdot \frac{1}{9{,}393} = 0{,}2385$$

$$P_2 = \frac{2{,}24^2}{2!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{5{,}0176}{2} \cdot \frac{1}{9{,}393} = 0{,}2671$$

$$P_3 = \frac{2{,}24^3}{3!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{11{,}2394}{6} \cdot \frac{1}{9{,}393} = 0{,}1994$$

$$P_4 = \frac{2{,}24^4}{4!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{25{,}1763}{24} \cdot \frac{1}{9{,}393} = 0{,}1117$$

$$P_5 = \frac{2{,}24^5}{5!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{56{,}3949}{120} \cdot \frac{1}{9{,}393} = 0{,}0500$$

$$P_6 = \frac{2{,}24^6}{6!} \cdot \frac{1}{2{,}71828^{2{,}24}} = \frac{126{,}3267}{720} \cdot \frac{1}{9{,}393} = 0{,}0187$$

L'ultima, indicata con $i \geq 7$ e uguale a 0,0081, è ricavata per differenza da 1,00 della somma di tutte le probabilità P_i precedenti (0,9919).

$$P_0 + P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = P_{\leq 6}$$
$$0{,}1065 + 0{,}2385 + 0{,}2671 + 0{,}1994 + 0{,}1117 + 0{,}0500 + 0{,}0187 = 0{,}9919$$

Infine,

- si ottiene la frequenza attesa per ogni i (riportata nella terza riga della tabella sottostante),
- moltiplicando la probabilità stimata P_i
- per il numero totale n di rilevazioni.

In questo caso, il numero totale di individui è $n = 25$.

Individui per quadrato	0	1	2	3	4	5	6	≥ 7	Totale
Frequenza **osservata**	4	8	2	5	2	3	1	0	25
Frequenza **attesa**	2,66	5,96	6,68	4,99	2,79	1,25	0,47	0,20	25,00

Osservando la tabella, la prima domanda che un ricercatore si deve porre è: "La distribuzione osservata (seconda riga) è in accordo con quella attesa? (terza riga)"
Per rispondere in modo scientifico, senza limitarsi alla semplice impressione, occorre applicare un test. Si dovrebbe quindi passare all'inferenza, i cui metodi sono presentati nei capitoli successivi.

ESEMPIO 2. Tassi elevati di inquinamento atmosferico possono determinare reazioni allergiche gravi. Durante un mese (30 giorni), nel pronto soccorso di un ospedale si sono avuti 27 ricoveri urgenti. Può essere ragionevole ipotizzare che gli eventi abbiano una distribuzione giornaliera costante, in accordo con la legge di Poisson.

Calcolare la probabilità (P_i) di avere i casi di allergia al giorno, per i che varia da 0 a 8.

Risposta. Dopo aver calcolato la media giornaliera 0,9 (27/30),
si applica la formula

$$P_{\mathrm{i}} = \frac{\mu^{\mathrm{i}}}{\mathrm{i!}} \mathrm{e}^{-\mu} = \frac{0{,}9^{i}}{i!} \cdot \frac{1}{2{,}71828^{0{,}9}}$$

e si ottengono i risultati riportati nella tabella

i	P_i
0	0.40657
1	0.365913
2	0.164661
3	0.049398
4	0.011115
5	0.002001
6	0.000300
7	0.000039
8	0.000004

Tabella 3. Distribuzione di Poisson con $\mu = 0.9$.

Se la distribuzione fosse esattamente casuale, nel 40% dei giorni non si dovrebbe avere nessun caso; si dovrebbe avere 1 solo ricovero nel 36,6% e 2 ricoveri nel 16,5% dei giorni.

La rappresentazione grafica

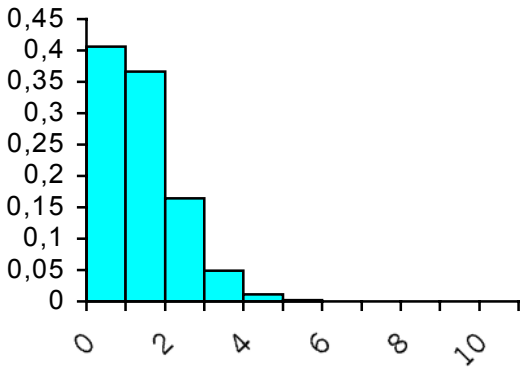


Figura 7. Distribuzione di Poisson con $\mu = 0{,}9$.

evidenzia come la distribuzione delle probabilità con μ = 0,9 sia fortemente asimmetrica, con asimmetria destra.

ESEMPIO 3. E' stato ipotizzato che gli individui della specie A abbiano una distribuzione poissoniana sul terreno. Dividendo l'area campionata in appezzamenti della stessa dimensione, per superficie unitaria si è ottenuto $\overline{X} = 2{,}0$.
Calcolare la frequenza attesa (P_i), per i che va da 0 a 5. in una distribuzione poissoniana.

Risposta: Il calcolo delle frequenze relative è

i	P_i
0	0.135335
1	0.270671
2	0.270671
3	0.180447
4	0.090224
5	0.036089

Tabella 4. Distribuzione di Poisson con $\mu = 2$

Come mostra il grafico successivo, la forma della distribuzione poissoniana con media uguale a 2 è ancora asimmetrica, seppure in modo molto meno accentuato della distribuzione precedente, che aveva una media inferiore. L'asimmetria è sempre destra o positiva, fino a quando la distribuzione diviene normale e simmetrica.

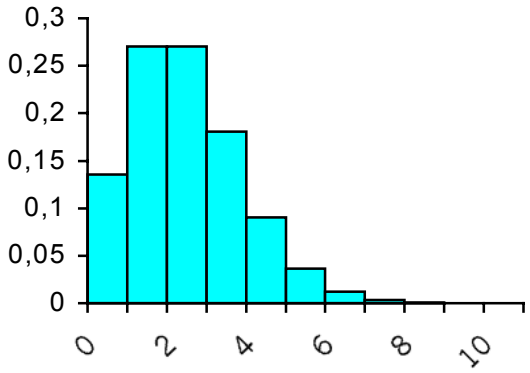


Figura 8. Distribuzione di Poisson, $\mu = 2$

ESEMPIO 4. In letteratura come esempio storico della distribuzione poissoniana, tratta da dati sperimentali, è famoso l'esempio di Ladislaus von **Bortkiewicz** (1868 – 1931, a volte scritto all'inglese come **Bortkiewitch)**. Prendendo i dati dell'armata prussiana del XIX secolo, per 20 anni ha contato in 10 corpi d'armata il numero di soldati che ogni anno morivano a causa di un calcio di mulo; ha quindi classificato i decessi nei 200 eventi (20 corpi d'armata per 10 anni), ottenendo la tabella sottostante:

numero di decessi **i**	0	1	2	3	4
Eventi osservati **r**	109	65	22	3	1

Tabella 6. Morti per corpo d'armata e per anno, rilevati da Bortkewicz.

Come riportato in essa,

- in 109 casi non si è avuto nessun morto,
- in 65 casi si è avuto 1 morto,
- in 22 casi sono stati contati 2 morti,
- in 3 casi 3 morti

- e in 1 caso 4 morti.

In totale, nei 200

$$109 + 65 + 22 + 3 + 1 = 200$$

casi esaminati il numero di morti è stato di 122

$$109 \text{ x } 0 + 65 \text{ x } 1 + 22 \text{ x } 2 + 3 \text{ x } 3 + 1 \text{ x } 4 = 122$$

Il calcolo della media e della varianza, secondo le formule presentate nel primo capitolo sulle distribuzioni di frequenza, fornisce i seguenti risultati:

media = μ = 122/200 = **0,6100**

varianza = σ^2 = **0,6079**

E' importante osservare che la varianza di questa distribuzione sperimentale è quasi identica alla sua media, come atteso in una distribuzione poissoniana teorica. E' una buona indicazione che questi eventi seguono la legge di Poisson; ma la dimostrazione più completa è fornita dal confronto tra la distribuzione attesa e quella osservata.

Applicando la distribuzione di Poisson, si determinano le probabilità teoriche di avere ogni anno per corpo d'armata **0** morti, **1** morto,..., **n** morti, eseguendo i calcoli sottostanti (approssimati rispetto alle possibilità attuali di calcolo)

$$P_0 = \frac{0,61^0}{0!} \cdot \frac{1}{2,71^{0,61}} = \frac{1}{1} \cdot \frac{1}{1,837} = 0,5440$$

$$P_1 = \frac{0,61^1}{1!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,61}{1} \cdot \frac{1}{1,837} = 0,3318$$

$$P_2 = \frac{0,61^2}{2!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,3721}{2} \cdot \frac{1}{1,837} = 0,1010$$

$$P_3 = \frac{0,61^3}{3!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,2270}{6} \cdot \frac{1}{1,837} = 0,0203$$

$$P_4 = \frac{0,61^4}{4!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,1385}{24} \cdot \frac{1}{1,837} = 0,0029$$

Le probabilità stimate sono riferite ad ogni corpo d'armata in un anno. Si ottengono i relativi eventi attesi rapportandole a 200 casi (20 corpi d'armata per 10 anni).

Numero di decessi	0	1	2	3	4
Eventi osservati	**109**	**65**	**22**	**3**	**1**
Frequenze relative attese	0,5440	0,3318	0,1010	0,0203	0,0029
Eventi attesi (su 200)	**108,80**	**66,36**	**20,20**	**4,06**	**0,58**

Tabella 7. Eventi osservati ed eventi attesi per corpo d'armata e per anno.

Il problema reale del confronto con un test tra le frequenze attese e quelle osservate consiste nel capire se le differenze sono di entità trascurabile, puramente imputabili al caso, oppure se sono di entità tale da lasciare presupporre l'intervento di leggi o fattori diversi da quelli ipotizzati nella distribuzione teorica.

2.3.4 DISTRIBUZIONE GEOMETRICA E DISTRIBUZIONE DI PASCAL

La **distribuzione binomiale** serve per valutare, entro un numero complessivo **n** di eventi, la probabilità **P** di avere **i** eventi favorevoli (P_i), ognuno dei quali ha una probabilità **p** costante. Con **n** prefissato (esempio: famiglie di 4 figli) e **p** costante (esempio: la probabilità p = 0.52 di avere un figlio maschio), permette di stimare la probabilità di avere **i** volte l'evento atteso (esempio: avere 0 oppure 1 oppure 2, 3, 4 figli maschi). Nel calcolo, **n e p sono costanti,** mentre **i varia.**

La **distribuzione geometrica** serve per valutare la probabilità di avere **un** evento favorevole (**i = 1**). Con **i** prefissato (uguale a 1) e **p** costante, permette di stimare la probabilità ($\mathbf{P_n}$) di avere l'evento desiderato all'aumentare del numero **n** di osservazioni. Nel calcolo della distribuzione geometrica, **i** e **p** sono **costanti** mentre **n varia**.

Ad esempio, si supponga che, in un'analisi complessa di laboratorio, un esperimento abbia il 30% di probabilità di dare una risposta positiva. Quante prove occorre fare per avere la probabilità **P** di avere una risposta positiva?

La probabilità che il **primo tentativo** sia **positivo** è **p**. Scritto in modo formale:

$$P(N = 1; p) = p$$

La probabilità che il **secondo tentativo** dia un risultato **positivo** è (1- p) x p:

$$P(N = 2; p) = (1 - p) \cdot p$$

La probabilità che sia il **terzo** a dare il risultato **positivo** è

$$P(N = 3; p) = (1 - p)^2 \cdot p$$

e **la probabilità che sia l'n-esimo è**

$$\mathbf{P(N = n; p) = (1 - p)^{n-1} \cdot p}$$

ESEMPIO. Con p = 0,3 calcolare la probabilità P di avere l'evento desiderato alle prove da 1 a n.

Risposta. Sviluppando

$$\mathbf{P(n) = (1 - p)^{n-1} \cdot p}$$

si ottengono le probabilità riportate nella tabella

Ad esempio, con n = 3 la probabilità è

$$P_3 = (1 - 0{,}3)^2 \cdot 0{,}3 = 0{,}147$$

n	P(n) con p = 0,3		Σ P(n)
1	0,3	0,3	0,300000
2	$0{,}7 \cdot 0{,}3$	0,21	0,510000
3	$0{,}7^2 \cdot 0{,}3$	0,147	0,657000
4	$0{,}7^3 \cdot 0{,}3$	0,1029	0,752900
5	$0{,}7^4 \cdot 0{,}3$	0,07203	0,83193
6	$0{,}7^5 \cdot 0{,}3$	0,050421	0,882351
7	$0{,}7^6 \cdot 0{,}3$	0,035295	0,917646
8	$0{,}7^7 \cdot 0{,}3$	0,024706	0,942352
9	$0{,}7^8 \cdot 0{,}3$	0,017294	0,959646
10	$0{,}7^9 \cdot 0{,}3$	0,012106	0,971752

Ottenuti mediante la formula sopra riportata, i valori della tabella sono le **probabilità** (i dati del calcolo nella colonna 2, i risultati nella colonna 3) che l'esperimento riesca al primo tentativo, al secondo e così via. Sono **probabilità esatte**: forniscono la **probabilità che ogni singolo tentativo dia la risposta desiderata**.

Se è necessario sapere **quanti tentativi occorre fare**, per avere una determinata probabilità che entro essi sia presente **almeno una risposta positiva**, si deve utilizzare la sommatoria dei singoli valori

esatti. L'elenco della tabella mostra che per avere una probabilità non inferiore al 90% di avere il primo risultato positivo occorre fare 7 tentativi, mentre per una probabilità non inferiore al 95% occorre fare 9 tentativi, sempre che la probabilità sia costante e gli eventi siano indipendenti.

La **distribuzione di Pascal** (Blaise Pascal matematico e filosofo francese, nato nel 1623 e morto nel 1662) **è la generalizzazione della distribuzione geometrica:** supponendo una **probabilità p costante** per ogni evento atteso, serve per valutare **la probabilità che entro n osservazioni sequenziali sia compreso i volte l'evento atteso** (con **i** che può essere uguale a 1, 2, 3, …, n) .
La precedente **distribuzione geometrica è un caso della distribuzione di Pascal, con i = 1**.

Se ogni evento ha una probabilità **p** costante ed indipendente, la probabilità che esso avvenga **i** volte all'aumentare del numero **n** di osservazioni è

$$P(N = n; i; p) = \frac{(n-1)!}{(i-1)!\cdot(n-i)!} \cdot p^{i} \cdot (1-p)^{n-i}$$

Dove, ovviamente, **n ≥ i** (la cui interpretazione è: per avere 3 risultati simultaneamente positivi, occorre fare almeno 3 prove).

ESEMPIO. Si supponga che, in una popolazione animale allo stato selvatico, il 40% degli individui (p = 0,4) sia colpito da una malattia; per una serie di analisi più approfondite e proporre la cura servano 3 animali ammalati. Quanti occorre catturarne, **con un campionamento sequenziale**, per avere i 3 individui ammalati ad una probabilità prefissata?

Risposta: Con i dati dell'esempio, **i** = 3 e **p** = 0,4
la formula generale diventa

$$P(N = n;\ \ 3;\ \ 0{,}4) = \frac{(n-1)!}{(3-1)!\cdot(n-3)!} \cdot 0{,}4^{3} \cdot 0{,}6^{n-3}$$

con **n** che aumenta progressivamente a partire da 3.

Con n = 7, il calcolo della probabilità di trovare tra essi 3 ammalati è

$$\frac{6!}{2!4!} \cdot 0{,}4^{3} \cdot 0{,}6^{4} = 15 \cdot 0{,}064 \cdot 0{,}1296 = 0{,}124416$$

uguale a 0,124416 o 12,44%
Nella tabella è stato riportato il calcolo delle probabilità con n che varia da 3 a 12:

n	P(n) con i = 3 e p = 0,4	Σ P(n)
3	0,064	0,064000
4	0,1152	0,179200
5	0,13824	0,317440
6	0,13824	0,455680
7	0,124416	0,580096
8	0,104509	0,684605
9	0,083607	0,768212
10	0,064497	0,832709
11	0,048373	0,881082
12	0,035473	0,916555

Se si catturano 3 animali, la probabilità che tutti e tre siano ammalati è uguale a 6,4 per cento.
Catturando un quarto animale, la probabilità che esso possa essere il terzo ammalato è di 11,52% e la probabilità complessiva di avere almeno i tre ammalati che servono diventa 17,92 %.
Infine, con il dodicesimo animale catturato, la probabilità di averne entro essi almeno tre ammalati è superiore al 90%, esattamente uguale a 91,6555 per cento.

L'esempio serve anche per risolvere altri tipi di problemi, quale la frequenza reale di ammalati nella popolazione da cui si estrae il campione.
Se, con un numero elevato di animali catturati (esempio 30), non è stato possibile selezionare i 3 individui ammalati, deve sorgere il sospetto che la percentuale di ammalati in realtà sia nettamente inferiore al 40 per cento stimato. Per esempio, si supponga che con 15 animali la probabilità cumulata superi il 95%; se con 15 animali catturati non si fosse ancora raggiunto il numero di 3 ammalati, si potrebbe ragionevolmente pensare che la frequenza **p** di ammalati sia minore di 0,4. Tale affermazione avrebbe una probabilità di oltre il 95% di essere vera e quindi inferiore al 5% di essere falsa.

Se con la distribuzione di **Pascal** si conta il numero di insuccessi avuti prima di arrivare al numero desiderato di casi postivi, si ottiene una distribuzione simile alla **distribuzione binomiale negativa** (che sarà presentata in un paragrafo successivo).

2.3.5 DISTRIBUZIONE IPERGEOMETRICA

Nella **distribuzione binomiale**, presa come riferimento per tutte le distribuzioni teoriche presentate, la probabilità **p** di un evento è costante. Quando la probabilità **p** di una estrazione casuale varia in funzione degli eventi precedenti, come succede in una **popolazione limitata** e di **piccole dimensioni**, si ha la **distribuzione ipergeometrica**.

Un modo semplice per chiarire la differenza tra distribuzione binomiale e distribuzione ipergeometrica è fornito dal gioco delle carte, con il calcolo delle probabilità nell'estrazione di un secondo re da un mazzo di 40 carte, in funzione delle regole stabilite.
Il gioco può avvenire in due modi: (A) **con reimmissione** o (B) **senza reimmissione** della carta estratta.

A- **Con reimmissione**: la probabilità di estrarre un re la prima volta è uguale a 4/40 o 1/10. Se la carta viene reintrodotta nel mazzo, la probabilità che la seconda carta sia un re rimane 1/10; in queste estrazioni, la probabilità di tutte le carte è sempre P = 0,1.

B- **Senza reimmissione**: la probabilità che la seconda carta sia un re dipende dalla prima estrazione.
a) Se la **prima** carta era un **re**, la probabilità che la seconda lo sia è di **3/39**, quindi **P = 0,077**;
b) se la **prima** carta **non** era un **re**, la probabilità che la seconda lo sia è di **4/39**, quindi **P = 0,103**
Per la seconda carta e quelle successive, la probabilità P varia in funzione di quanto è avvenuto nelle estrazioni precedenti.

Da questo esempio è facile comprendere che, se il mazzo di carte fosse molto grande (**n grande**), la probabilità **P** rimarrebbe approssimativamente costante. Ne deriva che quando il campione è grande, la distribuzione binomiale rappresenta una buona approssimazione della ipergeometrica, che è una distribuzione tipica delle popolazioni piccole.

Nella distribuzione ipergeometrica, la probabilità di un evento (**P**) dipende da vari parametri, che devono essere tenuti in considerazione nel rapporto tra combinazioni:

$$P_{(r/n)} = \frac{C_n^r \cdot C_{N-n}^{n_1-r}}{C_N^{n_1}}$$

dove:
- **N** = numero totale degli individui del campione (è un conteggio, quindi è un numero intero positivo);

- $\mathbf{n_1}$ = numero degli individui del campione che possiedono il carattere in oggetto (è un intero positivo, al massimo uguale a **N**);
- **n** = numero di individui estratti dal campione (è un numero intero non negativo, che al massimo può essere uguale a **N**);
- **r** = numero degli individui che presentano il carattere in oggetto tra quelli estratti (è un numero intero non negativo, che al massimo può essere uguale al minore tra **n** e $\mathbf{n_1}$).

La formula presentata può essere spiegata con un semplice ragionamento logico, fondato sul calcolo combinatorio.
Si supponga che un'urna contenga **N** biglie, delle quali $\mathbf{n_1}$ bianche e $\mathbf{N - n_1}$ nere; inoltre, si supponga che si estraggano dall'urna **n** biglie (con $n \leq N$) senza reintroduzione. Si vuole determinare la probabilità **P(r/n)** che delle **n** biglie estratte **r** siano bianche (con $r \leq n$).
Il calcolo delle varie probabilità richiede 4 passaggi logici:

1 - delle **N** biglie **n** possono essere estratte in C_N^n modi differenti,

2 - delle $\mathbf{n_1}$ biglie bianche **r** possono essere estratte in C_{n1}^r modi differenti,

3 - delle $\mathbf{N - n_1}$ biglie nere, **n - r** possono essere estratte in C_{N-n1}^{n-r} modi differenti,

4 - ognuna delle C_{n1}^r diverse possibilità di estrazione delle biglie bianche si combina con ognuna delle C_{N-n1}^{n-r} possibilità d'estrazione di biglie nere.

Da queste probabilità deriva la formula

$$P_{(r/n)} = \frac{C_n^r \cdot C_{N-n}^{n_1-r}}{C_N^{n_1}}$$

ESEMPIO 1. Per la cura di una malattia rara molto grave, presso una struttura ospedaliera sono stati ricoverati 12 pazienti, di cui 7 femmine e ovviamente 5 maschi. Dei 12 ammalati 6 sono guariti e 6 sono deceduti. L'osservazione dimostra che tra i 6 deceduti erano compresi tutti i maschi e una sola femmina.
E' statisticamente fondato il sospetto che tale cura possa essere idonea per le femmine, ma assolutamente dannosa per i maschi?

Risposta. Per impostare il calcolo combinatorio, è necessario attribuire i 4 gruppi in modo corretto:
1 – totale ammalati ricoverati: **N = 12**
2 – totale ricoverati morti: **n = 6**

3 – numero di femmine tra i ricoverati: $\mathbf{n_1 = 7}$

4 – numero di femmine tra i deceduti: $\mathbf{r} = 1$

Con la formula della distribuzione ipergeometrica

si ottiene

$$P_{(1/6)} = \frac{C_6^1 \cdot C_{12-6}^{7-1}}{C_{12}^7} = \frac{C_6^1 \cdot C_6^6}{C_{12}^7} = \frac{\frac{6!}{(6-1)!\cdot 1!} \cdot \frac{6!}{(6-6)!\cdot 6!}}{\frac{12!}{(12-7)!\cdot 7!}} = \frac{6 \cdot 1}{792} = 0{,}00757$$

che la probabilità che tra i 6 decessi su 12 ricoverati una sola fosse femmina per il solo effetto del caso è minore dell'otto su mille ($P_{1/6} = 0{,}00757$).

Lo stesso calcolo poteva essere impostato sui maschi. Quale è la probabilità che tra i sei deceduti fossero compresi i 5 maschi per solo effetto del caso?

I dati sarebbero diventati:

1 – totale ammalati ricoverati: $\mathbf{N = 12}$

2 – totale ricoverati morti: $\mathbf{n = 6}$

3 – numero di maschi tra i ricoverati: $\mathbf{n_1 = 5}$

4 – numero di maschi tra i deceduti: $\mathbf{r} = 5$

e la stima della probabilità

$$P_{(5/6)} = \frac{C_6^5 \cdot C_{12-6}^{5-5}}{C_{12}^5} = \frac{C_6^5 \cdot C_6^0}{C_{12}^5} = \frac{\frac{6!}{(6-5)!\cdot 5!} \cdot \frac{6!}{(6-0)!\cdot 0!}}{\frac{12!}{(12-5)!\cdot 5!}} = \frac{6 \cdot 1}{792} = 0{,}00757$$

avrebbe ovviamente dato un risultato identico ($P_{5/6} = 0{,}00757$), trattandosi dello stesso caso.

ESEMPIO 2. Con gli stessi dati dell'esempio 1, quale la probabilità se tra i decessi ci fossero stati 4 maschi? Inoltre stimare le probabilità per tutte le 6 possibili risposte.

Risposta. Con

1 – totale ammalati ricoverati: $\mathbf{N = 12}$

2 – totale ricoverati morti: $\mathbf{n = 6}$

3 – numero di maschi tra i ricoverati: $\mathbf{n_1 = 5}$

4 – numero di maschi tra i deceduti: $\mathbf{r} = 4$

$$P_{(4/6)} = \frac{C_6^4 \cdot C_{12-6}^{5-4}}{C_{12}^5} = \frac{C_6^4 \cdot C_6^1}{C_{12}^5} = \frac{\frac{6!}{(6-4)!\cdot 4!} \cdot \frac{6!}{(6-1)!\cdot 1!}}{\frac{12!}{(12-5)!\cdot 5!}} = \frac{15 \cdot 6}{792} = 0{,}1136$$

la probabilità sarebbe stata superiore a 11%.

Mentre nella prima risposta si poteva avere un dubbio ragionevole che il farmaco non fosse ugualmente efficace per femmine e maschi, in questo caso la probabilità ($P_{4/6}$ > 11%) è alta, quindi tale da poter essere ritenuta casuale.

Nello stesso modo si possono stimare tutte le 6 possibili risposte:

Maschi **r**	P
5	0,0076
4	0,1136
3	0,3788
2	0,3788
1	0,1136
0	0,0076

Tabella 8. Probabilità di avere per caso il numero r di maschi tra i 6 decessi.

In questo caso la distribuzione delle probabilità è simmetrica e il suo totale, come ovvio, è uguale a 1 non esistendo altre possibili risposte.

ESEMPIO 3. In una riserva di piccole dimensioni sono cresciuti 9 cinghiali, di cui 3 sono femmine (e ovviamente 6 sono maschi). Per ridurre il loro numero, è stata decisa una battuta di caccia, durante la quale sono stati abbattuti 5 animali, senza che vi sia stata la possibilità di fare attenzione al sesso. E' possibile stimare le diverse probabilità che nel gruppo degli animali uccisi siano compresi animali dei due sessi nei vari rapporti possibili.

Nel caso dell'esempio, le domande possono essere:

a) Quale è la probabilità che vengano uccise tutte le femmine?
b) Quale è la probabilità che resti una sola femmina?
c) Quale quella che sia uccisa una sola femmina?
d) Quale quella che sopravvivano tutte e 3 le femmine?

Risposta. Per impostare il calcolo combinatorio, è necessario attribuire i 4 gruppi in modo corretto:

1 - totale animali presenti: **N = 9**

2 - totale animali uccisi: **n = 5**

3 - femmine presenti: $\mathbf{n_1 = 3}$

4 - femmine uccise: **r** a) = **3**; b) = **2**; c) = **1**; d) = **0**.

Il calcolo delle probabilità è riportato nella tabella

R	P
0	0.047619
1	0.357143
2	0.47619
3	0.119048

Tabella 9. Probabilità di eliminare le r femmine su 9 cinghiali, dai quali ne sono stati uccisi 5.

Non esistono altri eventi possibili oltre a quelli stimati; di conseguenza, la somma delle loro probabilità è uguale a 1(o 100%).

La rappresentazione grafica delle probabilità mostra l'effetto del diverso rapporto tra maschi e femmine nel campione, per cui la distribuzione non è simmetrica

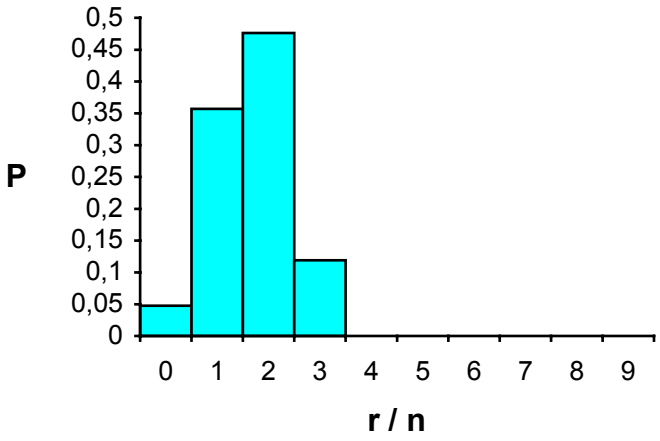


Figura 8. Probabilità di uccidere **r** femmine (da 0 a 3) sui 9 cinghiali.

La distribuzione ipergeometrica è **definita da 3 parametri** (**N, n_1** ed **n**, che rappresentano nell'ordine il numero totale di individui che formano la popolazione, il numero degli oggetti del gruppo considerato, il numero di individui estratti) **in funzione del quarto (r,** il numero di individui estratti appartenenti al gruppo considerato).

Per N che tende all'infinito, la distribuzione ipergeometrica converge verso la distribuzione binomiale, poiché le probabilità restano praticamente costanti. Di conseguenza, **la ipergeometrica è una distribuzione tipica di eventi che riguardano i gruppi formati da poche unità**.

Nella distribuzione ipergeometrica,

- la media μ è $\frac{n_1}{N} \cdot n$; poiché $\frac{n_1}{N} = p$,

essa risulta uguale a **n·p** e quindi alla media della distribuzione binomiale corrispondente.

La varianza σ^2 è uguale a

$$n \cdot p \cdot q \cdot \frac{N - n}{N - 1};$$

pertanto, (poiché $\frac{N-n}{N-1}$ è minore di 1)

è inferiore alla varianza della binomiale, per gli stessi valori di **N** e di **p.**

2.3.6 DISTRIBUZIONE BINOMIALE NEGATIVA

Tra le distribuzioni teoriche discrete, nella ricerca ambientale e biologica una delle più importanti è la distribuzione **binomiale negativa**. Ha forse un numero maggiore di applicazioni delle precedenti; ma è più complessa, sia per gli aspetti teorici e matematici, sia nei calcoli.

Come quella binomiale e quella poissoniana, la distribuzione binomiale negativa permette la stima delle **probabilità di eventi,** misurati mediante **un conteggio**.

In botanica e in ecologia, permette l'analisi della distribuzione territoriale di popolazioni animali e vegetali; in epidemiologia, l'analisi del numero di ammalati in periodi brevi di tempo e in popolazioni piccole, in intervalli abbastanza luoghi; nell'industria, quella del numero scarti o di errori nella produzione varia nel tempo o tra macchine e operatori. In generale, serve quando si elenca il numero di eventi avvenuti per unità temporali o spaziali. Ad esempio, con una rilevazione di 200 giorni su un lungo tratto stradale, può essere il conteggio di incidenti avvenuti ogni giorno; oppure, su 200 km di percorso, il conteggio di quelli avvenuti in un decennio, per ogni tratto lungo un Km.

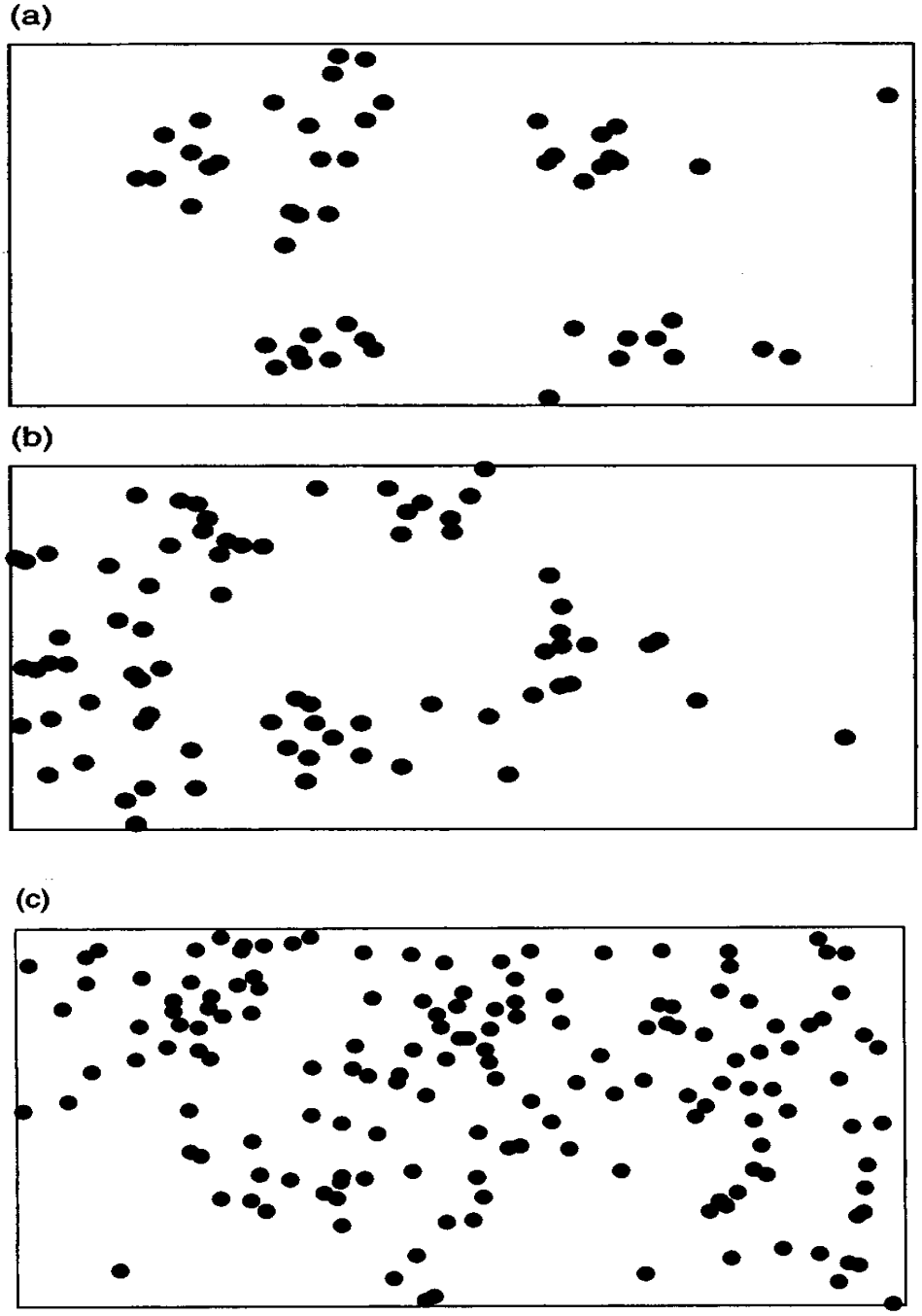


Figura 9. Tipi principali di aggregazione:

a) gruppi piccoli

b) gruppi larghi, con altri individui distribuiti a caso

c) gruppi larghi, con altri individui distribuiti in modo regolare

La distribuzione tende a essere binomiale negativa quando, come molto spesso succede,

- quando **la distribuzione degli eventi è aggregata**, vale a dire **le probabilità cambiano**. Nell'esempio dell'autostrada, si ha una distrbzuione binomiale negativa quando il numero di incidenti cambia perché sono diverse le condizioni meteorologiche oppure perché non tutti i tratti del percorso hanno la stessa pericolosità.

In modo didattico, l'infinita varietà di forme della distribuzione geografica di gruppi di individui è schematizzata in tre tipi principali, riportati graficamente nella figura precedente:

a – **gruppi piccoli**,

b – **gruppi larghi**, con altri individui distribuiti in modo **random**,

c – **gruppi larghi**, con altri individui distribuiti in modo **regolare**.

Per n grande e probabilità p basse, quando la media $\mu = np$ è piccola, unica o costante per tutta l'area o per tutto il periodo presi in considerazione, le frequenze attese sono fornite dalla **distribuzione poissoniana**. Come già illustrato, in essa la varianza $\sigma^2 = npq$ e la media $\mu = np$ sono uguali:

$$\sigma^2 = \mu$$

Ma quando il fenomeno è complesso e la distribuzione di frequenza è determinata da due o più fattori, ognuno con una **media μ diversa ma sempre piccola**, si ha la **distribuzione binomiale negativa**. Può essere vista come una **mistura** o **combinazione di altre distribuzioni**. Spesso è sinonimo di **distribuzioni aggregate.**

In essa, la varianza σ^2 è superiore alla media μ:

$$\sigma^2 > \mu$$

Nella presentazione delle distribuzioni precedenti, è stato evidenziato che

$$p + q = 1$$

dove p indica la probabilità del successo e q la probabilità dell'evento alternativo.

Le differenze fondamentali tra la distribuzione binomiale, poissoniana e binomiale negativa sono collegate ai loro **diversi rapporti tra media e varianza**.

Con n prove,

- la **distribuzione binomiale** è caratterizzata da una varianza $\sigma^2 = npq$ **inferiore** alla media ($\mu = np$), dato che $q < 1$;

- la **distribuzione poissoniana** ha una varianza **uguale** alla media ($\sigma^2 = \mu$) poiché $q \cong 1$**;**

- la **distribuzione binomiale negativa** è caratterizzata da una varianza superiore alla media ($\sigma^2 > \mu$); quindi dovrebbe avere $q > 1$

Per introdurre i concetti sui quali è fondata la distribuzione binomiale negativa, è possibile partire da una considerazione ovvia: quando si analizza la distribuzione territoriale di un gruppo di animali o di una specie vegetale, i dati sperimentali evidenziano spesso, come **dato di fatto**, che

- la varianza ($\sigma^2 = npq$) è superiore alla media ($\mu = np$).

Da questa semplice osservazione, si deducono alcune contraddizioni logiche:

a) poiché $npq > np$**,** deve essere necessariamente $q > 1$; **quindi** p (uguale a $1 - q$) **ha un valore negativo**; ma p rappresenta la probabilità che avvenga un evento e quindi **non può essere negativa**;

b) inoltre, poiché la media ($\mu = np$) deve essere positiva (in quanto media di eventi), con p **negativo**, anche n **deve essere negativo**; ma n è un conteggio e quindi al massimo può essere nullo, mai negativo.

E' una serie di illogicità che complica la soluzione matematica della distribuzione binomiale negativa. Le soluzioni proposte sono numerose.

In una di esse, si pone $n = -k$, con k intero positivo.

Quindi con $-p$ (che è mantenuto inalterato)

il binomio assume la forma

$$\left[q + \left(-p\right)\right]^{-k} = \left(q - p\right)^{-k}$$

da cui il nome di **binomiale negativa** (l'esponente $-k$ è chiamato ***negative-binomial*** k).

Usando una scala continua, quindi con valori indicati con X, il **termine generale** per la **distribuzione binomiale negativa** è

$$P_x = \left[\frac{\Gamma(k+x)}{x!\,\Gamma(k)}\right] \cdot \left(\frac{\mu}{\mu+k}\right)^x \cdot \left(\frac{k}{k+\mu}\right)^k$$

dove

- **P** = probabilità di un quadrato di contenere X individui o di una unità di tempo di avere X casi
- **x** = il conteggio delle unità di venti (0, 1, 2, 3, …)
- **μ** = media della distribuzione per unità di superficie o di tempo
- **k** = esponente binomiale-negativo
- **Γ** = funzione Gamma.

Come evidenzia la figura successiva,

la **distribuzione binomiale negativa è unimodale e ha forma simile alla distribuzione di Poisson**, ma con frequenze maggiori agli estremi e minori nella parte centrale.

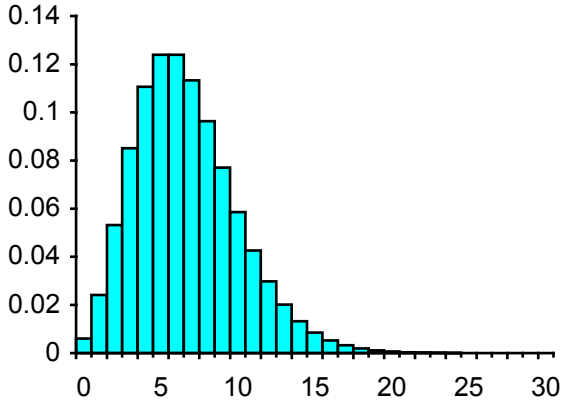


Figura 9. Distribuzione binomiale negativa ($\mu = 6.66, p = 0.6$)

La funzione Gamma può essere bene approssimata dalla **formula di Stirling**

$$\Gamma(z) \approx e^{-z} z^{(z-0,5)} \left(\sqrt{2\pi}\right)\left(1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51.840z^3} - \frac{571}{2.488.302z^4} + \ldots\right)$$

(la virgola e il punto sono usati con simbologia italiana).

I parametri essenziali, poiché gli altri possono essere derivati matematicamente da questi come precedentemente dimostrato, sono 2:

la media $\mu = np$ e l'esponente $k = -n$, da cui

$$\mu = kp$$

La **varianza teorica** o attesa σ^2 della binomiale negativa è

$$\sigma^2 = \mu + \frac{\mu^2}{k}$$

da cui si deduce che sempre $\sigma^2 > \mu$.

Con **dati sperimentali**, al posto di μ si usa $\bar{x}$ e, sviluppando la formula generale precedente,

- la probabilità **P** di osservare **0** individui per quadrato è

$$P_0 = \left(1 + \frac{\bar{x}}{k}\right)^{-k}$$

- la probabilità **P** di osservare **1** individuo per quadrato è

$$P_1 = \left(\frac{k}{1}\right) \cdot \left(\frac{\bar{x}}{\bar{x}+k}\right) \cdot \left(1 + \frac{\bar{x}}{k}\right)^{-k}$$

- la probabilità **P** di osservare **2** individui per quadrato è

$$P_2 = \left(\frac{k}{1}\right) \cdot \left(\frac{k+1}{2}\right) \cdot \left(\frac{\bar{x}}{\bar{x}+k}\right)^2 \cdot \left(1 + \frac{\bar{x}}{k}\right)^{-k}$$

- la probabilità **P** di osservare **3** individui per quadrato è

$$P_3 = \left(\frac{k}{1}\right) \cdot \left(\frac{k+1}{2}\right) \cdot \left(\frac{k+2}{3}\right) \cdot \left(\frac{\bar{x}}{\bar{x}+k}\right)^3 \cdot \left(1 + \frac{\bar{x}}{k}\right)^{-k}$$

- la probabilità **P** di osservare **4** individui per quadrato è

$$P_4 = \left(\frac{k}{1}\right) \cdot \left(\frac{k+1}{2}\right) \cdot \left(\frac{k+2}{3}\right) \cdot \left(\frac{k+3}{4}\right) \cdot \left(\frac{\bar{x}}{\bar{x}+k}\right)^4 \cdot \left(1 + \frac{\bar{x}}{k}\right)^{-k}$$

L'esponente **k** è il parametro più difficile da stimare dipendendo, con $\mu = kp$, sia da μ sia da k.

Inserendo la varianza campionaria (S^2) al posto di quella della popolazione (σ^2)
nella relazione

$$\sigma^2 = \mu + \frac{\mu^2}{k}$$

una prima stima approssimata di k è

$$\hat{k} = \frac{\bar{x}^2}{s^2 - \bar{x}}$$

Per il calcolo della distribuzione teorica a partire da dati campionari, sono stati proposti vari metodi, che dipendono

- sia dalle dimensioni del campione (ad esempio, in un conteggio su una superficie divisa in tanti quadrati uguali, da n = numero di quadrati),

- sia dal numero di individui per unità di rilevazione (vale a dire dalla media $\overline{X}$ e dalla varianza S^2 degli individui presenti nell'insieme dei quadrati).

Il testo di Charles J. **Krebs** del 1999 ***Ecological Methodology*** (2nd ed. Benjamin/ Cummings, Addison Wesley Longman, Menlo Park , California, p. X + 620) propone alcuni metodi approssimati, per risolvere il problema. Per ulteriori approfondimenti di questa metodologia e lo sviluppo di esempi, si rinvia ad esso.

Per **calcolare una distribuzione teorica binomiale negativa** a partire da una distribuzione osservata, esistono altre proposte.

La stima fornita dal metodo del **maximum likelohood** (***maximum likelihood estimation***) non porta a formule semplici, in questo caso.

Una delle soluzioni operativamente più semplici è fornita da Owen L. **Davies** e Peter L. **Goldsmith** nel testo del 1980 ***Statistical Methods in Research and Production with special regference the Chemical Industry*** (4th revised ed. Longman Group Limited, London, XIII + 478 p.). Per risolvere il problema delle contraddittorietà generate dal fatto che

- la varianza ($\sigma^2 = npq$) è superiore alla media ($\mu = np$),

- la media e la varianza sono date in termini di c e k

$$\mu = \frac{k}{c}$$

$$\sigma^2 = \frac{k}{c} + \frac{k}{c^2}$$

da cui si ricava

$$\sigma^2 = \mu\left(1 + \frac{1}{c}\right)$$

Con una **distribuzione campionaria**, i parametri μ e σ^2 sono stimati rispettivamente

- a partire **dalla media** $\overline{X}$ **e dalla varianza** S^2 **osservate.**

Da esse si ricavano

- c e k con

$$c = \frac{\overline{X}}{S^2 - \overline{X}} \quad \text{e} \quad k = c\overline{X}$$

- e le proporzioni p e q con

$$p = \frac{c}{c+1} \quad \text{e} \quad q = 1 - p = \frac{1}{c+1}$$

Infine le probabilità che i diversi eventi avvengano 0, 1, 2, 3, 4,ecc., volte
sono determinate mediante lo sviluppo di

$$p^k \cdot (1-q)^{-k}$$

che è appunto una **binomiale con un indice negativo**.

Nella pratica, per il calcolo della probabilità che l'evento succeda 0 volte, 1 volta, 2, 3 , 4, ecc. volte,

- la parte p^k resta costante, una volta stimate p e k,
- mentre la parte $(1-q)^{-k}$, una volta stimate q e k diventa:

i	P_i
0	1
1	kq
2	$\frac{k \cdot (k+1)}{2!} q^2$
3	$\frac{k \cdot (k+1) \cdot (k+2)}{3!} q^3$
4	$\frac{k \cdot (k+1) \cdot (k+2) \cdot (k+3)}{4!} q^4$

ESEMPIO. Su un tratto auto stradale di 269 Km, è stato contato il numero di tamponamenti avvenuti in cinque anni, per ogni tratto della lunghezza di un Km. Il conteggio del numero di eventi ha dato il risultato seguente

Classe – Eventi	0	1	2	3	4	5	6	7	8	≥9
Freq. Assol.	51	68	61	44	24	15	3	1	0	2

Calcolare la distribuzione binomiale negativa e quella possoniana equivalenti.
Alla fine, trarre le conclusioni dal loro confronto con la distribuzione osservata.

Risposta.

A- Per la **distribuzione binomiale negativa,**

1 - dapprima occorre calcolare la media delle m classi: $\overline{X} = \dfrac{\sum_{i=1}^{m} n_i X_i}{\sum_{i=1}^{m} n_i}$

Con la distribuzione osservata, si ottiene che

Classe – Eventi	X_i	0	1	2	3	4	5	6	7	8	≥9	Tot.
Freq. Assol.	n_i	51	68	61	44	24	15	3	1	0	2	269
Numero Eventi	$n_i X_i$	0	68	122	132	96	75	18	7	0	19*	537

(* Nella classe ≥ 9 è stimato che siano presenti un 9 e un 10, con totale 19 e valore medio X_i = 9,5)

- il numero di tratti o Km considerati è 269,
- il numero totale di tamponanti è 537
- la media per Km risulta $\overline{X}$ = 1,996

$$\overline{X} = \frac{537}{269} = 1{,}996$$

2 – Dalla distribuzione dei dati per classe e dalla media, si ricava la varianza

$$S^2 = \frac{\sum_{1=1}^{m} n_i \left(X_i - \overline{X}\right)^2}{n-1}$$

Con la distribuzione campionaria

X_i	0	1	2	3	4	5	6	7	8	≥9	Tot.
$\left\|Xi - \overline{X}\right\|$	1,996	0,996	0,004	1,004	2,004	3,004	4,004	5,004	6,004	7,504	----
$\left(X_i - \overline{X}\right)^2$	3,99	0,99	0,00	1,01	4,02	9,03	16,03	25,04	36,05	56,31	----
n_i	51	68	61	44	24	15	3	1	0	2	269
$n_i\left(X_i - \overline{X}\right)^2$	203,49	67,32	0,00	44,44	96,48	135,45	48,09	25,04	0,00	112,62	732,93

si ottiene $S^2 = 2{,}735$

$$S^2 = \frac{\sum_{1=1}^{m} n_i \left(X_i - \overline{X}\right)^2}{n-1} = \frac{732{,}93}{268} = 2{,}735$$

3 – Da $\overline{X} = 1{,}996$ e $S^2 = 2{,}735$ si derivano i **parametri** c, k, p, q della distribuzione binomiale negativa

$$c = \frac{\overline{X}}{S^2 - \overline{X}} = \frac{1{,}996}{2{,}735 - 1{,}996} = \frac{1{,}996}{0{,}739} = 2{,}701$$

$$k = c\overline{X} = 2{,}701x1{,}996 = 5{,}391$$

$$p = \frac{c}{c+1} = \frac{2{,}701}{2{,}701+1} = \frac{2{,}701}{3{,}701} = 0{,}730$$

$$q = 1 - p = \frac{1}{c+1} = 1 - 0{,}730 = \frac{1}{2{,}701+1} = \frac{1}{3{,}701} = 0{,}270$$

ricavando $c = 2{,}701$ $k = 5{,}391$ $p = 0{,}730$ $q = 0{,}270$.

4 – Con esse, si possono ricavare le frequenze attese

- Per $i = 0$, la probabilità è $P_0 = 0{,}1833$

$$P_0 = p^k \cdot 1 = 0{,}730^{5{,}391} \cdot 1 = 0{,}1833$$

- Per $i = 1$, la probabilità è $P_1 = 0{,}2668$

$$P_1 = p^k \cdot kq = \left(0{,}730^{5{,}391}\right) \cdot \left(5{,}391 \cdot 0{,}27\right) = 0{,}1833 \cdot 1{,}4556 = 0{,}2668$$

- Per $i = 2$, la probabilità è $P_2 = 0{,}2301$

$$P_2 = p^k \cdot \frac{k \cdot (k+1)}{2!} q^2 = \left(0{,}730^{5{,}391}\right) \cdot \left(\frac{5{,}391 \cdot 6{,}391}{2} \cdot 0{,}270^2\right)$$

$$P_2 = 0{,}1833 \cdot \left(17{,}2270 \cdot 0{,}0729\right) = 0{,}2301$$

- Per $i = 3$, la probabilità è $P_3 = 0{,}1533$

$$P_3 = p^k \cdot \frac{k \cdot (k+1) \cdot (k+2)}{3!} q^3 = \left(0{,}730^{5{,}391}\right) \cdot \left(\frac{5{,}391 \cdot 6{,}391 \cdot 7{,}391}{6} \cdot 0{,}270^3\right)$$

$$P_3 = 0{,}1833 \cdot \left(42{,}4415 \cdot 0{,}0197\right) = 0{,}1533$$

- Per $i = 4$, la probabilità è $P_4 = 0{,}0865$

$$P_4 = p^k \cdot \frac{k \cdot (k+1) \cdot (k+2) \cdot (k+3)}{4!} q^4 = \left(0{,}730^{5{,}391}\right) \cdot \left(\frac{5{,}391 \cdot 6{,}391 \cdot 7{,}391 \cdot 8{,}391}{24} \cdot 0{,}270^4\right)$$

$$P_4 = 0{,}1833 \cdot \left(89{,}0316 \cdot 0{,}0053\right) = 0{,}0865$$

- Per $i = 5$, la probabilità è $P_5 = 0{,}0460$

$$P_5 = p^k \cdot \frac{k \cdot (k+1) \cdot (k+2) \cdot (k+3) \cdot (k+4)}{5!} q^5$$

$$P_5 = \left(0{,}730^{5{,}391}\right) \cdot \left(\frac{5{,}391 \cdot 6{,}391 \cdot 7{,}391 \cdot 8{,}391 \cdot 9{,}391}{120} \cdot 0{,}270^5 \right)$$

$$P_5 = 0{,}1833 \cdot \left(167{,}2190 \cdot 0{,}0015\right) = 0{,}0460$$

- Per $i = 6$, la probabilità è $P_6 = 0{,}0212$

$$P_6 = p^k \cdot \frac{k \cdot (k+1) \cdot (k+2) \cdot (k+3) \cdot (k+4) \cdot (k+5)}{6!} q^6$$

$$P_6 = \left(0{,}730^{5{,}391}\right) \cdot \left(\frac{5{,}391 \cdot 6{,}391 \cdot 7{,}391 \cdot 8{,}391 \cdot 9{,}391 \cdot 10{,}391}{720} \cdot 0{,}270^6 \right)$$

$$P_6 = 0{,}1833 \cdot \left(289{,}5955 \cdot 0{,}0004\right) = 0{,}0212$$

Per la difficoltà pratica di continuare il calcolo con una semplice calcolatrice manuale e per l'approssimazione delle stime a causa degli arrotondamenti, è conveniente stimare la probabilità di avere 7 o più eventi, cumulando le probabilità più estreme.

In modo molto semplice può essere ottenuto per sottrazione da 1 delle probabilità già stimate per gli altri eventi:

0,1833 + 0,2668 + 0,2301 + 0,1533 + 0,0865 + 0,0460 + 0,0212 = 0,9872

- Per $i \geq 7$, la probabilità è $P_{\geq 7}$ = 1 – 0,9872 = 0,00128

In conclusione, le probabilità determinate con la distribuzione binomiale negativa sono

Classe – Eventi	0	1	2	3	4	5	6	≥7	Tot.
Bin. Negativa	,1833	,2668	,2301	,1533	,0865	,0460	,0212	,0128	1,000

B - Per la **distribuzione poissoniana**,

si utilizza la formula

$$P_{\mathrm{i}} = \frac{\overline{X}^i}{\mathrm{i}!} \cdot \mathrm{e}^{-\overline{\mathrm{X}}} = \frac{\overline{X}^i}{i!} \cdot \frac{1}{e^{\overline{X}}}$$

nella quale l'unico parametro che occorre calcolare dalla distribuzione osservata è la media.

Resa più semplice per i calcoli, con $e = 2{,}71828$ e con $\overline{X}$ = 1,996 può essere scritta

come

$$P_i = \frac{1{,}996^i}{i!} \cdot \frac{1}{2{,}71828^{1{,}996}}$$

- Per $i = 0$, la probabilità è $P_0 = 0{,}1359$

$$P_0 = \frac{1{,}996^0}{0!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{1}{1} \cdot \frac{1}{7{,}3595} = 1 \cdot 0{,}1359 = 0{,}1359$$

- Per $i = 1$, la probabilità è $P_1 = 0{,}2713$

$$P_1 = \frac{1{,}996^1}{1!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{1{,}996}{1} \cdot \frac{1}{7{,}3595} = 1{,}996 \cdot 0{,}1359 = 0{,}2713$$

- Per $i = 2$, la probabilità è $P_2 = 0{,}2707$

$$P_2 = \frac{1{,}996^2}{2!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{3{,}9840}{2} \cdot \frac{1}{7{,}3595} = 1{,}992 \cdot 0{,}1359 = 0{,}2707$$

- Per $i = 3$, la probabilità è $P_3 = 0{,}1801$

$$P_3 = \frac{1{,}996^3}{3!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{7{,}9521}{6} \cdot \frac{1}{7{,}3595} = 1{,}3254 \cdot 0{,}1359 = 0{,}1801$$

- Per $i = 4$, la probabilità è $P_4 = 0{,}0899$

$$P_4 = \frac{1{,}996^4}{4!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{15{,}8724}{24} \cdot \frac{1}{7{,}3595} = 0{,}6614 \cdot 0{,}1359 = 0{,}0899$$

- Per $i = 5$, la probabilità è $P_5 = 0{,}0359$

$$P_5 = \frac{1{,}996^5}{5!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{31{,}6813}{120} \cdot \frac{1}{7{,}3595} = 0{,}2640 \cdot 0{,}1359 = 0{,}0359$$

- Per $i = 6$, la probabilità è $P_6 = 0{,}0119$

$$P_6 = \frac{1{,}996^6}{6!} \cdot \frac{1}{2{,}71828^{1{,}996}} = \frac{63{,}2358}{720} \cdot \frac{1}{7{,}3595} = 0{,}0878 \cdot 0{,}1359 = 0{,}0119$$

Per $i \geq 7$ la probabilità può essere ottenuta per sottrazione da 1 delle probabilità già stimate per gli altri eventi:

$$0{,}1359 + 0{,}2713 + 0{,}2707 + 0{,}1801 + 0{,}0899 + 0{,}0359 + 0{,}0119 = 0{,}9957$$

- Per $i \geq 7$, la probabilità è $P_{\geq 7} = 1 - 0{,}9957 = 0{,}0043$

In conclusione, le probabilità determinate con la distribuzione Poissoniana sono

Classe – Eventi	0	1	2	3	4	5	6	≥7	Tot.
Poissoniana	,1359	,2713	,2707	,1801	,0899	,0359	,0119	,0043	1,000

Per un confronto tra la distribuzione osservata e le due distribuzioni teoriche calcolate, è utile una tabella riassuntiva

Classe – Eventi	0	1	2	3	4	5	6	≥7	Tot.
Distr. Osserv.	,1895	,2527	,2267	,1637	,0892	,0558	,0112	,0112	1,000
Bin. Negativa	,1833	,2668	,2301	,1533	,0865	,0460	,0212	,0128	1,000
Poissoniana	,1359	,2713	,2707	,1801	,0899	,0359	,0119	,0043	1,000

dove

- la probabilità P_0 osservata è calcolata da $\frac{51}{269} = 0{,}1895$
- la probabilità P_1 osservata è calcolata da $\frac{68}{269} = 0{,}2527$, ecc.
- la probabilità $P_{\geq 7}$ osservata e stimata da $\frac{3}{269} = 0{,}0112$, ecc.

Dal semplice confronto visivo, risulta con evidenza la maggiore affinità della distribuzione osservata con la distribuzione binomiale negativa. La distribuzione poissoniana ha valori più bassi agli estremi e più alti nelle classi centrali, vicini alla media $\overline{X} = 1{,}996$

Ma per valutare il grado di accordo tra la distribuzione osservata e una delle due distribuzioni teoriche, è necessario ricorre **a test sulla bontà dell'adattamento** (***goodness of fit test***), quali il test χ^2, il **test G^2** e il test di **Kolmogorov-Smirnov**, che sono illustrati nei capitoli successivi.

Per questi test, si devono **utilizzare le frequenze assolute**, non quelle relative.

Il numero totale di osservazioni effettuate è un parametro molto importante per la significatività di questi test. Con le stesse frequenze relative, la significatività dei test risulta tanto maggiore quanto più numeroso è il campione.
Rapportati al totale di 269 casi analizzati, la distribuzione del numero di eventi è

Classe – Eventi	0	1	2	3	4	5	6	≥7	Tot.
Distr. Osserv.	51	68	61	44	24	15	3	3	269,0
Bin. Negativa	49,3	71,8	61,9	41,2	23,3	12,4	5,7	3,4	269,0
Poissoniana	36,6	73,0	72,8	48,5	24,2	9,7	3,3	0,9	269,0

Ovviamente l'interpretazione visiva di quale delle sue distribuzioni teoriche sia in maggiore accordo con quella sperimentale non varia, rispetto alla descrizione effettuata con le proporzioni.
Anche in questo settore di applicazione della statica, **è sempre importante passare dalla interpretazione statistica a quella disciplinare**. E' questa la vera competenza richiesta nella statistica applicata: saper coniugare una buona preparazione statistica con la corretta e approfondita conoscenza dei problemi ai quali viene applicata.

L'**interpretazione disciplinare** dell'accordo tra la distribuzione osservata e quella binomiale è che nel fenomeno analizzato esiste non una media sola, ma sono presenti più medie.
Esse sono tra loro tanto più differenti, quanto maggiore è la varianza calcolata sul campione rilevato.
Riportata ai 269 Km analizzati, questa interpretazione significa che in certi tratti autostradali la frequenza media di incidenti è bassa, mentre in altri è sensibilmente più alta. Sotto l'aspetto operativo, per ridurre il numero di incidenti, è conveniente intervenire dove la pericolosità del tracciato è maggiore. Invece, se l'accordo maggiore fosse stato tra la distribuzione osservata e quella poissoniana, si sarebbe dovuto concludere che la probabilità di incidenti è costante su tutto il percorso. La conseguenza pratica sarebbe stata che una riduzione del numero di incidenti avrebbe potuto essere ottenuta solamente con misure generali, quali i limiti inferiori di velocità, estesi a tutto il tracciato.
Nell'industria, se il numero di scarti per giorno oppure per lotto segue una distribuzione poissoniana, significa che la probabilità di errori è casuale e costante; una riduzione degli scarti è ottenuta solo con un miglioramento delle macchine o un aggiornamento esteso a tutti gli addetti. Se invece segue una distribuzione binomiale negativa, significa che non tutte le macchine sono uguali e/o gli operatori non hanno tutti le stesse capacità; è conveniente intervenire su queste cause specifiche e non con provvedimenti generali.

2.3.7 DISTRIBUZIONE UNIFORME O RETTANGOLARE

La più semplice distribuzione discreta è quella uniforme; la sua caratteristica fondamentale è l'identica possibilità del verificarsi di tutti i risultati possibili. Per esempio, la probabilità che esca un numero da 1 a 6 con il lancio di un dado non truccato è uguale per ognuno dei 6 possibili risultati.

Ad essa si ricorre con frequenza quando, in assenza di ipotesi specifiche più dettagliate o di una conoscenza precisa del fenomeno, la verifica di una distribuzione uniforme rappresenta un punto di partenza minimo, che è quasi sempre accettabile.

L'espressione matematica con cui calcolare la probabilità che una variabile discreta **X**, che segue la distribuzione uniforme, assuma un particolare valore è stabilita da

$$P(x) = \frac{1}{(b-a)+1}$$

dove:

- b = risultato maggiore possibile di X
- a = risultato minore possibile di X.

Nell'esempio dei dadi, è semplice verificare che con **b = 6** e **a = 1**

si ottiene una probabilità

$$P(x) = \frac{1}{(6-1)+1} = \frac{1}{6}$$

pari a 1/6.

Nella distribuzione discreta uniforme la media è

$$\mu = \frac{a+b}{2}$$

e la deviazione standard σ è

$$\sigma = \sqrt{\frac{[(b-a)+1]^2 - 1}{12}}$$

La utilizzazione della distribuzione rettangolare è limitata quasi esclusivamente all'analisi di probabilità a priori. Un caso che ricorre con frequenza nella ricerca ambientale è l'analisi della frequenza di animali in varie aree, per verificare un'omogeneità di dispersione. L'ipotesi alternativa è l'esistenza di una associazione tra presenza (o assenza) della specie e la condizione ambientale dell'area.

2.4. ALCUNE DISTRIBUZIONI CONTINUE

Tutti i modelli precedenti forniscono la distribuzione teorica di variabili casuali discrete. Quando si devono descrivere **variabili casuali continue e positive**, come peso, altezza, reddito, tempo, i modelli più utili sono quelli di seguito riportati. Tra essi, la distribuzione più frequente e di maggiore utilità nella ricerca sperimentale, **fondamento della statistica parametrica, è la distribuzione normale o gaussiana**.

2.4.1 DISTRIBUZIONE NORMALE O DI GAUSS

La più importante distribuzione continua è la curva normale. E' stata individuata per la prima volta nel 1733 da Abraham **De Moivre** (nato in Francia nel 1667, vissuto in Inghilterra e morto nel 1754, la cui opera più importante ***The Doctrine of Chance*** contiene la teoria della probabilità enunciata nel 1718) ed è stata proposta nel 11797 da Karl Friedrich **Gauss** (tedesco, nato nel 1777 e morto nel 1855, ha scritto i suoi lavori più importanti su gravità, magnetismo e elettricità) nell'ambito della teoria degli errori.
Nella letteratura francese è attribuita anche a **Laplace** (1812), che ne avrebbe definito le proprietà principali prima della trattazione più completa fatta da Gauss in varie riprese, a partire dal 1809.

La teoria della stima degli errori è fondata empiricamente sul fatto che tutte le misure ripetute dello stesso fenomeno manifestano una variabilità, che è dovuta all'errore commesso ogni volta. Nei calcoli astronomici, l'indicazione della posizione di ogni stella è sottoposta a un errore e **la distribuzione delle medie campionarie**, secondo Gauss, **è di tipo normale**. Da tempo, in realtà viene chiamato il **dogma della normalità degli errori**, poiché questa supposta normalità della distribuzione non ha una dimostrazione, ma solo qualche verifica empirica.
D'altronde, se il problema è indicare esattamente **dove si trovi la stella**, la **soluzione è impossibile**, a meno che **esista una distribuzione nota degli errori**. Se essi hanno una **distribuzione normale**, è semplice indicare che la zona in cui con probabilità maggiore la stella si trova: è collocata simmetricamente intorno alla media di tutte le medie campionarie. **E' il centro della distribuzione di tutte le misure** e con le leggi della distribuzione normale è possibile **stimare anche la probabilità con la quale si trova entro un intervallo prestabilito**, collocato simmetricamente intorno alla media della distribuzione normale.

Il nome di **curva normale** deriva d alla convinzione, non sempre corretta, che molti fenomeni, da quelli biologici e quelli fisici, normalmente si distribuiscano secondo la **curva gaussiana**. La sua denominazione di **curva degli errori accidentali**, diffusa soprattutto nelle discipline fisiche, deriva

dall'osservazione sperimentale che la distribuzione degli errori, commessi quando si misura ripetutamente la stessa grandezza, è molto bene approssimata da tale curva.
Sotto l'aspetto matematico, la distribuzione gaussiana può essere considerata come il limite della distribuzione binomiale
- per **n** che tende all'infinito,
- mentre né **p** né **q** tendono a 0 (condizione che la differenzia dalla poissoniana).

Se **n** tende all'infinito e **p** resta costante, la media (**n·p**) a sua volta si approssima all'infinito e rende la distribuzione senza applicazioni pratiche. Per contro, la variabile considerata, che nel caso di pochi dati era quantificata per unità discrete, può essere espressa in unità sempre minori, tanto che diventa accettabile esprimerla come una grandezza continua.
La distribuzione gaussiana può essere considerata il limite anche della distribuzione poissoniana, quando **i** e μ diventano molto grandi.

Quando **n** tende all'infinito (in realtà quando **n**, **i** e μ sono molto grandi) a condizione che né **p** né **q** tendano a 0, secondo il teorema di **De Moivre** (1833) la probabilità $\mathbf{P_i}$ della distribuzione binomiale è sempre meglio approssimata
da

$$P(i) = \frac{1}{\sqrt{2\pi \cdot n \cdot p \cdot q}} e^{-\frac{(i - n \cdot p)^2}{2 \cdot n \cdot p \cdot q}}$$

Sostituendo **n·p** con la media μ della popolazione,
n·p·q con la varianza σ^2 della popolazione
e il conteggio **i** (indice di un valore discreto) con **x** (indice di una misura continua),
si ottiene

$$y = f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

che è l'espressione della **funzione di densità di probabilità** (o delle frequenze relative) della distribuzione normale. In termini meno matematici, **permette di stimare il valore di Y (il valore dell'ordinata o altezza della curva) per ogni valore di X (il valore della ascissa)**.
La curva normale ha la forma rappresentata nella figura 10.
La distribuzione normale con media μ**=0** e deviazione standard σ =1 è indicata con **N(0,1)**; al variare di questi due parametri che la definiscono compiutamente, si possono avere infinite curve normali.

Le caratteristiche più importanti della normale sono una frequenza relativamente più elevata dei valori centrali e frequenze progressivamente minori verso gli estremi. La funzione di densità è simmetrica rispetto alla media: cresce da zero fino alla media e poi decresce fino a $+\infty$. Ha due flessi: il primo, ascendente, nel punto $\boldsymbol{\mu}-\sigma$; il secondo, discendente, nel punto $\boldsymbol{\mu}+\sigma$.

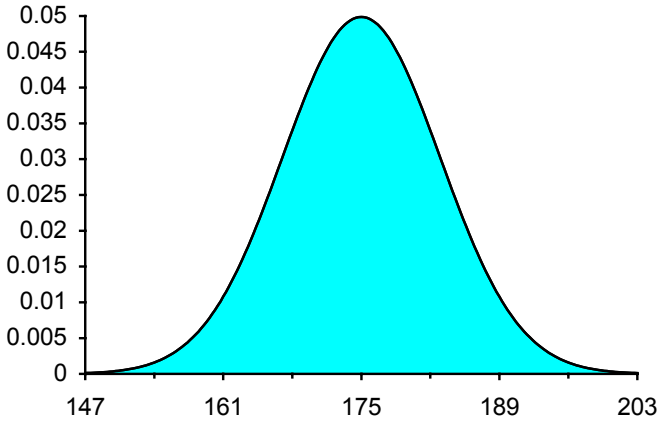


Figura 10. Distribuzione normale **μ=175,** σ=8.

In ogni curva normale, la media, la moda e la mediana sono coincidenti.

Se $\boldsymbol{\mu}$ varia e σ rimane costante, si hanno infinite curve normali con la stessa forma e la stessa dimensione, ma con l'asse di simmetria in un punto diverso. Quando due distribuzioni hanno media differente, è possibile ottenere l'una dall'altra mediante traslazione o trasformazione lineare dei dati.

Se invece $\boldsymbol{\mu}$ rimane costante e $\boldsymbol{\sigma}$ varia, tutte le infinite curve hanno lo stesso asse di simmetria; ma hanno forma più o meno appiattita, secondo il valore di $\boldsymbol{\sigma}$.

Le due curve della figura 11 hanno media $\boldsymbol{\mu}$ identica e deviazione standard $\boldsymbol{\sigma}$ differente.

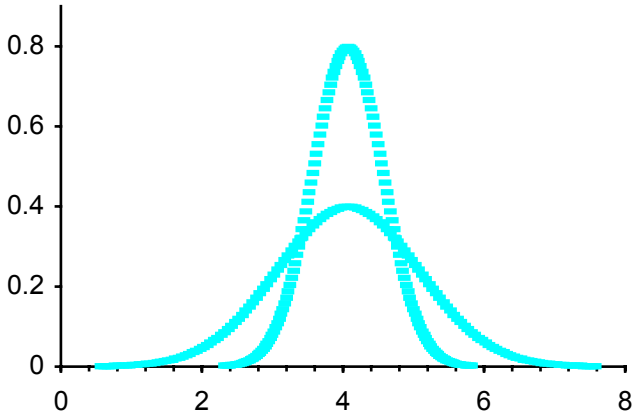


Figura 11. Curve normali con $\boldsymbol{\mu}$ uguale e $\boldsymbol{\sigma}$ diversa.

Una seconda serie di curve con medie diverse e dispersione uguale è quella riportata nella figura successiva (Fig.12).

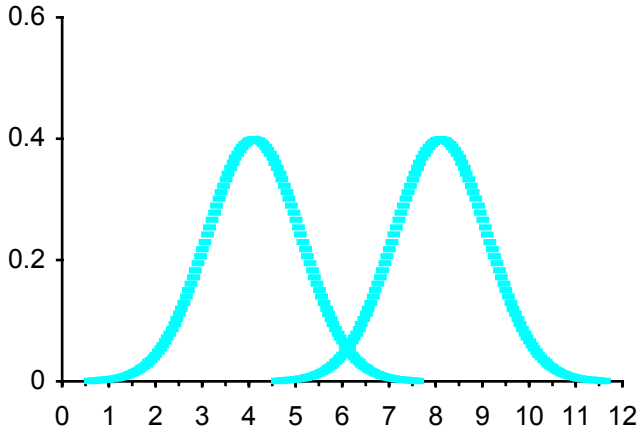


Figura 12. Curve normali con **μ** diversa e **σ** uguale.

Una terza serie è quella delle distribuzioni normali che differiscono sia per la media sia per la dispersione dei dati (figura 13).

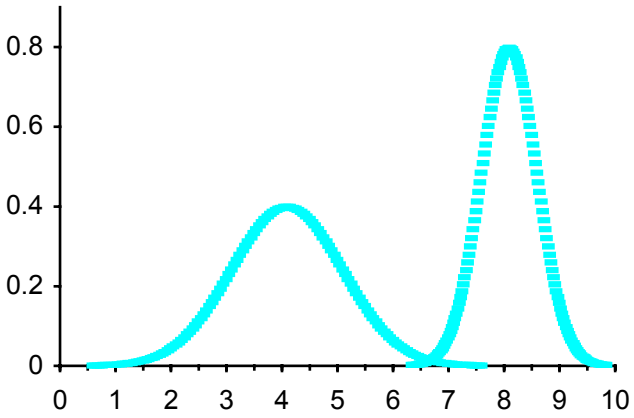


Figura 13. Curve normali **μ** e **σ** diverse.

I momenti e gli indici di forma della distribuzione normale (già presentati nel primo capitolo sulla statistica descrittiva) valutano in modo sintetico e mediante una rappresentazione numerica le sue caratteristiche essenziali, rese visibili e di più facile lettura, ma non quantificate, nella rappresentazione grafica. Poiché la distribuzione teorica normale è simmetrica, tutti i momenti di ordine dispari dalla media sono nulli.

Per i momenti di **ordine pari**, è bene ricordare che

- il momento di secondo ordine è uguale alla varianza ($\mu_2 = \sigma^2$)
- ed il momento di quarto ordine è uguale a 3 volte la varianza al quadrato ($\mu_4 = 3\sigma^4$).

L'indice di **simmetria di Pearson** risulta $\beta_1 = 0$.

L'indice di **curtosi di Pearson** è $\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{3\ \sigma^4}{\sigma^4} = 3.$

L'indice di **simmetria di Fisher** è $\gamma_1 = 0$.

L'indice di **curtosi di Fisher** è $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = 0$

Le infinite forme della distribuzione normale, determinate dalle combinazioni di differenze nella media e nella varianza, possono essere tutte ricondotte alla medesima forma. E' la **distribuzione normale standardizzata** o **normale ridotta**, che è ottenuta mediante il cambiamento di variabile dato da

$$Z = \frac{X - \mu}{\sigma}$$

La **standardizzazione** è una trasformazione che consiste nel:

- rendere la media nulla ($\mu = 0$), poiché ad ogni valore viene sottratta la media;
- prendere la deviazione standard σ come unità di misura ($\sigma = 1$) della nuova variabile.

Come conseguenza, si ottiene anche una trasformazione degli scarti $x-\mu$ in scarti ridotti,

$$Z' = \frac{X - \mu}{\sigma}$$

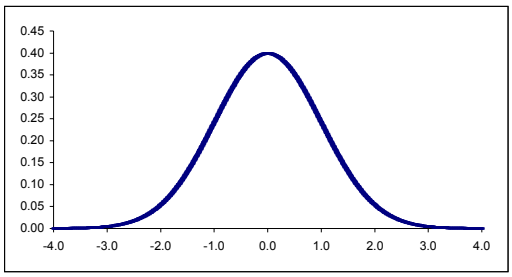
La distribuzione **normale ridotta** viene indicata con **N(0,1)**., che indica appunto una distribuzione normale con media 0 e varianza uguale a 1.

Dopo il cambiamento di variabile, nella normale ridotta la densità di probabilità è data da

$$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

dove

Z è il valore sull'asse delle ascisse, misurato in unità di deviazioni standard dalla media.



Y = Ordinata della curva normale standardizzata in z.

Z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Figura 14 con tabella delle ordinate per ascisse (z) della distribuzione normale standardizzata

Tale relazione evidenzia come **la forma della distribuzione non dipenda più né dalla sua media né dalla sua varianza: è sempre identica, qualunque sia la distribuzione gaussiana considerata**.

La tabella riportata serve solamente per rendere veloce la stima dell'ordinata. Una volta fissato il valore di z, è possibile calcolare il valore dell'ordinata, come nell'esempio successivo per il punto individuato in ascissa da **z** uguale a 1.
Utilizzando la formula

$$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

con i dati dell'esempio si ottiene

$$y = \frac{1}{\sqrt{2 \text{ x } 3{,}14}} \cdot \frac{1}{2{,}71828} = \frac{1}{2{,}5060} \cdot \frac{1}{1{,}6488} = 0{,}399 \text{ x } 0{,}6065 = 0{,}2420$$

un valore di Y uguale a 0,2420.

La tabella precedente riporta i valori della ordinata della curva normale standardizzata in **z**, per **z** che varia da 0 a 3.99.
In essa si trovano i valori delle ordinate, per valori di **z** riportati sommando la prima colonna (in grassetto) con la prima riga (in grassetto).

Ad esempio, per **Z** uguale a 1.00 si deve
- individuare nella prima colonna (in grassetto) il valore 1.0
- individuare nella prima riga (in grassetto) il valore 0.00,

perché la somma 1.0 + 0.00 è uguale a 1.00. Nel loro incrocio, è riportato .2420 che indica appunto il valore dell'ordinata.

Per **Z** uguale a 1,85 si deve
- individuare nella prima colonna (in grassetto) il valore di 1.8
- individuare nella prima riga (in grassetto) il valore .05

e leggere il valore riportato al loro incrocio, uguale a .0721 che indica appunto l'ordinata per **z** uguale a 1,85.

2.4.2 DISTRIBUZIONI ASINTOTICAMENTE NORMALI, CON APPROSSIMAZIONI E TRASFORMAZIONI

L'interesse per le applicazioni della distribuzione normale dipende dal fatto che molte variabili sono distribuite secondo questa legge; inoltre, è accresciuto dal fatto che varie distribuzioni, che non sono rigorosamente normali o addirittura lontane dalla normalità, possono divenirle od essere ritenute tali quando

1) certi loro parametri tendono all'infinito **(leggi asintoticamente normali)**,

2) sono quasi normali **(approssimazioni)**,

3) oppure mediante trasformazioni appropriate, che conducono a variabili distribuite normalmente almeno in modo approssimato **(trasformazioni)**.

1) Come esempi di **leggi asintoticamente normali** si possono ricordare 3 casi già presentati nei paragrafi precedenti.

a - **La distribuzione binomiale $(p + q)^n$ tende alla legge di distribuzione normale**, quando **n** tende all'infinito.

b - **La distribuzione poissoniana tende alla distribuzione gaussiana quando la media è elevata**; in pratica superiore a 6.

c - **La media di n variabili aleatorie indipendenti**, che singolarmente seguono una legge qualunque, **segue la legge normale quando n è grande**.

Non sono distribuzioni esattamente normali, ma sono considerate approssimativamente tali. E' il **teorema centrale** della statistica, noto anche come **teorema fondamentale della convergenza stocastica** o **teorema di Laplace-Chebyshev-Liapounoff** (spesso citato solo come teorema di Laplace):

- "**Qualunque sia la forma della distribuzione di n variabili casuali indipendenti (x_i), la loro somma X (con $X = x_1 + x_2 + x_3 + \ldots + x_n$) è asintoticamente normale, con media generale uguale alla somma delle singole medie e varianza generale uguale alla somma delle singole varianze"**.

Una dimostrazione semplice può essere fornita dal lancio dei dadi. Con un solo dado, i 6 numeri hanno la stessa probabilità ed hanno una distribuzione uniforme; ma con due o più dadi, la somma dei loro numeri tende ad essere sempre più simile alla normale, all'aumentare del numero dei dadi. Se invece della somma si considera la media di **n** lanci, si ottengono i medesimi risultati.

2) Come esempio di **approssimazione alla normale di una distribuzione non normale**, è possibile ricordare che nelle popolazioni animali e vegetali abitualmente la distribuzione normale viene usata

sia nello studio della massa o volume, come nello studio dell'altezza o delle dimensioni di singoli individui; ma tra essi il rapporto non è lineare e quindi queste variabili non potrebbero essere tutte contemporaneamente rappresentate con la stessa legge di distribuzione. **Nella pratica, sono tutte ugualmente bene approssimate dalla distribuzione normale**; quindi, per tutte indistintamente è prassi ricorrere all'uso della normale.

3) Quando i dati hanno una distribuzione differente dalla normale, **spesso una semplice trasformazione conduce ad una distribuzione normale**. E' il caso delle trasformazioni con la **radice quadrata o cubica**, oppure con il **reciproco**, l'**elevamento a potenza** o con **i logaritmi**. Oltre agli indici statistici sulla forma (varianza, simmetria e curtosi), che misurano come la distribuzione trasformata modifichi i suoi parametri, nella scelta del tipo di trasformazione occorre considerare anche la legge biologica o naturale che determina il fenomeno di dispersione dei dati.

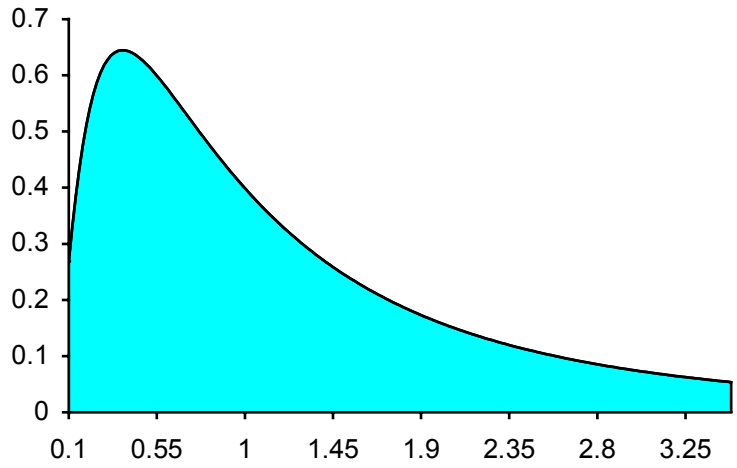


Figura 15. Distribuzione lognormale

Il caso di trasformazione che ricorre forse con frequenza maggiore in biologia e nelle scienze ambientali è quella logaritmica

$$X' = \log X$$

dove

X' diviene una serie di valori distribuiti in buon accordo con la normale.

Quando la distribuzione di una variabile X ha una forma simile a quella rappresentata nella precedente figura 15, con la trasformazione logaritmica in X' assume appunto una forma molto simile alla distribuzione normale.

Le trasformazioni dei dati, qui citate in modo estremamente sintetico con l'unico scopo di presentare il concetto di distribuzioni normali dopo trasformazione, sono numerose. Saranno presentate in modo più approfondito alla fine del secondo capitolo sull'analisi della varianza, quando si discuterà sulle condizioni di validità dei test di statistica parametrica.

La distribuzione normale di **Gauss** e **Laplace**, derivata dai loro studi sulla teoria della distribuzione degli errori d'osservazione, per lungo tempo ha occupato un posto di assoluta preminenza nella statistica. E' stato considerato assiomatico che, con un numero elevato di osservazioni, la variabile casuale avesse una distribuzione normale.
Più recentemente sono stati sollevati dubbi e critiche; alcuni critici e autori di testi di statistica sono giunti ad affermare che tale assunzione era accettata universalmente solo perché
- "**gli statistici sperimentali pensavano che essa derivasse da un teorema, mentre i matematici pensavano che essa fosse un fatto sperimentale**".

Benché la sua importanza sia stata ridimensionata, attualmente è diffusamente accettato che **moltissime distribuzioni sono <u>approssimativamente</u> normali.**

2.4.3 DALLA DISUGUAGLIANZA DI TCHEBYCHEFF ALL'USO DELLA DISTRIBUZIONE NORMALE

Nella pratica statistica, **le proprietà più utili della distribuzione normale** non sono i rapporti tra ascissa ed ordinata, presentati in precedenza**, ma le relazioni tra la distanza dalla media e la densità di probabilità sottesa dalla curva**. In modo più semplice, è possibile definire quanti sono i dati compresi tra la media ed un determinato valore, misurando la distanza dalla media μ in unità di deviazioni standard σ.

La frazione dei casi compresi
- fra $\mu+\sigma$ e $\mu-\sigma$ è uguale al 68,27% (in cifra tonda o in valore approssimato i 2/3),
- quella fra $\mu+2\sigma$ e $\mu-2\sigma$ è uguale 95,45% (in cifra tonda 95%),
- quella fra $\mu+3\sigma$ e $\mu-3\sigma$ è esattamente uguale al 99,73% (circa il 99,9%).

In pratica, nella curva normale la quasi totalità dei dati è compresa nell'intorno della media di ampiezza 3 σ**.**

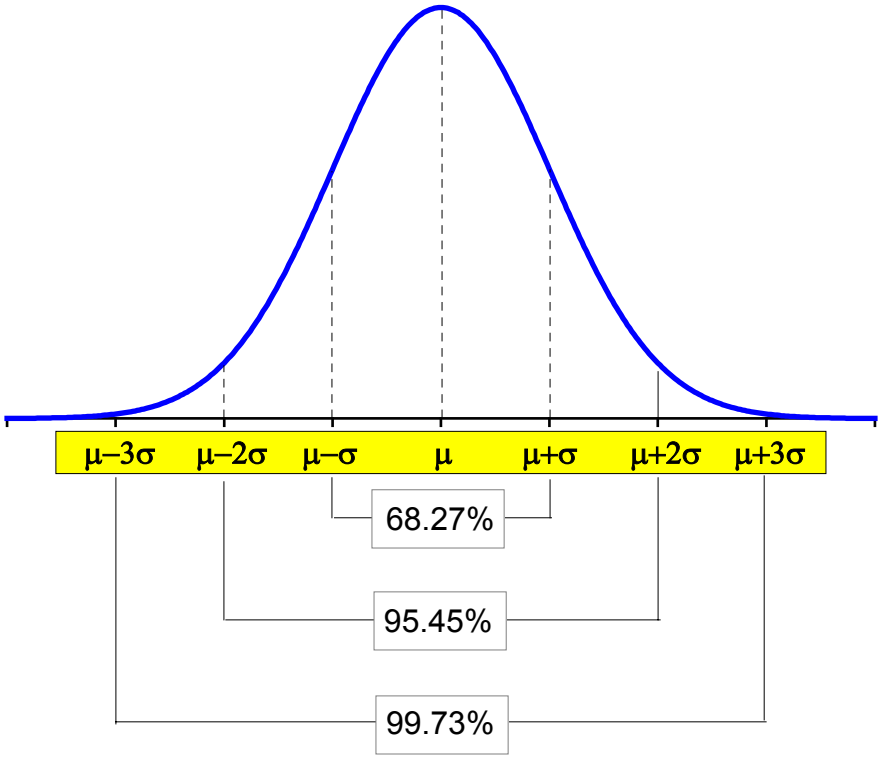


Figura 16. Relazioni tra distanza dalla μ in σ e densità di probabilità sottesa dalla curva.

La relazione tra la percentuale di dati sottesi dalla curva e le dimensioni dell'intervallo tra due valori è una caratteristica di rilevante importanza nella statistica applicata: **se la distribuzione è normale, è sufficiente conoscere due parametri di una serie di dati, la media μ e la varianza σ^2 (**o altro parametro da esso derivato come la deviazione standard **σ), per conoscere anche la sua distribuzione**.

Più di un secolo fa, a partire da dati sperimentali due matematici **Bienaymé e Chebyshev** (Jules Bienaymé francese, nato nel 1796 e morto nel 1878**;** Pahnuty Lvovich Chebyshev russo, nato nel 1821 e morto nel 1894, il cui cognome è ovviamente scritto in cirillico; qui è scritto secondo la pronuncia inglese, in altri testi è scritto **Tchebycheff** in tedesco e **Cebycev** secondo la pronuncia francese) avevano enunciato**: in un gruppo di dati comunque distribuito, la percentuale (P) di osservazioni comprese entro la distanza di k deviazioni standard (σ) intorno alla media μ sono almeno**

$$\mathrm{P}\left(k\frac{|X-\mu|}{\sigma}\right) \leq (1-\frac{1}{\mathrm{K}^2})\cdot 100 \qquad (\text{in}\%)$$

Per **dati distribuiti in qualsiasi modo,** secondo questo teorema noto come **disuguaglianza di Tchebycheff**, nell'intervallo compreso

tra $\pm 2\sigma$ rispetto alla media μ si ha

$$(1 - \frac{1}{2^2}) \cdot 100 = 75$$

almeno il 75% delle osservazioni,

mentre tra $\mu \pm 3\sigma$ si trova

$$(1 - \frac{1}{3^2}) \cdot 100 = \ 88,89$$

almeno l'88,89% dei dati.

e nell'intervallo $\mu \pm 4\sigma$ si trova

$$(1 - \frac{1}{2^4}) \cdot 100 = 93,75$$

almeno il 93,75% dei dati.

Con questa relazione, non è possibile calcolare la quantità di osservazioni compreso nell'intervallo $\mu \pm 1\sigma$.

Questo teorema, che **come principio** è di **notevole importanza nella statistica, in quanto giustifica perché la media e la varianza bastano nella maggior parte dei casi per descrivere la distribuzione di una variabile statistica**, fornisce una stima molto approssimata.

Nel 1946 Harald **Cramèr** (svedese, nato nel 1893 e morto nel 1985, chimico e matematico che ha dato un grande contributo allo studio del teorema del limite centrale e alle funzioni di distribuzione) ha dimostrato che, **se la distribuzione è simmetrica e unimodale,** la stima può essere molto più accurata. La relazione tra le dimensioni dell'intervallo intorno alla media e la distribuzione di frequenza o delle probabilità **P** diviene

$$\mathrm{P}\left(k\frac{|X-\mu|}{\sigma}\right) \leq \left[1 - \left(\frac{4}{9} \cdot \frac{1}{k^2}\right)\right] \cdot 100 \qquad (\text{in}\%)$$

La migliore è la distribuzione normale standardizzata (presentata in precedenza), che **permette i calcoli più precisi** e viene appunto utilizzata per questa sua caratteristica, quando la distribuzione dei dati ha tale forma.

Tabelle specifiche forniscono le frequenze sottese alla curva per ogni valore di **z**, ottenuto con la trasformazione in normale ridotta di qualsiasi distribuzione normale, mediante la relazione

$$Z = \frac{X - \mu}{\sigma}$$

Conoscendo **Z**, con esse è possibile stimare le frequenze o probabilità; con il percorso inverso, è possibile anche stimare il valore di **Z** partendo dalle frequenze. Per comprenderne l'uso, più di spiegazioni teoriche sono utili dimostrazioni pratiche, come quelle riportate negli esercizi seguenti.

Alla fine del capitolo ne sono state riportate 4, anche se i testi di statistica di norma ne riportano una sola. I modi con cui i valori della distribuzione normale sono pubblicati sono numerosi, ma tutti forniscono le stesse informazioni, permettono di derivare con facilità l'una dalle altre, servono per i medesimi scopi.

In tutte quattro le tabelle riportate, il valore di **z** è fornito con la precisione di 2 cifre decimali. Nella prima colonna è riportato la quota intera con la prima cifra decimale; spesso il valore si ferma a 3,0 e quasi mai supera 4,0.

Nella prima riga è riportato il secondo decimale, che ovviamente varia da 0 a 9 e spesso è indicato con 3 cifre da 0,00 a 0,09.

Entro la tabella, all'incrocio del valore dato dalla somma della prima colonna con la prima riga, è riportata la quota di probabilità o frequenza relativa sottesa alla curva entro un intervallo prestabilito, stimato in modo differente nelle 4 tabelle.

La **prima tabella** riporta la quota dell'area in una coda della curva normale standardizzata, non importa se destra o sinistra. Il primo valore, collocato in alto a sinistra, è uguale a 0.500 e corrisponde ad un valore di **Z** uguale a 0.00. Significa che la quota di probabilità sottesa alla curva normale, a destra di un valore che si discosta dalla media μ di una quantità pari a 0,00 σ (quindi coincidente con la media stessa), è pari a 0.500 oppure 50% dei valori, se espresso in percentuale. In altri termini, i dati con valore superiore alla media sono il 50%.

Per **Z** uguale a 1.00, la probabilità è 0.1587: significa che, considerando solo una coda della distribuzione, il 15,87 % dei dati ha un valore che si discosta dalla media di una quantità superiore a 1.00 volte σ. I valori di **Z** usati con frequenza maggiore sono:

- 1.645 perché delimita il 5% dei valori maggiori,
- 1.96 per la probabilità del 2,5% sempre in una coda della distribuzione,
- 2.328 per l'1%,
- 2.575 per il 5 per mille.

La **seconda tabella** riporta l'area sottostante la distribuzione normale standardizzata nell'intervallo compreso tra la media μ e il valore di **Z**. Considera le probabilità sottese in una coda della distribuzione, considerando la curva compresa tra la media e il valore che si discosta da essa di **Z** volte σ. In termini molto semplici, è la differenza a .5000 della tabella precedente: mentre la prima tabella variava da .5000 a 0, questa varia simmetricamente da 0 a .5000.
Per **Z** uguale a 0.00 è uguale a .0000, poiché nell'intervallo tra **Z** e **Z + 0.00**σ non è compreso alcun valore. Per **Z** uguale a 1 è 0.3413. Considerando i valori di **Z** più frequentemente utilizzati, si hanno le seguenti relazioni:
- 1.645 per il 45%,
- 1.96 per il 47,5%,
- 2.328 per il 49%,
- 2.575 per il 49,5%.

La **terza tabella** fornisce la probabilità di ottenere un valore dello scarto standardizzato minore di **Z**. Poiché i valori inferiori alla media sono il 50%, parte da 0.5000 e tende a 1. Rispetto alla prima tabella, può essere visto come il complemento ad 1; rispetto alla seconda, ha aggiunto .5000 ad ogni valore.
Per **Z** uguale a 0.00 è .5000, per **Z** uguale a 1.0 è uguale a 0.8413. Considerando i valori di **Z** più frequentemente utilizzati si hanno le seguenti relazioni:
- 1.645 per il 95%,
- 1.96 per il 97,5%,
- 2.328 per il 99%,
- 2.575 per il 99,5%.

La **quarta tabella** fornisce le probabilità nelle due code della distribuzione. Per ogni **Z**, il valore riportato è il doppio di quello della prima tabella. Parte da 1.000 e tende a 0 all'aumentare di z, con le seguenti relazioni:
- 1.645 per il 10%,
- 1.96 per il 5%,
- 2.328 per il 2%,
- 2.575 per l'1%.

Con una serie di esercizi, è possibile dimostrare in modo semplice e facilmente comprensibile l'utilizzazione pratica della distribuzione normale standardizzata. I calcoli utilizzano quasi esclusivamente la prima tabella, maggiormente diffusa nei testi di statistica applicata.

ESERCIZIO 1. Nella popolazione, la quantità della proteina A ha una media di 35 microgrammi e deviazione standard (σ) uguale 5. Quale è la probabilità di trovare:

a) individui con valori superiori a 40;

b) individui con valori inferiori a 40;

c) individui con valori inferiori a 25;

d) individui con valori compresi tra 40 e 50;

e) individui con valori tra 30 e 40.

Risposte:

a) Con

$$Z = \frac{X - \mu}{\sigma} = \frac{40 - 35}{5} = 1$$

Nella **prima tabella** a una coda, la probabilità esterna a Z = 1,00 è 0,1587; la frequenza di valori oltre 40 o la probabilità di trovare un valore oltre 40 è pari al 15,87%;

c) La probabilità di trovare individui con valori inferiori a 40 è 0,8413 (0,50000 prima della media e 0,3413 a Z = 1,00) corrispondente a 84,13%;

c) Con

$$Z = \frac{X - \mu}{\sigma} = \frac{25 - 35}{5} = -2{,}00$$

Nella prima tabella a Z = -2,00 corrisponde un'area esclusa a sinistra della media pari 0,0228 cioè 2,28%.

d) Il valore di 40 e il valore 50 corrispondono rispettivamente a Z = 1,00 e a Z = 3,00.
Nella prima tabella Z = 1,00 esclude il 15,87% delle misure mentre Z = 3,00 esclude il 0,01%; sono quindi compresi il 15,86%.

e) I due dati, 30 e 40 sono i valori compresi nell'intervallo Z = -1,00 e Z = +1,00; a sinistra e a destra della media l'area sottesa è in entrambi i casi pari a 0,3413 determinando un'area totale pari a 0,6826 (o 68,26%).

ESERCIZIO 2. E' possibile utilizzare le tabelle anche nel modo inverso; cioè leggere su di esse la probabilità e ricavare il valore di Z corrispondente

a) Quale è il valore minimo del 5% dei valori maggiori?

Nella prima tabella, la proporzione 0.05 non è riportata.

Sono presenti

- 0.051, che corrisponde a Z = 1,64 e
- 0,049 che corrisponde a Z = 1,65.

La loro media è 1,645. Essa indica che occorre spostarsi a destra della media (35) di una quantità pari a 1,645 volte la deviazione standard.

$$X = \mu + 1{,}645 \cdot \sigma = 35 + 1{,}645 \cdot 5 = 35 + 8{,}225 = 43{,}225$$

Il 5% dei valori più alti è oltre 43,225.

b) Quale è la quantità massima del 10% dei valori minori?

Nella prima tabella, alla proporzione 0.100 corrisponde Z = 1,28.

Essa indica che occorre spostarsi a sinistra della media (35) di una quantità pari a 1,28 volte la deviazione standard.

$$X = \mu - 1{,}28 \cdot \sigma = 35 - 1{,}28 \cdot 5 = 35 - 6{,}4 = 28{,}6$$

Il 10% dei valori più bassi è inferiore a 28,6.

ESERCIZIO 3. Un anestetico totale, somministrato prima di una operazione, ha una media di milligrammi 60 per Kg di peso, con una deviazione standard pari a 10.

A dose superiori, con media uguale a 120 e deviazione standard 20, esso determina conseguenze gravi sulla salute del paziente.

a) Se un individuo vuole il 90% di probabilità di dormire, di quanto anestetico ha bisogno? Ma con quella quantità di anestetico con quale probabilità può avere conseguenze gravi?

Sempre dalla prima tabella, rileviamo che il valore che esclude la proporzione 0,100 a destra della distribuzione è Z = 1,28 .

Pertanto da

$$X = \mu + 1{,}28 \cdot \sigma = 60 + 1{,}28 \cdot 10 = 60 + 12{,}8 = 72{,}8$$

ricaviamo che la quantità desiderata è 72,8 milligrammi per Kg di peso.

Per stimare il rischio che esso corre di avere conseguenze gravi, calcoliamo il valore di Z corrispondente alla seconda distribuzione normale

$$Z = \frac{X - \mu}{\sigma} = \frac{72{,}8 - 120}{20} = -2{,}36$$

Nella tabella della distribuzione normale, a Z = 2,36 nella coda sinistra della distribuzione corrisponde una probabilità pari a 0,009.
Se un paziente vuole la probabilità di dormire del 90% corre un rischio di avere conseguenze gravi pari al 9 per mille.

b) Ma il paziente ha molta paura e vuole il 99% di probabilità di dormire, di quanto anestetico ha bisogno? Ma con quella quantità di anestetico con quale probabilità può avere conseguenze gravi?

Sempre dalla prima tabella, rileviamo che il valore che esclude la proporzione 0,01 a destra della distribuzione è Z = 2,33 (ne compiono diversi, ma è la stima più precisa a due decimali).
Pertanto da

$$X = \mu + 2{,}33 \cdot \sigma = 60 + 2{,}33 \cdot 10 = 60 + 23{,}3 = 83{,}3$$

ricaviamo che la quantità desiderata è 83,3 milligrammi per Kg di peso.

Per stimare il rischio che esso corre di avere conseguenze gravi, calcoliamo il valore di Z corrispondente alla seconda distribuzione normale

$$Z = \frac{X - \mu}{\sigma} = \frac{83{,}3 - 120}{20} = -1{,}83$$

Nella tabella della distribuzione normale, a Z = 1,83 nella coda sinistra della distribuzione corrisponde una probabilità pari a 0,034.
Se un paziente vuole la probabilità di dormire del 99% corre un rischio di avere conseguenze gravi pari al 34 per mille.

La statistica fa solo i calcoli, non decide. Deve farli bene, perché compete poi al paziente decidere, ma su dati corretti.

Oltre a quantificare la distribuzione dei valori intorno alla media della popolazione, la distribuzione normale serve anche per

quantificare la dispersione delle medie campionarie ($\overline{X}$) intorno alla media della popolazione (μ).

L'unica differenza rispetto a prima è che non si utilizza la deviazione standard,

ma l'**errore standard**

$$\frac{\sigma}{\sqrt{n}}$$

che fornisce appunto la misura della dispersione delle medie di **n** dati, con frequenza prefissata :

$$\overline{X} = \mu \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Ad esempio, entro quale intervallo si troverà il 95% delle medie campionarie di 10 dati (n = 10), se la popolazione ha media μ = 30 con σ = 6.

Dopo aver rilevato nella tabella normale che per una probabilità α = 0.05 nelle due code della distribuzione è riportato Z = 1,96 da

$$\overline{X} = 30 \pm 1{,}96 \cdot \frac{6}{\sqrt{10}} = 30 \pm 1{,}96 \cdot 1{,}898 = 30 \pm 3{,}72$$

si ricava che il 95% delle medie di 10 dati estratte da quella popolazione avrà

- come limite superiore 33,72
- come limite inferiore 26,28.

2.4.4 APPROSSIMAZIONI E CORREZIONI PER LA CONTINUITA'

Molte distribuzioni discrete, quali **la binomiale, l'ipergeometrica e la normale, sono bene approssimate dalla distribuzione normale, per campioni sufficientemente grandi**. L'uso della normale è giustificata anche dalla impossibilità pratica di effettuare i calcoli con le formule delle distribuzioni discrete, a causa di elevamenti a potenze alte e del calcolo di fattoriali per numeri grandi. Nonostante la corrispondenza dei risultati, tra distribuzioni discrete e distribuzioni continue esistono differenze nel calcolo delle probabilità. Le prime forniscono le probabilità per **singoli valori** della variabile casuale, cioè la probabilità di ottenere **esattamente i** volte un determinato evento; le seconde forniscono la probabilità cumulata da un certo valore fino all'estremo della distribuzione.

Per calcolare la probabilità di un singolo valore, con la distribuzione normale si deve calcolare l'area sottesa all'intervallo **X ± 0,5**. In altri termini, per individuare un valore discreto **i** in una scala continua, occorre prendere non il valore esatto **X** ma l'intervallo unitario **X ± 0,5**.

ESEMPIO In una popolazione planctonica, la specie A ha una presenza del 10%; in un campionamento casuale di 120 individui quale è la probabilità di:

a) trovarne **esattamente** 15 della specie A?
b) trovarne **almeno** 15 della specie A?
c) trovarne **meno** di 15 della specie A?

Risposte: (ricordando che n = 120; x = 15; $\mu = n \cdot p = 120 \times 0,10 = 12$;
$\sigma^2 = n \cdot p \cdot q = 120 \times 0,10 \times 0,90 = 10,8$; $\sigma = 3,29$)

a) Per stimare la probabilità di avere esattamente 15, che è un valore discreto, in una scala continua si deve calcolare la probabilità compresa nell'intervallo tra 14,5 e 15,5. Poiché la tabella della distribuzione normale fornisce la probabilità cumulata da quel punto verso l'estremo, il calcolo richiede 3 passaggi:

1- la stima della probabilità da 15,5 verso destra

$$Z_1 = \frac{(X + 0,5) - \mu}{\sigma} = \frac{(15 + 0,5) - 12}{\sqrt{10,8}} = \frac{3,5}{3,29} = 1,06$$

corrispondente a $\mu + 1,06\sigma$ ed equivalente al **35,54%** delle osservazioni;

2- la stima della probabilità da 14,5 verso l'estremo nella stessa direzione precedente:

$$Z_2 = \frac{(X - 0,5) - \mu}{\sigma} = \frac{(15 - 0,5) - 12}{\sqrt{10,8}} = \frac{2,5}{3,29} = 0,76$$

corrispondente a $\mu + 0,76\sigma$ ed equivalente al **27,64%** delle osservazioni;

3 - la sottrazione della prima probabilità dalla seconda, per cui
la probabilità di trovare **esattamente** 15 individui della specie A è

35,54 - 27,64 = 7,90%

uguale a 7,9 %.

Con la distribuzione binomiale, per risolvere lo stesso esercizio il calcolo della probabilità avrebbe dovuto essere

$$P_{(15)} = C_{120}^{15}(0,10)^{15}(0,90)^{105}$$

che rappresenta un'operazione difficile da risolvere manualmente.

b) La probabilità di trovare almeno 15 individui della specie A potrebbe teoricamente essere calcolata con la distribuzione binomiale, sommando le singole probabilità esatte di trovare 15, 16, 17, ecc. fino a 120 individui. Alla precedente difficoltà di calcolo, si aggiunge quella del tempo richiesto per stimare queste 106 probabilità

$$P(X,n) = \sum_{x=0}^{n} C_n^x p^x q^{n-x}$$

da sommare per ottenere quella complessiva.

Con l'utilizzazione della distribuzione normale il calcolo diviene molto più semplice e rapido:

$$Z = \frac{(X-0,5)-\mu}{\sigma} = \frac{14,5-12}{\sqrt{10,8}} = \frac{2,5}{3,29} = 0,76$$

Si stima un valore della distribuzione normale standardizzata con z uguale a 0,76 equivalente ad una probabilità totale verso l'estremo uguale a 22,36%. Pertanto, la probabilità di trovare **almeno** 15 individui (15 o più individui) è uguale a 22,36%.

c) La probabilità di trovare **meno** di 15 individui, con la distribuzione binomiale è data dalla somma delle probabilità esatte di trovare da 0 a 14 individui:

$$C_{120}^{0}(0,1)^{0}(0,9)^{120} + C_{120}^{1}(0,1)^{1}(0,9)^{119} + \cdots + C_{120}^{15}(0,1)^{15}(0,9)^{105}$$

Con la distribuzione normale il calcolo diviene:

$$Z = \frac{(X - 0{,}5) - \mu}{\sigma} = \frac{14{,}5 - 12}{\sqrt{10{,}8}} = \frac{2{,}5}{3{,}29} = 0{,}76$$

Nella tavola della distribuzione normale standardizzata a $Z = 0{,}76$ corrisponde una probabilità, a partire dalla media, uguale a 27,64%. Ad essa va sommato 50 %, che corrisponde alla probabilità di trovare da 0 a 12 individui (la media). Pertanto, la probabilità complessiva di trovare **meno** di 15 individui è

$$50\% + 27{,}64\% = 77{,}64\%$$

uguale al 77,64%.

Allo stesso modo della distribuzione binomiale, è possibile **approssimare la distribuzione ipergeometrica** con la formula

$$Z = \frac{(X - n \cdot p) \pm 0{,}5}{\sqrt{n \cdot p \cdot q} \cdot \sqrt{\frac{N - n}{N - 1}}}$$

dove:

- **N** = numero totale di individui del campione,
- **n** = numero di individui estratti dal campione.

Per la **distribuzione di Poisson** il calcolo del valore di **z** diviene

$$Z = \frac{(X - \mu) \pm 0{,}5}{\sqrt{\mu}}$$

dove

μ è il valore della media che, come dimostrato, coincide con quello della varianza σ^2.

2.4.5 DISTRIBUZIONE RETTANGOLARE

Come nelle distribuzioni discrete, anche tra le distribuzioni continue la più semplice è la distribuzione rettangolare, detta anche **distribuzione uniforme continua**.

La **distribuzione rettangolare continua**, compresa nell'intervallo tra $x_1 = \alpha$ e $x_2 = \beta$, come densità di frequenze relative ha la funzione

$$f(x) = \frac{1}{\beta - \alpha} \qquad \text{con } (\alpha < x < \beta)$$

e pertanto è caratterizzata da una densità costante in tutto l'intervallo da α a β.

Nella rappresentazione grafica questa distribuzione ha la forma di un rettangolo, figura che giustifica il suo nome.

La **media** è

$$\mu = \frac{\alpha + \beta}{2}$$

e la **varianza**

$$\sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

Ovviamente questa distribuzione è l'equivalente dell'uniforme discreta, considerata nel continuo.

2.4.6 DISTRIBUZIONE ESPONENZIALE NEGATIVA

La **esponenziale negativa è una distribuzione continua** con funzione

$$f(x) = \alpha e^{-\alpha x} \qquad \text{con } (\alpha > 0 \ , x>0)$$

che prende il nome dall'esponente negativo che compare nella relazione. E' una funzione positiva o nulla continuamente decrescente, che tende a 0 per x che tende all'infinito. Nel discreto ha l'equivalente nella distribuzione geometrica decrescente.

Media e varianza sono rispettivamente:

media μ

$$\mu = \frac{1}{\alpha}$$

varianza σ^2

$$\sigma^2 = \frac{1}{\alpha^2} = \mu^2$$

E' di estremo interesse pratico, **per dedurre la curva sottostante una determinata distribuzione di dati sperimentali**, notare che in questa distribuzione **la varianza è uguale al quadrato della media**.

2.4.7 LE CURVE DI PEARSON

Karl Pearson ha proposto **non una curva sola ma una famiglia di curve**, utili per descrivere con elevata approssimazione molte distribuzioni empiriche, modificando i suoi parametri. Nonostante gli ottimi risultati ottenuti nell'approssimazione, ha il grave limite che i parametri che la definiscono non sono esplicativi del fenomeno e quindi si prestano male ad usi predittivi.

La forma esplicita della funzione può essere espressa come

$$\frac{dy}{dx} = \frac{y(x+c)}{b_0 + b_1 x + b_2 x^2}$$

che dipende dalle radici dell'espressione quadratica del denominatore, cioè dai valori dei parametri $\mathbf{b_0}$, $\mathbf{b_1}$ e $\mathbf{b_2}$, e dove **y** e **x** sono i valori degli assi e **c** è una costante.

Il sistema gode della proprietà di rappresentare molte curve, come quelle di seguito disegnate.

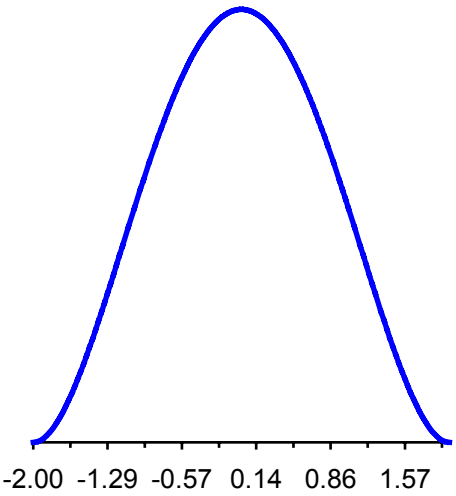


Fig. 17. Curva di Pearson : forma a campana simmetrica

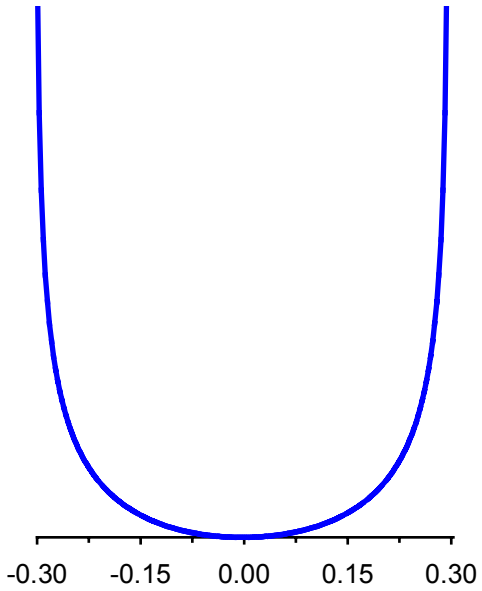


Fig.18. Curva di Pearson: forma a U simmetrica

Variando i parametri prima definiti, si passa dall'una all'altra forma di curva.

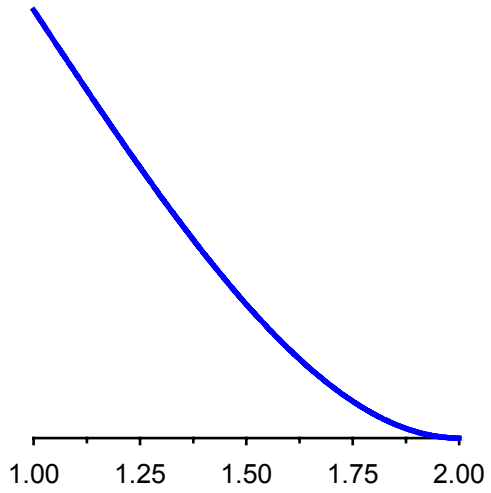


Figura19. Curva di Pearson : forma a J rovesciato.

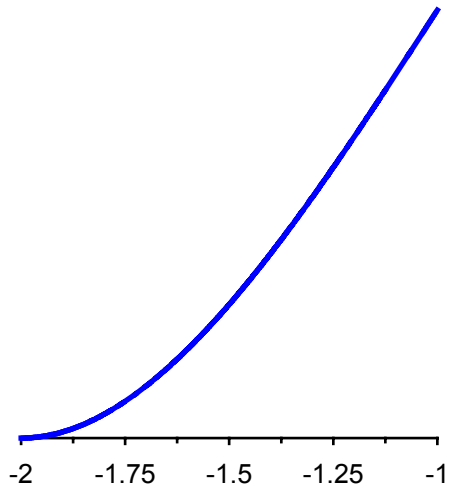


Figura 20. Curva di Pearson : forma a J .

Anche la distribuzione normale e le sue approssimazioni, con curtosi ed asimmetria variabili, possono essere rappresentate come una delle possibili curve di Pearson.

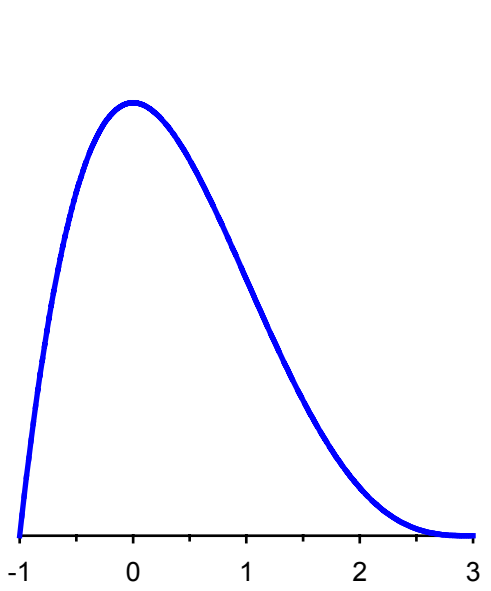


Fig. 21 Curva di Pearson a campana asimmetrica asimmetrica.

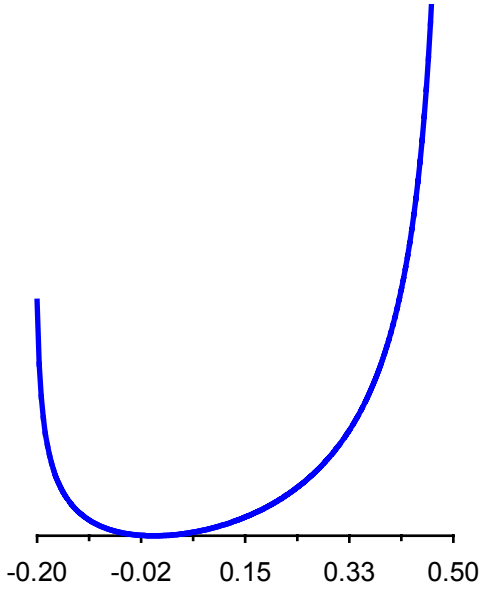


Fig. 22 Curva di Pearson: forma a U

Lo studio dettagliato delle diverse curve richiederebbe una trattazione complessa, che non viene affrontata in modo più dettagliato in questo corso, per le sue ridotte applicazioni nelle scienze ambientali ed ecologiche.

2.4.8 LA DISTRIBUZIONE GAMMA

Un altro modello per descrivere la distribuzione di variabili casuali continue e positive, come altezza o lunghezza, peso, tempo, densità e concentrazione, è la distribuzione Gamma (Γ).
La sua funzione di densità di probabilità è

$$f(x) = K \cdot (x^{\nu - 1} / \mu^{\nu}) \exp (-\nu x / \mu)$$

per $\mathbf{x} > 0$
e dove sia μ che ν sono maggiori di 0,
mentre **K** è una costante che rende unitaria l'area sottesa dalla curva.
Quando $\mathbf{x} \leq 0$ la funzione è uguale a 0.

I parametri che determinano la funzione di densità della curva Γ (indicata sui testi anche G da Gamma) sono
- la media μ ed ν (chiamato indice della distribuzione),
- mentre la costante **K**

è data da

$$K = \nu^{\nu} / \Gamma(\nu), \quad \text{con} \quad \Gamma(\nu) = \int_0^{\infty} x^{\nu-1} e^{-x} dx$$

Per il calcolo di $\Gamma(\nu)$ sono disponibili tavole apposite.
Per ν intero positivo, è possibile calcolare $\Gamma(\nu)$ mediante

$$\Gamma(\nu) = (\nu - 1) !$$

per i valori interi e

$$\Gamma(\nu + ½) = \sqrt{\pi} \ \{(1 \cdot 3 \cdot 5 \cdot 7 \cdot ... \cdot (2\nu-1)\} / 2^{\nu}$$

per valori interi + 0,5.

I casi particolari più importanti della distribuzione Gamma sono
- la distribuzione Esponenziale
- la distribuzione Chi-quadrato (χ^2).

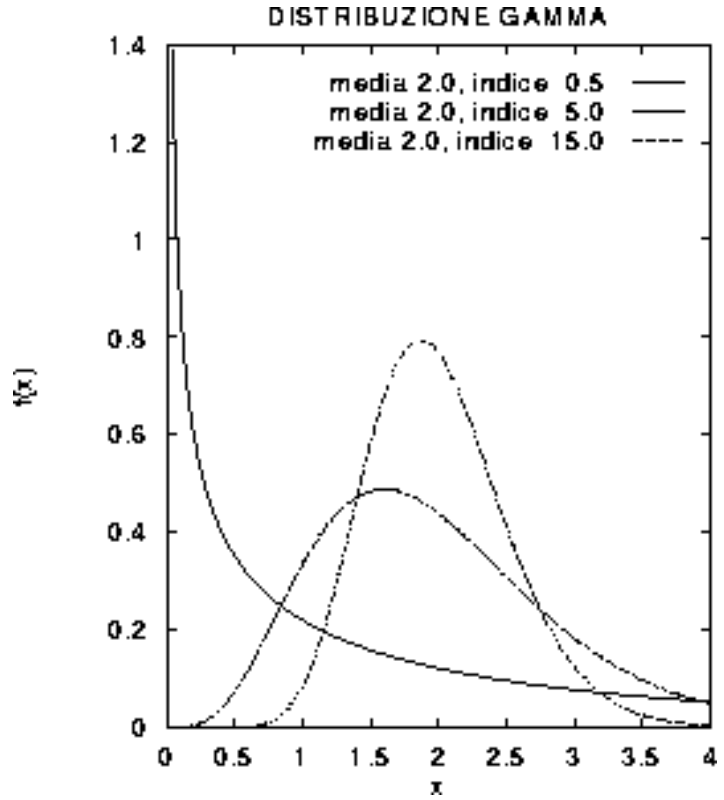


Figura 22. Alcune distribuzioni Gamma

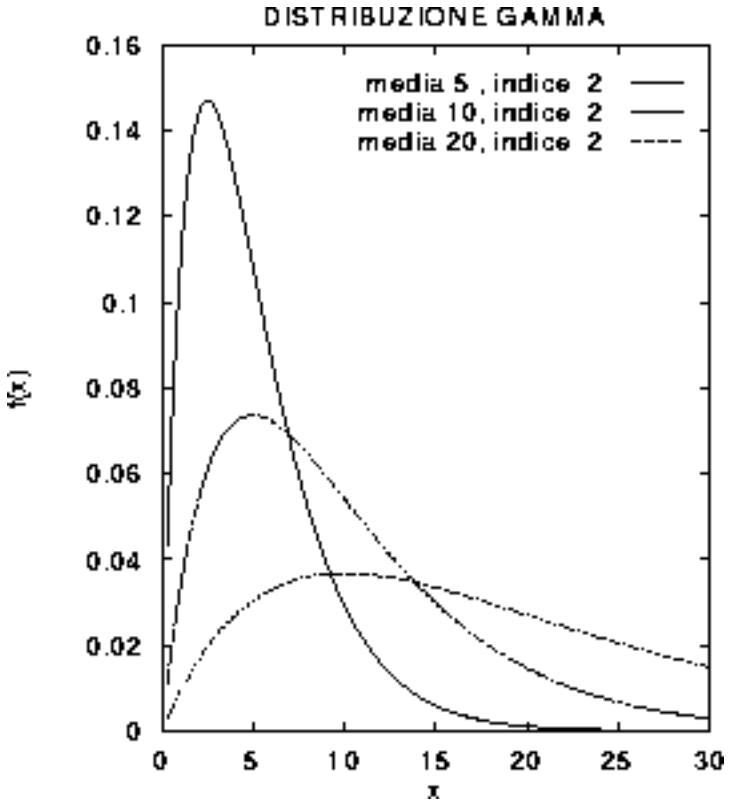


Figura 23. Altre forme della distribuzione Gamma.

La funzione di densità di probabilità della

distribuzione esponenziale è

$$f(x) = (1 / \mu) \exp (- x / \mu) \qquad \text{per } x>0$$

dove $\mu > 0$.

La esponenziale è utile per descrivere la distribuzione delle durate di intervalli di tempo, in cui si manifesta un fenomeno biologico od ambientale, calcolando la distribuzione degli intervalli di tempo tra il manifestarsi di due eventi successivi.

E' in relazione con la distribuzione di Poisson, che fornisce la distribuzione degli eventi in un determinato intervallo di tempo.

Se il numero di eventi **i,** che avvengono in un determinato intervallo di tempo **t**, segue la legge di Poisson con media λ,

- il tempo di attesa **X** intercorrente tra due eventi segue la legge esponenziale con parametro

$$\mu = 1 / \lambda.$$

Il tempo medio di attesa μ tra due eventi è costante ed è pari al reciproco della media utilizzata nella distribuzione binomiale (in epidemiologia, chiamato tasso d'incidenza ed uguale al numero di nuovi eventi contati in un periodo unitario, come ora, giorno, mese, anno).

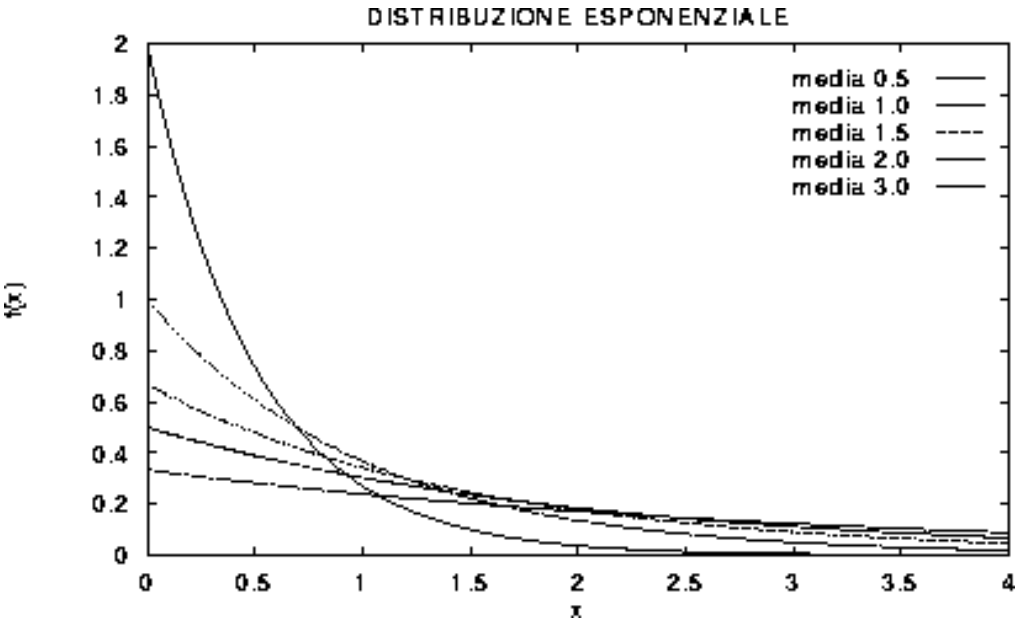


Figura 24. Distribuzione esponenziale (negativa).

2.5. DISTRIBUZIONI CAMPIONARIE DERIVATE DALLA NORMALE ED UTILI PER L'INFERENZA

La distribuzione normale è valida per campioni molto numerosi, teoricamente infiniti. Spesso è possibile disporne nella statistica economica e sociale, in cui si analizzano i dati personali di una regione o una nazione. Nella pratica della ricerca statistica biologica, naturalistica ed ambientale, per l'inferenza sono disponibili alcune decine, al massimo poche centinaia di osservazioni. In molti settori della ricerca applicata, **molto spesso i campioni hanno dimensioni ancora minori e la loro numerosità gioca un ruolo importante, nel determinare la forma della distribuzione**. Essa non può più essere considerata normale od approssimativamente tale, ma se ne discosta quanto più il campione è piccolo.

Per l'inferenza, nella statistica parametrica l'ipotesi fondamentale è che questi campioni siano estratti da una popolazione normalmente distribuita. E' un'ipotesi limitativa, ma basilare per le distribuzioni **t di Student e F di Fisher,** che insieme rappresentano le distribuzioni fondamentali dell'inferenza statistica parametrica. E' importante comprendere come le 3 distribuzioni più utilizzate nell'inferenza statistica, la distribuzione χ^2 di Pearson in quella non parametrica, la distribuzione **t di Student e la F di Fisher,** per la statistica parametrica, **siano legate logicamente e matematicamente con la distribuzione normale e tra loro.**

2.5.1 LA DISTRIBUZIONE χ^2

La distribuzione **Chi-quadrato (χ^2)**, il cui uso è stato introdotto dallo statistico inglese Karl Pearson (1857–1936), può essere fatta derivare dalla distribuzione normale. **Date n variabili casuali indipendenti $x_1, x_2, \ldots, x_n$,**

normalmente distribuite con $\mu = 0$ e $\sigma = 1$,

il χ^2 è una variabile casuale data dalla somma dei loro quadrati.

La funzione di densità di probabilità della distribuzione χ^2 è

$$f(x) = K \cdot x^{(\nu/2)-1} \exp(-x/2)$$

dove $\nu = 1, 2, \ldots$ e $\mathbf{K} = 2^{-\nu/2} / \Gamma(\nu/2)$.

La funzione di densità del χ^2 è determinata solo dal parametro ν, il numero di gradi di libertà, pertanto viene scritta come $\chi^2_{(\nu)}$.

La distribuzione χ^2 **parte da ν uguale a 1 e al suo aumentare assume forme sempre diverse, fino ad una forma approssimativamente normale per $\nu = 30$**

Una buona approssimazione è data dalla relazione

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1}$$

Con **ν molto grande** (oltre 200, per alcuni autori di testi) è possibile dimostrare che si ottiene una nuova variabile casuale (Z), normalmente distribuita, con media μ uguale a **0** e deviazione standard σ uguale a **1**

La distribuzione chi quadrato e le sue relazioni con la normale possono essere spiegate in modo semplice attraverso alcuni passaggi.

Supponendo di avere una popolazione di valori X, distribuita in modo normale,

$$Z = \sqrt{2\chi^2} - \sqrt{2\nu-1}$$

la media μ di questa distribuzione è

$$E(X) = \mu$$

e la varianza σ^2 è

$$E(X - \mu)^2 = \sigma^2$$

Se da questa distribuzione si estrae un valore **X** alla volta, per ogni singolo valore estratto si può stimare un **punteggio Z^2 standardizzato** attraverso la relazione

$$Z^2 = \frac{(X-\mu)^2}{\sigma^2}$$

Questo valore al quadrato, a differenza della **Z**,

- può essere **solo positivo** e variare da 0 all'infinito,

$$\chi^2_{(1)} = Z^2$$

Esso coincide con il chi quadrato con un grado di libertà.

Nella distribuzione **Z**, il 68% dei valori è compreso nell'intervallo tra –1 e +1; di conseguenza il chi quadrato con 1 gdl calcolato con

$$\chi^2_{(1)} = \frac{(X-\mu)^2}{\sigma^2}$$

ha una quantità equivalente di valori (il 68% appunto) tra 0 e 1.

Analizzando **non un caso solo ma due casi,** con la formula

$$Z_1^2 = \frac{(X_1 - \mu)^2}{\sigma^2} \quad \text{e} \quad Z_2^2 = \frac{(X_2 - \mu)^2}{\sigma^2}$$

si calcola un chi quadrato con 2 gradi di libertà

$$\chi^2_{(2)} = \frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} = z_1^2 + z_2^2$$

fondato su due osservazioni indipendenti, che avrà una forma meno asimmetrica del caso precedente e una quantità minore di valori compresi tra 0 e 1.

Con **n** osservazioni $\mathbf{X_i}$ indipendenti, estratte casualmente da una popolazione normale con media $\boldsymbol{\mu}$ e varianza $\boldsymbol{\sigma^2}$, si stima una variabile casuale chi quadrato

$$\chi^2_{(n)} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} z_i^2$$

con **n** gradi di libertà e uguale alla somma degli **n** valori $\mathbf{Z^2}$.

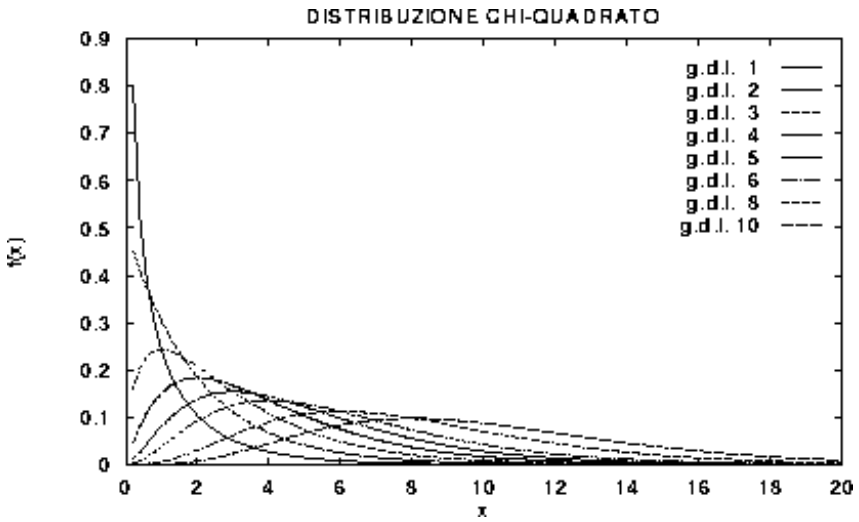


Figura 25. Alcune distribuzioni $\chi^2_{(\nu)}$, con gdl che variano da 1 a 10

La variabile casuale χ^2 **gode della proprietà additiva**: se due o più chi-quadrato, ognuno con i propri gdl sono indipendenti, dalla loro somma si ottiene un nuovo chi-quadrato con gdl uguale alla somma dei gdl.

$$\chi^2_{n1+n2+n3} = \chi^2_{n1} + \chi^2_{n2} + \chi^3_{n3}$$

Anche la **varianza campionaria s^2 ha una distribuzione chi quadrato,** come verrà in seguito approfondito. Il χ^2 può servire per valutare se la varianza σ^2 di una popolazione, dalla quale sia stato estratto un campione con varianza $\mathbf{S^2}$, sia uguale o diversa da un valore predeterminato σ_0^2. Questi concetti sono espressi nell'ipotesi nulla H_0

$$H_0: \sigma^2 = \sigma_0^2$$

con ipotesi alternativa H_1

$$H_1 = \sigma^2 \neq \sigma_0^2$$

Per decidere alla probabilità α tra le due ipotesi, si stima un valore del chi quadrato

$$\chi^2_{(n-1)} = \frac{(n-1)\cdot s^2}{\sigma_0^2}$$

determinato dal rapporto tra il prodotto degli **n-1** gradi di libertà con il valore sperimentale $\mathbf{s^2}$ e la varianza σ_0^2 attesa o predeterminata.

Per ogni grado di libertà, si dovrebbe avere una tabella dei valori del $\chi^2_{(v)}$, come già visto per la distribuzione normale. Per evitare di stampare tante pagine quante sono i gradi di libertà, di norma viene utilizzata una **tavola sinottica**, una pagina riassuntiva, che per ogni grado di libertà riporta solo i valori critici più importanti corrispondenti alla probabilità α del 5% (0.05), 1% (0.01), 5 per mille (0.005) e 1 per mille (0.001).

All'uso del $\chi^2_{(v)}$ è dedicato il successivo capitolo 3.

Non disponendo delle tabelle relative al chi quadrato e alla distribuzione normale, è possibile passare dall'una altra.

Per passare dai valori del χ^2 al valore **z**, ricordando che, con ν grande, la distribuzione $\chi^2_{(\nu)}$ è approssimativamente normale, è possibile ricorrere alla relazione

$$z_\alpha = \frac{\chi^2_{(\nu,\alpha)} - \nu}{\sqrt{2\nu}}$$

poiché quando i gradi di libertà sono molto più di 100
la media μ della distribuzione $\chi^2_{(\nu)}$ è uguale a ν e
e la varianza σ^2 è uguale a 2ν.

Per esempio, si abbia con ν **=100,** alla probabilità α **= 0.05** il valore di χ^2**= 124,342;**
mediante la relazione

$$\frac{124{,}342 - 100}{\sqrt{2 \cdot 100}} = \frac{24{,}342}{14{,}142} = 1{,}72$$

si ottiene un valore di **z** uguale a 1,72 mentre il valore corretto è 1,6449. L'errore relativo è del 4,5%.

Inversamente, dal valore di **Z** è possibile ricavare quello del $\chi^2_{(\nu)}$ alla stessa probabilità α. Quando ν è grande, maggiore di 100, per stimare il valore del chi quadrato che esclude una quota α di valori in una coda della distribuzione si ricorre alla relazione

$$\chi^2_{(\nu,\alpha)} = \frac{1}{2} \cdot \left[Z_\alpha + \sqrt{2\nu - 1} \right]^2$$

in cui Z_α è il valore di **Z** alla probabilità α prefissata.

Per esempio, con ν **=100,** alla probabilità α **= 0.05** con il valore di **Z = 1,6449** mediante la relazione

$$\chi^2_{(100,0.05)} = \frac{1}{2}\left[1{,}6449 + \sqrt{2 \cdot 100 - 1}\right]^2 = \frac{1}{2}\left(1{,}6449 + 14{,}1467\right)^2 = \frac{1}{2} \cdot 15{,}7516^2 = 124{,}056$$

si calcola un valore di $\chi^2_{(100,0.05)}$ uguale a 124,056 mentre il valore corretto alla terza cifra decimale, riportato nelle tabelle, è 124,342. Il processo inverso permette una stima migliore.

Una approssimazione ancora migliore, che fornisce una stima accurata anche con pochi gradi di libertà (**ν**), è stata proposta da **Wilson** e **Hilferty** nel 1931 con la relazione

$$\chi^2_{(\nu,\alpha)} = \nu \cdot \left[Z_\alpha \sqrt{\frac{2}{9\nu}} + 1 - \frac{2}{9\nu} \right]^3$$

Per esempio, con **ν =10,** alla probabilità **α = 0.05** con il valore di **Z = 1,6449** mediante la prima relazione, valida per **ν** grande

$$\chi^2_{(10,0.05)} = \frac{1}{2}\left[1{,}6449 + \sqrt{2 \cdot 10 - 1}\right]^2 = \frac{1}{2}\left(1{,}6449 + 4{,}3559\right)^2 = \frac{1}{2} \cdot 6{,}0008^2 = 18{,}0048$$

si trova un valore di $\chi^2_{(10,0.05)}$ uguale a 18,0048

mentre con la seconda relazione

$$\chi^2_{(10,0.05)} = 10 \cdot \left[1{,}6449 \cdot \sqrt{\frac{2}{9 \cdot 10}} + 1 - \frac{2}{9 \cdot 10} \right]^3 = 10 \cdot \left(1{,}6449 \cdot 0{,}1491 + 1 - 0{,}0222\right)^3 = 10 \cdot 1{,}2231^3$$

$$\chi^2_{(10,0.05)} = \mathbf{10 \cdot 1{,}8297 = 18{,}297}$$

si trova un valore di $\chi^2_{(10,0.05)}$ uguale a 18,297 che è molto più vicino al valore 18,3070 riportato nelle tabelle, appunto per **ν =10,** alla probabilità **α = 0.05.**

Nelle 2 tabelle successive, sono riporti i valori di **z** alle varie probabilità **α** per trovare il valore corrispondente del **χ^2** per i gradi di libertà **ν** prefissati

(la tabella del chi quadrato è riportata alla fine del terzo capitolo).

α	0.995	0.990	0.975	0.950	0.900	0.750	0.500
Z	-2,5758	-2,3263	-1,9600	-1,6449	1,2816	-0,6745	0,0000

α	0.250	0.100	0.050	0.025	0.010	0.005	0.001
Z	+0,6745	+1,2816	+1,6449	+1,9600	+2,3263	+2,5758	+3,0902

Occorre ricordare che anche la distribuzione chi quadrato è normale, quando **ν** è molto grande. Ciò spiega, in modo semplice ed intuitivo, perché in tale situazione quando **Z** è uguale a **0**, alla probabilità **α** corrispondente al **50%**, si abbia un valore del chi quadrato uguale alla sua media **ν**.
La tabella dei valori critici mostra che con gradi di libertà **ν = 100**, la media (corrispondente alla probabilità $\alpha = 0.500$) non è esattamente 100 ma 99,3341 a dimostrazione del fatto che non è una distribuzione perfettamente normale.

2.5.2 LA DISTRIBUZIONE t DI STUDENT

La distribuzione **t** di Student (pseudonimo del chimico inglese Gosset che ne propose l'applicazione al confronto tra medie campionarie) considera le relazioni tra media e varianza, **in campioni di piccole dimensioni**, quando si utilizza la varianza del campione. La scelta tra l'uso della normale o della distribuzione t di Student nel confronto tra medie deriva appunto dalla conoscenza della varianza σ^2 della popolazione o dal fatto che essa sia ignota e pertanto che, in sua vece, si debba utilizzare la varianza campionaria s^2.

Se una serie di medie campionarie ($\bar{X}$) è tratta da una distribuzione normale ridotta ($\mu = 0$, $\sigma = 1$) e la varianza del campione è s^2, con distribuzione χ^2 e **ν** gdl, è possibile derivare la v.c. **t di Student**, tramite la relazione

$$t^2 = \frac{Z^2}{\chi^2 / \nu}$$

dove

i gdl **ν** corrispondono a **N –1**, con **N** uguale al numero totale di dati.

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{\nu}}}$$

La curva corrispondente è simmetrica, leggermente più bassa della normale e con frequenze

$$f(t) = f_0\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

maggiori agli estremi, quando il numero di gdl (ν) è molto piccolo.

Per ν che tende all'infinito, la curva tende alla normale.

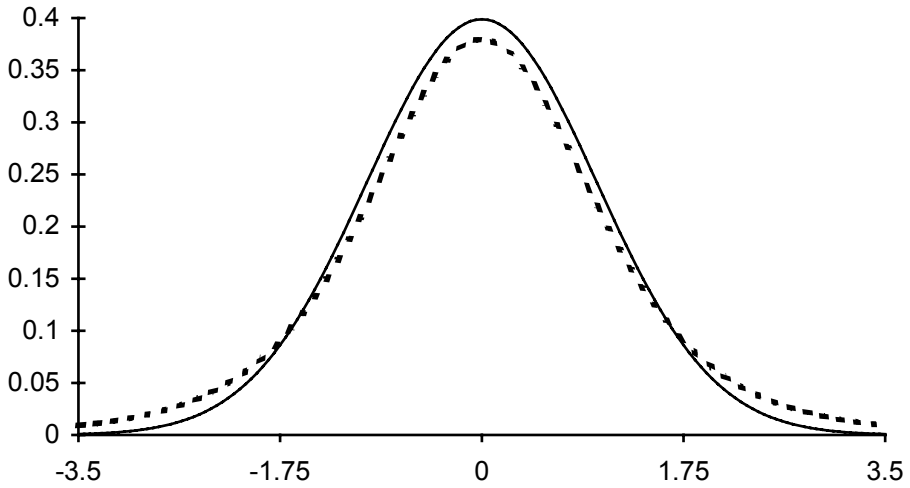


Figura 26. Confronto tra la densità di probabilità della v.c. t di Student con gdl 5 (linea tratteggiata) e la distribuzione normale corrispondente, con stessa media e stessa varianza (linea continua).

2.5.3 LA DISTRIBUZIONE F DI FISHER

Un'altra distribuzione di notevole interesse pratico, sulla quale è fondata l'inferenza di molta parte della statistica parametrica, è la **distribuzione F**.

Essa corrisponde alla distribuzione del

- **rapporto di 2 variabili casuali chi-quadrato indipendenti** (A e B), **divise per i rispettivi gradi di libertà** (indicata da m e da n).

$$F = (A/m) / (B/n)$$

$$f(F) = f_0\left(\nu_2 F^{\frac{\nu 1}{2}-1} + \nu_1 F^{-\frac{\nu 2}{2}-1} \right)$$

Questo **rapporto F è definito tra 0 e** $+\infty$.

La curva dipende sia dal valore di ν_1 e ν_2, tenendo conto delle probabilità α.

Di conseguenza, in quanto definita da tre parametri, la distribuzione dei valori di F ha tre dimensioni.

Il problema della rappresentazione dei valori di F è risolto praticamente con 2-4 pagine sinottiche, che riportano solo i valori più utilizzati, quelli che fanno riferimento in particolare alle probabilità 0.05, 0.01 e, più raramente, 0.005 e 0.001.

L'ordine con il quale sono riportati i due numeri che indicano i gradi di libertà è importante: la densità della distribuzione F non è simmetrica rispetto ad essi. Per convenzione, le tavole sono calcolate per avere F uguale o maggiore di 1.
Per primo si riporta sempre il numero di gradi di libertà del numeratore, che è sempre la varianza maggiore, e per secondo quello del denominatore, che è sempre la varianza minore.
Il valore di F in teoria può quindi variare da 1 a $+\infty$.
In realtà sono molto rari i casi in cui supera 10; avviene solo quando i gradi di libertà sono pochi.

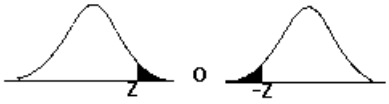
Storicamente,
- la distribuzione F è stata proposta dopo la distribuzione **t** e ne **rappresenta una generalizzazione**.

Tra esse esistono rapporti precisi:
- il quadrato di una variabile casuale t di Student con ν gradi di libertà è uguale ad una distribuzione F di Fisher con gradi di libertà 1 e ν.

$$\mathbf{t^2_{(\nu)} = F_{(1,\nu)}} \quad \text{oppure} \quad \mathbf{t_{(\nu)}} = \sqrt{F_{(1,\nu)}}$$

E' una relazione che sarà richiamata diverse volte nel corso, in particolare quando si tratterà di passare dal confronto tra più medie al confronto tra solo due.
Inoltre il test t di Student permette confronti unilaterali più semplici ed immediati, che in molti casi sono vantaggiosi rispetto a quelli bilaterali. Anche questi concett saranno sviluppati nella presentazione dei test d'inferenza.

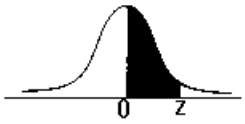
Aree in una coda della curva normale standardizzata



La tabella riporta la probabilità nell'area annerita

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0,1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0,2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0,3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0,4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0,5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0,6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0,7	0.242	0.239	0.236	0.233	0.230	0.227	0.224'	0.221	0.218	0.215
0,8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0,9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1,0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1,1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1,2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1,3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1,4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1,5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1,6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.048	0.046	0.046
1,7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.037	0.037
1,8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.030	0.029	0.029
1,9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2,0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2,1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2,2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2,3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2,4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2,5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2,6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2,7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2,8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2,9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3,0	0.001									

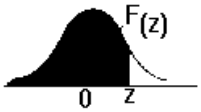
Valori della distribuzione normale standardizzata.



La parte annerita rappresenta l'area sottostante la distribuzione normale standardizzata dalla media aritmetica a z.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0.0	00000	00399	00792	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32';94	33147	33398	33646	33891
1,0	34134	34375	34614	34849	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46637	46712	46784	46856	46926	46995	47062
1.9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49745	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49991	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49995	49996	49996	49996	49996	49996	49997	49997

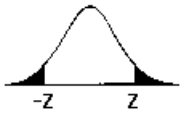
Valori dell'integrale di probabilità della distribuzione normale standardizzata



L'area annerita rappresenta la probabilità di ottenere un valore dello scarto standardizzato minore di z.

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0.5000	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0,1	0.5398	0.5438	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0,2	0.5792	0.5831	0.5870	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.6140
0,3	0.6179	0.6217	0.6255	0.6293	0.6330	0.6368	0.6405	0.6443	0.6480	0.6517
0,4	0.6554	0.6591	0.6627	0.6664	0.6700	0.6736	0.6772	0.6808	0.6843	0.6879
0,5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.7190	0.7224
0,6	0.7257	0.7290	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0,7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0,8	0.7881	0.7910	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0,9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1,0	0.8413	0.8437	0.8461	0.8485	0.8508	0.85.3	0.8554	0.8576	0.8599	0.8621
1,1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8790	0.8810	0.8829
1,2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1,3	0.9032	0.9049	0.9065	0.9082	0.9098	0.9114	0.9130	0.9146	0.9162	0.9177
1,4	0.9192	0.9207	0.9222	0.9236	0.9250	0.9264	0.9278	0.9292	0.9305	0.9318
1,5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.9440
1,6	0.9452	0.9463	0.9473	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9544
1,7	0.9554	0.9563	0.95.7	0.9581	0.9590	0.9599	0.9608	0.9616	0.9624	0.9632
1,8	0.9640	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1,9	0.9712	0.9719	0.9725	0.9732	0.9738	0.9744	0.9750	0.9755	0.9761	0.9767
2,0	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2,1	0.9821	0.9825	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9853	0.9857
2,2	0.9861	0.9864	0.9867	0.9871	0.9874	0.9877	0.9880	0.9884	0.9887	0.9889
2,3	0.9892	0.9895	0.9898	0.9901	0.9903	0.9906	0.9908	0.9911	0.9913	0.9915
2,4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2,5	0.9937	0.9939	0.9941	0.9943	0.9944	0.9946	0.9947	0.9949	0.9950	0.9952
2,6	0.9953	0.9954	0.9956	0.9957	0.9958	0.9959	0.9960	0.9962	0.9963	0.9964
2,7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9972	0.9973
2,8	0.9974	0.9975	0.9976	0.9976	0.9977	0.9978	0.9978	0.9979	0.9980	0.9980
2,9	0.9981	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3,0	0.9986	0.9986	0.9987	0.9987	0.9988	0.9988	0.9988	0.9989	0.9989	0.9990
3,1	0.9990	0.9990	0.9991	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992
3,2	0.9993	0.9993	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995
3,3	0.9995	01999	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996
3,4	0.9996	0.9996	0.9996	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997

Area nelle due code della distribuzione normale standardizzata



La tabella riporta le probabilità nelle aree annerite.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0.0	1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0.1	0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0.2	0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0.3	0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0.4	0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0.5	0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0.6	0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0.7	0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0.8	0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0.9	0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1.0	0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1.1	0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1.2	0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1.3	0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1.4	0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1.5	0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1.6	0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1.7	0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1.8	0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1.9	0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2.0	0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2.1	0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2.2	0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2.3	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.017	0.017
2.4	0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2.5	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2.6	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2.7	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2.8	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2.9	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3.0	0.003									