Capitolo IV

PRINCIPALI DISTRIBUZIONI TEORICHE

1. Distribuzioni continue e discontinue

Le misure di tendenza centrale o di posizione servono ad individuare il valore intorno al quale i dati sono raggruppati; se una distribuzione di dati dovesse essere descritta con un solo valore, questo è la misura più appropriata per sintetizzare l'insieme delle osservazioni. Esse, come prima indicazione, servono per dare la dimensione normale del fenomeno, ma la scelta della misura di tendenza centrale di una serie di dati dipende dalle caratteristiche della distribuzione del carattere e dal tipo di scala in cui sono espresse le misure.

I valori che una variabile può assumere sono compresi entro un *campo o intervallo di definizione* e, mentre alcune variabili possono assumere tutti i valori compresi tra l'estremo inferiore e l'estremo superiore del campo di definizione, altre variabili ne assumono solamente un numero finito e discontinuo. Per esempio, le misure che esprimono l'altezza di un individuo, il suo peso, il reddito, il tasso di natalità, eccetera, sono altrettante manifestazioni di *variabili continue* caratterizzate da valori compresi in certi intervalli di variazione al loro interno continui. Invece, altre variabili, come il numero dei semi di una certa pianta, il numero dei petali di certi fiori, il numero di figli nelle famiglie, il numero di pezzi meccanici prodotti, eccetera, sono per loro natura *discontinue o discrete*, perché esse possono assumere solo valori discontinui (isolati), cioè soltanto numeri interi e non frazioni di numero.

Il grafico di una distribuzione discontinua può essere tracciato sotto la forma di un *istogramma* o come un *diagramma di frequenza,* ma una distribuzione continua invece sarà rappresentata da una *curva*.

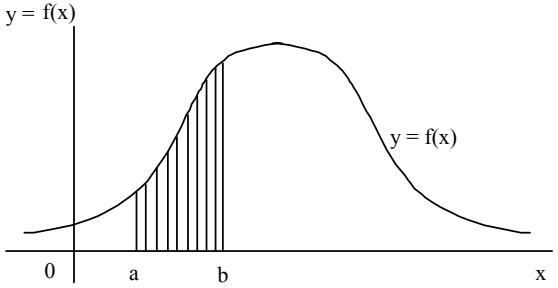
Ora, considerando una distribuzione continua della variabile x, possiamo definire una funzione f(x) tale che la probabilità di un qualunque valore di x compreso tra due numeri a e b, con a < b è data dall'integrale definito di f(x) tra i limiti a e b, quindi:

$$\int_a^b f(x).dx = P(a < x < b).$$

Da tale definizione segue che questa probabilità è uguale all'area tratteggiata che insiste sull'intervallo delimitato dalle ascisse a e b nel grafico che segue e che l'intera area sotto la curva di equazione y = f(x) è uguale a:

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

L'area complessiva racchiusa dalla curva y = f(x) e dall'asse delle ascisse descrive la probabilità che sia $-\infty < x < +\infty$, ma tale probabilità evidentemente coincide con la certezza. Inoltre, è $f(x) \geq 0$ e continua.



Quindi sono queste le condizioni che debbono soddisfatte affinché date funzioni possano essere scelte come modelli matematici delle distribuzioni dei valori osservati. In tal caso, la funzione f(x) è denominata "*funzione di frequenza o di densità della probabilità*".

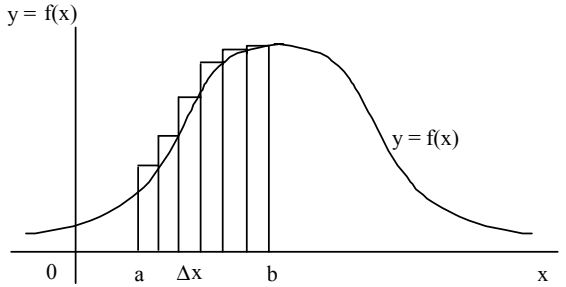
Poiché gli elementi statistici osservati in genere costituiscono successioni non continue (cioè discrete) di valori, la validità di quanto finora detto può essere mantenuta ragionando come segue.

Dal calcolo sappiamo che:

$$\lim_{\Delta x \to 0} \sum_{i=1}^{i=n} f_i.\Delta x = \int_a^b f(x).dx$$

cioè, la spezzata che figura nel grafico che segue e che rappresenta l'istogramma ottenibile con i dati osservati, via via che si riduce l'incremento Δx della variabile tenderà sempre più ad avvicinarsi alla curva di equazione y = f(x):

Mano a mano che Δx diventa sempre più piccolo, il numero dei rettangoli cresce senza limiti e la somma delle aree relative ai triangoli curvilinei che si trovano al di sopra della curva tende a zero, cosicché la somma delle aree dei rettangoli che hanno base sull'asse delle ascisse tende all'area sotto la curva compresa tra le ascisse a e b.



D'altra parte sappiamo che, se la base di ogni rettangolo si prende uguale all'unità, la somma delle aree dei rettangoli compresi tra i limiti a e b è uguale alla somma delle frequenze relative alle modalità dall'ascissa a all'ascissa b. Quindi, nel caso di distribuzioni continue, se f_i rappresenta la frequenza relativa della i-esima modalità del carattere esaminato, l'area tratteggiata compresa tra le ascisse a e b rappresenterà la probabilità che un qualsiasi valore della variabile x cada nell'intervallo compreso tra a e b .

A. DISTRIBUZIONI DISCRETE

2. Definizione della distribuzione binomiale

2.1 *Aspetti generali*

Quando si ha una distribuzione di frequenza formata da dati osservati spesso si può dedurre una distribuzione matematica di frequenza che servirà come se fosse un modello migliore della distribuzione data. In statistica, tra le distribuzioni teoriche correntemente usate per rappresentare serie osservate una delle principali è quella denominata "*distribuzionebinomiale*" o "*distribuzione di Bernoulli*", in onore del matematico svizzero J. Bernouilli (1654-1705), che ha fornito importanti contributi alla teoria della probabilità.

La binomiale è una distribuzione teorica discreta e finita, per eventi classificati con una *variabile binaria*, cioè caratteri che si manifestano solamente con due diverse modalità.

Le variabili casuali di tipo binario sono numerose: maschio/femmina, successo/insuccesso, malato/sano, inquinato/non inquinato, alto/basso, negativo/positivo. Si tenga conto, inoltre, che tutte le variabili, sia le multinomiali sia le continue, possono sempre essere ridotte ad una più semplice variabile *dicotomica*, sia pure con perdita di informazioni.

Ad esempio, una popolazione classificata in soggetti di specie diverse (A, B, C, D, E, ...) può sempre essere ricondotta ad una classificazione binaria (specie A, specie non-A). Come altro esempio,una serie di misure in scala discreta o continua, non importa se relative a ranghi, a intervalli o a rapporti, può sempre essere ricondotta ad una classificazione binaria del tipo: valori superiori (+) od inferiori (-) ad un limite prefissato.

Come si è visto nel secondo capitolo, ammettendo che p denoti la probabilità che si presenti un evento favorevole in una singola prova e ammettendo che sia q = 1 - p la probabilità dell'evento contrario, allora la probabilità che l'evento favorevole si presenti esattamente x volte in una serie di n prove indipendenti è data dalla espressione seguente:

$$P(x) = \frac{n!}{x!(n-x)!}.p^x.q^{n-x}.$$

Questa espressione costituisce una funzione che è chiamata "*funzione della distribuzione binomiale*". Essa deriva il suo nome dalla relazione seguente che esprime lo sviluppo della potenza n-esima di un binomio:

$$(p+q)^n = p^n + n.p^{n-1}.q + \frac{n(n-1)}{2!}.p^{n-2}.q^2 + \ldots + n.p.q^{n-1} + q^n =$$

$$= \sum_{x=0}^{x=n} \frac{n!}{x!(n-x)!}.p^x.q^{n-x} = \sum_{x=0}^{x=n} P(x).$$

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, si presenti 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.

La distribuzione binomiale è sovente utilizzata nella statistica non parametrica, perché molti tests ne fanno uso per il calcolo delle probabilità in piccoli campioni e perché in campioni con un numero di osservazioni sufficientemente alto è bene approssimata dalla distribuzione normale.

La distribuzione binomiale può essere oggetto di applicazione in molti problemi pratici nei quali si considerino numerose prove ripetute. Le frequenze teoriche corrispondenti ad x = 0, 1, 2, ..., n potranno essere ottenute considerando i termini successivi della seguente espressione, in cui N rappresenta il numero di repliche, ciascuna costituita di n prove:

$$N.(p+q)^n = N.p^n + N.n.p^{n-1}.q + N.\frac{n(n-1)}{2!}.p^{n-2}.q^2 + \ldots + N.n.p.q^{n-1} + N.q^n$$

Per esempio, si considerino 50 gruppi ciascuno di 4 lanci di una moneta perfetta. Il numero di casi in cui ci si può attendere di avere 0, 1, 2, 3 e 4 volte testa sono ottenuti dalla formula precedente nel modo seguente.

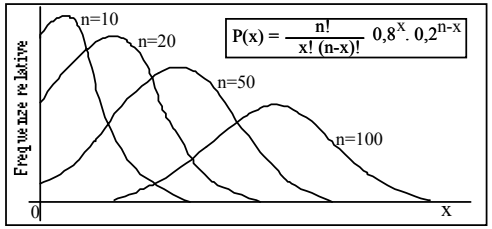
La probabilità di ottenere testa ad un lancio singolo è p = 1/2, mentre la probabilità di non ottenerla e, cioè, di ottenere croce è q = 1 - p = 1/2. Inoltre, abbiamo n = 4 ed N = 50. Ora se si applica la formula binomiale troviamo i numeri che rappresentano le frequenze teoriche attese e che corrispondono alla uscita di 0, 1, 2, 3 e 4 volte testa su 50 insiemi di 4 lanci ciascuno, con il seguente risultato.:

$$50\left(\frac{1}{2}+\frac{1}{2}\right)^4 = 50\left[\left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + \frac{4x3}{2x1}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 + \frac{4x3x2}{3x2x1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4\right] =$$

$$= 50.\frac{1}{2^4}(1+4+6+4+1) = \frac{50}{16}+\frac{200}{16}+\frac{300}{16}+\frac{200}{16}+\frac{50}{16} =$$

$$= 3,125 + 12,5 + 18,75 + 12,5 + 3,125 = 50$$

L'equazione della funzione binomiale mostra che la forma della curva binomiale dipende dai valori n, p e q. Per valori di n molto grandi, anche quando p e q sono molto diversi tra loro, la curva binomiale assume una forma approssimativamente simmetrica, ma quando n è piccolo e p e q sono molto diversi tra loro, la curva appare asimmetrica, con asimmetria tanto maggiore quanto più diversi sono i valori di p e di q e quanto più piccolo è il numero delle osservazioni n.



Nella figura riportata viene mostrato come si modifica approssimativamente la forma della curva binomiale per diversi valori di n nel caso in cui, per esempio, si abbia p = 0,8 e q = 0,2.

Tuttavia, man mano che i valori di p e di q si avvicinano l'un l'altro, la distribuzione binomiale diviene via via più simmetrica e, nel caso in cui p = q, la curva mostra una simmetria perfetta.

Tutte queste annotazioni sono utilissime in sede di determinazione della distribuzione che si vuole applicare ai dati osservati.

2.2 *Momenti della distribuzione binomiale*

2.2.1 Media aritmetica

Il *primo momento rispetto all'origine* è definito come segue:

$$M = \frac{1}{N}\sum_{i=0}^{i=n} x_i . f_i = \sum_{i=0}^{i=n} x_i . P(x_i)$$

e poiché

$$P(x) = \frac{n!}{x!(n-x)!}.p^x.q^{n-x}$$

si avrà che la media aritmetica è uguale a:

$$M = \sum_{x=0}^{x=n} \frac{n!}{x!(n-x)!}.p^x.q^{n-x}.x =$$

$$= np.\sum_{x=0}^{x=n-1} \frac{(n-1)!}{(x-1)!(n-x)!}.p^{x-1}.q^{n-x} =$$

$$= np.(p+q)^{n-1} = n.p \qquad \text{per } (p+q)=1.$$

La media aritmetica della distribuzione binomiale è quindi data dalla relazione M = n.p .

2.2.2 Deviazione standard

La deviazione standard non è altro che la radice quadrata del momento secondo rispetto alla media ossia della varianza.

Per la determinazione del *momento secondo rispetto all'origine* valgono le seguenti relazioni:

$$M_2 = \sum_{i=0}^{i=n} x_i^2 . P(x_i) = \sum_{x=0}^{x=n} x^2 . \frac{n!}{x!(n-x)!} . p^x . q^{n-x} =$$

$$= \sum_{x=0}^{x=n} \left[x + x(x-1)\right] \frac{n!}{x!(n-x)!} . p^x . q^{n-x} =$$

$$= \sum_{x=0}^{x=n} x . \frac{n!}{x!(n-x)!} . p^x . q^{n-x} + \sum_{x=0}^{x=n} x(x-1) . \frac{n!}{x!(n-x)!} . p^x . q^{n-x} =$$

$$= np + n(n-1)p^2 . \sum_{x=0}^{x=n-2} \frac{(n-2)!}{(x-2)!(n-x)!} . p^{x-2} . q^{n-x} =$$

$$= np + n(n-1)p^2 (p+q)^{n-2} = np + n(n-1)p^2 = np + n^2 p^2 - np^2$$

Ma, poiché la varianza o momento secondo rispetto alla media, se espressa in funzione dei momenti rispetto all'origine, è data dalla relazione:

$\mu_2 = \sigma^2 = M_2 - M^2 = (np + n^2 p^2 - np^2) - n^2 p^2 = np - np^2 = np(1-p) = npq$

come risultato si ottiene per la deviazione standard la relazione:

$$\sigma = \sqrt{\mu_2} = \sqrt{npq} .$$

Nella distribuzione binomiale la varianza è inferiore alla media; infatti essa è uguale alla media n.p moltiplicata per un numero che è inferiore all'unità [q=(1-p)].

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

Se i dati sono espressi non come numero di successi o frequenze assolute, come finora abbiamo fatto, ma come proporzioni o frequenze relative, allora media e varianza possono essere calcolati in base alle seguenti relazioni:

$$\mu = p \; ; \sigma^2 = \frac{p.q}{n} ; \qquad \sigma = \sqrt{\frac{pq}{n}} .$$

I momenti di ordine maggiore possono essere trovati con ragionamenti simili.

2.2.3 Momento terzo e misure di asimmetria

Il *terzo momento rispetto all'origine* è definito dalla relazione:

$$M_3 = \sum_{i=0}^{i=n} x_i^3 . P(x_i) = \sum_{x=0}^{x=n} x^3 . \frac{n!}{x!(n-x)!} . p^x . q^{n-x}$$

e tenendo presente che vale la uguaglianza $x^3 = x[x+x(x-1)]$, con gli opportuni passaggi si ha anche:

$$M_3 = np + 3n(n-1)p^2 + n(n-1)(n-2)p^3$$

per cui il *terzo momento rispetto alla media* risulterà essere:

$$\mu_3 = M_3 - 3 . M_2 . M_1 + 2 . M_1^3 = npq(q-p).$$

Per misurare l'asimmetria possiamo usare l'indice α_3 che risulta uguale a:

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{npq(q-p)}{npq\sqrt{npq}} = \frac{q-p}{\sqrt{npq}} .$$

Il quadrato dell'indice α_3 è generalmente indicato dalla lettera β_1 e rappresenta il *coefficiente di asimmetria del Pearson*; esso sarà quindi definito dalla relazione:

$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq} .$$

Questa espressione pone a confronto il quadrato del terzo momento con il cubo del secondo momento ambedue rispetto alla media, come si è visto nel capitolo III, paragrafo 6. Poiché tale espressione ha segno positivo risulta, dunque, che la distribuzione binomiale ha asimmetria positiva, cioè il ramo della curva situato a destra della moda è più allungato del ramo che si trova a sinistra della stessa moda.

2.2.4 Momento quarto e misure di appiattimento

Il *quarto momento rispetto all'origine* è definito dalla espressione seguente:

$$M_4 = \sum_{i=0}^{i=n} x_i^4 . P(x_i) = \sum_{x=0}^{x=n} x^4 . \frac{n!}{x!(n-x)!} . p^x . q^{n-x}$$

e tenendo presente che valgono le relazioni:

$$x^4 = x(x-1)(x-2)(x-3) + 6x^3 - 11x^2 + 6x$$

e

$$\mu_4 = M_4 - 4.M_3.M_1 + 6.M_2.M_1^2 - 3.M_1^4$$

il *quarto momento rispetto alla media* si può scrivere:

$$\mu_4 = 3n^2p^2q^2 + npq(1-6pq) = n^2p^2q^2\left[3 + \frac{(1-6pq)}{npq}\right] = \mu_2^2\left[3 + \frac{(1-6pq)}{npq}\right]$$

ed inoltre come misura dell'appiattimento si avrà:

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = \frac{1}{(npq)^2}\left[3n^2p^2q^2 + npq(1-6pq)\right] = 3 + \frac{(1-6pq)}{npq}$$

Come si è visto nel capitolo III, paragrafo 7, l'indice α_4 viene indicato spesso anche dalla lettera β_2 e rappresenta *il coefficiente di*

appiattimento del Pearson. Esso è definito dalla relazione ottenuta rapportando tra loro il momento quarto rispetto alla media ed il quadrato della varianza:

$$\beta_2 = \alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = 3 + \frac{(1-6pq)}{npq}.$$

Ricordando che per la curva rappresentatrice della distribuzione normale questo indice assume il valore 3 si vede che la curva binomiale è più aguzza della curva normale o, come si dice, è *leptokurtica* quando è p.q < 1/6, mentre è più appiattita della curva normale, cioè *platikurtica*, quando è p.q > 1/6.

Esempio IV.1

Come applicazione consideriamo il seguente esempio.

In un villaggio viene eseguita un'inchiesta sul numero dei campi che sono gestiti direttamente dal proprietario. Da precedenti esperienze si sa che la proporzione dei campi gestiti direttamente dal proprietario è pari a 2/3. Vengono estratti 100 campioni di 6 campi ciascuno con i seguenti risultati relativi al numero dei campi che risultano gestiti dal proprietario:

Numero di campi gestiti dal proprietario sui 6 campi esaminati	Proporzione dei campi gestiti dal proprietario	Numero di campioni che danno tale risultato
0	0	1
1	1/6	2
2	2/6	10
3	3/6	22
4	4/6	35
5	5/6	24
6	6/6	6
		100

Adattare ai suddetti dati una distribuzione binomiale e confrontare le frequenze teoriche calcolate con i risultati ottenuti dalla rilevazione campionaria. Inoltre, trovare la media e la deviazione standard di questa distribuzione e tracciare un grafico che mostri le frequenze osservate e quelle teoriche.

In questo esempio, p = 2/3, q = 1/3, n = 6 ed N = 100. La tavola che segue mostra le frequenze teoriche trovate calcolando i termini nell'espansione del binomio:

$$100\left(\frac{2}{3}+\frac{1}{3}\right)^6 = 100\left[\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)^0 + 6\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right) + 15\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^2 + 20\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^3 + \right.$$
$$\left. + 15\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^4 + 66\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^5 + \left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^6\right] =$$

$$= \frac{100}{3^6}(64 + 6x32 + 15x16 + 20x8 + 15x4 + 6x2 + 1) =$$

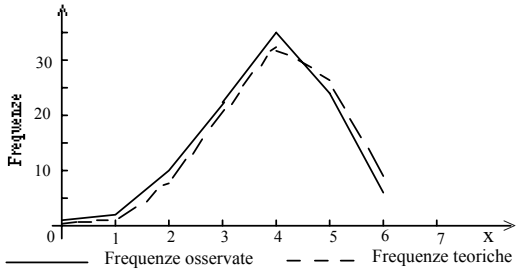
$$= 0,137174211(64 + 192 + 240 + 160 + 60 + 12 + 1) =$$

$$= 8,779 + 26,338 + 32,922 + 21,948 + 8,230 + 1,646 + 0,137 = 100$$

Avremo, quindi la tavola di confronto seguente:

x	Frequenze osservate	Frequenze teoriche
0	1	0,137
1	2	1,646
2	10	8,230
3	22	21,948
4	35	32,922
5	24	26,338
6	6	8,779
Totale	100	100,000

La rappresentazione grafica dei dati suddetti è, quindi, la seguente:



I momenti, la media e la deviazione standard della distribuzione osservata sono calcolati in base alla tabella che segue:

x	f	xf	x^2f
0	1	0	0
1	2	2	2
2	10	20	40
3	22	66	198
4	35	140	560
5	24	120	600
6	6	36	216
	100	384	1616

per cui sarà:

M = 384/100 = 3,84; M_2 = 1616/100 = 16,16

$$s^2 = M_2 - M^2 = 16,16 - (3,84)^2 = 16,16 - 14,75 = 1,41$$

$$s = \sqrt{1,41} = 1,187.$$

La media e la deviazione standard dei valori teorici, invece, saranno:

M = np = 6. 2/3 = 4 , $s = \sqrt{6.2/3.1/3} = \sqrt{4/3} = 1,153$.

3. 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 ad 1, che possono comparire nel corso di N prove, indipendenti, successive o simultanee, al fine di calcolare la *probabilità complessiva che si presenti la combinazione di tutti gli eventi scelti nelle proporzioni stabilite*.

Se X_1, X_2, ..., X_k sono variabili casuali che indicano rispettivamente il numero delle volte n_1, n_2, ..., n_k in cui i k eventi indipendenti si verificano in *n* prove, in modo che $X_1 + X_2 + ... + X_k = n$, la probabilità che si verifichino tutti gli eventi indicati è determinata dallo sviluppo del multinomio:

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

Non esiste un limite al numero di fattori da considerare insieme; la probabilità con cui ognuno di essi può presentarsi varia da 0 a 1, considerando che il totale di tutte le singole probabilità alternative p_1, p_2, ..., p_i, ..., p_n è sempre uguale all'unità.

Esempio IV.2

Se un dado non truccato viene lanciato 12 volte, la probabilità di ottenere 1, 2, 3, 4, 5 e 6 esattamente due volte ciascuno è

$$P = \frac{12!}{2!2!2!2!2!2!}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2 = \frac{1.925}{559.872} = 0,00344$$

Il numero probabile di volte in cui ciascuno dei *k* eventi indipendenti si può verificare in *n* prove è definito, per ogni evento dalle relazioni:

$$E(X_1)=np_1;\ E(X_2)=np_2;\ .....;\ E(X_k)=np_k\ .$$

La distribuzione multinomiale ha un uso limitato, circoscritto alla stima della probabilità complessiva di più eventi *concomitanti e indipendenti*. La sua applicazione è quindi tipica per i fenomeni che si presentano con più di due modalità distinte, per ciascuna delle quali esiste una determinata probabilità di manifestarsi in un esperimento campionario.

4. Distribuzione di Poisson

4.1 *Formulazione della poissoniana*

La funzione binomiale si approssima alla curva normale per valori molto grandi di *n*. Tuttavia, se *p* è molto piccolo anche se *n* è molto grande, la approssimazione alla forma normale fornisce un adattamento molto scarso. Inoltre, in questo caso l'uso della binomiale presenta vari inconvenienti pratici dato che l'innalzamento di frequenze molto basse a potenze elevate ed il calcolo di fattoriali per numeri grandi rendono il calcolo manuale praticamente impossibile.

Quando *n* tende all'infinito (cioè il numero dei dati è molto grande), se *p* si avvicina allo zero (cioè la probabilità che l'evento ha di verificarsi è molto piccola) in modo che il prodotto *n.p* rimanga costante, allora una buona approssimazione alla curva binomiale è possibile tramite l'uso della seguente forma analitica chiamata *funzione di distribuzione di Poisson*:, secondo la quale

$$P(x)=\frac{e^{-M}.M^x}{x!}\ \text{se}\begin{cases} n\to\infty \\ p\to 0 \\ n.p=\cos t. \end{cases}.$$

Essa fu individuata per primo dal matematico francese S. D. Poisson (1781-1840) nel 1837 e fu applicata da Bortkiewicz a dati riferiti ad un periodo di 20 anni, che esprimevano il numero di uomini dell'Armata Prussiana che, in ogni anno del periodo e per ognuno dei 10 Corpi di appartenenza, erano stati uccisi dal calcio di un cavallo. Questa applicazione della funzione di Poisson ad un caso in cui la probabilità di verificarsi dell'evento è molto piccola (probabilità che un uomo muoia per il calcio di un cavallo) è divenuta classica.

Infatti, la funzione di distribuzione di Poisson fornisce spesso un adattamento molto buono in molti problemi pratici in cui si consideri il verificarsi di un *evento raro*. Ad esempio, le manifestazioni patologiche particolarissime nell'uomo e negli animali (la distribuzione del numero di morti provocati da una malattia rara), gli eventi catastrofici in natura (il numero di inondazioni in un paese) oppure negli aggregati sociali o naturali (il numero di microrganismi di una certa specie che si vedono in una unità di superficie o di volume) sono altrettanti esempi di eventi poco frequenti o addirittura rari.

4.2. *La poissoniana come approssimazione della binomiale*

La funzione di distribuzione di Poisson può essere vista come approssimazione di una funzione di probabilità binomiale per *n* tendente ad infinito e *p* prossimo allo zero in modo che il prodotto *np* rimanga costante.

Infatti, si consideri l'espressione binomiale:

$$P(x)=\frac{n!}{x!(n-x)!}.p^x.q^{n-x} \qquad (1)$$

Sia *n.p = a*, ossia *p = a/n.* e *q = (1 - a/n)* . Sostituendo *p* e *q* nella formula (1) con le quantità indicate, si ottiene:

$$P(x)=\frac{n!}{x!(n-x)!}.\left(\frac{a}{n}\right)^x.\left(1-\frac{a}{n}\right)^{n-x}=$$
$$=\frac{n(n-1)(n-2)....(n-x+1)}{n^x}.\frac{a^x}{x!}.\left(1-\frac{a}{n}\right)^{n-x} \qquad (2)$$

Ma, secondo l'analisi matematica, valgono le seguenti relazioni:

$$\underset{n\to+\infty}{\text{Limite}}\ \frac{n(n-1)(n-2)\ldots(n-x+1)}{n^x}=1 \qquad (3a)$$

ed anche:

$$\underset{n\to+\infty}{\text{Limite}}\left(1-\frac{a}{n}\right)^{n-x}=e^{-a} \qquad (3b)$$

E' noto che il numero e è definito come il limite della espressione $(1 + 1/n)^n$ quando n tende all'infinito. Per cui, in base a questa definizione possiamo anche scrivere le seguenti relazioni:

$$\underset{n\to+\infty}{\text{Limite}}\left(1+\frac{1}{n}\right)^n=e=2,718281829\ldots\ldots$$

$$\underset{n\to+\infty}{\text{Limite}}\left(1+\frac{a}{n}\right)^n=e^a \qquad (a = \text{costante})$$

$$\underset{n\to+\infty}{\text{Limite}}\left(1-\frac{a}{n}\right)^n=e^{-a}$$

D'altro canto il numero e^a può anche essere espresso sotto forma di una serie infinita come segue:

$$e^a=1+a+\frac{a^2}{2!}+\frac{a^3}{3!}+\ldots\ldots+\frac{a^k}{k!}+\ldots\ldots$$

In particolare, per $a = 1$ avremo:

$$e^a=1+1+\frac{1}{2!}+\frac{1}{3!}+\ldots\ldots+\frac{1}{k!}+\ldots\ldots=2,718281829$$

Questo numero rappresenta la base dei logaritmi naturali ed ha molte ed importanti proprietà che ne determinano un largo uso nella teoria matematica, per i vantaggi che si possono trarre dalla semplicità del suo trattamento nelle questioni di natura formale, tanto dell'analisi algebrica quanto di quella infinitesimale.

Ora tornando alla formula (2), se consideriamo i limiti definiti dalle relazioni (3a) e (3b), è possibile scrivere la formula (2) come segue:

$$P(x)=\frac{a^x}{x!}.e^{-a}$$

Da questa espressione si vede che *la distribuzione di Poisson è una distribuzione teorica discreta che dipende o è totalmente definita da un solo parametro* e precisamente da *a*, che, come si vedrà nel paragrafo seguente, non è altro che la *media aritmetica* della distribuzione di Poisson.

L'approssimazione di Poisson è buona per grandi valori di n e piccoli valori di p. Spesso essa si considera una buona approssimazione quando il valore di p calcolato da un dato insieme di dati è $p < 0,03$ o quando $np < 5$.

Per questa ragione tale tipo di funzione viene considerata in particolare per studiare il manifestarsi di *eventi rari*, in quanto sono assai più frequenti le classi con zero eventi o con pochi eventi rispetto alle classi con numerosi eventi. Essa viene anche chiamata *legge dei piccoli numeri*, perchè la frequenza assoluta di questi eventi è espressa da un numero piccolo, anche in un numero elevato di prove.

4.3 *Momenti della distribuzione di Poisson*

4.3.1 Media o primo momento rispetto all'origine

In base alla definizione di media aritmetica, se la frequenza è definita da una distribuzione poissoniana si avrà:

$$M=\sum_{x_i=o}^{x_i=+\infty}x_i.P(x_i)=\sum_{x_i=o}^{x_i=+\infty}x_i.\frac{a^{x_i}}{x_i!}.e^{-a}= \qquad [\text{in cui } P(x_i)=\frac{a^{x_i}}{x_i!}.e^{-a}]$$

$$=e^{-a}\left(0+a+2\frac{a^2}{2!}+3\frac{a^3}{3!}+4\frac{a^4}{4!}+\ldots\ldots\right)=$$

$$=a.e^{-a}\left(1+a+\frac{a^2}{2!}+\frac{a^3}{3!}+\frac{a^4}{4!}+\ldots\ldots\right)=a.e^{-a}.e^a=a$$

Quindi, poiché l'unico parametro della distribuzione di Poisson risulta essere $a = M$, la funzione prende la forma:

$$P(x) = \frac{e^{-M}.M^x}{x!}$$

Si conclude, quindi, che *la funzione di distribuzione di Poisson dipende da un solo parametro che coincide con la media aritmetica della distribuzione*.

4.3.2 Relazione tra M_2 e σ

Vediamo ora come sono definiti e in che relazione stanno tra loro il momento secondo rispetto all'origine (M_2) e la deviazione standard (σ) di una distribuzione di Poisson.

Il momento secondo rispetto all'origine si otterrà, come si è già detto per la media aritmetica, con l'espressione:

$$M_2 = \sum_{x_i=o}^{x_i=+\infty} x_i^2 . P(x_i) = \sum_{x_i=o}^{x_i=+\infty} x_i^2 . \frac{a^{x_i}}{x_i!} . e^{-a} =$$

$$= e^{-a}\left(0 + a + 2^2\frac{a^2}{2!} + 3^2\frac{a^3}{3!} + 4^2\frac{a^4}{4!} + \ldots\ldots\right) = = a.e^{-a}\left(1 + 2a + 3\frac{a^2}{2!} + 4\frac{a^3}{3!} + 5\frac{a^4}{4!} + \ldots\ldots\right) =$$

$$= a.e^{-a}\left(1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \frac{a^4}{4!} + \ldots\ldots\right) + a.e^{-a}\left(a + 2\frac{a^2}{2!} + 3\frac{a^3}{3!} + 4\frac{a^4}{4!} + \ldots\ldots\right) =$$

$$= a + a^2.e^{-a}\left(1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \frac{a^4}{4!} + \ldots\ldots\right) = a + a^2$$

Se, inoltre, si considera che la varianza o momento secondo rispetto alla media è uguale alla differenza tra il momento secondo rispetto all'origine ed il quadrato del momento primo, si avrà che *la varianza della distribuzione di Poisson è esattamente uguale alla media della distribuzione stessa*, cioè:

$$\mu_2 = \sigma^2 = M = a.$$

La deviazione standard della distribuzione di Poisson può essere ottenuta con facilità come radice quadrata della varianza. Quindi, *la deviazione standard della distribuzione di Poisson è uguale alla radice quadrata della media aritmetica della distribuzione*.

Essa può essere determinata anche cercando la forma limite della deviazione standard della funzione binomiale man mano che n tende all'infinito ed il prodotto $np = M$. Infatti, quando p tende a zero e $q=(1 - p)$ tende all'unità, si avrà per risultato che la deviazione standard tende alla radice quadrata della media, cioè vale la relazione:

$$\sigma = \sqrt{npq} = \sqrt{np} = \sqrt{M} = \sqrt{a}.$$

4.3.3 Momento terzo e misura dell'asimmetria

Per il calcolo del terzo momento sia rispetto all'origine sia rispetto alla media si può seguire lo stesso procedimento già mostrato per il calcolo del momento secondo. Una volta ottenuti i risultati, si potrà calcolare anche l'asimmetria della distribuzione ricorrendo agli indici già noti. Per p tendente a zero, $np = M$, q tendente all'unità ed n tendente all'infinito si hanno le seguenti misure di asimmetria:

$$\alpha_3 = \underset{n\to+\infty}{\text{Limite}} \frac{q - p}{\sqrt{npq}} = \frac{1}{\sqrt{M}}$$

$$\beta_1 = \alpha_3^2 = \frac{1}{M}$$

La distribuzione di Poisson ha una forma molto asimmetrica per valori piccoli della media aritmetica, come sono considerati quelli inferiori a 3; ma una media uguale a 7 è già considerata grande e determina una distribuzione delle probabilità di ricorrenza dei vari eventi che tende ad essere simmetrica ed è bene approssimata dalla distribuzione normale o gaussiana.

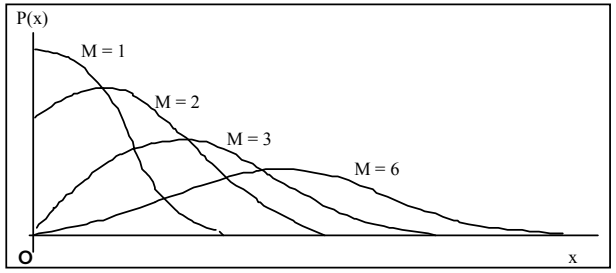
4.3.4 Momento quarto e misura dell'appiattimento

Pure per i momenti del quarto ordine può ripetersi quanto detto per i momenti di ordine inferiore ed una volta che siano noti si potranno calcolare gli indici di appiattimento. Per p tendente a zero, $np = M$, q tendente all'unità ed n tendente all'infinito l'appiattimento sarà misurato dal seguente valore dell'indice, che definisce la poissoniana come curva leptokurtica:

$$\alpha_4 = \underset{n\to+\infty}{\text{Limite}}\left(3+\frac{1-6pq}{npq}\right) = 3+\frac{1}{M} = \beta_2$$

4.4 *Forma generale del poligono di Poisson*

La forma del *poligono di Poisson* cambia al variare di M e di x. Per bassi valori di M il poligono assume una forma molto asimmetrica, ma man mano che M diventa più grande, essa diviene più simmetrica. Il grafico che segue può dare una idea generale della forma del poligono di Poisson per valori diversi della media:



Il valore dell'espressione:

$$P(x) = \frac{e^{-M}.M^x}{x!}$$

corrispondente a diversi valori di x e di M può essere trovato in molte tavole statistiche. Le frequenze teoriche sono ottenute moltiplicando queste probabilità per N.

Le singole frequenze possono anche essere calcolate rapidamente derivando ciascuna frequenza dalla precedente, come sarà mostrato nella soluzione dell'esempio che segue.

Esempio IV.3

Viene qui riportata la distribuzione di frequenza dei decessi di mucche per una malattia rara in 50 provincie durante un periodo di 10 anni:

Numero di mucche decedute	Numero di provincie in 10 anni
0	240
1	150
2	60
3	25
4	17
5	8
Totale	500

Su questi dati calcoliamo le frequenze della distribuzione di Poisson che ha la stessa media e compariamo i risultati con le frequenze osservate.

La tavola che segue mostra i calcoli da eseguire:

x	f	xf	P(x)	N.P(x)
0	240	0	0,4044	202,20
1	150	150	0,3664	183,20
2	60	120	0,1660	83,00
3	25	75	0,0501	25,05
4	17	68	0,0113	5,65
5	8	40	0,0020	1,00
	500	453		

$np = M = 453 / 500 = 0,906$; $p = 0,906 / 500 = 0,0018 < 0,03$

$\text{Log } e^{-0,906} = - 0,906 \,.\, 0,434 = - 0,39321$; $e^{-0,906} = 0,4044$

$P_0 = 0,4044$; $P_1 = 0,4044 \,.\, 0,906 = 0,3664$

$P_2 = 0,3664 \,.\, 0,906 / 2 = 0,1660$; $P_3 = 0,1660 \,.\, 0,906 / 3 = 0,0501$

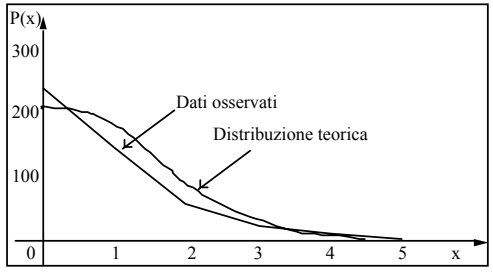
$P_4 = 0,0501 \,.\, 0,906 / 4 = 0,0113$; $P_5 = 0,0113 \,.\, 0,906 / 5 = 0,0020$

Se ora vogliamo misurare l'adattamento della curva poissoniana ai dati osservati e, nello stesso tempo, vogliamo anche valutarne l'attendibilità, possiamo calcolare una media quadratica degli scarti tra i valori osservati ed i valori teorici ottenuti, operando nel modo indicato nella tabella che segue:

x	f	N.P(x)	Δ=[N.P(x) - f]	Δ^2	Δ^2/N.P(x)
0	240	202,20	- 37,8	1428,84	7,0665
1	150	183,20	+ 33,2	1102,24	6,0166
2	60	83,00	+ 23,0	529,00	6,3735
3	25	25,05	0	0	0
4	17	5,65	- 11,4	129,96	23,2071
5	8	1,00	- 7,0	49,00	49,0000
	500	500,00	0	3239,04	91,6637

Come si vede, se si considera il test del χ^2 (Cfr. capitolo IV, paragrafo 13) per misurare la bontà dell'adattamento, si ha un valore di $\chi^2 = 91,6637$ cioè un valore che può essere superato per effetto del caso solo con una probabilità molto piccola. Ciò vuol dire che l'adattamento ottenuto con la funzione di Poisson non è molto buono e che si dovrebbe procedere alla selezione di una funzione di distribuzione di altro tipo per rappresentare il fenomeno studiato.

Rappresentando in un grafico i risultati ottenuti per i valori teorici e, insieme ad essi, anche i dati osservati si ottiene la distribuzione di Poisson adattata al diagramma di frequenza dato, che si riporta nel grafico seguente:



5. Distribuzione ipergeometrica

Nella distribuzione binomiale la probabilità di un evento si mantiene sempre costante. Quando essa varia in funzione degli eventi precedenti, come succede nell'estrazione *senza ripetizione* di alcuni oggetti da un campione di piccole dimensioni, si ha la *distribuzione ipergeometrica*.

Un esempio semplice è fornito dal gioco delle carte, con il calcolo delle probabilità nell'estrazione di un secondo re da un mazzo di 40 carte. Se il gioco avviene con reimmissione della carta già estratta, la probabilità di estrarre un re è costantemente pari a 4/40. Ma se il gioco si svolge senza reintroduzione nel mazzo della carta estratta per prima, la probabilità che la seconda carta sia un re varia in rapporto all'estrazione della prima: se la prima era un re, la probabilità per la seconda estrazione è pari a 3/39; se la prima era una carta diversa, la probabilità che la seconda sia un re è 4/39.

In quest'ultima condizione, le probabilità dei vari eventi sono stimate dalla distribuzione ipergeometrica, che viene espressa come rapporto tra combinazioni:

$$P_{(r,n)} = \frac{C^{r}_{n_1} \cdot C^{n-r}_{N-n_1}}{C^{n}_{N}} = \frac{\binom{n_1}{r}\binom{N-n_1}{n-r}}{\binom{N}{n}}$$

in cui N è un numero intero positivo che rappresenta il numero totale degli elementi della popolazione; n è un numero intero non negativo ed al massimo uguale ad N, che rappresenta il numero totale degli elementi del campione; n_1 è un intero positivo che al massimo è uguale ad N e che rappresenta il numero degli individui del campione che posseggono la caratteristica che si vuole esaminare; r è il numero degli individui che presentano la caratteristica in esame tra quelli estratti.

La distribuzione ipergeometrica è definita da *tre parametri* (N, n_1, n) che rappresentano nell'ordine: il numero totale di unità che formano la popolazione, il numero di unità del gruppo considerato, il numero delle unità estratte) *in funzione di r* (il numero di unità estratte appartenenti al gruppo considerato).

Se poniamo $p=n_1/N$ e $q = (N-n_1)/N$ la probabilità suddetta potrà scriversi come segue:

$$P_{(r,n)} = \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}}$$

La media e la varianza di questa distribuzione sono, rispettivamente:

$$M = np \qquad e \qquad \sigma^2 = \frac{npq(N-n)}{N-1}.$$

Quindi la media è uguale a quella della distribuzione binomiale corrispondente, mentre la varianza è inferiore. La distribuzione ipergeometrica, per *N* tendente ad infinito, ossia per *N* grande relativamente ad *n*, tende alla distribuzione binomiale.

Esempio IV.4

Un'urna contiene N biglie, delle quali n_1 bianche e N-n_1 nere. Si estraggono dall'urna n biglie (con n≤N) senza reintroduzione; si vuole determinare la probabilità che delle n biglie estratte r siano bianche (con r≤n).

Delle N biglie, n possono essere estratte in $\binom{N}{n}$ modi differenti; delle n_1 biglie bianche, r possono essere estratte in $\binom{n_1}{r}$ modi differenti; delle (N-n_1) biglie nere, (n-r) possono essere estratte in $\binom{N-n_1}{n-r}$ modi differenti. Ognuna delle $\binom{n_1}{r}$ possibilità di estrazione delle biglie bianche si combina con ognuna delle $\binom{N-n_1}{n-r}$ possibilità di estrazione delle biglie nere. La soluzione si trova in base alla relazione:

$$P_{(r/n)} = \frac{\binom{n_1}{r}\binom{N-n_1}{n-r}}{\binom{N}{n}} = \frac{n_1!.(N-n_1)!.n!.(N-n)}{r!.(n_1-r)!.(n-r)!.(N-n_1-n+r)!.N!}$$

6. Distribuzione binomiale negativa

Tra le distribuzioni teoriche discrete ha una certa importanza, particolarmente nella ricerca ambientale, la distribuzione *binomiale negativa*, la quale si utilizza nelle circostanze che sono qui appresso descritte.

La distribuzione binomiale, nella quale è *(p+q)=1,* con *p* che indica la probabilità di verificarsi di un certo evento e *q* quella del verificarsi dell'evento contrario, con *n* prove le probabilità dei diversi eventi sono determinate dallo sviluppo dl binomio $(p+q)^n=1$. Tale distribuzione, in particolare, è caratterizzata dall'avere la varianza *npq* inferiore alla media *np*, dato che è *q<1*.

La distribuzione binomiale negativa si applica quando la varianza ha un valore superiore a quello della media; in una distribuzione di dati sperimentali, che dovrebbe seguire la distribuzione binomiale positiva o quella poissoniana, può risultare che la varianza *npq* sia superiore alla media *np*, dal ché si deduce quanto segue:

- poiché *npq>np* deve essere *q>1*; necessariamente *p (=1-q)* ha un valore negativo;

- con *p* negativo, poiché la media (*np*) deve essere positiva in quanto dipendente da un conteggio di eventi, anche *n* è negativo. Ma *n* è un conteggio e non può essere negativo. Quindi, ponendo *n = -k,* con *k* intero positivo, la potenza del binomio assume la forma

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

Come quella binomiale e quella poissoniana, la distribuzione binomiale negativa si presta per calcolare le probabilità di eventi misurati mediante conteggio.

Negli studi di epidemiologia, ad esempio, è usata nell'analisi della letalità di una malattia, per stimare quanti sono i periodi (giorni, settimane, o mesi) con 0 morti, 1 morto, 2 morti e le frequenze successive, che sono ovviamente contati con numeri interi positivi. Nella ricerca ambientale è usata nei conteggi delle popolazioni animali, come, ad esempio, quando nella distribuzione di afidi, si contano le foglie che hanno la presenza di 0 animali, di 1 animale, di 2 animali, ecc.

Per *n* grande e probabilità basse, quando la media è unica le frequenze attese sono fornite dalla distribuzione poissoniana, ma quando il fenomeno è complesso, la distribuzione di frequenza è spesso determinata da due o più fattori, ognuno con media diversa, ne deriva che la variabilità aumenta e la

varianza è superiore alla media. In questi casi, la distribuzione delle frequenze può essere stimata in modo appropriato dalla distribuzione binomiale negativa.

Il numero *n* di prove bernoulliane necessarie per ottenere *x* successi e fornito dalla relazione

$$P(x) = \binom{n+x-1}{x} p^n q^x, \text{ con } x=0, 1, 2, ..., n$$

La media e la varianza della distribuzione binomiale negativa sono, rispettivamente

$$M = \frac{nq}{p} \text{ e } \sigma^2 = \frac{nq}{p^2} = \frac{M}{p}.$$

Questa distribuzione, nel caso n=1 si riduce alla distribuzione geometrica, la cui probabilità è definita dalla relazione

$$P(x) = pq^x \text{ con } x=0, 1, 2, ...$$

7. Distribuzione uniforme

La distribuzione uniforme è la più semplice delle distribuzioni discrete. La sua caratteristica fondamentale è che tutti i risultati hanno l'identica possibilità di verificarsi. 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 analitica che rappresenta la probabilità che una variabile discreta X di distribuzione uniforme assuma un particolare valore è stabilita dalla espressione:

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

nella quale *b*=risultato maggiore possibile di X; *a*= risultato minore possibile di X

La probabilità che esca un numero da 1 a 6 con il lancio di un dado non truccato, per ognuno dei 6 possibili risultati, è uguale a P(x)= 1/[(6-1)+1]= 1/6.

Nella distribuzione discreta uniforme la media e la deviazione standard sono, rispettivamente:

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

La utilizzazione di questa distribuzione è limitata quasi esclusivamente all'analisi di probabilità a priori.

Un caso di frequente applicazione, ad esempio, è la distribuzione di animali in appezzamenti di dimensioni uguali, per verificare una omogeneità di dispersione nelle varie condizioni o situazioni in cui sono collocate le aree campionate. L'ipotesi alternativa è l'esistenza di una associazione tra presenza (o assenza) della specie e condizione ambientale.

B. <u>PRINCIPALI DISTRIBUZIONI CONTINUE</u>

8. La distribuzione normale o di Gauss

Nello studio della distribuzione binomiale si è fatto riferimento al verificarsi di eventi distinti, quali il numero di volte che si ha testa nel lancio di una moneta, il numero dei campi che sono gestiti direttamente dal proprietario, eccetera. Si sono, cioè, considerate *variabili discrete* o capaci di assumere valori discontinui. Ora studieremo, invece, una distribuzione matematica in cui le variabili possono assumere valori continui.

La funzione di distribuzione continua più usata sia nelle applicazioni pratiche sia in quelle di teoria è una distribuzione simmetrica di forma campanulare, nota come "*funzione di distribuzione normale*". La curva che

rappresenta la distribuzione normale è talvolta chiamata "*curva degli errori accidentali*", perché essa rappresenta graficamente il numero degli scarti delle osservazioni reali dal loro valore vero, quando tali scarti si verificano casualmente, cioè non sotto l'effetto di cause sistematiche di differenziazione. I matematici i cui nomi sono associati a questa legge sono De Moivre, Gauss e Laplace. Per questa ragione, la curva normale o degli errori accidentali è anche chiamata qualche volta "*curva gaussiana*" o "*curva di Laplace*".

E' possibile dimostrare che la distribuzione degli errori casuali attorno al valore vero - ovvero delle stime campionarie di un parametro rispetto al parametro della popolazione da cui proviene il campione - è descritta molto bene da una funzione esponenziale del tipo indicato dalla seguente espressione algebrica:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} . e^{-\frac{z^2}{2\sigma^2}}$$

In tale espressione z rappresenta l'errore del valore osservato rispetto al suo valore vero preso come origine e σ non è altro che la deviazione standard o scostamento quadratico medio dell'intera distribuzione dei dati rispetto al loro valore medio.

In pratica, molte distribuzioni osservate mostrano la forma di una curva normale. Specialmente le distribuzioni che si riferiscono a misure effettuate su caratteri antropometrici e biometrici sono di forma normale. Per di più, si può mostrare che la curva normale può essere usata anche per le distribuzioni che non sono normali, ma che possono approssimarsi a tale forma attraverso opportune trasformazioni di variabile.

Per esempio; una distribuzione che sembri molto asimmetrica rispetto alla variabile x può diventare quasi normale quando invece della variabile x si consideri una sua trasformata del tipo $\sqrt{x}$, oppure x^2, oppure $\log x$, eccetera. In tali casi, quindi, la distribuzione normale può essere presa come modello matematico dopo aver operato una semplice trasformazione della variabile originaria.

La grande importanza che ha la curva normale nella teoria statistica deriva anche dalle sue numerose proprietà matematiche.

Per esempio, nella teoria dei campioni, si dimostra che, anche quando la variabile di base non ha una distribuzione normale, la media campionaria, attraverso la quale si cerca di stimare la media vera della variabile di base, segue approssimativamente una distribuzione normale.

Se si indicano in forma esplicita gli scarti o differenze dei singoli valori rispetto alla loro media, l'espressione generale della funzione che descrive la distribuzione normale può essere scritta in forma più chiara come segue:

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

nella quale x indica la variabile di base, M la sua media e σ^2 la varianza della distribuzione. Se si introduce una nuova variabile t chiamata "unità standard" o "scarto ridotto" oppure "scarto standardizzato" e definita dalla relazione:

$$t = \frac{x - M}{\sigma}$$

possiamo scrivere l'equazione della "*curva normale standardizzata*" come segue:

$$f(t) = \frac{N}{\sqrt{2\pi}} . e^{-\frac{t^2}{2}}$$

La distribuzione della variabile normale standardizzata t si ottiene dunque dalla distribuzione della variabile di base x dopo avere assoggettato i valori di quest'ultima ad una trasformazione che trasferisce l'origine dei valori nel punto medio della distribuzione di base ed assume come nuova unità di misura la deviazione standard della variabile di base. Quindi *la trasformata t è una variabile caratterizzata dal fatto di avere media uguale a zero e deviazione standard uguale all'unità*.

8.1. *La normale come approssimazione della binomiale*

L'equazione della curva della distribuzione normale può essere trovata anche come forma limite della funzione binomiale per valori di n

sufficientemente grandi. Per dimostrare quanto detto, si consideri la funzione di probabilità binomiale:

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

e si indichi con u la deviazione della variabile x dalla sua media np, cioè potremo scrivere $u = x - np$ e quindi $x = np + u$. Sostituendo quest'ultima espressione di x nella relazione precedente, si ottiene:

$$P(x) = \frac{n!}{(np+u)!(n-np-u)!} p^{np+u} q^{n-np-u} =$$

$$P(x) = \frac{n!}{(np+u)!(nq-u)!} p^{np+u} q^{nq-u}.$$

Ora, la formula di De Moivre-Stirling definisce il fattoriale di n con lo sviluppo in serie seguente:

$$n! = n^n e^{-n} \sqrt{2\pi n} \left(1 + \frac{1}{12.n} + \frac{1}{288.n^2} + \ldots\ldots\right)$$

Se in essa, per valori di n sufficientemente elevati, si trascurano i termini tra parentesi, si ottiene una espressione sufficientemente approssimata del fattoriale di n secondo la quale risulta:

$$n! = n^n e^{-n} \sqrt{2\pi n}$$

Se nell'ultima espressione ottenuta per la P(x) si sostituisce questo valore approssimato di n! con alcune semplificazioni si ottiene per P(x) la nuova formulazione che si presenta come segue:

$$P(x) = \frac{1}{\sqrt{2\pi npq}} \left(1 + \frac{u}{np}\right)^{-np-u-\frac{1}{2}} \left(1 - \frac{u}{nq}\right)^{-nq+u-\frac{1}{2}}$$

Di questa espressione si considerino ora i logaritmi naturali di ambedue i membri in modo da ottenere:

$$\log\left[P(x).\sqrt{2\pi npq}\right] = -\left(np + u + \frac{1}{2}\right) \log\left(1 + \frac{u}{np}\right) - \left(nq - u + \frac{1}{2}\right) \log\left(1 - \frac{u}{nq}\right)$$

Espandendo le espressioni logaritmiche in serie di potenza e trascurando i termini di ordine (1/n), si perviene, infine, al seguente risultato:

$$\log\left[P(x).\sqrt{2\pi npq}\right] = -\frac{u^2}{2npq}$$

dal quale si ricava:

$$P(x) = \frac{1}{\sqrt{2\pi npq}}.e^{-\frac{u^2}{2npq}}$$

In questa espressione la lettera *e* rappresenta la base dei logaritmi naturali; inoltre sappiamo che $\sigma = \sqrt{npq}$ e che $u = x - np = x - M$. Quindi l'equazione assume la forma:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{(x-M)^2}{2\sigma^2}}$$

In essa P(x) misura la ordinata della curva di probabilità nota come *densità della probabilità* della variabile x. In altre parole, P(x) rappresenta la probabilità che la variabile descritta dalla distribuzione assuma il valore x e la funzione precedente è l'espressione analitica della *curva normale della probabilità*.

Tale curva possiede alcune proprietà notevoli che sono di grande ausilio nello studio delle caratteristiche distributive di alcuni caratteri statistici più rilevanti sia nell'ambito biometrico, sia nel campo demografico, economico-produttivo e ambito sociale.

8.2 *Alcune proprietà della curva normale*

Dato che la somma di tutte le probabilità è pari all'unità, la somma di tutti i valori di P(x) compresi tra $x = -\infty$ ed $x = +\infty$ sarà pari all'unità e quindi vale la relazione:

$$\int_{-\infty}^{+\infty} P(x).dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = \frac{1}{\sigma\sqrt{2\pi}}.\int_{-\infty}^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = 1$$

oppure, nella forma standard in cui $t = (x-M)/\sigma$ e $dx = \sigma dt$:

$$\int_{-\infty}^{+\infty} P(t).dt = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}.e^{-\frac{t^2}{2}}.dt = \frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}.dt = 1$$

Poiché *la curva normale è simmetrica rispetto alla verticale passante per la media della distribuzione*, l'intera area sotto la curva compresa tra le ascisse $-\infty$ e $+\infty$ può essere considerata uguale al doppio dell'area compresa tra zero e $+\infty$ per cui si può scrivere la seguente relazione:

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = \frac{2}{\sigma\sqrt{2\pi}}.\int_{-\infty}^{0} e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = \frac{2}{\sigma\sqrt{2\pi}}.\int_{0}^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = 1$$

e similmente, facendo riferimento alla variabile ridotta standardizzata, anche per essa si avrà che l'ordinata in corrispondenza dell'ascissa 0 rappresenta l'elemento di separazione in due parti uguali dell'intera area racchiusa tra la curva e l'asse delle ascisse, cioè:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}.e^{-\frac{t^2}{2}}.dt = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}}.e^{-\frac{t^2}{2}}.dt = \frac{2}{\sqrt{2\pi}}.\int_{0}^{+\infty} e^{-\frac{t^2}{2}}.dt = 1$$

D'altra parte, se vogliamo che l'area sotto la curva normale sia uguale alla frequenza totale dei casi N, dovremo utilizzare la relazione N.P(x) = f(x) e quindi sarà:

$$\int_{-\infty}^{+\infty} N.P(x).dx = \frac{N}{\sigma\sqrt{2\pi}}.\int_{-\infty}^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}}.dx = N.$$

In questo caso le ordinate della curva sono uguali alle frequenze assolute ed infatti sarà:

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

Come può vedersi dalla sua equazione, *una distribuzione normale è completamente determinata dalla sua media e dalla sua deviazione standard*, in quanto essi sono i due soli parametri necessari per descriverla completamente.

Prendendo le derivate prima e seconda della equazione suddetta si può vedere che la curva ha le seguenti caratteristiche:

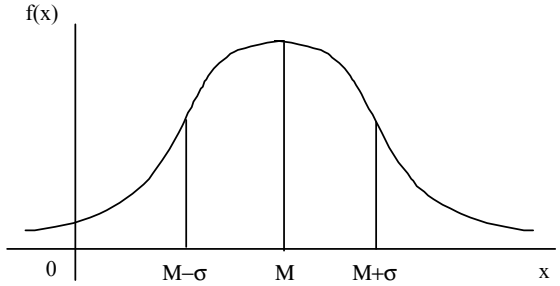
- possiede un solo massimo in corrispondenza del valore medio, cioè del punto $x = M$;

- essa ha anche due punti di flesso in corrispondenza dei valori $x = M \pm \sigma$;

- inoltre il valore della funzione $f(x)$ *tende a zero man mano che* $f(x)$ *si approssima a* $\pm \infty$, *cosicché l'asse delle* x *ne rappresenta un asintoto;*

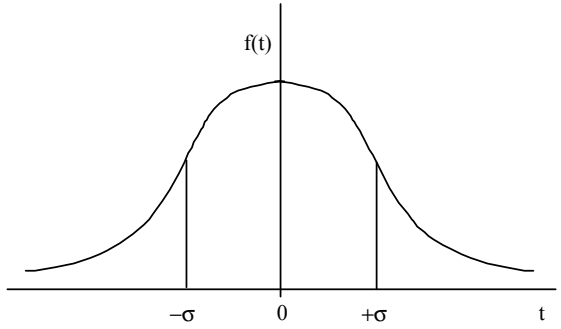
- infine, media, mediana e moda della distribuzione normale coincidono.

La forma generale della curva normale è descritta nei grafici che seguono:



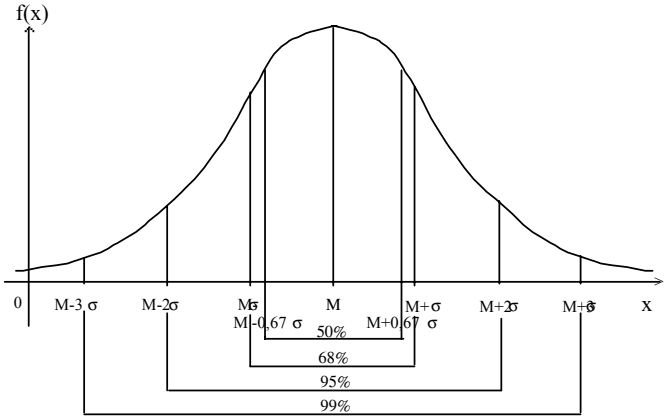
I. Rappresentazione rispetto alla variabile di origine

Data la grande importanza della curva normale, sono state predisposte diverse tavole numeriche che forniscono i valori delle ordinate della curva corrispondenti ad un valore determinato dell'ascissa e che danno anche la misura delle aree comprese sotto la curva a sinistra ed a destra di certe ordinate.



II. Rappresentazione rispetto alla variabile standardizzata

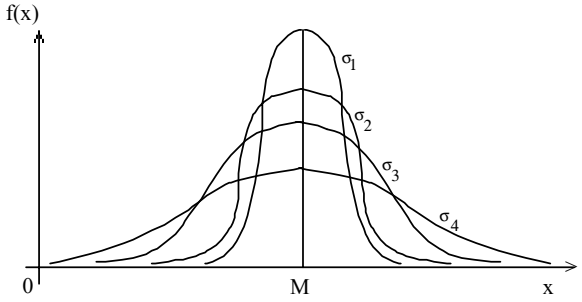
Poiché la forma standard della curva normale è molto più semplice, in generale le tavole sono preparate utilizzando i valori della variabile espressi in unità standard.



Da queste tavole si può vedere che circa il 68% dell'area complessiva sotto la curva normale è compresa nell'intervallo $M \pm \sigma$, il 95% insiste sull'intervallo $M \pm 2.\sigma$, mentre il 99% circa ricade nell'intervallo $M \pm 3.\sigma$. La metà dell'intera area è compresa nell'intervallo $M \pm 0{,}676.\sigma$, il quale è anche

chiamato "*errore probabile*". Possiamo stabilire le medesime relazioni con parole leggermente diverse: 50%, 68%, 95% e 99% di tutti i valori di una distribuzione normale sono inclusi rispettivamente negli intervalli $M\pm 0{,}676.\sigma$, $M\pm \sigma$, $M\pm 2.\sigma$, $M\pm 3.\sigma$. Nel grafico che precede questo capoverso sono evidenziate le relazioni ora dette.

Via via che il valore della deviazione standard diviene più piccolo la distribuzione normale tende a concentrarsi sempre più attorno alla sua media. Questa proprietà, riferita a distribuzioni normali che hanno la stessa media, ma deviazioni standard diverse, viene descritta graficamente nella figura che segue, ove si può osservare come, *a parità di valore medio, la forma campanulare è tanto più appiattita quanto più grande è la variabilità*. Si ricorda, infatti, che la variabilità è misurata attraverso opportune medie di potenze degli scarti, le quali proprio per effetto dell'elevamento a potenza dei vari dati, fanno dipendere il risultato in misura maggiore dagli scarti più elevati in valore assoluto (cioè quelli situati alle estremità della curva) ed in misura minore da quelli più piccoli (cioè quelli dei valori centrali o in prossimità della media).



$$\sigma_1 < \sigma_2 < \sigma_3 < \sigma_4$$
$$M_1 = M_2 = M_3 = M_4$$

Inoltre, *date n variabili indipendenti, se la loro distribuzione è di tipo normale, allora la somma di queste variabili è anch'essa distribuita normalmente con media e varianza uguali alla somma delle medie ed alla somma delle varianze delle variabili originarie*.

8.3 *Momenti della distribuzione normale*

La media e la deviazione standard della distribuzione normale sono uguali alla M ed al σ che compaiono nell'equazione normale. I momenti teorici di ordine più elevato possono essere calcolati tramite l'operazione di integrazione.

Il momento rispetto alla media di ordine k è definito dalla relazione:

$$\mu_k = \int_{-\infty}^{+\infty} \frac{(x-M)^k . f(x)}{N} . dx = \frac{1}{\sigma\sqrt{2\pi}} . \int_{-\infty}^{+\infty} (x-M)^k . e^{-\frac{(x-M)^2}{2\sigma^2}} . dx$$

Allora avremo:

$$\alpha_k = \frac{\mu_k}{\sigma^k} = \int_{-\infty}^{+\infty} \frac{(x-M)^k}{\sigma^k} . \frac{1}{\sigma\sqrt{2\pi}} . e^{-\frac{(x-M)^2}{2\sigma^2}} . dx$$

Considerando che

$$\frac{(x-M)^k}{\sigma^k} = t^k \qquad \text{e che} \qquad \frac{dx}{\sigma} = dt$$

si ottiene:

$$\alpha_k = \int_{-\infty}^{+\infty} t^k . \frac{1}{\sqrt{2\pi}} . e^{-\frac{t^2}{2}} . dt = \frac{1}{\sqrt{2\pi}} . \int_{-\infty}^{+\infty} t^k . e^{-\frac{t^2}{2}} . dt$$

Le principali caratteristiche della legge normale possono essere riassunte come segue.

I° <u>*La distribuzione normale è simmetrica rispetto ad M*</u>. Le frequenze relative su due segmenti della stessa lunghezza dx, di centro x, corrispondenti agli scarti -(M-x) e +(M-x) sono uguali e pari a:

$$f(x).dx = \frac{1}{\sigma\sqrt{2\pi}} . e^{-\frac{(x-M)^2}{2\sigma^2}} . dx$$

Ciò risulta dalla simmetria della curva normale, di cui più volte si è discusso in precedenza per evidenziare il grande interesse di questa curva nelle applicazioni della Statistica per lo studio dei fenomeni collettivi.

II° *Il valore modale di x, M_d, tale che la densità di ripartizione sia massima, è $M_d = M$*. La densità o frequenza relativa corrispondente è:

$$f(M_d) = \frac{1}{\sigma\sqrt{2\pi}}$$

III° *La mediana di x, M_e, tale che le frequenze relative delle unità per le quali $x \leq M_e$ è 1/2, è $M_e = M$*. Questo risulta dalla simmetria della distribuzione.

IV° *La media di x é M.* Essa è per definizione uguale a:

$$M = \int_{-\infty}^{+\infty} x.f(x).dx = \frac{1}{\sigma\sqrt{2\pi}}.\int_{-\infty}^{+\infty} x.e^{-\frac{(x-M)^2}{2\sigma^2}}.dx$$

cioè la somma dei prodotti dei valori assunti da x ciascuno moltiplicato per la rispettiva frequenza relativa. *Il parametro M della legge normale è uguale alla media aritmetica della distribuzione*.

V° *Il momento di ordine k rispetto alla media è μ_k* Esso è definito dalla relazione:

:

$$\mu_k = \int_{-\infty}^{+\infty}(x-M)^k.f(x).dx = \frac{1}{\sigma\sqrt{2\pi}}.\int_{-\infty}^{+\infty}(x-M)^k.e^{-\frac{(x-M)^2}{2\sigma^2}}.dx$$

Tenuto conto della simmetria della distribuzione, per il calcolo dei momenti di ordine pari e di quelli di ordine dispari si dovrà seguire una procedura differenziata, della quale si offre una corretta descrizione in ciò che segue.

Momenti di ordine dispari. Quando k è dispari, i prodotti sotto il segno di integrale si annullano due a due, dato che le frequenze per i due scarti (x - M) di pari valore assoluto sono uguali, ma di segno contrario. *Tutti i momenti di ordine dispari rispetto alla media sono nulli*. Questa proprietà non vale solo per la legge normale, ma è valida per tutte le funzioni simmetriche e, pertanto, per questo tipo di funzioni i momenti di ordine dispari non si calcolano.

Momenti di ordine pari. Per calcolarli, ricordiamo la relazione:

$$\mu_k = \frac{\sigma^k}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} t^k.e^{-\frac{t^2}{2}}.dt$$

Tenuto conto della relazione differenziale:

$$d\left(t^{k-1}.e^{-\frac{t^2}{2}}\right) = (k-1).t^{k-2}.e^{-\frac{t^2}{2}}.dt - t^k.e^{-\frac{t^2}{2}}.dt$$

la quale per integrazione fornisce il seguente risultato:

$$\left[t^{k-1}.e^{-\frac{t^2}{2}}\right]_{-\infty}^{+\infty} = (k-1).\int_{-\infty}^{+\infty} t^{k-2}.e^{-\frac{t^2}{2}}.dt - \int_{-\infty}^{+\infty} t^k.e^{-\frac{t^2}{2}}.dt$$

ossia, in funzione dei momenti μ:

$$0 = (k-1)\frac{\sqrt{2\pi}}{\sigma^{k-2}}.\mu_{k-2} - \frac{\sqrt{2\pi}}{\sigma^k}.\mu_k$$

si ricava la seguente espressione finale del momento di ordine k:

$$\mu_k = (k-1).\sigma^2.\mu_{k-2}.$$

Questa relazione ricorrente permette di ottenere tutti i momenti. Dato che è $\mu_0 = 1$ e $\mu_1 = 0$, si hanno successivamente tutti i momenti di ordine pari:

$$\mu_2 = \sigma^2 \; ; \quad \mu_4 = 3.\sigma^4 \; ; \quad \mu_6 = 15.\sigma^6 \quad \text{e così via.}$$

Il parametro σ *della legge normale è dunque lo scarto tipo o deviazione standard di* x che è già stato definito.

Si ha così una interpretazione semplice dei due parametri che figurano nell'espressione della legge normale, in quanto, come si è già detto, l'uno è la *media aritmetica* e l'altro è la *deviazione standard* della distribuzione.

La variabile ausiliaria t, ottenuta con una trasformazione dalla variabile originaria x e più precisamente portando l'origine degli assi nel punto medio e sostituendo all'unità di misura originaria un nuovo sistema di misura che assume come unità di riferimento la deviazione standard della variabile originaria, viene denominata anche *variabile normalizzata o scarto ridotto* e la espressione della sua legge di distribuzione viene detta *forma ridotta della legge normale*.

Essa non dipende da alcun parametro o meglio ha la *media uguale a zero* e la *deviazione standard uguale all'unità*. Risulta inoltre che per la curva normale standardizzata gli indici ottenuti considerando i momenti dei successivi ordini fino al quarto assumono i valori $\alpha_3 = 0$ ed $\alpha_4 = 3$ e lo stesso dicasi per β_1 e β_2. Questo risultato è molto importante, in quanto è a questi valori che vengono paragonati i corrispondenti valori calcolati per le altre distribuzioni di tipo campanulare, per valutarne il grado di asimmetria e di appiattimento.

8.4 *Adattamento della distribuzione normale a dati empirici*

Si supponga di avere una distribuzione di frequenza di tipo campanulare di una variabile che si ritiene di poter rappresentare analiticamente con una forma uguale a quella che definisce la curva normale. Per vedere in concreto come si può operare per effettuare questo tipo di interpolazione si ragioni come segue.

Sappiamo che la densità di frequenza della distribuzione normale nel punto x è definita dalla relazione:

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

Prendendo i logaritmi dei due i membri della relazione si ha:

$$\log f(x) = -\frac{(x - M)^2}{2\sigma^2} + \log \frac{N}{\sigma\sqrt{2\pi}}$$

Se in questa relazione definiamo $Y = \log f(x)$ ed $X = (x - M)^2$, si avrà:

$$Y = -\frac{1}{2\sigma^2} . X + \log \frac{N}{\sigma\sqrt{2\pi}}$$

ed essendo costanti sia il coefficiente della X, sia il termine logaritmico situato nel secondo termine della relazione, i punti che hanno coordinate (X,Y) devono giacere su una linea retta.

Ne segue che, se ai dati osservati di cui si dispone può essere adattata una curva normale, si potrà utilizzare la tecnica di adattamento che viene descritta qui di seguito.

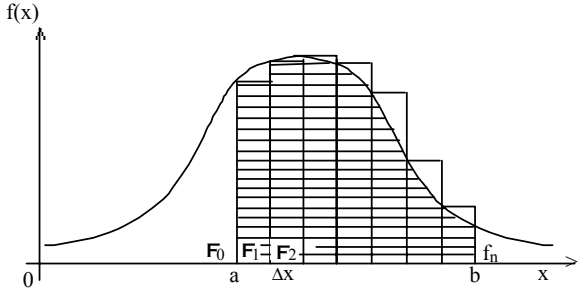
L'interpolazione di una curva normale può essere fatta usando le ordinate della curva o le percentuali che rappresentano le aree sotto la curva per i diversi intervalli di classe.

Molti libri di Statistica forniscono le tavole della curva normale standardizzata che danno le ordinate o le aree in corrispondenza di determinati valori delle ascisse. Va osservato che, in generale, *è più conveniente lavorare con le percentuali che rappresentano le aree piuttosto che con le frequenze assolute*, malgrado ciò si debba fare con molta attenzione.

Nel grafico che segue l'area tratteggiata rappresenta la probabilità che la variabile x assuma uno qualunque dei valori compresi tra a e

b. Sappiamo che questa probabilità è uguale alla somma delle frequenze relative, la prima frequenza essendo $f(a) = f_0$ e l'ultima essendo $f(b) = f_n$.

Ma dal grafico si vede facilmente che, se la base di ogni piccolo rettangolo è $\Delta x = 1$, l'area tra *a* e *b* è uguale alla somma delle frequenze da f_0 fino a f_{n-1}, l'ultima frequenza , infatti, deve essere trascurata per evitare che si abbiano duplicazioni di valori che possano alterare sostanzialmente il risultato finale.



Cioè, avremo:

$$\int_a^b f(x).dx = \sum_{i=o}^{i=n-1} f_i.\Delta x_i = \sum_{i=o}^{i=n} f_i.\Delta x_i - f_n$$

ovvero:

$$\sum_{i=o}^{i=n} f_i.\Delta x_i = \int_a^b f(x).dx + f(x) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}}.\int_a^b e^{-\frac{(x-M)^2}{2\sigma^2}}.dx + \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{(x-M)^2}{2\sigma^2}}$$

Per esprimere la sezione a destra del segno di uguaglianza in un solo termine, si è soliti aggiungere un intervallo in più a quello definito dalla distanza (a,b); questa operazione viene eseguita aggiungendo un intervallo pari alla metà della classe all'estremo superiore del campo di variazione definito da (a,b) e togliendo metà dell'intervallo di classe dall'estremo inferiore dello stesso campo di variazione.

Il passo successivo per procedere all'adattamento della curva è quello di calcolare le unità standardizzate corrispondenti a ciascun estremo di classe in base alla trasformazione:

$$t = \frac{x - M}{\sigma}$$

Quindi le aree a sinistra di ogni valore di t possono essere trovate nelle tavole già predisposte da vari autori per la variabile che rappresenta lo *scarto ridotto*. Le differenze tra due successive aree di base unitaria, moltiplicate per la somma delle frequenze N, fornisce le frequenze teoriche della distribuzione normale che meglio si adatta alla seriazione di dati statistici considerati.

Esempio IV.5

Si consideri la seguente tabella delle frequenze dell'altezza in centimetri di 78 piante di una determinata specie. Ci si propone di:

a) verificare se la curva normale può essere scelta come una buona curva interpolatrice per tale distribuzione di frequenze;

b) calcolare le frequenze teoriche della curva normale che ha la stessa media e la stessa varianza della seriazione data.

Altezza (cm)	Frequenza
10 ¬ 15	4
15 ¬ 20	20
20 ¬ 25	28
25 ¬ 30	12
30 ¬ 35	8
35 ¬ 40	6
	78

Calcoliamo anzitutto i momenti della distribuzione data.

Classi di altezza(cm)	Media(v_i) di classe	f_i	$x_i=(v_i-22,5)$	x_if_i	$x_i^2f_i$	$x^3{}_if_i$	$x_i^4f_i$
10 ¬ 15	12,5	4	-2	-8	16	-32	64
15 ¬ 20	17,5	20	-1	-20	20	-20	20
20 ¬ 25	22,5	28	0	0	0	0	0
25 ¬ 30	27,5	12	1	12	12	12	12
30 ¬ 35	32,5	8	2	16	32	64	128
35 ¬ 40	37,5	6	3	18	54	162	486
Totali		78	3	18	134	186	710

$$M = 18/78 = 0,23;$$

$$M_2 = 134/78 = 1,72 ;$$

$$M_3 = 186/78 = 2,38;$$

$$M_4 = 710/78 = 9,10 ;$$

$$\mu_2 = M_2 - M^2 = 1,718 - 0,053 = 1,665 .$$

Tenendo conto della correzione di Sheppard sul momento secondo avremo:

$$\mu'_2 = \mu_2 - 1/12 = 1,66 - 0,08 = 1,58 \text{ e quindi } \sigma = \sqrt{1,58} = 1,26.$$

Per i momenti terzo e quarto avremo:

$$\mu'_3 = M_3 - 3.M_2.M + 2.M^3 = 2,38 - 1,19 + 0,024 = 1,21$$

$$\alpha_3 = 1,21/(1,26)^3 = 1,21/1,99 = 0,6$$

$$\mu_4 = M_4 - 4.M_3.M + 6.M_2.M^2 - 3.M^4 = 9,10 - 2,19 + 0,55 - 0,01 = 7,44$$

e tenendo conto della correzione di Sheppard su μ_4 avremo:

$$\mu'_4 = \mu_4 - 1/2\mu_2 + 7/240 = 7,44 - 0,83 + 0,03 = 6,64$$

$$\alpha_4 = 6,64/(1,26)^4 = 6,64/2,50 = 2,6.$$

Come si vede, i valori di α_3 e α_4 sono quasi prossimi a zero ed a 3; si può quindi concludere che la curva normale può essere adottata come curva interpolatrice, malgrado essa non dia un perfetto adattamento ai dati.

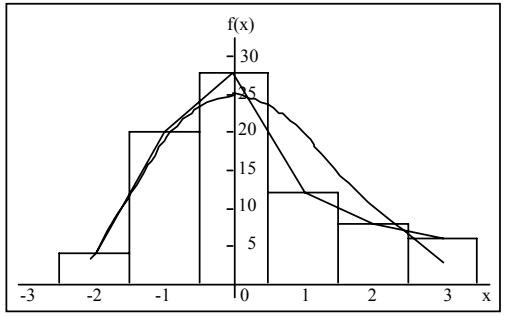
La curva della distribuzione data ha asimmetria positiva e quindi essa sarà più allungata verso destra e più inclinata verso sinistra. Inoltre, poichè $\alpha_4 = 2,6$ è minore di 3, la curva sarà più appiattita della curva normale.

Per trovare le frequenze teoriche usando le aree della curva normale, si dovranno effettuare le elaborazioni che sono riportate nella tabella che segue, nella quale figurano i singoli passaggi necessari, ivi inclusa l'utilizzazione di una variabile di comodo.

v_i	$x_i=(v_i-22,5)$	f_i	(x_i-M)	$t_i=(x_i-M)/\sigma$	Integrale tra $-\infty$ e 0 di P(t)	Aree unitarie $\Delta.A$	Frequenze teoriche N. Δ A
12,5	-2,5	4	-2,73	-2,18	0,015	0,027	2,106
17,5	-1,5	20	-1,73	-1,38	0,042	0,239	18,642
22,5	-0,5	28	-0,73	-0,58	0,281	0,306	23,868
27,5	0,5	12	0,27	0,22	0,587	0,259	20,202
32,5	1,5	8	1,27	1,02	0,846	0,120	9,360
37,5	2,5	6	2,27	1,82	0,966	0,030	2,340
	3,5		3,27	2,62	0,996		

Totali	78						

La rappresentazione grafica della distribuzione osservata e della curva normale ad essa adattata risulta essere la seguente:



9. Distribuzione rettangolare

La *distribuzione rettangolare* è la più semplice tra le distribuzioni continue. Essa è anche detta *distribuzione uniforme continua*. Essa, nell'intervallo tra $x_1=\alpha$ e $x_2=\beta$ ha una densità di frequenza relativa definita dalla espressione

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

ed è pertanto caratterizzata da una densità costante in tutto l'intervallo compreso tra α e β. La rappresentazione grafica di questa distribuzione ha la forma di un rettangolo, che giustifica il suo nome. Essa è l'equivalente della distribuzione uniforme discreta considerata nel continuo e la sua media e varianza sono:

$$M = \frac{\alpha + \beta}{2} \text{ e } \sigma^2 = \frac{(\beta - \alpha)^2}{12}.$$

10. Distribuzione esponenziale negativa

La *esponenziale negativa* è una distribuzione continua descritta dalla relazione

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

Essa prende il nome dall'esponente negativo che compare nella relazione, è una *funzione positiva o nulla continuamente decrescente*, che tende a 0 per x tendente all'infinito. Nel discreto ha l'equivalente nella distribuzione geometrica decrescente. La media e la varianza sono, rispettivamente:

$$M = \frac{1}{\alpha} \text{ e } \sigma^2 = \frac{1}{\alpha^2} = M^2 .$$

E' di estremo interesse notare che in questa distribuzione *la varianza è il quadrato della media*, dato che le relazioni tra media e varianza delle distribuzioni teoriche sono utili per capire, da una serie di dati sperimentali sulla quale siano state calcolate le due statistiche, quale possa essere la distribuzione teorica corrispondente.

11. Le curve di Pearson

Il sistema di curve di frequenza di Karl Pearson riesce a descrivere con elevata approssimazione molte distribuzioni empiriche; si adatta molto bene a rappresentare la distribuzione dei valori in molti fenomeni reali, ma ha il grave limite che i parametri che la definiscono non sono esplicativi, spesso non forniscono alcun significato per l'interpretazione del fenomeno studiato e di conseguenza non si prestano ad usi predittivi.

La forma esplicita della funzione può essere espressa dalla seguente equazione differenziale:

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

che dipende dalle radici dell'espressione quadratica del denominatore, cioè dai valori dei parametri b_0, b_1, b_2, essendo x ed y i valori degli assi coordinati e c una costante.

Il sistema gode della proprietà di rappresentare molte curve e di poter passare dall'una all'altra variando i parametri prima definiti. Anche la

distribuzione normale e le sue approssimazioni con asimmetria ed appiattimento variabili possono essere rappresentate come una delle possibili curve del Pearson e possono essere riguardate come una parte delle distribuzioni che è possibile descrivere con elevata precisione. La curva specifica è scelta in funzione dei momenti delle distribuzioni, con un criterio che è definito dagli indici di asimmetria e di appiattimento di Pearson. Le forme tipiche di curve che possono essere rappresentate sono sette e vanno dalla forma campanulare simmetrica e asimmetrica , alla curva ad U simmetrica e asimmetrica, dalla forma a J a quella a J rovesciato.

12. La distribuzione Gamma (III tipo del Pearson)

Un altro modello per descrivere la distribuzione di variabili casuali continue e positive è la distribuzione Gamma (Γ). Se una variabile continua x è distribuita con densità espressa dalla relazione:

$$f(x) = \frac{1}{\Gamma(m)} x^{m-1} e^{-x} \quad (0 \le x < \infty;\ m > 0)$$

in cui la funzione gamma è descritta dalla relazione:

$$\Gamma(m) = \int_0^{\infty} x^{m-1} e^{-x} dx \qquad \text{[per m intero è } \Gamma(m)=(m-1)!\text{]}$$

si dice che essa distribuita secondo la funzione Gamma, con parametro m, o più brevemente si chiama una variabile $\gamma(m)$.

Tale distribuzione è asintotica rispetto all'asse delle x e si annulla per x=0 se m>2. Essa fu chiamata da Legendre, agli inizi dell'800, integrale euleriano di seconda specie, in onore di Eulero che la studiò per primo nella seconda metà del '700, e costituisce una curva del III tipo del Pearson.

La distribuzione Gamma è di notevole importanza in Statistica poiché si dimostra che, *quando una variabile statistica è distribuita normalmente, il suo quadrato è distribuito secondo una distribuzione Gamma*. Inoltre, anche *la*

distribuzione della varianza dei campioni è ancora una distribuzione Gamma e, in generale, *la somma dei quadrati di variabili normali è distribuita secondo una distribuzione Gamma*.

13. La distribuzione Beta (I tipo del Pearson)

Se una variabile continua x è distribuita con densità definita dalla relazione:

$$f(x) = \frac{x^{k-1}(1-x)^{m-1}}{B(k,m)} \qquad (0 \le x \le 1;\ k > 0;\ m > 0)$$

in cui

$$B(k,m) = \frac{\Gamma(k).\Gamma(m)}{\Gamma(k+m)} = \int_0^1 x^{k-1}(1-x)^{m-1}dx$$

essa si dice distribuita secondo la funzione Beta con parametri k ed m o più brevemente chiamasi variabile β(k,m). La f(x) sopra definita costituisce una curva del I tipo del Pearson.

La distribuzione Beta è strettamente collegata con la distribuzione Gamma. Infatti, si dimostra che *se due variabili indipendenti sono distribuite rispettivamente secondo una γ(k) e una γ(m), il loro rapporto è distribuito secondo una β(k,m)*. Ciò è molto importante. Infatti, la distribuzione del rapporto delle varianze è una distribuzione Beta, in quanto è un rapporto di due variabili distribuite ciascuna secondo una Gamma. Essa è la distribuzione F, che si vedrà in seguito, di capitale importanza per l'analisi della varianza.

14. La distribuzione Chi-quadrato $\chi^2_{(n)}$

Se $X_1, X_2,..., X_n$ sono n variabili casuali indipendenti distribuite normalmente con media zero e varianza uno, la variabile casuale:

$$\chi^2 = X_1^2 + X_2^2 + ... + X_n^2$$

per x≥0 è distribuita secondo la funzione

$$P(\chi^2 \le x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}.\Gamma\left(\frac{n}{2}\right)} = \displaystyle\int_0^x u^{\frac{n}{2}-1} e^{-\frac{u}{2}} du & \text{per } x \ge 0 \\ 0 & \text{per } x < 0 \end{cases}.$$

Questa è chiamata la *distribuzione chi-quadrato* e varia in funzione del parametro *n*; essa è un caso particolare della distribuzione Gamma.

Valgono i seguenti teoremi:

- *Se X_1, X_2,...., X_n sono variabili casuali indipendenti e normalmente distribuite con media 0 e varianza 1, allora $\chi^2 = X_1^2 + X_2^2 + ... + X_n^2$ è la distribuzione chi-quadrato con n gradi di libertà.*

- *Se U_1, U_2,...., U_k sono k variabili casuali indipendenti con distribuzione chi-quadrato ed n_1, n_2,..., n_k gradi di libertà rispettivamente, allora la loro somma $W = U_1 + U_2 + ... + U_k$ ha una distribuzione di tipo chi-quadrato con $n_1+n_2+....+n_k$ gradi di libertà.*

- *Se V_1 e V_2 sono due variabili casuali indipendenti e se V_1 ha distribuzione chi-quadrato con n_1 gradi di libertà e $V = V_1 + V_2$ è una chi-quadrato con $n<n_1$ gradi di libertà, allora la distribuzione di V_2 è una distribuzione chi-quadrato con $n_2 = n - n_1$ gradi di libertà.*

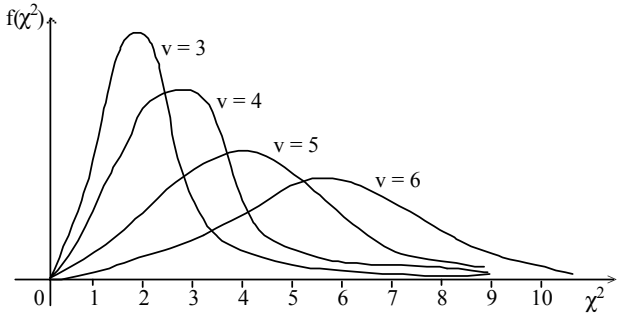
Questa distribuzione assume importanza nell'inferenza statistica, quando si confronta una distribuzione osservata con una specifica distribuzione teorica, oppure quando si confrontano 2 o più distribuzioni osservate. I suoi valori sono tabulati in una tavola sinottica che per ogni grado di libertà riporta i valori critici più importanti (P=0,05; P=0,01; P=0,005; P=0,001).

La variabile χ^2 si usa anche quando si vuole misurare la discrepanza tra frequenze osservate e frequenze teoriche o attese in base ad una data ipotesi. In questo caso la quantità χ^2 è definita dalla formula seguente:

$$\chi^2_{(\nu)} = \sum \left[\frac{(f_o - f_t)^2}{f_t} \right]$$

in cui: ν rappresenta il numero dei gradi di libertà, f_0 rappresenta le frequenze osservate ed f_t rappresenta le frequenze teoriche o attese. Queste ultime, in generale, sono determinate in base all'ipotesi di indipendenza statistica.

Per la definizione di χ^2, si sa che maggiore è il valore di χ^2, maggiore sarà la probabilità di una divergenza reale delle frequenze osservate dalle frequenze teoriche. In altre parole, un grande valore di χ^2 indica che l'ipotesi di indipendenza non può essere accettata. Ora, per determinare quali valori di χ^2, bisogna considerare significativamente grandi con un certo limite di confidenza, abbiamo bisogno di conoscere la distribuzione campionaria di χ^2, come sopra definita, tenendo presente che, secondo i diversi valori di n , la curva di distribuzione del χ^2 assume forme diverse, generalmente di tipo campanulare asimmetrico a destra, delle quali alcune sono mostrate nel grafico che segue. Da esso si vede chiaramente che al crescere del numero dei gradi di libertà la forma della distribuzione tende a divenire più simmetrica e ad abbassare il valore dell'ordinata massima.



Il grafico, inoltre, mostra che la curva del χ^2 si estende da 0 a $+\infty$. Le aree sotto la curva del χ^2 per diversi valori del χ^2 e di ν sono state tabulate appositamente per facilitarne la lettura e l'uso, ossia per scopi pratici.

Queste aree danno le probabilità che in un campionamento casuale si abbia un valore di χ^2 maggiore del valore ottenuto dai dati osservati.

La discussione su questa funzione sarà ripresa nel capitolo dedicato allo studio della associazione tra caratteri non necessariamente quantitativi. Si tratta, infatti, di una distribuzione particolarmente utile per le analisi di statistica non parametrica.

15. La distribuzione F di Fisher

Un'altra distribuzione di notevole interesse pratico, sulla quale si basa molta parte della statistica parametrica è la *distribuzione F*. Essa corrisponde alla distribuzione del rapporto di due variabili casuali chi-quadrato indipendenti (A e B), divise per i loro rispettivi parametri (μ ed ν), cioè:

$$F(\mu, \nu) = \frac{\frac{A}{\mu}}{\frac{B}{\nu}}$$

L'ordine con il quale sono riportati i due numeri che indicano i gradi di libertà di F è importante, perché la densità della distribuzione di F non è simmetrica rispetto ad essi. Per primo si riporta sempre il numero di gradi di libertà del numeratore e per secondo quello del denominatore (si veda anche quanto è detto nel capitolo VIII, paragrafo 3).

16. La distribuzione t di Student

La *distribuzione t di Student*, con ν gradi di libertà è data dal rapporto tra una variabile casuale con distribuzione normale standardizzata (Z) e la radice quadrata di una variabile casuale con distribuzione chi-quadrato (A), indipendente dalla Z e divisa per i suoi gradi di libertà (ν), cioè:

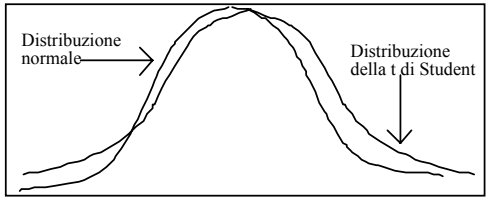
$$\mathbf{t}(\nu)=\frac{Z}{\frac{\sqrt{A}}{\nu}}.$$

Essa è definita dalla funzione seguente:

$$f(t/\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right).\sqrt{\nu\pi}}.\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}=C.\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

la quale mostra che le ordinate della curva assumono valori diversi per valori diversi del parametro ν che appare nella equazione. Questo parametro è chiamato "*Gradi di libertà*".

Il grafico seguente mostra un confronto fra la forma della distribuzione t di Student con ν gradi di libertà e quella della curva normale standard.



In molti testi di statistica si possono trovare tavole speciali che danno il livello delle probabilità corrispondente ai diversi valori di t e di ν. Via via che il valore di ν diviene più grande la distribuzione della t di Student si avvicina sempre più alla distribuzione normale. Per $\nu > 30$ al posto della distribuzione della t di Student, si può usare la distribuzione normale.

All'aumentare di ν, la distribuzione t di Student si avvicina alla distribuzione normale standardizzata.

Il quadrato di una t di Student con ν gradi di libertà è uguale ad una distribuzione F di Fisher con gradi di libertà 1 ed ν, cioè:

$$\mathbf{t}^2_{(\nu)}=F_{(\mathbf{1},\nu)}, \text{ oppure } \mathbf{t}_{(\nu)}=\sqrt{F_{(\mathbf{1},\nu)}}.$$

Anche queste due funzioni saranno riprese in considerazione al momento in cui si considererà il problema della inferenza statistica (analisi delle medie ed analisi della varianza).