Statistica Matematica

Lo studio dei fenomeni complessi, tipici della realtà industriale moderna, comporta l'adozione di opportuni modelli matematici che ne descrivano i caratteri essenziali, funzionali agli obiettivi che il suddetto studio si prefigge. Nei riguardi di un sistema di produzione di beni e/o servizi, è compito dell'ingegneria gestionale definire strategie di intervento che determinino il miglioramento del ciclo produttivo, l'uso efficiente di risorse, la sicurezza e manutenzione degli impianti, il rispetto dell'ambiente.

Il processo di formazione delle decisioni fa uso appunto di modelli descrittivi e di predizione. Questi si ottengono individuando le variabili di interesse e le relazioni che intervengono tra esse. Quest'ultimo passaggio è abbastanza complicato nei sistemi complessi, dove il numero delle variabili che intervengono nel processo è elevato; di conseguenza si dispone anche di un insieme di dati sperimentali, ottenuto dalle misure di dette grandezze, molto grande.

L'analisi dei dati sperimentali ha il compito di individuare quali tra le variabili misurate siano effettivamente significative nel processo in questione, e la complessità delle relazioni che intervengono tra esse. In tale ambito, la statistica gioca un ruolo fondamentale: nel suo aspetto descrittivo e induttivo permette di dedurre i caratteri essenziali di una distribuzione di valori dall'esame di un campione di essi, per poi fornire, unitamente ai risultati propri della teoria della probabilità, una metodologia per la formazione delle decisioni. I metodi statistici intervengono nell'identificazione e nella verifica dei modelli: la prima permette di selezionare il miglior modello di una data classe e la seconda permette di convalidare il modello identificato in termini di rappresentatività dei dati e di potere predittivo.

Statistica descrittiva

Consideriamo un apparato di produzione di supporti in ferro; si vuole caratterizzare la qualità del prodotto finito in termini di *carico di rottura* (Kg/cm^2). Si esamina un lotto di $N = 100$ pezzi ottenendo l' insieme di dati sperimentali raccolti in Tab.1. Visti così, i dati sembrano presentare delle fluttuazioni del tutto arbitrarie uno dall'altro da far sospettare che le caratteristiche meccaniche del pezzo prodotto siano accidentali. Ma questo è frutto di un atteggiamento errato, che consiste nel confrontare i singoli dati tra loro; se analizziamo l'insieme da un punto di vista più generale, ad una scala più larga per così dire, potremmo individuare una certa regolarità che,

opportunamente caratterizzata, permetterà di definire un numero finito di parametri caratteristici del lotto in questione.

Tab. 1. Unità statistiche. Carico di rottura dei pezzi di un lotto (*kg/cm²*)

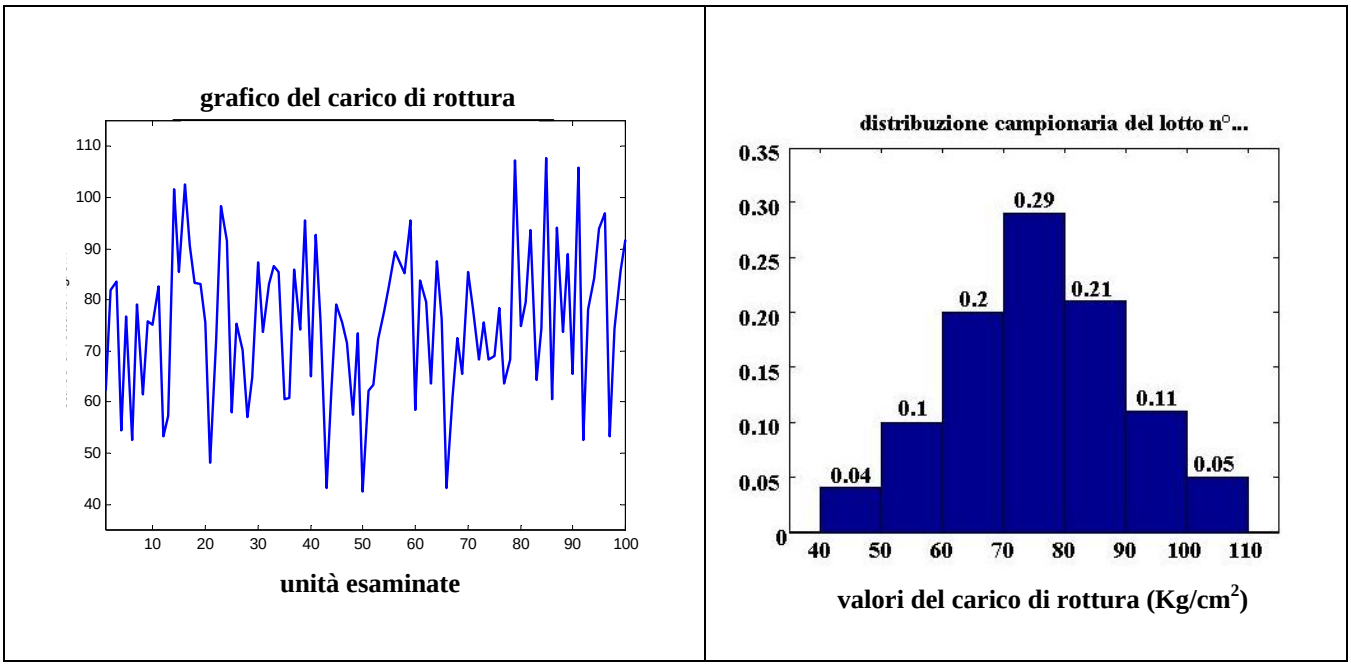
62.3	82.0	83.5	54.4	76.8	52.6	79.1	61.5	75.8	75.2
82.7	53.2	57.3	101.6	85.4	102.6	90.5	83.4	83.2	75.6
48.1	72.2	98.4	91.5	58.1	75.4	70.2	57.1	64.8	87.3
73.8	83.0	86.5	85.4	60.5	60.8	85.8	74.2	95.5	65.1
92.6	75.8	43.1	62.5	79.2	75.6	71.5	57.6	73.5	42.6
62.1	63.4	72.3	77.6	82.6	89.4	87.3	85.1	95.6	58.4
83.7	79.6	63.7	87.5	76.3	43.2	61.0	72.6	65.6	85.5
78.5	68.2	75.6	68.3	69.1	78.4	63.6	68.2	107.2	74.8
79.5	93.6	64.4	74.3	107.8	60.5	94.2	73.6	89.0	65.4
105.8	52.6	78.2	84.3	93.8	97.0	53.3	74.3	85.6	91.8

Il carico di rottura minimo è di 42.6 ed il massimo è di 107.8 *kg/cm²*

Cominciamo con l'osservare che tutti i valori cadono nell'intervallo Ω dell'asse reale $[40,110]$, in effetti il minimo valore del carico di rottura è 42.6 Kg/cm² mentre il valore massimo è di 107.8 Kg/cm² per cui, per questo esperimento consideriamo il carico di rottura come una variabile aleatoria continua X, con intervallo di definizione Ω. Dividiamo l'intervallo Ω in sottointervalli contigui E_i, $\Omega = \bigcup E_i$ (in questo esempio sono della stessa ampiezza, ma in generale non devono necessariamente esserlo): in questo caso si scelgano $M = 7$ sottointervalli di uguale ampiezza pari a 10; per ogni E_i si conti il numero n_i di dati che vi cadono all'interno. Il numero n_i prende il nome di *frequenza assoluta* dell'evento che il generico risultato X cada nell'intervallo E_i, mentre il rapporto $\pi_i = n_i / N$ prende il nome di *frequenza relativa*, o *rapporto di frequenza* dell'evento $X \in E_i$. Come è noto dalla teoria della probabilità, se N è sufficientemente grande, il rapporto di frequenza π_i è una buona approssimazione della probabilità p_i dell'evento $\{X \in E_i\}$. Si noti che, ovviamente deve risultare che

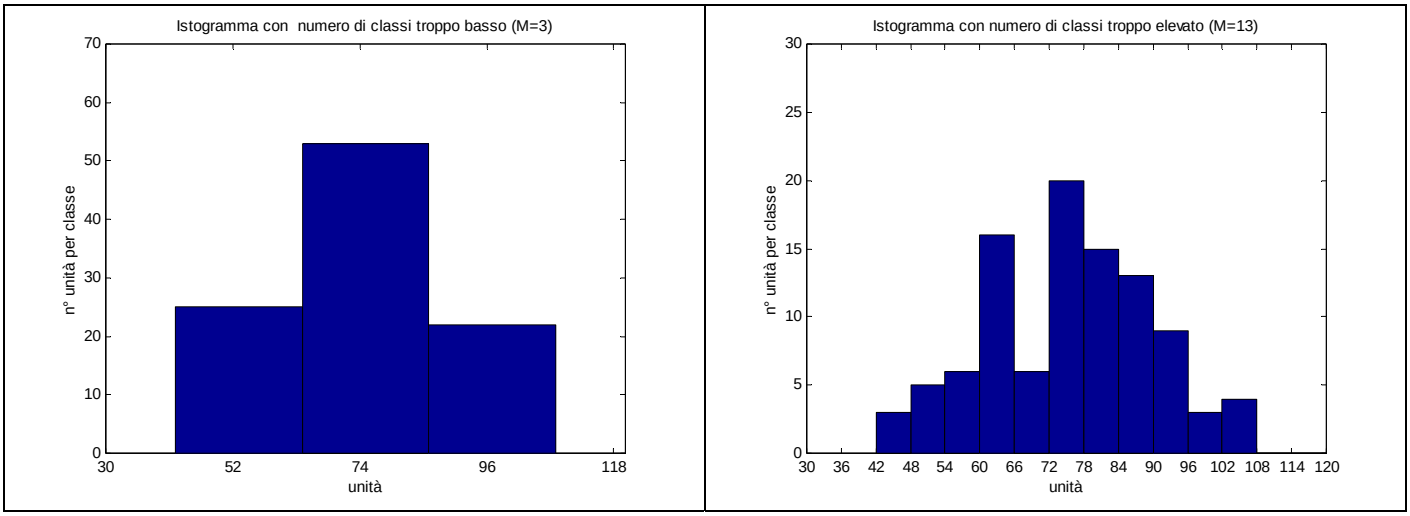
$$\sum_{i=1}^{M} n_i = N, \quad \sum_{i=1}^{M} \pi_i = 1$$

Riportando su un grafico in ascisse i valori della variabile X ed in ordinate i valori delle frequenze relative otteniamo il classico diagramma a barre



che viene detto *istogramma della distribuzione campionaria di* X .

Il numero e l'ampiezza dei sottointervalli devono essere tali che in ognuno di essi cada un numero sufficiente di dati che presentino dei valori sostanzialmente omogenei. Da un lato, un basso numero di sottointervalli lascia i dati ancora troppo raggruppati insieme, dall'altro un'ampiezza troppo piccola del generico sottointervallo non garantisce che questo possa intercettare un numero significativo di dati all'interno dell'insieme dato.



L'istogramma determina una rappresentazione compatta dei dati di partenza, dalla quale è possibile estrarre utili informazioni. Da una semplice ispezione visiva possiamo dire che per il lotto di prodotti considerato, la distribuzione dei valori del carico di rottura si localizza nella classe [70, 80], dove si presenta il massimo della frequenza relativa pari a 0.29. Osserviamo inoltre che i valori di X sono abbastanza addensati intorno alla classe centrale (più dello 0.7 di frequenza relativa nella classe centrale più le due ad essa contigue) e che si distribuiscono in modo simmetrico rispetto ad essa. Quindi, come si vede, possiamo in definitiva individuare una certa regolarità nel processo di produzione che ad un primo esame dei dati non era parsa evidente; l'istogramma è una rappresentazione più maneggevole dei dati iniziali, fornendone una classificazione significativa.

Le qualità dell'istogramma che naturalmente sono risultate rilevanti ai fini della caratterizzazione dell'insieme di dati analizzato, sono suscettibili di una precisa definizione analitica, e quindi di una valutazione quantitativa. Ai fini del calcolo, la variabile aleatoria dell'esempio trattato che riguardava una grandezza a valori nel continuo, può essere approssimata da una variabile aleatoria, che chiameremo ancora X, a valori discreti $\{x_i\}$ rappresentati dalle ascisse dei punti centrali delle classi $\{E_i\}$, assunti con valori di probabilità $\{\pi_i\}$ dati dalle frequenze relative delle classi suddette.

Il valore medio

Questo parametro costituisce una misura della *localizzazione* della distribuzione dei valori, in quanto determina quel valore rispetto al quale si distribuiscono meglio tutti gli altri

$$\mu = \sum_{i=1}^{M} x_i \pi_i$$

dove x_i è l'ascissa del punto centrale della classe, o sottointervallo, E_i. Nell'esempio considerato avremmo

Tab.2. Calcolo della media del carico di rottura μ (kg/cm^2)

i	classe	x_i	π_i	$x_i\,\pi_i$	$x_i - \mu$	$(x_i - \mu)\pi_i$
1	40-50	45	0.04	1.80	-30.6	-1.224
2	50-60	55	0.1	5.50	-20.6	-2.060
3	60-70	65	0.2	13.00	-10.6	-2.120
4	70-80	75	0.29	21.75	-0.6	-0.174
5	80-90	85	0.21	17.85	9.4	1.974
6	90-100	95	0.11	10.45	19.4	2.134

7	100-110	105	0.05	5.25	29.4	1.470
Totale			1	75.60		0.0

ottenendo $\mu = 75.60$ Kg/cm^2.

Altre misure di localizzazione sono la *moda* e la *mediana*. La prima definisce il valore della X per cui si ha un massimo locale della frequenza relativa; se si hanno più massimi locali si parla di distribuzione multimodale. La mediana invece fornisce il valore della X che divide la distribuzione in due classi contigue di frequenza relativa pari 0.5. Nel caso dell'esempio, dalla Tab. 2 notiamo che il valore 0.5 di frequenza relativa viene raggiunto nella classe [70, 80]; infatti la somma delle unità fino alla classe [60, 70] è di 34, per cui la 50-esima unità (cioè l'unità mediana nei 100 pezzi) è la 16-esima della classe [70, 80]; il calcolo della mediana (più facile a farsi che a dirsi) si ottiene nel seguente modo

$$m = 70 + \frac{16}{29}(80 - 70) = 75.517$$

dove 29 sono le unità che compongono la classe [70, 80]. La mediana risulta essere poco sensibile alla variazioni dei termini estremi, in quanto se ai termini della prima metà si sostituiscono termini con valore minore e a quelli della seconda metà termini con valore maggiore, la mediana non cambia.

La varianza

Questa è una misura di quanto i valori della distribuzione siano più o meno dispersi intorno al valor medio

$$\sigma^2 = \sum_{i=1}^{M} (x_i - \mu)^2 \pi_i$$

e si ottiene come valore medio degli scarti al quadrato, per cui *è sempre una quantità positiva*. Un basso valore di σ^2 denota che le determinazioni della X sono addensate intorno al valor medio, e che quindi il processo descritto dalla X ha una bassa variabilità; al contrario un grande valore della varianza, significa che si hanno scarti dal valor medio grandi, e che quindi la X è molto dispersa denotando grande variabilità nel fenomeno allo studio.

Si noti come la varianza sia una grandezza del secondo ordine, per cui in termini di unità di misura non è omogenea alla X ; a tale scopo si è soliti considerare la radice quadrata (positiva) della varianza

$$\sigma = \sqrt{\sigma^2}$$

che prende il nome di *deviazione standard*, od in inglese *root mean square* (rms). Per l'insieme di dati dell'esempio considerato si ha

Tab.3. Calcolo della deviazione standard del carico di rottura σ (kg/cm^2)

i	**classe**	x_i	π_i	$(x_i-\mu)^2$	$(x_i-\mu)^2\pi_i$	***r.m.s***
1	40-50	45	0.04	936.36	37.4544	
2	50-60	55	0.1	424.36	42.4360	
3	60-70	65	0.2	112.36	22.4720	
4	70-80	75	0.29	0.36	0.1044	
5	80-90	85	0.21	88.36	18.5556	
6	90-100	95	0.11	376.36	41.3996	
7	100-110	105	0.05	864.36	43.2180	
Totale			1	75.60	205.64	14.34

ottenendo $\sigma^2 = 205.64$ ($Kg/cm^2)^2$, e $\sigma = 14.34$ Kg/cm^2.

Prima di esaminare altri parametri, vediamo come si modificano valor medio e varianza quando la variabile aleatoria cui si riferiscono subisce delle semplici trasformazioni.

Somma di una costante: $Y = X + c$

$$\mu_Y = \sum_{i=1}^{M} y_i\pi_i = \sum_{i=1}^{M}(x_i + c)\pi_i = \sum_{i=1}^{M} x_i\pi_i + c\sum_{i=1}^{M}\pi_i = \mu_X + c$$

$$\sigma_Y^2 = \sum_{i=1}^{M}(y_i - \mu_y)^2\pi_i = \sum_{i=1}^{M}(x_i + c - (\mu_X + c))^2\pi_i = \sum_{i=1}^{M}(x_i - \mu_X)^2\pi_i = \sigma_X^2$$

da cui si vede che il valor medio varia proprio della costante c addizionata, mentre la varianza resta inalterata.

Prodotto per una costante: $Y = cX$

$$\mu_Y = \sum_{i=1}^{M} y_i \pi_i = \sum_{i=1}^{M} c x_i \pi_i = c \sum_{i=1}^{M} x_i \pi_i = c \mu_X$$

$$\sigma_Y^2 = \sum_{i=1}^{M} \left(y_i - \mu_y \right)^2 \pi_i = \sum_{i=1}^{M} \left(c x_i - c \mu_X \right)^2 \pi_i = c^2 \sum_{i=1}^{M} \left(x_i - \mu_X \right)^2 \pi_i = c^2 \sigma_X^2$$

per cui il valor medio risulta moltiplicato per la stessa costante, mentre la varianza è moltiplicata per la costante al quadrato.

Trasformazione affine: $Y = aX + b$

$$\mu_Y = \sum_{i=1}^{M} y_i \pi_i = \sum_{i=1}^{M} \left(a x_i + b \right) \pi_i = a \sum_{i=1}^{M} x_i \pi_i + b \sum_{i=1}^{M} \pi_i = a \mu_X + b$$

$$\sigma_Y^2 = \sum_{i=1}^{M} \left(y_i - \mu_y \right)^2 \pi_i = \sum_{i=1}^{M} \left((a x_i + b) - (a \mu_X + b) \right)^2 \pi_i = a^2 \sum_{i=1}^{M} \left(x_i - \mu_X \right)^2 \pi_i = a^2 \sigma_X^2$$

L'ultima trasformazione riassume il risultato ottenuto nelle prime due! Questa è importante perché permette di effettuare la standardizzazione di una variabile aleatoria X, ovvero la trasformazione in una variabile aleatoria X' con lo stesso tipo di distribuzione, ma con valor medio nullo e varianza pari ad uno

$$X' = \frac{X - \mu_X}{\sigma_X}$$

che corrisponde ad una trasformazione affine con $a = 1/\sigma_X$ e $b = -\mu_X / \sigma_X$; infatti si ha

$$\mu_{X'} = a\mu_X + b = \frac{1}{\sigma_X}\mu_X + \left(-\frac{\mu_X}{\sigma_X}\right) = 0$$

$$\sigma^2_{X'} = a^2\sigma^2_X = \frac{1}{\sigma^2_X}\sigma^2_X = 1$$

L'importanza della standardizzazione sarà chiara in seguito.

Dissimmetria

Questo parametro dà una misura della dissimmetria della curva della distribuzione rispetto al valore medio, ed è definita nel seguente modo

$$\sum_{i=1}^{M}(x_i - \mu)^3 \pi_i$$

Tuttavia per ottenere un indice adimensionale come indice di dissimmetria (o skewness) si considera la seguente grandezza

$$d = \frac{\sqrt[3]{\sum_{i=1}^{M}(x_i - \mu)^3 \pi_i}}{\sigma}$$

Valori positivi dell'indice denotano che nella distribuzione sono più frequenti scarti positivi dal valor medio; il viceversa vale nel caso di valori di *d* negativi. Per l'esempio trattato si ottiene

Tab.3. Calcolo della skewness del carico di rottura $\boldsymbol{d}$ (*kg/cm*^{2})

i	**classe**	x_i	π_i	$(x_i-\mu)^3$	$(x_i-\mu)^3\pi_i$	**d**
1	40-50	45	0.04	-28652.626	-1146.1046	
2	50-60	55	0.1	-8741.826	-874.1816	
3	60-70	65	0.2	-1191.026	-238.2032	
4	70-80	75	0.29	-0.216	-0.0626	
5	80-90	85	0.21	830.584	174.4226	
6	90-100	95	0.11	7301.384	803.1522	

7	100-110	105	0.05	25412.184	1270.6092	
Totale			1	75.60	-10.3680	-0.1521

con $d = -2.18/14.34 = -0.152\,1$.

Curtosi (o indice di eccesso)

Questo parametro non è molto usato nel nostro ambito, viene riportato per completezza. Per distribuzioni unimodali, e simmetriche misura il grado di appiattimento della distribuzione intorno al valor medio o, corrispondentemente, l'assottigliamento delle code della distribuzione. E' definito nel seguente modo

$$C = \frac{\sum_{i=1}^{M} (x_i - \mu)^4 \pi_i}{\sigma^4} - 3$$

e vale zero per una distribuzione gaussiana. Una distribuzione con Curtosi positiva vuol dire che ha una distribuzione più appuntita e concentrata intorno al valor medio rispetto ad una gaussiana di pari valor medio e varianza; si dice anche in questo caso che la distribuzione presenta un eccesso positivo rispetto alla gaussiana. Nel caso opposto la distribuzione apparirà più appiattita e dispersa in corrispondenza del valor medio rispetto ad una gaussiana.

Vediamo come la considerazione di questi semplici elementi descrittivi di una distribuzione possa permettere di orientarci tra varie strategie di intervento. Consideriamo un tipico esempio in cui un'azienda debba decidere tra diversi tipi di investimento nei riguardi della produzione e vendita di un certo prodotto. Nella seguente tabella si riportano, per ognuna delle cinque strategie, le previsioni per gli utili annui x_i (in euro) e la distribuzione delle frequenze relative $\{\pi_i\}$

1		**2**		**3**		**4**		**5**	
X_1	$\{\pi_i\}_1$	X_2	$\{\pi_i\}_2$	X_3	$\{\pi_i\}_3$	X_4	$\{\pi_i\}_4$	X_5	$\{\pi_i\}_5$
0	0.8	2000	0.1	0	0.3	4000	1	1000	0.2
20000	0.2	3000	0.3	3000	0.4			2000	0.7
		4000	0.4	5000	0.2			8000	0.1
		5000	0.2	7000	0.1				
1			1		1		1		1

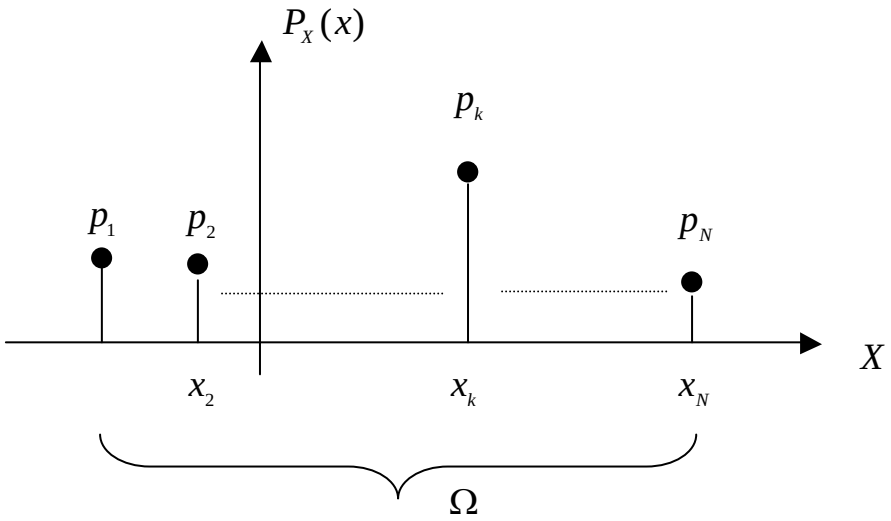
Dalla semplice ispezione dei dati non riusciamo facilmente a decidere quale strategia sia preferibile alle altre. Una buona politica è quella di scegliere l'investimento che mediamente comporti i maggiori guadagni. Se calcoliamo il valore medio dei dati presenti in ciascuna colonna della tabella, si ottengono i seguenti *guadagni medi annui*

Strategia 1	$\mu_{X_1} = 4000$
Strategia 2	$\mu_{X_2} = 3700$
Strategia 3	$\mu_{X_3} = 2900$
Strategia 4	$\mu_{X_4} = 4000$
Strategia 5	$\mu_{X_5} = 2400$

Si vede che le strategie migliori dal punto di vista dei guadagni medi annui previsti sono la prima e la quarta, ma quale scegliere tra le due? A questo punto dobbiamo valutare la variabilità dei dati relativi alle due politiche di investimento. Nel caso della strategia n° 4 la variabilità è nulla: quindi in questo caso si guadagna mediamente 4000 euro senza alcun rischio. Per la strategia n° 1 si ottiene $\sigma_{X_1} = 8000$, pari al doppio del valor medio; questo indica una estrema variabilità che rende questa strategia molto rischiosa: nel 20% dei casi potremmo guadagnare molto, 20000 euro, ma nell'80% dei casi i guadagni attesi potrebbero essere nulli. Chi ama rischiare sceglierà la strategia n°1, mentre chi vuole assicurarsi sceglierà la strategia n° 4.

Questo semplice esempio mostra come i parametri della distribuzione campionaria dei dati, determinino una rappresentazione concisa dell'informazione contenuta in essi, ed utile alla definizione di opportune alternative di decisione nei riguardi di un dato problema.
In molti casi pratici tuttavia l'uso dell'istogramma non è molto agevole e risulta più utile poter sostituire alla distribuzione empirica un'opportuna distribuzione analitica ad essa equivalente.
Riportiamo quindi di seguito le distribuzioni più utilizzate e le loro proprietà!

Nella introduzione della distribuzione campionaria dei dati abbiamo visto come per una v.a. discreta con un numero finito N di possibili valori $x_1,\ldots,x_N$, la distribuzione è rappresentata da N valori (masse concentrate) $p_1,\ldots,p_N$



Naturalmente gli N valori p_k devono essere tali che la massa totale valga 1, cioè $\sum_{k=1}^{N} p_k = 1$.

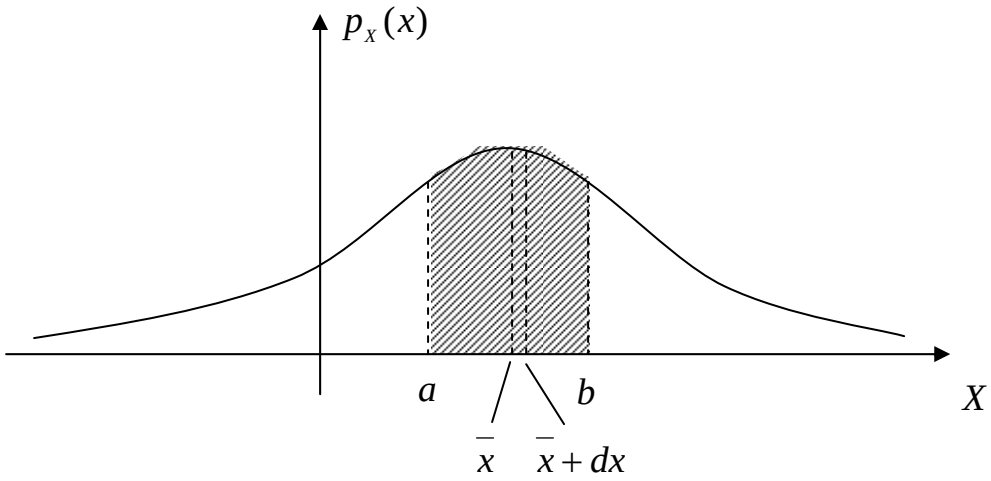
Nel caso di v.a. continua la legge di distribuzione è data da una funzione $p_X(x)$ che chiameremo funzione di *densità di probabilità* . Tale denominazione deriva dal fatto che per il generico valore ammissibile $\bar{x}$ essa fornisce la probabilità dell'evento $E = \{X \in (\bar{x}, \bar{x} + dx)\}$ di lunghezza infinitesima dx intorno al punto considerato

$$P_X(E) = p_X(\bar{x})dx$$

Considerato poi un qualunque altro evento A rappresentato ad esempio da un intervallo (a,b) di lunghezza finita, la sua probabilità si ottiene "sommando" tutti i termini infinitesimi del tipo precedente relativi ai punti che compongono tale intervallo

$$P_X(A) = \int_a^b p_X(x)dx$$

Da un punto di vista geometrico il calcolo precedente corrisponde a calcolare l'area campeggiata in figura contenuta tra l'intervallo (a,b) ed il tratto della curva $p_X(x)$ da esso individuato!



Come più volte precisato, non è tanto importante poter calcolare la probabilità di un qualsiasi evento legato ad una v.a. quanto caratterizzare la legge di distribuzione con un numero limitato di parametri che ne descrivano il carattere globale. Questi parametri sono dati dai *momenti* della distribuzione, e sono una misura delle seguenti caratteristiche.

1. **Valor medio.** E' il momento del primo ordine e si calcola nel seguente modo

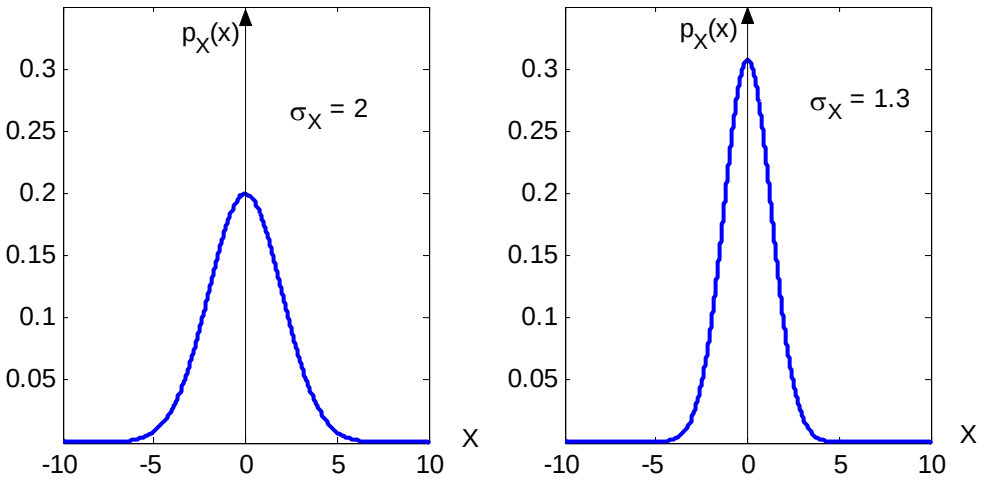
$$\mu_X = E(X) = \int_\Omega x\, p_X(x)\, dx, \quad \mu_X = \sum_{k=1}^{N} x_k\, p_k$$

nel caso continuo e nel caso discreto. Fornisce il baricentro della distribuzione, ovvero quel valore compreso in Ω rispetto al quale si ripartiscono in modo equilibrato i valori assunti dalla X .

2. **Varianza.** E' il momento centrato del secondo ordine: ovvero detta $\overline{X} = X - E(X)$ la v.a. centrata, ovvero lo scarto della v.a. rispetto al suo valor medio, la varianza è data da

$$\sigma_X^2 = E(\overline{X}^2) = \int_\Omega (x-\mu_X)^2\, p_X(x) dx, \quad \sigma_X^2 = \sum_{k=1}^{N} (x_k-\mu_X)^2\, p_k$$

In altre parole la varianza è lo scarto quadratico medio rispetto a μ_X ed è una misura della dispersione dei valori della X : una varianza grande indica che possiamo trovare con buona probabilità valori della X lontani dal valor medio, mentre una bassa varianza vuol dire che i



valori della v.a. sono addensati intorno al valor medio e valori distanti da esso occorrono con bassa probabilità. Dalle figure vediamo che la prima distribuzione è molto più dispersa intorno al valor medio della seconda distribuzione; la prima ha certamente una varianza maggiore della seconda.

Tuttavia osserviamo che parlare di entità della dispersione in assoluto non ha molto senso; infatti occorre rapportare il valore della varianza all'entità del valor medio: per esempio se considerassimo due distribuzioni con stessa varianza pari 10, ma una con valor medio pari a 20 e l'altra con valor medio pari 1000 vedremmo che l'entità dello scarto rispetto al valor medio sarebbe nel primo caso del 50%, nel secondo caso del 1%. Si è soliti quindi introdurre un fattore di forma della curva della distribuzione che valuta l'entità della varianza rispetto al valor medio

$$cv_X = \frac{\sqrt{\sigma_X^2}}{\mu_X} = \frac{\sigma_X}{\mu_X}, \quad \mu_X \neq 0$$

che viene detto coefficiente di variazione della v.a. X ; in esso compare la radice quadrata della varianza per poter confrontare grandezze omogenee, e prende il nome di *deviazione standard*. Il cv_X in definitiva fornisce lo scarto medio dei valori della X rispetto alla media in percentuale del valore della media stessa: un valore del 1% indica una distribuzione molto

concentrata intorno al valor medio, mentre un valore del 50% indica che i valori di X possono essere anche abbastanza lontani da μ_X .

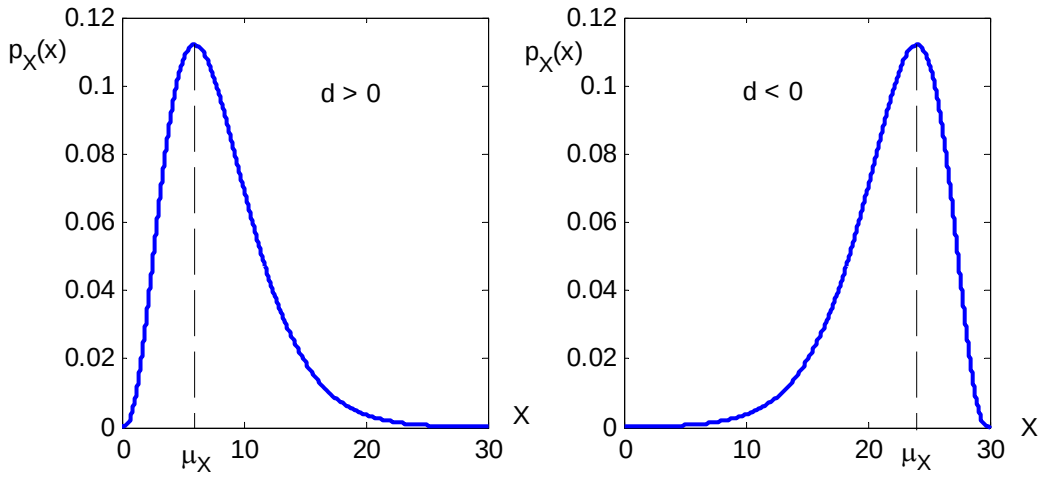
3. **Skewness (simmetria).** Un altro elemento importante del carattere di una distribuzione è costituito dal fatto che i valori della v.a. X si distribuiscano in modo simmetrico rispetto alla media. Questo comporta che si debbano ritenere equiprobabili sia scarti positivi che scarti negativi rispetto al valor medio. In caso contrario significa che preferibilmente i valori della v.a. saranno a destra o a sinistra della media. Il grado di simmetria si misura considerando il valor medio centrato del terzo ordine

$$E(\overline{X}^3) = \int_{\Omega} (x - \mu_X)^3 p_X(x) dx, \quad E(\overline{X}^3) = \sum_{k=1}^{N} (x_k - \mu_X)^3 p_k$$

ed è dato dal seguente parametro

$$d = \frac{\sqrt[3]{E(\overline{X}^3)}}{\sigma_X}$$

che prende appunto il nome di skewness. Le distribuzioni simmetriche hanno skewness nulla; un valore positivo di d indica che i valori della X si distribuiscono principalmente a destra della media m_X , cioè sono più frequenti scarti positivi rispetto a μ_X ; al contrario se d è negativa significa che sono più frequenti scarti negativi rispetto a μ_X

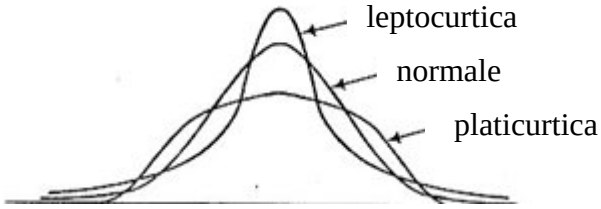


4. **Tightness (Curtosi).** Questo parametro misura quanto una distribuzione si discosti da una gaussiana; in particolare dice se le code della distribuzione vadano zero più o meno rapidamente di quelle di una gaussiana. Si deve considerare il momento centrato del quarto ordine

$$E(\overline{X}^4) = \int_{\Omega} (x - \mu_X)^4 p_X(x) dx, \quad E(\overline{X}^4) = \sum_{k=1}^{N} (x_k - \mu_X)^4 p_k$$

da cui

$$C = \frac{E(\overline{X}^4)}{\sigma_X^4} - 3$$



Per una gaussiana C vale 0; se una distribuzione ha $C > 0$ si dice ipergaussiana (leptocurtica) ed intorno alla media è più appuntita di una gaussiana ed ha code più alte, cioè che vanno meno rapidamente a zero di quelle di una gaussiana ; viceversa essa si dice ipogaussiana (platicurtica) e risulta di andamento più dolce intorno alla media, ma con code che vanno rapidamente a zero.

5. **Percentili.** Si è detto che nota la distribuzione di una v.a. X è possibile calcolare la probabilità di un qualunque evento legato ad essa. Tuttavia interessano in pratica solo alcuni tipi di eventi, che vengono utilizzati nei test di ipotesi. In particolare si è interessati a eventi del tipo

$$\left\{ \frac{|X - m_x|}{\sigma_X} > \lambda_\varepsilon \right\}$$

con probabilità

$$P\left(\frac{|X - m_x|}{\sigma_X} > \lambda_\varepsilon \right) = \varepsilon\%$$

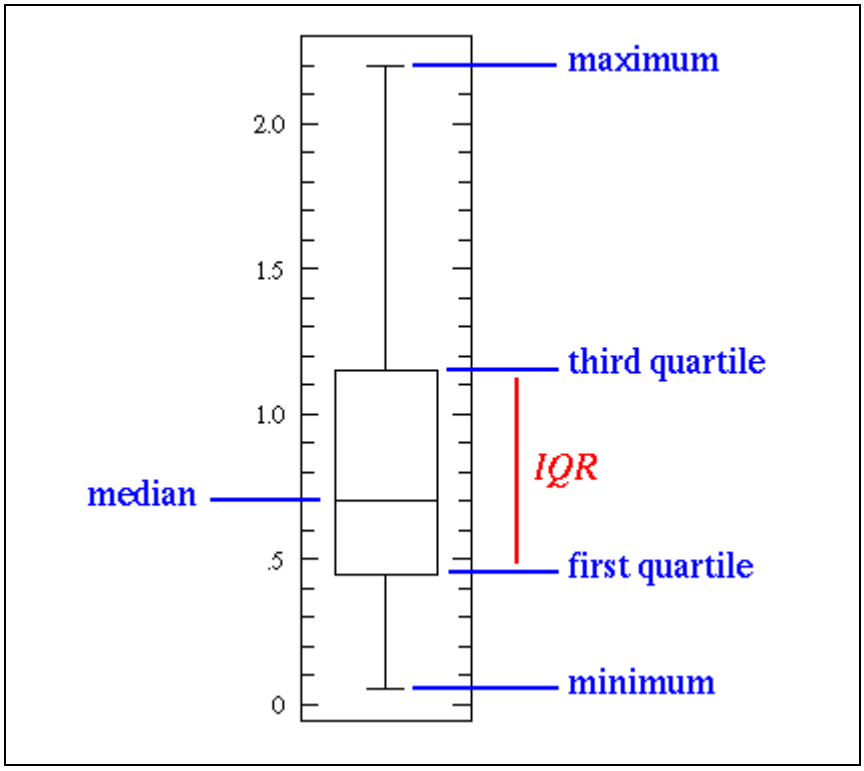
La precedente relazione va utilizzata specificando il valore $\varepsilon\%$ della probabilità e calcolando il valore dell'ascissa λ_ε per cui l'evento considerato ha probabilità appunto $\varepsilon\%$. Si noti che il λ_ε viene calcolato sempre con riferimento alla v.a. standardizzata

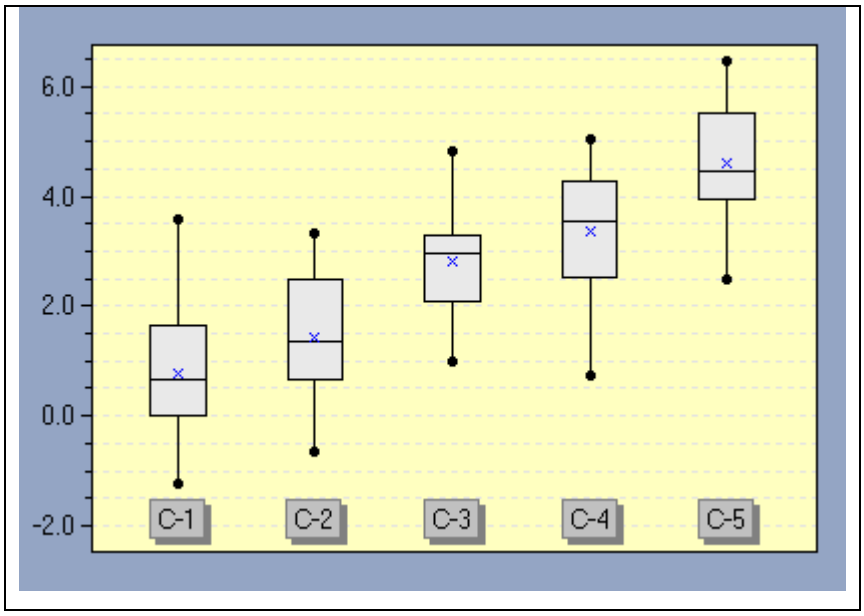
$$\frac{X - m_X}{\sigma_X}$$

Tali ascisse vengono dette *percentili* della distribuzione in quanto sono i valori della X per cui l'evento del tipo considerato ha un assegnato $\varepsilon\%$ di probabilità! Questi vengono forniti in tabelle disponibili su tutti i testi di statistica per le distribuzioni normalmente usate nei test statistici, come la gaussiana, la χ^2, la t-Student e la F-Fisher.

Prima di passare in rassegna le distribuzioni di più largo uso, illustriamo un altro metodo per rappresentare in maniera concisa le proprietà statistiche di un insieme di dati .

Box Plot. In tale rappresentazione vengono riportati: la mediana, i percentili $\lambda_{0.25}$ e $\lambda_{0.75}$ (si chiamano anche *quartili* per via che corrispondono a valori di probabilità multipli di 0.25) i valori massimo e minimo dei dati, eventuali outliers.





Se la mediana non è equidistante dal primo e terzo quartile la distribuzione non è simmetrica.
Il Box Plot permette un rapido confronto tra le proprietà statistiche principali di più insiemi di dati che possano riguardare uno stesso fenomeno.

La distribuzione gaussiana.

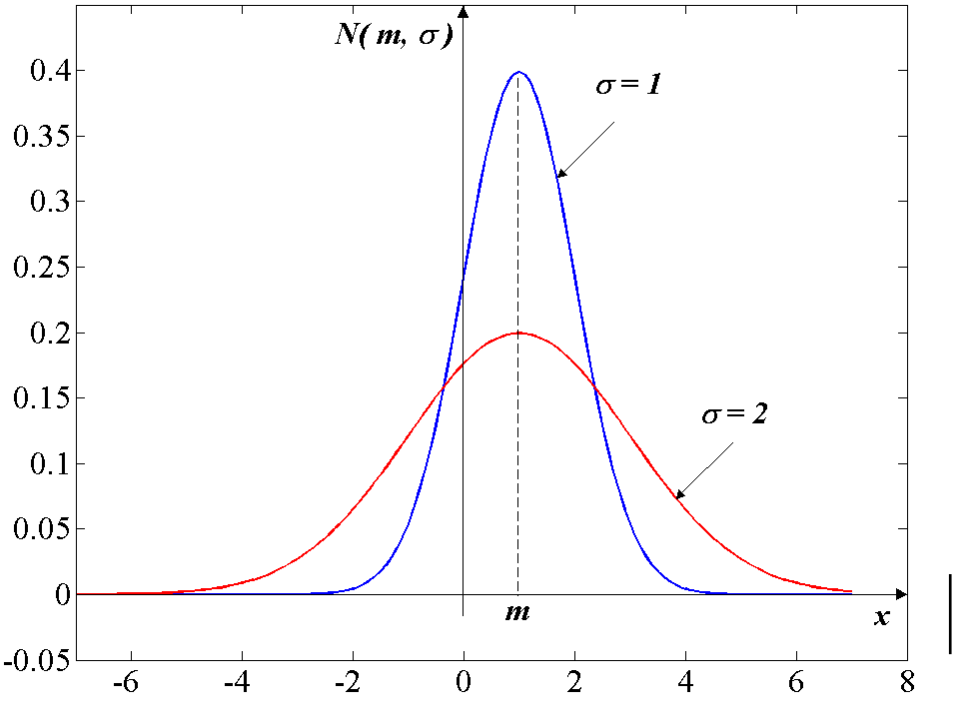
Tale distribuzione assume valori in $(-\infty, \infty)$ ed è completamente caratterizzata dal valor medio m e dalla varianza σ^2; viene detta anche distribuzione normale ed indicata con il simbolo $N(m,\sigma)$

$$N(m,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

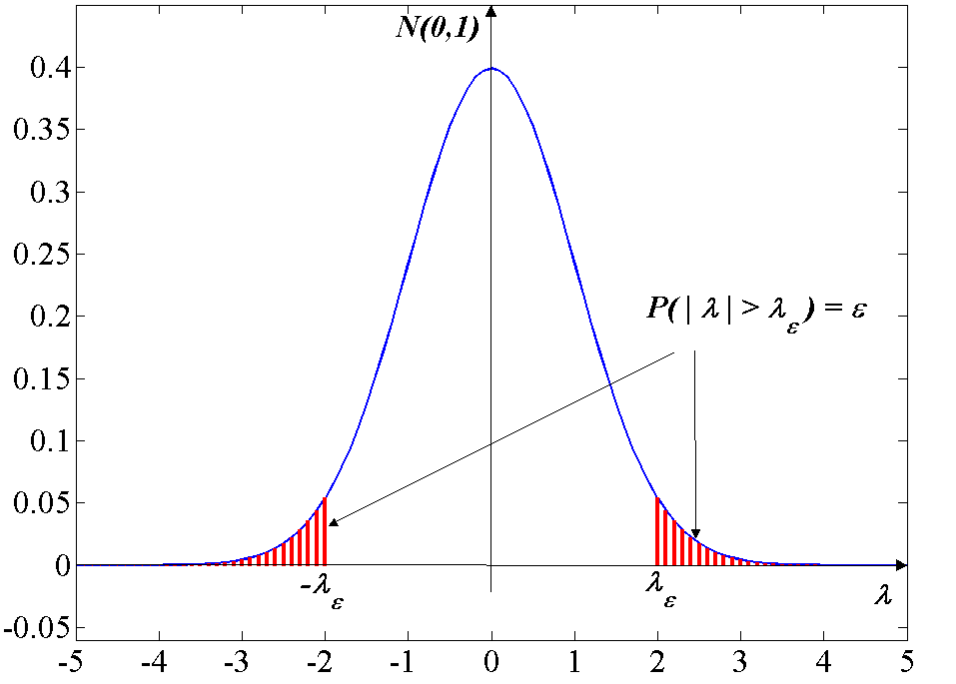
E' una distribuzione simmetrica ($\mu_3 = 0$); in particolare tutti i momenti centrati dispari sono nulli, mentre per quelli pari si ha

$$E(\overline{X}^{2k}) = 1\cdot 3\cdot 5\cdot\cdots\cdot(2k-1)\cdot(\sigma^2)^k, \quad k = 1,2,3,\ldots$$

da cui si vede subito che $E[\overline{X}^4] = 3.$



I percentili vengono tabulati in riferimento alla v.a. standardizzata $N(0,1)$



La distribuzione χ^2.

Consideriamo n v.a. ξ_i gaussiane standard $N(0,1)$ indipendenti; la v.a. χ^2 è definita nel seguente modo

$$\chi^2 = \sum_{i=1}^{n} \xi_i^2$$

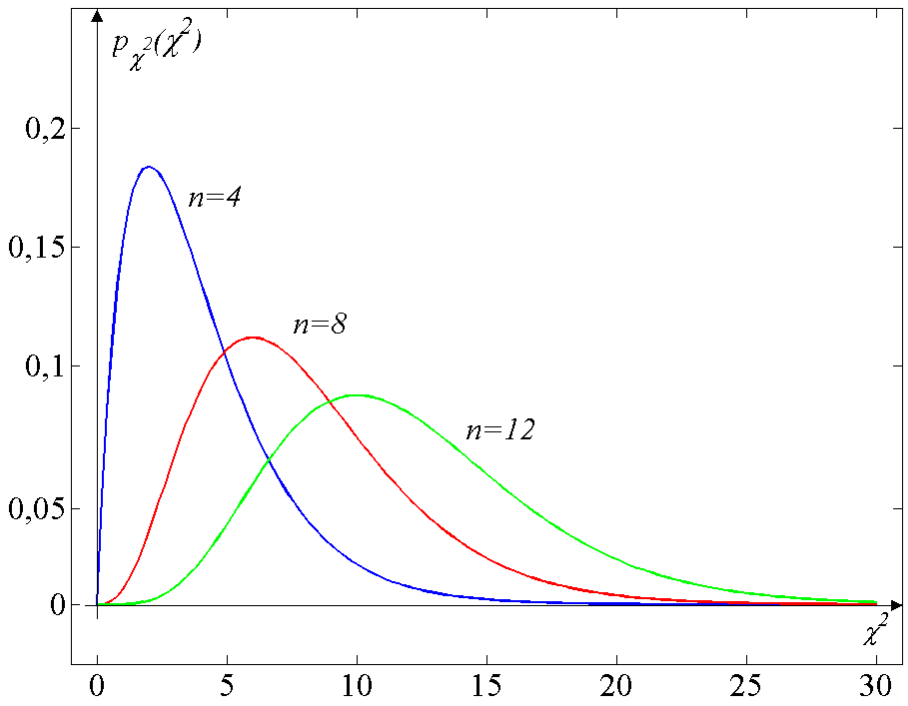
Il parametro n definisce il numero dei gradi di libertà della v.a., che assume valori in $(0, \infty)$. E' generalmente una distribuzione non simmetrica, che tende a diventare simmetrica all'aumentare di n. Essa ha andamento monotono decrescente per $n \le 2$, mentre per $n > 2$ è unimodale con il massimo di ascissa $(n-2)$

$$p_{\chi^2}(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, \quad y > 0$$

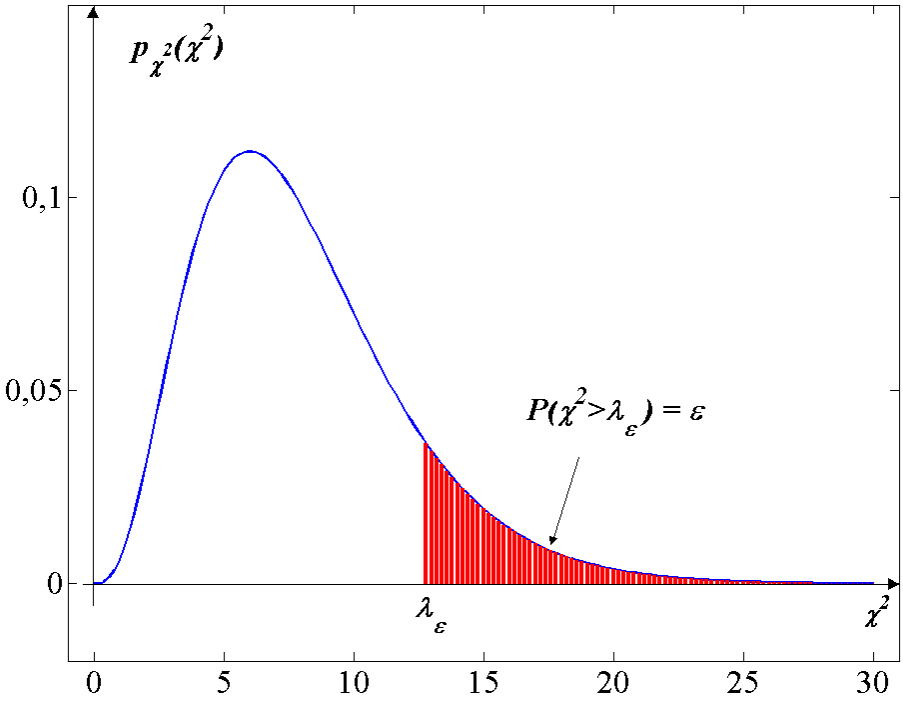
dove $\Gamma(\cdot)$ è la funzione speciale *gamma-euleriana*. Essa possiede i momenti di qualunque ordine

$$E(y^k) = n(n+2)\cdots\cdots(n+2k-2)$$

per cui il valor medio è $m = n$ e la varianza $\sigma^2 = 2n$



I percentili vengono tabulati per numero crescente di gradi di libertà e si riferiscono ad eventi del tipo $(\chi^2 \geq \lambda_\varepsilon)$



La distribuzione di Student.

Consideriamo $n+1$ v.a. gaussiane indipendenti $x, \xi_1, \xi_2, \cdots \xi_n$ tutte $N(0,\sigma)$ e costruiamo la seguente v.a

$$t = \frac{x}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \xi_i^2}}$$

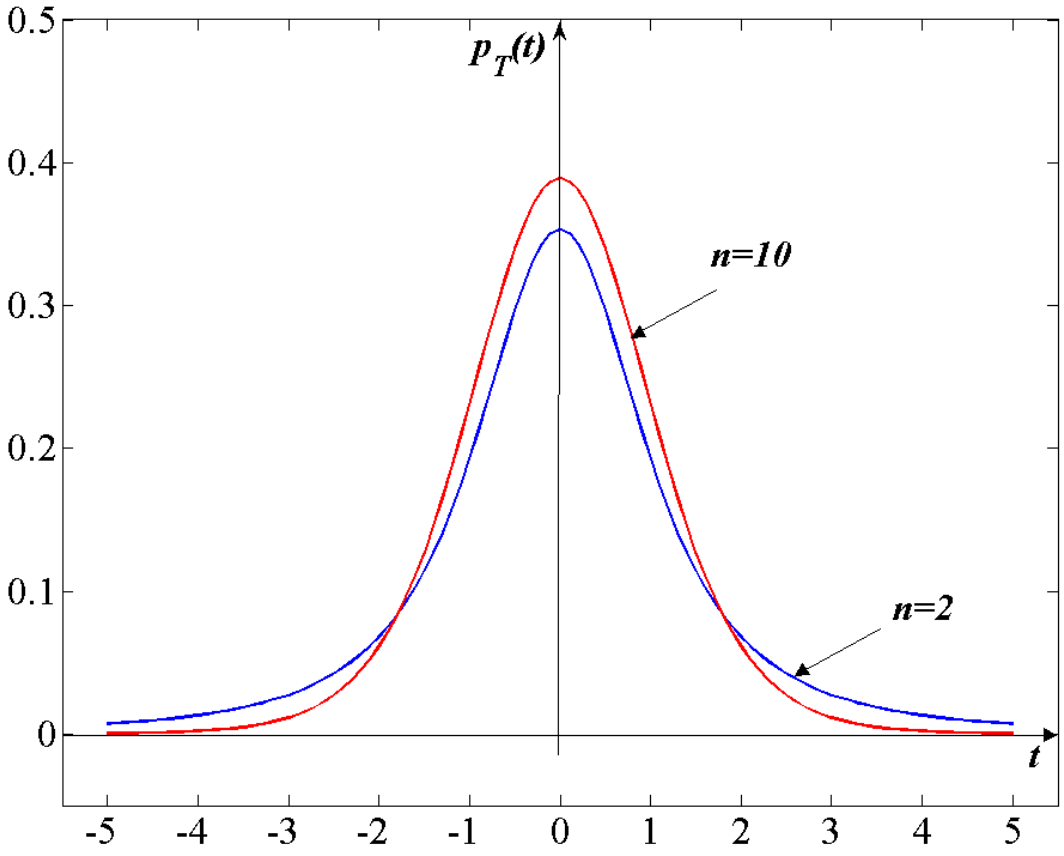
che prende il nome di distribuzione $t-$Student a n gradi di libertà ed ha la seguente densità di probabilità

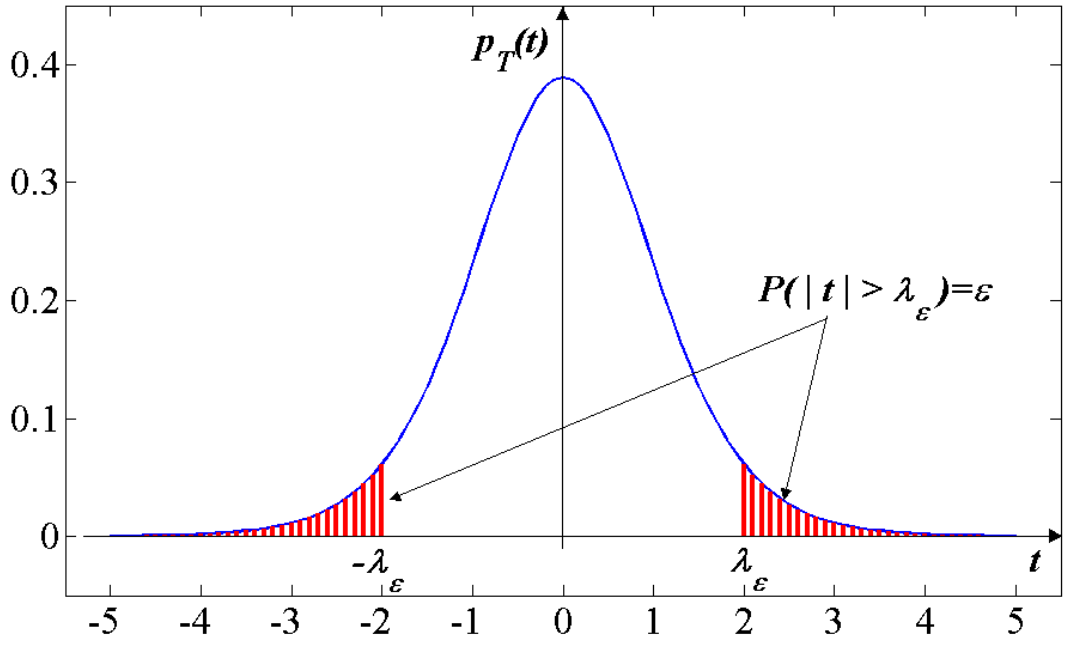
$$p_T(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \geq 0$$

E' importante notare che la distribuzione è indipendente dalla varianza σ^2 delle componenti. Essa ha valor medio nullo ed è simmetrica, quindi con tutti i momenti dispari nulli, con i momenti pari (per $n > 2$) dati da

$$E(t^{2k}) = \frac{1 \cdot 3 \cdot \cdots \cdot (2k-1) n^k}{(n-2)(n-4)\cdots(n-2k)}$$

per cui la varianza vale $\sigma^2 = n/(n-2)$





Al solito i percentili vengono tabulati per numero crescente di gradi di libertà e si riferiscono ad eventi del tipo $(|t| > \lambda_\varepsilon)$.

La distribuzione di Fisher.

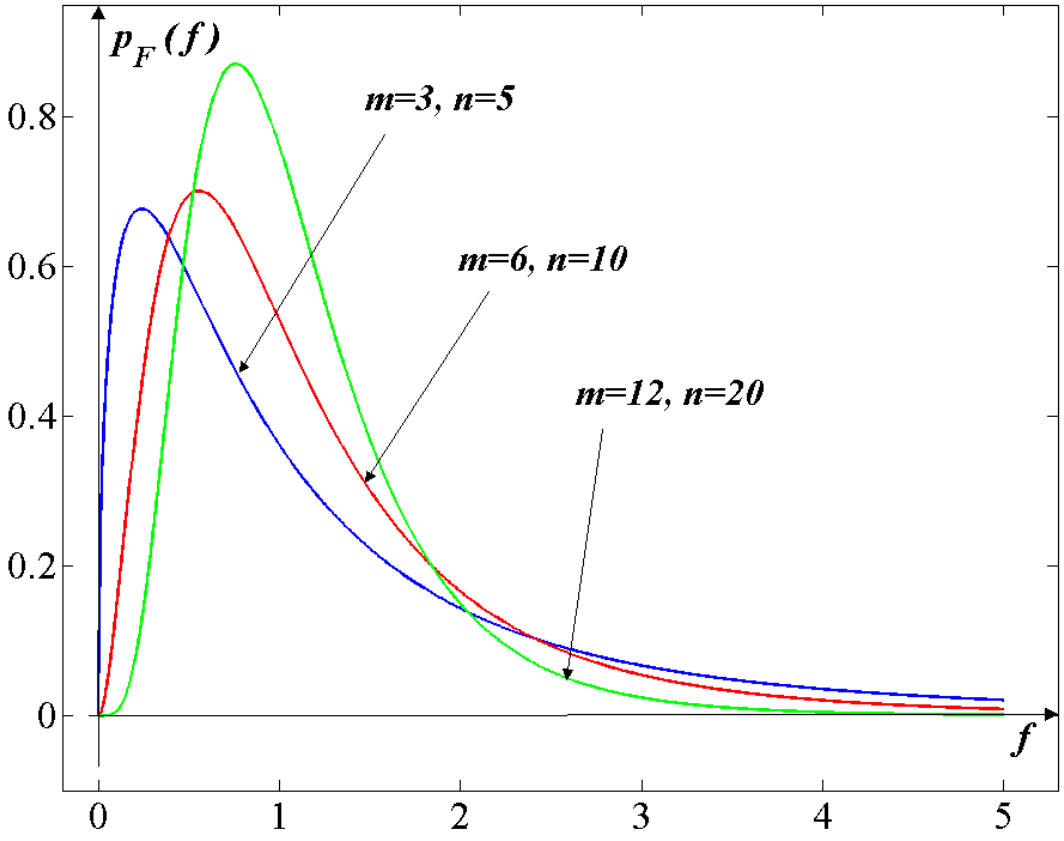
Consideriamo $m+n$ v.a. gaussiane indipendenti $\xi_1,\cdots,\xi_m,\eta_1,\cdots,\eta_n$ tutte $N(0,\sigma)$. La v.a.

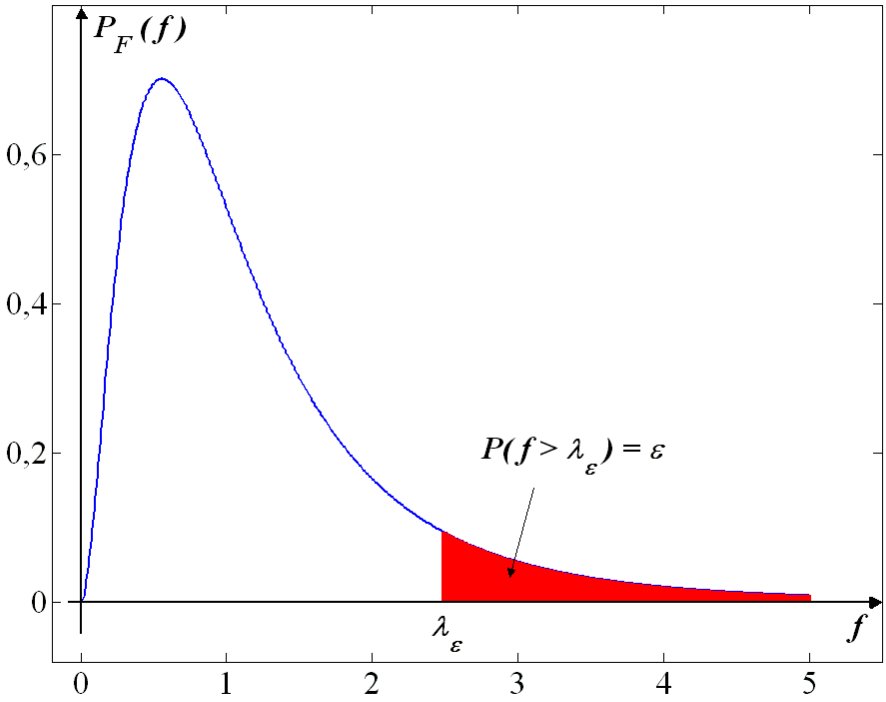
$$F = \frac{\frac{1}{m}\sum_{i=1}^{m}\xi_i^2}{\frac{1}{n}\sum_{j=1}^{n}\eta_j^2}$$

ha distribuzione che prende il nome di distribuzione di Fisher, che risulta indipendente dalla varianza delle componenti. Assume valori in $(0,\infty)$ con valor medio e varianza dati da

$$E(F)=\frac{n}{n-2},\qquad n>2$$

$$\sigma_F^2=\frac{2n^2(n+m-2)}{m(n-2)^2(n-4)},\quad n>4$$





I percentili vengono tabulati per diversi valori di m ed n.

Il teorema fondamentale della convergenza stocastica

Questo teorema è anche noto con il nome di “teorema del limite centrale”. Sia $\{X_k\}$ una successione di v.a. *indipendenti* con

$$E[X_k] = \mu_k, \qquad E\left[\overline{X_k}^2\right] = \sigma_k^2$$

e si consideri la seguente v.a.

$$S_n = \sum_{k=1}^{n} X_k \qquad con \qquad \mu_{S_n} = \sum_{k=1}^{n} \mu_k, \quad \sigma_{S_n}^2 = \sum_{k=1}^{n} \sigma_k^2$$

Se valgono le seguenti ipotesi

- $\lim_{n\to\infty} \sigma_{S_n}^2 = \infty$
- $E\left[X_k^{\alpha}\right] \le C, \quad \alpha > 2$

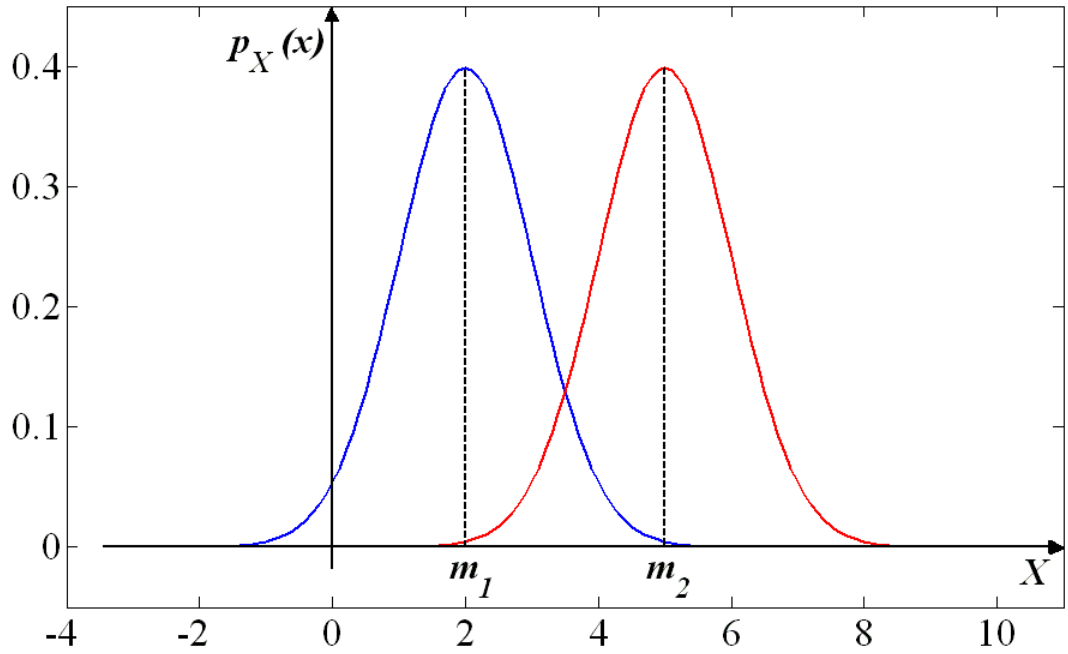
allora la distribuzione della v.a. standardizzata $(S_n - \mu_{S_n})/\sigma_{S_n}$ per $n \to \infty$ tende ad una gaussiana standard N(0,1).

Il risultato di questo teorema è di notevole interesse in quanto stabilisce che, sotto le due ipotesi fatte, la somma di un numero elevato di v.a. indipendenti tende a distribuirsi come una gaussiana, indipendentemente dalla distribuzione delle singole componenti. Le due ipotesi in pratica stabiliscono che la somma deve essere determinata da variabili indipendenti (prima ipotesi) nessuna predominante rispetto alle altre (seconda ipotesi). In particolare la seconda ipotesi è verificata se le v.a. componenti hanno tutte la stessa distribuzione.

Questo teorema assegna una importanza centrale della distribuzione gaussiana nell’insieme delle leggi di distribuzione della variabili aleatorie e spiega il largo impiego della distribuzione gaussiana come modello statistico nella maggior parte dei processi considerati nell’ambito ingegneristico!

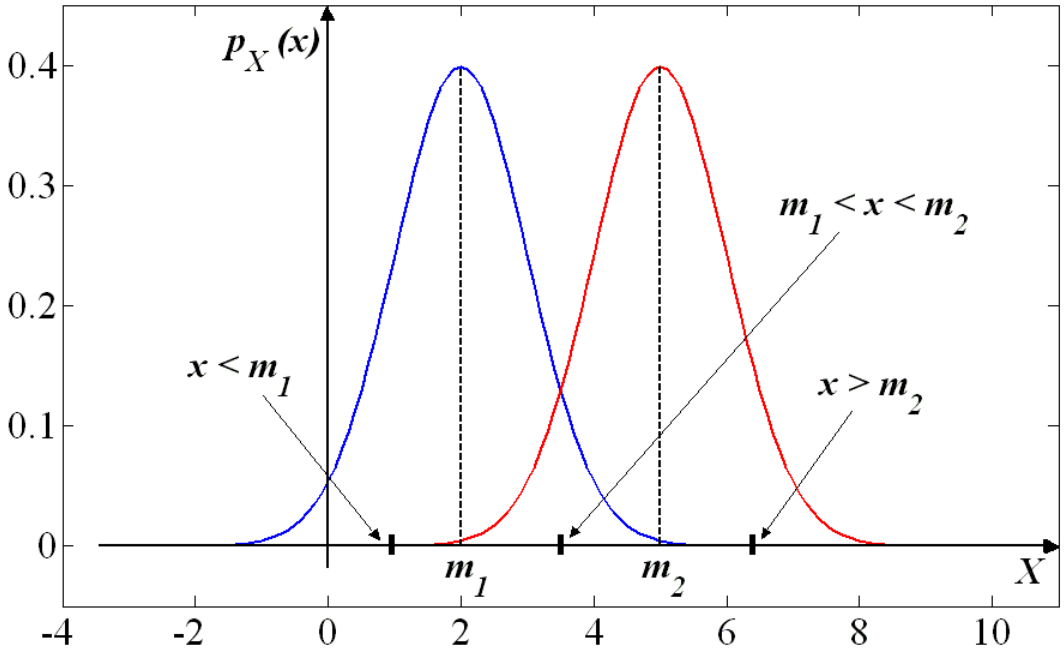
Test statistici di ipotesi

I test statistici consistono in procedure per validare ipotesi di modello riguardanti le caratteristiche statistiche di dati sperimentali ottenuti come risultati di un dato fenomeno aleatorio. Un esempio chiarirà meglio il senso del problema. Sia X una variabile aleatoria che descriva i valori di un certo attributo di un prodotto o di un servizio, ed abbia distribuzione gaussiana con varianza nota σ^2 e valor medio m incognito. Tutto quello che si sa a proposito del valor medio è che potrebbe avere o un valore m_1 oppure un valore m_2. Un caso come questo si può presentare ad esempio se si valuta la risposta di un test per un tipo di virus influenzale in una popolazione: la risposta dei soggetti sani e dei soggetti infetti avrà grosso modo la stessa variabilità, ma le risposte dei soggetti infetti devono localizzarsi intorno ad un valore medio significativamente differente dal valor medio della risposta dei soggetti sani (il test è tanto più discriminante quanto più questa differenza è accentuata). Ora andiamo noi stessi a fare l'analisi per vedere se abbiamo preso l'influenza: il risultato del test fornirà quindi un dato sperimentale x (il nostro!) della variabile aleatoria X.



Si vuole decidere se sia più verosimile che il dato osservato derivi dalla distribuzione con media m_1 (e quindi saremmo sani) oppure dalla distribuzione con media m_2 (e quindi saremmo infettati dal virus). L'ipotesi che $m = m_1$ viene chiamata *ipotesi nulla*, ed indicata con H_0, mentre l'ipotesi che $m = m_2$ viene chiamata *ipotesi alternativa*, ed indicata con H_1 (ovviamente si poteva

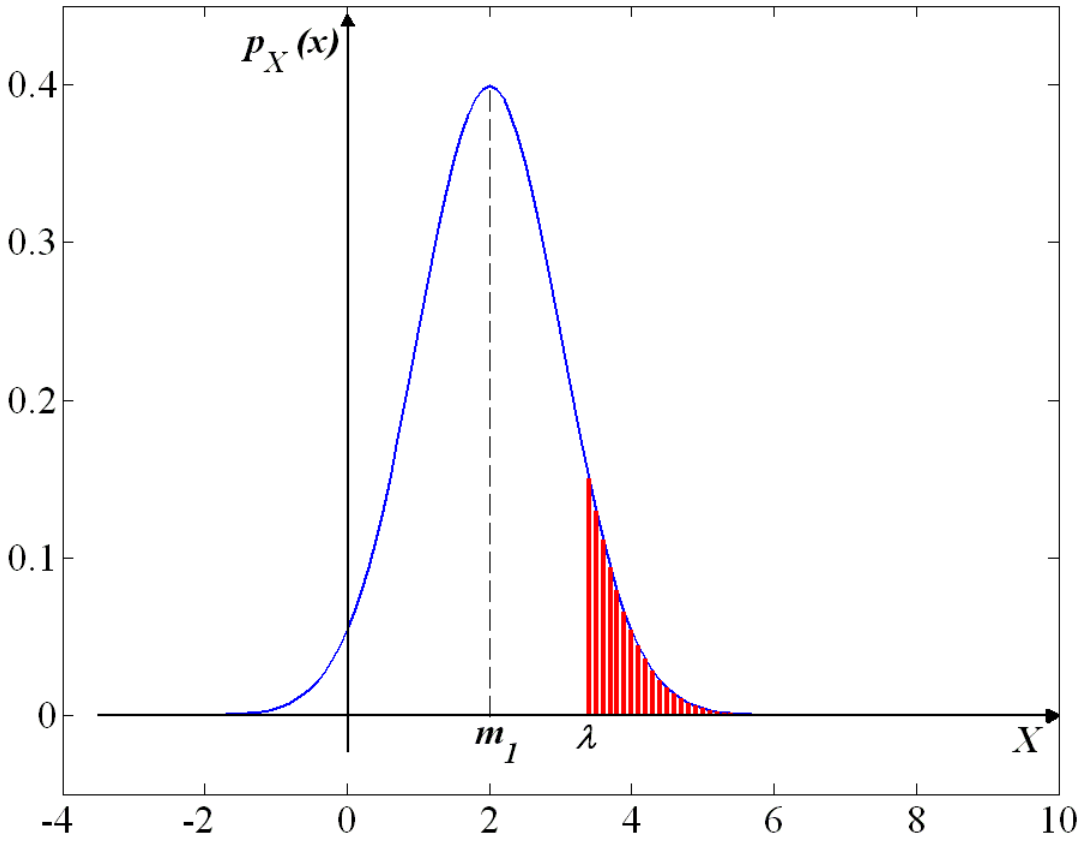
scegliere anche nell'altro modo). Come facciamo a prendere tale decisione? Tutto dipende da dove si localizza il dato osservato.



Nel caso in cui $x < m_1$ potremmo senza dubbio accettare l'ipotesi H_0 e ritenerci sani perché la distribuzione con media m_2 assegna ad x un valore di probabilità veramente trascurabile (è ben dentro la coda della distribuzione in rosso) rispetto a quello assegnato dalla distribuzione con media m_1. Ricordiamo che questo valore di probabilità è approssimato come $p_X(x) \cdot \Delta$, dove Δ è un piccolo intorno di x. Allo stesso modo, nel caso in cui fosse $x > m_2$, con analogo ragionamento, potremmo certamente accettare l'ipotesi H_1 e ritenerci infetti.

Nel caso intermedio $m_1 < x < m_2$ le cose sono meno ovvie; entrambi le distribuzioni assegnano valori di probabilità confrontabili per cui dobbiamo stabilire un valore λ di X compreso tra m_1 e m_2 per cui se $x \leq \lambda$ accettiamo H_0 e rifiutiamo H_1 (notiamo che questa regola comprende anche i valori $x < m_1$), se $x > \lambda$ rifiutiamo H_0 e accettiamo H_1 (questa regola include anche i valori $x > m_2$). Tuttavia, comunque venga scelto λ, si può notare che le regole precedenti comportano il rischio di prendere una decisione sbagliata. Infatti, se risulta $x > \lambda$ si rifiuta H_0; ma osservando la figura seguente si vede come la distribuzione che corrisponde all'ipotesi nulla assegna all'evento $x > \lambda$ una probabilità finita data dall'area della zona campeggiata in rosso. Questo

significa che se l'ipotesi H_0 è vera, per cui la media della distribuzione è effettivamente m_1, c'è comunque una probabilità non trascurabile di poter ottenere dati sperimentali di ampiezza più grande del valore λ, che quindi ci farebbero rifiutare l'ipotesi H_0. Tale errore viene detto *errore di tipo 1*: si rifiuta H_0 quando è vera.



La probabilità di commettere questo errore è data dalla probabilità dell'evento $x>\lambda$ sotto l'ipotesi H_0

$$P(x>\lambda\,|\,H_0)=\frac{1}{\sqrt{2\pi}\sigma}\int_{\lambda}^{\infty} e^{\frac{-(x-m_1)^2}{2\sigma^2}}\,dx$$

e viene detta *livello di significatività del test*. L'insieme dei valori $\{x>\lambda\}$ per cui si rifiuta l'ipotesi nulla prende il nome di *set critico del test*.

Normalmente in un test di ipotesi si fissa il livello di significatività ε del test (usualmente 0.05, in alcuni casi 0.01), quindi si determina il set critico. Dobbiamo quindi trovare il valore di λ per cui risulti

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\lambda}^{\infty} e^{\frac{-(x-m_1)^2}{2\sigma^2}}\,dx=\varepsilon$$

Questo può essere ottenuto facilmente esprimendo la distanza $\lambda - m_1$ secondo la scala tipica di variazione di X, e cioè in unità di deviazione standard

$$\lambda - m_1 = \sigma\lambda_0$$

Ora quindi dobbiamo trovare il valore di λ_ε per cui risulti $P(x>\lambda\,|H_0)=P(x>m_1+\sigma\lambda_0)=\varepsilon$. Ma questo è facilmente ottenibile dalla tabella dei percentili di una gaussiana. Infatti l'evento

$$x>m_1+\sigma\lambda_0$$

equivale al seguente

$$\frac{x-m_1}{\sigma}>\ \lambda_0$$

e quindi il valore di λ_0 per cui

$$P(x>m_1+\sigma\lambda_0)=P(\frac{x-m_1}{\sigma}>\lambda_0)=\varepsilon$$

è proprio il percentile $\lambda_{2\varepsilon}$ della gaussiana (si ricordi che per le distribuzioni simmetriche usualmente i percentili sono tabulati in corrispondenza agli eventi bilaterali $\frac{|x-m_1|}{\sigma}>\ \lambda_0$).

In definitiva se la risposta x della nostra analisi per l'influenza supera il valore $m_1+\sigma\lambda_{2\varepsilon}$ dobbiamo concludere di essere infetti, e quindi seguiremo la profilassi per l'influenza, sapendo che con una probabilità pari ad ε siamo invece non affetti dal virus e prenderemmo delle medicine inutilmente.

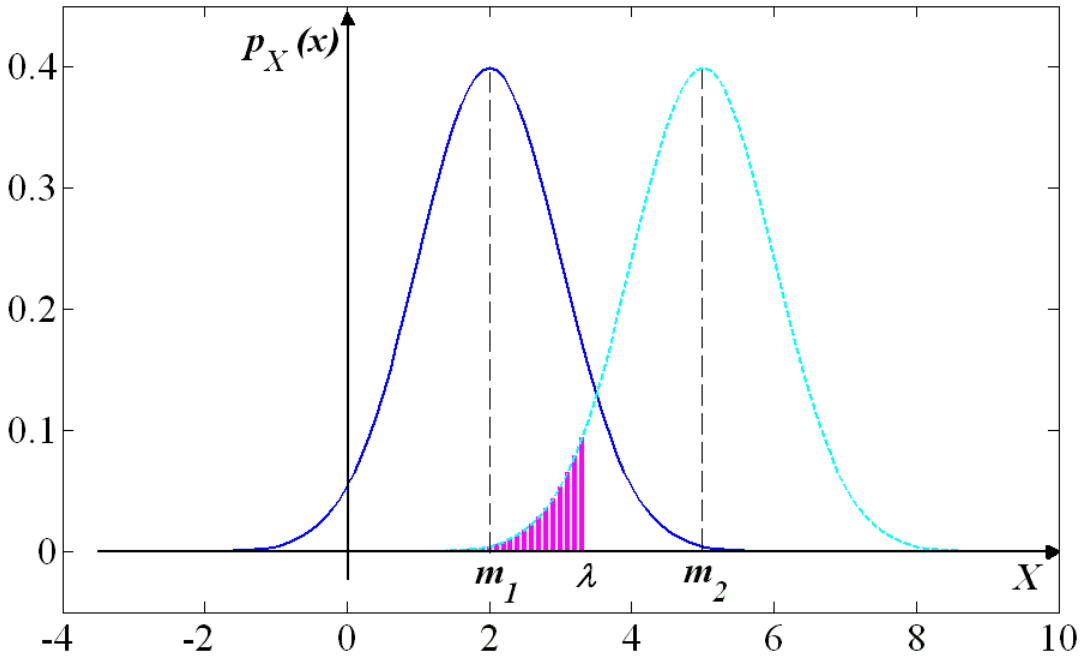
Facciamo ora un passo ulteriore: che fine ha fatto l'ipotesi alternativa? Questa in effetti entra in gioco quando $x\le\lambda$ per cui accettiamo H_0. Come si vede dalla figura seguente, la distribuzione che corrisponde all'ipotesi alternativa assegna una probabilità finita all'evento $x\le\lambda$ data da

$$P(x\le\lambda\,|H_1)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\lambda} e^{\frac{-(x-m_2)^2}{2\sigma^2}}\,dx$$

Per cui, se l'ipotesi H_0 è falsa per cui la media della distribuzione è effettivamente pari a m_2, c'è comunque una probabilità non nulla di osservare valori sperimentali x che siano minori di λ e per i quali effettivamente accetteremmo H_0. Tale errore prende il nome di *errore di tipo 2*: si accetta H_0 quando è falsa (attenzione non è il complementare dell'errore di tipo 1). La quantità

$$1-P(x\leq\lambda\,|\,H_1)=\frac{1}{\sqrt{2\pi}\sigma}\int_{\lambda}^{\infty}e^{\frac{-(x-m_2)^2}{2\sigma^2}}dx$$

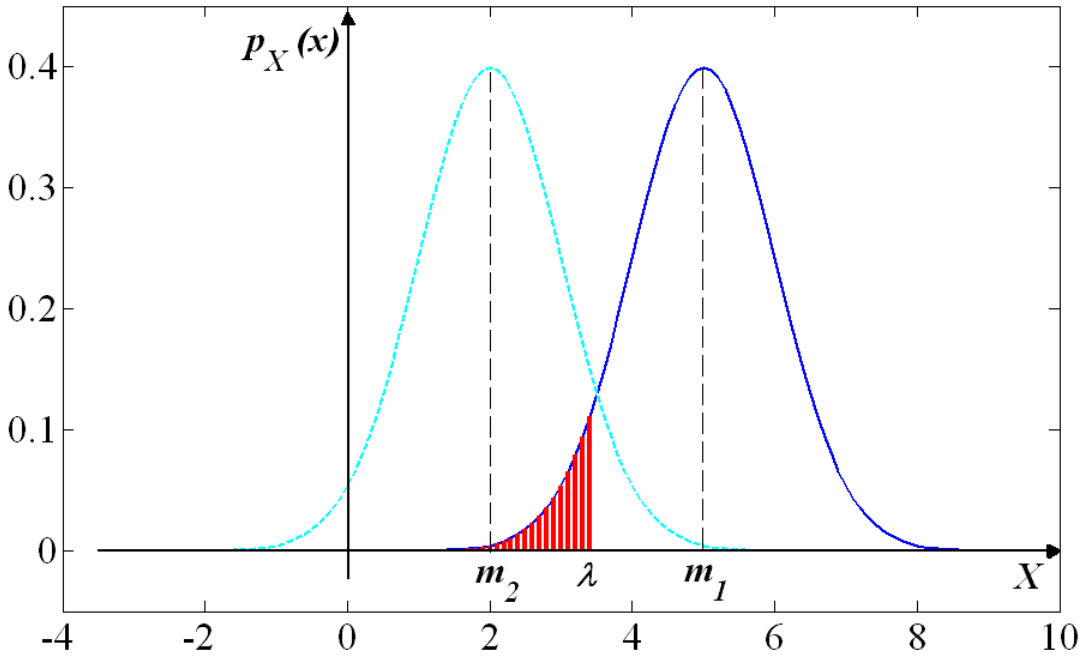
prende il nome di *potenza del test*, e corrisponde alla probabilità del set critico sotto l'ipotesi alternativa $H_1:\ P(x>\lambda\,|\,H1)$.



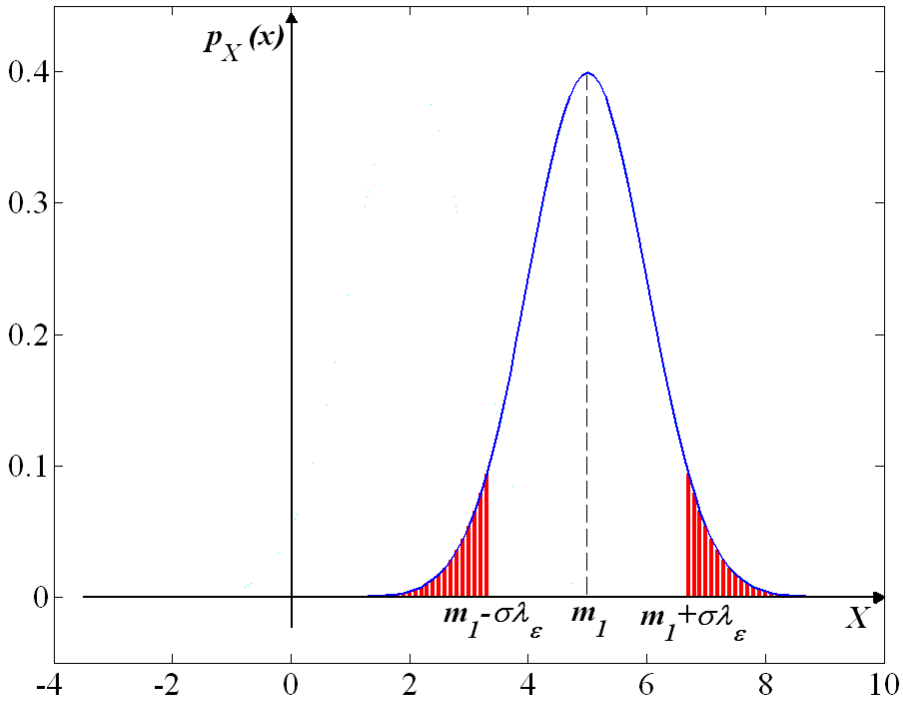
Si può dimostrare (lemma di Neyman-Pearson) che, assegnato il livello ε di significatività del test, il set critico scelto $x>m_1+\sigma\lambda_{2\varepsilon}$ è quello a cui corrisponde la minima probabilità di commettere l'errore di tipo 2, e quindi la massima potenza; per cui potremmo anche dire che il set critico scelto fornisce il test più potente di livello ε .

Nell'esempio trattato abbiamo considerato il caso che m_2 fosse maggiore di m_1; nella situazione complementare in cui risulti m_2 minore di m_1, con ragionamenti analoghi a quelli fatti si otterrebbe il seguente test più potente di livello ε

$$x<m_1-\lambda=m_1-\sigma\lambda_{2\varepsilon}$$



Se infine l'ipotesi alternativa fosse stata $H_1 : m \neq m_1$, avremmo dovuto contemplare contemporaneamente le due situazioni precedenti. Con facili ragionamenti si arriverebbe al seguente set critico bilaterale

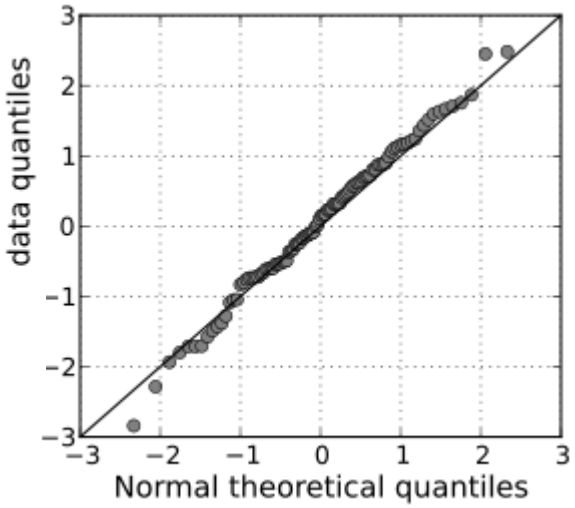


$$\{x < m_1 - \sigma\lambda_\varepsilon\} \cup \{x > m_1 + \sigma\lambda_\varepsilon\}$$

ottenendo ancora il set critico più potente di livello ε.

I test di ipotesi del tipo di quello analizzato vengono detti *test di ipotesi semplice*: in questi la forma della distribuzione è nota, il vettore θ dei suoi parametri è incognito ma può assumere valore solo in un punto θ_0 per l'ipotesi nulla e un punto θ_1 per l'ipotesi alternativa. Nel caso in cui, per almeno una delle due ipotesi, il vettore dei parametri della distribuzione potesse assumere valori in un insieme di punti si parlerebbe di *test di ipotesi composta*. Altri tipi di test di ipotesi semplice e composta saranno affrontati nel seguito del corso.

In molte situazioni, l'ipotesi da verificare riguarda proprio la forma della distribuzione nel suo insieme. Quindi, da un insieme di dati, si vuole stabilire se la variabile aleatoria cui essi si riferiscono abbia o meno una distribuzione $p_X(x)$ assegnata. Nel caso che la forma ipotizzata della distribuzione sia gaussiana $N(m,\sigma^2)$, esiste un semplice metodo grafico per testare questa ipotesi, il Q-Q Plot (Quantile-Quantile Plot). Il grafico riporta in ascisse i quantili della $N(m,\sigma^2)$ (usualmente ad intervalli di 0.1) ed in ordinate gli stessi quantili ottenuti dalla distribuzione campionaria dei dati.



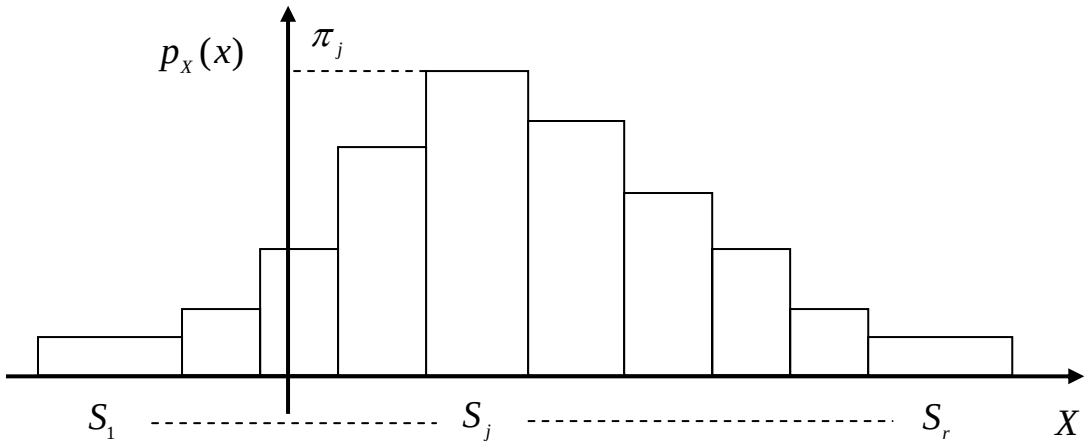
Quanto più i quantili sono uguali tanto più i dati confermano l'ipotesi di gaussianità. In questo caso il Q-Q plot si presenta come un insieme di punti abbastanza allineati lungo la bisettrice (vedi figura). Nel caso in cui i punti non fossero adeguatamente allineati lungo la bisettrice si dovrebbe rifiutare l'ipotesi che la loro distribuzione sia $N(m,\sigma^2)$. E' un metodo diciamo abbastanza euristico in quanto si basa su un giudizio soggettivo, ma è di rapida applicazione e di immediata interpretazione. Il Q-Q plot è presente in un qualsiasi applicativo, come ad es. Matlab.

Test χ^2 di Pearson.

In questo test l'ipotesi nulla H_0 consiste nello specificare la distribuzione $p_X(x)$ di una variabile aleatoria X. Dividiamo l'insieme ammissibile dei valori della distribuzione allo studio in r sottointervalli $S_1,\cdots,S_r$ disgiunti, non necessariamente tutti uguali. Questi sono eventi elementari di cui si possono definire le probabilità secondo l'ipotesi H_0

$$p_j = \int_{S_j} p_X(x)dx, \quad j = 1,\cdots,r$$

risultando peraltro che $\sum_{j=1}^{r} p_j = 1$. In congruenza con la decomposizione dell'insieme dei possibili risultati effettuata, dividiamo ora il campione osservato in gruppi di dati ottenuti contando per ogni sottointervallo S_j il numero n_j di risultati fra gli N possibili che appartengono ad esso. Possiamo a questo punto costruire un istogramma della distribuzione campionaria



riportando in corrispondenza degli S_j i valori di frequenza realtiva $\pi_j = n_j / N$, che ovviamente verificano $\sum_{j=1}^{r} \pi_j = 1$. Da un punto di vista intuitivo se l'ipotesi H_0 è vera, per N abbastanza grande le frequenze relative dovrebbero non discostarsi molto dai valori di probabilità p_j, calcolati con la distribuzione ipotizzata.

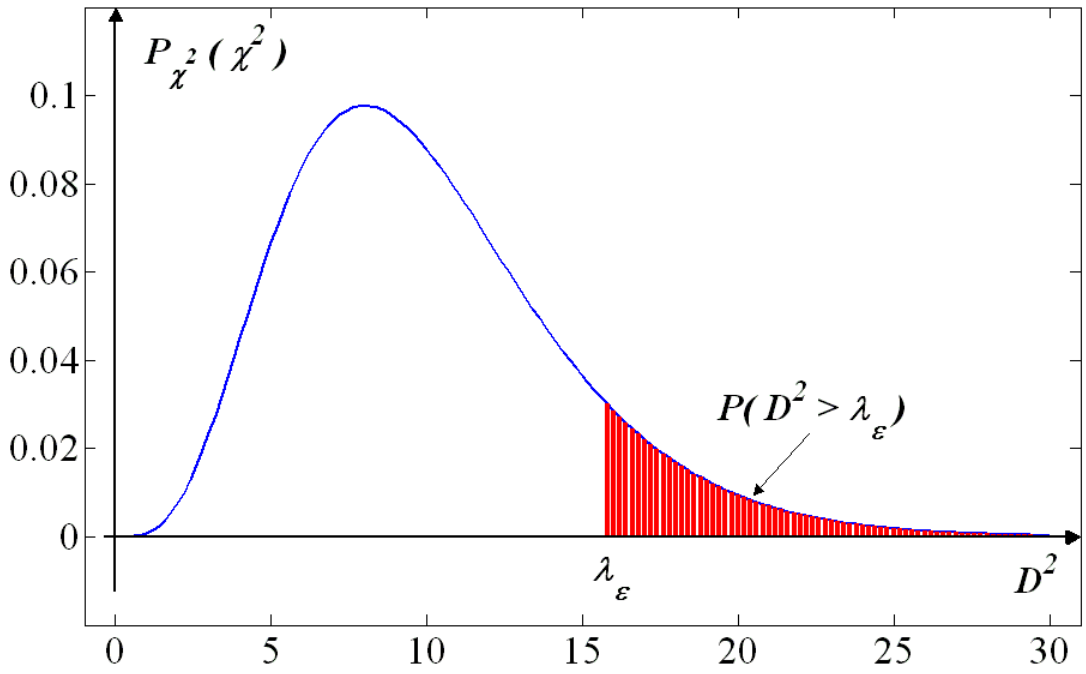
Una misura della deviazione della distribuzione campionaria (istogramma) dalla distribuzione ipotizzata può essere la seguente

$$D^2 = \sum_{j=1}^{r} \frac{N(\pi_j - p_j)^2}{p_j} = \sum_{j=1}^{r} \frac{(n_j - Np_j)^2}{Np_j}$$

dove vengono considerati gli scarti al quadrato tra i valori ipotizzati di probabilità e quelli determinati dai dati sperimentali.

Il risultato notevole ottenuto da Pearson consiste nell'aver dimostrato che al crescere di N, la distribuzione di D^2 tende ad una distribuzione limite che è indipendente da quella ipotizzata . In effetti si dimostra che tale distribuzione limite corrisponde ad una χ^2 con $r-1$ gradi di libertà. Si noti che questo non vuol dire che la v.a. diventi una χ^2 (in particolare non è vero che risulti essere la somma dei quadrati di gaussiane standard indipendenti) ma solo che la probabilità degli eventi legati alla D^2 può essere valutata mediante la distribuzione limite, con approssimazione tanto migliore quando maggiore è N.

A questo punto si può determinare quel valore percentile λ_ε a cui corrisponde una probabilità $\varepsilon\%$ di ottenere una deviazione $D^2 \geq \lambda_\varepsilon$



(area della coda della distribuzione a destra di λ_ε). Quindi se nel nostro esperimento, a fronte degli N dati prelevati, con N sufficientemente grande, dovessimo ottenere un valore D^2 minore del λ_ε prescelto, dovremmo ritenere la distribuzione campionaria consistente con l'ipotesi , con un livello di significatività pari a $\varepsilon\%$; un valore superiore a sarebbe considerato un valore di deviazione troppo grande, tale da ritenere che l'evidenza sperimentale non supporti l'ipotesi .

Nell'applicazione pratica del test di Pearson, bisogna saper scegliere opportunamente il numero r dei sottointervalli in cui è decomposto l'insieme ammissibile, ed il numero N che stabilisce la dimensione del campione di dati da analizzare. Il test di Pearson si basa sul confronto tra l'istogramma della distribuzione campionaria del campione di N dati ed il profilo della distribuzione ipotizzata. Dipendentemente da quest'ultimo, un istogramma con un numero troppo basso di sottointervalli darebbe luogo comunque ad una grossa deviazione indipendentemente dalla numerosità del campione; la pratica suggerisce di scegliere un numero di sottointervalli non inferiore a 5. Per ogni sottointervallo S_j poi deve risultare $Np_j \geq 10$ che permette di scegliere N. E' ovvio che dovendo ottenere una informazione molto sofisticata quale l'andamento della distribuzione, ci si debba aspettare valori di N molto grandi.

Valori grandi di N sono anche richiesti dal fatto che il test di Pearson non è un test esatto, l'approssimazione alla distribuzione limite è tanto migliore quanto più N è grande. Il vantaggio di questo test è che la statistica del test non dipende dall'ipotesi da testare, e che inoltre i parametri dell'ipotesi da testare possono anche essere stimati dai dati; in questo caso la statistica limite sarà una χ^2_{r-c} dove c è pari al numero dei parametri da stimare aumentato di uno.

Test di Kolomogorov-Smirnov (K-S test)

Questo test esegue il confronto tra la distribuzione cumulativa ipotizzata e quella ottenuta dai dati. Come è noto, i valori della prima si ottengono nel modo seguente

$$F(x) = \int_{-\infty}^{x} p_X(t)\,dt$$

per cui, se $x_1, x_2, \ldots, x_N$ sono i dati raccolti, calcoleremo N valori della distribuzione cumulativa di probabilità ipotizzata

$$F(x_i) = \int_{-\infty}^{x_i} p_X(t)\,dt, \quad i = 1, \ldots, N$$

Per calcolare i valori corrispondenti della distribuzione cumulativa campionaria basta ordinare i dati in valore crescente $x_1 < x_2 < \ldots < x_N$, si ottiene subito

$$F_c(x_i) = \frac{i-1}{N}$$

che corrisponde al numero dei dati che hanno valore minore di x_i diviso il numero totale dei dati.

A questo punto la statistica del test è ottenuta nel modo seguente

$$\begin{aligned} D^2 &= \max_{1 \le i \le N} \left(F(x_i) - F_c(x_i), F_c(x_{i+1}) - F(x_i) \right) \\ &= \max_{1 \le i \le N} \left(F(x_i) - \frac{i-1}{N}, \frac{i}{N} - F(x_i) \right) \end{aligned}$$

I percentili che corrispondono alla statistica del test sono forniti da opportune tabelle. Ogni tabella è costruita rispetto ad opportune variazioni di scala della statistica; quindi bisogna fare attenzione, quando si usa una di queste tabelle, di scalare la D^2 come previsto dalla tabella. C'è da dire che anche questo test si trova già implementato (tabelle incluse) in tutti gli applicativi di analisi dei dati in commercio. Il test quindi va eseguito nel solito modo, si fissa il livello di significatività ε , dalle opportune tabelle si ricava il percentile corrispondente λ_ε , l'ipotesi viene rifiutata se $D^2 > \lambda_\varepsilon$.

Come il test del χ^2 , anche il K-S test è indipendente dall'ipotesi da testare. Esso è però un test esatto in quanto la sua statistica non è una statistica limite, per cui normalmente è richiesto un numero N di dati moderato (qualche decina).

Vediamo i difetti. Il test si può eseguire solo per distribuzioni continue e la distribuzione deve essere completamente specificata, cioè i suoi parametri debbono essere noti e non stimati dai dati. Quest'ultimo in effetti rappresenta un difetto sostanziale. Inoltre il test è maggiormente sensibile ai valori centrali della distribuzione e meno a quelli sulle code.

Test di Anderson -Darling (A-D test)

Questo test rimuove tutti i difetti del K-S test ed è una sua estensione. E' un test esatto che da più peso ai valori della distribuzione sulle code, i parametri della distribuzione ipotizzata possono essere stimati dai dati, va bene anche per le distribuzioni discrete. L'unico difetto è che il test dipende dall'ipotesi da testare, per cui avremo tabelle dei percentili differenti a seconda della distribuzione ipotizzata da testare (gaussiana, log-normale, esponenziale, Weibull, logistica, ...). Anche qui le tabelle possono riferirsi ad opportune variazioni di scale della statistica del test.

Con le stesse notazioni introdotte nel K-S test, la statistica del test di Anderson-Darling è data da

$$D^2 = -N - \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln\left(F(x_i)\right) + \ln\left(1 - F(x_{N+1-i})\right) \right]$$

dove i dati sono ovviamente ordinati per valori crescenti. Il test è affidabile anche con un numero di dati esiguo, intorno a 20.