Distribuzione normale o gaussiana

Una variabile random si dice distribuita normalmente, o secondo una curva gaussiana, se la sua funzione di densità di probabilità è del tipo:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad \text{con } -\infty \leq x \leq +\infty$$

μ rappresenta il valore centrale, che in questo caso corrisponde anche alla media della distribuzione, mentre σ^2 è la varianza.

In molti casi si preferisce introdurre per praticità la variabile random Z, espressa dalla relazione:

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

Se X è distribuita normalmente anche Z lo è e la sua funzione di densità di probabilità è data da:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{z^2}{2} \right\} \quad \text{con } -\infty \leq z \leq +\infty$$

nota come funzione di densità di probabilità normale standard.

f(z) è una curva simmetrica centrata sul valore 0 e con varianza 1 e spesso viene indicata con la notazione **N(0,1)**.

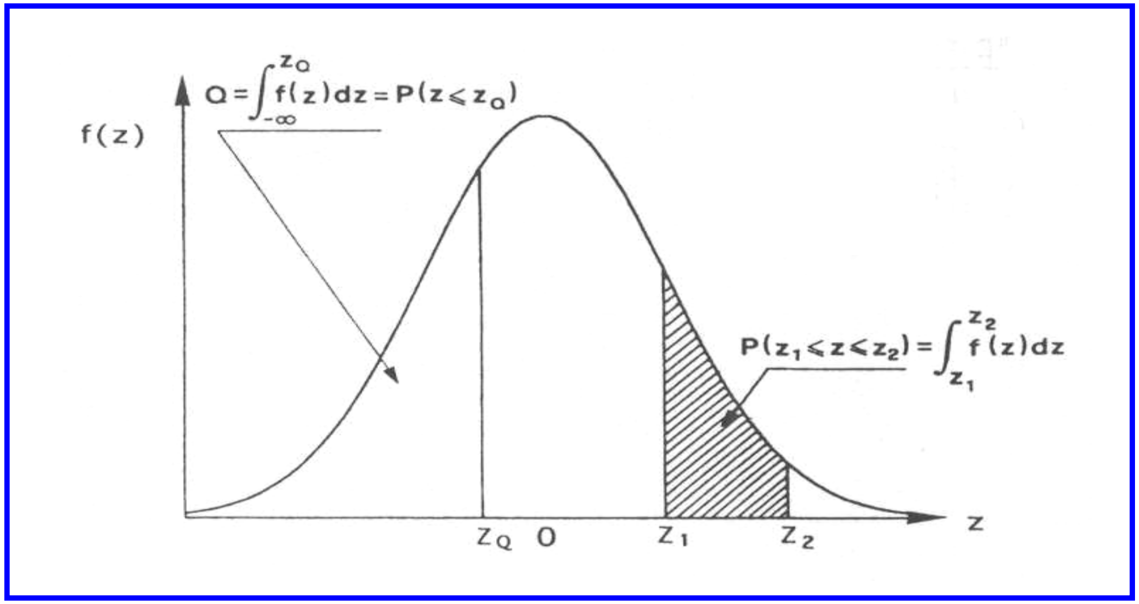
Come per ogni funzione densità di probabilità anche per f(z) vale la relazione:

$$\boxed{\int_{-\infty}^{+\infty} f(z)dz = 1}$$

La probabilità che la variabile random Z sia compresa fra due valori z_1 e z_2, $P(z_1 \leq z \leq z_2)$, corrisponde all'area sottesa dalla curva gaussiana fra i due valori ed è calcolabile dalla relazione:

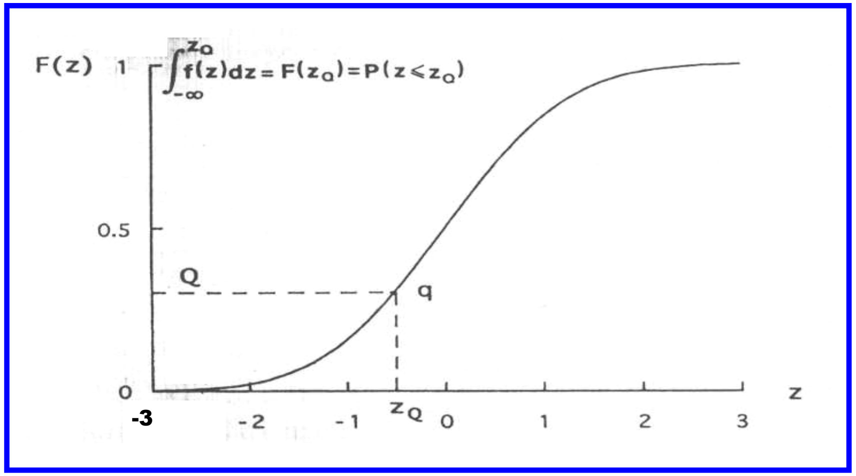
$$\boxed{\int_{z_1}^{z_2} f(z)dz}$$

L'integrale, detto Q, di f(z) fra -∞ e un certo z_Q corrisponde alla frequenza cumulativa relativa per tale z_Q e viene definito quantile.



Il calcolo di P può essere effettuato anche a partire dalla curva di distribuzione normale standard, F(z):

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz$$



infatti la probabilità che la variabile Z sia inferiore ad un generico valore z_Q è data direttamente da $F(z_Q)$,

mentre $P(z_1 \leq z \leq z_2) = F(z_2) - F(z_1)$

Gli intervalli di valori di z del tipo $-n \rightarrow n$, con n intero, delimitano porzioni caratteristiche dell'area sottesa alla curva gaussiana standard.
In particolare:

fra z = -1 e z = 1 è racchiuso il 68.3 % dell'area sottesa totale;
fra z = -2 e z = 2 il 95.4 %
fra z = -3 e z = 3 il 99.7 %

Ragioni dell'importanza della curva gaussiana

Ci sono almeno quattro ragioni che giustificano l'uso esteso della curva gaussiana per le variabili random:

- ✓ l'esperienza pratica mostra che la curva gaussiana è la più appropriata per descrivere la variazione della misura di molte grandezze chimico-fisiche;
- ✓ la distribuzione gaussiana è stata studiata a fondo ed i suoi valori sono facilmente accessibili in forma di tavole o direttamente inseriti all'interno di software di tipo statistico;
- ✓ molte tecniche statistiche basate sulla distribuzione normale sono statisticamente *robuste*, ossia forniscono un esito approssimativamente corretto anche in presenza di scostamenti ragionevolmente grandi dalla normalità;
- ✓ in virtù del Teorema del Limite Centrale, molte distribuzioni campionarie appaiono gaussiane a prescindere dalla distribuzione effettiva della popolazione a cui si riferiscono, purché le dimensioni del campione (ossia il numero dei dati considerati) siano sufficientemente elevate.

Il teorema del limite centrale, dimostrato nel 1922 dal matematico e statistico finlandese Jarl Waldemar Lindeberg, afferma che:

"date le variabili random X_1, X_2, …X_n, ciascuna delle quali caratterizzata da una media μ_i e varianza σ_i^2, la variabile data dalla loro somma tende ad una distribuzione normale di media $\Sigma_i \mu_i$ e varianza $\Sigma_i \sigma_i^2$ al tendere di n ad infinito".

Si noti che le variabili X_i del teorema potrebbero essere rappresentate anche da valori derivanti da una stessa popolazione, dunque essere distribuite allo stesso modo.

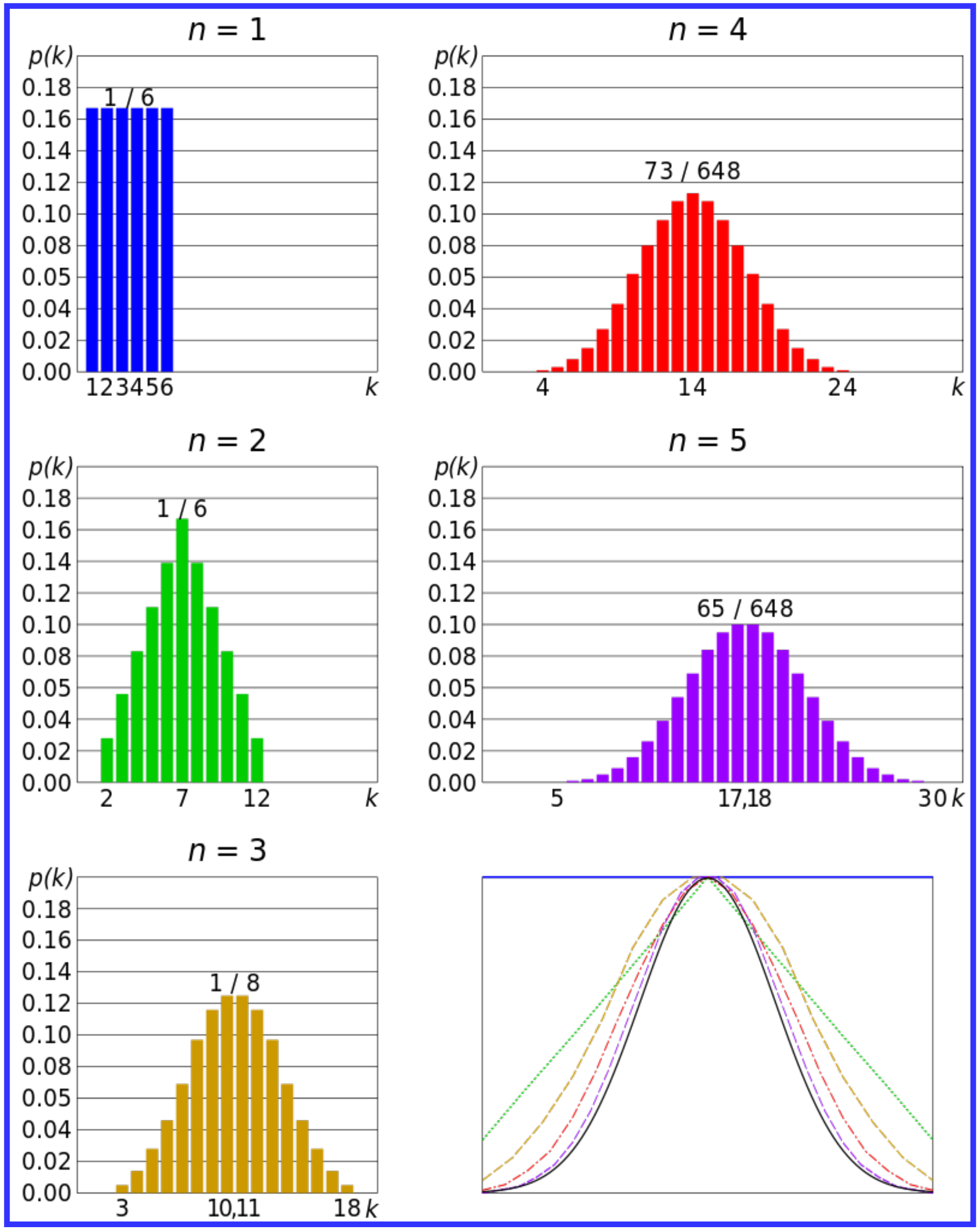
Le fluttuazioni nelle misure analitiche, dovute a fonti diverse (strumentali, ambientali, umane, ecc.), possono essere considerate derivanti dalla combinazione lineare di componenti diverse aventi distribuzioni indipendenti.

Per il Teorema del Limite Centrale tale combinazione, e quindi l'errore random ad essa legato, può avere una distribuzione normale.

Nella figura mostrata a lato si evidenzia l'effetto della combinazione fra un certo numero di valori ottenuti lanciando un singolo dato una o più volte.

In questo caso la variabile random (discreta) è rappresentata dal punteggio ottenuto lanciando il dado, che per ogni lancio può andare da 1 a 6, con la stessa probabilità (1/6), se il dado non è truccato.

Come si può vedere, via via che si aumenta il numero di punteggi sommati, l'istogramma che rappresenta la densità di probabilità si avvicina sempre più ad una funzione gaussiana.



Distribuzione chi-quadro (χ^2)

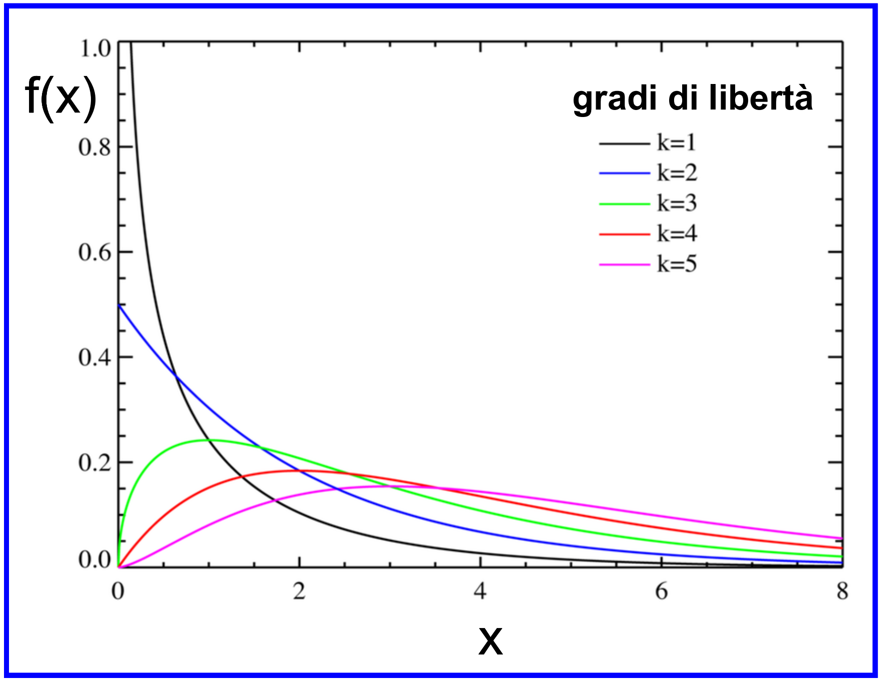
Si dice che una variabile random X è distribuita secondo una distribuzione χ^2 con ν gradi di libertà (una funzione descritta dall'ottico Ernst Abbe nel 1863) se la sua curva di densità di probabilità è data da:

$$f(x) = \frac{x^{\nu/2-1}\, e^{-x/2}}{2^{\nu/2}\,\Gamma(\nu/2)}$$

con $\nu > 0$ e $0 \leq x \leq +\infty$

Γ è la funzione gamma:

$$\Gamma(m) = \int_0^{\infty} e^{-x} x^{m-1} dx$$



- ✓ Media di una distribuzione χ^2 : ν
- ✓ Varianza di una distribuzione χ^2: 2ν

Proprietà principali della distribuzione chi-quadro

✓ date le variabili random X_1, X_2, …,X_n, ciascuna distribuita secondo una distribuzione normale standard N(0,1),
la variabile data dalla somma dei loro quadrati è distribuita secondo una distribuzione χ^2 con n-1 gradi di libertà (χ^2_{n-1})

✓ date le variabili random X_1, X_2, …,X_n, ciascuna distribuita secondo una distribuzione normale $N(\mu,\sigma^2)$,
la variabile $\Sigma_i (X_i - \overline{X})^2/\sigma^2 = (n-1)s^2/\sigma^2$ è distribuita anch'essa come una χ^2 con n-1 gradi di libertà

✓ se le variabili X_1 e X_2 sono distribuite indipendentemente come $\chi^2_{\nu 1}$ e $\chi^2_{\nu 2}$, la variabile data dalla loro somma è distribuita come $\chi^2_{\nu 1+\nu 2}$.

Distribuzione t di Student

Una distribuzione t di Student (William S. Gosset, 1908) a ν gradi di libertà è descritta matematicamente dalla seguente funzione di densità di probabilità:

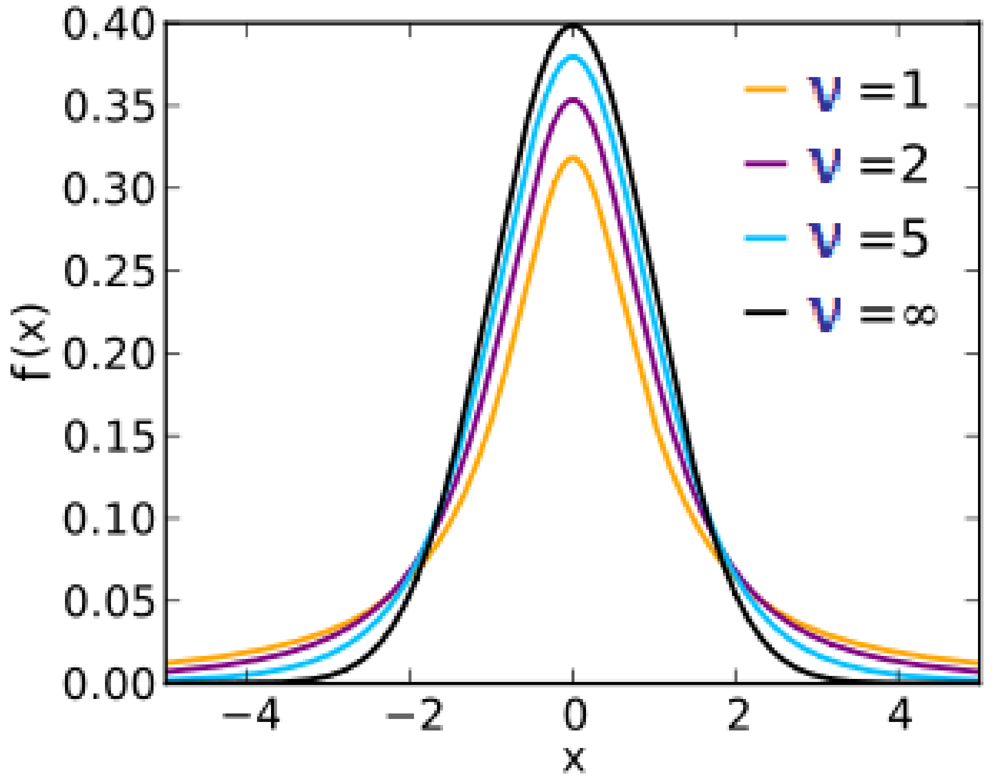
$$f(x) = \frac{(\pi\upsilon)^{-1/2}\,\Gamma\{(\upsilon+1)/2\}}{\Gamma(\upsilon/2)}\left(1+\frac{x^2}{\upsilon}\right)^{-(\upsilon+1)/2}$$

con:

$$\Gamma(m) = \int_0^{\infty} e^{-x} x^{m-1}\,dx$$

La distribuzione t di Student

- ✓ è simmetrica intorno al valore x = 0
- ✓ tende ad una distribuzione gaussiana per ν che tende ad infinito.



Proprietà fondamentale della distribuzione t di Student:

se A e B sono due variabili random indipendenti, distribuite rispettivamente come N(0,1) e χ^2_ν, la variabile random $Z = A/(B/\nu)^{1/2}$ è distribuita secondo una funzione t di Student con ν gradi di libertà.

Poiché per una variabile $X \sim N(\mu,\sigma^2)$, la variabile random:

$$\boxed{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}$$

è distribuita secondo una curva normale standard, N(0,1),

mentre la variabile random:

$$\boxed{\frac{(n-1)s^2}{\sigma^2}}$$

è distribuita secondo una distribuzione χ^2_{n-1},

la variabile random:

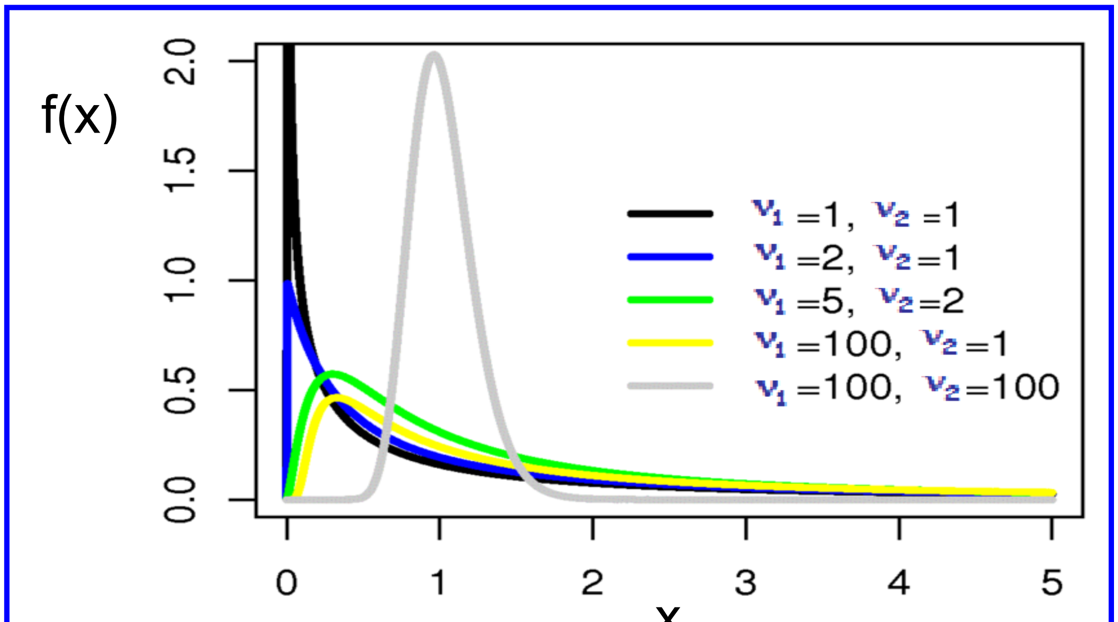
$$\frac{\boxed{(\overline{X}-\mu)/(\sigma/\sqrt{n})}}{\left\{\boxed{\cancel{(n-1)}s^2/\sigma^2}\,\cancel{(n-1)}\right\}^{1/2}} = \boxed{\frac{\overline{X}-\mu}{s/\sqrt{n}}}$$

è distribuita secondo t_{n-1}.

Distribuzione F

Si dice che una variabile random è distribuita secondo una distribuzione F (nome attribuitole in onore dello statistico inglese Sir Ronald Fisher, che la introdusse nell'analisi della varianza nel 1924) con gradi di libertà ν_1 ed ν_2 se la sua curva di densità di probabilità è del tipo:

$$f(x) = \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right) x^{\frac{\nu_1-2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(\nu_1 x+\nu_2\right)^{\frac{\nu_1+\nu_2}{2}}}$$



Date le variabili random A e B, distribuite rispettivamente secondo $\chi^2_{\nu 1}$ e $\chi^2_{\nu 2}$, la variabile **$(A/\nu_1)/(B/\nu_2)$** è distribuita secondo una distribuzione **$F\nu_1,\nu_2$**.

Se dunque X e Y sono variabili random distribuite rispettivamente secondo curve gaussiane $N(\mu_1,\sigma_1^2)$ e $N(\mu_2,\sigma_2^2)$

e, quindi, le variabili $(n_1-1)s_1^2/\sigma_1^2$ e $(n_2-1)s_2^2/\sigma_2^2$ sono distribuite rispettivamente secondo χ^2_{n1-1} e χ^2_{n2-1},

la variabile $(s_1^2/\ \sigma_1^2)/(s_2^2/\ \sigma_2^2)$ è distribuita secondo una funzione F con gradi di libertà n_1-1, n_2-1.