CDL in MEDICINA & CHIRURGIA

Statistica Medica

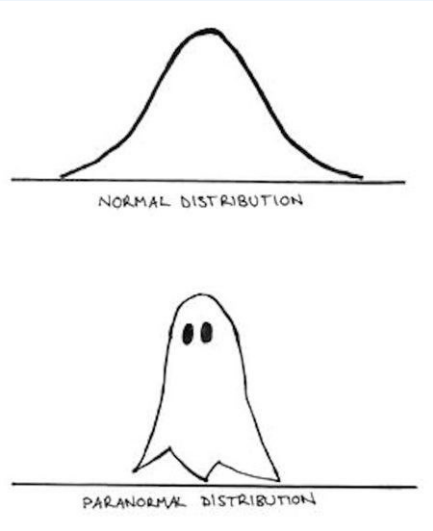
gbarbati@units.it

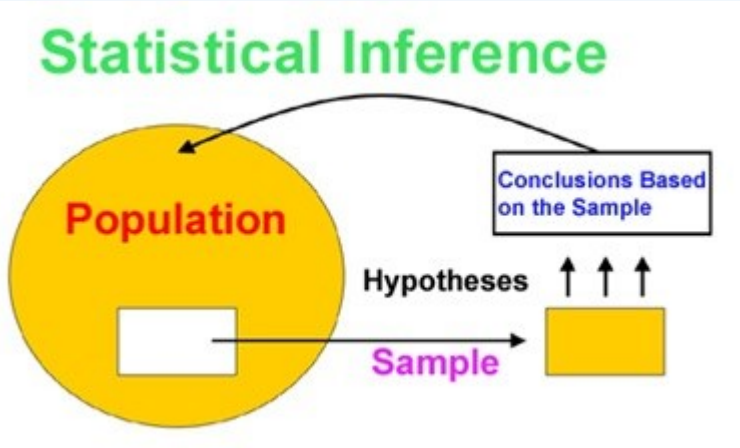
A.A. 2024-25

UNITÀ DI BIOSTATISTICA

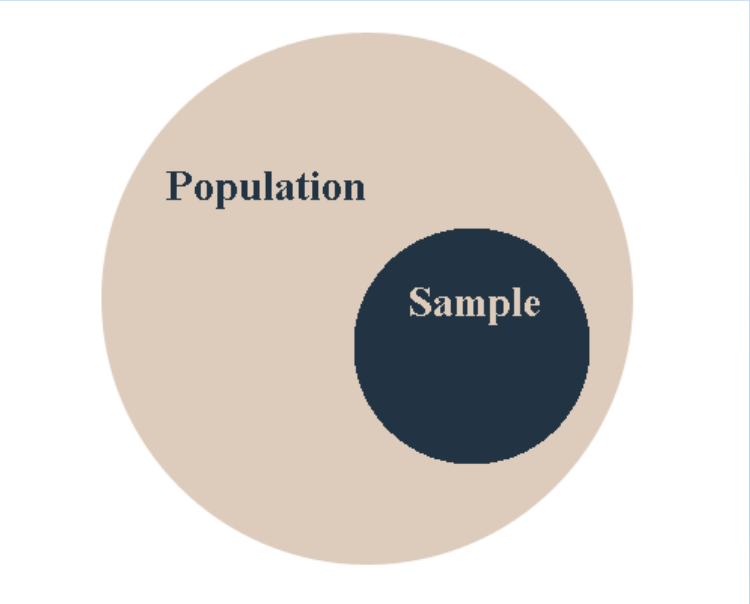
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Sommario:





- Introduzione all'inferenza
- Definizioni di probabilità
- Cenni di calcolo delle probabilità
- Variabili Aleatorie & Distribuzioni di Probabilità



Un *trial* clinico ha mostrato che 50 pazienti che ricevono il farmaco A per una malattia guariscono più velocemente (in media) di 50 pazienti che ricevono il farmaco B.

(1) E' giusto affermare, **in generale**, che A è migliore di B per curare questa malattia?

(2) Il medico dovrà usare in futuro A piuttosto che B per trattare al meglio i suoi pazienti?

(1) Domanda di tipo "***inferenziale***": quali conclusioni possono essere tratte dal **campione** rispetto alla **popolazione** da cui è stato estratto?

(2) Domanda è di tipo "***decisionale***": qual è la scelta più razionale per il futuro trattamento, tenendo conto dell'informazione offerta dal trial?

Le risposte ad entrambe le domande, e a tutte le domande riguardanti studi ***campionari***, sono caratterizzate da un certo livello di ***incertezza***.

L' indicazione dal campione è che il farmaco A sia migliore di B, ma…:

- possiamo essere **certi** che i pazienti che hanno ricevuto B non fossero più gravemente malati di quelli che hanno ricevuto A?
- e che questa ***variabilità*** tra i due gruppi non fosse il motivo per le loro differenti risposte?

Le risposte devono quindi essere formulate in termini di '***incertezza***':

incertezza bassa -> conclusione campionaria affidabile, decisione «sicura»

incertezza alta -> l'esperimento sarà considerato non conclusivo e non sarà di aiuto nella decisione.

Per avere una ***'misura dell'incertezza***' di un risultato campionario, lo strumento matematico appropriato è la ***teoria della probabilità***.

Dalla Statistica Descrittiva alla Statistica Inferenziale

Popolazione

Fenomeni X, Y
Descritti da: μ_X, μ_Y
Med_X, Med_Y
$\rho, \Sigma, \ldots$

Parametri (non noti)

Estrazione [casuale…] di un campione di dimensione *n*

Campione

Statistiche $\bar{x}, \bar{y}$
med_x, med_y
$r, S, \ldots$

sui dati campionari (noti)

Inferenza
Le "statistiche" campionarie vengono usate per trarre delle conclusioni sui **parametri** della popolazione, basandosi sulle osservazioni del campione **casuale**

Quindi:

- stime/conclusioni soggette a **incertezza** (basate su un **campione casuale** della popolazione…)
- **teoria della probabilità** è lo strumento matematico per effettuare l'inferenza

Definizione di probabilità

La probabilità rappresenta in termini matematici i fenomeni caratterizzati da comportamenti **variabili**

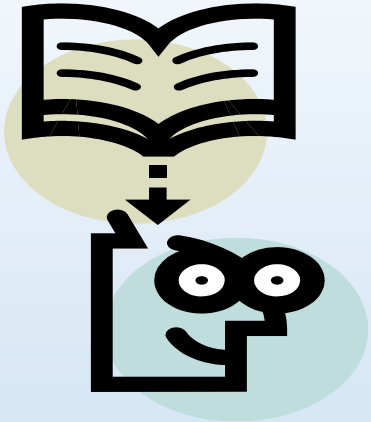
Se una moneta viene lanciata tante volte e si guarda l'esito di ogni lancio:

TTCTCCTCTTTCTCCTCCCCTTC....
T=testa; C=croce

Tale sequenza è definita una ***sequenza casuale*** o una ***sequenza random***; ogni posto nella sequenza rappresenta una ***prova*** o ***estrazione*** (***trial***) ed ogni risultato del lancio viene definito come un ***evento***.

Una sequenza random è caratterizzata dal fatto che non è possibile predire («*indovinare*») l'evento successivo sulla base dei precedenti.

La probabilità di avere 'croce' in un certo passo è la stessa che ad ogni altro, e non è influenzata da ciò che è successo prima



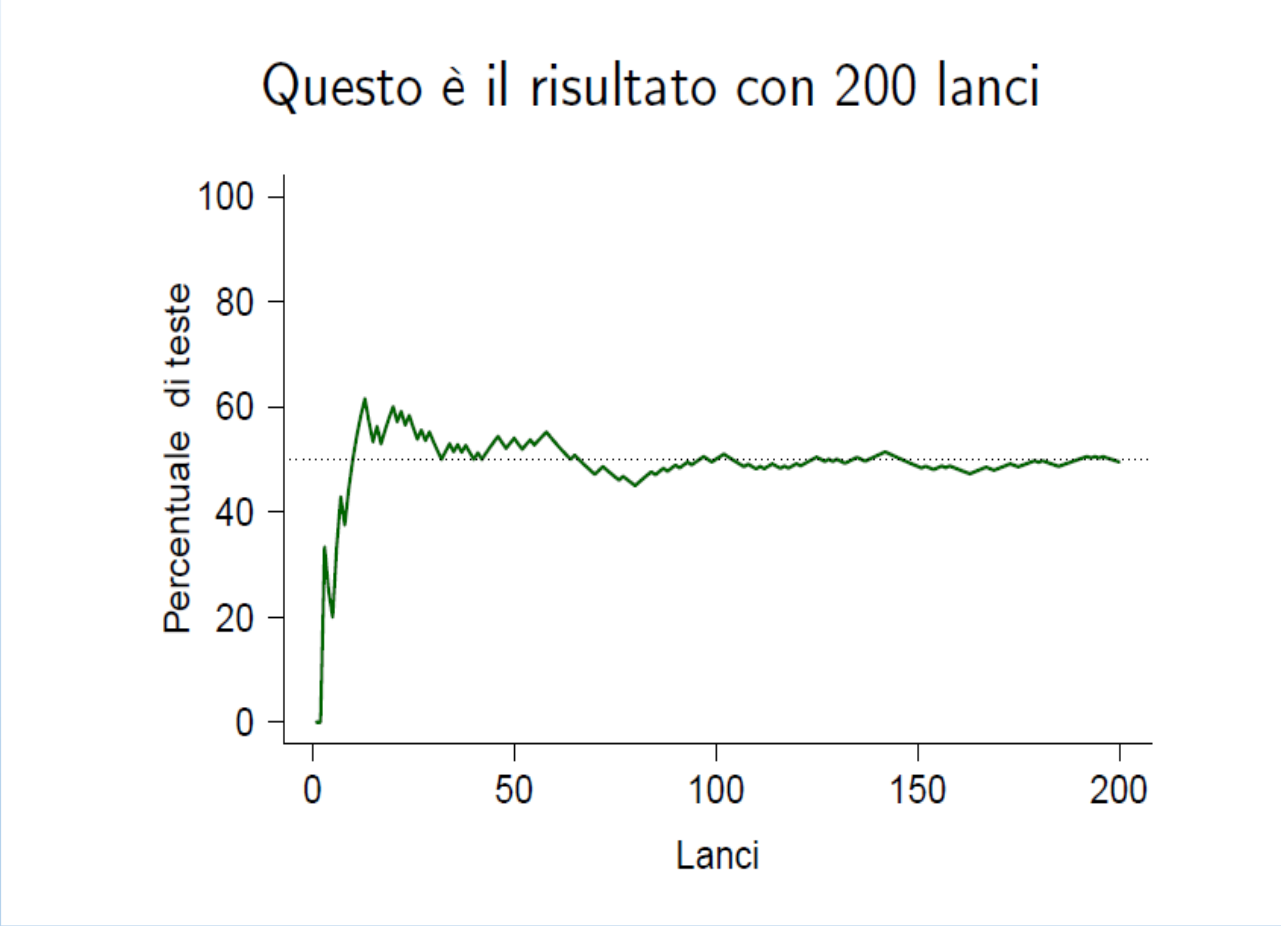
La definizione “classica” di probabilità è:

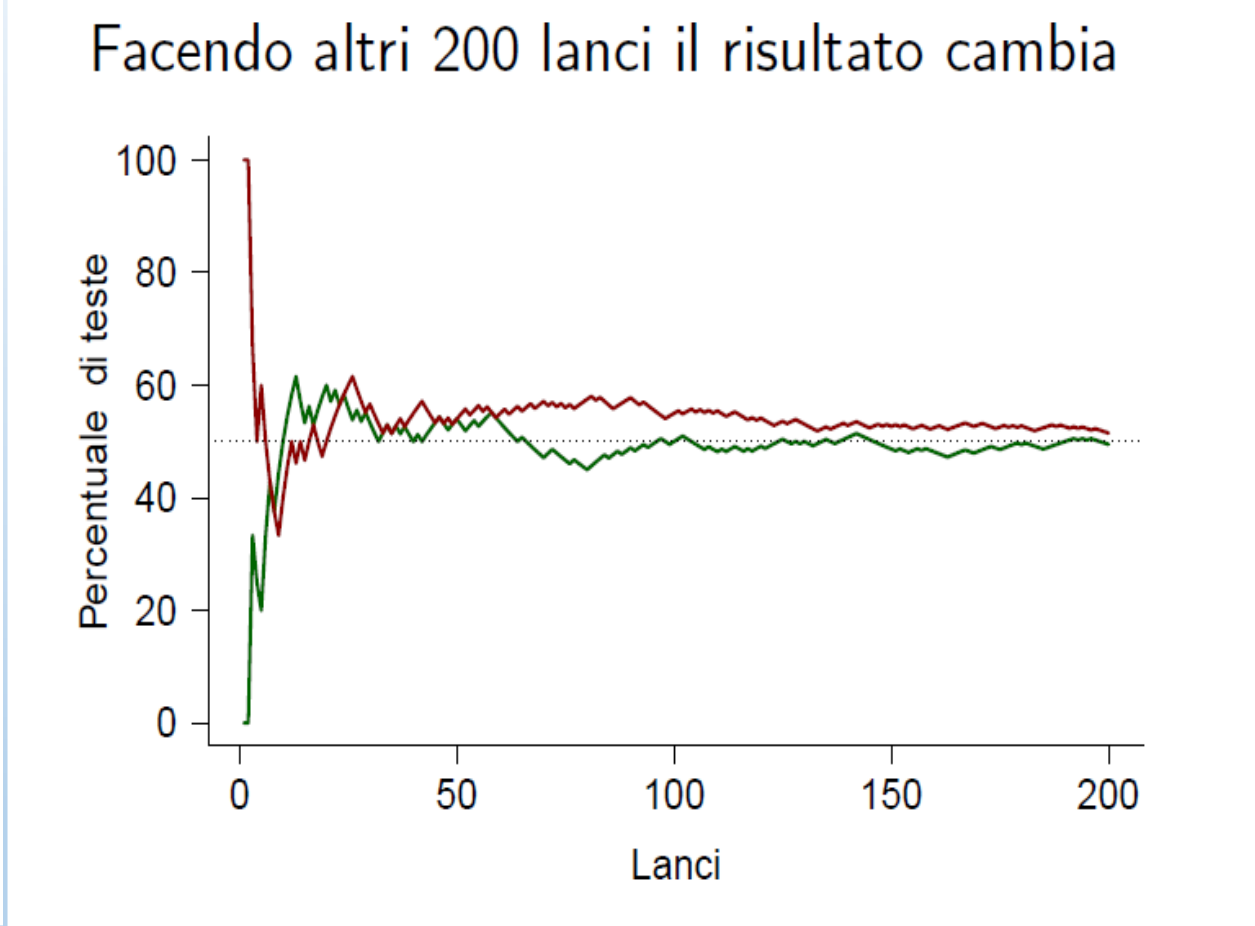
La probabilità di un evento è data dal rapporto tra il numero dei casi *favorevoli* all’evento ed il numero dei casi *possibili*, purchè questi ultimi siano tutti *ugualmente probabili*.

Per la moneta i casi possibili sono 2 (testa o croce) e quindi la probabilità che esca testa (o che esca croce) è pari ad 1/2.

Limite “*concettuale*” di questa definizione:
per definire la probabilità occorre sapere preliminarmente che cosa significa che due casi sono ugualmente probabili, cioè sapere già che cosa è la probabilità!

All’aumentare dei lanci, la proporzione (*frequenza*) di un evento diventa sempre meno variabile e sempre più vicina ad un valore limite: tale proporzione *'di lungo periodo*' viene definita *probabilità* dell'evento:



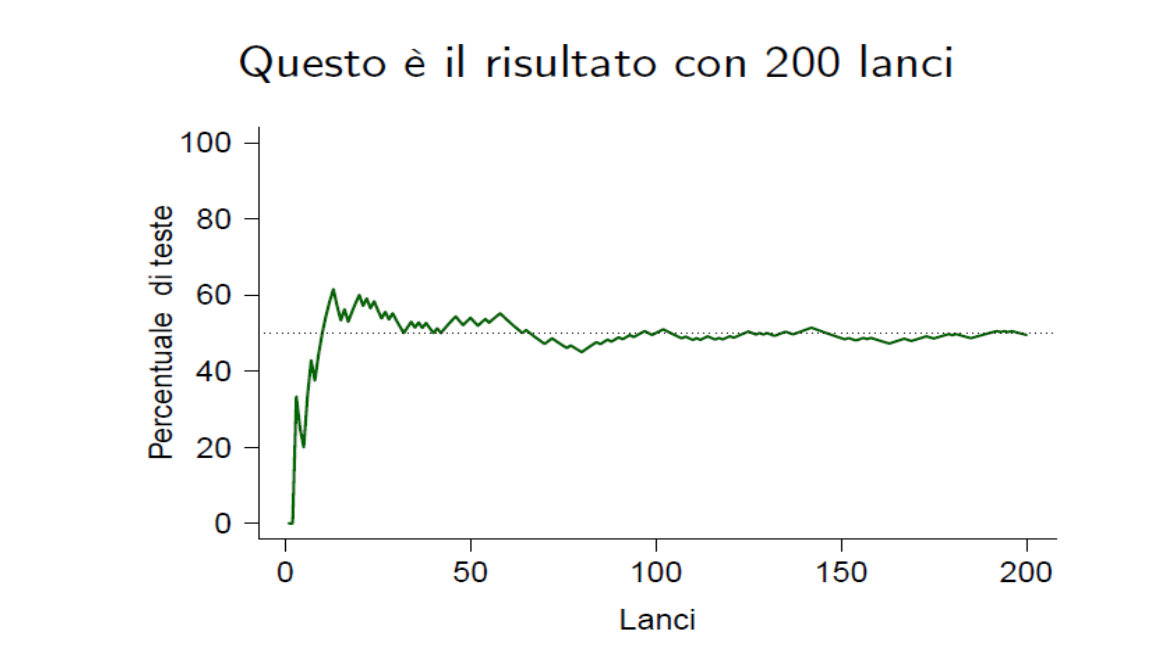


Man mano che si va avanti **il risultato si stabilizza intorno al 50%,** che è la probabilità che, intuitivamente, attribuiremmo all’evento ‘esce testa’.

In una successione di prove fatte nelle stesse condizioni, la **frequenza** di un evento si avvicina alla probabilità dell'evento stesso, e l'approssimazione tende a migliorare con l'aumentare del numero delle prove*.

E quindi, secondo la definizione "***frequentista***" di probabilità:

La probabilità di un evento è il limite della frequenza (relativa) dei successi (cioè delle prove in cui l'evento si verifica) quando il numero delle prove tende all'infinito.



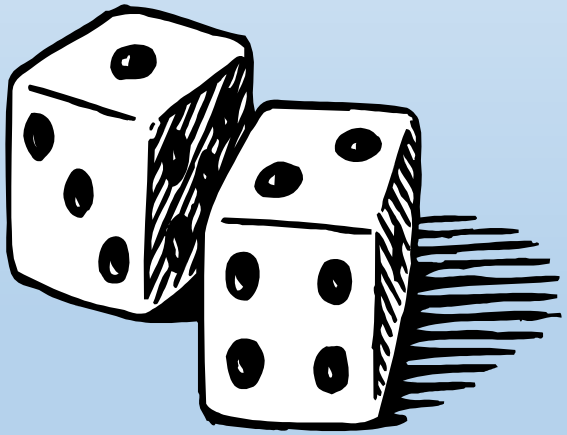
La probabilità dell'evento «esce Testa» corrisponde alla proporzione (frequenza relativa) della faccia testa: all'aumentare del numero dei lanci tale frequenza tende a ½ .

*Legge empirica del caso

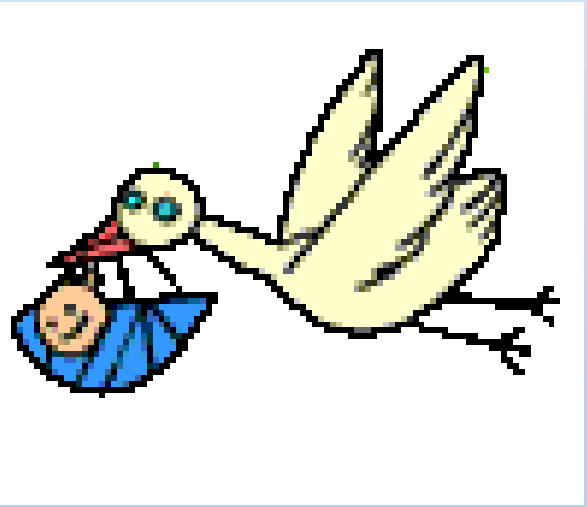
Anche questa definizione di probabilità è però alquanto criticabile: in generale, non potremo mai osservare una sequenza ***infinita*** di prove !!

..e se osserviamo solo una 'parte' di sequenza non possiamo affermare *con precisione* la probabilità di un certo evento, perchè la probabilità, come si è detto, è una proprietà ***a lungo termine***…

… il concetto di sequenza '***infinita***' deve essere interpretato come una **cornice teorica ideale**
(noi osserveremo sempre la frequenza degli eventi in **campioni di dimensione finita**).



Se si lancia un dado la frequenza relativa di ogni singola faccia del dado è circa pari ad 1/6 in una *lunga successione di lanci;*
in una coorte di neonati la frequenza del sesso maschile (o femminile) oscilla intorno a ½…



La probabilità si misura con un numero tra 0 e 1:

- se un evento *non capita mai* - in nessuno dei *trials* della sequenza random- la sua probabilità è 0
- se un evento *capita sempre* - in tutti i *trials* della sequenza random- la sua probabilità è 1

Complichiamo un po' la faccenda…

Qual è la probabilità che il fumo contribuisca al cancro al polmone ?

Difficile rispondere in termini di una sequenza casuale di 'prove' in cui si osserva se in alcune il fumo è una causa di cancro al polmone ed in altre no…

B. De Finetti 1906-1985

Definizione "***soggettiva***"* di probabilità:

La probabilità di un evento è il **grado di fiducia** - espresso tra 0 e 1 - che un individuo ha nel verificarsi dell'evento.
(Bruno De Finetti)

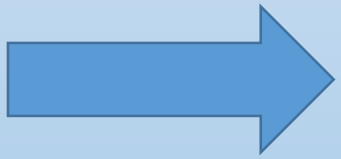
In altri termini, e usando il linguaggio delle scommesse:

La probabilità di un evento è il prezzo che un individuo ritiene ***equo*** pagare per ricevere **1** se l'evento si verifica e **0** se l'evento non si verifica. Le probabilità degli eventi devono essere attribuite in modo tale che non sia possibile ottenere con un insieme di scommesse una vincita *certa* o una perdita *certa*.

* Fondamento della impostazione «**bayesiana**» dell'inferenza statistica

Probabilità come frequenza:
L'approccio classico alla statistica definisce la probabilità di un evento come "il numero di volte che l'evento si verifica sul totale dei tentativi, nel limite di una serie infinita di ripetizioni equiprobabili".

Probabilità come grado di fiducia:
Il punto di vista soggettivo si basa sul principio semplice e intuitivo che: "la probabilità è una misura del grado di fiducia in una proposizione".



Vedremo come si forma questo *grado di fiducia* quando parleremo del Teorema di Bayes

R.A. Fisher 1890 – 1962

Fisher riteneva che la statistica dovesse essere ***oggettiva***, basata esclusivamente sui dati osservati e non influenzata da opinioni o *credenze* soggettive.

Ripetibilità: Fisher sottolineava l'importanza della **ripetibilità** degli esperimenti.

Interpretazione: Fisher era molto critico nei confronti dell'interpretazione soggettiva delle probabilità. A suo avviso, la probabilità doveva essere intesa esclusivamente come **frequenza relativa** a lungo termine.

B. De Finetti 1906-1985

Secondo De Finetti, la probabilità ***non è un fatto oggettivo***, ma riflette il **grado di fiducia** che una persona ha in un certo evento.

Coerenza: Le *credenze* devono essere logicamente coerenti tra loro, evitando contraddizioni.

Aggiornamento: Le probabilità non sono fisse, ma possono essere aggiornate alla luce di nuove informazioni.

Parametri : sono costanti **fisse** ma sconosciute.

Le **probabilità** sono sempre interpretate come **frequenza relativa** a lungo termine.

Le procedure statistiche sono giudicate in base a come si comportano a lungo termine su un *numero infinito* di *ripetizioni ipotetiche* dell'esperimento.

- Intervalli di confidenza
- Test di ipotesi

IMPOSTAZIONE FREQUENTISTA

Poiché siamo incerti sul valore vero dei **parametri**, li considereremo come **variabili casuali**.

Le regole della probabilità vengono utilizzate *direttamente* per fare inferenze sui parametri.

Probabilità a priori + dati => Probabilità a posteriori

- Intervalli di credibilità
- Fattore di Bayes

IMPOSTAZIONE BAYESIANA

Cenni di calcolo delle probabilità

Le operazioni di base del calcolo delle probabilità sono l'addizione e la moltiplicazione.

Esempio: qual è la probabilità che esca 1 oppure 3 in un lancio di un dado?

Se il dado è bilanciato, la probabilità di 1 è 1/6 e la probabilità di 3 è 1/6.

In nessun lancio potranno uscire ***contemporaneamente*** 1 e 3.

L'evento "***composto***" definito come: «esce la faccia 1 oppure esce la faccia 3»:

P(esce 1 oppure esce 3)=1/6 **+**1/6=1/3.

La faccia 1 e la faccia 3 non possono uscire contemporaneamente: sono due eventi *mutualmente esclusivi (incompatibili)*.

Altrimenti, la ***regola di addizione*** non sarebbe stata valida.

Complichiamoci un po' la vita…

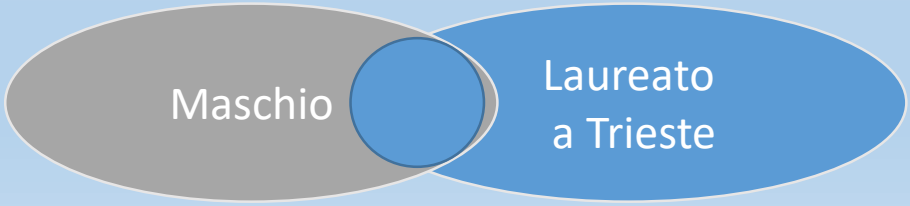
Supponiamo che il nome di uno specializzando di medicina venga estratto **casualmente** da un registro, e qualcuno ci dica che la probabilità che sia maschio è di 0.4 (40%) mentre la probabilità che si sia laureato a Trieste è di 0.8 (80%)…

…qual è la probabilità che lo specializzando che abbiamo estratto sia maschio ***oppure*** si sia laureato a Trieste, ***oppure*** entrambe le cose?

Se le due probabilità venissero addizionate, il risultato sarebbe: (0.4+0.8)=1.2,
→ sbagliato poichè la probabilità di un evento, anche composto, non puo' mai superare 1 !!!

L'errore risiede nel fatto che la probabilità del doppio evento: «specializzando maschio e laureato a Trieste" è stata contata 2 volte, una volta come parte della probabilità di essere maschio e una volta come parte della probabilità di essere laureato a Trieste.

i due eventi NON SONO mutualmente esclusivi sono «compatibili»



Indichiamo con:
P = probabilità dell'evento;
A = evento ['specializzando maschio'];
B=evento ['specializzando laureato a Trieste']
P(A e B) = probabilità dell'evento **congiunto** ['specializzando maschio e laureato a Trieste']

La forma piu' generale della *regola di addizione* è:

P(A o B) = P(A)+P(B)-P(A e B)

Supponiamo che la probabilità dell'evento **congiunto** P(A e B) è pari a 0.3; allora:

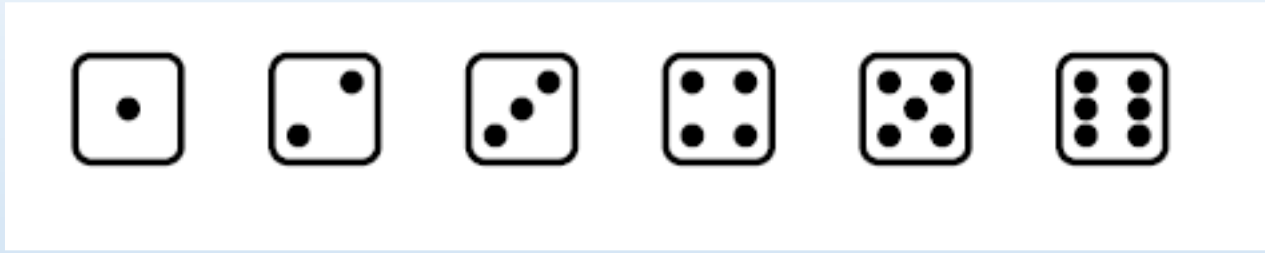
P(A o B)=(0.4+0.8) - 0.3=0.9.

Se i due eventi A e B sono **mutualmente esclusivi**, allora:

P(A e B)=0, e cosi' si ottiene la *forma semplice* della regola di addizione:

P(A o B)= P(A) + P(B)

Supponiamo adesso che ad ogni prova vengano lanciati **contemporaneamente** una moneta ed un dado: qual è la *probabilità congiunta* di testa (T) sulla moneta e 5 come faccia sul dado?



Regola di moltiplicazione:

P(T e 5)=P(T) * P(5, dato testa)=P(5)*P(T, dato 5)

P(5, dato testa) = '***probabilità condizionata***'

In questo particolare esempio, non c'è ragione di supporre che la probabilità di 5 sul dado sia in qualche modo 'condizionata' dall'evento Testa sulla moneta; in altre parole:

P(5, dato testa) = P(5)
P(T, dato 5)=P(T)

Notazione matematica: P(A, dato B) = P(A|B)

Se la probabilità condizionata è uguale alla probabilità non condizionata, i due eventi sono definiti ***indipendenti***, e si ottiene ***la forma semplice*** della regola di moltiplicazione:

$$P(T \text{ e } 5)=P(T)*P(5)=1/2 * 1/6=1/12$$

Nell'esempio dello specializzando maschio (A) e laureato a Trieste (B), se questi due eventi fossero indipendenti avremmo:

$$P(A \text{ e } B)=P(A)*P(B)=0.4*0.8=0.3$$

Se questi due eventi invece non fossero indipendenti, perchè per esempio i corsi di laurea a Trieste accettano *come regola* più donne che uomini, il valore corretto per *P(A e B)* andrebbe ottenuto accertando il valore della probabilità condizionata:

$$P(B=laureato\ a\ Trieste|A=maschio)$$

Variabili Aleatorie/Casuali e Distribuzioni di probabilità

Prima di lanciare una moneta possiamo solo «predire» il risultato con una certa probabilità.
La *funzione* che associa la probabilità agli eventi è una variabile aleatoria.

il risultato di un esperimento che non può essere determinato con certezza (solo la probabilità di ottenere un certo risultato può essere stimata)

Le VA sono descritte da *distribuzioni di probabilità*

Es: Lancio di una moneta: testa o croce hanno probabilità P=1/2

Distribuzioni di Probabilità

Che cosa è una distribuzione di probabilità ?

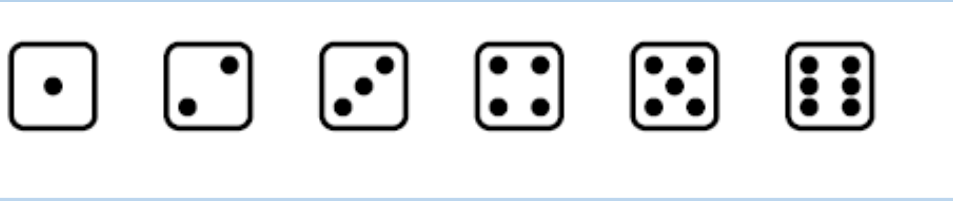
Outcome X	Probability of outcome $P(X)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

E’ una tabella (equazione) che ci dice quale è la probabilità di ciascun evento o risultato di un «esperimento»

Questa tabella ci dice quale è la distribuzione di probabilità che otteniamo per i vari risultati dei lanci ripetuti di un dado.

L’equazione che genera questa distribuzione di probabilità è:

$P(X)=1/6$



Variabili casuali discrete

Le VA si dicono **DISCRETE** se il meccanismo che genera i dati è un conteggio oppure del tipo «presenza/assenza» (il *range di valori* è un insieme finito o numerabile).

Immaginiamo un mazzo di carte con i quattro assi, tre due, due tre e un quattro (10 carte) e definiamo la variabile casuale:

$$P(x = j) = \begin{cases} 0.4\ se\ x = 1 \\ 0.3\ se\ x = 2 \\ 0.2\ se\ x = 3 \\ 0.1\ se\ x = 4 \end{cases}$$

La funzione o *variabile casuale P(x)* assegna una probabilità ad ogni carta; la somma delle probabilità di tutti gli eventi è sempre pari ad 1.

In questo caso i valori che la v.c. può assumere sono discreti e quindi si ottiene una distribuzione di probabilità discreta [una distribuzione di frequenza]

Si può definire anche la ***funzione cumulata di frequenza***, detta anche ***funzione di ripartizione***:

$$P(x \leq j) = \begin{cases} 0.4 \; se \; x \leq 1 \\ 0.7 \; se \; x \leq 2 \\ 0.9 \; se \; x \leq 3 \\ 1.0 \; se \; x \leq 4 \end{cases}$$

Possiamo definire poi la **media** (***valore atteso***) di una variabile casuale discreta come:

Expected Value: $$\mu = E(X) = \sum x_i P(X = x_i)$$

Possiamo infine definire anche la **varianza** di una variabile casuale discreta come:

$$\sigma^2 = Var(X) = E[X - E(X)]^2 = \sum (x_i - \mu)^2 P(X = x_i)$$

VA di Bernoulli

Jakob Bernoulli, 1654-1705

Una V.A. X di Bernoulli assume solo due valori: 0 o 1
fallimento/successo ; assenza/presenza…

La distribuzione di X è caratterizzata da una probabilità costante **p** di «successo» :

P(X=1)=p ; P(X=0)=1-p

Esempio: lancio di una moneta: X=1 se esce testa; X=0 esce croce -> p(X=1)=p(X=0)=1/2

$$X \approx Bernoulli(p)$$

Una sequenza di lanci di una moneta è una sequenza di realizzazioni di una VA di Bernoulli

VA Binomiale

Una VA X Binomiale «conta» il numero *k* dei successi di *n* VA di Bernoulli.

$$Y = X_1 + X_2 + \ldots + X_n \qquad Y \approx Bin(n, p)$$

Es: Quale è la probabilità che in *n* lanci di una moneta esca *k* volte testa?

Due parametri descrivono la distribuzione di X:

- il numero delle prove ***n***
- la probabilità di successo ***p***

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

Coefficiente binomiale:

Conta tutti i possibili modi in cui possiamo estrarre k elementi da n

$$\binom{n}{k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{1*2*3\ldots*k} = \frac{n!}{(n-k)!k!}$$

[OPZIONALE]

$$E(X) = \sum_{i=1}^{k} x_i * P(x_i)$$ Valore atteso (media) di una v.c. discreta

Mean of binomial. The mean of the *binomial*(n, π) distribution is the sample size times the probability of success since

$$\begin{aligned} \mathrm{E}[Y|\pi] &= \sum_{y=0}^{n} y \times f(y|\pi) \\ &= \sum_{y=0}^{n} y \times \binom{n}{y} \pi^y (1-\pi)^{n-y} . \end{aligned}$$

We write this as a conditional mean because it is the mean of Y given the value of the parameter π. The first term in the sum is 0, so we can start the sum at $y = 1$. We cancel y in the remaining terms, and factor out $n\pi$. This gives

$$\mathrm{E}[Y|\pi] = \sum_{y=1}^{n} n\pi \binom{n-1}{y-1} \pi^{y-1} (1-\pi)^{n-y} .$$

Factoring $n\pi$ out of the sum and substituting $n' = n - 1$ and $y' = y - 1$, we get

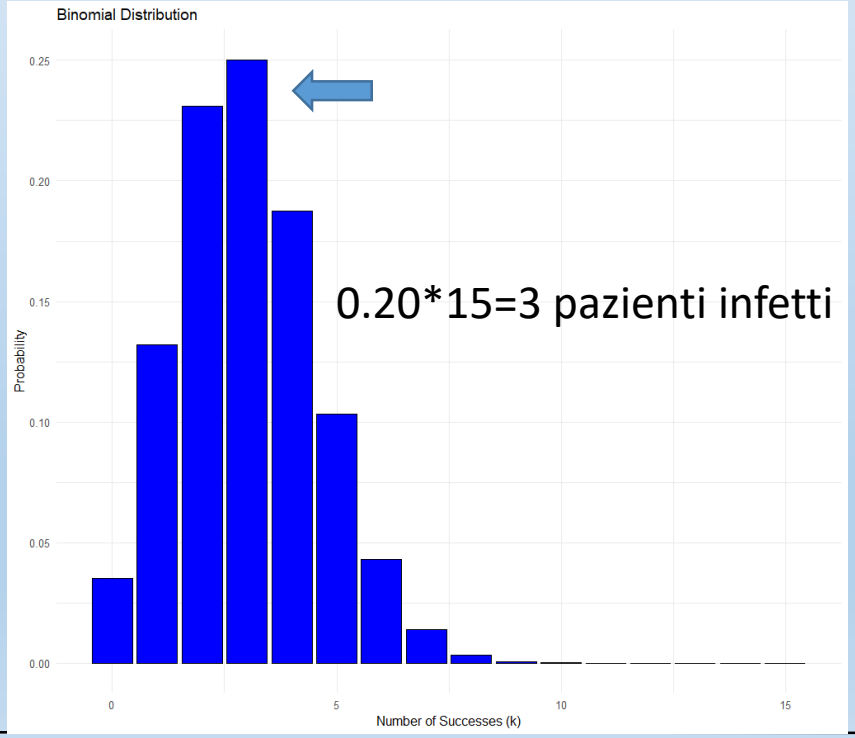
$$\mathrm{E}[Y|\pi] = n\pi \sum_{y'=0}^{n'} \binom{n'}{y'} \pi^{y'} (1-\pi)^{n'-y'} .$$

We see the sum is a binomial probability function summed over all possible values. Hence it equals one, and the mean of the binomial is

$$\mathrm{E}[Y|\pi] = n\pi . \qquad (5.7)$$

Il valore atteso della distribuzione di probabilità di una variabile casuale binomiale è il prodotto tra il numero di «prove» per la probabilità di successo in ogni prova.

La probabilità di contrarre una infezione in ospedale è una v.c. binomiale con p=0.20.
Abbiamo 15 pazienti ricoverati nel nostro reparto.
Quanti pazienti infetti «ci aspettiamo» di osservare (valore atteso?)



La media di una variabile casuale binomiale è il prodotto **n*p**

[OPZIONALE]

Varianza della variabile aleatoria binomiale:

Variance of binomial. The variance is the sample size times the probability of success times the probability of failure. We write this as a conditional variance since it is the variance of Y given the value of the parameter π. Note that

$$\begin{aligned}\mathrm{E}[Y(Y-1)|\pi] &= \sum_{y=0}^{n} y(y-1) \times f(y|\pi) \\ &= \sum_{y=0}^{n} y(y-1) \times \binom{n}{y} \pi^y (1-\pi)^{n-y} .\end{aligned}$$

The first two terms in the sum equal 0, so we can start summing at $y = 2$. We cancel $y(y-1)$ out of the remaining terms and factor out $n(n-1)\pi^2$ to get

$$\mathrm{E}[Y(Y-1)|\pi] = \sum_{y=2}^{n} n(n-1)\pi^2 \binom{n-2}{y-2} \pi^{y-2}(1-\pi)^{n-y} .$$

Substituting $y' = y - 2$ and $n' = n - 2$, we get

$$\begin{aligned}\mathrm{E}[Y(Y-1)|\pi] &= n(n-1)\pi^2 \sum_{y'=0}^{n-2} \binom{n'}{y'} \pi^{y'}(1-\pi)^{n'} \\ &= n(n-1)\pi^2\end{aligned}$$

since we are summing a binomial distribution over all possible values. The variance can be found by

$$\begin{aligned}\mathrm{Var}[Y|\pi] &= \mathrm{E}[Y^2|\pi] - [\mathrm{E}[Y|\pi]]^2 \\ &= \mathrm{E}[Y(Y-1)|\pi] + \mathrm{E}[Y|\pi] - [\mathrm{E}[Y|\pi]]^2 \\ &= n(n-1)\pi^2 + n\pi - [n\pi]^2 .\end{aligned}$$

Hence the variance of the binomial is the sample size times the probability of success times the probability of failure.

$$\mathrm{Var}[Y|\pi] = n\pi(1-\pi) . \qquad (5.8)$$

La varianza di una variabile casuale binomiale è il prodotto **n*p*(1-p)**

Applicazioni della VA Binomiale

La distribuzione binomiale descrive il **numero** di volte in cui un particolare evento si verifica in una sequenza (casuale) di osservazioni (=campione).

Questa distribuzione può essere utilizzata quando l'***outcome*** di interesse è l'occorrenza di un evento, non la sua "***intensità***".

• In un trial clinico, la condizione di un paziente può migliorare oppure no. Si intende valutare la probabilità di miglioramento, non "*di quanto*" i pazienti siano migliorati.

• Quante persone "ansiose" ci sono in questa aula ? La distribuzione binomiale potrebbe valutare il numero di persone ansiose, non il loro "grado di ansia".

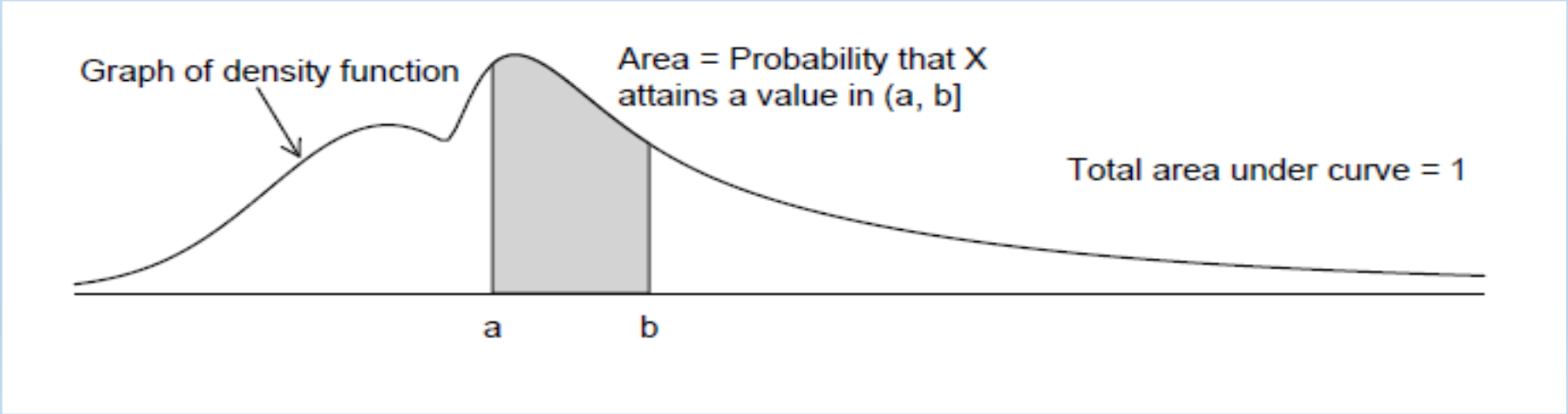
• Nel controllo di qualità, si vuole identificare il numero di lotti difettati in una filiera produttiva, *indipendentemente* dal tipo di difetto.

VA continua

- Le VA sono **CONTINUE** se il meccanismo generatore dei dati è un processo di misurazione su una scala numerica: peso corporeo, glicemia…

- Essendo la scala di misura continua ha senso chiedersi con quale probabilità i valori cadano in un certo **intervallo** più che identificare dei valori «puntuali»*…

Distribuzione di probabilità di una VA X continua: viene descritta tramite una ***funzione di densità*** la cui area sotto la curva rappresenta la probabilità:



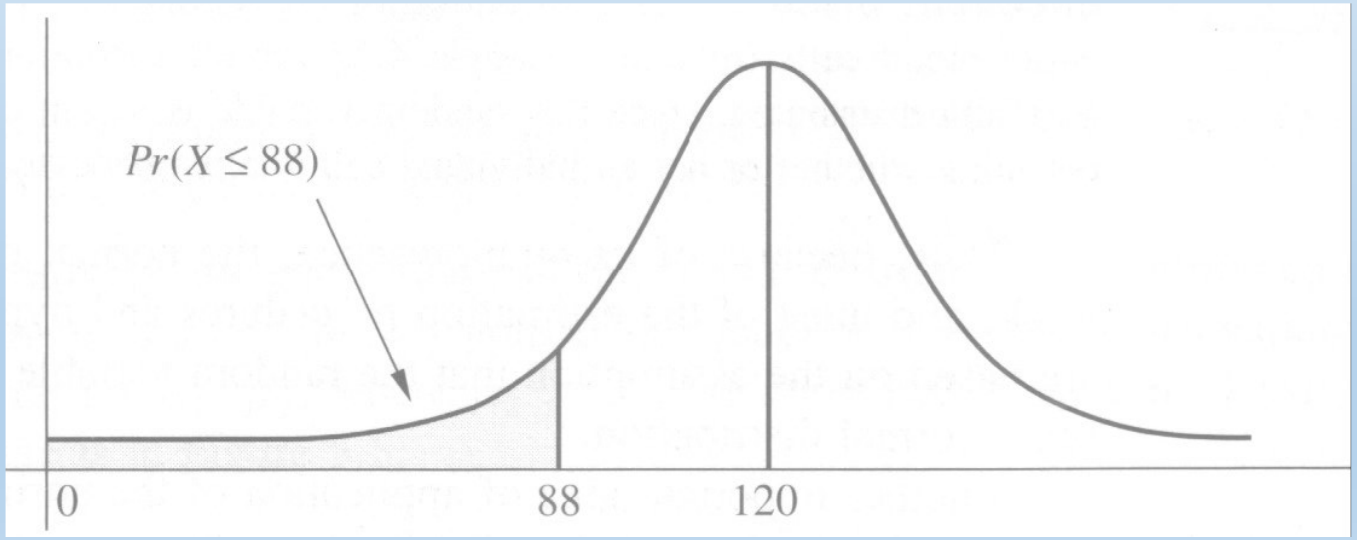
*essendoci potenzialmente infiniti valori…

La distribuzione di **probabilità cumulativa** di una variabile casuale continua **X** valutata in un punto ***a*** è **P(X<=a)**

Rappresenta l'area sotto la curva ‘alla sinistra’ del valore ***a***

Es: la distribuzione cumulativa del peso alla nascita valutata in 88 once (˜2,5 kg) è pari a P(X<= 88) ed corrisponde all'area sotto la curva a sinistra di 88 once.

Questo valore è usato come limite per possibili outcomes sfavorevoli nel primo anno di vita:



Per calcolare le probabilità cumulative al posto di utilizzare delle **sommatorie**
nel caso delle variabili casuali continue si utilizzando gli **integrali.**

Per una variabile casuale *continua* (dove ciò che si ottiene è detto *funzione di densità* o *densità di frequenza*) la funzione di ripartizione (=probabilità cumulata), la media e la devianza sono definite ricorrendo al concetto di *integrale*:

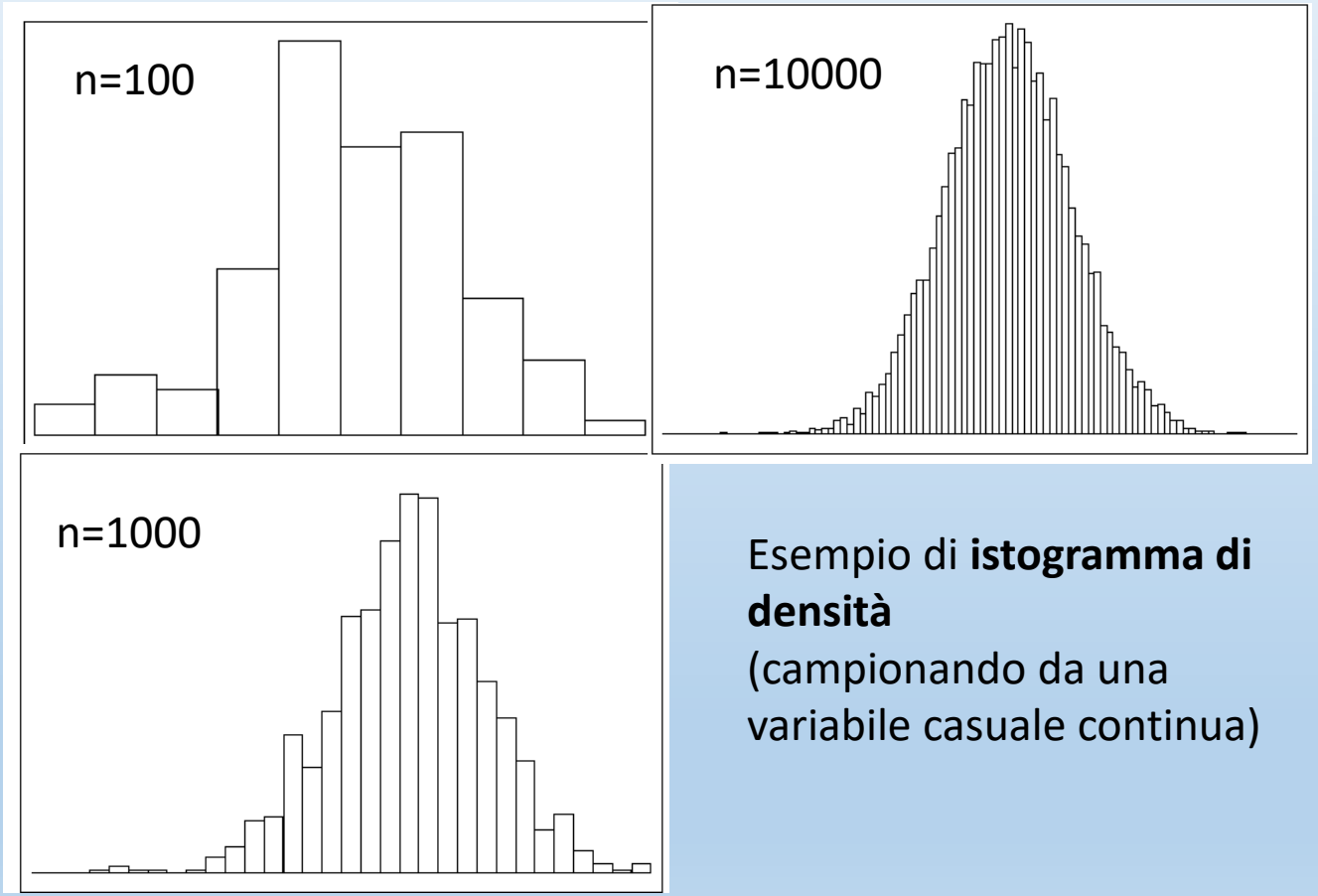
Funzione di densità $P(X) = f(x)$ $\int_{-\infty}^{\infty} f(x)dx = 1$

Funzione di ripartizione $P(X < x) = \int_{-\infty}^{x} f(x)dx$

Media o valore atteso $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

Varianza $\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

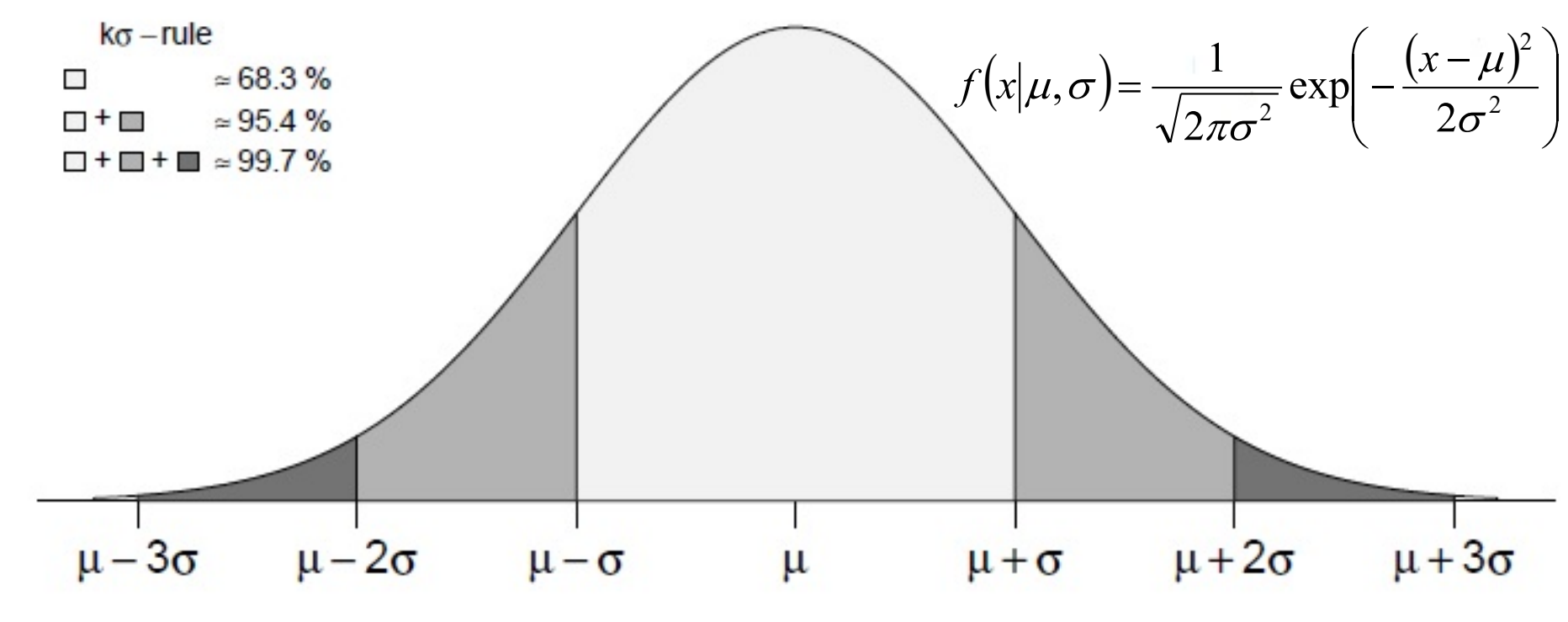
$P(a < X < b) = \int_{a}^{b} f(x)dx$



Esempio di **istogramma di densità** (campionando da una variabile casuale continua)

VA GAUSSIANA (o Normale)

Alcuni fenomeni (continui) possono essere descritti da una distribuzione di probabilità a forma di «**campana**»:



C.F. Gauss (1777-1855)

$N(\mu,\sigma^2)$ è definita dai parametri μ e σ (media e deviazione standard)

$N(0,1)$ distribuzione normale «standardizzata» (μ=0 e σ=1) $\longrightarrow$ $\frac{X-\mu}{\sigma}=Z\approx N(0,1)$

[OPZIONALE]

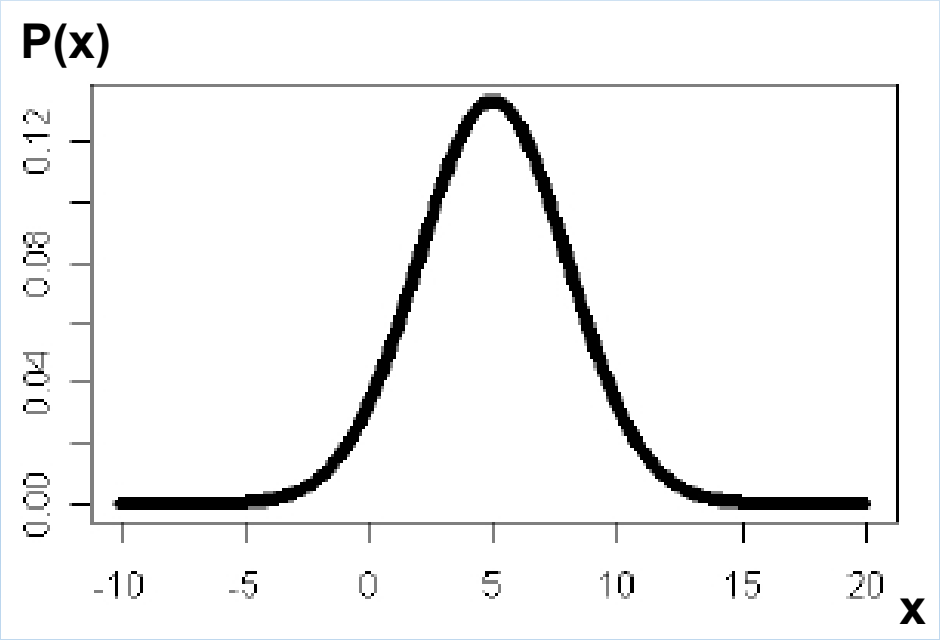
Densità della V.C. Gaussiana: exp=e

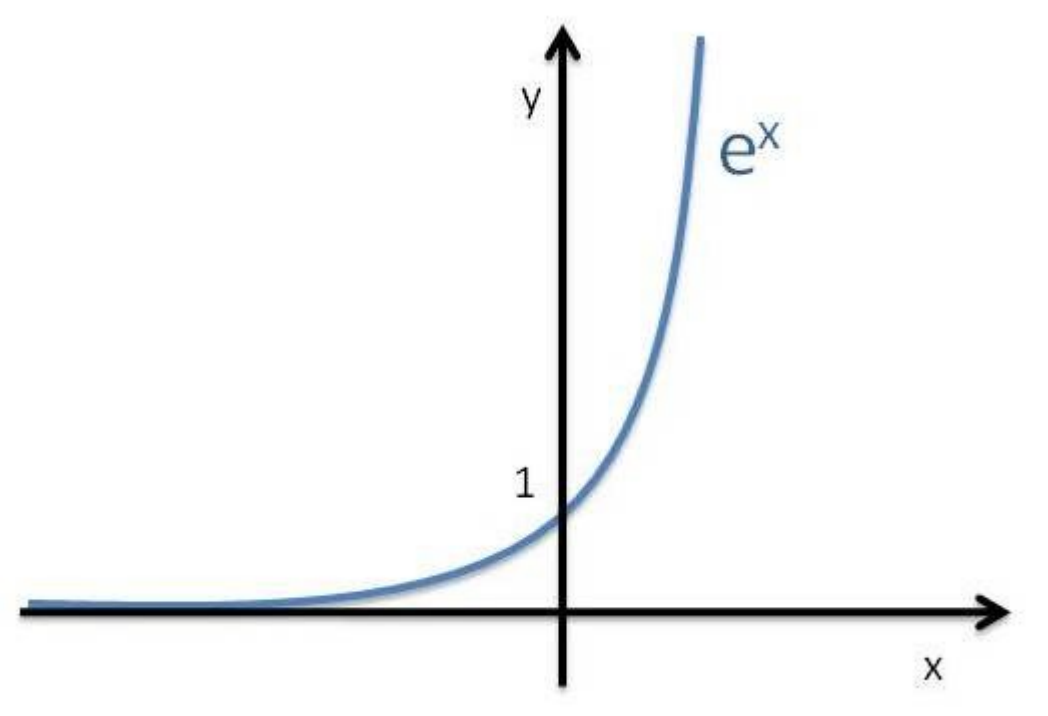
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

π=3.14

In matematica, la funzione esponenziale è la funzione che associa a un valore ***x*** l'elevamento a potenza con base il numero di Eulero ***e*** (una costante matematica il cui valore approssimato è 2.718) e ed esponente ***x***. Viene solitamente rappresentata come e^x, oppure $\exp(x)$.

N(μ;σ)=N(5;3)





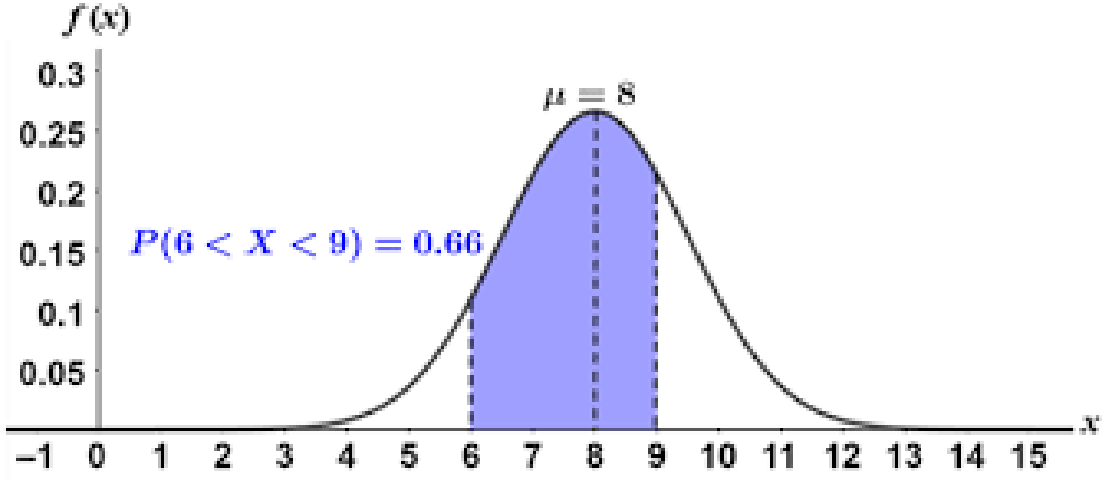
Funzione di densità della V.C. Gaussiana standard: **N(μ;σ)=N(0;1)**

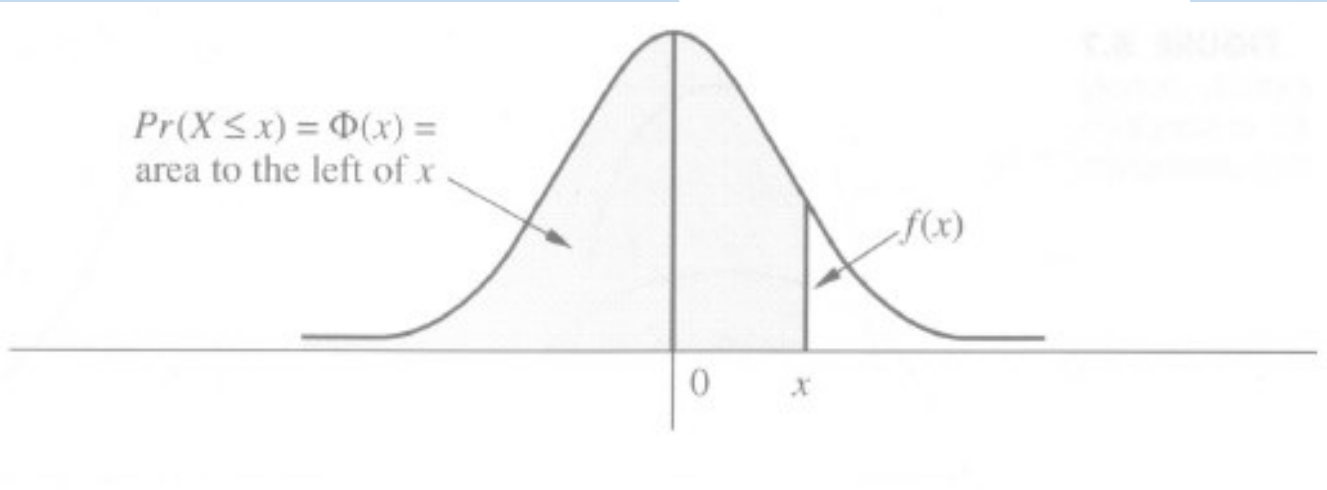
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty$$

Possiamo definire **la probabilità di *qualsiasi intervallo*** utilizzando il calcolo integrale basato sulla distribuzione standardizzata.

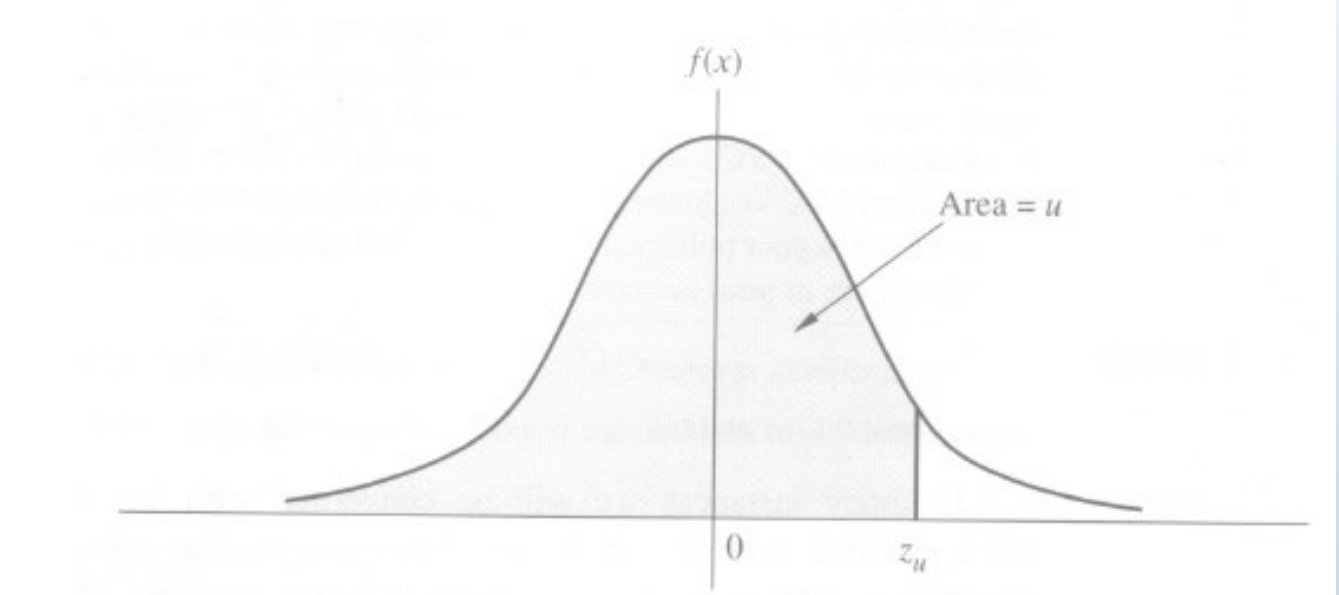
Funzione di ripartizione

$$\Phi(x) = P(X \leq x)$$





Calcolo dei *percentili* di una Normale standard



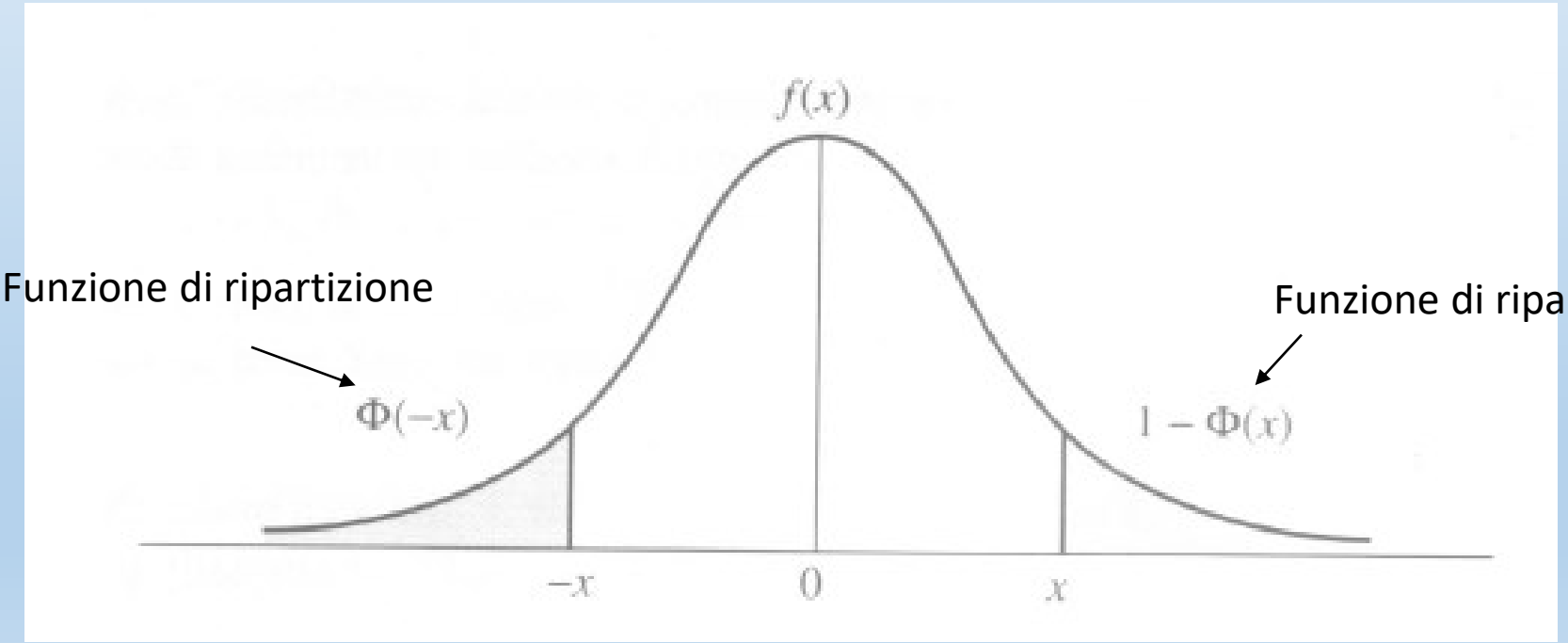
Il (100-*u*) percentile di una normale standard è indicato con z_u ed è definito dalla relazione:

$$P(X < z_u) = u$$

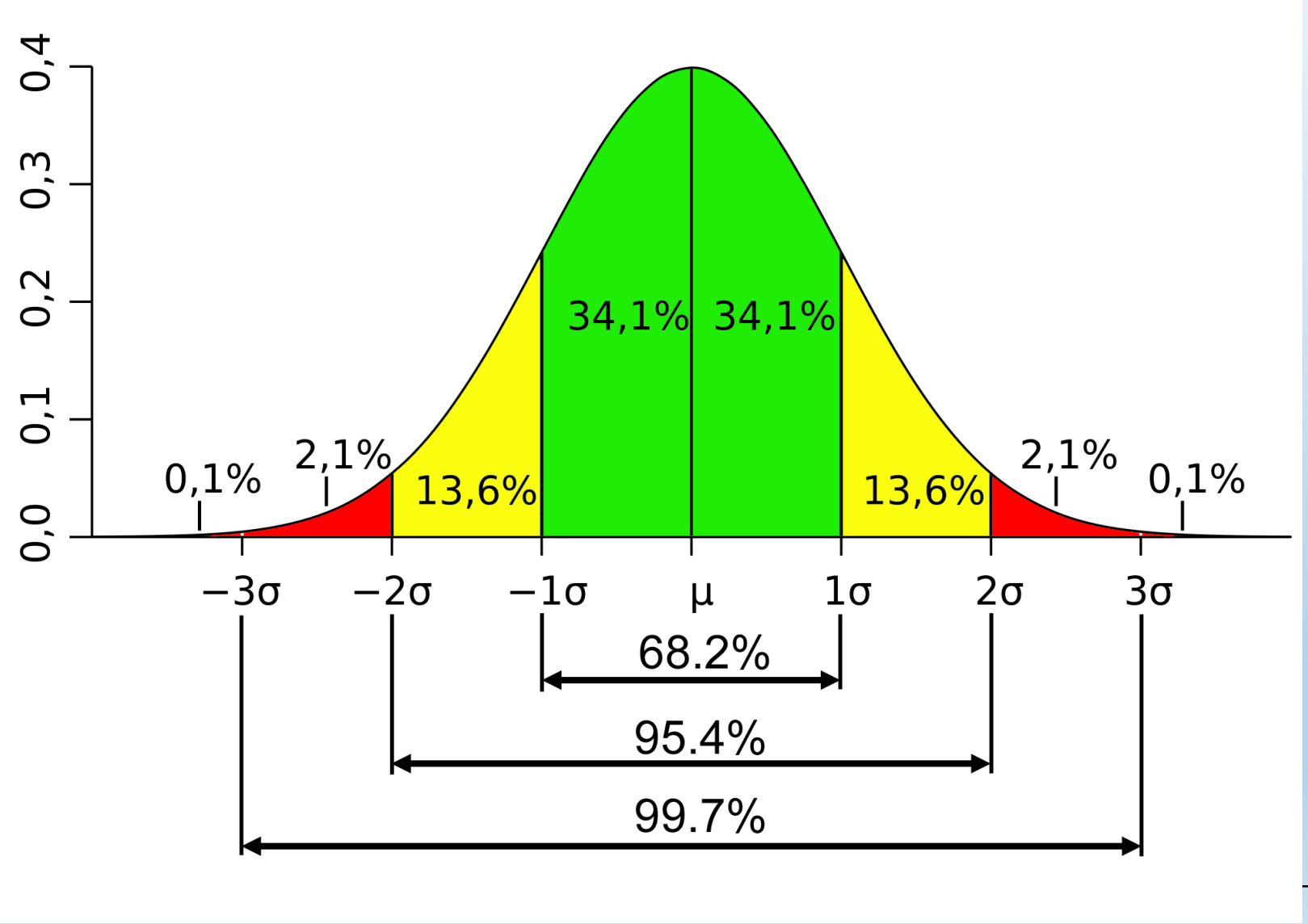
Per trovare z_u si deve trovare l'area u e quindi trovare il valore che corrisponde a tale area.

Si può usare la **simmetria** della distribuzione: $z_u = -z_{1-u}$

$$z_{.975} = 1.96 \quad z_{.95} = 1.645 \quad z_{.5} = 0 \quad z_{.025} = -1.96$$

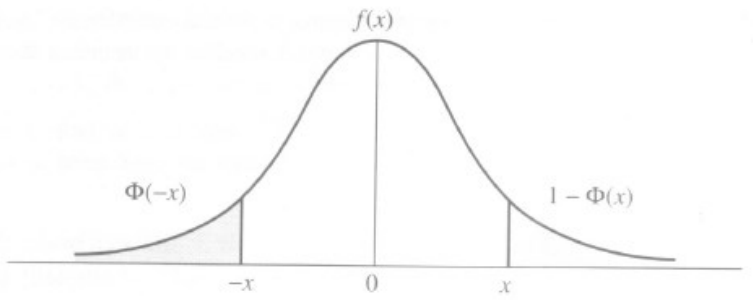


La «**forma**» della distribuzione normale (curva gaussiana)



Date le proprietà della curva normale, sappiamo ad esempio anche la % di unità statistiche che cadono nelle «*code*» della distribuzione.

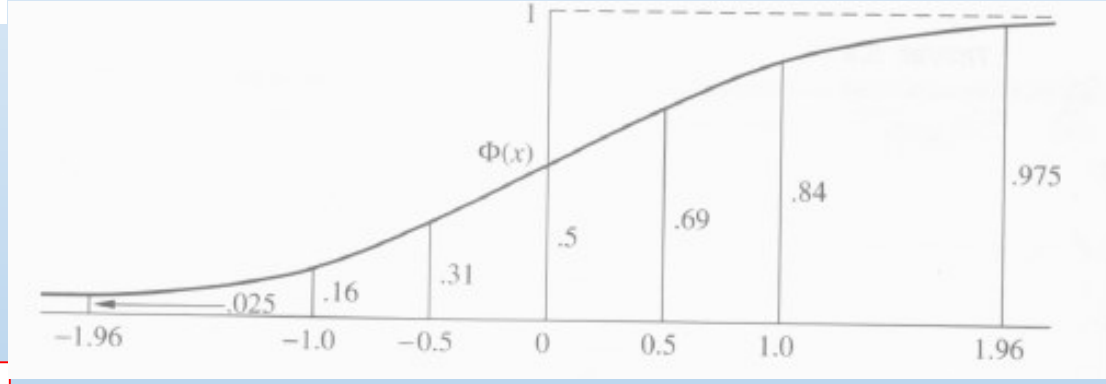
Code: definite da «**salti**» rispetto alla media di entità pari alla deviazione standard.



$$\Phi(-x) = P(X \leq -x) = P(X \geq x) = 1 - P(X \leq x) = 1 - \Phi(x)$$

Funzione di ripartizione

$$\Phi(x) = P(X \leq x)$$



Si calcoli $P(X \leq 1.96)$ e $P(\leq 1)$. Dalle tavole si ha che

$$\Phi(1.96) = 0.975 \quad \Phi(1) = 0.8413$$

Si calcoli $P(X \leq -1.96)$ Dalle tavole si ha che

$$\Phi(-1.96) = P(X \geq 1.96) = 0.0250$$

Ricapitolando:

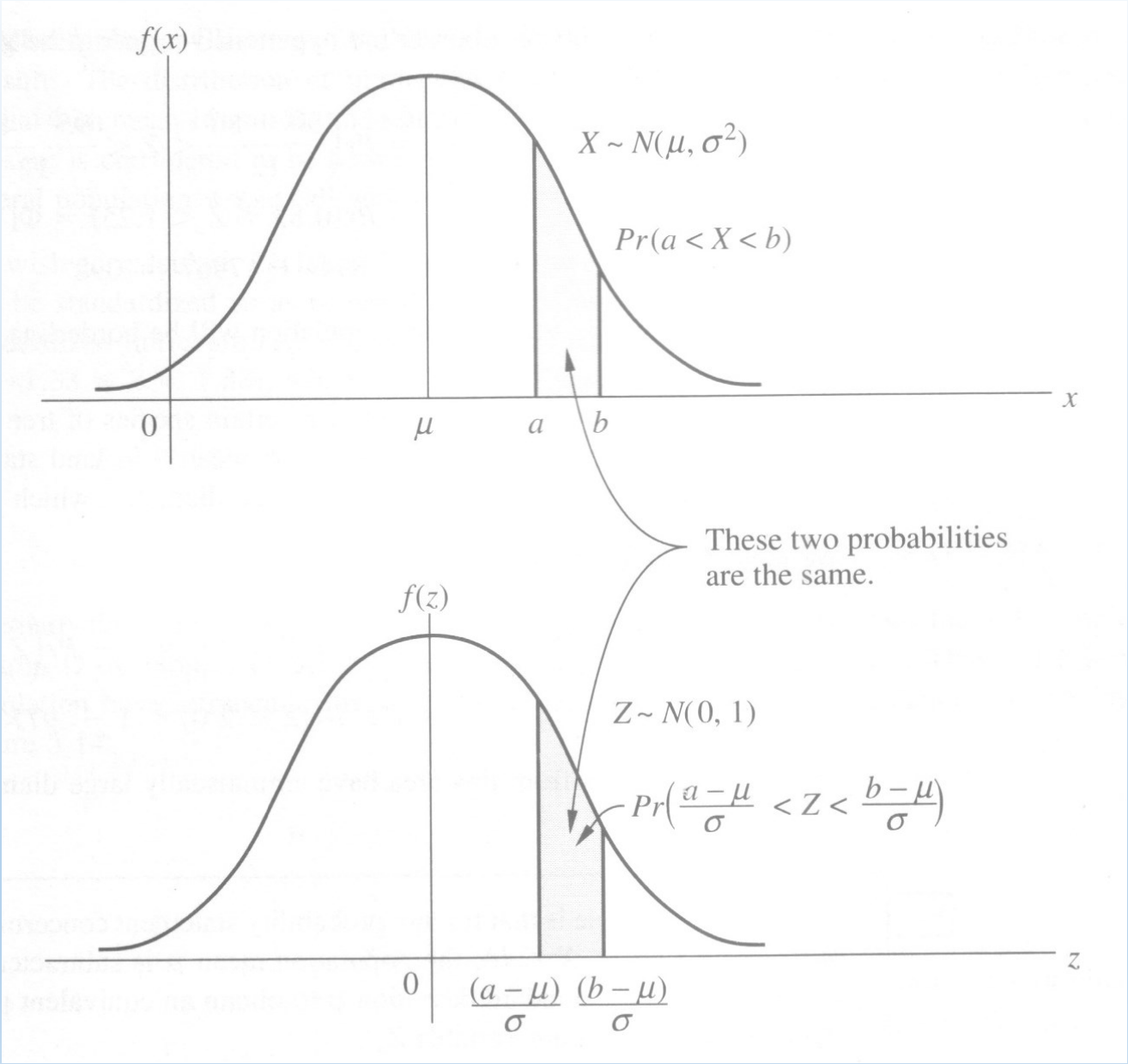
Dalla Normale alla Normale Standard:

Se: $$X \sim N(\mu, \sigma^2) \quad e \quad Z = \frac{X - \mu}{\sigma}$$

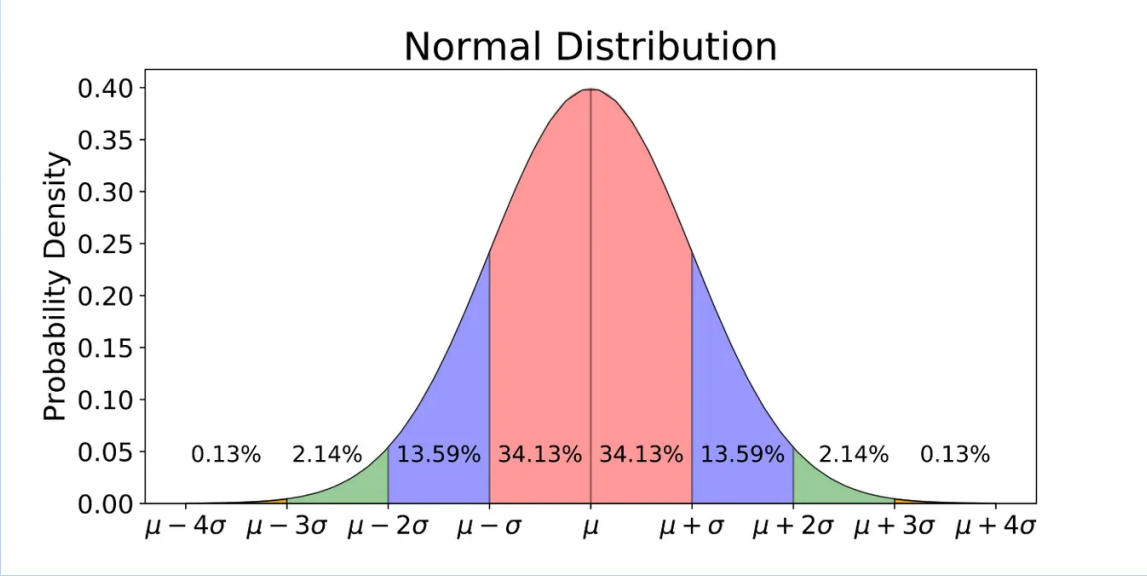
Allora: $$Z \sim N(0, 1)$$

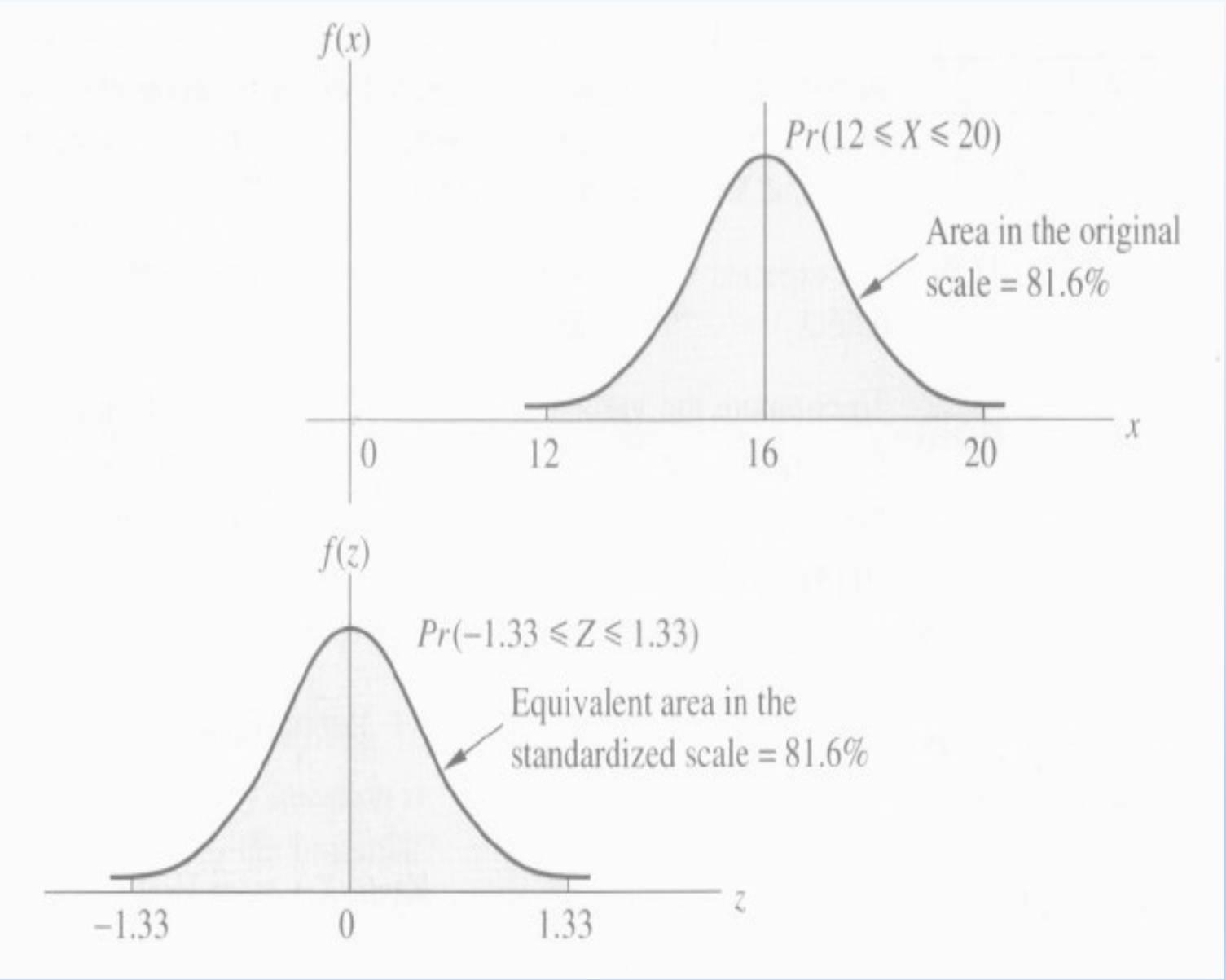
in tali condizioni si ha che:

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi[(b - \mu)/\sigma] - \Phi[(a - \mu)/\sigma]$$



Il motivo per cui Gauss decise di utilizzare la v.c. Normale Standardizzata è legato al fatto che così gli integrali da risolvere per determinare le aree di probabilità in certi intervalli sono ***generali*** e non dipendono da μ da σ, valori che possono invece variare da applicazione ad applicazione.





Le distribuzioni normali sono **infinite** (infiniti sono i valori possibili per μ e σ) ma possono tutte essere ricondotte **ad una sola distribuzione** di riferimento con $\mu = 0$ e $\sigma = 1$: la distribuzione **normale standardizzata**.

Ciò permette di risolvere il problema del calcolo di frequenza o di probabilità semplicemente ricorrendo alle **tavole degli integrali** della distribuzione normale standardizzata.

Una persona **borderline** per ipertensione è definita come una DBP compresa tra 90 e 95 mm Hg.
Nei maschi tra i 35 ed i 44 anni la DBP **è distribuita normalmente** con media 80 e varianza 144 (dev std=12).
Quale è la probabilità di essere borderline per un maschio in quella classe di età ??

$$P(90 < X < 95) = P\left(\frac{90-80}{12} \leq Z \leq \frac{95-80}{12}\right)$$

$$= \Phi[1.25] - \Phi[0.83] = 0.8944 - 0.7967 = 0.098$$

La probabilità di essere borderline per un maschio tra 35 e 44 anni è pari a 0.098 (circa 10%)

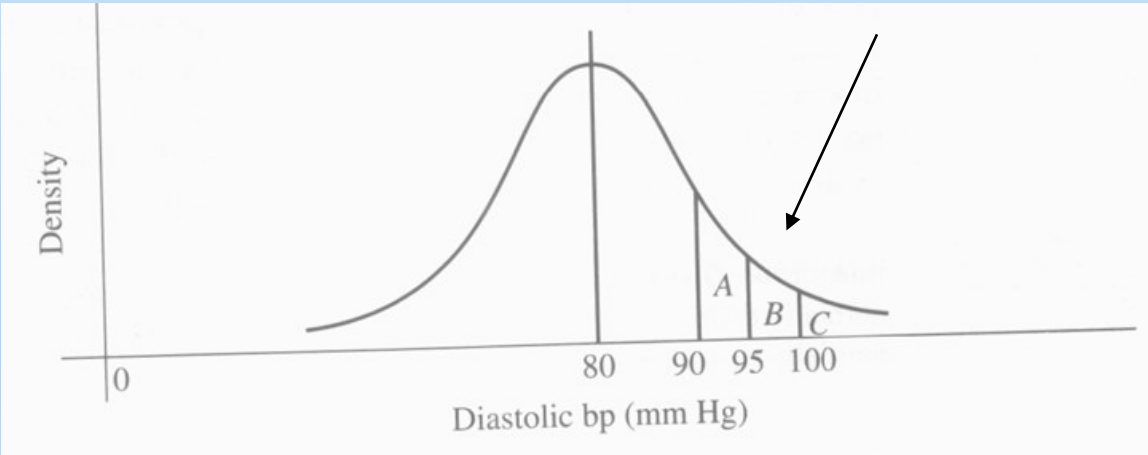
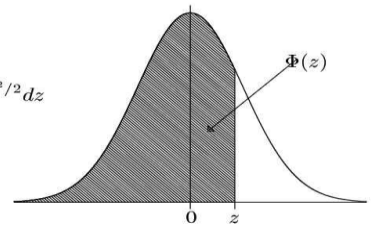


Tavola 1: Funzione di ripartizione della Variabile Casuale Normale Standardizzata

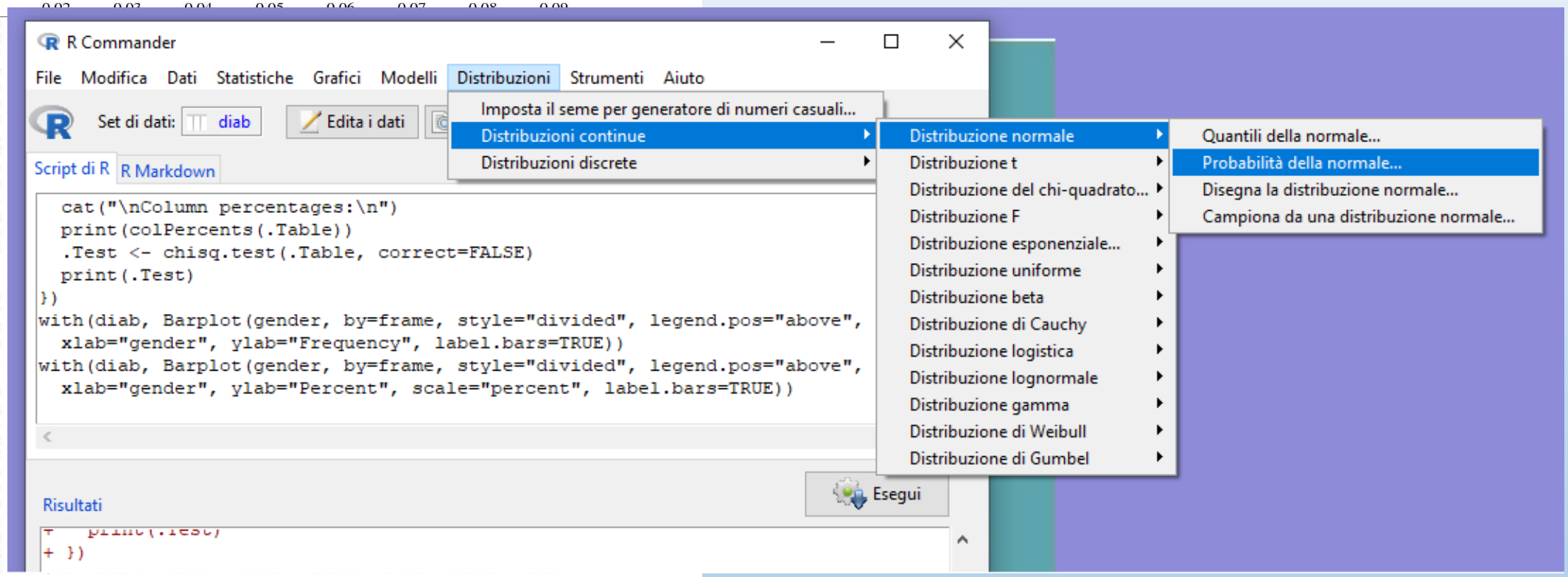
$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003								
-3.3	0.0005	0.0005								
-3.2	0.0007	0.0007								
-3.1	0.0010	0.0009								
-3.0	0.0013	0.0013								
-2.9	0.0019	0.0018								
-2.8	0.0026	0.0025								
-2.7	0.0035	0.0034								
-2.6	0.0047	0.0045								
-2.5	0.0062	0.0060								
-2.4	0.0082	0.0080								
-2.3	0.0107	0.0104								
-2.2	0.0139	0.0136								
-2.1	0.0179	0.0174								
-2.0	0.0228	0.0222								
-1.9	0.0287	0.0281								
-1.8	0.0359	0.0351								
-1.7	0.0446	0.0436								
-1.6	0.0548	0.0537								
-1.5	0.0668	0.0655								
-1.4	0.0808	0.0793								
-1.3	0.0968	0.0951								
-1.2	0.1151	0.1131								
-1.1	0.1357	0.1335								
-1.0	0.1587	0.1562								
-0.9	0.1841	0.1814								
-0.8	0.2119	0.2090								
-0.7	0.2420	0.2389								
-0.6	0.2743	0.2709								
-0.5	0.3085	0.3050								
-0.4	0.3446	0.3409								
-0.3	0.3821	0.3783								
-0.2	0.4207	0.4168								
-0.1	0.4602	0.4562								
-0.0	0.5000	0.4960								

Oggi non usiamo più le tavole di Gauss!

Ci viene in aiuto il computer, e i software di statistica ….



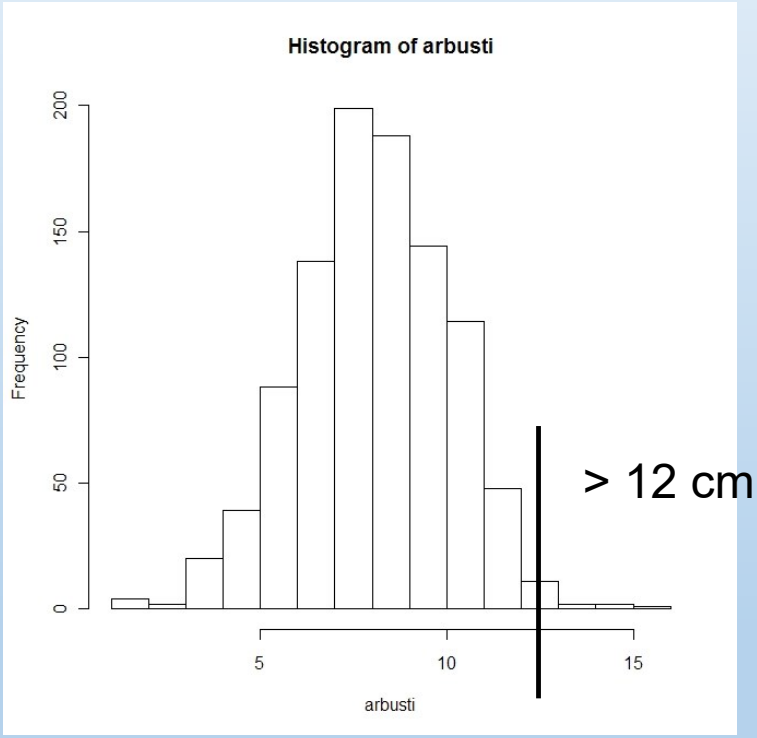
1- 0.1056=0.8944

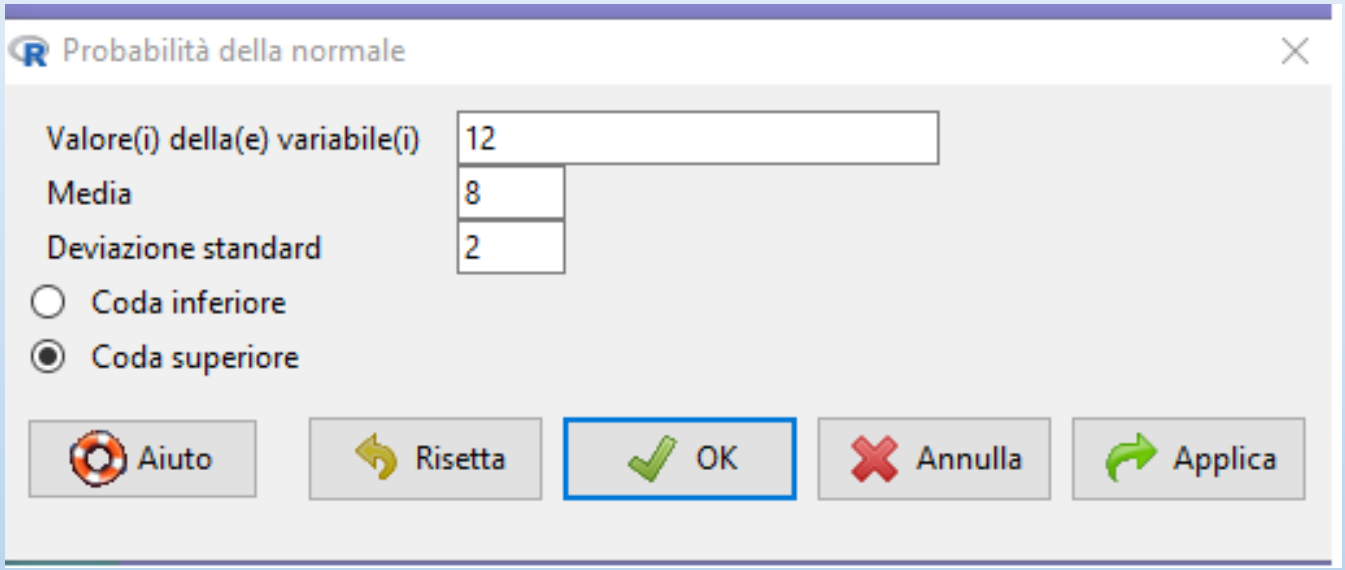
1- 0.2033=0.7967

$\Phi(1.25) - \Phi(0.83) = 0.098$

Es: Si supponga che i diametri di alcuni arbusti siano distribuiti normalmente con media 8 cm e deviazione standard 2 cm.

La probabilità per un arbusto di avere un diametro eccezionalmente grande (> 12cm) è pari a:





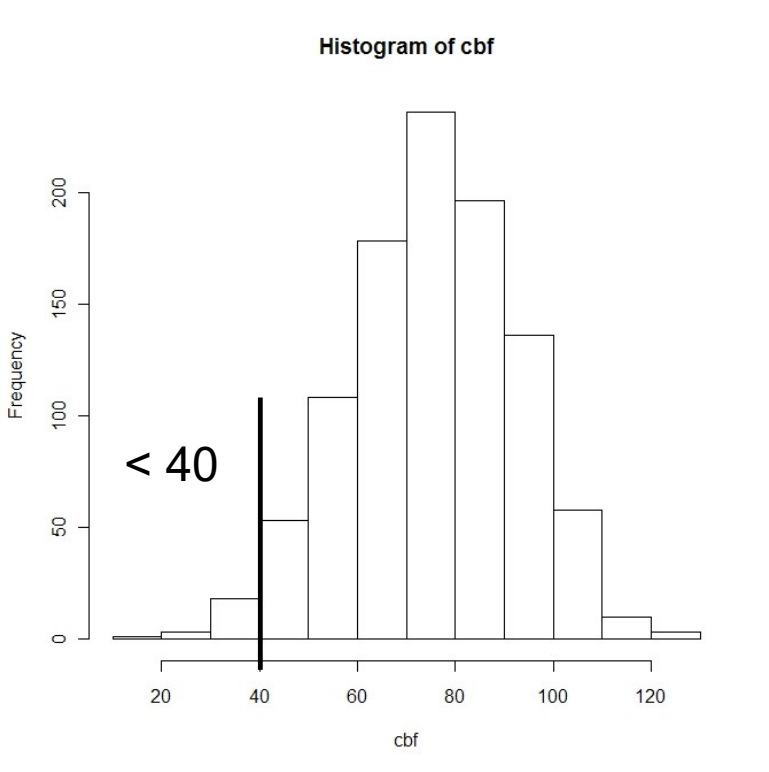
pnorm(c(12), mean=8, sd=2, lower.tail=FALSE)

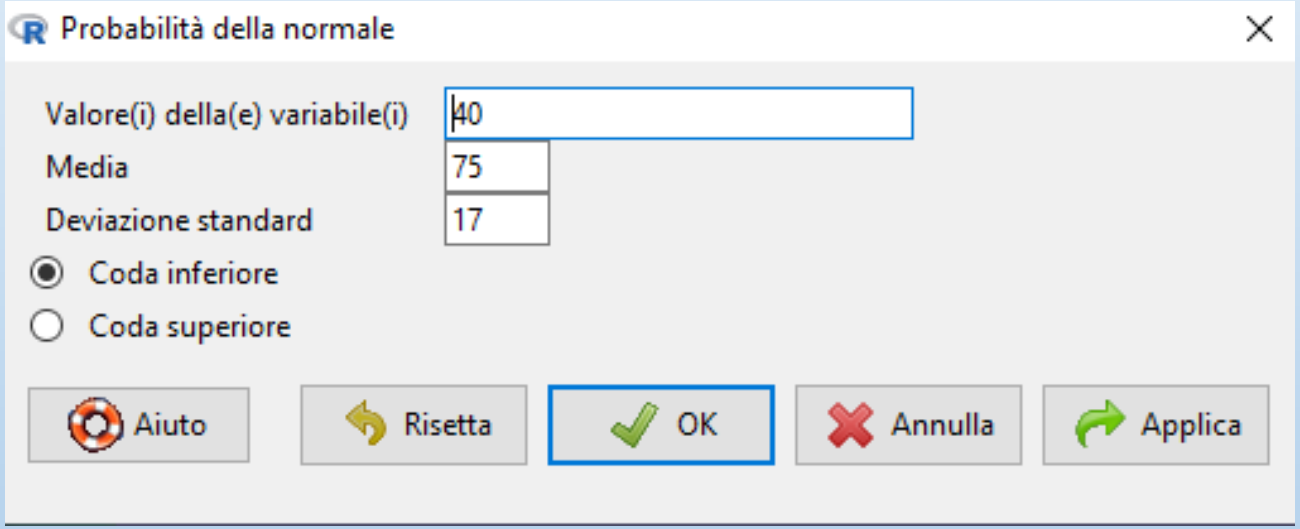
0.02275013 circa il 2.2%

Es: Nella popolazione generale la pressione vascolare cerebrale CBF è distribuita con media 75 e deviazione standard 17.

Un paziente è definito a rischio di stroke se ha una CBF < 40.

La probabilità di essere a rischio [provenendo dalla popolazione generale] è pari a:





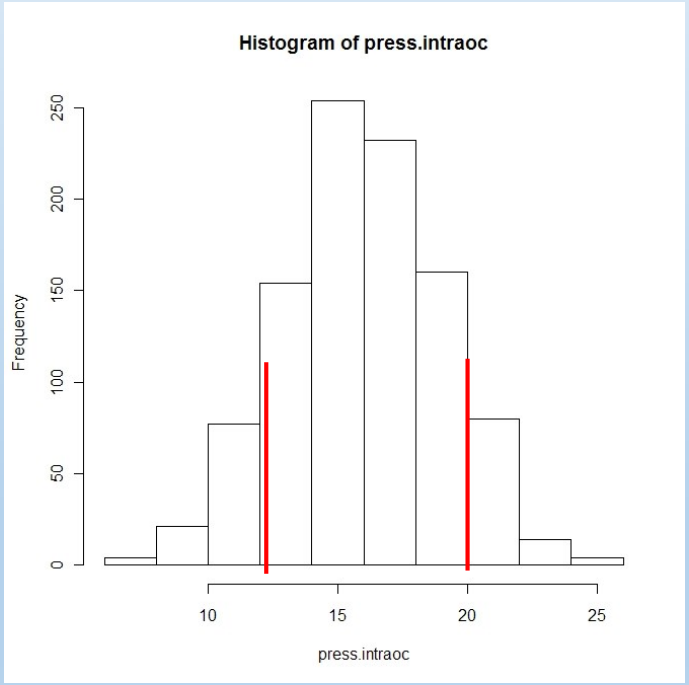
pnorm(c(40), mean=75, sd=17, lower.tail=TRUE)

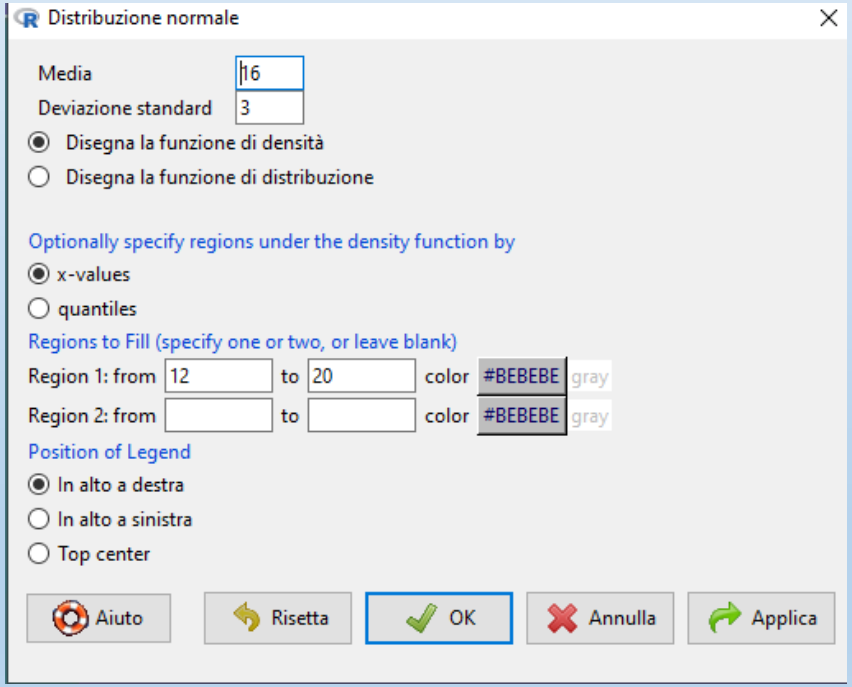
0.01975557 circa il 2%

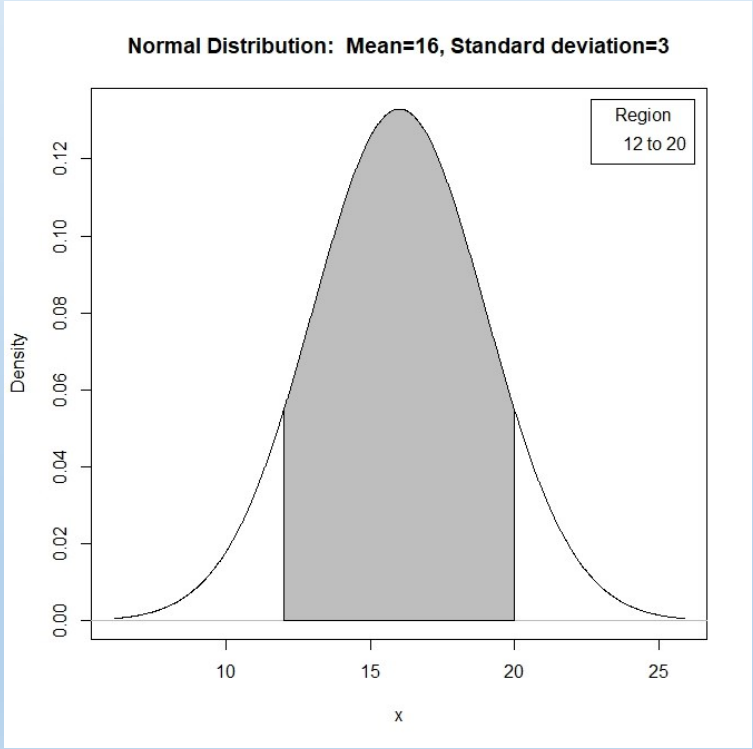
Es: Il glaucoma è una malattia che si presenta con una pressione intraoculare eccessiva.

Nella popolazione generale, la pressione intraoculare è normale con media 16 mm Hg e deviazione standard 3 mm Hg.

Se il range di normalità è tra i 12 ed i 20 mm Hg, la probabilità di essere normali è pari a:







0.8175776

La probabilità di ricadere nel range di normalità è pari a 0.82 (82%).

MATERIALE OPZIONALE

Cenni di calcolo combinatorio

Cenni di calcolo combinatorio: le permutazioni

Le permutazioni sono come gli anagrammi ed indicano **in quanti modi** (*ordinamenti*) **diversi** possono essere presi *n* oggetti [conta l'ordine].

Le permutazioni di *n* elementi sono date dallo *sviluppo fattoriale* di *n* (ossia *n* è moltiplicato per tutti i numeri interi inferiori ad *n*).

Ad esempio, date 4 lettere (ABCD), in quanti ordini si possono presentare?
(ABCD, BACD, DACB, ...)?

$$n! = n*(n-1)*(n-2)*(n-3)*\ldots$$

$$4! = 4*3*2*1 = 24$$

$$n! = \begin{cases} 1 & \text{se } n = 0 \\ n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 & \text{se } n \in \mathbb{N},\ n \neq 0 \end{cases}$$

Cenni di calcolo combinatorio: le disposizioni (I)

Disposizioni semplici:

Sono le possibili scelte di ***k* elementi *ordinati* [conta l'ordine]** da un insieme composto da *n* oggetti (disposizioni di *n* elementi di classe *k*).

Ad esempio, quali sono i possibili podi (1,2,3) in una gara tra 8 atleti?

Al primo posto possono esserci 8 persone diverse, al secondo posto ce ne possono essere 7 e al terzo 6; le disposizioni possibili possono essere 8 x 7 x 6 = 336.

In generale:

$$Disposizioni = \frac{n!}{(n-k)!}$$

$$\frac{8!}{(8-3)!} = \frac{8*7*6*5*4*3*2*1}{5*4*3*2*1} = 8*7*6 = 336$$

Le disposizioni coincidono con le permutazioni se $n = k$

Nota bene: **0!=1**

Cenni di calcolo combinatorio: disposizioni (II)

Disposizioni con ripetizione:

Le disposizioni **con ripetizione** (disposizioni con ripetizione di *n* elementi ordinati di classe *k*) sono come le disposizioni, ma ogni oggetto, dopo essere stato scelto **viene rimesso** nell'insieme di partenza.

Date 10 lettere (da A a L), quante *disposizioni con ripetizione* da 4 lettere posso effettuare?

Per la prima lettera ho dieci possibilità, altrettante per la seconda e così via, cioè:

$$10 \times 10 \times 10 \times 10 = 10^4 = 10.000$$

In generale, le disposizioni con ripetizione di *n* elementi di classe *k* sono: n^k

NOTA: nelle disposizioni semplici abbiamo sempre k elementi distinti (k <=n), mentre nelle disposizioni con ripetizione lo stesso elemento può essere ripetuto fino a k volte.

Cenni di calcolo combinatorio: le combinazioni

Nelle combinazioni (di *n* elementi di classe *k*, con *k* <= *n*) **non viene considerato l'ordine** con cui gli oggetti si presentano.

In quanti modi si possono selezionare k oggetti da un insieme di n quando non si tiene conto dell'ordine di estrazione?

$$combinazioni = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

Coefficiente binomiale

e per convenzione si ha che $\binom{n}{0} = 1$ e che vale la simmetria

$$\binom{n}{k} = \binom{n}{n-k}$$

Coefficiente binomiale

Il coefficiente binomiale rappresenta **il numero di sottoinsiemi** di *k* elementi che si possono estrarre da un insieme di *n* elementi.

Date le 10 (*n*) lettere da A ad L, le combinazioni a gruppi di 4 (*k*) che posso avere sono:

$$\frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} = \frac{10*9*8*7*6*5*4*3*2*1}{6*5*4*3*2*1*4*3*2*1} = 10*3*7 = 210$$

In un trial clinico di tipo caso-controllo, su 10 controlli *eleggibili* se ne devono scegliere 5. Si hanno:

$$\binom{10}{5} = \frac{10 \times 9 \times 8 \times 7 \times 6}{5!} = 252$$

252 possibili campioni di 5 controlli.

OPZIONALE: il concetto di *valore atteso (variabile casuale discreta) dal punto di vista bayesiano*

Nell'impostazione **bayesiana** la probabilità P(A) di un evento A viene definita come il prezzo «equo» da pagare per ricevere 1 se A si verifica, e 0 se A non si verifica.

Si può scommettere un importo *x* pagando *x**P(A) per ricevere *x* se A si verifica.

Siano: $\{A_r; r = 1,2, \dots, n\}$ n eventi (mutualmente esclusivi)

Siano: $p_r = P(A_r)$ le probabilità di tali eventi.

Se su ciascun evento A_r scommettiamo l'importo x_r paghiamo complessivamente per le *n* scommesse la somma: $\sum_{i=1}^{r} x_r * p_r$

In cambio riceviamo un importo ***variabile*** a seconda dell'evento che si verifica, cioè una variabile aleatoria X che ha distribuzione:

$$\{x_r, p_r; r = 1,2, \dots, n\}$$

Il numero: $\sum_{i=1}^{r} x_r * p_r$ viene chiamato media o speranza matematica o valore atteso, E(X), ed è quindi il ***prezzo «equo»**** da pagare per ricevere l'importo aleatorio X.

*equo: ***dipende*** dall'atteggiamento dell'individuo di fronte al rischio