Elementi di probabilità

Argomenti:

- Introduzione
- Elementi di teoria delle probabilità
  - Variabili deterministiche e casuali
  - Istogrammi
  - Funzione Distribuzione di probabilità
  - Funzione di Distribuzione cumulativa
  - Modello Gaussiano

Introduzione

Caratteristiche di ***casualità***, cioè di **assoluta imprevedibilità**, possono essere osservate praticamente in qualsiasi esperimento

Ripetizioni di misure di una grandezza possono fornire **risultati leggermente diversi,** anche quando immediate

Questa "*casualità*" è dovuta a **variabili ambientali non controllate o non controllabili**, oppure a una mancanza di precisione nel processo di misura (sempre riconducibili ad elementi non controllati/controllabili)

In alcune attività la casualità (meglio dire le variabili non controllate) sono dominanti e diventa difficile individuare con la semplice ispezione dei dati un andamento

Probabilità e Statistica ci aiuteranno ad individuare questo andamento sommerso da dati confusi

La teoria delle probabilità ci permetterà di costruire modelli dei fenomeni casuali utili anche al di fuori della sperimentazione

Introduzione

Cenni sugli elementi fondamentali necessari per la gestione del problema

E' necessario risalire dalle informazioni relative ad un campione alle caratteristiche della popolazione della quale il campione fa parte

Problemi:

- Grandezza priva di modello matematico atto alla previsione
- Ho eseguito un certo numero di misure, quale è la possibilità che ripetendo la misura ottenga valori compresi tra due estremi assegnati?
- Ho caricato un provino al disotto dello sforzo nominale di snervamento ma si è snervato; come mai?
- …

Richiami

Alcuni elementi interessanti dal punto di vista sperimentale sono un patrimonio relativamente assodato; per esempio le calcolatrici tascabili sono in grado di effettuare i calcolo di alcune grandezze statistiche di comune impiego (media, varianza, regressione lineare e coefficiente di correlazione).

Cerchiamo di razionalizzare il loro impiego

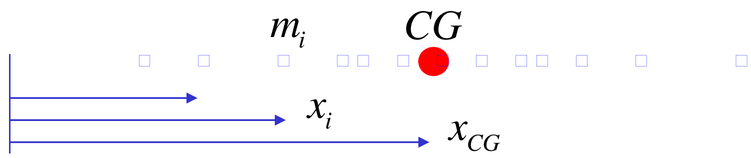
Media Date N osservazioni x_i la media vale:

$$x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Dal punto di vista intuitivo appare più vantaggioso utilizzare la media su un certo numero di misure, piuttosto che una singola misura.

La media è considerata **la miglior stima del valore vero** in assenza di altre informazioni; vedremo in quali condizioni ciò rappresenta una corretta ipotesi.

Possiamo vedere il calcolo della media di una serie di valori come il calcolo della posizione del baricentro di un corpo monodimensionale costituito da particelle, tutte di massa uguale, posizionate in corrispondenza di ogni dato:



$$x_{CG} = \sum_{i=1}^{N} x_i m_i / \sum_{i=1}^{N} m_i = \left(m \sum_{i=1}^{N} x_i \right) / \left(m \sum_{i=1}^{N} 1 \right) = \sum_{i=1}^{N} x_i / N$$

Coordinata CG come rapporto tra i momenti statici di ordine 1 e di ordine 0 delle masse

Per quanto riguarda il problema delle misure è importante valutare quanto le singole misure si discostino dalla media, cioè la loro **dispersione**

Possiamo definire la **Deviazione** (visto come massimo errore di precisione) come differenza di una misura con il valore medio

$$d_i = x_i - \overline{x}$$

Non rappresenta l’errore sulla singola misura, dato che il valore medio **NON** è il valore medio reale, ma solo la sua migliore stima.

La **media delle deviazioni** non è una opzione praticabile :

$$\frac{1}{N}\sum_{i=1}^{N} d_i = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\overline{x}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i - \overline{x} = 0$$

Risultato intuitivo, dato che i residui saranno sia positivi che negativi.

Possiamo introdurre altri due indici di dispersione:

Deviazione media (mean deviation):

la media della somma delle deviazioni in valore assoluto

$$\alpha = \frac{1}{N}\sum_{1}^{N}{}_i |d_i| = \frac{1}{N}\sum_{1}^{N}{}_i |x_i - \bar{x}|$$

Stima eccessivamente pessimistica

E' universalmente adottata la **Deviazione standard delle misure (root mean square deviation):** definita come radice quadrata della media degli scarti quadratici (varianza):

$$S = \sqrt{\frac{1}{N}\sum_{1}^{N}{}_i (x_i - \bar{x})^2}$$

S^2 = varianza

Ha la stessa dimensione fisica della misura

(meglio dividere per N-1, soprattutto per poche misure, se le misure sono numerose la differenza è trascurabile: tra 10 e 11 la differenza è meno del 5% tra 20 e 21 scende a poco più di 2%)

Interpretazione fisica in termini di caratteristiche di inerzia: consideriamo il momento di inerzia baricentrico del sistema di masse concentrate tutte uguali

$$I_{CG} = \sum_{i=1}^{N} (x_i - x_{CG})^2 m_i$$

La varianza è il momento di inerzia baricentrico rapportato alla massa totale

$$I_{CG} / M = \sum_{i=1}^{N} (x_i - x_{CG})^2 m_i / \sum_{i=1}^{N} m_i =$$

$$\left(m\sum_{i=1}^{N} (x_i - x_{CG})^2 \right) / \left(m\sum_{i=1}^{N} 1 \right) = \frac{1}{N}\sum_{i=1}^{N} (x_i - x_{CG})^2 = S^2$$

A parità di massa un momento di inerzia maggiore indica una dimensione maggiore del corpo

Nelle misure una varianza maggiore è sintomo, a parità di valore medio, di una maggiore dispersione dei dati

La deviazione standard delle misure è una valida indicazione di quanto disperse sono le misure rispetto al valore medio

$$\text{Misura} = \overline{x} \pm S_x$$

Viene utilizzato per descrivere la dispersione di una grandezza (in relazione con l'incertezza per le misure) tipicamente con l'associazione di una valutazione del livello di confidenza all'intervallo di variabilità che definisce.

Questa definizione ammette un significato probabilistico, quello di avere il 68% circa di probabilità che una successiva misura rientri nell'intervallo così definito

MA SOLO SE la grandezza di riferimento possiede alcune caratteristiche.

In generale (in assenza di errori sistematici) più misure vengono effettuate più accurata è la stima del valor medio.

Ma non è sufficiente effettuare più misure per essere certi di migliorare la qualità del dato: si riducono solo gli errori casuali

Occorre capire se e quando è conveniente aumentare il numero di misure piuttosto che intervenire su altri elementi della procedura di misura

Media e deviazione standard sono valori dipendenti dai dati misurati: diverse serie di misure daranno diversi valori medi e diverse deviazioni standard.

I motivi di "differenze" non sono prevedibili: necessità di tecniche oggettive di analisi

Cominciamo da qualche elemento di ***probabilità***

Variabili Deterministiche e Casuali

Variabili Deterministiche (prevedibili) e Casuali (non prevedibili).

Una grandezza il cui valore è prevedibile e ripetibile, quindi descrivibile mediante modelli deterministici, viene classificata come variabile Deterministica

Una grandezza che non rientri in questa categoria viene classificata come Casuale

Le grandezze misurate sono considerate come *variabili casuali* (*random*).

Una variabile casuale può essere continua (es. La vita di una lampadina o la misura della velocità) o discreta (es. i risultati di un lancio di dadi o l'esito positivo/negativo di un controllo di qualità

Come costruire le informazioni

Una variabile casuale non è prevedibile. Esempi tipici: intensità della turbolenza atmosferica, altezza delle onde

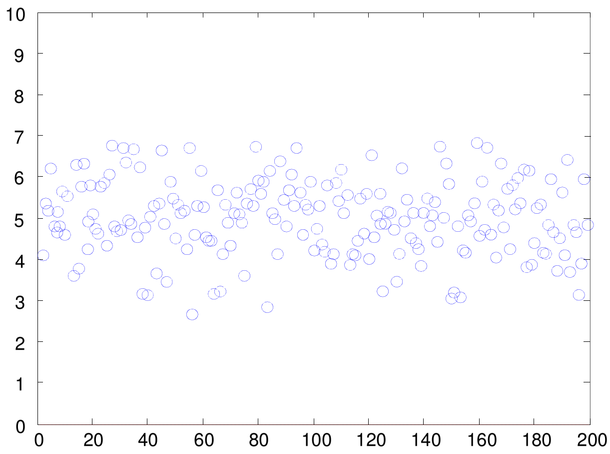
Ammettiamo di sapere come determinare le forze su di una struttura off-shore a seguito di ondate di altezza nota; come prevedere l'altezza in maniera ragionevole?

Ci servono modelli che forniscano dati "ragionevoli", cioè "probabili" anche se NON il valore specifico di un evento

"Misurare" e "Elaborare"

Supponiamo di costruire un sistema in grado di misurare l'altezza delle onde nel sito di installazione di una piattaforma con una assegnata cadenza temprale

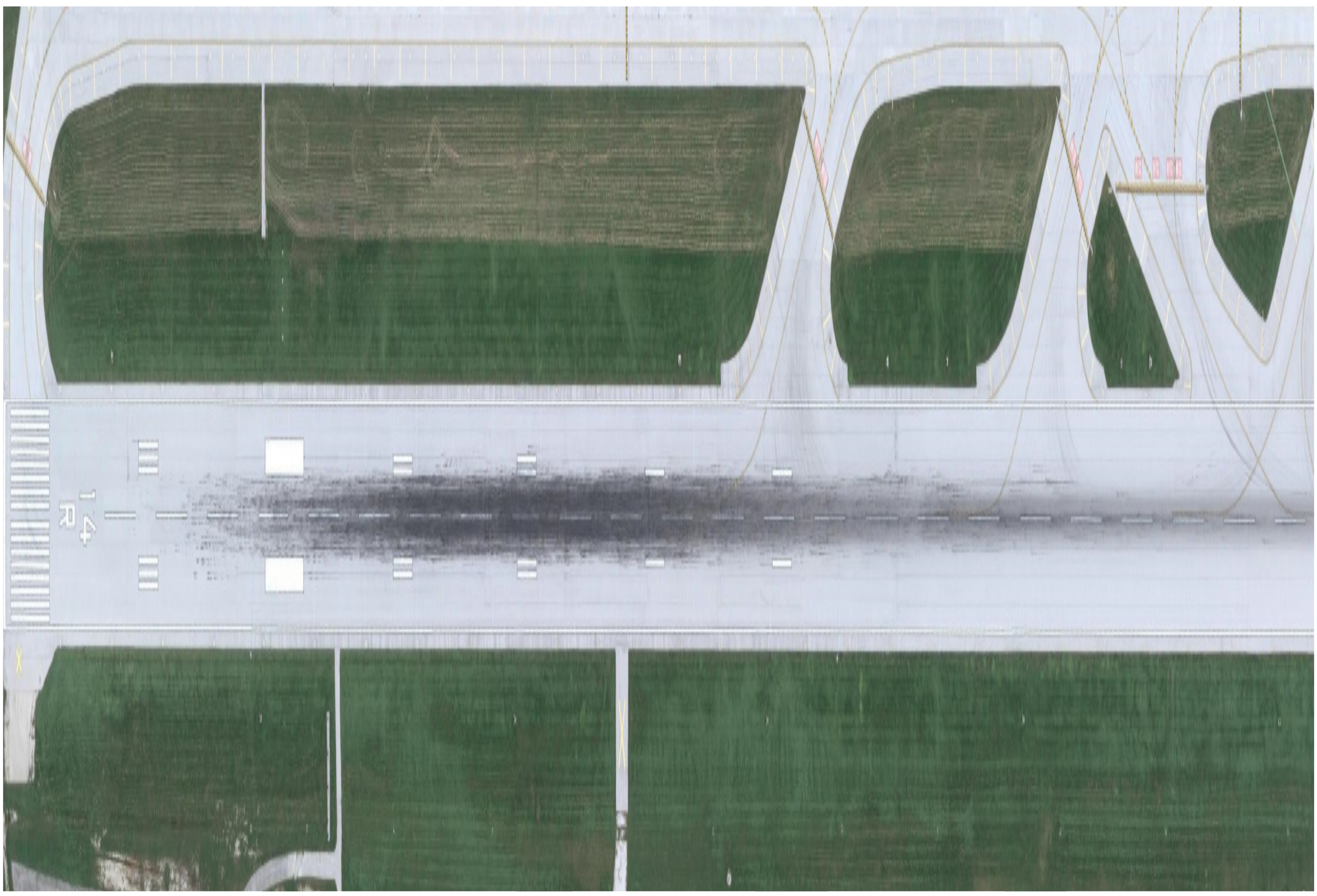
Al termine di un certo periodo di osservazione avremo qualcosa del genere:



Il diagramma mostra come i dati sono distribuiti nel tempo

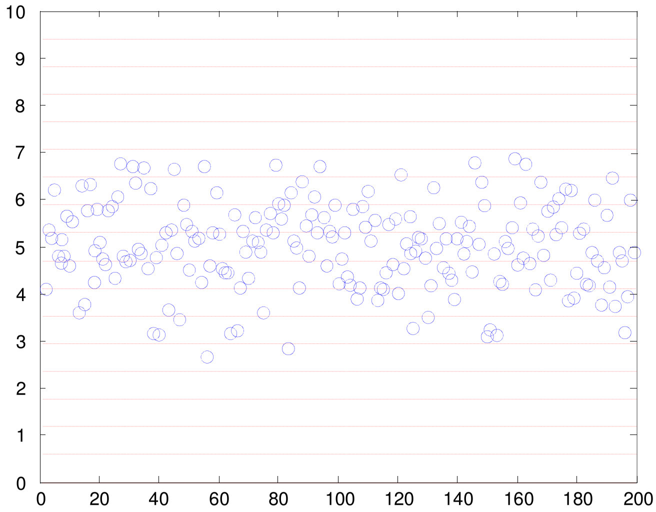
Ci fornisce indicazioni realmente utili? Evidentemente no se non c'è una relazione con il tempo (frequenza degli eventi...)

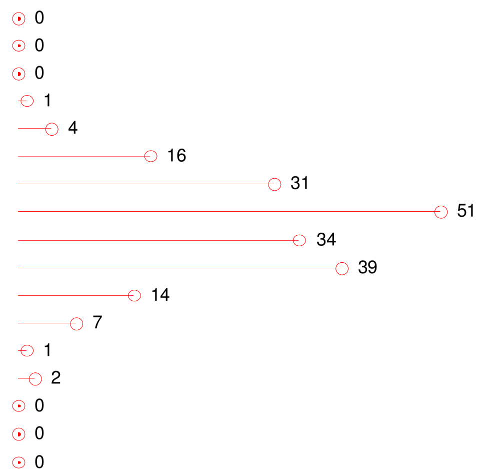
Come costruire le informazioni



Supponiamo di costruire uno strumento che conti quante volte vengono raggiunti assegnati valori di altezza dell’onda

Al termine di un periodo di osservazione avremo il grafico di destra anziché quello di sinistra





Il diagramma, un “Istogramma”, mostra come i dati sono distribuiti sul campo di misura e **indica le zone più “probabili”**

Gli istogrammi

Per costruire un istogramma:

• si osserva il fenomeno acquisendo l’informazione numerica

• si divide il campo di osservazione in intervalli

• per ognuno si contano le occorrenze/eventi (per segnali continui si misura il tempo di permanenza del segnale nell’intervallo)

• sopra ogni intervallo si traccia un segmento di lunghezza pari al numero di occorrenze (se si utilizza un rettangolo si deve ricordare che la base non ha significato, per es. non è pari all’intervallo)

Per la determinazione del numero di intervalli si può utilizzare la formula (N>40; N < 40 istogramma non significativo)

$$N_{Bins} = 1.87(N-1)^{0.40} + 1$$

Il numero di eventi in un intervallo deve essere significativo (≥ 5) altrimenti la distribuzione risulterebbe dipendente dai dati disponibili

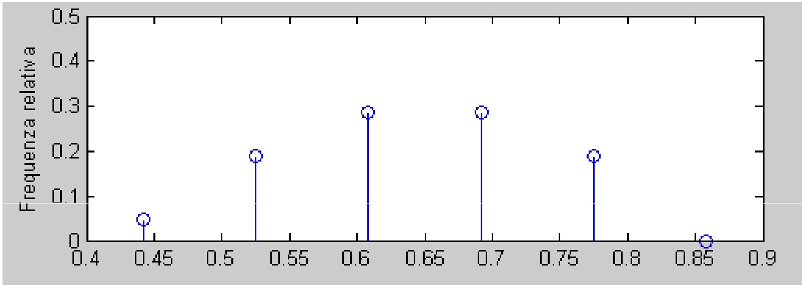
In caso contrario è meglio diminuire il numero di intervalli

Gli istogrammi

Il numero di eventi non è significativo

Normalizzando il numero di occorrenze con il numero di misure totali si esprimono i dati in termini di *frequenza relativa*

$$p_i = \frac{n_i}{N}$$



si hanno n_i possibilità su N che un nuovo dato cada nell'intervallo i-esimo; quindi c'è corrispondenza tra l'istogramma e la probabilità di effettuare una particolare osservazione o che si verifichi un preciso evento

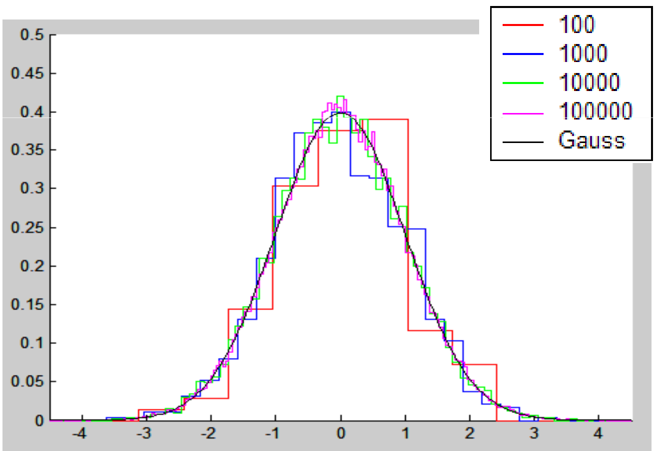
La somma delle frequenze relative fornisce la probabilità per un intervallo più ampio:

$$P(x_i \le x \le x_{j+1}) = \sum_{k=i}^{j} \frac{n_k}{N}$$

Gli istogrammi

All'aumentare del numero di dati è ragionevole aumentare il numero di intervalli risolvendo meglio il campo di variabilità

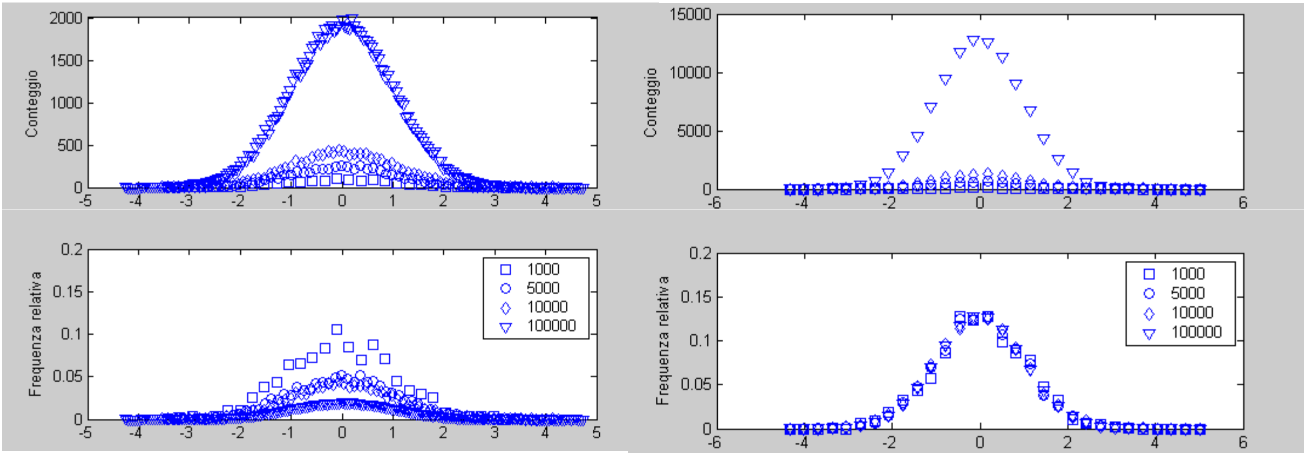
La rappresentazione migliora tendendo ad una funzione apparentemente continua



Rimane l'arbitrarietà legata al numero di valori e di intervalli

Gli istogrammi

Il risultato infatti dipende dai dati (a meno di mantenere fisso il numero di intervalli, cosa che vanifica l'aumento del numero di dati)



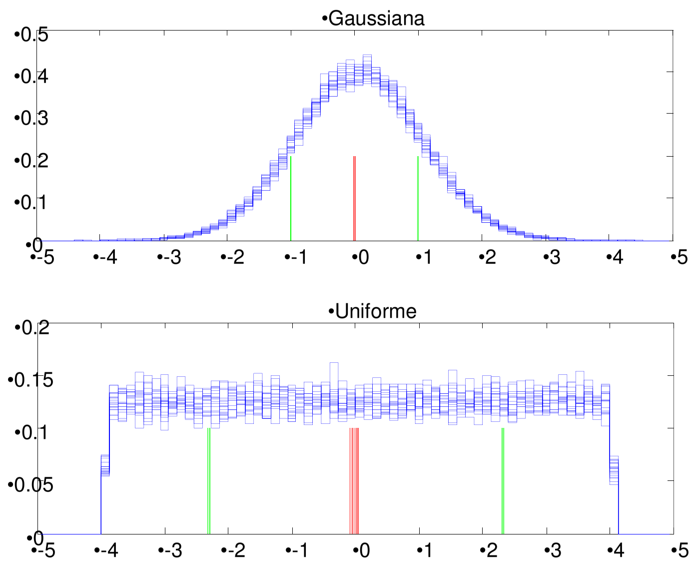
N intervalli variabile *N intervalli fisso*

Necessario passare ad un modello "continuo"

Diverse serie di dati relativi allo stesso processo daranno distribuzioni di probabilità leggermente diversi

20 serie di 10000 dati

Per ciascuna riportati anche valore medio e deviazione standard



Osservazione: Media e DevStd della serie sono molto STABILI

Una funzione di questi dati consentirebbe di costruire un modello matematico probabilistico della grandezza casuale

Dall'istogramma alla Densità di probabilità

Si ottiene una funzione indipendente dalla struttura delle osservazioni normalizzando la frequenza relativa rispetto all'ampiezza degli intervalli: ciò equivale a definire una **funzione** continua, costante nell'intervallo, a parità di integrale (probabilità)

Uguaglianza delle probabilità: $$P(x_i \le x \le x_{i+1}) = \frac{n_i}{N} = \int_{x_i}^{x_{i+1}} p(x)dx$$

Se Densità costante **a tratti:** $$\int_{x_i}^{x_{i+1}} p(x)dx = p(x_i)\int_{x_i}^{x_{i+1}} dx = p_i \Delta x$$

Valore della densità dell'intervallo: $$p_i = \frac{n_i}{N} / \Delta x$$

I risultati di osservazioni diverse sono comunque **confrontabili**, indipendentemente dal numero di eventi osservati e/o dal numero e dall'ampiezza degli intervalli

La funzione $p(x)$ ha il significato di **Densità di Probabilità** e **caratterizza** il fenomeno casuale

Utilizzando gli istogrammi la probabilità è data dalla sommatoria: $$P(x_i \le x \le x_j) = \sum_{k=i}^{j} \frac{n_k}{N}$$

Utilizzando la funzione di densità di probabilità si dovrà integrare: $$P(x_i \le x \le x_j) = \int_{x_i}^{x_j} p(x)dx$$

La probabilità che si faccia una misura che appartenga ad un intervallo è data dall'integrale della funzione densità di probabilità esteso a quell'intervallo: $$P(a \le x \le b) = \int_a^b p(x)dx$$

In particolare la probabilità di un valore assegnato è nulla: $$P(x = a) = \int_a^a p(x)dx = 0$$

L'integrale della funzione densità su tutto il dominio di misura è unitario: la probabilità che una misura vi rientri deve infatti essere 100%. $$\int_{-\infty}^{\infty} p(x)dx = P(-\infty < x < \infty) = 1$$

Definizione di probabilità: caso discreto

Probabilità (Probability). E' un valore numerico che esprime la stima della possibilità di occorrenza di un determinato evento, tra tutti quelli possibili espressi da un numero finito

Quindi è il numero che esprime la verosimiglianza dell'occorrenza di un evento relativamente a tutte le possibilità dello spazio relativo.

La probabilità (*A*) di un evento **discreto** è ottenuta dividendo il numero di evenienze del tipo di interesse (*m*), o esiti positivi, per il numero totale di soluzioni possibili (*n*), cioè dei possibili esiti.

Probabilità dell'evento *A*=m/n

Es. c'è il 50% di probabilità che il lancio di una moneta dia testa.

Es. se lanciamo due dadi la probabilità di ottenere un doppio 1 è 1/36 in quanto c'è una sola possibilità dell'esito voluto a fronte di 36 possibili eventi (6 indipendenti per ciascun dado)

Definizione di probabilità: caso continuo

Probabilità (Probability). E' un valore numerico che esprime la stima della possibilità di occorrenza di un valore della funzione compreso tra due estremi (l'evento è quindi espresso in relazione ad un intervallo)

Definizione di evento $(a \le x \le b)$

Quindi è il numero che esprime la verosimiglianza dell'occorrenza di un evento relativamente a tutte le possibilità dello spazio relativo (±∞) cioè ad un valore unitario

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

La probabilità di una grandezza casuale continua è ottenuta integrando la sua densità di probabilità sull'intervallo di interesse:

Probabilità dell'evento $P(a \le x \le b) = \int_a^b p(x)dx$

Dalla Densità di probabilità a parametri sintetici

Attraverso la sua **funzione di densità di probabilità** possiamo **prevedere i valori di una grandezza**, non soltanto la probabilità che un certo evento avvenga

L'integrale della funzione di densità di probabilità $p(x)$ della grandezza x pesata per la grandezza stessa ne fornisce il ***valore atteso*** (*expected value*) (in alcuni casi coincide con il valore più probabile)

Momento di ordine 1 $E(x) = \mu = \int_{-\infty}^{\infty} x\ p(x)\ dx$

Analogamente la sua varianza viene ottenuta integrando la densità di probabilità pesata per il quadrato della distanza tra la variabile x e il suo valore atteso:

Momento di ordine 2 $V = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\ p(x)\ dx$

$$\boxed{\bar{x} = \frac{1}{N}\sum_i x_i = \sum_i \frac{x_i}{N} = \frac{n_1 x_1}{N} + \frac{n_2 x_2}{N} + \cdots = \sum_j \frac{n_j}{N} x_j = \sum_j f_j x_j \rightarrow \int_x p(x) x\, dx}$$

Definizione di Variabile Casuale

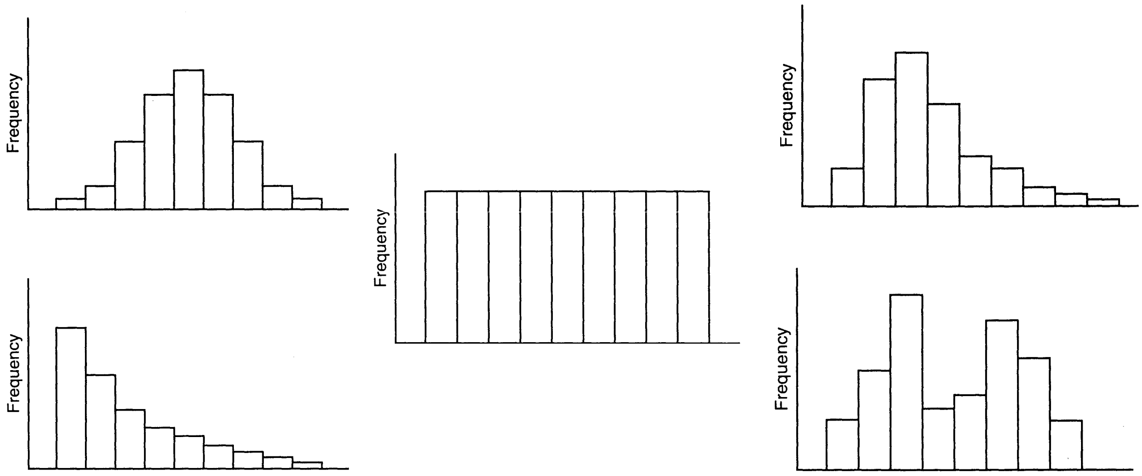
Si è così arrivati alla definizione di ***Variabile casuale (Random variable)***: un parametro il cui evento non è descrivibile mediante modelli deterministici, quindi non direttamente prevedibile; un parametro che può essere previsto solo in forma statistica attraverso una ***distribuzione di probabilità***.

Attraverso la ***distribuzione di probabilità*** è possibile risalire ad alcuni parametri sintetici che caratterizzano la variabile casuale

Le grandezze misurate sono considerate come *variabili casuali* (*random*).

Matematicamente, una variabile random è una funzione valutata numericamente, mediante le tecniche probabilistiche, per una data popolazione elaborando le informazioni ottenute con una collezione di valutazioni (campione)

La forma della distribuzione della densità di probabilità è una caratteristica del fenomeno.



E’ stato osservato che il comportamento di alcune classi di problemi è abbastanza bene descritto ad alcune leggi

Distribuzioni di interesse ingegneristico e applicazioni tipiche

Binomia (Binomial): distribuzione discreta. Utilizzata nel controllo di qualità (scarto di prodotti difettosi), nelle analisi successo/ insuccesso buono/cattivo.

Normale (Normal) o Gaussiana: continua e simmetrica; largamente impiegata nell’analisi sperimentale e nella fisica. Utilizzata per la spiegazione del comportamento di variabili casuali nella sperimentazione.

Student's *t:* continua e simmetrica; utilizzata per l’analisi della varianza quando il numero di campioni è limitato (<30). Per un numero di campioni superiore si comporta come la distribuzione normale.

χ^2: continua, nonsimmetrica; usata per l’analisi della varianza di campioni di popolazione e per valutare la qualità del fitting di una distribuzione

Weibul: continua, non simmetrica; usata per descrivere i fenomeni di durata di parti e componenti.

Esponenziale: continua, non simmetrica; usata per l'analisi affidabilistica di sistemi completi e assemblaggi di parti.

Lognormal: continua, non simmetrica; utilizzata per lo studio della vita e della durata di parti e componenti.

Uniform: continua, simmetrica; usata per la stima delle probabilità di variabili random per la simulazione.

Poisson: discreta; utilizzata quando le occorrenze degli eventi sono osservabili ma non lo sono le non occorrenze (es i guasti in un impianto o i black-out in una zona)

Riassumendo

Calcolo di probabilità

$$P(x \leq a) = \int_{-\infty}^{a} p(x)dx$$

$$P(x > a) = 1 - P(x \leq a)$$

$$P(x = a) = 0$$

Calcolo del valore atteso

$$E(x) = \mu = \int_{-\infty}^{\infty} x\ p(x)\ dx$$

Calcolo della varianza prevista

$$V = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2\ p(x)\ dx$$

Proprietà della probabilità

L'evento viene rappresentato da una *variabile casuale continua* (x) e la sua probabilità rappresentata da una funzione $P(x)$. Per variabili casuali discrete Xi, la probabilità verrà data da $P(Xi)$.

La previsione della probabilità di un evento è uno degli scopi dell'analisi statistica.

Esaminiamo le proprietà più significative relative alla probabilità:

1. La Probabilità è sempre positiva con valore massimo unitario:

$$0 \leq P(x \text{ or } Xi) \leq 1.$$

2. Se l'occorrenza di un evento, A, è certa:

$$P(A) = 1.$$

3. Se è certo che un evento, A, non avviene allora:

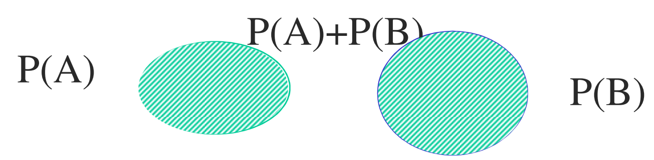
$$P(A) = 0.$$

4. Se esiste un evento A^* complementare all'evento A (cioè se avviene l'evento A, l'evento A^* non avviene):

$$P(A)=1-P(A^*)$$

5. Se due eventi A e B sono mutualmente esclusivi (cioè se la probabilità di evenienza simultanea è nulla), la probabilità di occorrenza dell'evento A o B è:

$$P(A \text{ or } B)=P(A)+P(B)$$

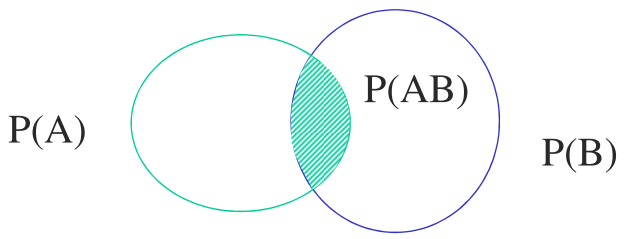


Per esempio, nel caso di un lancio di dadi la possibilità di ottenere 3 o 6 è

$$P(3 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

6. Se gli eventi *A* e *B* sono indipendenti (cioè se il loro accadimento non dipende uno dall'altro), la probabilità che entrambi avvengano è

$$P(AB)=P(A)P(B)$$

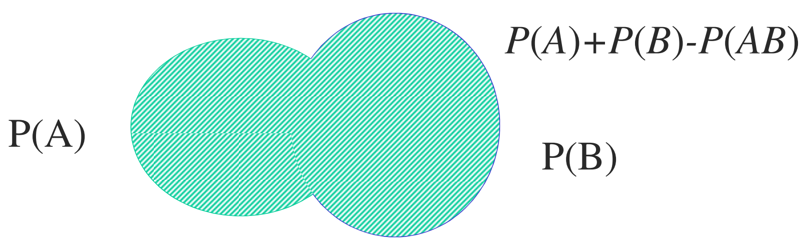


Es. Consideriamo l'assemblaggio di due pezzi, *A* e *B,* prodotti da differenti aziende. Sappiamo che c'è la possibilità del 5% che il pezzo *A* sia difettoso contro il 2% del pezzo *B*. La probabilità che il pezzo assemblato contenga sia *A* che *B difettosi* è:

$$P(AB)=0.05 \times 0.02 = 0.001 \text{ o } 0.1\%$$

7. La probabilità di evenienza di *A* o *B* o entrambi, rappresentata da $P(A \cup B)$ ("probabilità dell'unione di *A* e *B"),* è:

$$P(A \cup B)=P(A)+P(B)-P(AB)$$



La probabilità è quindi inferiore alla somma delle due singole probabilità in quanto si deve tenere conto della probabilità che un pezzo presenti entrambi i difetti

Nell'es. precedente, la probabilità di avere *A* o *B* o entrambi difettosi è:

$$P(A \cup B)=0.05 + 0.02 - 0.001 = 0.069 \text{ cioè } 6.9\%$$

Definizioni

E' utile introdurre qualche definizione:

Popolazione (Population). La *popolazione* è l'insieme che contiene tutti gli oggetti, misure, osservazioni, ecc.., analizzate, per le quali possono essere espresse alcune generalizzazioni.

Esempi: tutte le lampadine di un lotto di produzione
tutti gli abitanti di una città
tutte le misure della velocità del vento di un anno in un sito

Se consideriamo una serie di misurazioni della stessa grandezza cosa si può dire? Si può sostenere che la *popolazione* sia costituita da tutte le misure disponibili, ma avremmo potuto farne altre, o meno, allora quale sarebbe la *popolazione*?

Campione (Sample). Un *campione* è un sottoinsieme rappresentativo di una popolazione per la quale sono stati ottenuti dati numerici, per esempio con misure.

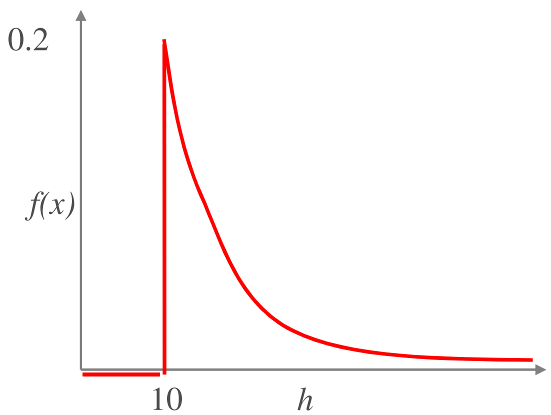
Es.: 10 lampadine selezionate in un lotto di produzione di 10000 oppure la velocità del vento può essere misurata una volta all'ora per 24 ore.

Evento (Event). Un *evento* è un possibile esito di un esperimento su di una variabile casuale. Es. la misura di una grandezza.

Parametro (Parameter). Un *parametro* è un attributo numerico utilizzato per caratterizzare **l'intera popolazione**. Per esempio il valore medio della vita delle lampadine di un intero lotto di produzione.

Statistica (Statistic). Una *statistica* è un attributo numerico di un **campione**. Per esempio il valor medio della vita delle lampadine di un campione di un lotto di produzione

Esempio: Si consideri un cuscinetto la cui vita è esprimibile mediante la seguente legge di densità di distribuzione di probabilità di rottura:



$$f(x) = \begin{cases} 0 & x < 10h \\ \dfrac{200}{x^3} & x \geq 10h \end{cases}$$

Se preleviamo un cuscinetto a caso da un lotto di produzione, qual è la probabilità che la sua vita sia, rispettivamente: inferiore a 20h, maggiore di 20h, uguale a 20h ?

Soluzione: $$P(x<20)=\int_{-\infty}^{20} f(x)dx=\int_{-\infty}^{10} 0dx+\int_{10}^{20}\frac{200}{x^3}dx=0.75$$

$$P(x\ge 20)=1-P(x\le 20)=0.25$$

$$P(x=20)=0$$

Ma allora cosa ho misurato se la vita è risultata essere 20 h?

$$P(x=20) \quad \Rightarrow \quad P(20-\delta\le x\le 20+\delta)$$

$$P(20-\delta\le x\le 20+\delta)=\int_{20-\delta}^{20+\delta}\frac{200}{x^3}dx=-\left.\frac{100}{x^2}\right|_{20-\delta}^{20+\delta}$$

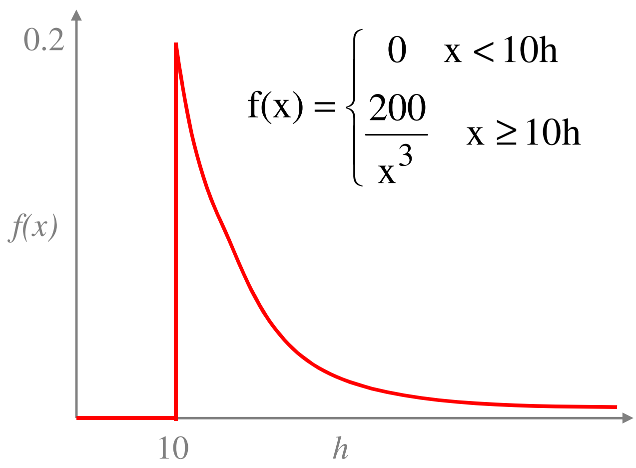
Se δ è pari ad un minuto, allora: $P(x\simeq 20)=0.0833\%$

Esempio: Per il cuscinetto dell'esempio precedente è possibile calcolare la vita attesa? Ci interessa cioè quel valore che dovremmo ottenere come media di un rilevamento sperimentale

Soluzione:

$$E(x)=\mu=\int_{0}^{+\infty} x\,f(x)dx=$$

$$\int_{10}^{\infty} x\frac{200}{x^3}dx=-\left.\frac{200}{x}\right|_{10}^{\infty}=20h$$



$$f(x)=\begin{cases}0 & x<10h\\ \dfrac{200}{x^3} & x\ge 10h\end{cases}$$

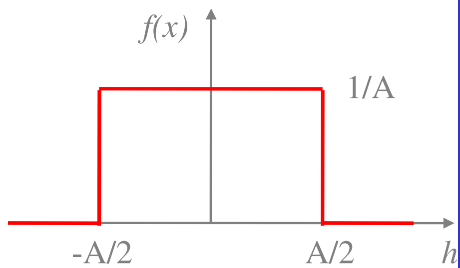
Esempio: Consideriamo il caso di una variabile a distribuzione di densità di probabilità costante, di valore 1/A, nell'intervallo ±A/2.

Quali sono il valore atteso e la deviazione della variabile?

Soluzione:

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx = \int_{-A/2}^{A/2} x \frac{1}{A} dx = \left. \frac{x^2}{2A} \right|_{-A/2}^{A/2} = 0$$

$$f(x) = \begin{cases} 0 & x < A/2 \\ 1/A & -A/2 \le x \le A/2 \\ 0 & x > A/2 \end{cases}$$



$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = \int_{-A/2}^{A/2} (x)^2 \frac{1}{A} dx =$$

$$\left. \frac{x^3}{3A} \right|_{-A/2}^{A/2} = \frac{1}{12} A^2 = \frac{1}{3}\left(\frac{A}{2}\right)^2 \quad \Rightarrow \quad \sigma = \frac{1}{\sqrt{3}} \frac{A}{2}$$

Funzione di distribuzione cumulativa

La **Funzione di Distribuzione Cumulativa** (Cumulative Distribution Function) permette di valutare la probabilità che una variabile *v* descritta dalla Distribuzione di Probabilità *f*(*v*) assuma valori minori od uguali ad un valore assegnato $\bar{x}$ avendo eseguito l'integrazione una volta per tutte in forma analitica

La Funzione di Distribuzione Cumulativa è definita come:

$$F(v \le \bar{x}) = F(\bar{x}) = \int_{-\infty}^{\bar{x}} f(x) dx \quad \equiv \quad P(v \le \bar{x})$$

Nel caso di un istogramma si avrebbe: $F(v \le x_i) = \sum_{j=1}^{i} P(x_j)$

Se $P(v \le a) = F(a)$ si possono dimostrare le seguenti relazioni:

$$P(v > a) = 1 - F(a)$$

$$P(a < x \le b) = F(b) - F(a)$$

Se disponibile non serve effettuare un integrale: è sufficiente la differenza di due valutazioni della funzione

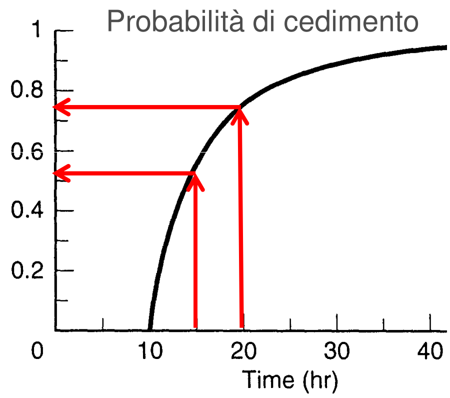
Esempio: Per il lotto di cuscinetti dell'esempio precedente, qual è la probabilità che un cuscinetto abbia una vita inferiore a 15 o 20 ore?

Soluzione: si utilizza la funzione di distribuzione cumulativa

$$F(x) = \int_{-\infty}^{x} f(s)ds = \int_{-\infty}^{x} 0\,ds = 0 \qquad \text{per } x \le 10$$

$$F(x) = \int_{-\infty}^{x} \frac{200}{s^3} ds = 1 - \frac{100}{x^2} \qquad \text{per } x > 10$$

Per 15 ore si ha una probabilità di 0.55, per 20 ore diventa 0.75 (lo stesso risultato ottenuto per P(x<20))



Con questo approccio non occorre effettuare gli integrali ogni volta: è sufficiente calcolare la funzione per il valore di interesse

Esempio: Volendo garantire con l'80% di probabilità la vita di un cuscinetto dello stesso lotto, quanto devo dichiarare?

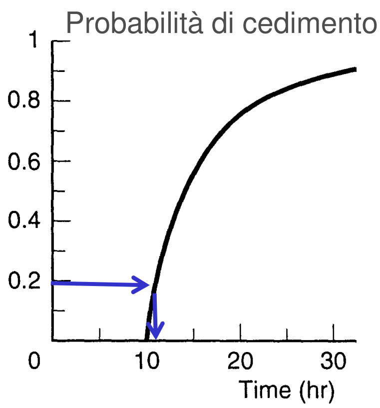
Soluzione: si utilizza ancora la funzione di distribuzione cumulativa

$$F(x) = 0 \qquad \text{per } x \le 10 \qquad\qquad F(x) = 1 - \frac{100}{x^2} \qquad \text{per } x > 10$$

La funzione fornisce la probabilità di cedimento dopo un certo tempo

Per garantire l'80% di probabilità la vita di un cuscinetto devo cercare il tempo per cui se ne sono rotti solo il 20% o complementare la funzione (1-F(x))

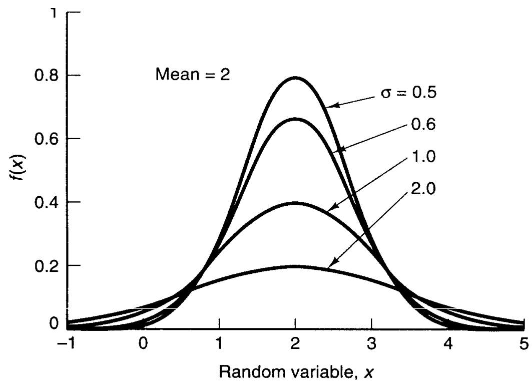
Risolvendo la distribuzione cumulativa per F(x)=20% o il suo complemento 1-F(x)=80% ottengo che dovrò dichiarare 11 ore (11.18)



Funzione di densità di probabilità Gaussiana

Una distribuzione di particolare interesse è quella Gaussiana:

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



μ valore centrale

σ distribuzione attorno al valor medio

Il coefficiente moltiplicativo rende unitario l'integrale

E' simmetrica centrata sul valore medio μ dove assume valore massimo pari a

$$p_{\max} = p(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

A distanza $\pm\sigma$ dalla media la funzione presenta un flesso

Perché ci interessa?

Descrive bene l'andamento di una parte degli errori presenti nelle misure: essa descrive la dispersione di dati sperimentali conseguenza di fattori casuali, quando l'evenienza di deviazioni positive e negative di una certa entità è egualmente probabile con misure vicino al valore medio più probabili.

Gli errori casuali evidenziano una distribuzione abbastanza bene descrivibile con una distribuzione Gaussiana.

Anche se gli errori individuali non seguono questa distribuzione, **le medie di gruppi di queste misure hanno una distribuzione tendente a quella gaussiana** per un numero elevato di gruppi

Si tratta di un modello non fisico: è possibile (anche se poco probabile) un valore estremamente lontano dal valore medio

Funzione di densità di probabilità normale

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

Notando la struttura dell'esponente, è naturale introdurre la variabile z come rapporto tra la distanza dal valor medio e il parametro di distribuzione:

$$z = \frac{x-\mu}{\sigma}$$

Per la probabilità di evenienza di un intervallo si ottiene: $\left[dx = \sigma dz \right]$

$$P(x_0 < x < x_1) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_0}^{x_1} e^{-\frac{(x-\mu)^2}{\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{z_0}^{z_1} e^{-\frac{1}{2}z^2} \sigma dz = \int_{z_0}^{z_1} f(z) dz$$

Avendo definito la funzione *densità di probabilità normale* o *standard* come:

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

La variabile z esprime la distanza di x dal valore medio in multipli non interi del parametro di distribuzione σ

Si può verificare che questo parametro coincide con la deviazione standard della variabile x

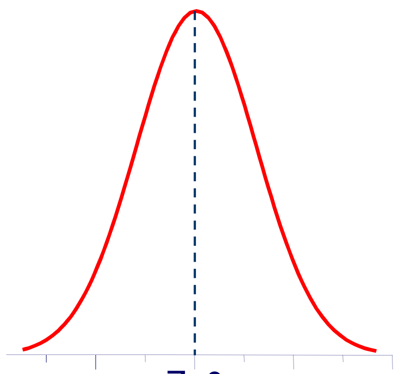
Diremo quindi che la variabile casuale x , a distribuzione Gaussiana, è caratterizzata da due parametri: il valore medio μ e la deviazione standard σ

La funzione di densità normalizzata

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

Ha media nulla (distanza dal valor medio della funzione gaussiana)

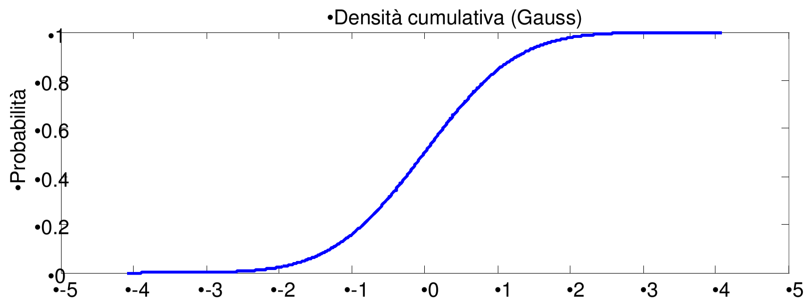
Deviazione standard unitaria



Applicando la definizione di probabilità cumulativa (integrale indefinito):

$$F(z) = \int_{-\infty}^{z} f(z)dz =$$

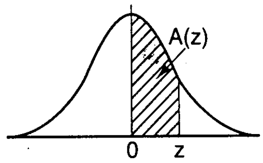
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{z^2}{2}\right) dz$$



Il cambio di variabile risulta utile perché consente di esprimere la probabilità come differenza dei valori che la densità cumulativa assume ai due estremi dell'intervallo di interesse

$$P(x_0 < x < x_1) = \int_{z_0}^{z_1} f(z)dz = F(z_1) - F(z_0)$$

La funzione è disponibile nel semipiano $0 \le z \le \infty$ definendo l'area sottesa dalla curva di densità di probabilità come tratteggiata in figura



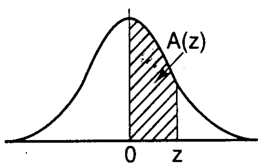
quindi $$P(x_0 < x < x_1) = I\left(\frac{x_1 - \bar{x}}{\sigma_x}\right) - I\left(\frac{x_0 - \bar{x}}{\sigma_x}\right) = I(z_1) - I(z_0)$$

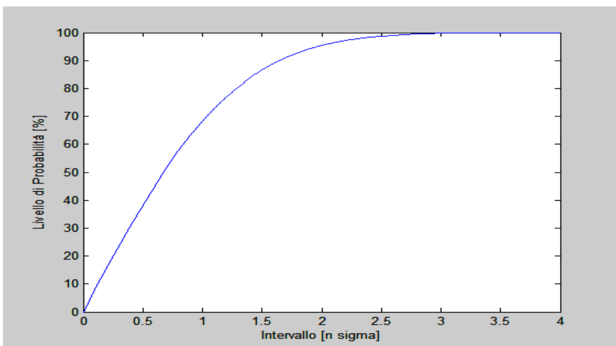
La forma integrale è disponibile in forma tabulare con passo z/100

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07355
0.2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
…										
2.7	49653	49664	49674	49683	[illegible]	[illegible]	49711	49720	49728	49736
2.8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
…										
5.0	50000	50000	50000	50000	50000	50000	50000	50000	50000	50000

z=2.83

Tabella della funzione integrale





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1571
0.4	.1554	.1591	.1628	.1604	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817

Esempio: un sensore ha dato le misure di seguito riportate.

.512	.734	.627	.477	.771	.701	.794
.486	.573	.672	.559	.721	.713	.614
.802	.588	.687	.553	.621	.722	.605

Assumendo una distribuzione normale qual'è la probabilità che una lettura fatta a caso fornisca una misura compresa tra 0.5 e 0.7?

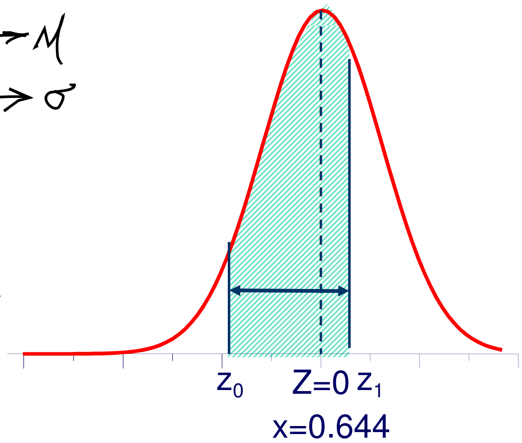
La media delle misure è: 0.644 $\rightarrow \mu$

La deviazione standard è: 0.098 $\rightarrow \sigma$

x_1=0.7-0.644=0.056 z_1=0.57

x_0=0.5-0.644=-0.144 z_0=-1.47

Con questi dati possiamo impostare il calcolo della probabilità



L'espressione della probabilità scritta per i dati specifici è:

$$P(0.5 \le x \le 0.7) = P(z_1 \le z \le z_2) =$$

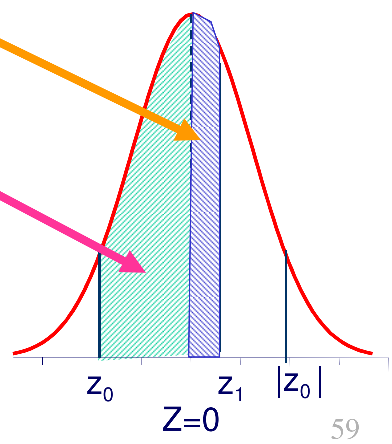
$$P\left(\left\langle \frac{x_1 - \bar{x}}{\sigma_x} \right\rangle \frac{0.5-0.644}{0.098} \le z \le \frac{0.7-0.644}{0.098}\right) = P\left(-1.47 \le z \le 0.57\right) =$$

$$P(-1.47 \le z \le 0) + P(0 \le z \le 0.57) =$$

Dovendo operare a cavallo dello zero si sommano i termini di probabilità per valori positivi e negativi di z

La tabella è scritta per valori di scostamento positivi, ma vale anche per valori negativi essendo la funzione densità di probabilità simmetrica

$$P(-1.47 \le z \le 0) = P(0 \le z \le 1.47)$$



Quindi: $P(0 \le z \le 1.47) + P(0 \le z \le 0.57) =$

$$I(z_1) + I(z_2) = I(1.47) + I(0.57) = 0.4292 + 0.2157 = 0.6449$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
...	...	...	...	...	...	...	...	...	...	...
0.4	.1554	.1591	.1628	.1604	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	z=0.57	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	z=1.47	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
...	...	...	...	...	...	...	...	...	...	...

Quindi la probabilità che si verifichi la condizione $0.5 \le x \le 0.7$ è del 64%.

Esempio 1

Qual è la probabilità che nuove misure, di una grandezza a distribuzione gaussiana, a media 0.644 e deviazione standard 0.098, siano maggiori di 0.8?

$$P(0.8 \le x \le \infty) = I(\infty) - I\left(\frac{0.8 - 0.644}{0.098}\right) =$$

$$I(\infty) - I(1.5918) = 0.5 - 0.44408 = 0.056$$

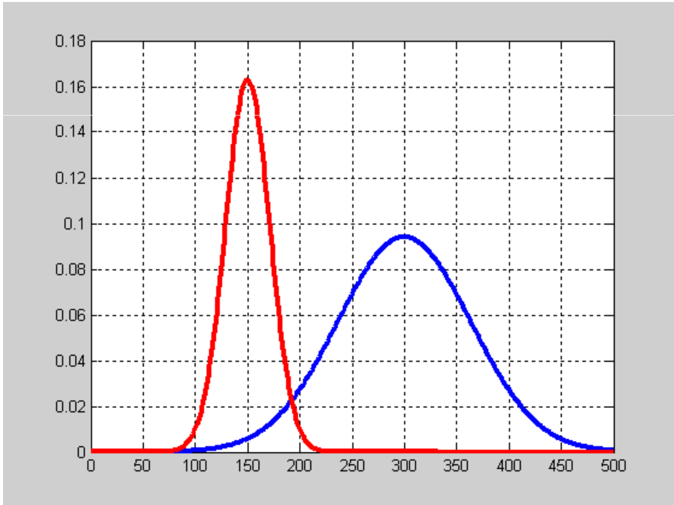
Quindi la probabilità che si verifichi la condizione $x \ge 0.8$ è del 5.6%.

Esempio 2

La sperimentazione su di un lotto di provini dello stesso materiale ha portato a definire lo sforzo di rottura medio e la sua deviazione standard. Se il carico applicato è anch'esso noto in termini di valore medio e di deviazione standard e ipotizzando che entrambi abbiano distribuzione casuale, quale è la probabilità che un provino si rompa?

$$\sigma_{R\_medio} = 300 \pm 45MPa$$

$$\sigma_{A\_medio} = 150 \pm 15MPa$$



Possiamo definire l'indice di sollecitazione come differenza tra lo sforzo di rottura e quello applicato: $I_S = \sigma_{R\_medio} - \sigma_{A\_medio}$

La differenza di due variabili casuali ha come valore medio la differenza dei valori medi e come varianza la somma delle varianze:

$$d_{medio} = x_{1\_medio} - x_{2\_medio}$$

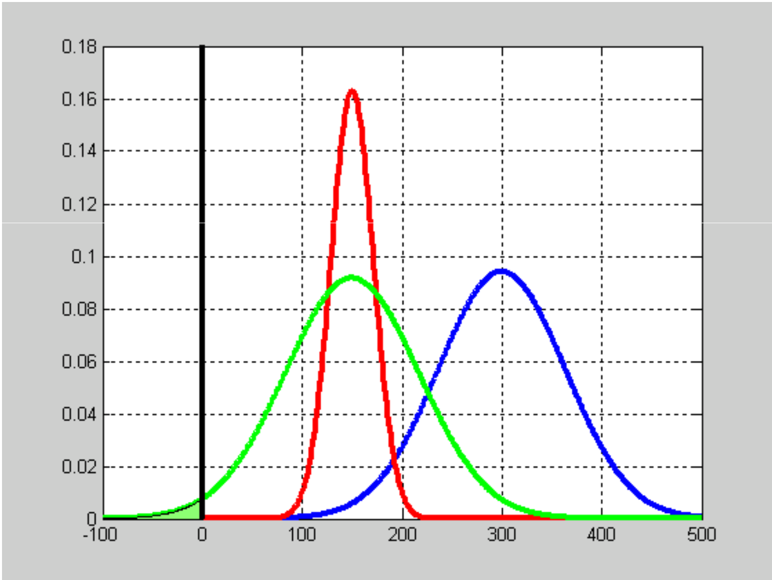
$$S^2_{d\_medio} = S^2_{x1\_medio} + S^2_{x2\_medio}$$

Quindi: $I_S = \sigma_{R\_medio} - \sigma_{A\_medio} = 300 - 150 = 150MPa$

$$S_{Is} = \sqrt{S^2_{\sigma_R} + S^2_{\sigma_A}} = \sqrt{45^2 + 15^2} = 47.43MPa$$

La variabile casuale di interesse, distanza tra lo sforzo di rottura e quello applicato, è completamente definita come: $I_S = 150 \pm 47.43MPa$

La probabilità di rottura si esprime come: probabilità che l'indice di sollecitazione sia minore di zero: $P(I_S < 0)$



Trasformiamo in forma normale: $z = \frac{x - \overline{x}}{\sigma_x} = \frac{I_S - I_{S\_Medio}}{S_{Is}}$

E definiamo gli estremi dall'intervallo nel quale : $P(I_S < 0)$

$$I_S = -\infty \quad \Rightarrow \quad z_1 = -\infty$$

$$I_S = 0 \quad \Rightarrow \quad z_2 = \frac{0-150}{47.43} = -3.1626$$

Quindi: $P(-\infty < z < -3.1626) = I(-\infty) - I(-3.1626) =$

$$I(\infty) - I(3.1626) = 0.5 - 0.4992 = 0.0008 = 0.08\%$$

C'è quindi una probabilità, piccola ma non nulla, che il materiale venga sollecitato ad un livello superiore al suo limite di resistenza nonostante lo sforzo nominale applicato sia metà di quello di rottura

Esempio 3

Qual è la probabilità che un numero elevato di misure cada in un intervallo attorno al valor medio pari a due volte la deviazione standard?

$$P(\bar{x} - 2\sigma_x < x < \bar{x} + 2\sigma_x)\,?$$

Dalla definizione della variabile z si ottiene che essa definisce un intervallo attorno alla media in termini di deviazione standard

$$z = \frac{x - \bar{x}}{\sigma_x} \qquad \Rightarrow \qquad x = \sigma_x z + \bar{x}$$

Sostituendo e con semplici passaggi:

$$P(\bar{x} - 2\sigma_x < x < \bar{x} + 2\sigma_x) = P(\bar{x} - 2\sigma_x < \sigma_x z + \bar{x} < \bar{x} + 2\sigma_x) =$$
$$= P(\cancel{\bar{x}} - 2\cancel{\sigma_x} < \cancel{\sigma_x} z + \cancel{\bar{x}} < \cancel{\bar{x}} + 2\cancel{\sigma_x}) = P(-2 < z < +2) =$$
$$= I(2) - I(-2) = 0.4772 - 0.4772 = 0.9544$$

Quindi la probabilità che la condizione si verifichi è del 95.44%.

Esempio 4

<u>Esempio</u>: Nella produzione di banchi motore, la tolleranza sul diametro D ha distribuzione gaussiana. Il D medio misurato è μ=4 in con una deviazione σ=0.002 in. Qual è la probabilità delle seguenti condizioni:

a) Un cilindro sia misurato con D $\leq$ 4.002 in;

b) D $\geq$ 4.005 in;

c) 3.993 $\leq$ D $\leq$4.003 in;

d) Se i cilindri con deviazione >2σ sono eliminati, che percentuale totale verrà eliminata?

Soluzione:

a) $$z = \frac{D-\mu}{\sigma} = \frac{4.002-4.000}{0.002} = 1$$

$$P(-\infty \le z \le 1) = P(-\infty \le z \le 0) + P(0 \le z \le 1) =$$
$$= .5 + .3413 = .8413 = 84.3\%$$

b) $$z = \frac{D-\mu}{\sigma} = \frac{4.005-4.000}{0.002} = 2.5$$

$$P(2.5 \le z \le +\infty) = P(0 \le z \le +\infty) - P(0 \le z \le 2.5) =$$
$$= .5 - .4938 = .0062 = 0.62\%$$

c) $$z_1 = \frac{3.993-4.00}{0.002} = -3.5 \quad ; \quad z_2 = \frac{4.003-4.00}{0.002} = +1.5$$

$$P(-3.5 \le z \le +1.5) = P(-3.5 \le z \le 0) + P(0 \le z \le 1.5) =$$
$$= .4998 + .4332 = .933 = 93.3\%$$

d) Se il criterio di rigetto è pari a 2σ , i banchi che vengono accettati sono:

$$P(-2 \le z \le +2) = 2P(0 \le z \le +2) = 95.44\%$$

mentre quelli che vengono eliminati sono:

$$1 - .9544 = .0456 = 4.56\%$$

Intervallo di confidenza

Da una FDP abbiamo ottenuto media e deviazione standard

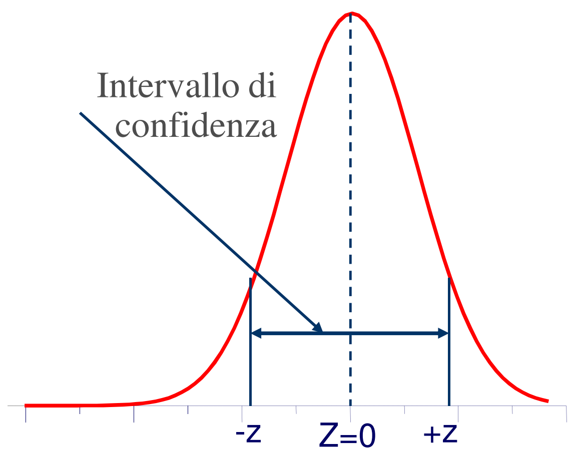
Abbiamo visto che la quest'ultima può essere utilizzata per definire un intervallo attorno alla media dentro il quale ci si può aspettare che rientrino nuovi rilievi

A questo intervallo si associa un livello di probabilità definito dall'area sottesa dalla funzione densità di probabilità tra i valori limite $\pm k\ \sigma$ rispetto alla media

z è la semi ampiezza dell'intervallo di confidenza: è definito in modo da avere garantita la probabilità che una successiva misura vi rientri

$$z = \frac{x - \bar{x}}{\sigma_x}$$

$$P(\bar{x} - k\sigma \leq x \leq \bar{x} + k\sigma) = \text{Valore}$$



Il problema dell'intervallo di confidenza si può porre nelle due forme:

a) che intervallo occorre assumere se è necessario conseguire un certo livello di confidenza?

Per rispondere a questa domanda occorre trovare k tale per cui a $\pm k\sigma$ corrisponde una probabilità $integrale(p(x), \pm k) = \%$

b) quale livello di confidenza è assicurato da un dato intervallo, generalmente espresso in rapporto alla deviazione standard?

In questi caso occorre valutare quale probabilità è associata all'intervallo $\pm k\ \sigma$

Problema a) che intervallo occorre assumere se è necessario conseguire un certo livello di confidenza?

Se l'intervallo di confidenza è simmetrico occorre rintracciare la metà del livello richiesto nella tabella e dedurre il parametro z corrispondente

Se ci interessa un livello di probabilità del 90% dobbiamo cercare 0.45 che si ottiene per $1.64 \leq z \leq 1.65$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
…										
1.4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1.5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
…										

b) Cosa significa assumere un intervallo di confidenza k volte la deviazione standard?

Utilizzando i valori integrati numericamente riportati in tabella è possibile valutare i **livelli di confidenza** (intervalli di confidenza), ovvero la probabilità che i valori della misura cadano negli **intervalli di confidenza** definiti da una semiampiezza pari a **n** volte la deviazione standard attorno al valor medio:

Intervallo di confidenza	**Livello di confidenza**
$\pm 1\,\sigma$	2 x .3413 = 68.26%
$\pm 2\,\sigma$	2 x .4772 = 95.44%
$\pm 3\,\sigma$	2 x .4987 = 99.74%
$\pm 3.5\,\sigma$	2 x .4998 = 99.96%

L'intervallo 2 σ (95%) è frequentemente utilizzato: significa che ci dobbiamo aspettare che ogni 20 misure solo una cada all'esterno dell'intervallo di confidenza

Intervallo di confidenza	Livello di confidenza
$\pm 1\,\sigma$	68.26%
$\pm 2\,\sigma$	95.44%
$\pm 3\,\sigma$	99.74%