**"Metodi statistici per l'analisi dei dati sperimentali"**
**DIMTI - Facoltà di Ingegneria - Università di Trento**
**Corso per il personale tecnico (08.01 - 22.02.2008)**

## Elementi di probabilità e statistica

**Prof. Stefano Siboni**

Argomenti del corso:

(1) Misure sperimentali ed errori

(2) Elementi di probabilità

(3) Variabili casuali e distribuzioni di probabilità

(4) Valori medi

(5) Distribuzioni di probabilità speciali

(6) Funzioni di variabili casuali e misure indirette

(7) Teoria dei campioni (stime e intervalli di confidenza)

(8) Test delle ipotesi (parametrici e non parametrici)

(9) Interpolazione, regressione e correlazione

(10) Analisi della varianza

# 1. MISURE SPERIMENTALI ED ERRORI

### □ Errore di misura

Differenza fra il valore di una grandezza misurato sperimentalmente ed il valore "vero" della stessa grandezza.
Mai certo, può solo essere stimato

### □ Errori sistematici

Errori riproducibili imputabili a tecniche di misura scorrette o all'impiego di apparati di misura non correttamente funzionanti o tarati.
Se individuati, possono essere rimossi o corretti

### □ Errori casuali

Esprimono le fluttuazioni nei risultati di misure sperimentali ripetute. Non sono associati a cause ben definite e individuabili. Non si riproducono

### □ Errore assoluto

Stima dell'errore di misura di una grandezza (di questa ha la stessa unità di misura)

### □ Errore relativo

Quoziente fra errore assoluto e valore vero stimato (presunto) di una grandezza. Si tratta di un numero puro, sovente espresso in forma percentuale. Quantifica la **precisione** di una misura

**Esempi di errori sistematici**:

strumento di misura non correttamente tarato:

- cronometro che ritarda o anticipa
- regolo deformato
- bilancia con masse campione alterate
- voltmetro-amperometro con resistenze alterate
- .....

procedura sperimentale non corretta, che in realtà non valuta la grandezza desiderata

errori di parallasse (procedura di lettura non corretta)

**Esempi di errori casuali**:

tempo di reazione dell'operatore nell'avvio/arresto della misura

disturbi meccanici (vibrazioni) sull'apparato sperimentale

disturbi elettromagnetici sull'apparato sperimentale

errori di parallasse (procedura di lettura corretta)

### Accuratezza di una misura

Indica quanto il risultato della misura sperimentale di una grandezza può ritenersi prossimo al valore "vero" della grandezza stessa

$$\text{accuratezza} \iff \text{irrilevanza degli errori sistematici}$$

### Precisione di una misura

Esprime quanto esattamente il risultato di una misura è determinato, senza riferimento ad alcun valore "vero" della grandezza misurata. È sinonimo di **riproducibilità**

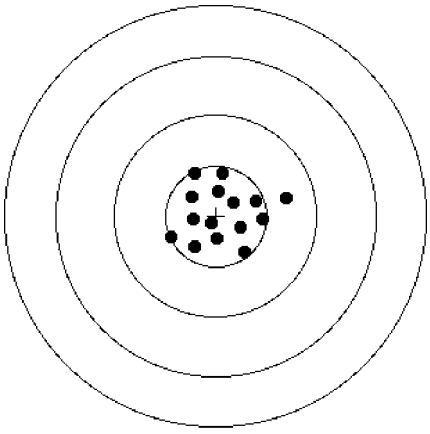
$$\begin{aligned} \text{precisione} &\iff \text{riproducibilità} \\ &\iff \text{irrilevanza degli errori casuali} \\ &\iff \text{irrilevanza dell}'\text{errore relativo} \end{aligned}$$

### Postulato della popolazione statistica

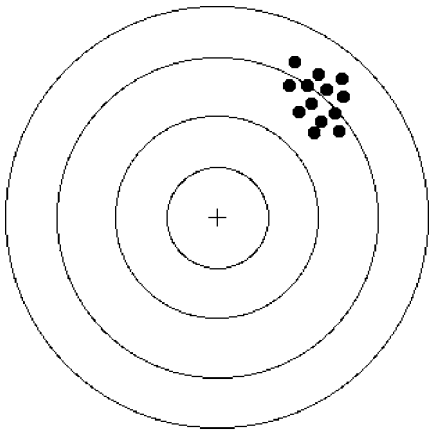
In presenza di errori casuali, il risultato di una misura viene interpretato come la realizzazione di una variabile casuale, che descrive una **popolazione statistica**

### Popolazione statistica

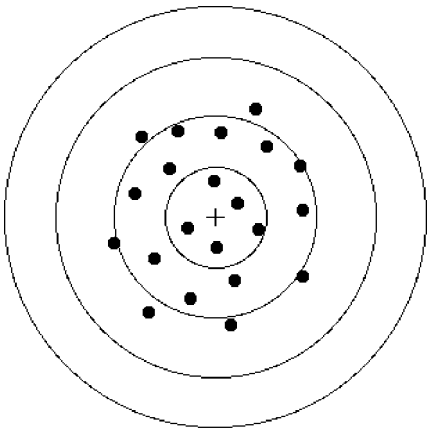
Insieme infinito ed ipotetico di punti (valori) di cui i punti sperimentali si assumono essere un campione casuale



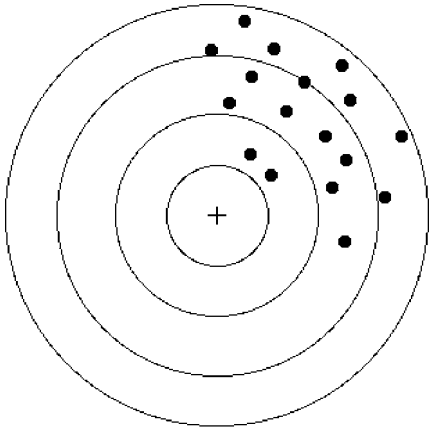
good precision
good accuracy



good precision
bad accuracy



bad precision
good accuracy



bad precision
bad accuracy

**Confronto fra precisione e accuratezza**

# 2. ELEMENTI DI PROBABILITÀ

## □ Esperimento casuale

Una misura sperimentale soggetta ad errori casuali;

– il risultato della misura non è riproducibile, ovvero

– la ripetizione dell'esperimento conduce a risultati diversi.

## □ Spazio campione (sample space)

L'insieme $\mathbb{S}$ di tutti i possibili risultati di un esperimento casuale.

## □ Punto campione (sample point)

Ciascun possibile risultato di un esperimento casuale.

Lo spazio campione che descrive i risultati di un esperimento può non essere unico (secondo il diverso livello di dettaglio della descrizione).

• **Esempio:** Per il lancio di un dado può aversi:

$$\mathbb{S} = \{1, 2, 3, 4, 5, 6\}$$

$$\mathbb{S} = \{\text{risultato pari}, \text{risultato dispari}\}$$

Quando possibile i punti campione vengono rappresentati in forma numerica (sempre, nel caso di un esperimento casuale). Lo spazio campione $\mathbb{S}$ è quindi un insieme numerico.

## □ Tipologie di spazi campione

– $\mathbb{S}$ **finito**: i punti campione sono in numero finito.

• **Esempio:** risultati del lancio di un dado

– $\mathbb{S}$ **numerabile**: i punti campione costituiscono una successione, o comunque sono in corrispondenza biunivoca con l'insieme $\mathbb{N}$ dei numeri naturali

• **Esempio:** conteggio dei decadimenti radioattivi in un campione di radionuclide

– $\mathbb{S}$ **non numerabile**: i punti campione sono in corrispondenza biunivoca con un intervallo reale

• **Esempio:** valore misurato di una grandezza continua soggetta a fluttuazioni o variazioni casuali

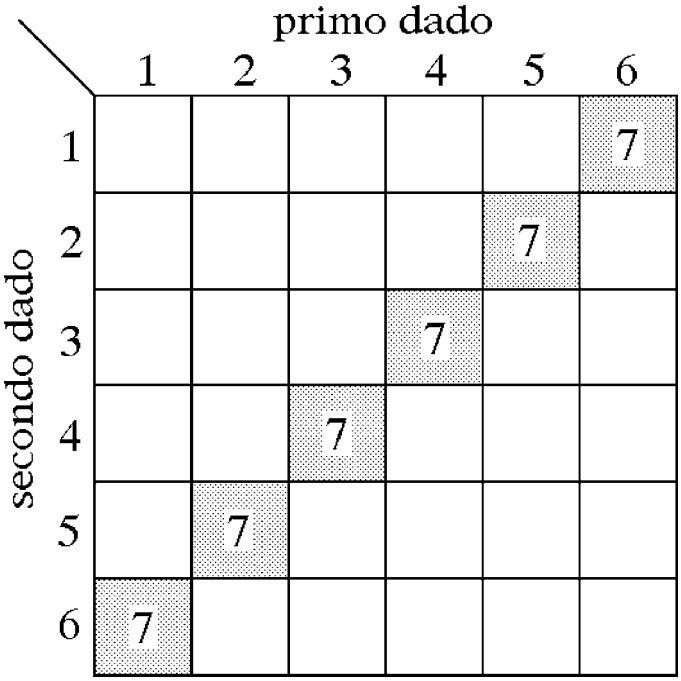
– $\mathbb{S}$ **discreto**: spazio campione finito o numerabile

– $\mathbb{S}$ **continuo**: spazio campione non numerabile

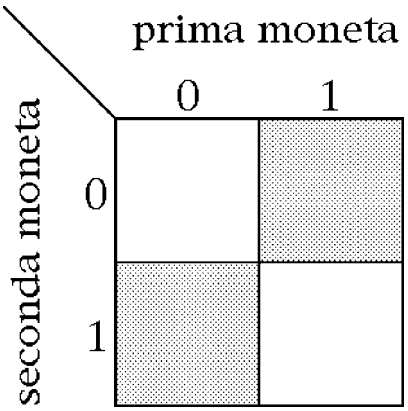
## □ Evento

Un sottoinsieme $A$ di $\mathbb{S}$, ossia un insieme di risultati possibili di un esperimento casuale

• **Esempio**: il risultato che nel lancio di due dadi la somma dei punteggi sia uguale a 7




• **Esempio**: l'evento che nel lancio di due monete esca testa una sola volta (testa=0, croce=1)



### □ Realizzazione di un evento

Se il risultato di un esperimento casuale è un elemento dell'evento $A$ si dice che $A$ si è verificato (o realizzato)

### □ Eventi notevoli:

– **semplice (o elementare)**: evento consistente di un solo punto campione

– **certo**: l'evento costituito dall'intero spazio campione $\mathbb{S}$ (l'evento è realizzato qualunque sia il risultato)

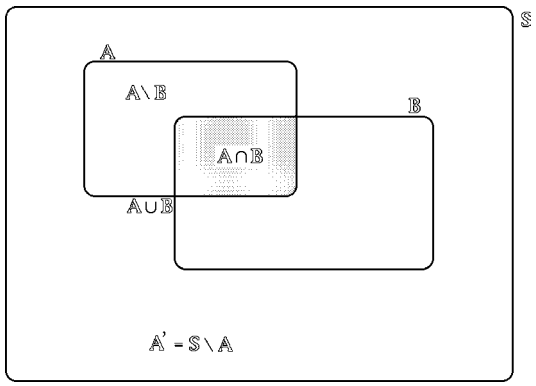
– **impossibile**: l'evento costituito dall'insieme vuoto $\emptyset$ in $\mathbb{S}$ (l'evento si realizza quando l'esperimento non produce alcun risultato, il che è impossibile)

– $A \cup B$**, unione degli eventi** $A$ **e** $B$: è verificato quando si realizza $A$ oppure $B$

– $A \cap B$**, intersezione degli eventi** $A$ **e** $B$: verificato quando si verificano sia $A$ che $B$

– $A'$**, complementare di** $A$: si verifica quando $A$ non si realizza

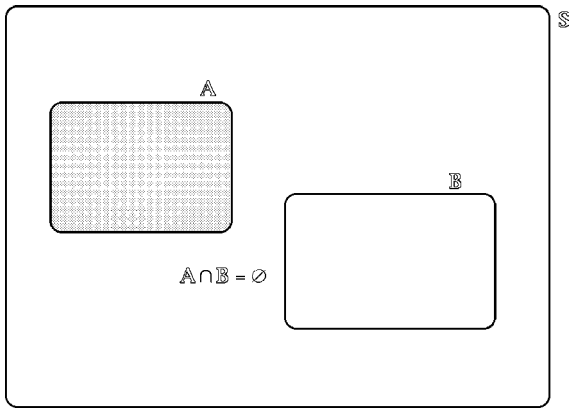
– $A \setminus B$**,** $A$ **ma non** $B$: verificato quando si realizza $A$ ma non $B$ (da notare che $A' = \mathbb{S} \setminus A$)



### □ Coppia di eventi mutualmente esclusivi

Due eventi $A$ e $B$ si dicono mutualmente esclusivi se essi non possono verificarsi contemporaneamente, se cioè l'evento intersezione è impossibile:

$$A \cap B = \emptyset$$




• **Esempio**: nel lancio di un dado, gli eventi:

$$E_1 = \{\text{il risultato è pari}\} \quad \text{e} \quad E_2 = \{\text{il risultato è dispari}\}$$

sono mutualmente esclusivi in quanto

$$E_1 \cap E_2 = \emptyset$$

### □ Collezione finita di eventi mutualmente esclusivi

Gli eventi $A_1, A_2, \ldots, A_n$ si dicono mutualmente esclusivi se essi lo sono due a due:

$$A_i \cap A_j = \emptyset \qquad \forall\, i, j = 1, \ldots, n$$

### □ Nozione di probabilità

Ad ogni evento di un esperimento casuale si assegna una **probabilità**, un numero compreso fra 0 e 1

Tale numero quantifica la possibilità (chance) che l'evento si verifichi contro la possibilità che esso non si verifichi

All'**evento certo** si assegna una probabilità 1 (o del 100%)

All'**evento impossibile** è attribuita una probabilità 0 (ovvero dello 0%)

### □ Stima della probabilità di un evento

Si possono considerare tre diversi approcci:

($i$) **approccio classico**

($ii$) **approccio frequentistico**

($iii$) **approccio assiomatico**

($i$) **Approccio classico alla definizione della probabilità**
Se lo spazio campione è costituito da un **numero finito** $n$ di punti campione **ugualmente probabili**, un evento composto da $f$ punti campione ha probabilità $f/n$ (probabilità **a priori**)

• **Esempio**: nel lancio di una moneta si hanno due soli punti campione, corrispondenti al risultato testa o al risultato croce. I risultati si assumono ugualmente probabili (moneta non truccata):

– probabilità che esca testa $= 1/2$

– probabilità che esca croce $= 1/2$

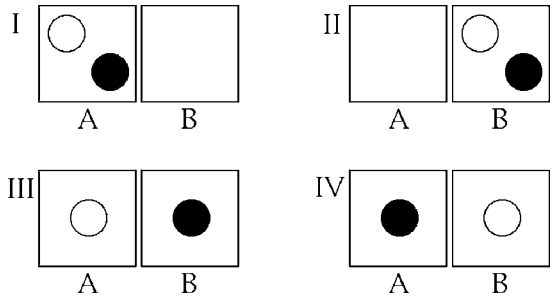
• **Esempio**: nell'estrazione di una carta da un mazzo di 52 carte, si può assumere che tutte le carte abbiano la stessa probabilità di essere estratte (mazzo ben mescolato). Si avrà allora:

– probabilità di estrarre una carta di quadri $= 13/52 = 1/4$

– probabilità di estrarre un re $= 4/52 = 1/13$

**Problema di buona posizione:**
come riconoscere punti campione **equiprobabili**?

• **Esempio**: due palline, bianca e nera, collocate a caso in due contenitori A e B, possono assumere quattro configurazioni diverse, che si possono assumere equiprobabili

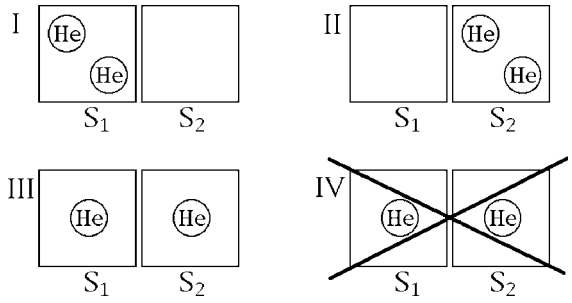



La probabilità che in ciascun contenitore sia contenuta una pallina è quindi

$$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

le configurazioni III e IV avendo uguale probabilità 1/4

• **Esempio**: due atomi di elio collocati a caso in due stati quantomeccanici S1 ed S2 possono assumere **tre** configurazioni, tutte equiprobabili, e non quattro



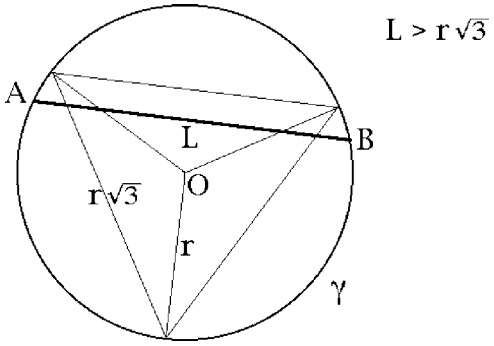

Gli atomi di elio sono infatti particelle identiche di Bose (o **bosoni**): le configurazioni III e IV sono **indistinguibili** e rappresentano lo **stesso stato quantomeccanico** dei due atomi

**Ulteriore problema di buona posizione:**
come estendere la definizione al caso di spazi campione infiniti?

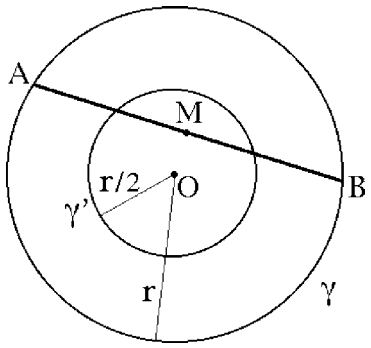
Per applicare la definizione classica si devono usare lunghezze, aree, o qualche altra **misura di insiemi infiniti** per determinare il quoziente $f/n$

In generale l'estensione è tutt'altro che ovvia e può condurre a risultati contraddittori

• **Esempio**: **Paradosso di Bertrand**
Dato un cerchio $\gamma$ di centro $O$ e raggio $r$, calcolare la probabilità $p$ che una corda $AB$ scelta a caso abbia una lunghezza $L$ maggiore del lato $r\sqrt{3}$ del triangolo equilatero inscritto

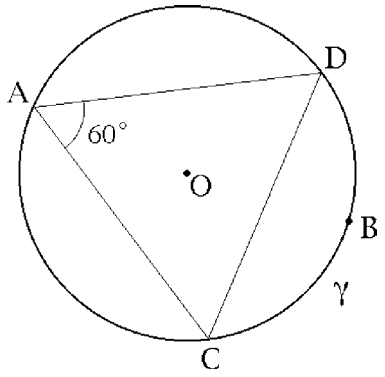



**(a)** se il punto medio $M$ della corda $AB$ giace dentro il cerchio $\gamma'$ di centro $O$ e raggio $r/2$, allora $L > r\sqrt{3}$. Si può così porre:




$$p = \frac{\text{area}(\gamma')}{\text{area}(\gamma)} = \frac{\pi(r/2)^2}{\pi r^2} = \frac{1}{4}$$

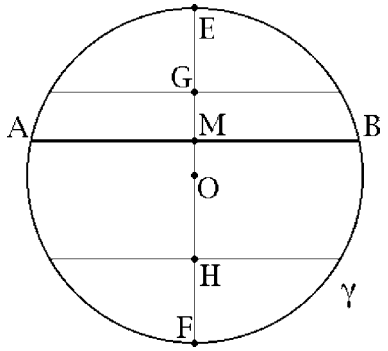
**(b)** si fissi l'estremo $A$ della corda. Ciò riduce il numero dei casi possibili ma anche — e nella stessa proporzione — il numero di posizioni utili del secondo estremo $B$ (nessun effetto su $p$).




Qualsiasi estremo $B$ compreso entro l'arco $CD$ di $\gamma$ tale che $C\hat{A}D = 60^o$ assicura $L > r\sqrt{3}$. Vale allora

$$p = \frac{\text{lunghezza arco } CD}{\text{lunghezza circonferenza } \gamma} = \frac{2\pi r/3}{2\pi r} = \frac{1}{3}$$

**(c)** si scelga la corda $AB$ perpendicolare a un diametro $EF$ fissato. La restrizione non ha alcun effetto su $p$. Indicati con $G$ e $H$ i punti medi dei raggi $OE$ ed $OF$, risulta $L > r\sqrt{3}$ se e solo se il punto medio $M$ di $AB$ appartiene al segmento $GH$. Si pone perciò:




$$p = \frac{\text{lunghezza}(GH)}{\text{lunghezza}(EF)} = \frac{r}{2r} = \frac{1}{2}$$

**TRE DEFINIZIONI DIVERSE**, TUTTE PLAUSIBILI!

($ii$) **Approccio frequentistico alla probabilità**
Se, avendo ripetuto un esperimento un numero **molto grande** $n$ di volte, un evento si è verificato $f$ volte, allora la probabilità dell'evento è $f/n$ (**probabilità empirica**, o **a posteriori**)

• **Esempio**: una moneta viene lanciata 1000 volte e si trova che il risultato testa si verifica 531 volte. La probabilità di ottenere testa è stimata come $531/1000 = 0.531$

• **Esempio**: si lancia un dado 600 volte, ottenendo i seguenti risultati per le rispettive facce $(1, 2, 3, 4, 5, 6)$:

$$49 \qquad 102 \qquad 118 \qquad 141 \qquad 111 \qquad 79$$

Le probabilità empiriche stimate sono pertanto:

$$\frac{49}{600} \qquad \frac{102}{600} \qquad \frac{118}{600} \qquad \frac{141}{600} \qquad \frac{111}{600} \qquad \frac{79}{600}$$

Il risultato mette in discussione l'assunto della **equiprobabilità** delle facce, su cui si basa l'approccio classico alla probabilità: **il dado potrebbe essere truccato**

**Problema di buona posizione**:
come stabilire se il numero di prove ripetute è **abbastanza grande**?

(*iii*) **Approccio assiomatico alla probabilità**
Per ovviare ai problemi di buona posizione, impliciti nell'approccio classico ed in quello frequentistico, si preferisce ricorrere alla **definizione assiomatica** della probabilità

Gli elementi fondamentali della definizione assiomatica sono i concetti di **insieme misurabile** e di **misura di probabilità**

**Insiemi misurabili**
Nello spazio campione $\mathbb{S}$ gli eventi non possono essere sottoinsiemi qualsiasi, ma soltanto sottoinsiemi "misurabili"

$$\text{eventi} \qquad \Longleftrightarrow \qquad \text{sottoinsiemi "misurabili" di } \mathbb{S}$$

La famiglia $\mathfrak{F}$ dei sottoinsiemi misurabili di $\mathbb{S}$ soddisfa i seguenti requisiti:

- $A \in \mathfrak{F} \Rightarrow A' \in \mathfrak{F}$ (il complementare di un insieme misurabile è a sua volta misurabile)
- l'unione numerabile di insiemi misurabili è sempre un insieme misurabile

$$A_n \in \mathfrak{F} \quad \forall\, n \in \mathbb{N} \qquad \Longrightarrow \qquad \bigcup_{n \in \mathbb{N}} A_n \in \mathfrak{F}$$

In particolare si ha che lo spazio campione è misurabile

$$A \in \mathfrak{F} \quad \Longrightarrow \quad A' \in \mathfrak{F} \quad \Longrightarrow \quad \mathbb{S} = A \cup A' \in \mathfrak{F}$$

e di conseguenza anche l'insieme vuoto

$$\emptyset = \mathbb{S}' \in \mathfrak{F}$$

**Nota**
Una famiglia di sottoinsiemi misurabili di $\mathbb{S}$ si dice una **$\sigma$-algebra** di $\mathbb{S}$

• **Esempio:** per **spazi campione continui**, che sono o $\mathbb{R}^n$ o sottoinsiemi di $\mathbb{R}^n$, di solito si considera la **$\sigma$-algebra di Borel** generata dagli usuali sottoinsiemi aperti dello spazio campione (topologia indotta dalla distanza euclidea)

$$\{\text{famiglia degli aperti di } \mathbb{R}^n\} \subset \mathfrak{F}$$

La $\sigma$-algebra di Borel in $\mathbb{R}^n$ è la più piccola $\sigma$-algebra che contiene tutti i sottoinsiemi aperti di $\mathbb{R}^n$

• **Esempio:** per **spazi campione discreti**, di solito fa uso della $\sigma$-algebra di Borel rispetto alla **topologia discreta**

topologia discreta in $\mathbb{S}$

$\Updownarrow$

tutti i sottoinsiemi di $\mathbb{S}$ sono aperti

$\Updownarrow$

tutti i sottoinsiemi di $\mathbb{S}$ appartengono alla $\sigma$-algebra di Borel

In uno spazio campione discreto, di regola, tutti i sottoinsiemi sono dunque considerati (aperti e) misurabili: essi costituiscono dunque degli **eventi**

**Misura di probabilità**
Una misura di probabilità, o **funzione di probabilità**, è una funzione reale $P$ definita sulla famiglia degli insiemi misurabili

$$P \ : \ \mathfrak{F} \longrightarrow \mathbb{R}$$

che soddisfa le seguenti proprietà:

- $\forall\, A \in \mathfrak{F}$ vale $P(A) \geq 0$ (**assioma di positività**)
- $P(\mathbb{S}) = 1$, la probabilità dell'evento certo è uguale a 1 (**assioma di normalizzazione**)
- se $A_n \in \mathfrak{F}\ \forall\, n \in \mathbb{N}$ e $A_i \cap A_j = \emptyset\ \forall\, i, j \in \mathbb{N}$, $i \neq j$, si ha

$$P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} P(A_n)$$

(**assioma di additività numerabile**)

In particolare, per una **famiglia finita** $(A_n)_{n=1,\ldots,k}$ di insiemi misurabili due a due disgiunti (eventi mutualmente esclusivi) vale

$$P\Big(\bigcup_{n=1}^{k} A_n\Big) = \sum_{n=1}^{k} P(A_n)$$

Molte altre proprietà utili seguono direttamente dagli assiomi

## Proprietà notevoli della misura di probabilità

- se l'evento $A_1$ può verificarsi solo essendosi verificato l'evento $A_2$, la probabilità dell'evento "$A_2$ ma non $A_1$" è uguale alla differenza delle probabilità di $A_2$ e $A_1$

$$A_1 \subset A_2 \quad \Rightarrow \quad P(A_2 \setminus A_1) = P(A_2) - P(A_1)$$

- se l'evento $A_1$ può verificarsi solo essendosi verificato l'evento $A_2$, la probabilità del primo non eccede quella del secondo

$$A_1 \subset A_2 \quad \Rightarrow \quad P(A_1) \leq P(A_2)$$

- la probabilità di qualsiasi evento è un numero compreso fra 0 e 1

$$A \in \mathfrak{F} \quad \Rightarrow \quad 0 \leq P(A) \leq 1$$

- la probabilità dell'evento impossibile è nulla

$$P(\emptyset) = 0$$

- per un evento $A$ la probabilità dell'evento complementare $A'$ è il complemento a 1 della probabilità di $A$

$$A \in \mathfrak{F} \quad \Rightarrow \quad P(A') = 1 - P(A)$$

- per ogni coppia di eventi, la probabilità dell'evento unione è uguale alla somma delle probabilità degli eventi parziali ridotta della probabilità dell'evento intersezione

$$A, B \in \mathfrak{F} \quad \Rightarrow \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- per ogni coppia di eventi $A$ e $B$, la probabilità di $A$ è la somma delle probabilità di $A \cap B$ e di $A \cap B'$

$$A, B \in \mathfrak{F} \quad \Rightarrow \quad P(A) = P(A \cap B) + P(A \cap B')$$

**Assegnazione della funzione di probabilità**
Se lo spazio campione $\mathbb{S}$ è **finito**

$$\mathbb{S} = \{a_1, a_2, \ldots, a_n\}$$

allora le probabilità degli eventi semplici (punti campione)

$$A_1 = \{a_1\} \qquad A_2 = \{a_2\} \qquad \ldots\ldots \qquad A_n = \{a_n\}$$

devono soddisfare la condizione di normalizzazione

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1$$

L'assegnazione delle probabilità $P(A_1), \ldots, P(A_n) \in [0, 1]$ deve rispettare questo requisito

Se gli eventi semplici si assumono **equiprobabili** si ottiene la definizione classica della probabilità

$$P(A_i) = \frac{1}{n} \qquad \forall\, i = 1, \ldots, n$$

Una condizione analoga deve essere rispettata nel caso di uno spazio campione **infinito** (discreto o continuo)

L'assegnazione della probabilità costituisce un **modello matematico**, la cui validità va verificata sperimentalmente

Un modello probabilistico può quindi risultare **più o meno adeguato** a descrivere gli esiti di un esperimento casuale

### □ Probabilità condizionata

Dati due eventi $A$ e $B$, con $P(A) > 0$, si indica con

$$P(B|A)$$

la probabilità di $B$ quando si suppone che $A$ si sia verificato, nota come la **probabilità di $B$ condizionata ad $A$**

Essendosi realizzato $A$, questo evento diventa il nuovo **spazio campione** in luogo di $\mathbb{S}$.

Ciò giustifica la definizione

$$P(B|A) \;=\; \frac{P(A \cap B)}{P(A)}$$

che equivale alla relazione

$$P(A \cap B) \;=\; P(A)\, P(B|A)$$

• **Esempio**: nel lancio di un dado non truccato si vuole determinare la probabilità di ottenere un numero minore di 4:

($a$) qualora non si disponga di ulteriori informazioni

$$P(<4) \;=\; P(1) + P(2) + P(3) \;=\; \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \;=\; \frac{1}{2}\,;$$

($b$) sapendo che il risultato è un numero dispari

$$P(<4|1,3,5) \;=\; \frac{P(1,3)}{P(1,3,5)} \;=\; \frac{\dfrac{1}{6} + \dfrac{1}{6}}{\dfrac{1}{6} + \dfrac{1}{6} + \dfrac{1}{6}} \;=\; \frac{2}{3}\,.$$

**Primo teorema sulla probabilità condizionata**
Per tre eventi $A_1, A_2, A_3$ qualsiasi vale la relazione

$$P(A_1 \cap A_2 \cap A_3) \;=\; P(A_1)\;\; P(A_2|A_1)\;\; P(A_3|A_1 \cap A_2)$$

La probabilità che si verifichino tutti e tre gli eventi è cioè data da una catena di prodotti:

la probabilità $P(A_1)$ dell'evento $A_1$

**moltiplicata per**

la probabilità $P(A_2|A_1)$ dell'evento $A_2$
supponendo che $A_1$ si sia verificato

**moltiplicata per**

la probabilità $P(A_3|A_1 \cap A_2)$ dell'evento $A_3$
supponendo che $A_1$ ed $A_2$ si siano verificati
(ossia che si sia realizzato l'evento $A_1 \cap A_2$)

Il risultato si estende immediatamente al caso di $n$ eventi:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) \;=\; P(A_1)\; P(A_2|A_1)\; P(A_3|A_1 \cap A_2)\; \ldots$$

$$\ldots\;\; P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

come è facile verificare per induzione

• **Esempio**: due carte sono estratte da un mazzo ben mescolato di 52 carte. Calcolare la probabilità di estrarre due assi se la prima carta ($a$) è rimessa nel mazzo ($b$) non è rimessa nel mazzo

**Soluzione**
Introdotti gli eventi

$$A_1 = \{\text{asso alla } 1^a \text{ carta}\} \qquad A_2 = \{\text{asso alla } 2^a \text{ carta}\}$$

si deve calcolare la probabilità

$$P(A_1 \cap A_2) = P(A_1)\ P(A_2|A_1)$$

osservando che in ambo i casi risulta $P(A_1) = 4/52 = 1/13$

($a$) se la prima carta è rimessa nel mazzo, la seconda estrazione avviene su un mazzo di 52 carte con 4 assi disponibili, per cui si ha

$$P(A_2|A_1) = \frac{4}{52} = \frac{1}{13}$$

e quindi

$$P(A_1 \cap A_2) = \frac{1}{13} \cdot \frac{1}{13} = \frac{1}{169}$$

($b$) se la prima carta non è rimessa nel mazzo, alla seconda estrazione il mazzo contiene 3 soli assi su un totale di 51 carte e risulta pertanto

$$P(A_2|A_1) = \frac{3}{51} = \frac{1}{17}$$

in modo che la probabilità richiesta vale

$$P(A_1 \cap A_2) = \frac{1}{13} \cdot \frac{1}{17} = \frac{1}{221}$$

**Secondo teorema sulla probabilità condizionata**
Se l'evento $A$ può verificarsi solo essendosi verificato uno degli eventi $A_1, A_2, \ldots, A_n$, tra loro mutualmente esclusivi, ossia

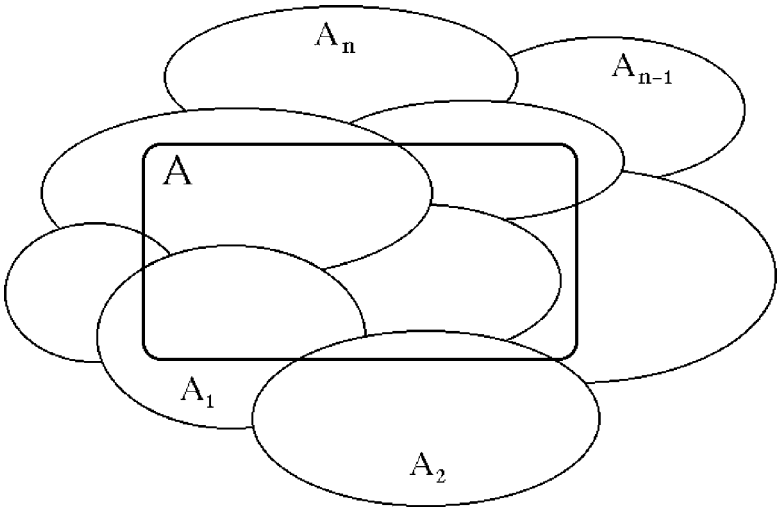
$$A \subseteq \bigcup_{i=1}^{n} A_i \qquad A_i \cap A_j = \emptyset \quad \forall\, i, j = 1, \ldots, n\,,$$

vale la relazione

$$P(A) = \sum_{i=1}^{n} P(A \cap A_i)$$

che, usando la definizione di probabilità condizionata e assumendo $P(A_i) > 0\ \forall\, i = 1, \ldots, n$, si riduce alla forma equivalente

$$P(A) = \sum_{i=1}^{n} P(A|A_i)\, P(A_i)$$



### □ Eventi indipendenti

**Due eventi** $A$ e $B$ si dicono **indipendenti** se la probabilità del verificarsi di $B$ non è influenzata dal fatto che $A$ si sia verificato o meno, se cioè

$$P(B|A) = P(B)$$

In base alla definizione di probabilità condizionata, questa condizione equivale a

$$P(A \cap B) = P(A)\ P(B)$$

per cui può anche esprimersi (simmetricamente) nella forma

$$P(A|B) = P(A)$$

**Tre eventi** $A_1, A_2, A_3$ si definiscono indipendenti se la probabilità di ciascuno di essi non dipende dal fatto che si siano verificati gli altri due (uno solo o entrambi). Ciò equivale a richiedere che gli eventi siano due a due indipendenti

$$P(A_i \cap A_j) = P(A_i)\ P(A_j) \qquad \forall\, i, j = 1, 2, 3$$

e che risulti inoltre

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)\ P(A_2)\ P(A_3)$$

L'estensione al caso di $n$ **eventi indipendenti** è analoga

**Osservazione**

Gli eventi mutualmente esclusivi si possono riconoscere prima ancora di introdurre una probabilità (eventi per i quali l'evento intersezione si identifica con l'evento impossibile)

Gli eventi indipendenti sono riconoscibili solo una volta assegnata la misura di probabilità

• **Esempio**: tre palline sono estratte successivamente, a caso, da un'urna che ne contiene 6 rosse, 4 bianche e 5 azzurre. Si calcoli la probabilità che esse siano, nell'ordine, rossa, bianca e azzurra quando ciascuna pallina $(a)$ è rimessa nell'urna dopo l'estrazione, ovvero $(b)$ non è rimessa nell'urna.

**Soluzione**
Si introducano gli eventi:

$$\begin{aligned} R_1 &= \{\text{pallina rossa alla } 1^a \text{ estrazione}\} \\ B_2 &= \{\text{pallina bianca alla } 2^a \text{ estrazione}\} \\ A_3 &= \{\text{pallina azzurra alla } 3^a \text{ estrazione}\} \end{aligned}$$

La probabilità da calcolare è perciò

$$P(R_1 \cap B_2 \cap A_3) = P(R_1)\ P(B_2|R_1)\ P(A_3|R_1 \cap B_2)$$

$(a)$ se le palline estratte sono rimesse nell'urna **gli eventi sono indipendenti** e la probabilità cercata diventa

$$\begin{aligned} P(R_1 \cap B_2 \cap A_3) &= P(R_1)\ P(B_2|R_1)\ P(A_3|R_1 \cap B_2) = \\ &= P(R_1)\ P(B_2)\ P(A_3) = \\ &= \frac{6}{6+4+5}\ \frac{4}{6+4+5}\ \frac{5}{6+4+5} = \frac{8}{225} \end{aligned}$$

$(b)$ se le palline estratte non sono reimbussolate **gli eventi non sono indipendenti** e la probabilità richiesta vale

$$\begin{aligned} P(R_1 \cap B_2 \cap A_3) &= P(R_1)\ P(B_2|R_1)\ P(A_3|R_1 \cap B_2) = \\ &= \frac{6}{6+4+5}\ \frac{4}{5+4+5}\ \frac{5}{5+3+5} = \frac{4}{91} \end{aligned}$$

**□ Teorema di Bayes (o regola di Bayes)**
Siano $A_1, A_2, \ldots, A_n$ eventi mutualmente esclusivi la cui unione coincida con lo spazio campione $\mathbb{S}$

$$\bigcup_{i=1}^{n} A_i = \mathbb{S} \qquad A_i \cap A_j = \emptyset \quad \forall\, i, j = 1, 2, \ldots, n$$

e le cui probabilità siano strettamente positive

$$P(A_i) > 0 \qquad \forall\, i = 1, 2, \ldots, n$$

Per ogni evento $A$ vale allora le relazione seguente

$$P(A_k|A) = \frac{P(A_k)\ P(A|A_k)}{\sum_{i=1}^{n} P(A_i)\ P(A|A_i)}$$

nota come **regola di Bayes**.

Il teorema permette di determinare le probabilità degli eventi $A_1, \ldots, A_n$ che possono essere la causa di $A$, **una volta che** $A$ **si sia verificato**

$$P(A_k|A) \qquad \forall\, k = 1, 2, \ldots, n$$

purchè si conoscano le probabilità $P(A_i)$ delle singole cause e le probabilità $P(A|A_i)$ che l'evento $A$ sia determinato dalla causa $A_i,\ \forall\, i = 1, 2, \ldots, n$

Per questo il teorema di Bayes è anche noto come **teorema della probabilità delle cause**

**Dimostrazione**
Basta ricordare che, per definizione,

$$P(A_k|A) \;=\; \frac{P(A_k \cap A)}{P(A)}$$

e osservare che

$$P(A_k \cap A) \;=\; P(A|A_k)\; P(A_k)$$

mentre

$$\begin{aligned} P(A) \;&=\; P(A \cap \mathbb{S}) \;= \\ &=\; P\Big(A \cap \bigcup_{i=1}^{n} A_i\Big) \;= \\ &=\; \sum_{i=1}^{n} P(A \cap A_i) \;= \\ &=\; \sum_{i=1}^{n} \frac{P(A \cap A_i)}{P(A_i)}\; P(A_i) \;= \\ &=\; \sum_{i=1}^{n} P(A|A_i)\; P(A_i) \end{aligned}$$

Per ottenere il risultato non si deve che sostituire queste espressioni nel numeratore e nel denominatore della definizione di $P(A_k|A)$

• **Esempio**: l'incidenza media di una malattia in una certa popolazione è pari al 10%. Si dispone di un test che segnala la presenza della malattia:

- con una probabilità del 95% su un soggetto malato;
- con una probabilità del 3% su un soggetto sano.

Nel caso che il test dia esito positivo su un soggetto, si vuole determinare la probabilità che questo sia effettivamente malato.

**Soluzione**
Si introducono i seguenti eventi mutualmente esclusivi e la cui unione è certa:

$$A_1 = \{\text{il soggetto è sano}\} \qquad A_2 = \{\text{il soggetto è malato}\}$$

con probabilità rispettive:

$$P(A_1) = 1 - 0.10 = 0.90 \qquad P(A_2) = 0.10$$

dove $P(A_2) = 10\%$ si identifica con l'incidenza media della malattia sulla popolazione, mentre $P(A_1) = 1 - P(A_2)$.

L'evento di interesse è infine

$$A = \{\text{il test dà esito positivo su un soggetto}\}$$

con le probabilità condizionate:

$P(A|A_1) = 0.03$, probabilità che il test dia esito positivo su un soggetto sano;

$P(A|A_2) = 0.95$, probabilità che il test dia esito positivo su un soggetto malato.

Si vuole calcolare la probabilità che un soggetto sia malato sapendo su di esso che il test ha dato un risultato positivo, vale a dire la probabilità condizionata

$$P(A_2|A)$$

Il teorema di Bayes porge

$$P(A_2|A) = \frac{P(A|A_2)\ P(A_2)}{P(A|A_1)\ P(A_1) + P(A|A_2)\ P(A_2)}$$

ovvero, sostituendo i valori numerici delle probabilità,

$$P(A_2|A) = \frac{0.95 \cdot 0.10}{0.03 \cdot 0.90 + 0.95 \cdot 0.10}$$

La probabilità richiesta vale pertanto

$$P(A_2|A) = 0.778688525$$

La probabilità che sia effettivamente malato un soggetto sul quale il test ha dato esito positivo è quindi pari al 77.87% circa

• **Esempio**: nello stesso problema precedente, si valuti la probabilità che sia effettivamente sano un soggetto per il quale il test ha dato risultato negativo.

**Soluzione**
L'evento interessante in questo caso è

$$B = \mathbb{S} \setminus A = \{\text{il test dà esito negativo su un soggetto}\}$$

con le probabilità condizionate:

- $P(B|A_1) = 1 - P(A|A_1) = 0.97$, prob. che il test risulti negativo su un soggetto sano;
- $P(B|A_2) = 1 - P(A|A_2) = 0.05$, prob. che il test risulti negativo su un soggetto malato.

La probabilità che sia effettivamente sano un soggetto per il quale il test abbia dato risultato negativo è data da

$$P(A_1|B)$$

che il teorema di Bayes permette di scrivere nella forma

$$P(A_1|B) = \frac{P(B|A_1)\ P(A_1)}{P(B|A_1)\ P(A_1) + P(B|A_2)\ P(A_2)}$$

Non rimane che sostituire i valori numerici delle probabilità $P(A_1)$, $P(A_2)$ e delle prob. condizionate $P(B|A_1)$, $P(B|A_2)$

$$P(A_1|B) = \frac{0.97 \cdot 0.90}{0.97 \cdot 0.90 + 0.05 \cdot 0.10} = 0.994305239$$

La probabilità che sia sano un soggetto sul quale il test ha avuto esito negativo è quindi del 99.43% circa

• **Esempio**: l'urna 1 contiene 3 palline blu e 2 rosse, mentre l'urna 2 ne contiene 2 blu e 5 rosse. Con uguale probabilità si può accedere all'una o all'altra urna per estrarne una pallina. Viene estratta una pallina blu; determinare la probabilità che l'estrazione abbia avuto luogo dall'urna 1.

**Soluzione**
Si definiscono gli eventi di interesse:

$A_1$ = scelta dell'urna 1 per l'estrazione

$A_2$ = scelta dell'urna 2 per l'estrazione

$A$ = estrazione di una pallina blu dall'urna prescelta

Per ipotesi le probabilità di accesso alle due urne sono uguali

$$P(A_1) = \frac{1}{2} \qquad P(A_2) = \frac{1}{2}$$

mentre le probabilità di estrazione di una pallina blu da ciascuna urna valgono

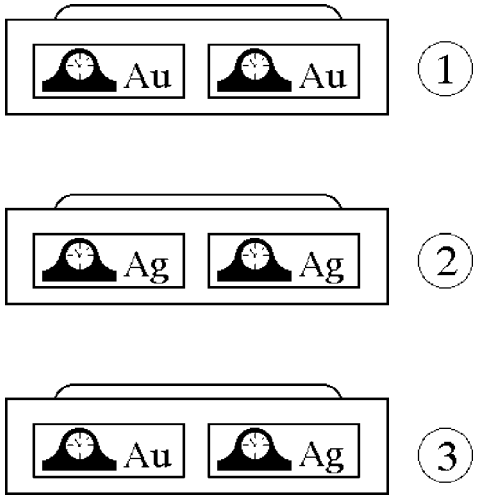
$$P(A|A_1) = \frac{3}{3+2} = \frac{3}{5} \qquad P(A|A_2) = \frac{2}{2+5} = \frac{2}{7}$$

La probabilità che l'urna 1 sia quella scelta per l'estrazione è allora data dalla formula di Bayes

$$P(A_1|A) = \frac{P(A|A_1)\ P(A_1)}{P(A|A_1)\ P(A_1) + P(A|A_2)\ P(A_2)} =$$

$$= \frac{\frac{3}{5} \cdot \frac{1}{2}}{\frac{3}{5} \cdot \frac{1}{2} + \frac{2}{7} \cdot \frac{1}{2}} = \frac{3}{10} \cdot \frac{70}{31} = \frac{21}{31}$$

• **Esempio**: tre cofanetti identici hanno due cassetti ciascuno. Il cofanetto 1 contiene un orologio d'oro in ciascun cassetto; i cassetti del cofanetto 2 contengono un orologio d'argento ciascuno; nel cofanetto 3 ci sono un cassetto con un orologio d'oro e un cassetto con un orologio d'argento.




Si sceglie a caso un cofanetto, si apre un cassetto e si trova un orologio d'argento. Determinare la probabilità che nel secondo cassetto dello stesso cofanetto si trovi un orologio d'oro.

**Soluzione**

Se in un cassetto è presente un orologio d'argento, il cofanetto utile può essere soltanto il numero 3, quello che contiene un orologio d'oro e uno d'argento.

Occorre quindi determinare la probabilità che il cofanetto prescelto sia il numero 3, sapendo che un cassetto contiene un orologio d'argento e che la scelta del cofanetto è stata del tutto casuale.

Si introducono gli eventi:

$$\begin{aligned} A_1 &= \{\text{scelto il cofanetto numero 1}\} \\ A_2 &= \{\text{scelto il cofanetto numero 2}\} \\ A_3 &= \{\text{scelto il cofanetto numero 3}\} \\ A &= \{\text{ aperto un cassetto con orologio d'argento }\} \end{aligned}$$

con le probabilità

$$P(A_1) = \frac{1}{3} \qquad P(A_2) = \frac{1}{3} \qquad P(A_3) = \frac{1}{3}$$

e

$$P(A|A_1) = 0 \qquad P(A|A_2) = 1 \qquad P(A|A_3) = \frac{1}{2}$$

La regola di Bayes fornisce allora

$$P(A_3|A) =$$

$$= \frac{P(A|A_3)\ P(A_3)}{P(A|A_1)\ P(A_1) + P(A|A_2)\ P(A_2) + P(A|A_3)\ P(A_3)} =$$

$$= \frac{\frac{1}{2} \cdot \frac{1}{3}}{0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} = \frac{1}{3}$$

La probabilità richiesta è quindi 1/3

# 3. VARIABILI CASUALI E DISTRIBUZIONI DI PROBABILITÀ

## □ Variabile casuale (Random variable, RV)

Una funzione reale definita su uno spazio campione $\mathbb{S}$, munito di una misura di probabilità

$$X \ : \ \xi \in \mathbb{S} \longrightarrow X(\xi) \in \mathbb{R}$$

La funzione deve essere **misurabile**, deve cioè aversi che $\forall\, x \in \mathbb{R}$ assegnato il sottoinsieme di $\mathbb{S}$

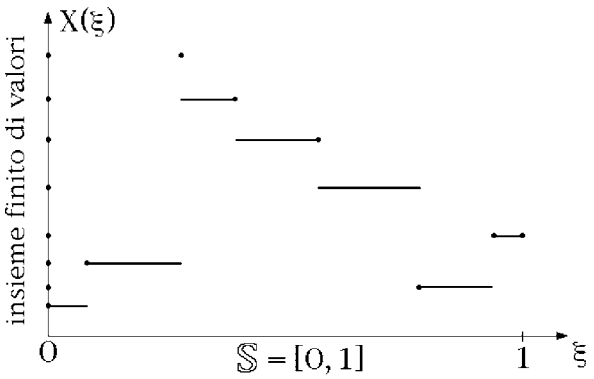
$$\{\xi \in \mathbb{S} \ : \ X(\xi) \leq x\}$$

sia misurabile in $\mathbb{S}$, e ne sia pertanto definita la misura di probabilità

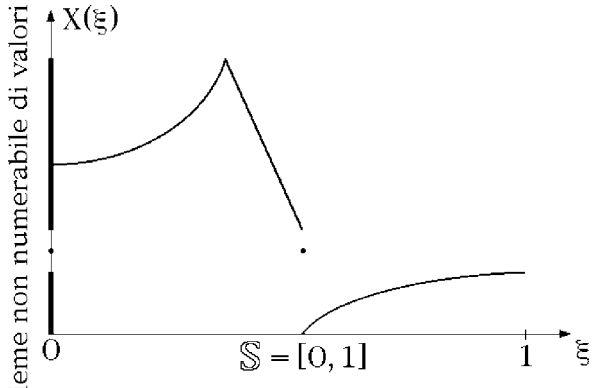
$$P\Big(\{\xi \in \mathbb{S} \ : \ X(\xi) \leq x\}\Big)$$

Si distinguono:

- **variabili casuali discrete**, suscettibili di assumere un numero finito o una infinità numerabile di valori
- **variabili casuali continue**, capaci di assumere una infinità non numerabile di valori




variabile casuale discreta




variabile casuale continua

### □ Distribuzione cumulativa di una variabile casuale

La funzione reale di una variabile reale $P : \mathbb{R} \to \mathbb{R}$ definita da

$$P(x) = P\Big(\{\xi \in \mathbb{S} \ : \ X(\xi) \leq x\}\Big) \qquad \forall\, x \in \mathbb{R}$$

fornisce per una dato $x \in \mathbb{R}$ la probabilità che la variabile casuale $X$ assuma un valore minore o uguale a $x$. Essa è detta **distribuzione cumulativa** della variabile casuale. Si ha che:

$(i)$ $P(x)$ è non decrescente

$$x < y \qquad \Longrightarrow \qquad P(x) \leq P(y)$$

in quanto $x < y$ implica che si abbia

$$\{\xi \in \mathbb{S} \ : \ X(\xi) \leq x\} \subseteq \{\xi \in \mathbb{S} \ : \ X(\xi) \leq y\}\,;$$

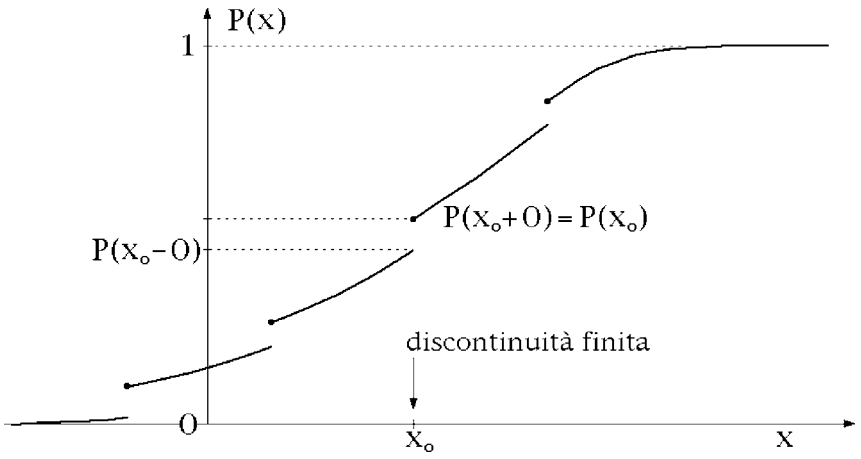
$(ii)$ il codominio di $P(x)$ è l'intervallo $[0, 1]$, in quanto

$$\lim_{x \to -\infty} P(x) = 0 \qquad \lim_{x \to +\infty} P(x) = 1\,;$$

$(iii)$ $P(x)$ è continua a destra

$$\lim_{h \to 0+} P(x + h) = P(x) \qquad \forall\, x \in \mathbb{R}\,.$$

Il grafico ha l'andamento illustrato nella figura seguente:



**□ Distribuzione di probabilità di una RV discreta**
Data una variabile casuale discreta $X : \xi \in \mathbb{S} \rightarrow X(\xi) \in \mathbb{R}$, con codominio

$$X(\mathbb{S}) = \{x_1, x_2, x_3, \ldots\} \subset \mathbb{R}$$

per definizione finito o numerabile, si definisce **distribuzione di probabilità di** $X$ la funzione reale di una variabile reale

$$p(x) = \begin{cases} P\big(\{\xi \in \mathbb{S} : X(\xi) = x\}\big) & \text{se } x \in X(\mathbb{S}) \\ \\ 0 & \text{se } x \notin X(\mathbb{S}) \end{cases}$$

La funzione $p(x)$ fornisce dunque la probabilità che la RV discreta $X$ assuma il valore $x \in \mathbb{R}$

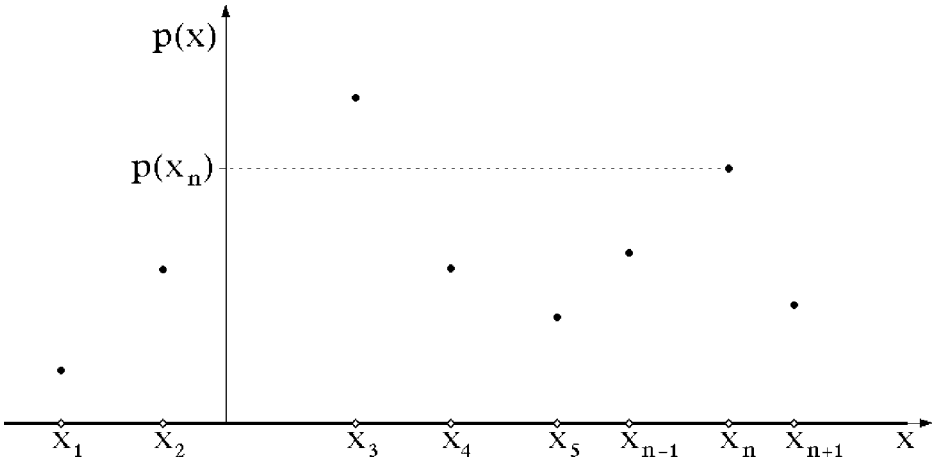
Le proprietà seguenti sono ovvie:

$(i)$ $p(x) \geq 0 \; \forall \, x \in \mathbb{R}$ (positività)

$(ii)$ $\displaystyle\sum_{x \in \mathbb{R}} p(x) = \sum_{x \in X(\mathbb{S})} p(x) = 1$ (normalizzazione)

$(iii)$ nei punti $x$ che non appartengono a $X(\mathbb{S})$ la funzione è nulla, in quanto tali valori di $x$ sono impossibili

Il grafico della distribuzione di probabilità ha quindi l'andamento illustrato nella figura seguente




## Relazione fra distribuzione di probabilità e distribuzione cumulativa di una RV discreta

Nota che sia la distribuzione di probabilità $p(x)$ di una RV discreta è facile ricavare la distribuzione cumulativa $P(x)$ corrispondente

$$P(x) = \sum_{y \leq x} p(y) \qquad \forall\, x \in \mathbb{R}$$

Per ogni $x \in \mathbb{R}$ si ha infatti

$$\begin{aligned} P(x) &= P\Big(\{\xi \in \mathbb{S} : X(\xi) \leq x\}\Big) = \\ &= P\Big(\bigcup_{y \leq x} \{\xi \in \mathbb{S} : X(\xi) = y\}\Big) = \\ &= \sum_{y \leq x} P\Big(\{\xi \in \mathbb{S} : X(\xi) = y\}\Big) = \sum_{y \leq x} p(y) \end{aligned}$$

• **Esempio**: si lancia due volte, consecutivamente e in modo indipendente, una moneta non truccata (con faccia "testa" $T$ e faccia "croce" $C$)

Lo spazio campione è finito, consistendo di 4 eventi soltanto

$$\mathbb{S} = \{TT\,,\ TC\,,\ CT\,,\ CC\}$$

che per l'ipotesi di lanci indipendenti e moneta non truccata devono assumersi equiprobabili

$$P(TT) = \frac{1}{4} \quad P(TC) = \frac{1}{4} \quad P(CT) = \frac{1}{4} \quad P(CC) = \frac{1}{4}$$

Come **variabile casuale** $X : \mathbb{S} \to \mathbb{R}$ si considera il numero di risultati "testa" ottenuti nei due lanci:

$$X(TT) = 2 \quad X(TC) = 1 \quad X(CT) = 1 \quad X(CC) = 0$$

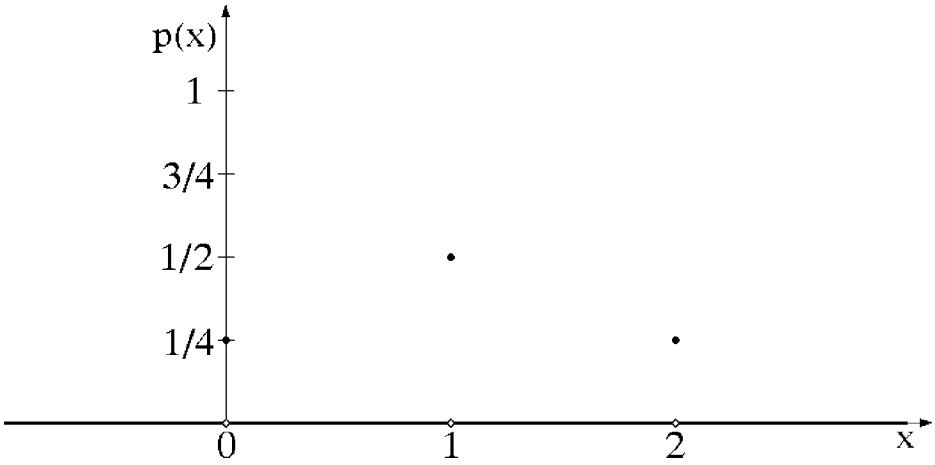
Potendo assumere soltanto tre valori (0, 1 e 2), $X$ è una RV discreta con **distribuzione di probabilità**

$$p(x) = \begin{cases} 1/4 & \text{per } x = 0 \\ 1/2 & \text{per } x = 1 \\ 1/4 & \text{per } x = 2 \end{cases}$$

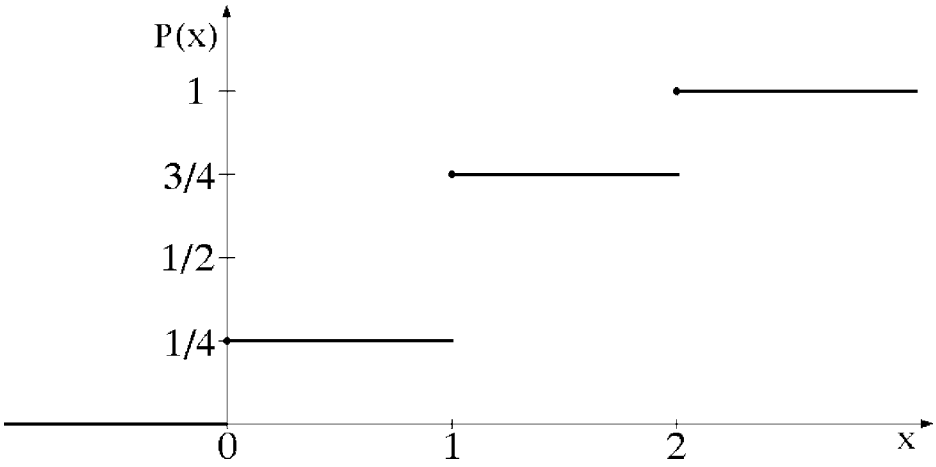
e **distribuzione cumulativa**

$$P(x) = \begin{cases} 0 & \text{per } x < 0 \\ 1/4 & \text{per } 0 \leq x < 1 \\ 3/4 & \text{per } 1 \leq x < 2 \\ 1 & \text{per } 2 \leq x \end{cases}$$

Il grafico della distribuzione di probabilità di $X$ è quindi




mentre quello della distribuzione cumulativa corrispondente assume la forma



### □ Densità di probabilità di una RV continua

Una variabile casuale continua $X$ si dice **assolutamente continua** se la sua distribuzione cumulativa può esprimersi nella forma

$$P(x) \,=\, P\big(\{\xi \in \mathbb{S} \,:\, X(\xi) \,\leq\, x\}\big) \,=\, \int_{-\infty}^{x} p(y)\, dy$$

dove la funzione $p(x)$ è detta **densità di probabilità** della variabile casuale continua e soddisfa le proprietà:

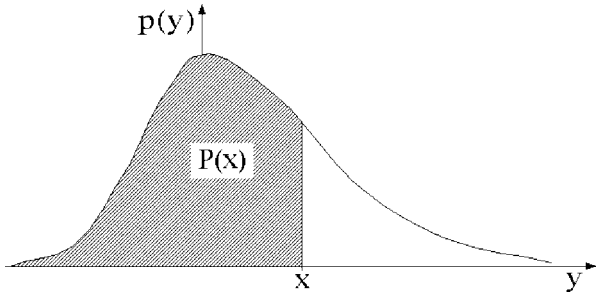
$(i)$ $p(x) \geq 0 \;\forall\, x \in \mathbb{R}$ (positività)

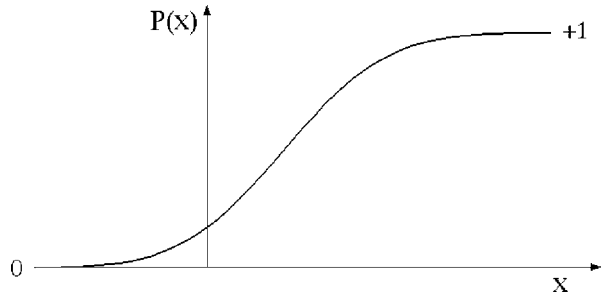
$(ii)$ $p(x)$ è integrabile in $\mathbb{R}$ (sommabilità)

$(iii)$ $\displaystyle\int_{-\infty}^{+\infty} p(x)\, dx \,=\, 1$ (normalizzazione)

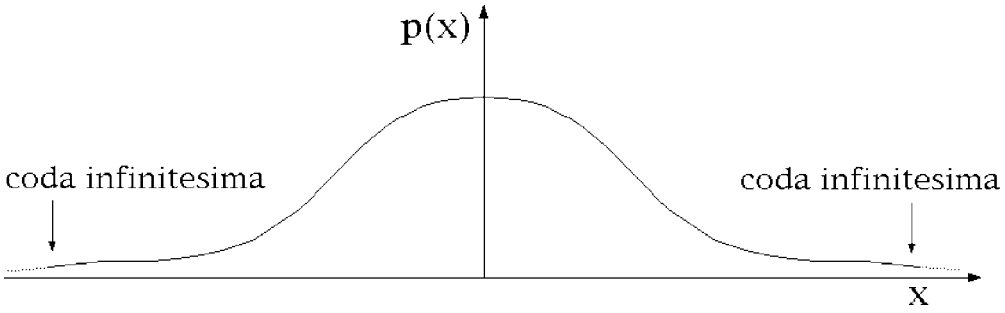
**Osservazioni**

- per definizione di integrale, $P(x) \,=\, \displaystyle\int_{-\infty}^{x} p(y)\, dy$ rappresenta l'area della regione di piano compresa fra il grafico di $p(y)$ e l'intervallo $(-\infty, x]$ dell'asse delle ascisse



- per poter soddisfare la condizione di normalizzazione $p(x)$ deve tendere a zero per $x \to -\infty$ e $x \to +\infty$ (le "code" della distribuzione sono infinitesime o nulle)



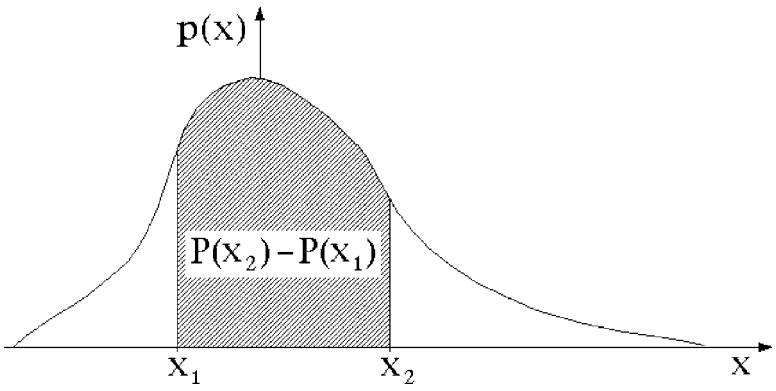

- per ogni $x_1, x_2 \in \mathbb{R}$ fissati, con $x_1 < x_2$, la differenza

$$P(x_2) - P(x_1) = \int_{x_1}^{x_2} p(y)\, dy$$

rappresenta la probabilità che la variabile casuale assuma un valore compreso fra $x_1$ e $X_2$

$$P\big(\{\xi \in \mathbb{S} : x_1 < X(\xi) \leq x_2\}\big)$$

e si identifica con l'area del trapezoide compreso fra il grafico di $p(x)$ e l'intervallo $[x_1, x_2]$ dell'asse delle ascisse



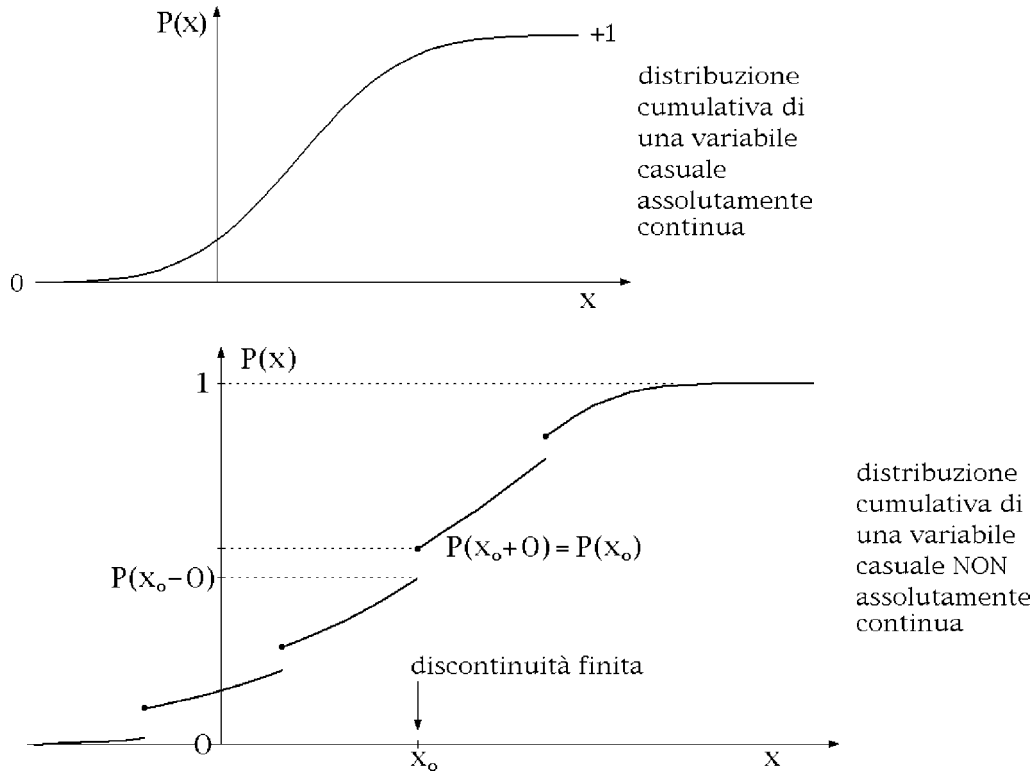
- il prodotto $p(x)\,\delta x$, con $\delta x > 0$ molto piccolo, rappresenta approssimativamente la probabilità che la variabile casuale $X$ assuma un valore compreso fra $x$ e $x + \delta x$

- il numero $p(x)$ **non esprime** quindi la probabilità che $X$ assuma il valore $x$, ma soltanto la **probabilità specifica** per unità di $x$ nell'intorno di tale valore

$$p(x) = \frac{p(x)\,\delta x}{\delta x} = \frac{P\big(\{\xi \in \mathbb{S} : x < X(\xi) \leq x + \delta x\}\big)}{\delta x}$$

  Ciò giustifica la denominazione di **densità di probabilità**

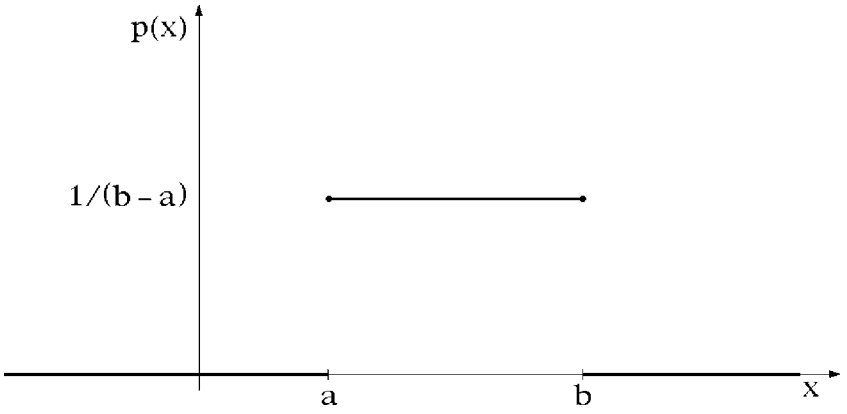
- se una variabile casuale è assolutamente continua, ammette cioè una densità di probabilità, allora la relativa distribuzione cumulativa è **continua** (non può presentare salti nel proprio grafico)
  $\Rightarrow$ esistono RV continue **non assolutamente continue**



**Esempio**: una variabile casuale assolutamente continua ha la densità di probabilità seguente

$$p(x) = \begin{cases} \dfrac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{se } x < a \text{ o } x > b \end{cases}$$

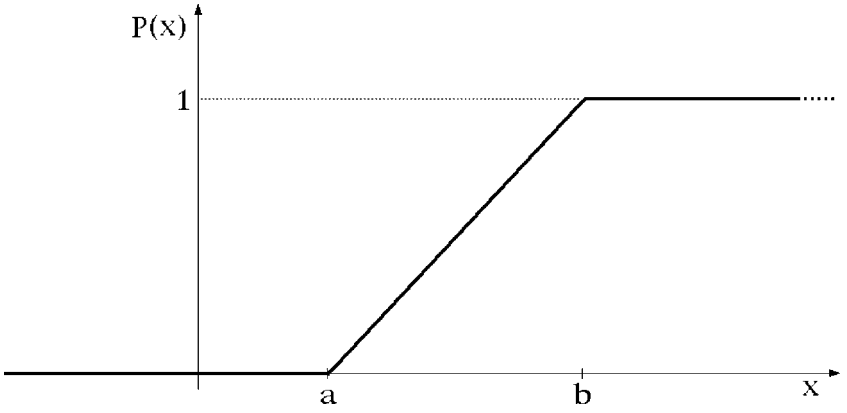
con $a, b \in \mathbb{R}$, $a < b$. Il grafico della densità di probabilità è




La distribuzione cumulativa vale allora

$$P(x) = \begin{cases} 0 & \text{se } x \leq a \\ \displaystyle\int_a^x \frac{1}{b-a}\, dy = \frac{x-a}{b-a} & \text{se } x \in [a, b] \\ 1 & \text{se } x \geq b \end{cases}$$

con il grafico



## ☐ Variabili casuali in più dimensioni (multivariate random variables, MRVs)

Una funzione **a valori vettoriali in** $\mathbb{R}^n$ definita su uno spazio campione $\mathbb{S}$, munito di una misura di probabilità,

$$X \ : \ \xi \in \mathbb{S} \longrightarrow X(\xi) \in \mathbb{R}^n$$

con la condizione che tutte le componenti

$$X_1(\xi) \ , \quad X_2(\xi) \ , \quad \dots \ , \quad X_n(\xi)$$

siano funzioni **misurabili** di $\mathbb{S}$

$X$ è una RV multivariata
**se e solo se**
tutte le sue componenti $X_1, X_2, \dots, X_n$ sono RV

Ci si può riferire a $X$, indifferentemente, come a:

- una variabile casuale $n$-dimensionale
- una variabile casuale multidimensionale
- una variabile casuale multivariata
- un set di $n$ variabili casuali
- un sistema di $n$ variabili casuali

Le MRVs possono essere **discrete** o **continue**

## □ Distribuzione di probabilità congiunta di una MRV

La distribuzione di probabilità congiunta (o $n$-dimensionale)

$$p(x_1, x_2, \ldots, x_n)$$

delle $n$ variabili casuali $X_1, X_2, \ldots, X_n$:

($i$) **nel caso discreto** descrive la probabilità che le $n$ variabili assumano **simultaneamente** i valori indicati in argomento (evento composto);

($ii$) **nel caso continuo** individua, per mezzo dell'integrale $n$-dimensionale

$$\int_\Omega p(x_1, x_2, \ldots, x_n)\ dx_1 dx_2 \ldots dx_n$$

la probabilità che il vettore di variabili casuali $(x_1, x_2, \ldots, x_n)$ assuma un qualsiasi valore compreso entro la regione di integrazione $\Omega \subseteq \mathbb{R}^n$.

In ambo i casi la distribuzione soddisfa una condizione di **normalizzazione**, che nel caso continuo assume la forma

$$\int_{\mathbb{R}^n} p(x_1, x_2, \ldots, x_n)\ dx_1 dx_2 \ldots dx_n\ =\ 1$$

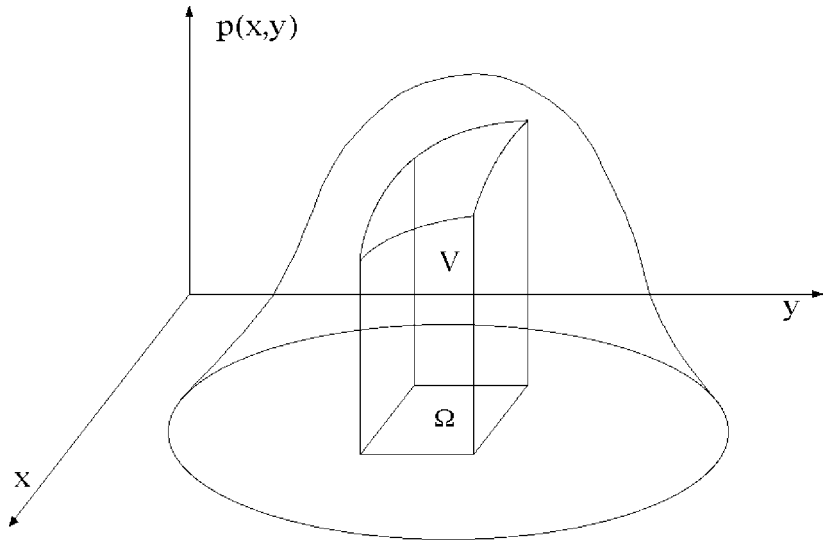
mentre una espressione analoga vale per variabili discrete

Nel **caso continuo** la distribuzione di probabilità congiunta ha il significato di una **densità di probabilità**

La probabilità che una coppia di variabili casuali $(X, Y)$ assuma un qualsiasi valore nell'insieme $\Omega$ del piano $\mathbb{R}^2$

$$P\big(\{\xi \in \mathbb{S} \,:\, (X(\xi), Y(\xi)) \in \Omega\}\big) \;=\; \int\limits_{\Omega} p(x,y)\, dxdy$$

rappresenta il volume $V$ della regione di spazio $\mathbb{R}^3$ compresa fra il grafico di $p(x,y)$ e il sottoinsieme $\Omega$ nel piano $xy$




In particolare, per $\delta x$ e $\delta y$ positivi e piccoli il prodotto

$$p(x,y)\, \delta x\, \delta y$$

esprime **approssimativamente** la probabilità che le variabili casuali $X, Y$ assumano un valore compreso nel rettangolo

$$x \,\leq\, X \,\leq\, x + \delta x \qquad\qquad y \,\leq\, Y \,\leq\, y + \delta y$$

### □ Variabili casuali (stocasticamente) indipendenti

Le $n$ variabili casuali $X_1, X_2, \ldots, X_n$ si dicono **indipendenti** se la loro distribuzione di probabilità congiunta si fattorizza in un prodotto di $n$ distribuzioni ciascuna relativa ad una sola variabile

$$p(x_1, x_2, \ldots, x_n) \;=\; p_1(x_1)\, p_2(x_2) \;\ldots\; p_n(x_n)$$

Il realizzarsi di un determinato valore di una variabile del sistema non influenza la distribuzione di probabilità congiunta delle variabili residue (probabilità condizionata)

**Esempio:** la distribuzione di probabilità congiunta

$$p(x_1, x_2) \;=\; \frac{1}{2\pi\sigma_1\sigma_2}\, e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}} \;=$$

$$= \;\frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \;\cdot\; \frac{1}{\sqrt{2\pi}\,\sigma_2}\, e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}$$

descrive una coppia di variabili casuali indipendenti, $X_1$ e $X_2$

### □ Variabili casuali (stocasticamente) dipendenti

Sono variabili casuali non indipendenti. La loro distribuzione di probabilità congiunta non è fattorizzabile come nel caso delle variabili indipendenti.

**Esempio:**
le variabili $X$ e $Y$ aventi la distribuzione congiunta

$$p(x, y) \;=\; \frac{3}{2\pi}\, e^{-x^2 - xy - y^2}$$

sono stocasticamente dipendenti

# 4. VALORI MEDI E DISPERSIONE

In molte applicazioni delle RV ha interesse determinare alcuni **parametri numerici caratteristici** atti ad esprimere:

- l'ubicazione, o collocazione, complessiva
- il livello di dispersione
- il grado di simmetria
- altre caratteristiche generali

della relativa distribuzione di probabilità

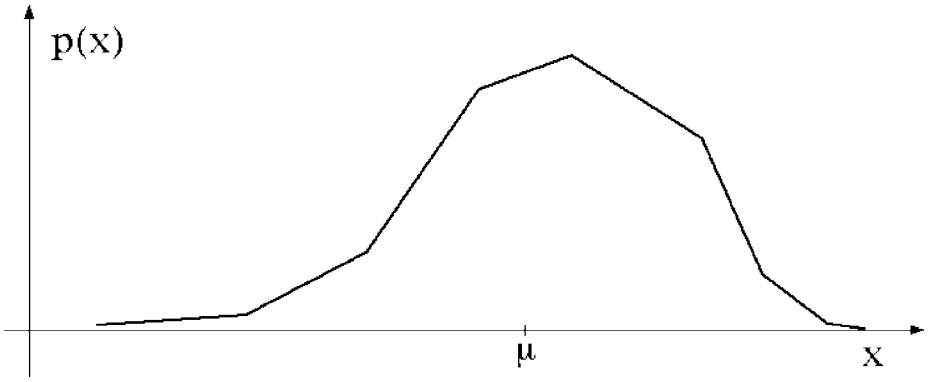
### □ Media di una variabile casuale (distribuzione)

Specifica il c.d. valor medio della variabile/distribuzione

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{+\infty} p(x)\, x\, dx \qquad \text{RV continua}$$

$$= \sum_{x_i \in X(\mathbb{S})} x_i\, p(x_i) \qquad \text{RV discreta}$$

ossia il valore della retta reale attorno al quale la distribuzione può considerarsi centrata

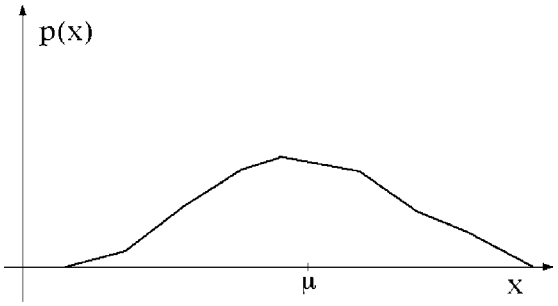


### □ Varianza di una variabile casuale (distribuzione)

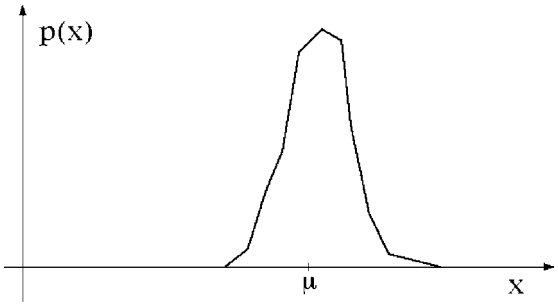
Indica quanto la distribuzione della variabile casuale sia "concentrata" nell'intorno della sua media

$$\sigma^2 = \mathbb{E}[(X - \mathbb{E}(X))^2] = \int_{-\infty}^{+\infty} p(x)\,(x-\mu)^2\,dx \quad \text{RV continua}$$

$$= \sum_{x_i \in X(\mathbb{S})} p(x_i)\,(x_i - \mu)^2 \quad \text{RV discreta}$$




grande varianza



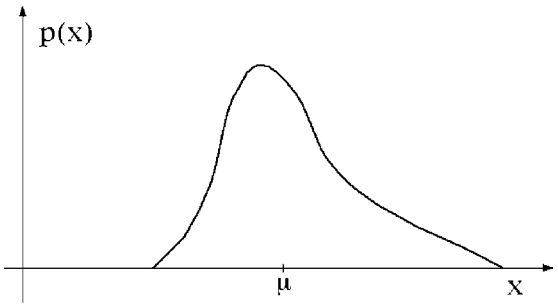

piccola varianza

### □ Skewness di una variabile casuale (distribuzione)

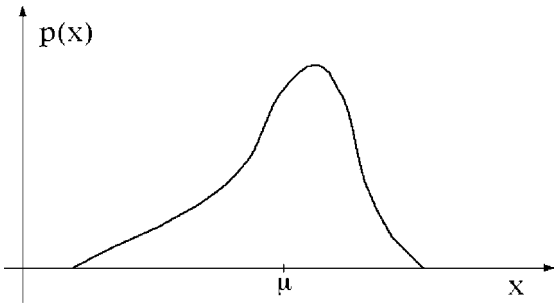
Misura il grado di simmetria di $p(x)$ attorno alla media

$$\alpha_3 = \frac{\mathbb{E}[(X - \mathbb{E}(X))^3]}{\sigma^3} = \frac{1}{\sigma^3} \int_{-\infty}^{+\infty} p(x)\,(x-\mu)^3\,dx$$

$$= \frac{1}{\sigma^3} \sum_{x_i \in X(\mathbb{S})} p(x_i)\,(x_i - \mu)^3$$




skewness positiva




skewness negativa

### □ Deviazione standard di una variabile casuale (distribuzione)

È la radice quadrata della varianza

$$\sigma \;=\; \sqrt{\sigma^2} \;=\; \mathbb{E}[(X-\mathbb{E}(X))^2]^{1/2} \;=\; \left\{\int_{-\infty}^{+\infty} p(x)\,(x-\mu)^2\,dx\right\}^{1/2}$$

Ha un significato analogo a quello della varianza, ma diversamente da questa è una **grandezza omogenea** a $x$ e $\mu$

### □ Teorema di Tchebyshev:

La probabilità che una variabile casuale continua assuma valori in un intervallo centrato sulla media $\mu$ e di semiampiezza pari a $k\sigma$, con $k > 0$ fissato non è mai inferiore a $1 - (1/k^2)$

$$p[\mu - k\sigma \le x \le \mu + k\sigma] \;=\; \int_{\mu-k\sigma}^{\mu+k\sigma} p(x)\,dx \ge 1 - \frac{1}{k^2}$$

**Dim.**

$$\begin{aligned}
\sigma^2 \;&=\; \int_{-\infty}^{+\infty} p(x)(x-\mu)^2 dx \ge \\
&\ge\; \int_{-\infty}^{\mu-k\sigma} p(x)(x-\mu)^2 dx + \int_{\mu+k\sigma}^{+\infty} p(x)(x-\mu)^2 dx \ge \\
&\ge\; \int_{-\infty}^{\mu-k\sigma} p(x)k^2\sigma^2 dx + \int_{\mu+k\sigma}^{+\infty} p(x)k^2\sigma^2 dx \;= \\
&=\; k^2\sigma^2\left[1 - \int_{\mu-k\sigma}^{\mu+k\sigma} p(x)\,dx\right] \\
&\Longrightarrow \qquad \frac{1}{k^2} \ge 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} p(x)\,dx \qquad \square
\end{aligned}$$

### Moda

Per una RV discreta è il valore che ha localmente la massima probabilità di verificarsi

Per una RV continua è il valore dove la densità di probabilità presenta un massimo relativo

Una distribuzione/variabile casuale può presentare **più mode** (distribuzione bimodale, trimodale, multimodale)

### Mediana

È il valore $m$ di una variabile casuale continua $X$ tale che

$$P\big(\{\xi \in \mathbb{S} : X(\xi) \leq m\}\big) = \frac{1}{2} = P\big(\{\xi \in \mathbb{S} : X(\xi) > m\}\big)$$

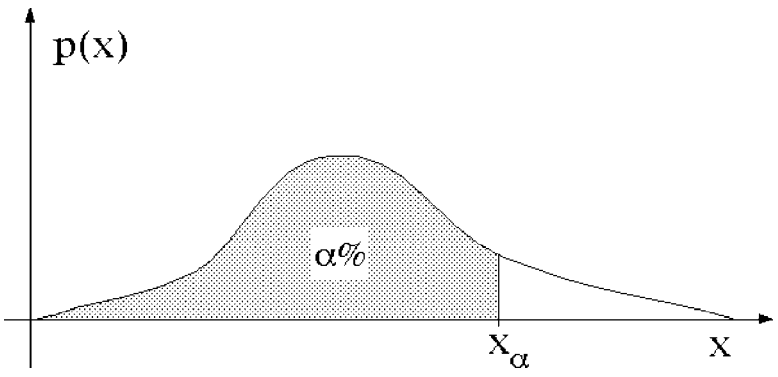
La RV ha quindi la stessa probabilità — $1/2$ — di assumere un valore $\leq m$ oppure $> m$

### Percentili

L'$\alpha$-esimo percentile di una RV continua $X$, con $\alpha$ intero fra 0 e 100, è il valore $x_\alpha$ della variabile casuale per il quale risulta

$$P\big(\{\xi \in \mathbb{S} : X(\xi) \leq x_\alpha\}\big) = \alpha/100$$

L'area compresa fra il grafico della distribuzione di probabilità e l'intervallo $(-\infty, x_\alpha]$ dell'asse delle ascisse, vale quindi $\alpha/100$



### Media, covarianza e correlazione di un set di variabili casuali

Dato un sistema $(X_1, X_2, \ldots, X_n)$ di variabili casuali, si definiscono le grandezze di seguito specificate

### Media di $X_i$

È il valor medio della variabile casuale $X_i$ del sistema

$$\mu_i \;=\; \mathbb{E}(X_i) \;=\; \int_{\mathbb{R}^n} x_i \, p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n$$

### Varianza di $X_i$

È la varianza della variabile casuale $X_i$ del sistema

$$\begin{aligned} \mathrm{var}(X_i) \;&=\; \mathbb{E}[(X_i - \mu_i)^2] \;= \\ &= \int_{\mathbb{R}^n} (x_i - \mu_i)^2 \, p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n \end{aligned}$$

La sua radice quadrata definisce la **deviazione standard** della variabile $X_i$

### Covarianza di $X_i$ e $X_j$, $i \neq j$

È l'aspettazione (valor medio) del prodotto degli scarti delle variabili casuali $X_i$ e $X_j$ rispetto alle relative medie

$$\begin{aligned} \mathrm{cov}(X_i, X_j) \;&=\; \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \;= \\ &= \int_{\mathbb{R}^n} (x_i - \mu_i)(x_j - \mu_j)\, p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n \end{aligned}$$

### □ Matrice di covarianza del sistema

È la matrice $C$, $n \times n$, reale e simmetrica, definita da

$$C_{ij} = \begin{cases} \text{cov}(X_i, X_j) & \text{se } i \neq j \\ \text{var}(X_i) & \text{se } i = j \end{cases} \qquad i, j = 1, 2, \ldots, n$$

I suoi elementi diagonali si identificano quindi con le varianze delle singole variabili casuali del sistema, mentre gli elementi non diagonali corrispondono alle covarianze fra tutte le possibili coppie delle stesse variabili

### □ Correlazione di $X_i$ e $X_j$, $i \neq j$

È la covarianza fra le variabili $X_i$ e $X_j$ "normalizzata" secondo le deviazioni standard delle stesse variabili

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)}\sqrt{\text{var}(X_j)}}$$

in modo da definire una grandezza adimensionale, indipendente dalle unità di misura in cui si esprimono $X_i$ e $X_j$

Dalla diseguaglianza di Cauchy-Schwarz segue che $\text{corr}(X_i, X_j)$ è sempre un numero compreso fra $-1$ e $+1$

La correlazione di una variabile con sè stessa non è significativa, risultando identicamente $\text{corr}(X_i, X_i) = 1 \; \forall \, i = 1, \ldots, n$

### □ Matrice di correlazione del sistema

È la matrice $V$, reale, simmetrica, $n \times n$, data da

$$V_{ij} = \text{corr}(X_i, X_j) \qquad i, j = 1, 2, \ldots, n$$

con gli elementi diagonali tutti uguali a 1

### □ Variabili casuali non correlate (o scorrelate)

Sono variabili casuali a correlazione nulla

$$\begin{aligned} X_i \text{ e } X_j \text{ scorrelate} &\iff \operatorname{corr}(X_i, X_j) = 0 \\ (X_1, \ldots, X_n) \text{ scorrelate} &\iff V \text{ diagonale} \\ &\iff C \text{ diagonale} \end{aligned}$$

**Teorema**

Due variabili casuali $X_i$ e $X_j$ sono scorrelate se e soltanto se il valor medio del loro prodotto è uguale al prodotto dei valori medi delle singole variabili.

Il risultato segue dall'identità

$$\begin{aligned} \operatorname{cov}(X_i, X_j) &= \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \\ &= \mathbb{E}[X_i X_j - \mu_i X_j - X_i \mu_j + \mu_i \mu_j] = \\ &= \mathbb{E}[X_i X_j] - \mathbb{E}[\mu_i X_j] - \mathbb{E}[X_i \mu_j] + \mathbb{E}[\mu_i \mu_j] = \\ &= \mathbb{E}[X_i X_j] - \mu_i \mathbb{E}[X_j] - \mathbb{E}[X_i] \mu_j + \mathbb{E}[\mu_i \mu_j] = \\ &= \mathbb{E}[X_i X_j] - \mu_i \mu_j - \mu_i \mu_j + \mu_i \mu_j = \\ &= \mathbb{E}[X_i X_j] - \mu_i \mu_j \end{aligned}$$

**Teorema**

L'indipendenza stocastica è **condizione sufficiente** ma **non necessaria** alla mancanza di correlazione fra due o più variabili casuali.

Se $X_i$ e $X_j$ sono indipendenti si ha infatti

$$\begin{aligned} \operatorname{cov}(X_i, X_j) &= \\ &= \int_{\mathbb{R}^2} (x_i - \mu_i)(x_j - \mu_j)\, p_i(x_i)\, p_j(x_j)\, dx_i dx_j = \\ &= \int_{\mathbb{R}} (x_i - \mu_i)\, p_i(x_i)\, dx_i \int_{\mathbb{R}} (x_j - \mu_j)\, p_j(x_j)\, dx_j = 0 \end{aligned}$$

# 5. DISTRIBUZIONI DI PROBABILITÀ SPECIALI

## 5.1 Esempi di distribuzioni di probabilità discrete

### □ Distribuzione di Bernoulli

La variabile casuale può assumere soltanto un numero finito di valori reali $x_i$, $i = 1, \ldots, n$ con le rispettive probabilità $w_i = p(x_i)$. Vale la condizione di normalizzazione $\sum_{i=1}^{n} w_i = 1$. Inoltre:

$$\mu = \sum_{i=1}^{n} w_i x_i \qquad \sigma^2 = \sum_{i=1}^{n} w_i (x_i - \mu)^2$$

**Esempio**: la variabile casuale che nel lancio di una moneta non truccata assume il valore 0 in caso esca testa e il valore 1 in caso esca croce è una variabile di Bernoulli con valori $(0, 1)$ e probabilità $(1/2, 1/2)$

### □ Distribuzione binomiale

È una distribuzione a due parametri della forma

$$p_{n,p}(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

con $n \in \mathbb{N}$ e $p \in (0, 1)$ assegnati. Media e varianza sono date da:

$$\mu = np \qquad \sigma^2 = np(1-p)$$

Nasce in modo naturale dal considerare $n$ variabili Bernoulli indipendenti $(x_1, x_2, \ldots, x_n)$, identicamente distribuite, con valori $\{1, 0\}$ e probabilità rispettive $p$, $1-p$. In tal caso la somma $x = \sum_{i=1}^{n} x_i$ segue la distribuzione binomiale suindicata.

Equivalentemente, si possono considerare $n$ prove ripetute indipendenti, ciascuna delle quali può produrre due soli risultati:

- un **risultato utile**, con probabilità $p$,
- e un **risultato alternativo**, con probabilità $1 - p$

La probabilità che si verifichi **una sequenza assegnata** di $x$ risultati utili e $n - x$ alternativi è data dal prodotto delle probabilità degli eventi parziali

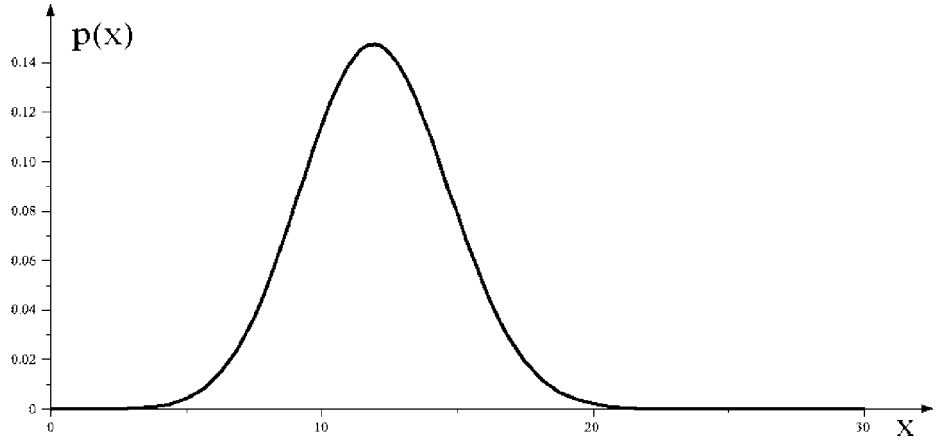
$$\underbrace{p \cdot p \cdot \ldots p}_{x \text{ volte}} \cdot \underbrace{(1-p) \cdot (1-p) \cdot \ldots \cdot (1-p)}_{n-x \text{ volte}}$$

mentre l'arbitrarietà dell'ordine con cui detti eventi possono verificarsi giustifica l'introduzione del coefficiente binomiale

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

che esprime il numero delle possibili permutazioni degli $x$ eventi utili sulle $n$ prove

Il grafico della distribuzione binomiale per $n = 30$ e $p = 0.4$ è illustrato nella figura seguente



### □ Distribuzione di Poisson

È una distribuzione discreta ad un solo parametro

$$p_m(x) \;=\; e^{-m}\,\frac{m^x}{x!} \qquad x = 0, 1, 2, \ldots$$

con $m$ costante positiva assegnata. Uguali valori di media e varianza:

$$\mu \;=\; m \qquad\qquad \sigma^2 \;=\; m$$

Può ricavarsi come **limite della distribuzione binomiale** ponendo

$$np \;=\; m\;, \quad \text{costante}\,,$$

e prendendo $n \to +\infty$

Si può considerare perciò come la distribuzione di probabilità di ottenere $x$ risultati utili su una sequenza molto lunga di $n \gg 0$ prove, nell'ipotesi che il singolo evento utile avvia una probabilità $p$ molto piccola

**Esempio**: in un campione di radionuclide la probabilità $p$ che un nucleo decada in un intervallo di tempo assegnato $\Delta T$ è **costante** e tipicamente **molto piccola**. I decadimenti dei vari nuclei sono inoltre eventi indipendenti.
Il numero $x$ di decadimenti osservato nell'intervallo di tempo $\Delta T$ su un campione di $n \gg 0$ nuclei radioattivi è quindi una variabile di Poisson con valor medio

$$m \;=\; np$$

## 5.2 Esempi di distribuzioni di probabilità continue

### □ Distribuzione uniforme

È descritta da una densità di probabilità della forma

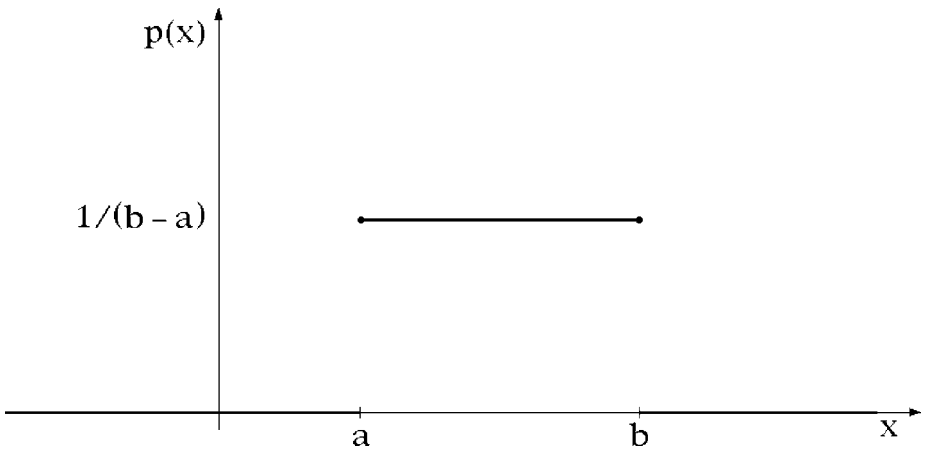
$$p(x) = \frac{1}{b-a} \begin{cases} 1 & \text{se } x \in [a,b] \\ 0 & \text{altrimenti} \end{cases} \qquad x \in \mathbb{R}$$

Tutti i valori della variabile casuale $X$ entro l'intervallo $[a,b]$ sono ugualmente probabili, gli altri impossibili.

Media e varianza sono immediate:

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

Il grafico della distribuzione uniforme è la funzione caratteristica dell'intervallo $(a,b)$



### □ Distribuzione normale o gaussiana

È caratterizzata dalla densità di probabilità

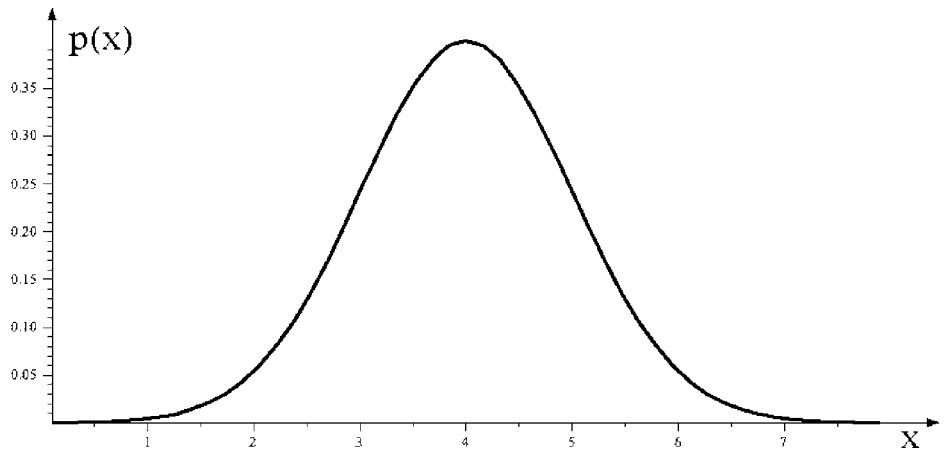
$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad x \in \mathbb{R}$$

con $\mu \in \mathbb{R}$ e $\sigma > 0$ costanti assegnate.

Il parametro $\mu$ si identifica con la media,
mentre $\sigma^2$ è la varianza e $\sigma$ la deviazione standard.

La media $\mu$ coincide con la mediana e con la moda

Il grafico ha l'andamento illustrato nella figura seguente




Una variabile normale di media $\mu$ e deviazione standard $\sigma$ viene talora indicata con $N(\mu, \sigma)$

La variabile casuale $z = (x - \mu)/\sigma$ segue ancora una distribuzione gaussiana, ma con media nulla e varianza unitaria:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad z \in \mathbb{R}$$

Tale distribuzione (o variabile) normale si dice **standard**

**Osservazione. Rilevanza delle distribuzioni gaussiane**

Gli errori casuali che affliggono molti tipi di misure sperimentali possono assumersi seguire una distribuzione gaussiana di media e varianza opportune.

Questa relativa ubiquità delle distribuzioni gaussiane può essere in parte giustificata per mezzo del noto:

**Teorema del limite centrale (Lindeberg-Lévy-Turing)**
Sia $(X_n)_{n\in\mathbb{N}} = X_1, X_2, \ldots$ una successione di variabili casuali indipendenti, identicamente distribuite, con varianza finita $\sigma^2$ e media $\mu$.
Per ogni $n \in \mathbb{N}$ fissato si consideri la variabile casuale

$$Z_n \;=\; \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

la cui distribuzione di probabilità si indicherà con $p_n(z_n)$.
Vale allora che

$$p_n(z) \quad \xrightarrow[n\to+\infty]{} \quad N(0,1)(z) \qquad \forall z \in \mathbb{R}$$

essendo $N(0,1)(z)$ la distribuzione gaussiana standard

$$N(0,1)(z) \;=\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad z \in \mathbb{Z}$$

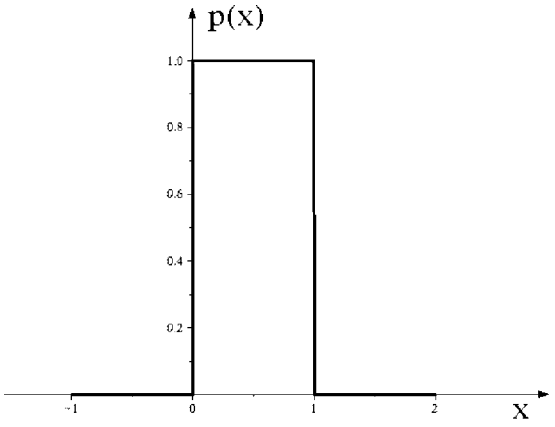
**Note**:
Si dice che la successione di variabili casuali $(Z_n)_{n\in\mathbb{N}}$ **converge in distribuzione** alla variabile gaussiana standard $Z$.

La condizione che le variabili siano identicamente distribuite può essere indebolita ($\Longrightarrow$ risultato più generale).

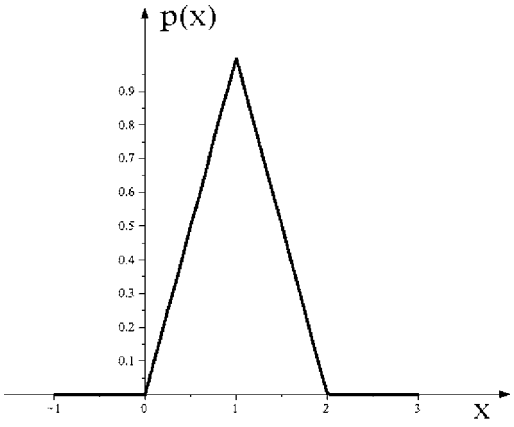
Una somma di $n$ RV continue indipendenti e identicamente distribuite segue una distribuzione **approssimativamente normale** per $n$ abbastanza grande

Tipicamente ciò avviene per $n$ dell'ordine di qualche decina, **ma può bastare molto meno**
(dipende dalla comune distribuzione di tutte le variabili)

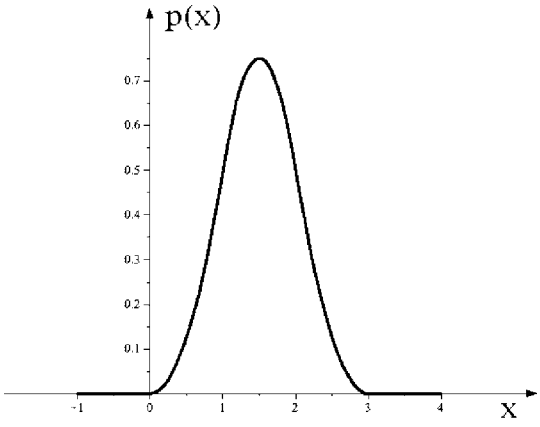
**Esempio**: RV con distribuzione uniforme in $[0, 1]$
Sommando 1, 2, 3, 4 variabili indipendenti dello stesso tipo si ottengono RV con le seguenti distribuzioni di probabilità



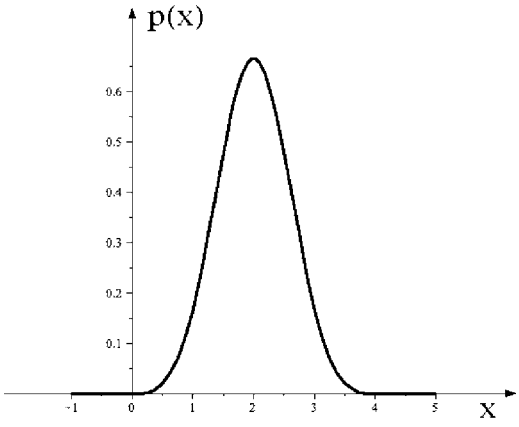

1 variabile




2 variabili




3 variabili




4 variabili

### □ Distribuzione di chi quadrato a $n$ gradi di libertà

Una variabile di chi quadrato a $n$ gradi di libertà viene indicata convenzionalmente con $\mathcal{X}^2$ e la sua distribuzione è

$$p_n(\mathcal{X}^2) = \frac{1}{\Gamma(n/2)\,2^{n/2}} e^{-\mathcal{X}^2/2} (\mathcal{X}^2)^{\frac{n}{2}-1} \qquad \mathcal{X}^2 \geq 0$$

dove $\Gamma$ indica la **funzione Gamma di Eulero**:

$$\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1}\, dx \qquad a > 0$$

a sua volta legata alla **funzione fattoriale**:

$$a! = \Gamma(a+1) = \int_0^{+\infty} e^{-x} x^{a}\, dx \qquad a > -1$$

Dalla relazione di ricorrenza caratteristica della funzione $\Gamma$

$$\Gamma(a+1) = a\Gamma(a) \quad \forall a > 0$$

e dai seguenti valori notevoli

$$\Gamma(1) = 1 \qquad \Gamma(1/2) = \sqrt{\pi}$$

si deduce che

$$\Gamma(n+1) = n! \qquad n = 0, 1, 2, \ldots$$

$$\Gamma(n+1) = n(n-1)(n-2)\ldots(3/2)(1/2)\sqrt{\pi} \quad n = 1/2, 3/2, \ldots$$

Una variabile di $\mathcal{X}^2$ a $n$ gradi di libertà è la somma dei quadrati di $n$ variabili gaussiane standard indipendenti $z_1, z_2, \ldots, z_n$

$$\mathcal{X}^2 = \sum_{i=1}^{n} z_i^2 \qquad p(z_i) = \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \quad i = 1, \ldots, n$$

### □ Distribuzione di Student a $n$ gradi di libertà

Una variabile di Student a $n$ gradi di libertà viene indicata con $t$ ($t$ di Student). La sua distribuzione vale

$$p_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \qquad t \in \mathbb{R}$$

La distribuzione ha media nulla e risulta simmetrica rispetto al valor medio.

Per $n$ grandi la distribuzione si approssima ad una gaussiana standard, cui tende nel limite di $n \to +\infty$.

La $t$ di Student è esprimibile come rapporto

$$t = \sqrt{n}\frac{z}{\sqrt{\mathcal{X}^2}}$$

essendo:

$z$ una **variabile gaussiana standard**;

$\mathcal{X}^2$ una **variabile di chi quadrato** a $n$ gradi di libertà;

le due variabili $z$ e $\mathcal{X}^2$ **stocasticamente indipendenti**

$$p(z, \mathcal{X}^2) = p(z)\, p_n(\mathcal{X}^2)$$

### □ Distribuzione di Fisher a $(n_1, n_2)$ gradi di libertà

La $F$ di Fisher a $(n_1, n_2)$ gradi di libertà segue la distribuzione di probabilità

$$p_{(n_1,n_2)}(F) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}}$$

con $F \geq 0$.

La variabile si scrive nella forma

$$F = \frac{n_2}{n_1}\frac{\mathcal{X}_1^2}{\mathcal{X}_2^2}$$

dove:

$\mathcal{X}_1^2$ è una variabile di **chi quadrato a $n_1$ gradi di libertà**;

$\mathcal{X}_2^2$ è una variabile di **chi quadrato a $n_2$ gradi di libertà**;

le variabili $\mathcal{X}_1^2$ e $\mathcal{X}_2^2$ sono **stocasticamente indipendenti**

$$p(\mathcal{X}_1^2, \mathcal{X}_2^2) = p_{n_1}(\mathcal{X}_1^2)\, p_{n_2}(\mathcal{X}_2^2)$$

Si ha, in particolare, che $\forall\,(n_1, n_2) \in \mathbb{N}^2$ il reciproco di una $F$ a $(n_1, n_2)$ gradi di libertà si distribuisce come una $F$ di Fisher a $(n_2, n_1)$ gradi di libertà:

$$\frac{1}{F_{(n_1,n_2)}} = F_{(n_2,n_1)}$$

### □ Variabili gaussiane (o normali) multidimensionali

Un vettore $X^T = (X_1, X_2, \ldots, X_n)$ di variabili casuali costituisce un sistema di variabili gaussiane multidimensionali se la distribuzione di probabilità congiunta è del tipo

$$p(x_1, \ldots, x_n) \;=\; \frac{(\det A)^{1/2}}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}(x-\mu)^T A (x-\mu)} \;, \quad x \;=\; \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

essendo $A$ una matrice $n \times n$ reale simmetrica definita positiva (matrice di struttura), $\det A > 0$ il suo determinante e $\mu \in \mathbb{R}^n$ un vettore colonna arbitrario (l'esponente $^T$ marca i trasposti di matrici e vettori colonna).

**Media di $x$:**

$$\mathbb{E}(x) \;=\; \int_{\mathbb{R}^n} x\, \frac{(\det A)^{1/2}}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}(x-\mu)^T A (x-\mu)}\, dx_1 \ldots dx_n \;=\; \mu$$

**Matrice di covarianza di $x$:**
coincide con l'inversa della matrice di struttura

$$C \;=\; A^{-1}$$

**Teorema**
Per variabili gaussiane multidimensionali, l'indipendenza stocastica è condizione equivalente alla mancanza di correlazione

$$(x_1, \ldots, x_n) \text{ scorrelate} \quad \Longleftrightarrow \quad C \text{ diagonale} \quad \Longleftrightarrow$$

$$\Longleftrightarrow \quad A \text{ diagonale} \quad \Longleftrightarrow \quad (x_1, \ldots, x_n) \text{ indipendenti}$$

**Esempio**: Si introducono il vettore dei valori medi e la matrice di struttura

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2 \qquad A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

la seconda essendo reale, simmetrica e definita positiva (entrambi gli autovalori sono positivi)

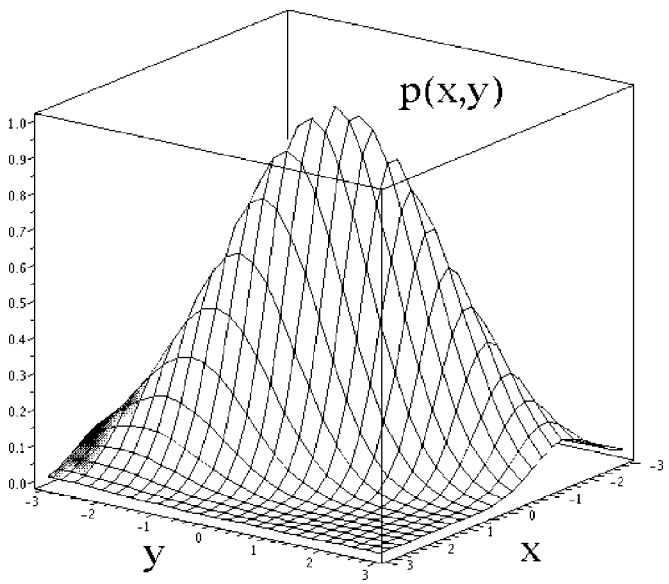
$$\det(A - \lambda \mathbb{I}) = \det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 1-\lambda \end{pmatrix} = \lambda^2 - 3\lambda + 1 = 0$$

$$\Longrightarrow \qquad \lambda = \lambda_\pm = \frac{3 \pm \sqrt{5}}{2} > 0$$

La distribuzione della RV normale bivariata $(X, Y)$ si scrive

$$p(x,y) = \frac{\sqrt{1}}{(2\pi)^{2/2}} \, e^{-\frac{1}{2}(x \;\; y) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}} = \frac{e^{-\frac{1}{2}(2x^2 + 2xy + y^2)}}{2\pi}$$

e ha il grafico mostrato in figura



## Forme quadratiche di variabili normali indipendenti standard

**Teorema di Craig**
Sia $X$ un vettore di $n$ variabili gaussiane indipendenti e standard. Date due matrici reali simmetriche $A$ e $B$ si considerino le forme quadratiche

$$Q_1 = X^T AX \qquad Q_2 = X^T BX$$

Allora le variabili casuali $Q_1$ e $Q_2$ sono stocasticamente indipendenti se e solo se $AB = 0$.

**Teorema**
**di caratterizzazione delle forme quadratiche di variabili normali indipendenti e standard che presentano una distribuzione di chi quadrato**
Sia $X^T AX$ una forma quadratica semidefinita positiva delle variabili normali indipendenti e standard $(X_1, X_2, \ldots, X_n) = X$. Allora la variabile casuale $X^T AX$ risulta una variabile di $\mathcal{X}^2$ se e soltanto se la matrice $A$ è idempotente

$$A^2 = A$$

In tal caso, indicato con $p$ il rango della matrice $A$, il numero di gradi di libertà di $X^T AX$ è pari a $p$.

**Esempio**: siano date le variabili $X_1$ e $X_2$, normali standard e indipendenti. Si considerano le RV

$$Q_1 = 2X_1X_2 = (X_1\ X_2)\, A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$Q_2 = X_1^2 - 4X_1X_2 + X_2^2 = (X_1\ X_2)\, B \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

con

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{e} \qquad B = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$$

Si ha che:

$$AB = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \neq 0$$

e quindi le RV $Q_1$ e $Q_2$ sono **stocasticamente dipendenti** per il teorema di Craig.

D'altra parte vale

$$A^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \neq A$$

e analogamente

$$B^2 = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix} \neq B$$

per cui **né** $Q_1$ **né** $Q_2$ **sono variabili di** $\mathcal{X}^2$

**Teorema di Fisher-Cochran**
Sia $(x_1, x_2, \ldots, x_n) = x$ un insieme di variabili normali indipendenti e standard. Siano $Q_1, Q_2, \ldots, Q_k$ forme quadratiche semidefinite positive delle variabili $(x_1, x_2, \ldots, x_n)$

$$Q_r = x^T A_r x = \sum_{i,h=1}^{n} (A_r)_{ih} x_i x_h \qquad r = 1, 2, \ldots, k$$

con matrici rappresentative $A_1, A_2, \ldots, A_k$, di ranghi rispettivi $n_1, n_2, \ldots, n_k$. Valga inoltre

$$\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{k} Q_j$$

Allora le $Q_1, Q_2, \ldots, Q_k$ sono variabili di $\mathcal{X}^2$ mutualmente indipendenti se e soltanto se

$$\sum_{j=1}^{k} n_j = n$$

e in tal caso $Q_r$ ha $n_r$ gradi di libertà, $\forall\, r = 1, \ldots k$

**Casi notevoli a due variabili**

($i$) Siano

$$Q_1 = x^T A_1 x \qquad Q_2 = x^T A_2 x$$

due forme quadratiche semidefinite positive delle variabili gaussiane indipendenti e standard $(x_1, x_2, \ldots, x_n) = x$. Le variabili casuali $Q_1$ e $Q_2$ soddisfino la condizione

$$Q_1 + Q_2 = \sum_{i=1}^{n} x_i^2$$

Allora se $Q_1$ è una variabile di $\mathcal{X}^2$ a $p < n$ gradi di libertà, la $Q_2$ è una variabile di $\mathcal{X}^2$ a $n - p < n$ gradi di libertà, e le due variabili sono stocasticamente indipendenti

**Dim.**
Segue immediatamente dai teoremi di Craig e di caratterizzazione delle variabili $\mathcal{X}^2$, osservando che $A_1 + A_2 = \mathbb{I}$ □

($ii$) Siano $Q$, $Q_1$ e $Q_2$ tre forme quadratiche semidefinite positive delle variabili gaussiane $(x_1, x_2, \ldots, x_n) = x$, indipendenti e standard. Le variabili casuali $Q$, $Q_1$ e $Q_2$ soddisfino la condizione

$$Q_1 + Q_2 = Q$$

Allora se $Q$ e $Q_1$ sono variabili di $\mathcal{X}^2$ rispettivamente a $n$ e $p < n$ gradi di libertà, la $Q_2$ è una variabile di $\mathcal{X}^2$ a $n - p < n$ gradi di libertà, e risulta stocasticamente indipendente da $Q_1$

**Dim.**
Si può verificare che questo teorema è facilmente riconducibile al precedente (avremo occasione di parlarne in seguito)

# 6. FUNZIONI DI VARIABILI CASUALI E MISURE INDIRETTE

**Misura diretta**
Viene ottenuta mediante il confronto con un campione o l'uso di uno strumento tarato

**Misura indiretta**
È desunta da un calcolo, da una espressione di grandezze a loro volta misurate direttamente o indirettamente

In presenza di errori casuali, il risultato di una misura indiretta si deve intendere come una variabile casuale **funzione di** altre **variabili casuali**

$$X = f(X_1, X_2, \ldots, X_n)$$

Nota la distribuzione congiunta di probabilità delle variabili $(X_1, X_2, \ldots, X_n)$, si vuole determinare quella di $X$

| | | |
|---|---|---|
| distribuzione congiunta di $(X_1, X_2, \ldots, X_n)$ | $\Longrightarrow$ | distribuzione di $X = f(X_1, X_2, \ldots, X_n)$ |

La distribuzione di probabilità della funzione

$$X = f(X_1, X_2, \ldots, X_n)$$

delle variabili casuali $X_1, X_2, \ldots, X_n$ può essere **calcolata esplicitamente** solo in casi molto particolari, assegnata che sia la distribuzione di probabilità congiunta $p(x_1, x_2, \ldots, x_n)$

Càpita sovente che la distribuzione di probabilità congiunta delle $X_1, X_2, \ldots, X_n$ non sia nota, ma che si conoscano soltanto (stime di) **medie, varianze e covarianze**

Può persino verificarsi che delle variabili $X_1, X_2, \ldots, X_n$ si abbia soltanto una stima dei valori veri presunti $\overline{x}_1, \ldots, \overline{x}_n$ e degli errori $\Delta x_1, \ldots, \Delta x_n$

$$\Downarrow$$

Ci si può trovare nella **impossibilità di calcolare esplicitamente la distribuzione di probabilità di** $X$.

In tal caso ci si deve limitare a obiettivi più modesti:

($i$) calcolare **media e varianza** della funzione casuale $x$;

($ii$) stimare soltanto il **valore vero presunto e l'errore** di $x$;

($iii$) ricavare un'**approssimazione numerica** della distribuzione di probabilità di $X$ con un **metodo di Monte-Carlo**

## □ Combinazione lineare di variabili casuali Calcolo della media e della varianza

Date le variabili casuali $(X_1, X_2, \ldots, X_n) = X$, si consideri la variabile casuale definita dalla combinazione lineare

$$Z \;=\; \sum_{i=1}^{n} a_i X_i \;=\; a^T X$$

con le $a_i$ costanti reali assegnate e $a^T = (a_1, \ldots, a_n)$.
Si ha allora che:

($i$) la media di $Z$ è la combinazione lineare, di uguali coefficienti $a_i$, delle medie delle singole variabili

$$\mathbb{E}(Z) \;=\; \mathbb{E}\Big[\sum_{i=1}^{n} a_i X_i\Big] \;=\; \sum_{i=1}^{n} \mathbb{E}(a_i X_i) \;=\; \sum_{i=1}^{n} a_i \mathbb{E}(X_i)$$

($ii$) la varianza di $Z$ è data dall'espressione

$$\begin{aligned}
\mathbb{E}[(Z - \mathbb{E}(Z))^2] \;&=\; \mathbb{E}\Big[\sum_{i,j=1}^{n} a_i [X_i - \mathbb{E}(X_i)]\, a_j [X_j - \mathbb{E}(X_j)]\Big] \;= \\
&=\; \sum_{i,j=1}^{n} a_i a_j \mathbb{E}\big[[X_i - \mathbb{E}(X_i)][X_j - \mathbb{E}(X_j)]\big] \;= \\
&=\; \sum_{i,j=1}^{n} a_i a_j C_{ij} \;=\; a^T C a
\end{aligned}$$

in cui $C$ indica la matrice di covarianza delle variabili $X$.
Per **variabili scorrelate** (a maggior ragione se indipendenti)

$$\mathbb{E}[(Z - \mathbb{E}(Z))^2] \;=\; \sum_{i=1}^{n} a_i^2 \,\mathrm{var}(X_i)$$

## □ Combinazione lineare di variabili normali

Sia $(X_1, X_2, \ldots, X_n) = X$ un sistema di variabili gaussiane con una distribuzione di probabilità congiunta individuata dalla matrice di struttura $A$ e dal valor medio $\mu \in \mathbb{R}^n$.
Allora:

($i$) data una qualsiasi matrice $n \times n$ $R$, reale e non singolare, l'insieme di variabili casuali definito da $(Y_1, \ldots, Y_n) = Y = RX$ costituisce ancora un sistema di variabili gaussiane, con distribuzione congiunta

$$p(y) \;=\; \frac{[\det[(R^{-1})^T\, A\, R^{-1}]]^{1/2}}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}(y-R\mu)^T (R^{-1})^T A R^{-1}(y-R\mu)}$$

e quindi media $R\mu$ e matrice di struttura $(R^{-1})^T A R^{-1}$;

($ii$) con riferimento al risultato precedente, se le variabili gaussiane $X$ sono stocasticamente indipendenti e standard, e inoltre la matrice $R$ è ortogonale ($R^T = R^{-1}$), le variabili $Y = RX$ risultano a loro volta gaussiane, indipendenti e standard;

($iii$) per un qualsiasi vettore costante $a^T = (a_1, \ldots, a_n) \in \mathbb{R}^n$, la **combinazione lineare**

$$Z \;=\; a^T X \;=\; \sum_{i=1}^{n} a_i X_i$$

**è una variabile gaussiana** di media e varianza rispettive

$$\mathbb{E}(Z) \;=\; a^T \mu \qquad\qquad \mathbb{E}[(Z - \mathbb{E}(Z))^2] \;=\; a^T C a \;=\; a^T A^{-1} a$$

## □ Propagazione degli errori nelle misure indirette

### ($i$) Legge di Gauss

Se:

- varianze e covarianze delle variabili casuali $X_1, X_2, \ldots, X_n$ sono note e abbastanza piccole,
- si conoscono le medie $\mu_1, \ldots, \mu_n$ delle stesse variabili,
- la funzione $f$ risulta sufficientemente regolare (es. $C^2$),

allora la media e la varianza di $X = f(X_1, X_2, \ldots, X_n)$ possono essere stimate usando l'approssimazione di Taylor:

$$X \;=\; f(\mu_1, \ldots, \mu_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n)(X_i - \mu_i)$$

e valgono perciò

$$\mathbb{E}(X) \;=\; f(\mu_1, \ldots, \mu_n)$$

$$\mathbb{E}[(X - \mathbb{E}(X))^2] \;=\; \sum_{i,j=1}^{n} \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n) \frac{\partial f}{\partial x_j}(\mu_1, \ldots, \mu_n) C_{ij}$$

essendo $C$ la matrice di covarianza

Per **variabili scorrelate** (o a maggior ragione se indipendenti) vale la relazione

$$\mathbb{E}[(X - \mathbb{E}(X))^2] \;=\; \sum_{i=1}^{n} \Big[ \frac{\partial f}{\partial x_i}(\mu_1, \ldots, \mu_n) \Big]^2 \mathrm{var}(X_i)$$

che esprime la c.d. **legge di propagazione degli errori casuali**, o di Gauss

**Esempio**: siano date due variabili casuali indipendenti $X$ e $Y$ tali che:

$$X \text{ ha media } \mu_X = 0.5 \text{ e varianza } \sigma_X^2 = 0.1$$
$$Y \text{ ha media } \mu_Y = 0.2 \text{ e varianza } \sigma_Y^2 = 0.5$$

Si consideri la variabile casuale

$$Z \;=\; f(X,Y) \;=\; X \,\sin Y + X^2 Y$$

Le derivate parziali prime della funzione $f(X,Y)$ sono

$$\frac{\partial f}{\partial X}(X,Y) \;=\; \sin Y + 2XY$$

$$\frac{\partial f}{\partial Y}(X,Y) \;=\; X \cos Y + X^2$$

e nei valori medi $(X,Y) = (\mu_X, \mu_Y) = (0.5, 0.2)$ valgono

$$\frac{\partial f}{\partial X}(0.5,0.2) \;=\; \sin 0.2 + 2 \cdot 0.5 \cdot 0.2 \;=\; 0.398669331$$

$$\frac{\partial f}{\partial Y}(0.5,0.2) \;=\; 0.5 \cdot \cos 0.2 + 0.5^2 \;=\; 0.740033289$$

La variabile casuale $Z = f(X,Y)$ ha valor medio

$$f(0.5,0.2) \;=\; 0.5 \cdot \,\sin 0.2 + 0.5^2 \cdot 0.2 \;=\; 0.1493346654$$

e varianza

$$\sigma_Z^2 \;=\; 0.398669331^2 \cdot 0.1 + 0.740033289^2 \cdot 0.5 \;=\; 0.2897183579$$

### (*ii*) **Differenziale logaritmico**

Se sulle grandezze $X_1, X_2, \ldots, X_n$ non si dispone di alcuna informazione statistica, ma sono noti soltanto i loro valori veri presunti $\overline{x}_1, \ldots, \overline{x}_n$ e gli errori stimati $\Delta x_1, \ldots, \Delta x_n$

$$\Downarrow$$

si valuta il valore vero presunto di $x = f(x_1, x_2, \ldots, x_n)$ come

$$\overline{x} \;=\; f(\overline{x}_1, \ldots, \overline{x}_n)$$

e l'errore $\Delta x$ nella forma data dal differenziale

$$\Delta x \;=\; \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\overline{x}_1, \ldots, \overline{x}_n) \right| \Delta x_i$$

Questa relazione viene spesso ricavata calcolando il differenziale del **logaritmo naturale** di $f$

$$\frac{\Delta x}{\overline{x}} \;=\; \sum_{i=1}^{n} \left| \frac{\partial \ln f}{\partial x_i}(\overline{x}_1, \ldots, \overline{x}_n) \right| \Delta x_i \qquad (1)$$

molto comodo quando $f$ è un prodotto di fattori. La procedura è allora nota come **metodo del differenziale logaritmico**

Nella progettazione di un esperimento il metodo del differenziale logaritmico viene spesso usato attenendosi al cosiddetto **principio di uguaglianza degli effetti**.

Ci si adopera cioè per ridurre le incertezze $\Delta x_i$ in modo che i termini della somma (1) siano **tutti dello stesso ordine di grandezza**

**Esempio**: applicazione del differenziale logaritmico
Il numero di Reynolds di un sistema fluidodinamico è dato dalla relazione

$$R = \frac{\rho v \ell}{\eta}$$

dove:

- $\eta = (1.05 \pm 0.02) \cdot 10^{-3}$ kg m$^{-1}$s$^{-1}$ (coefficiente di viscosità dinamica del liquido)
- $\rho = (0.999 \pm 0.001) \cdot 10^{3}$ kg m$^{-3}$ (densità del liquido)
- $\ell = (0.950 \pm 0.005)$ m (lunghezza caratteristica)
- $v = (2.5 \pm 0.1)$ m s$^{-1}$ (velocità caratteristica)

Il valore vero presunto di $R$ si calcola con i valori presunti di $\rho, v, \ell, \eta$:

$$\overline{R} = \frac{0.999 \cdot 10^{3} \cdot 2.5 \cdot 0.950}{1.05 \cdot 10^{-3}} = 2.259642857 \cdot 10^{6}$$

mentre l'errore relativo si stima con il differenziale logaritmico:

$$\begin{aligned} \frac{\Delta R}{\overline{R}} &= \frac{\Delta \rho}{\overline{\rho}} + \frac{\Delta v}{\overline{v}} + \frac{\Delta \ell}{\overline{\ell}} + \frac{\Delta \eta}{\overline{\eta}} = \\ &= \frac{0.001}{0.999} + \frac{0.1}{2.5} + \frac{0.005}{0.950} + \frac{0.02}{1.05} = 0.06531177795 \end{aligned}$$

L'errore relativo sul numero di Reynolds è quindi del 6.5% e corrisponde ad un errore assoluto

$$\begin{aligned} \Delta R &= \overline{R} \cdot \frac{\Delta R}{\overline{R}} = 2.259642857 \cdot 10^{6} \cdot 0.06531177795 = \\ &= 0.1475812925 \cdot 10^{6} \end{aligned}$$

**Esempio**: applicazione del differenziale logaritmico
Il volume del cilindro circolare retto è dato da

$$V = \pi r^2 h$$

con:

- $r = (2.15 \pm 0.02)$ m (raggio di base)
- $h = (3.45 \pm 0.05)$ m (altezza)

Il volume del cilindro è stimato da

$$\overline{V} = 3.141592653 \cdot 2.15^2 \cdot 3.45 = 50.10094154 \text{ m}^3$$

valore al quale è associato l'errore relativo

$$\frac{\Delta V}{\overline{V}} = 2\frac{\Delta r}{\overline{r}} + \frac{\Delta h}{\overline{h}} = 2 \cdot \frac{0.02}{2.15} + \frac{0.05}{3.45} = 0.03309740478$$

in cui l'errore sulla costante numerica $\pi$ si è assunto trascurabile
L'errore assoluto vale infine

$$\Delta V = \overline{V} \cdot \frac{\Delta V}{\overline{V}} = 50.10094154 \cdot 0.03309740478 = \\ = 1.658211142 \text{ m}^3$$

Da notare che la precisione del risultato finale non può essere superiore a quella dei fattori $r$ ed $h$ (l'errore relativo sul risultato finale è la somma degli errori relativi)

**Esempio**: applicazione del differenziale logaritmico
Il periodo delle piccole oscillazioni di un pendolo semplice è espresso dalla legge di Galileo:

$$T = 2\pi\sqrt{\frac{L}{g}}$$

in termini di:

- $L = (109.0 \pm 0.2)$ cm (lunghezza)
- $g = (980.5 \pm 0.5)$ cm s$^{-2}$ (accelerazione gravitazionale)

L'errore relativo su $T$ si calcola con il differenziale logaritmico

$$\frac{\delta T}{\overline{T}} = \frac{1}{2}\frac{\Delta L}{\overline{L}} + \frac{1}{2}\frac{\Delta g}{\overline{g}}$$

trascurando l'errore su $\pi$

Si ha così

$$\overline{T} = 2 \cdot 3.141592653 \cdot \sqrt{\frac{109.0}{980.5}} = 2.094929046 \text{ s}$$

e

$$\frac{\delta T}{\overline{T}} = \frac{1}{2}\frac{0.2}{109.0} + \frac{1}{2}\frac{0.5}{980.5} = 0.1172403146\,\%$$

per cui

$$\Delta T = \overline{T} \cdot \frac{\delta T}{\overline{T}} = 2.094929046 \cdot 0.001172403146 = \\ = 0.002456101404 \text{ s}$$

### (*iii*) Propagazione dell'errore nella soluzione di un sistema di equazioni algebriche lineari

Sia dato il sistema algebrico lineare non-omogeneo

$$Ax = b$$

in cui $A$ è una matrice invertibile $n \times n$ e $b$ un vettore colonna di $\mathbb{R}^n$.

Si supponga che gli elementi di $A$ e di $b$ siano soggetti a certi errori, esprimibili per mezzo di appropriate matrici $\Delta A$ e $\Delta b$.

Di conseguenza, la soluzione $x \in \mathbb{R}^n$ del sistema sarà a sua volta soggetta ad un certo errore $\Delta x$.

Nell'ipotesi (usuale) che gli errori $\Delta A$ su $A$ siano piccoli, vale la seguente stima dell'errore $\Delta x$ su $x$

$$\frac{|\Delta x|}{|x|} \leq \mathrm{cond}(A)\Big(\frac{|\Delta b|}{|b|} + \frac{\|\Delta A\|}{\|A\|}\Big)\frac{1}{1 - \mathrm{cond}(A)\dfrac{\|\Delta A\|}{\|A\|}}$$

dove:

$|\cdot|$ indica una qualsiasi **norma vettoriale** di $\mathbb{R}^n$;

$\|\cdot\|$ è la **norma matriciale subordinata** alla norma vettoriale precedente $|\cdot|$;

$\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|$ rappresenta il c.d. **numero di condizionamento** (condition number) della matrice $A$

**Esempi notevoli di norme vettoriali di uso corrente**

Per un arbitrario vettore $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$
si definiscono:

– la **norma uno** o $\ell^1$, data da

$$|x|_1 = \sum_{i=1}^{n} |x_i|$$

– la **norma due** o $\ell^2$, che è la consueta norma euclidea

$$|x|_2 = \left[ \sum_{i=1}^{n} x_i^2 \right]^{1/2}$$

– la **norma infinito** o $\ell^\infty$, espressa come

$$|x|_\infty = \max_{i=1,\ldots,n} |x_i|$$

La norma matriciale $\|\cdot\|$ subordinata ad una norma vettoriale $|\cdot|$ è definita formalmente come

$$\|A\| \;=\; \inf\bigl\{c \in \mathbb{R}\,;\, |Ax| \le c|x| \,\forall\, x \in \mathbb{R}^n\bigr\}$$

Niente paura! Per le norme vettoriali precedenti il calcolo delle norme matriciali subordinate è abbastanza semplice. Infatti, per una arbitraria matrice quadrata $A$:

- la norma matriciale subordinata alla norma vettoriale $|\cdot|_1$ risulta
$$\|A\|_1 \;=\; \max_{j=1,\ldots,n} \sum_{i=1}^{n} |A_{ij}|$$
e dunque si ottiene calcolando la somma dei valori assoluti degli elementi di ogni singola colonna di $A$ e considerando poi il maggiore dei risultati;

- la norma matriciale subordinata alla norma vettoriale $|\cdot|_2$ si scrive
$$\|A\|_2 \;=\; \sqrt{\lambda}$$
essendo $\lambda$ il massimo autovalore della matrice $A^T A$, reale, simmetrica e semidefinita positiva;

- la norma matriciale subordinata alla norma vettoriale $|\cdot|_\infty$ è
$$\|A\|_\infty \;=\; \max_{i=1,\ldots,n} \sum_{j=1}^{n} |A_{ij}|$$
e si ricava pertanto determinando la somma dei valori assoluti degli elementi di ogni riga e prendendo quindi la maggiore di tali somme

**Esempio**: propagazione di errore nella soluzione di un sistema lineare di due equazioni algebriche in due incognite. Si consideri il sistema

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases}$$

i cui coefficienti numerici sono affetti da una incertezza:

$$\begin{aligned} a_{11} &= 5.0 \pm 0.1 & a_{12} &= 6.2 \pm 0.1 \\ a_{21} &= 3.5 \pm 0.1 & a_{22} &= 10.5 \pm 0.2 \\ b_1 &= 12.0 \pm 0.2 & b_2 &= 15.2 \pm 0.1 \end{aligned}$$

Le matrici da prendere in esame sono le seguenti:

$$A = \begin{pmatrix} 5.0 & 6.2 \\ 3.5 & 10.5 \end{pmatrix} \qquad \Delta A = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$$

$$b = \begin{pmatrix} 12.0 \\ 15.2 \end{pmatrix} \qquad \Delta b = \begin{pmatrix} 0.2 \\ 0.1 \end{pmatrix}$$

Occorre calcolare l'inversa della matrice $A$

$$\begin{aligned} A^{-1} &= \frac{1}{5.0 \cdot 10.5 - 3.5 \cdot 6.2} \begin{pmatrix} 10.5 & -6.2 \\ -3.5 & 5.0 \end{pmatrix} = \\ &= \begin{pmatrix} 0.34090909 & -0.20129870 \\ -0.11363636 & 0.16233766 \end{pmatrix} \end{aligned}$$

per mezzo del determinante

$$\det A = 5.0 \cdot 10.5 - 3.5 \cdot 6.2 = 30.80$$

La soluzione stimata del sistema lineare si calcola usando i valori medi dei vari coefficienti:

$$\begin{aligned}\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} &= A^{-1}b = \\ &= \begin{pmatrix} 0.34090909 & -0.20129870 \\ -0.11363636 & 0.16233766 \end{pmatrix} \begin{pmatrix} 12.0 \\ 15.2 \end{pmatrix} = \\ &= \begin{pmatrix} 1.03116884 \\ 1.10389611 \end{pmatrix}\end{aligned}$$

Valutiamo la propagazione dell'errore usando le norme $\ell^\infty$.
Si ha:

- $|b|_\infty = \max\{12.0, 15.2\} = 15.2$
- $|\Delta b|_\infty = \max\{0.2, 0.1\} = 0.2$
- $\|A\|_\infty = \max\{11.2, 14.0\} = 14.0$, in quanto

$$\begin{pmatrix} |5.0| & |6.2| \\ |3.5| & |10.5| \end{pmatrix} \begin{matrix} \rightarrow \text{ somma } 11.2 \\ \rightarrow \text{ somma } 14.0 \end{matrix}$$

- $\|\Delta A\|_\infty = \max\{0.2, 0.3\} = 0.3$, essendo

$$\begin{pmatrix} |0.1| & |0.1| \\ |0.1| & |0.2| \end{pmatrix} \begin{matrix} \rightarrow \text{ somma } 0.2 \\ \rightarrow \text{ somma } 0.3 \end{matrix}$$

- $\|A^{-1}\|_\infty = \max\{0.54220779, 0.27597402\} = 0.54220779$, poichè

$$\begin{pmatrix} |+0.34090909| & |-0.20129870| \\ |-0.11363636| & |+0.16233766| \end{pmatrix} \begin{matrix} \rightarrow \text{ somma } 0.54220779 \\ \rightarrow \text{ somma } 0.27597402 \end{matrix}$$

Di conseguenza, il condition number risulta

$$\begin{aligned} \text{cond}(A) &= \|A\|_\infty \cdot \|A^{-1}\|_\infty = \\ &= 14.0 \cdot 0.54220779 = 7.59090906 \end{aligned}$$

La norma della soluzione vale infine

$$|\mathbf{x}|_\infty = \max\{1.03116884, 1.10389611\} = 1.10389611$$

La propagazione dell'errore sulla soluzione è allora stimata da

$$\frac{|\Delta\mathbf{x}|_\infty}{|\mathbf{x}|_\infty} = \frac{\max\{|\Delta x|, |\Delta y|\}}{|\mathbf{x}|_\infty} \leq$$

$$\leq \text{cond}(A)\Big(\frac{|\Delta b|_\infty}{|b|_\infty} + \frac{\|\Delta A\|_\infty}{\|A\|_\infty}\Big)\frac{1}{1 - \text{cond}(A)\dfrac{\|\Delta A\|_\infty}{\|A\|_\infty}} =$$

$$= 7.59090906 \cdot \Big(\frac{0.2}{15.2} + \frac{0.3}{14.0}\Big) \cdot \frac{1}{1 - 7.59090906 \cdot \dfrac{0.3}{14.0}} =$$

$$= 0.31354462$$

L'errore relativo sulla soluzione è circa del 31% !

Errore assoluto:

$$\max\{|\Delta x|, |\Delta y|\} \leq 0.31354462 \cdot 1.10389611 = 0.34612068$$

### (*iv*) **Stima della distribuzione di probabilità di una funzione di variabili casuali con il metodo di Monte-Carlo**

È dato costruire una distribuzione di probabilità approssimata della funzione $x = f(x_1, x_2, \ldots, x_n)$ **generando a caso** i valori delle variabili casuali $x_1, \ldots, x_n$ secondo la relativa distribuzione cumulativa di probabilità, che si suppone nota

La raccolta e istogrammazione dei risultati fornisce la distribuzione di probabilità approssimata della variabile casuale $x$

Il caso più semplice è quello delle variabili **stocasticamente indipendenti**, con distribuzioni di probabilità assegnate

Si rende necessario un algoritmo per la **generazione di numeri casuali**

In realtà è sufficiente un algoritmo per la generazione di numeri casuali **con distribuzione uniforme** nell'intervallo $[0, 1]$

Se infatti $p(x_i)$ è la distribuzione di probabilità della variabile $x_i \in \mathbb{R}$ e $P(x_i)$ quella cumulativa corrispondente, si dimostra che la variabile casuale

$$Z \;=\; P(X) \;=\; \int_{-\infty}^{X} p(x_i)\, dx_i \;\in\; (0, 1)$$

segue una **distribuzione uniforme in** $[0, 1]$

**Teorema**
Sia $x$ una variabile casuale in $\mathbb{R}$ con distribuzione di probabilità $p(x)$ e $g : \mathbb{R} \longrightarrow \mathbb{R}$ una funzione $C^1$ monotòna crescente. La variabile casuale

$$y = g(x)$$

assume allora valori nell'intervallo $(g(-\infty), g(+\infty))$, con distribuzione di probabilità

$$\hat{p}(y) = p[g^{-1}(y)] \frac{1}{\left.\frac{dg}{dx}(x)\right|_{x=g^{-1}(y)}}$$

essendo $g(\pm\infty) = \lim_{x\to\pm\infty} g(x)$ e $g^{-1}$ l'inversa di $g$.

**Dim.**
Basta introdurre il cambiamento di variabile $x = g^{-1}(y)$ nell'integrale di normalizzazione e fare uso del teorema di derivazione delle funzioni inverse

$$1 = \int_{-\infty}^{+\infty} p(x)\,dx = \int_{g(-\infty)}^{g(+\infty)} p\big[g^{-1}(y)\big] \frac{d}{dy}\big[g^{-1}(y)\big]\,dy =$$

$$= \int_{g(-\infty)}^{g(+\infty)} p[g^{-1}(y)] \frac{1}{\left.\frac{dg}{dx}(x)\right|_{x=g^{-1}(y)}}\,dy \qquad \square$$

**N.B.**
Se $y = g(x) = P(x)$ si ha

$$\hat{p}(y) = p[g^{-1}(y)] \frac{1}{\left.p(x)\right|_{x=g^{-1}(y)}} = 1$$

per $y \in (P(-\infty), P(+\infty)) = (0, 1)$

È dunque possibile procedere nel modo seguente:

(1) Tabulazione della distribuzione cumulativa $P(x)$, per $x \in (-\infty, +\infty)$ (o intervallo abbastanza grande)

(2) Generazione dei valori della variabile $z = P(x)$, uniformemente distribuita in $[0, 1]$, mediante un **generatore di numeri casuali uniformi in** $[0, 1]$

(3) Calcolo dei corrispondenti valori di

$$x = P^{-1}(z)$$

grazie all'invertibilità della funzione $P$

Il calcolo avviene di solito per interpolazione sui valori tabulati di $P$

La variabile $x$ risulterà distribuita secondo $p(x)$!

**Rimane il problema del generatore di numeri casuali!**

## Generatori di numeri casuali uniformi in $[0, 1]$

Si tratta sempre di **algoritmi deterministici**,
che richiedono un input iniziale (seed)

I valori generati non sono realmente casuali,
ma soltanto **pseudocasuali**

Gli algoritmi assicurano che le sequenze di valori ottenibili soddisfino alcuni **criteri di stocasticità**

Per esempio, che generando un numero di valori sufficientemente elevato, il relativo istogramma di frequenza risulti pressoché costante per larghezze di bin relativamente piccole e indipendentemente dall'input iniziale

Gli algoritmi di generazione casuale possono essere più o meno sofisticati

Una tipologia di generatori piuttosto semplice è costituita dagli algoritmi **moltiplicazione-modulo**

Un esempio standard è il seguente

$$y_{n+1} = 16807\, y_n \mod 2147483647 \qquad \forall\, n = 0, 1, 2, \ldots$$

Se il seme iniziale $y_0$ è un intero dispari grande, la sequenza

$$x_n = y_n / 2147483647$$

simula una successione di estrazioni casuali di una variabile casuale uniformemente distribuita in $[0, 1)$

# 7. TEORIA DEI CAMPIONI

I risultati di una misura ripetuta soggetta ad errori casuali
si assumono come elementi estratti da una
**popolazione statistica**

Questa è descritta da un'appropriata
**distribuzione di probabilità**

Sorge il problema di ricavare, dal campione ridotto disponibile,
le **caratteristiche** di questa distribuzione di probabilità

⇕

**Problema fondamentale dell′inferenza statistica**

In particolare:

**Stima di media e varianza** della distribuzione

In generale:

**Test delle ipotesi** sulla distribuzione e sui parametri che in essa compaiono

## 7.1 Stima puntuale della media

Se $x_1, x_2, \ldots, x_n$ sono i risultati di $n$ ripetizioni di una **stessa misura sperimentale**, essi si possono intendere come le realizzazioni di $n$ variabili casuali indipendenti **identicamente distribuite**, secondo una **distribuzione incognita**

Conviene indicare con gli stessi simboli $x_1, x_2, \ldots, x_n$ tali variabili casuali identicamente distribuite (per semplicità)

La **media** $\mu$ **della distribuzione** di probabilità incognita viene stimata, sulla base del campione ridotto disponibile, per mezzo della **media aritmetica**

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

La media aritmetica va intesa come una **variabile casuale**, funzione delle variabili casuali $x_1, x_2, \ldots, x_n$

La stima è **corretta** (o **consistente**). Infatti:

($i$) la media di $\overline{x}$ è $\mu$, per ogni $n$

$$\mathbb{E}(\overline{x}) \;=\; \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mu \;=\; \mu$$

($ii$) se $\sigma^2$ è la varianza di ciascuna variabile $x_i$, la variabile casuale $\overline{x}$ ha varianza $\sigma^2/n$ (**formula di de Moivre**)

$$\mathbb{E}[(\overline{x}-\mu)^2] = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[(x_i-\mu)(x_j-\mu)] = \frac{1}{n^2}\sum_{i,j=1}^{n}\delta_{ij}\sigma^2 = \frac{\sigma^2}{n}$$

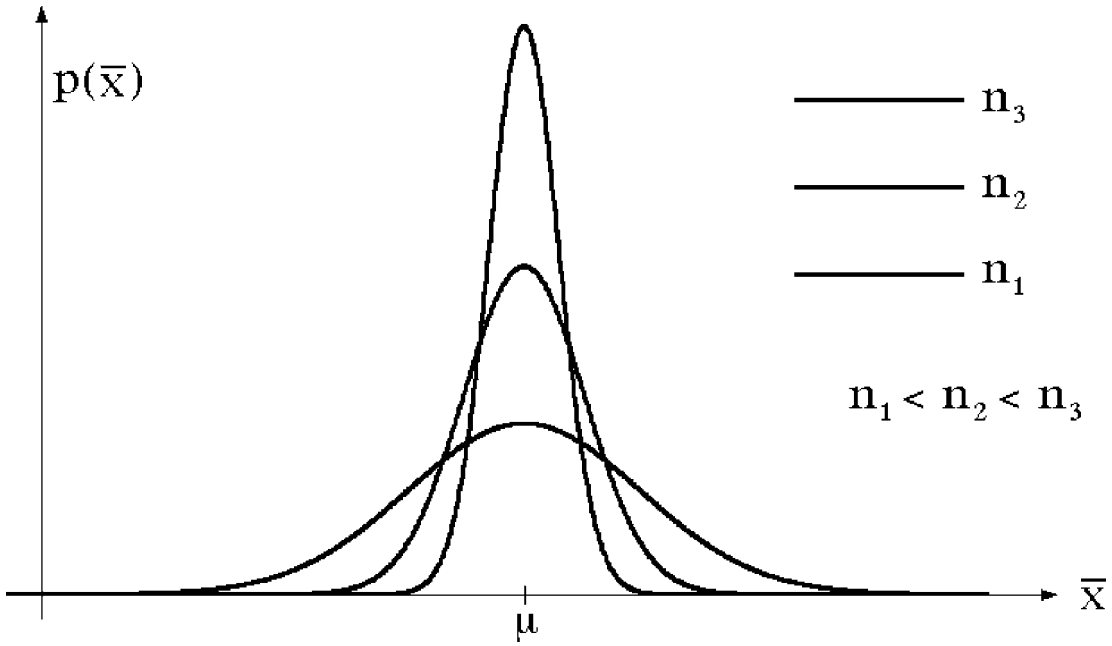
($iii$) più in generale, vale il **Teorema di Khintchine** o **Legge debole dei grandi numeri**
Se $(x_1, x_2, \ldots, x_n)$ è un campione estratto da una popolazione statistica la cui distribuzione di probabilità abbia media $\mu$, allora la media aritmetica $\overline{x} = \overline{x}_n$ **converge in probabilità** a $\mu$ per $n \to \infty$

$$\forall\, \varepsilon > 0 \qquad \lim_{n\to\infty} p[\mu - \varepsilon \le \overline{x}_n \le \mu + \varepsilon] \;=\; 1$$

**Dim.** Per $\sigma^2$ finita, il risultato segue immediatamente dal teorema di Tchebyshev

Questo significa che per $n$ crescenti la media aritmetica $\overline{x}$ tende a fornire una approssimazione sempre migliore della media $\mu$ della distribuzione di probabilità

A illustrazione di quanto affermato, la figura seguente mostra come all'aumentare del numero $n$ di dati su cui la media campionaria è calcolata, la distribuzione di probabilità di questa tende sempre più a **concentrarsi a ridosso del valor medio** $\mu$




Per quanto riguarda la **forma** della distribuzione, si può notare quanto segue:

- se i dati $x_1, \ldots, x_n$ del campione **hanno distribuzione normale**, anche la media $\overline{x}$ risulta una variabile normale, quale combinazione lineare di variabili normali;

- anche **se la distribuzione dei dati non è normale**, tuttavia, il teorema del limite centrale assicura che per $n$ sufficientemente grande la distribuzione di $\overline{x}$ è normale con ottima approssimazione

## 7.2 Stima puntuale della varianza

Nelle stesse ipotesi, la **varianza della distribuzione** di probabilità viene stimata per mezzo dell'espressione

$$s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

che deve sempre essere considerata come una variabile casuale funzione di $x_1, x_2, \ldots, x_n$

Anche in questo caso **la stima è corretta**, in quanto:

($i$) la media $\mathbb{E}(s^2)$ di $s^2$ coincide con $\sigma^2$

$$\frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^{n}(x_i-\overline{x})^2\Big] = \frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^{n}(x_i-\mu)^2 - n(\overline{x}-\mu)^2\Big]$$

$$= \frac{1}{n-1}\Big[\sum_{i=1}^{n}\mathbb{E}[(x_i-\mu)^2] - n\mathbb{E}[(\overline{x}-\mu)^2]\Big] \;=\; \sigma^2$$

($ii$) se il momento centrale quarto $\mathbb{E}[(x_i-\mu)^4]$ delle variabili $x_i$ è finito, la varianza di $s^2$ risulta

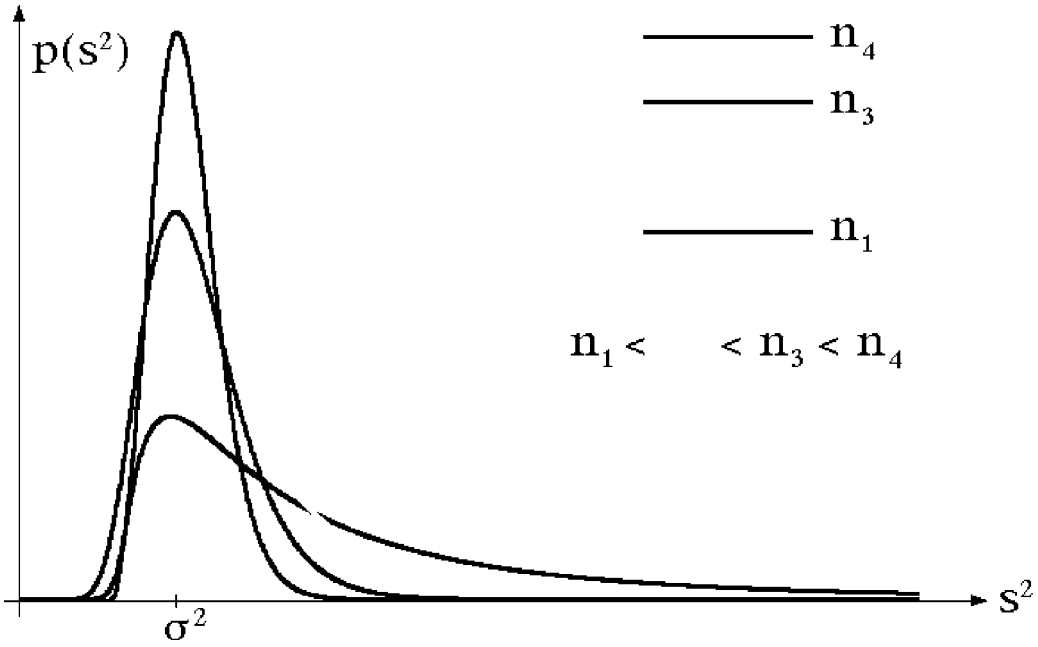
$$\mathbb{E}[(s^2-\sigma^2)^2] \;=\; \frac{1}{n}\mathbb{E}[(x_i-\mu)^4] + \frac{3-n}{n(n-1)}\sigma^4$$

tendendo a zero per $n \to +\infty$. Conseguentemente $s^2 = s_n^2$ converge in probabilità al valore $\sigma^2$

$$\forall\,\varepsilon > 0 \qquad \lim_{n\to\infty} p[\sigma^2-\varepsilon \le s_n^2 \le \sigma^2+\varepsilon] \;=\; 1$$

**N.B.** La stima con $1/n$ in luogo di $1/(n-1)$ non è corretta

Anche per la stima campionaria della varianza, dunque, la distribuzione tende a concentrarsi intorno alla varianza $\sigma^2$




Quanto alla **forma** della distribuzione, si ha che:

- se i dati $x_1, \ldots, x_n$ del campione **hanno distribuzione normale con varianza** $\sigma^2$, la variabile casuale

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

  segue una distribuzione di $\mathcal{X}^2$ a $n-1$ gradi di libertà (questo aspetto verrà discusso nel seguito);

- **se la distribuzione dei dati non è normale**, per $n$ abbastanza grande ($n > 30$) la distribuzione di $s^2$ può ritenersi pressochè normale, con la media $\sigma^2$ e la varianza $\mathbb{E}[(s^2-\sigma^2)^2]$ già citate. Si noti che il risultato **non segue** dal teorema del limite centrale, poichè le variabili $x_i - \overline{x}$ in $s^2$ non sono indipendenti.

## 7.3 Intervalli di confidenza di $\mu$ e $\sigma$

Nelle applicazioni statistiche di regola non è sufficiente disporre di una stima **puntuale** della media $\mu$ e della deviazione standard $\sigma$, cioè di un semplice **valore numerico**

Occorre specificare un **intervallo** nel quale $\mu$ o $\sigma$ sia contenuto con una probabilità assegnata (**intervallo di confidenza**, o **intervallo fiduciario (confidence interval, CI)**

La probabilità che il parametro sia effettivamente contenuto nel relativo intervallo di confidenza viene detta **livello di confidenza** dell'intervallo

Gli intervalli di confidenza si calcolano facilmente in due casi:

- qualora il campione dei dati $x_1, \ldots, x_n$ sia estratto da una **popolazione arbitraria**, purchè il numero $n$ di individui sia abbastanza grande, tipicamente $n > 30$. Si parla in questo caso di **grandi campioni**;
- se i dati sono estratti da una **popolazione normale**, qualunque sia il loro numero (quindi anche per piccoli campioni, o **campioni ridotti**)

Nel primo caso gli intervalli di confidenza calcolati sono **solo approssimati**: per il CI calcolato non è noto esattamente il livello di confidenza o viceversa, per un livello di confidenza assegnato il CI è determinabile solo approssimativamente

Nel secondo caso, al contrario, si dispone di una **teoria esatta**

## 7.4 Grandi campioni di una popolazione arbitraria. Intervalli di confidenza approssimati di $\mu$ e $\sigma$

In questo caso la definizione degli intervalli di confidenza è possibile perchè per grandi campioni la distribuzione di $\overline{x}$ e $s^2$ risulta approssimativamente normale per qualsiasi popolazione

### 7.4.1 Intervallo di confidenza per la media $\mu$

Per qualsiasi popolazione di media $\mu$ e varianza $\sigma^2$ finite, il teorema del limite centrale assicura che per $n$ abbastanza grande la variabile casuale

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} \;=\; \sqrt{n}\, \frac{\overline{x} - \mu}{\sigma} \;=\; \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

è approssimativamente normale standard (e l'approssimazione è tanto migliore al crescere di $n$): le variabili $x_1, \ldots, x_n$ sono infatti indipendenti e identicamente distribuite

La deviazione standard $\sigma$ non è nota, ma $s^2$ è una RV di media $\sigma^2$ e la sua deviazione standard tende a zero come $1/\sqrt{n}$: per campioni molto grandi risulta molto probabile che si abbia $s^2 \simeq \sigma^2$. Si può così assumere che la variabile casuale

$$z_n \;=\; \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

sia **approssimativamente normale standard** per **campioni abbastanza grandi**

Questa approssimazione fornisce la chiave per costruire il CI del valor medio $\mu$

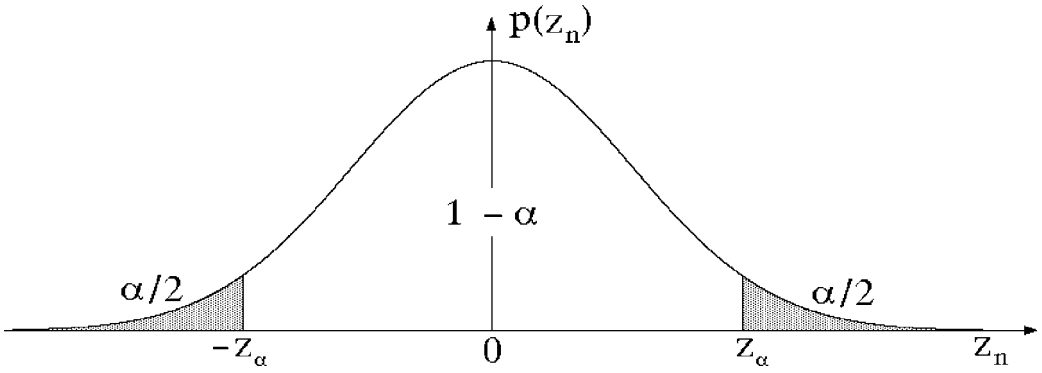
Fissato un $z_\alpha > 0$, la probabilità che la variabile $z_n$ assuma un qualsiasi valore nell'intervallo (simmetrico rispetto a $z_n = 0$)

$$-z_\alpha \leq z_n \leq z_\alpha$$

è approssimativamente data dall'integrale fra $-z_\alpha$ e $z_\alpha$ della distribuzione normale standard

$$\int_{-z_\alpha}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

e si indica con $1 - \alpha$. In realtà si fa il contrario: si assegna il **livello di confidenza** $1 - \alpha$ e si determina il valore di $z_\alpha$ corrispondente. Basta notare che, per la simmetria della distribuzione, le "code" $\{z < -z_\alpha\}$ e $\{z > z_\alpha\}$ devono avere la stessa probabilità $\alpha/2$




il valore critico $z_\alpha > 0$ essendo così definito dall'equazione

$$\frac{\alpha}{2} = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^{z_\alpha} e^{-\xi^2/2} d\xi$$

L'integrale fra 0 e $z_\alpha$ della distribuzione normale standard (o viceversa, il valore di $z_\alpha$ noto che sia il valore $(1-\alpha)/2$ dell'integrale) può essere calcolato ricorrendo direttamente ad una tabella del tipo qui parzialmente riprodotto

Area under the Normal Curve from 0 to X

| X | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.00000 | 0.00399 | 0.00798 | 0.01197 | 0.01595 | 0.01994 | 0.02392 | 0.02790 | 0.03188 | 0.03586 |
| 0.1 | 0.03983 | 0.04380 | 0.04776 | 0.05172 | 0.05567 | 0.05962 | 0.06356 | 0.06749 | 0.07142 | 0.07535 |
| 0.2 | 0.07926 | 0.08317 | 0.08706 | 0.09095 | 0.09483 | 0.09871 | 0.10257 | 0.10642 | 0.11026 | 0.11409 |
| 0.3 | 0.11791 | 0.12172 | 0.12552 | 0.12930 | 0.13307 | 0.13683 | 0.14058 | 0.14431 | 0.14803 | 0.15173 |
| 0.4 | 0.15542 | 0.15910 | 0.16276 | 0.16640 | 0.17003 | 0.17364 | 0.17724 | 0.18082 | 0.18439 | 0.18793 |
| 0.5 | 0.19146 | 0.19497 | 0.19847 | 0.20194 | 0.20540 | 0.20884 | 0.21226 | 0.21566 | 0.21904 | 0.22240 |
| 0.6 | 0.22575 | 0.22907 | 0.23237 | 0.23565 | 0.23891 | 0.24215 | 0.24537 | 0.24857 | 0.25175 | 0.25490 |
| 0.7 | 0.25804 | 0.26115 | 0.26424 | 0.26730 | 0.27035 | 0.27337 | 0.27637 | 0.27935 | 0.28230 | 0.28524 |
| 0.8 | 0.28814 | 0.29103 | 0.29389 | 0.29673 | 0.29955 | 0.30234 | 0.30511 | 0.30785 | 0.31057 | 0.31327 |
| 0.9 | 0.31594 | 0.31859 | 0.32121 | 0.32381 | 0.32639 | 0.32894 | 0.33147 | 0.33398 | 0.33646 | 0.33891 |
| 1.0 | 0.34134 | 0.34375 | 0.34614 | 0.34849 | 0.35083 | 0.35314 | 0.35543 | 0.35769 | 0.35993 | 0.36214 |
| 1.1 | 0.36433 | 0.36650 | 0.36864 | 0.37076 | 0.37286 | 0.37493 | 0.37698 | 0.37900 | 0.38100 | 0.38298 |
| 1.2 | 0.38493 | 0.38686 | 0.38877 | 0.39065 | 0.39251 | 0.39435 | 0.39617 | 0.39796 | 0.39973 | 0.40147 |

(che fornisce i valori dell'integrale per $z_\alpha$ compreso fra 0.00 e 1.29, campionati a intervalli di 0.01)

In alternativa, si può riesprimere l'integrale nella forma

$$\frac{1}{\sqrt{2\pi}} \int_0^{z_\alpha} e^{-\xi^2/2} d\xi \;=\; \frac{1}{2}\mathrm{erf}\Big(\frac{z_\alpha}{\sqrt{2}}\Big)$$

in termini della **funzione degli errori** (error function):

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_0^{x} e^{-t^2} dt$$

i cui valori sono tabulati in modo analogo

Per ottenere l'intervallo di confidenza della media $\mu$ è sufficiente ricordare che la doppia disequazione

$$-z_\alpha \leq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leq z_\alpha$$

è soddisfatta con probabilità $1 - \alpha$, per cui vale

$$\overline{x} - z_\alpha \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + z_\alpha \frac{s}{\sqrt{n}}$$

con la stessa probabilità $1 - \alpha$. Quello ottenuto è il CI della media con **livello di confidenza** $1 - \alpha$.

### 7.4.2 Intervallo di confidenza per la varianza

Per grandi campioni ($n > 30$) la varianza campionaria $s^2$ può considerarsi una variabile normale di media $\sigma^2$ e varianza

$$\mathbb{E}[(s^2 - \sigma^2)^2] = \frac{1}{n}\mathbb{E}[(x_i - \mu)^4] + \frac{3 - n}{n(n-1)}\sigma^4 .$$

La varianza si stima, al solito, con la varianza campionaria

$$\sigma^2 \simeq s^2$$

mentre il momento quarto può essere valutato usando la corrispondente stima campionaria

$$\mathbb{E}[(x_i - \mu)^4] \simeq m_4 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^4$$

o, per popolazioni di tipo noto, eventuali relazioni che leghino il momento quarto alla varianza

È dunque lecito affermare che per grandi campioni la variabile casuale

$$z_n = \frac{s^2 - \sigma^2}{\sqrt{\frac{1}{n}m_4 + \frac{3-n}{n(n-1)}s^4}}$$

è approssimativamente normale e standard

Si può così procedere al calcolo dell'intervallo di confidenza approssimato per la varianza $\sigma^2$ in modo analogo a quanto visto per la media $\mu$.

L'intervallo di confidenza per la varianza risulta:

$$s^2 - z_\alpha\sqrt{\frac{1}{n}m_4 + \frac{3-n}{n(n-1)}s^4} \leq \sigma^2 \leq \\ \leq s^2 + z_\alpha\sqrt{\frac{1}{n}m_4 + \frac{3-n}{n(n-1)}s^4}$$

e quello della deviazione standard $\sigma$ è di conseguenza:

$$\sqrt{s^2 - z_\alpha\sqrt{\frac{1}{n}m_4 + \frac{3-n}{n(n-1)}s^4}} \leq \sigma \leq \\ \leq \sqrt{s^2 + z_\alpha\sqrt{\frac{1}{n}m_4 + \frac{3-n}{n(n-1)}s^4}},$$

entrambi con livello di confidenza $1 - \alpha$.

**Esempio**: **CI per la media di una popolazione arbitraria**
Le misure dei diametri di un campione casuale di 200 sferette da cuscinetto, costruite da una macchina in una settimana, mostrano una media di 0.824 cm e una deviazione standard di 0.042 cm. Determinare, per il diametro medio delle sferette:

(**a**) l'intervallo di confidenza al 95%;

(**b**) l'intervallo di confidenza al 99%.

**Soluzione**
Poichè $n = 200 > 30$ si può applicare la teoria dei grandi campioni e non è necessario assumere che la popolazione dei dati sia normale.

(**a**) Il livello di confidenza è $1 - \alpha = 0.95$, per cui $\alpha = 0.05$ e

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \quad \Longrightarrow \quad \frac{1}{2} - \frac{\alpha}{2} = 0.5 - 0.025 = 0.475$$

Dalla tabella della distribuzione normale standard si ha allora

$$z_\alpha = z_{0.05} = 1.96$$

e l'intervallo di confidenza risulta

$$\overline{x} - z_{0.05}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{0.05}\frac{s}{\sqrt{n}}$$

ossia

$$0.824 - 1.96 \cdot \frac{0.042}{\sqrt{200}} \leq \mu \leq 0.824 + 1.96 \cdot \frac{0.042}{\sqrt{200}}$$

ed infine

$$0.81812 \text{ cm} \leq \mu \leq 0.8298 \text{ cm}\,.$$

Lo stesso intervallo di confidenza può esprimersi nella forma equivalente:

$$\mu \,=\, 0.824 \pm 0.0058 \text{ cm}\,.$$

(**b**) Nella fattispecie il livello di confidenza vale $1 - \alpha = 0.99$, in modo che $\alpha/2 = 0.005$ e

$$\frac{1-\alpha}{2} \,=\, \frac{0.99}{2} \,=\, 0.495$$

La tabella della distribuzione normale standard fornisce così

$$z_\alpha \,=\, z_{0.01} \,=\, 2.58$$

e l'intervallo di confidenza della media diventa

$$\overline{x} - z_{0.01}\frac{s}{\sqrt{n}} \,\leq\, \mu \,\leq\, \overline{x} + z_{0.01}\frac{s}{\sqrt{n}}$$

ovvero

$$0.824 \,-\, 2.58 \cdot \frac{0.042}{\sqrt{200}} \,\leq\, \mu \,\leq\, 0.824 \,+\, 2.58 \cdot \frac{0.042}{\sqrt{200}}$$

ed infine

$$0.8163 \text{ cm} \,\leq\, \mu \,\leq\, 0.8317 \text{ cm}\,.$$

L'espressione alternativa è quella che mette in evidenza l'errore assoluto:

$$\mu \,=\, 0.824 \pm 0.0077 \text{ cm}\,.$$

## 7.5 Variabili normali. Intervalli di confidenza di $\mu$ e $\sigma$

Nel caso che la popolazione statistica sia **normale**, è possibile determinare **intervalli di confidenza esatti** della media $\mu$ e della varianza $\sigma^2$ anche per **piccoli campioni**

Basta osservare che:

($i$) la variabile casuale, funzione di $x_1, x_2, \ldots, x_n$,

$$z \;=\; \sqrt{n}\frac{\overline{x}-\mu}{\sigma}$$

è normale e standard

**Dim.** $\overline{x}$ è normale, con media $\mu$ e varianza $\sigma^2/n$ □

($ii$) la variabile casuale

$$\mathcal{X}^2 \;=\; \frac{(n-1)s^2}{\sigma^2} \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

segue una distribuzione di $\mathcal{X}^2$ a $n-1$ gradi di libertà

**Dim.**
$\mathcal{X}^2$ è una forma quadratica semidefinita positiva delle variabili gaussiane indipendenti e standard $(x_i-\mu)/\sigma$

$$\mathcal{X}^2 = \sum_{i=1}^{n}\Big[\frac{x_i-\mu}{\sigma}\Big]^2 - n\Big[\frac{\overline{x}-\mu}{\sigma}\Big]^2 = \sum_{ij=1}^{n}\Big[\delta_{ij}-\frac{1}{n}\Big]\frac{x_i-\mu}{\sigma}\frac{x_j-\mu}{\sigma}$$

con matrice rappresentativa idempotente

$$\sum_{j=1}^{n}\Big[\delta_{ij}-\frac{1}{n}\Big]\Big[\delta_{jk}-\frac{1}{n}\Big] = \sum_{j=1}^{n}\Big[\delta_{ij}\delta_{jk}-\frac{1}{n}\delta_{ij}-\frac{1}{n}\delta_{jk}+\frac{1}{n^2}\Big] = \delta_{ik}-\frac{1}{n}$$

e rango $n-1$ (autovalori 1 e 0, di molteplicità $n-1$ e 1) □

(*iii*) le variabili $z$ e $\mathcal{X}^2$ sono stocasticamente indipendenti

**Dim.**
Basta provare che le forme quadratiche $z^2$ e $\mathcal{X}^2$

$$z^2 = \sum_{ij=1}^{n} \frac{1}{n} \frac{x_i - \mu}{\sigma} \frac{x_j - \mu}{\sigma} = \sum_{ij=1}^{n} A_{ij} \frac{x_i - \mu}{\sigma} \frac{x_j - \mu}{\sigma}$$

$$\mathcal{X}^2 = \sum_{ij=1}^{n} \Big[\delta_{ij} - \frac{1}{n}\Big] \frac{x_i - \mu}{\sigma} \frac{x_j - \mu}{\sigma} = \sum_{ij=1}^{n} B_{ij} \frac{x_i - \mu}{\sigma} \frac{x_j - \mu}{\sigma}$$

sono indipendenti, facendo uso del teorema di Craig. In effetti

$$(AB)_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk} = \sum_{j=1}^{n} \frac{1}{n} \Big[\delta_{jk} - \frac{1}{n}\Big] = \frac{1}{n} \Big[1 - \frac{1}{n} n\Big] = 0$$

$\forall\, ik = 1, \ldots, n$. Pertanto $AB = 0$ □

(*iv*) la variabile casuale

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

è una $t$ di Student a $n-1$ gradi di libertà

**Dim.**
Vale

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \sqrt{n} \frac{\overline{x} - \mu}{s} = \sqrt{n-1}\, \sqrt{n} \frac{\overline{x} - \mu}{\sigma} \frac{1}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}}}$$

per cui

$$t = \sqrt{n-1} \frac{z}{\sqrt{\mathcal{X}^2}}$$

con $t$ e $\mathcal{X}^2$ stocasticamente indipendenti, la prima gaussiana standard e la seconda di $\mathcal{X}^2$ a $n-1$ gradi di libertà □

### 7.5.1 Intervallo di confidenza per la media $\mu$

La distribuzione cumulativa della $t$ di Student a $n$ g.d.l.

$$P(\tau) = p[t \leq \tau] = \int_{-\infty}^{\tau} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}\,dt \qquad \tau \in \mathbb{R}$$

rappresenta la probabilità che la variabile $t$ assuma valori $\leq \tau$

Per ogni $\alpha \in (0,1)$ si pone $t_{[\alpha](n)} = P^{-1}(\alpha)$, in modo che

$$P\left(t_{[\alpha](n)}\right) = \alpha$$

e la simmetria della distribuzione di Student implica

$$t_{[1-\alpha](n)} = -t_{[\alpha](n)}$$

Fissato a piacere $\alpha \leq 1/2$, la disequazione

$$t_{[\frac{\alpha}{2}](n-1)} \leq \frac{\overline{x}-\mu}{s/\sqrt{n}} \leq t_{[1-\frac{\alpha}{2}](n-1)}$$

è allora soddisfatta con probabilità $1-\alpha$ e definisce l'**intervallo di confidenza** (bilaterale) per la media $\mu$, nella forma

$$\overline{x} - t_{[1-\frac{\alpha}{2}](n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} - t_{[\frac{\alpha}{2}](n-1)}\frac{s}{\sqrt{n}}$$

ovvero

$$\overline{x} - t_{[1-\frac{\alpha}{2}](n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{[1-\frac{\alpha}{2}](n-1)}\frac{s}{\sqrt{n}}$$

$1-\alpha$ si dice il **livello di confidenza** dell'intervallo

### 7.5.2 Intervallo di confidenza per la varianza $\sigma^2$

La distribuzione cumulativa della variabile di $\mathcal{X}^2$ a $n$ g.d.l.

$$P(\tau) \,=\, p[\mathcal{X}^2 \leq \tau] \,=\, \int_0^{\tau} \frac{1}{\Gamma(n/2)\, 2^{n/2}} e^{-\mathcal{X}^2/2} (\mathcal{X}^2)^{\frac{n}{2}-1}\, d\mathcal{X}^2$$

individua la probabilità che la variabile assuma valori $\leq \tau$

Per ogni $\alpha \in (0,1)$ si pone ${\mathcal{X}^2}_{[\alpha](n)} \,=\, P^{-1}(\alpha)$, in modo che

$$P\big({\mathcal{X}^2}_{[\alpha](n)}\big) \,=\, \alpha$$

La diseguaglianza

$${\mathcal{X}^2}_{[\frac{\alpha}{2}](n-1)} \leq \frac{(n-1)s^2}{\sigma^2} \leq {\mathcal{X}^2}_{[1-\frac{\alpha}{2}](n-1)}$$

è perciò verificata con probabilità $1-\alpha$ e definisce l'**intervallo di confidenza (bilaterale) della varianza** come

$$\frac{1}{{\mathcal{X}^2}_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{{\mathcal{X}^2}_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

In alternativa, si può introdurre un **intervallo di confidenza unilaterale**, per mezzo delle diseguaglianze

$$\frac{(n-1)s^2}{\sigma^2} \geq {\mathcal{X}^2}_{[\alpha](n-1)} \quad \Longleftrightarrow \quad \sigma^2 \leq (n-1)s^2\, \frac{1}{{\mathcal{X}^2}_{[\alpha](n-1)}}$$

In ambo i casi, il **livello di confidenza** dell'intervallo è dato da $1-\alpha$

**Esempio**: **intervalli di confidenza per dati normali**
Si consideri la seguente tabella di dati, relativi ad alcune misure ripetute di una certa grandezza (in unità arbitrarie). Il campione si assume normale.

| $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|
| 1 | 1.0851 | 13 | 1.0768 |
| 2 | 834 | 14 | 842 |
| 3 | 782 | 15 | 811 |
| 4 | 818 | 16 | 829 |
| 5 | 810 | 17 | 803 |
| 6 | 837 | 18 | 811 |
| 7 | 857 | 19 | 789 |
| 8 | 768 | 20 | 831 |
| 9 | 842 | 21 | 829 |
| 10 | 786 | 22 | 825 |
| 11 | 812 | 23 | 796 |
| 12 | 784 | 24 | 841 |

Determinare l'intervallo di confidenza (CI) della media e quello della varianza, entrambi con livello di confidenza $1 - \alpha = 0.95$

**Soluzione**
Numero dei dati: $n = 24$

Media campionaria: $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = 1.08148$

Varianza campionaria:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 6.6745 \cdot 10^{-6}$$

Deviazione standard campionaria: $s = \sqrt{s^2} = 0.002584$

• Il CI della media è

$$\overline{x} - t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}}$$

e per $\alpha = 0.05$, $n = 24$ ha perciò i limiti

$$\overline{x} - t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 - 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08039$$

$$\overline{x} + t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 + 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08257$$

in modo che il CI si scrive

$$1.08039 \leq \mu \leq 1.08257$$

o, equivalentemente,

$$[1.08148 \pm 0.00109]$$

• Il CI della varianza assume la forma

$$\frac{1}{\mathcal{X}^2_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{\mathcal{X}^2_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

con $\alpha = 0.05$ e $n = 24$. Pertanto

$$\frac{1}{\mathcal{X}^2_{[0.975](23)}} 23\, s^2 = \frac{1}{38.076} 23 \cdot 6.6745 \cdot 10^{-6} = 4.03177 \cdot 10^{-6}$$

$$\frac{1}{\mathcal{X}^2_{[0.025](23)}} 23\, s^2 = \frac{1}{11.689} 23 \cdot 6.6745 \cdot 10^{-6} = 13.1332 \cdot 10^{-6}$$

e il CI diventa

$$4.03177 \cdot 10^{-6} \leq \sigma^2 \leq 13.1332 \cdot 10^{-6}$$

# 8. TEST DELLE IPOTESI

Lo studio delle caratteristiche di una (o più) distribuzioni di probabilità avviene attraverso la formulazione di appropriate ipotesi (tipologia della distribuzione, valore dei parametri, ecc.)

La veridicità dell'ipotesi **non è certa**, ma deve essere testata

L'ipotesi di cui si vuole testare la veridicità è detta **ipotesi nulla**, di solito indicata con il simbolo $H_0$
L'ipotesi alternativa è indicata con $H_1$

L'accettazione o il rigetto dell'ipotesi nulla può risolversi in un **errore**:

– **del I tipo**, qualora l'ipotesi, in realtà corretta, sia rigettata;

– **del II tipo**, qualora venga accettata una ipotesi nulla in realtà scorretta

Il test deve poter condurre all'accettazione o al rigetto della ipotesi nulla, quantificando la relativa probabilità di errore (**livello di significatività del test**)

Molti test sono basati sul **rigetto dell'ipotesi nulla**, specificando la probabilità di un errore del primo tipo

## Test basati sul rigetto dell'ipotesi nulla

La base del test è una opportuna variabile casuale (**variable del test**), funzione degli individui del campione ridotto.
La scelta della variabile del test dipende dalla natura dell'ipotesi nulla $H_0$ che si vuole testare

L'insieme dei valori assunti dalla variabile del test viene ripartito in due regioni: una **regione di rigetto** e una **regione di accettazione**. Di regola si tratta di intervalli reali

Se il valore della variabile del test per il campione considerato cade nella regione di rigetto, l'ipotesi nulla viene rigettata

L'idea chiave è che la variabile del test sia scelta in modo da seguire una **distribuzione di probabilità nota qualora l'ipotesi nulla sia corretta**

È così possibile calcolare la **probabilità** che, nonostante la correttezza dell'ipotesi nulla, la variabile del test assuma un valore compreso entro la regione di rigetto, conducendo pertanto ad un errore del I tipo

I test basati sul rigetto dell'ipotesi nulla permettono di **quantificare la probabilità di commettere un errore del I tipo**, ossia di rigettare come falsa una ipotesi nulla in realtà corretta

## □ Primo esempio illustrativo

• Processo da studiare:
**lancio ripetuto di un dado** (lanci indipendenti)

$H_0$: la faccia 1 ha una probabilità di uscita $\pi = 1/6$
(dado regolare)

$H_1$: la faccia 1 ha una probabilità di uscita $\pi = 1/4$
(dado truccato in modo da favorire l'uscita della faccia 1)

• Probabilità di $x$ uscite della faccia 1 su $n$ lanci

$$p_{n,\pi}(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} \qquad (x = 0, 1, \ldots, n)$$

con media $n\pi$ e varianza $n\pi(1-\pi)$

• Frequenza della faccia 1 sugli $n$ lanci

$$f = \frac{x}{n} \qquad \left(f = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right)$$

• Probabilità di una frequenza $f$ osservata su $n$ lanci

$$p_{n,\pi}(f) = \frac{n!}{(nf)!\,[n(1-f)]!}\pi^{nf}(1-\pi)^{n(1-f)}$$

con media $\frac{1}{n}n\pi = \pi$ e varianza $\frac{1}{n^2}n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$

$$H_0 \text{ vera} \quad \Longrightarrow \quad \pi = 1/6$$

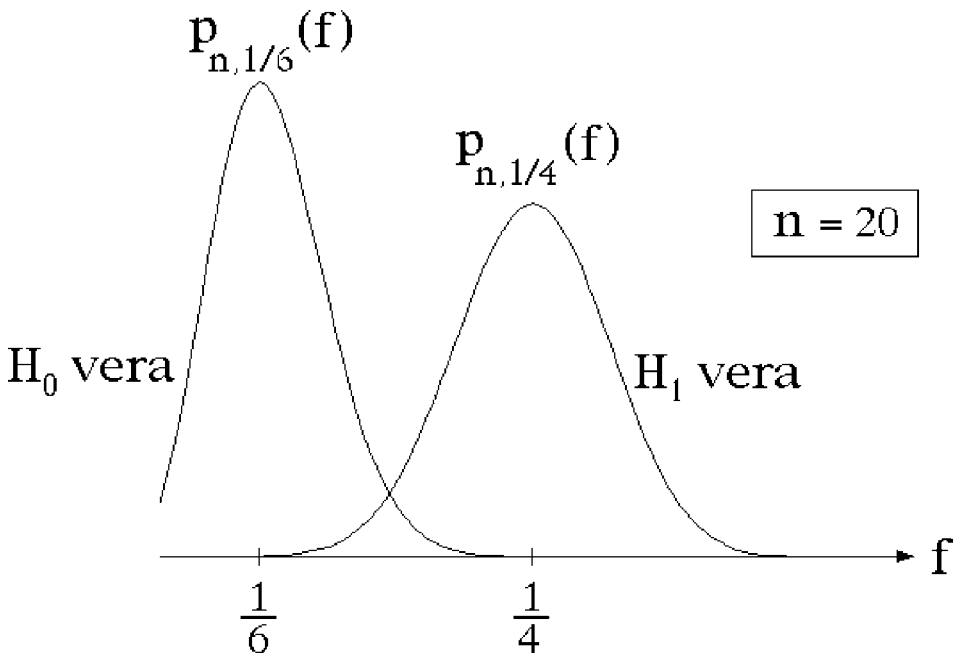
$$H_1 \text{ vera} \quad \Longrightarrow \quad \pi = 1/4$$

$$\Downarrow$$

$p_{n,1/6}(f)$: distribuzione di probabilità di $f$ se è vera $H_0$

$p_{n,1/4}(f)$: distribuzione di probabilità di $f$ se è vera $H_1$

Andamento delle due distribuzioni per $n = 20$



Strategia del test

$f$ è la **variabile del test**

se $f$ è abbastanza piccola **si accetta** $H_0$ (e rigetta $H_1$)

se $f$ è abbastanza grande **si rigetta** $H_0$ (e accetta $H_1$)

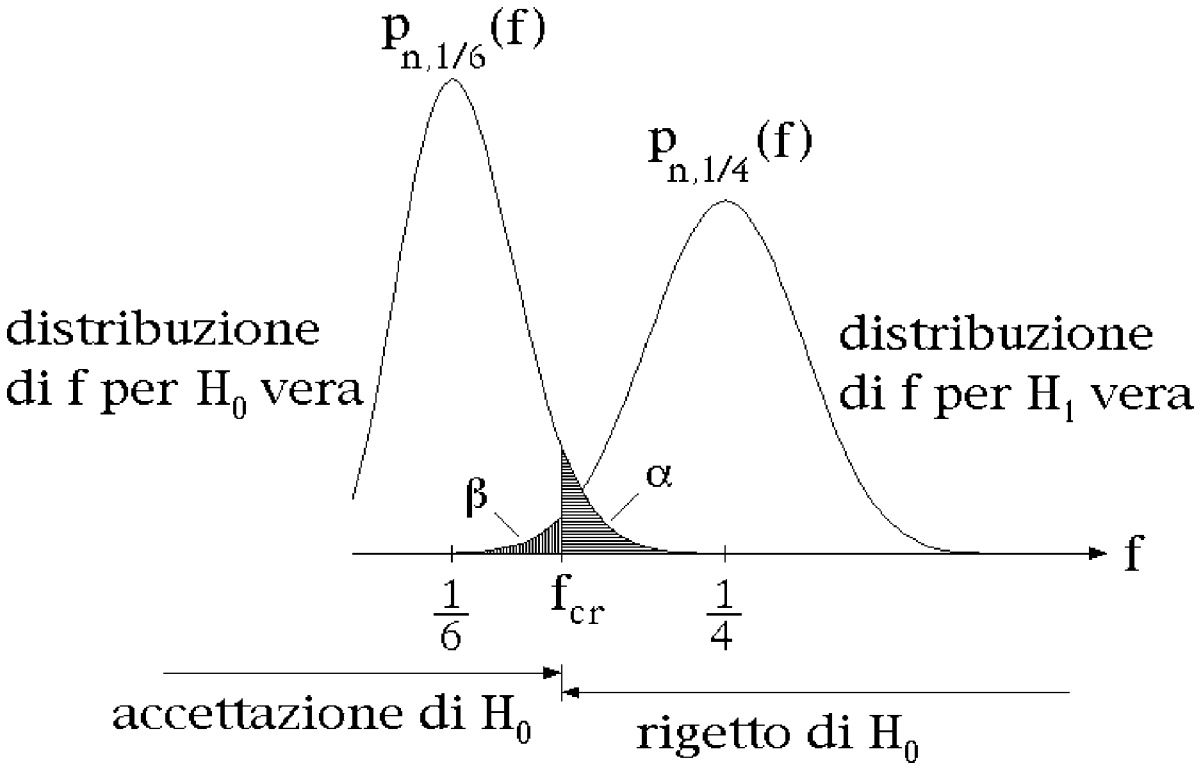
Si definisce un valore critico $f_{\text{cr}}$ di $f$ tale che

$f \leq f_{\text{cr}} \quad \Longrightarrow \quad$ **accettazione di** $H_0$

$f > f_{\text{cr}} \quad \Longrightarrow \quad$ **rigetto di** $H_0$

$\{f \in [0,1] : \ f \leq f_{\text{cr}}\}$ è la **regione di accettazione**

$\{f \in [0,1] : \ f > f_{\text{cr}}\}$ è la **regione di rigetto**



Interpretazione delle aree tratteggiate

$\alpha$: probabilità di rigettare $H_0$ nonostante sia vera
— probabilità di **errore del I tipo**

$\beta$: probabilità di accettare $H_0$ benché falsa
— probabilità di **errore del II tipo**

$\alpha$: **"livello di significatività"** del test

$1 - \beta$: **"potenza"** del test
(probabilità di riconoscere come falsa l'ipotesi nulla, qualora questa lo sia effettivamente)

**N.B.** Conviene che $\alpha$ e $\beta$ siano entrambi piccoli

In tal modo:

la probabilità $\alpha$ di errore del I tipo è piccola;

la "potenza" $1 - \beta$ del test è elevata ($\sim 1$).

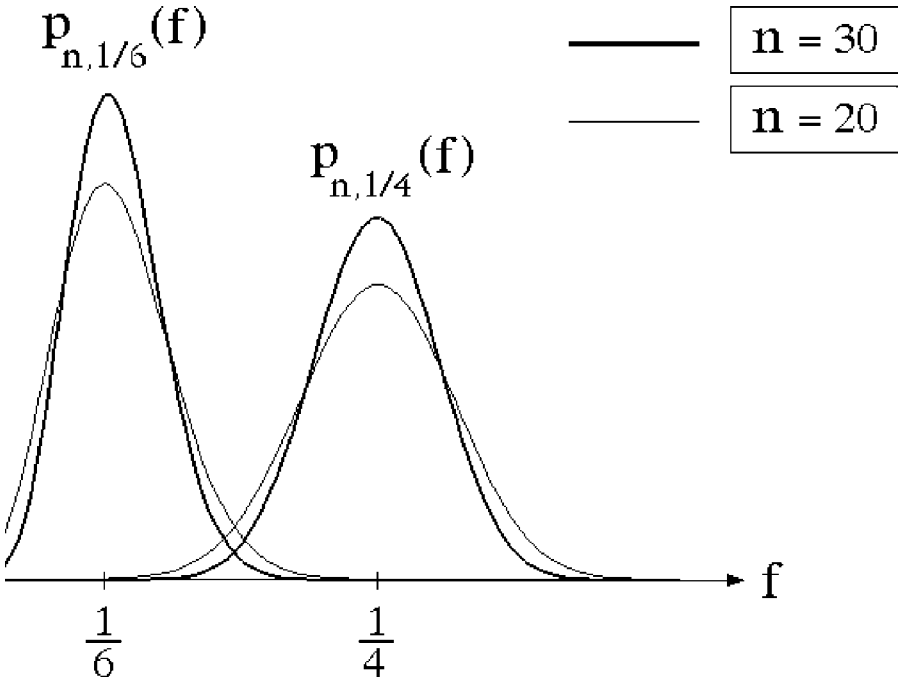
Non è però possibile ridurre tanto $\alpha$ quanto $\beta$ modificando il solo valore di $f_{\text{cr}}$

se $f_{\text{cr}}$ cresce $\implies$ $\alpha$ decresce, ma $\beta$ cresce;

se $f_{\text{cr}}$ decresce $\implies$ $\alpha$ cresce e $\beta$ diminuisce;

Per ridurre $\alpha$ **e** $\beta$ occorre considerare un **campione più ampio** (incrementare $n$)

Al crescere di $n$ la varianza $\pi(1-\pi)/n$ di $f$ si riduce e la distribuzione $p_{n,\pi}(f)$ si concentra maggiormente attorno al valor medio $\pi$




In questo esempio **il verificarsi dell'ipotesi alternativa** $H_1$ : $\pi = 1/4$ **caratterizza completamente la distribuzione della popolazione**

Un'ipotesi che al proprio verificarsi specifica completamente la distribuzione di probabilità della variabile del test si dice **semplice**; in caso contrario si parla di **ipotesi composta**

**Il calcolo di** $\alpha$ **e** $\beta$ **è quindi possibile solo se entrambe le ipotesi** $H_0$ **e** $H_1$ **sono semplici**

## □ Secondo esempio illustrativo

• Processo da studiare:
**lancio ripetuto di un dado** (lanci indipendenti)

$H_0$: la faccia 1 ha una probabilità di uscita $\pi = 1/6$
(dado regolare)

$H_1$: la faccia 1 ha una probabilità di uscita $\pi > 1/6$
(dado truccato in modo da favorire l'uscita della faccia 1)

• Se $\pi$ è la reale probabilità di uscita della faccia 1 in un singolo lancio, la probabilità che su $n$ lanci si osservi la faccia 1 con una frequenza $f = 0, 1/n, 2/n, \ldots, (n-1)/n, 1$ vale sempre

$$p_{n,\pi}(f) \;=\; \frac{n!}{(nf)!\,[n(1-f)]!}\pi^{nf}(1-\pi)^{n(1-f)}$$

con media $\pi$ e varianza $\pi(1-\pi)/n$. La frequenza $f$ viene scelta ancora come **variabile del test**

• Le regioni di accettazione e rigetto di $H_0$ saranno comunque del tipo:

$$\{f \in [0,1] \,:\, f \leq f_{\rm cr}\} \qquad \text{e} \qquad \{f \in [0,1] \,:\, f > f_{\rm cr}\}$$

• In questo caso però $H_1 \,:\, \pi > 1/6$ **è un'ipotesi composta**: al suo verificarsi **la distribuzione di $f$ non è specificata completamente**

• Ne segue che **la potenza $1-\beta$ del test non può essere calcolata**

**Per $n$ ed $\alpha$ fissati** ($\Leftrightarrow$ $f_{cr}$ assegnato) si possono però determinare:

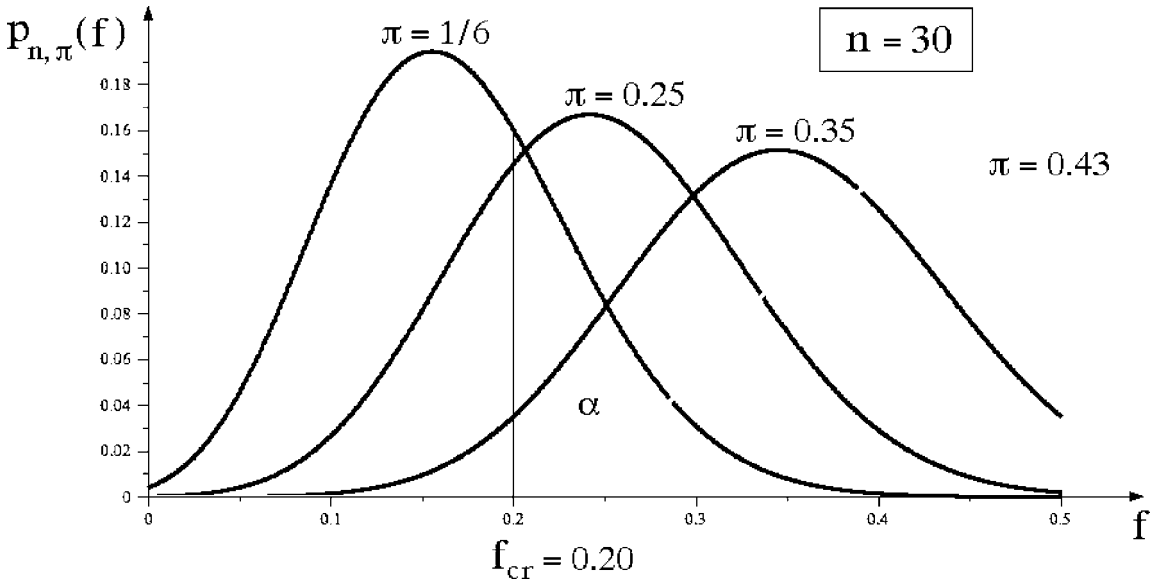
- la **curva operativa caratteristica** del test, e
- la **funzione di potenza** del test

**• Curva operativa caratteristica (curva OC)**
È il grafico della funzione

$$\pi \in (0,1) \longrightarrow \beta(\pi)$$

ottenuta calcolando la probabilità $\beta$ per una distribuzione alternativa corrispondente a un qualsiasi valore assegnato di $\pi$

Dal grafico della funzione $p_{n,\pi}(f)$ è evidente che all'aumentare di $\pi$ il picco della distribuzione si sposta sempre più verso destra ed il valore di $\beta(\pi)$ diminuisce progressivamente




rappresentando l'area compresa fra il grafico della distribuzione e l'intervallo di accettazione $\{0 \leq f \leq f_{cr}\}$. **La funzione $\beta(\pi)$ è monotòna decrescente**

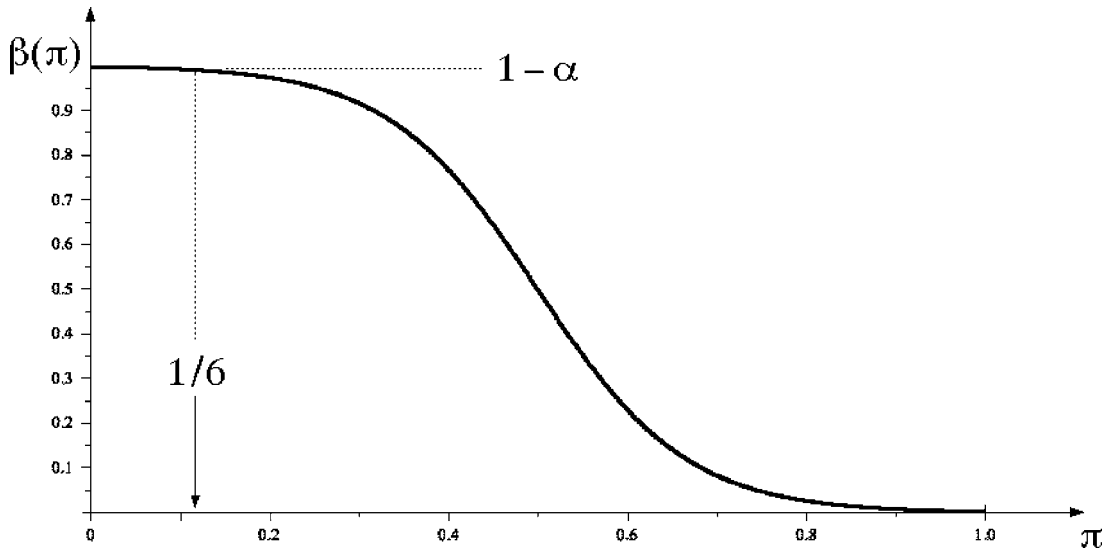
È chiaro inoltre che:

- $\beta(\pi)$ tende a 1 per $\pi$ tendente a 0, in quanto la distribuzione di picca completamente a sinistra di $f_{cr}$;

- $\beta(\pi)$ tende a 0 quando $\pi$ si approssima a 1, perchè in tal caso la distribuzione si colloca completamente a destra del valore critico $f_{cr}$;

- $\beta(1/6) = 1-\alpha$, poichè la distribuzione alternativa coincide con quella per $H_0$ vera e quindi $\beta + \alpha = 1$.

La curva operativa caratteristica presenta dunque l'andamento illustrato nella figura seguente:



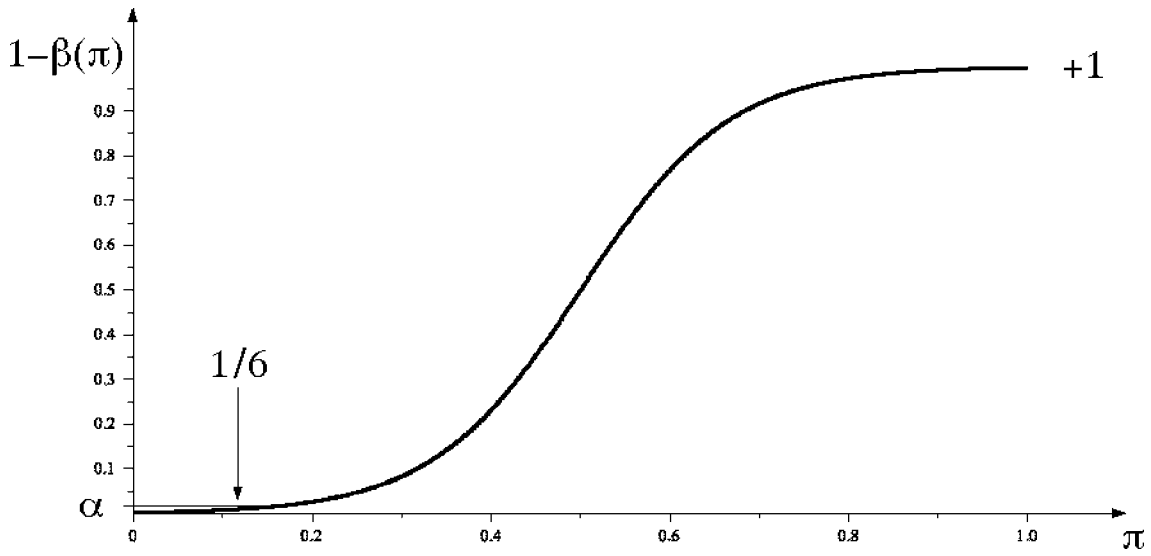
• **Funzione di potenza (curva di potenza)**
È definita come il grafico della funzione

$$\pi \in (0,1) \longrightarrow 1 - \beta(\pi)$$

ed il suo andamento si desume immediatamente da quello della curva operativa caratteristica. In particolare, la funzione:

- è monotòna crescente;
- tende a 0 per $\pi = 0$;
- soddisfa $1 - \beta(1/6) = \alpha$;
- converge a 1 in $\pi = 1$,

per cui presenta un grafico della forma seguente




Si osservi come la potenza del test cresca all'aumentare del valore vero di $\pi$: il test diventa molto efficiente nel riconoscere l'ipotesi $H_0$ quando $\pi$ è significativamente maggiore del valore testato 1/6, e dunque meglio distinguibile da questo.

## □ Terzo esempio illustrativo

• Processo da studiare:
**lancio ripetuto di un dado** (lanci indipendenti)

$H_0$: la faccia 1 ha una probabilità di uscita $\pi = 1/6$
(dado regolare)

$H_1$: la faccia 1 ha una probabilità di uscita $\pi \neq 1/6$
(dado irregolare, ma non truccato deliberatamente)

• La variabile del test è ancora la frequenza $f$ di uscita della faccia 1 su $n$ lanci. Se $\pi$ è la probabilità effettiva di uscita della faccia 1, $f$ segue sempre la distribuzione $p_{n,\pi}(f)$.

• Valori di $f$ sufficientemente prossimi a $1/6$ inducono a ritenere $H_0$ corretta; valori lontani da $1/6$ portano invece a preferire $H_1$. La regione di accettazione di $H_0$ sarà dunque un intervallo del tipo:

$$\{f \in [0,1] : \ f_{\mathrm{LCR}} \leq f \leq f_{\mathrm{UCR}}\}$$

contenente il punto $f = 1/6$ e quella di rigetto:

$$\{f \in [0,1] : \ f < f_{\mathrm{LCR}}\} \cup \{f \in [0,1] : \ f > f_{\mathrm{UCR}}\}$$

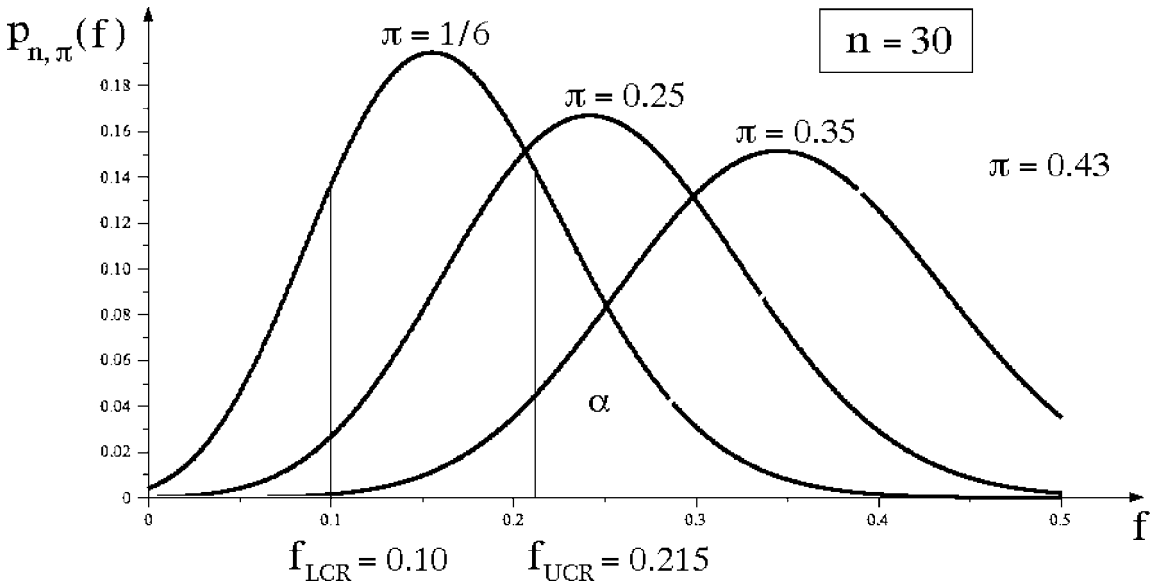
essendo $f_{\mathrm{LCR}}$, $f_{\mathrm{UCR}}$ due valori critici scelti in modo che il livello di significatività del test sia $\alpha$. **Il test è bilaterale**, o **a due code**, in quanto la regione di rigetto si compone di due intervalli disgiunti ("code").

• Anche in questo caso $H_1 : \pi \neq 1/6$ **è un'ipotesi composta**

**Per $n$ ed $\alpha$ fissati** ($\Leftrightarrow$ $f_{\text{LCR}}$ e $f_{\text{UCR}}$ assegnati) si possono ancora determinare la **curva operativa caratteristica** e la **funzione di potenza** del test, ma entrambe hanno un andamento diverso rispetto al caso del test unilaterale considerato nell'esempio precedente.

### • Curva operativa caratteristica

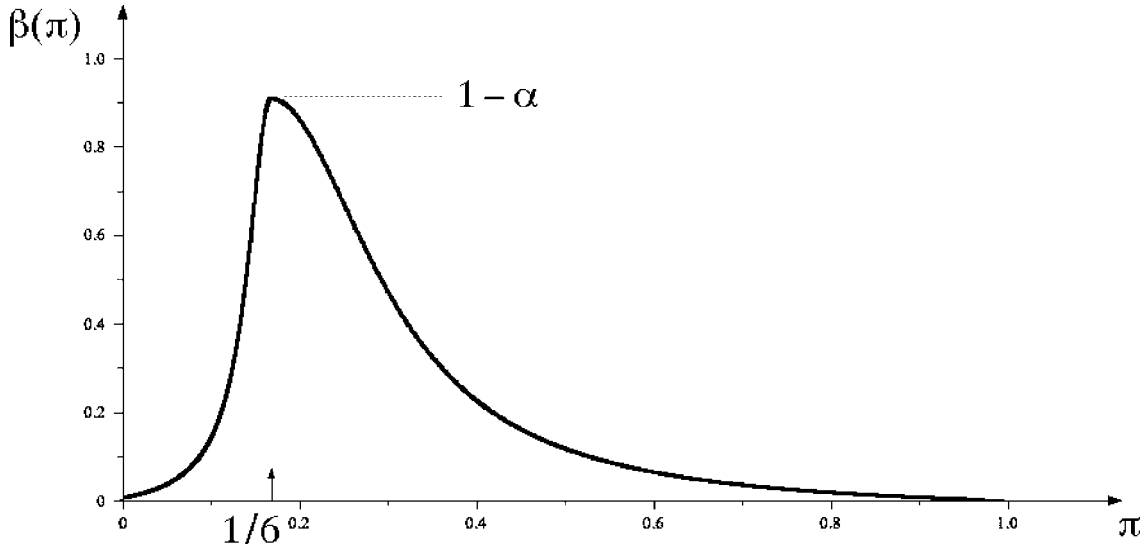
Il grafico di $p_{n,\pi}(f)$ mostra ancora che al crescere di $\pi$ il picco della distribuzione si sposta progressivamente verso destra. In questo caso, tuttavia, a causa della struttura a doppia coda della regione di rigetto la funzione $\beta(\pi)$ tende a 0 tanto per $\pi$ prossimo a 0 che per $\pi$ tendente a 1:



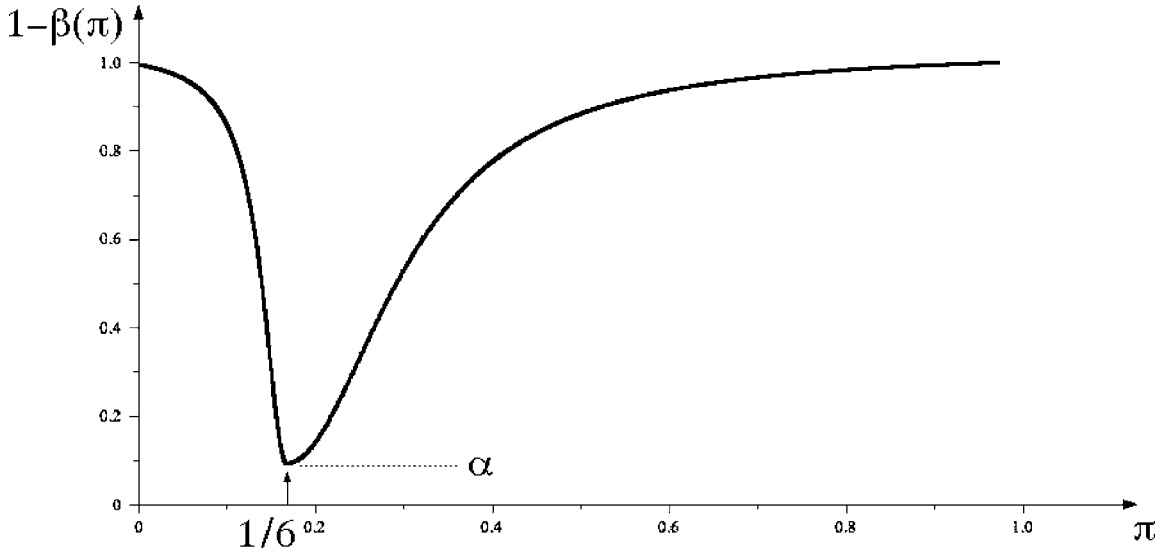

in quanto $\beta(\pi)$ rappresenta sempre l'area compresa fra il grafico della distribuzione e l'intervallo di accettazione $\{f_{\text{LCR}} \leq f \leq f_{\text{UCR}}\}$. Si ha inoltre, come nel caso precedente, $\beta(1/6) = 1-\alpha$.

Infine, la curva è crescente per $\pi \in (0, 1/6)$ e decrescente nell'intervallo $\pi \in (1/6, 1)$.

La curva operativa caratteristica presenta dunque l'andamento illustrato nella figura seguente:




• **Funzione di potenza**
Si ricava immediatamente dalla curva operativa caratteristica:




La potenza del test cresce quando il $\pi$ effettivo è lontano dal valore testato 1/6, e dunque meglio distinguibile da questo.

**Osservazione**: **Ipotesi nulla $H_0$ composta**

Potrebbe avere interesse considerare una **ipotesi nulla composta**, come ad esempio

$$H_0 \ : \ \pi \leq 1/6$$

contro l'ipotesi alternativa $H_1 \ : \ \pi > 1/6$.

In questo caso **se $H_0$ è vera, la distribuzione della variabile del test $f$ non è specificata in modo univoco**, potendo risultare corretto qualsiasi valore di $\pi \leq 1/6$.

Al livello di significatività $\alpha$ deve essere allora attribuito un significato diverso rispetto al caso di $H_0$ semplice.

Le regioni di accettazione e di rigetto assumono la forma

$$\{f \ \leq \ f_{cr}\} \qquad \text{e} \qquad \{f \ > \ f_{cr}\}$$

dove $f_{cr}$ è un valore critico scelto appropriatamente (un poco superiore a 1/6, per esempio).

Ad $n$ fissato la probabilità che per $H_0$ vera la variabile $f$ appartenga alla regione di rigetto $\{f > f_{cr}\}$ dipenderà allora dal valore effettivo di $\pi$:

$$\alpha(\pi) \ = \ \sum_{f > f_{cr}} p_{n,\pi}(f)$$

Il livello di significatività del test $\alpha$ viene identificato con il massimo delle precedenti probabilità:

$$\alpha \ = \ \max_{\pi \leq 1/6} \sum_{f > f_{cr}} p_{n,\pi}(f)$$

**Il livello di significatività $\alpha$ rappresenta la massima probabilità di rifiutare l'ipotesi $H_0$ in realtà corretta.**

## 8.1 t-test sulla media di una popolazione normale

Ha lo scopo di verificare se la media $\mu$ di una **popolazione normale** sia rispettivamente uguale, minore o maggiore di un valore prefissato $\mu_0$

Trattandosi di popolazione normale, la variabile del test è

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

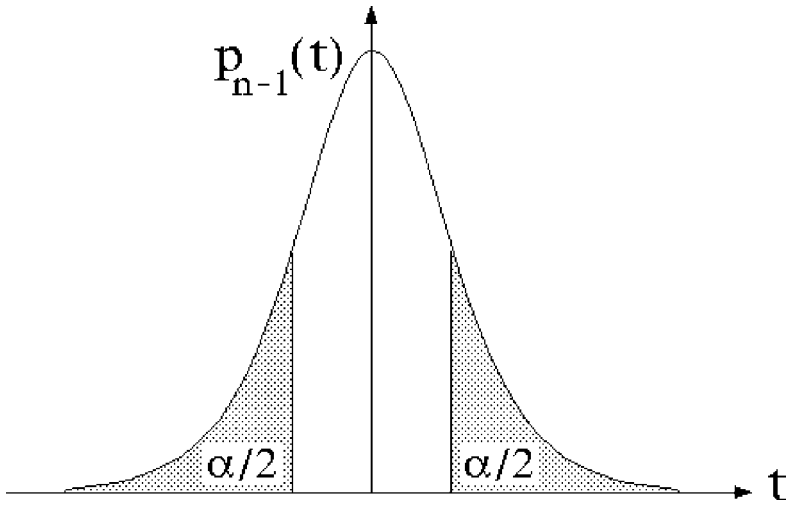
che per $\mu = \mu_0$ segue una distribuzione di Student a $n-1$ g.d.l.

Si possono distinguere tre casi

$(i)$ $H_0$: $\mu = \mu_0$ contro $H_1$: $\mu \neq \mu_0$

Il test è bilaterale. La regione di rigetto è l'unione di due intervalli simmetricamente disposti rispetto all'origine $t = 0$

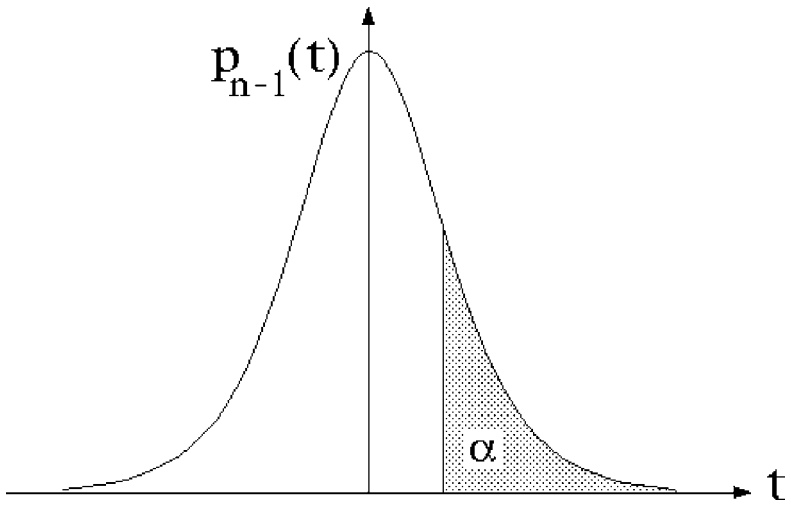
$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](n-1)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n-1)}\right\}$$



(*ii*) $H_0$: $\mu = \mu_0$ contro $H_1$: $\mu > \mu_0$

Nella fattispecie il test è unilaterale. La regione di rigetto consiste in un intervallo di valori sufficientemente grandi di $t$. Precisamente

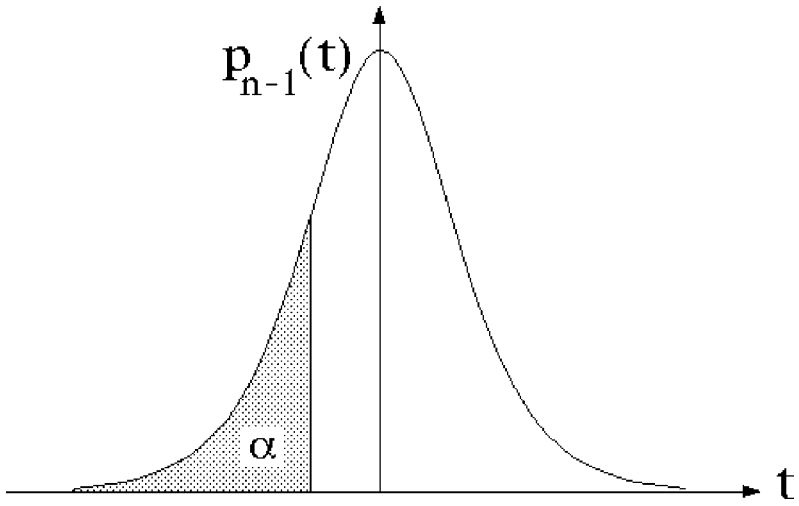
$$\left\{ t \geq t_{[1-\alpha](n-1)} \right\}$$




(*iii*) $H_0$: $\mu = \mu_0$ contro $H_1$: $\mu < \mu_0$

Altro test unilaterale, ma con una regione di rigetto costituita da un intervallo di valori sufficientemente piccoli di $t$

$$t \leq t_{[\alpha](n-1)} = -t_{[1-\alpha](n-1)}$$



**Esempio: CI e test delle ipotesi sulla media**

Si considerino i dati della tabella seguente:

| $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|
| 1 | 449 | 16 | 398 |
| 2 | 391 | 17 | 472 |
| 3 | 432 | 18 | 449 |
| 4 | 459 | 19 | 435 |
| 5 | 389 | 20 | 386 |
| 6 | 435 | 21 | 388 |
| 7 | 430 | 22 | 414 |
| 8 | 416 | 23 | 376 |
| 9 | 420 | 24 | 463 |
| 10 | 381 | 25 | 344 |
| 11 | 417 | 26 | 353 |
| 12 | 407 | 27 | 400 |
| 13 | 447 | 28 | 438 |
| 14 | 391 | 29 | 437 |
| 15 | 480 | 30 | 373 |

Assumendo la popolazione normale, si vuole verificare l'ipotesi nulla che la media $\mu$ sia $\mu_0 = 425$. contro l'ipotesi alternativa che $\mu \neq 425$, con un livello di significatività $\alpha = 2\%$.

**Soluzione**

Numero dei dati: $n = 30$

Media campionaria: $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = 415.66$

Varianza campionaria: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1196.44$

Deviazione standard campionaria: $s = \sqrt{s^2} = 34.59$

La variabile del test è

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

e la regione di rigetto bilaterale si scrive

$$\{t < -t_{[1-\frac{\alpha}{2}](n-1)}\} \cup \{t > t_{[1-\frac{\alpha}{2}](n-1)}\}$$

con $\alpha = 0.02$, $n = 30$.

Si calcola il valore della variabile del test

$$t = \frac{415.66 - 425}{34.59/\sqrt{30}} = -1.47896$$

ed il $t$-value della stessa variabile per l'$\alpha$ assegnato

$$t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.99](29)} = 2.462$$

cosicchè la regione di rigetto diventa

$$\{t < -2.462\} \cup \{t > 2.462\}$$

e **non contiene** il valore di $t$.

**L'ipotesi $\mu = 425$ non può essere rigettata** sulla base del campione disponibile e con livello di significatività 0.02.

**Osservazione - Intervallo di confidenza per $\mu$**

Il CI della media presenta i limiti inferiore e superiore:

$$\overline{x} - t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 - 2.462 \cdot \frac{34.59}{\sqrt{30}} = 400.11$$
$$\overline{x} + t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 + 2.462 \cdot \frac{34.59}{\sqrt{30}} = 431.21$$

e si riduce quindi a

$$400.11 \leq \mu \leq 431.21$$

La media ipotizzata $\mu_0 = 425$ **appartiene effettivamente** all'intervallo di confidenza di $\mu$.

**In generale:**

La variabile del test assume valore nella regione di accettazione di $H_0 : \mu = \mu_0$ se e soltanto se il valore testato $\mu_0$ della media appartiene all'intervallo di confidenza di $\mu$ (il livello di significatività del test, $\alpha$, e il livello di confidenza del CI, $1-\alpha$, sono complementari a 1).

## 8.2 $\mathcal{X}^2$-test sulla varianza di una popolazione normale

Serve ad accertare se la varianza $\sigma^2$ di una **popolazione normale** sia rispettivamente uguale, minore o maggiore di un valore $\sigma_0^2$ assegnato

La variabile del test è data da

$$\mathcal{X}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

che per una popolazione normale e $\sigma^2 = \sigma_0^2$ segue una distribuzione di $\mathcal{X}^2$ a $n-1$ g.d.l.

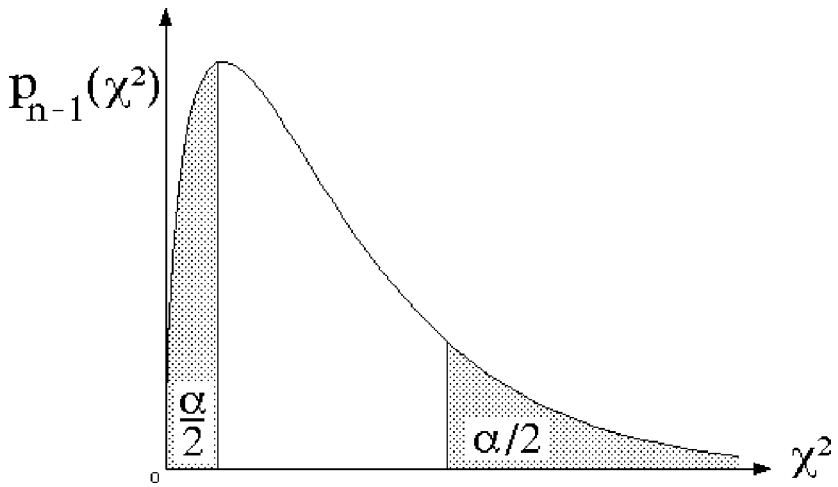
Di nuovo conviene distinguere tre casi

$(i)$ $H_0$: $\sigma^2 = \sigma_0^2$ contro $H_1$: $\sigma^2 \neq \sigma_0^2$

Test bilaterale con regione di rigetto di $H_0$

$$\left\{\mathcal{X}^2 \leq \mathcal{X}^2_{[\frac{\alpha}{2}](n-1)}\right\} \cup \left\{\mathcal{X}^2 \geq \mathcal{X}^2_{[1-\frac{\alpha}{2}](n-1)}\right\}$$

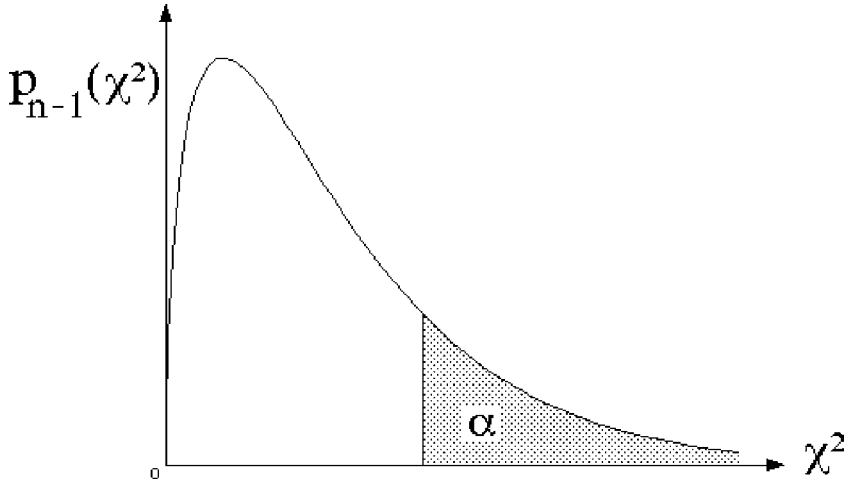
per cui l'ipotesi viene rigettata qualora $\mathcal{X}^2$ assuma valori o troppo piccoli o troppo grandi



(*ii*) $H_0$: $\sigma^2 = \sigma_0^2$ contro $H_1$: $\sigma^2 > \sigma_0^2$

Il test è unilaterale e la regione di rigetto corrisponde al ricorrere di valori molto grandi della variabile del test

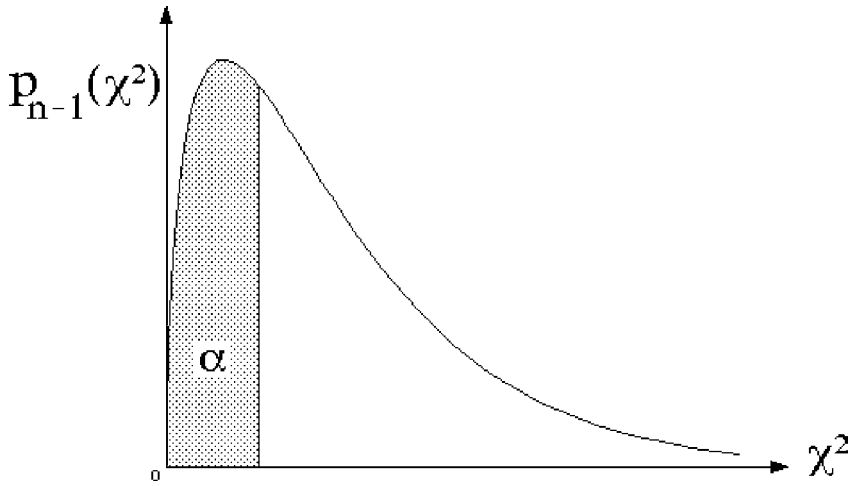
$$\left\{\mathcal{X}^2 \geq \mathcal{X}^2{}_{[1-\alpha](n-1)}\right\}$$




(*iii*) $H_0$: $\sigma^2 = \sigma_0^2$ contro $H_1$: $\sigma^2 < \sigma_0^2$

Anche in questo caso il test è di tipo unilaterale, la regione di rigetto corrispondendo a valori molto **piccoli** della variabile del test

$$\left\{\mathcal{X}^2 \leq \mathcal{X}^2{}_{[\alpha](n-1)}\right\}$$



**Esempio**: **CI e test delle ipotesi sulla varianza**
Misure ripetute della forza elettromotrice ai capi di una cella fotovoltaica, esposta a radiazione di intensità e spettro costanti, hanno prodotto la seguente tabella di risultati (in V), che possono assumersi estratti da una popolazione normale:

| $i$ | $V_i$ | $i$ | $V_i$ |
|---|---|---|---|
| 1 | 1.426 | 16 | 1.497 |
| 2 | 1.353 | 17 | 1.484 |
| 3 | 1.468 | 18 | 1.466 |
| 4 | 1.379 | 19 | 1.383 |
| 5 | 1.481 | 20 | 1.411 |
| 6 | 1.413 | 21 | 1.428 |
| 7 | 1.485 | 22 | 1.358 |
| 8 | 1.476 | 23 | 1.483 |
| 9 | 1.498 | 24 | 1.473 |
| 10 | 1.475 | 25 | 1.482 |
| 11 | 1.467 | 26 | 1.435 |
| 12 | 1.478 | 27 | 1.424 |
| 13 | 1.401 | 28 | 1.521 |
| 14 | 1.492 | 29 | 1.399 |
| 15 | 1.376 | 30 | 1.489 |

**(a)** Determinare l'intervallo di confidenza (CI) della varianza, con un livello di confidenza del 98%. **(b)** Verificare inoltre, con un livello di significatività del 2%, l'ipotesi che sia $\sigma^2 = \sigma_0^2 = 32.0 \cdot 10^{-4}$.

**Soluzione**
Per prima cosa si determinano le stime campionarie della media, della varianza e della deviazione standard.

Numero dei dati: $n = 30$

Media campionaria: $\overline{V} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} V_i = 1.4467$

Varianza campionaria:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^2 = 0.00221318$$

Deviazione standard campionaria: $s = \sqrt{s^2} = 0.04704448$

Si può ora procedere a determinare il CI della varianza ed a verificare l'ipotesi che sia $\sigma^2 = \sigma_0^2 = 32.0 \cdot 10^{-4}$.

**(a)** Il CI della varianza assume la forma

$$\frac{1}{\mathcal{X}^2_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{\mathcal{X}^2_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

con $\alpha = 0.02$ e $n = 30$. Pertanto

$$\frac{29 \cdot s^2}{\mathcal{X}^2_{[0.99](29)}} = \frac{29 \cdot 0.00221318}{49.588} = 12.94311 \cdot 10^{-4}$$

mentre

$$\frac{29\, s^2}{\mathcal{X}^2_{[0.01](29)}} = \frac{29 \cdot 0.00221318}{14.256} = 45.02125 \cdot 10^{-4}$$

e il CI diventa

$$12.94311 \cdot 10^{-4} \leq \sigma^2 \leq 45.02125 \cdot 10^{-4}\,.$$

**(b)** La variabile appropriata per effettuare il test è

$$\mathcal{X}^2 \;=\; (n-1)s^2/\sigma_0^2$$

che se l'ipotesi nulla $H_0 \,:\, \sigma^2 = \sigma_0^2$ risulta corretta segue una distribuzione di $\mathcal{X}^2$ a $n-1$ g.d.l.

Il livello di significatività richiesto è $\alpha = 2\% = 0.02$

Assumendo come ipotesi alternativa $H_1 : \sigma^2 \neq \sigma_0^2$, la regione di rigetto assume la forma bilaterale

$$\left\{\mathcal{X}^2 \leq {\mathcal{X}^2}_{[\frac{\alpha}{2}](n-1)}\right\} \cup \left\{\mathcal{X}^2 \geq {\mathcal{X}^2}_{[1-\frac{\alpha}{2}](n-1)}\right\}$$

con $\alpha = 0.02$ e $n = 30$. Come prima, la tabella del $\mathcal{X}^2$ porge

$${\mathcal{X}^2}_{[\frac{\alpha}{2}](n-1)} \;=\; {\mathcal{X}^2}_{[0.01](29)} \;=\; 14.256$$

$${\mathcal{X}^2}_{[1-\frac{\alpha}{2}](n-1)} \;=\; {\mathcal{X}^2}_{[0.99](29)} \;=\; 49.588$$

in modo che la regione di rigetto di $H_0$ diventa

$$\left\{\mathcal{X}^2 \leq 14.256\right\} \cup \left\{\mathcal{X}^2 \geq 49.588\right\}$$

Nella fattispecie, la variabile del test prende il valore

$$\mathcal{X}^2 \;=\; \frac{(n-1)s^2}{\sigma_0^2} \;=\; \frac{29 \cdot 0.002213183}{32 \cdot 10^{-4}} \;=\; 20.057$$

che **non è contenuto** nella regione di rigetto: $H_0$ **non può essere rifiutata**.

Si osservi che accettare l'ipotesi $\sigma^2 = \sigma_0^2$ con livello di significatività $\alpha$ equivale a verificare che $\sigma_0^2$ appartiene all'intervallo di confidenza di $\sigma^2$ con livello di confidenza $1-\alpha$

## 8.3 F-test sul confronto delle varianze di due popolazioni normali indipendenti

Ha lo scopo di stabilire se due popolazioni normali indipendenti di varianze rispettive $\sigma_1^2$ e $\sigma_2^2$ abbiano o meno la stessa varianza

Si tratta quindi di testare l'ipotesi nulla $H_0$: $\sigma_1^2 = \sigma_2^2$ contro l'ipotesi alternativa $H_1$: $\sigma_1^2 \neq \sigma_2^2$

Si suppone di avere a disposizione due campioni ridotti, rispettivamente di $p$ e $q$ individui,

$$(y_1, y_2, \ldots, y_p) \qquad (z_1, z_2, \ldots, z_q)$$

estratti dalle due popolazioni indipendenti

Le medie stimate delle due popolazioni sono date da

$$\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i \qquad \overline{z} = \frac{1}{q}\sum_{j=1}^{q} z_j$$

mentre le corrispondenti stime delle varianze risultano

$$s_y^2 = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \overline{y})^2 \qquad s_z^2 = \frac{1}{q-1}\sum_{j=1}^{q}(z_j - \overline{z})^2$$

Trattandosi di campioni estratti da popolazioni normali **indipendenti**, le variabili casuali

$$\frac{(p-1)s_y^2}{\sigma_1^2} = \frac{1}{\sigma_1^2}\sum_{i=1}^{p}(y_i-\overline{y})^2 \qquad \frac{(q-1)s_z^2}{\sigma_2^2} = \frac{1}{\sigma_2^2}\sum_{j=1}^{q}(z_j-\overline{z})^2$$

costituiscono variabili di $\mathcal{X}^2$ **indipendenti** rispettivamente a $p-1$ e $q-1$ gradi di libertà

Se ne deduce che il quoziente

$$F = \frac{q-1}{p-1}\frac{(p-1)s_y^2}{\sigma_1^2}\Big(\frac{(q-1)s_z^2}{\sigma_2^2}\Big)^{-1} = \frac{\sigma_2^2}{\sigma_1^2}\frac{s_y^2}{s_z^2}$$

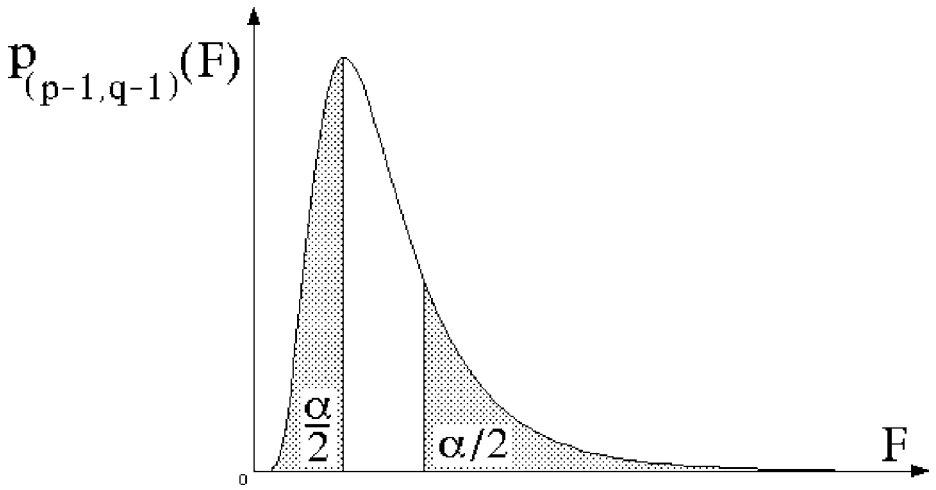
è per definizione una variabile $F$ di Fisher a $p-1$, $q-1$ gradi di libertà

Qualora l'ipotesi nulla $\sigma_1^2 = \sigma_2^2$ sia corretta, la variabile campionaria si riduce a

$$F = \frac{s_y^2}{s_z^2}$$

Valori molto piccoli o molto grandi della $F$ devo essere ritenuti indice di un quoziente $\sigma_1^2/\sigma_2^2$ significativamente diverso da 1, per cui la regione di rigetto di $H_0$ contro $H_1$ viene definita bilateralmente come

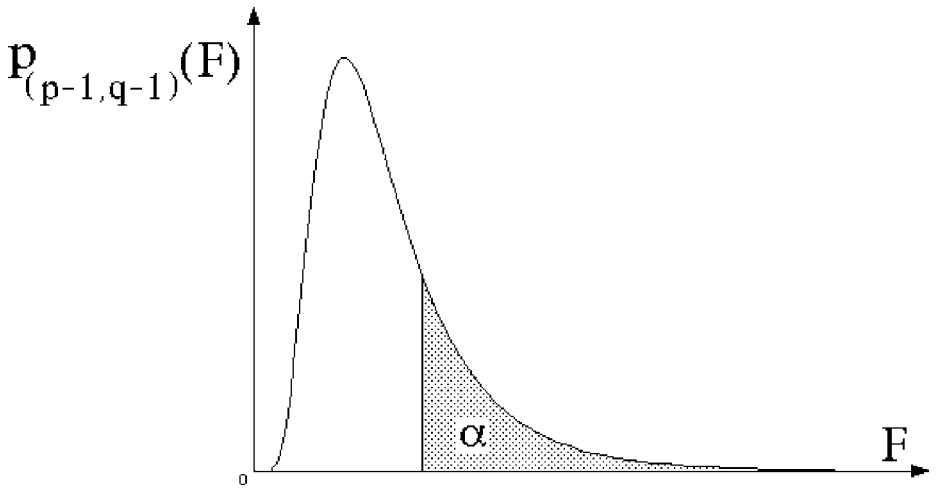
$$\Big\{F \le F_{[\frac{\alpha}{2}](p-1,q-1)}\Big\} \cup \Big\{F \ge F_{[1-\frac{\alpha}{2}](p-1,q-1)}\Big\}$$




Se l'ipotesi alternativa è $H_1$: $\sigma_1^2 > \sigma_2^2$ la regione di rigetto viene definita unilateralmente

$$\left\{F \geq F_{[1-\alpha](p-1,q-1)}\right\}$$

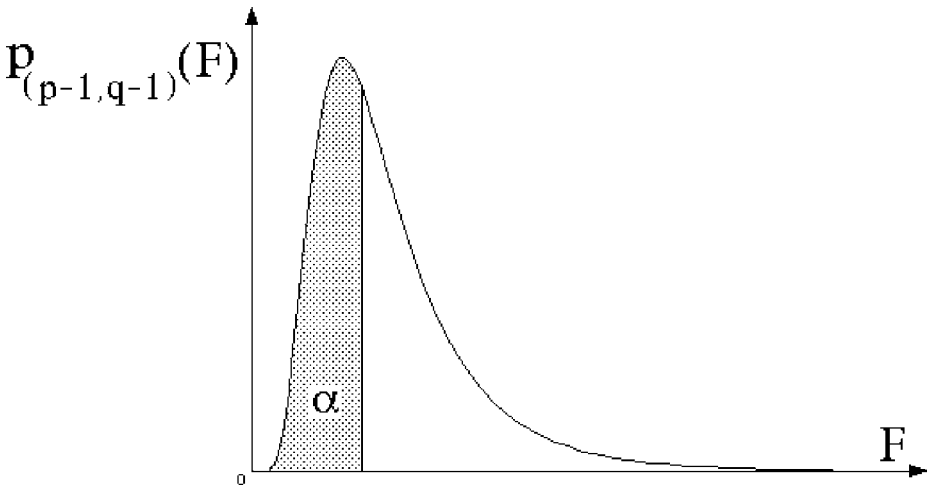
poiché valori elevati del quoziente $s_y^2/s_z^2$ devono essere associati al verificarsi dell'ipotesi $H_1$



Nel caso infine che l'ipotesi alternativa sia $H_1$: $\sigma_1^2 < \sigma_2^2$ la regione di rigetto è ancora definita unilateralmente

$$\left\{F \leq F_{[\alpha](p-1,q-1)}\right\}$$

e saranno valori piccoli del quoziente $s_y^2/s_z^2$ ad indicare il verificarsi dell'ipotesi $H_1$



**Esempio**: **F-test sul confronto delle varianze di due popolazioni normali indipendenti**

Per catalizzare una reazione chimica si deve scegliere fra due diversi catalizzatori, indicati schematicamente con $A$ e $B$. Al fine di controllare se la varianza sul quantitativo del prodotto di reazione sia la stessa o meno, per i due catalizzatori, se eseguono $p = 10$ esperimenti con $A$ e $q = 12$ esperimenti con $B$, ottenendo le seguenti stime campionarie:

$$s_A^2 = 0.1405 \qquad s_B^2 = 0.2820$$

Ad un livello di significatività del 5%, è possibile rigettare l'ipotesi che $\sigma_A^2 = \sigma_B^2$? Le due popolazioni di dati si assumono indipendenti e normali.

**Soluzione**

Si può utilizzare la variabile del test

$$F = s_A^2/s_B^2$$

che segue una distribuzione di Fisher con $(p-1, q-1) = (9, 11)$ g.d.l. nel caso che l'ipotesi nulla $H_0 : \sigma_A^2 = \sigma_B^2$ sia soddisfatta. Si ha

$$F = 0.1405/0.2820 = 0.49822695$$

La regioe di rigetto si scrive

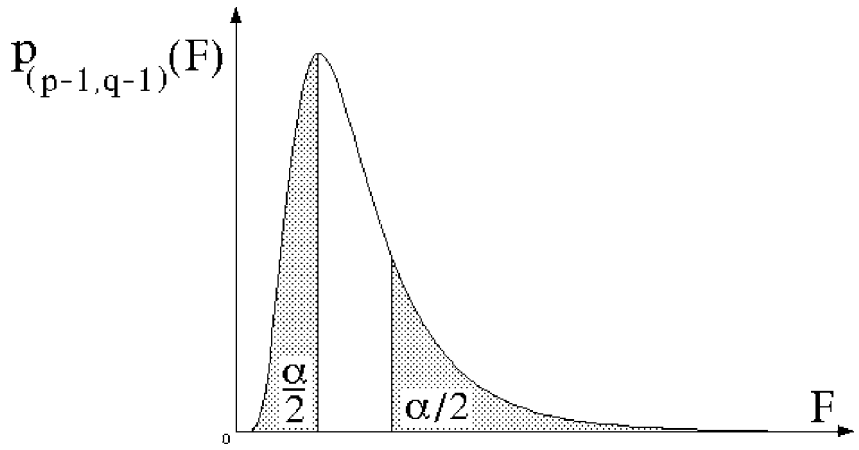
$$\{F \leq F_{[\frac{\alpha}{2}](p-1,q-1)}\} \cup \{F \geq F_{[1-\frac{\alpha}{2}](p-1,q-1)}\}$$

ossia (poichè $p = 10$, $q = 12$ e $\alpha = 0.05$)

$$\{F \leq F_{[0.025](9,11)}\} \cup \{F \geq F_{[0.975](9,11)}\}$$

dove

$$F_{[0.025](9,11)} = 0.2556188 \qquad F_{[0.975](9,11)} = 3.587898$$




Gli $F$-value non sono disponibili sulla tabella delle distribuzioni cumulative di $F$, ma possono essere calcolati facilmente per via numerica. Ad esempio, mediante i seguenti comandi Maple$^{\text{TM}}$

$$stateval f[icdf, fratio[9, 11]](0.025)$$

$$stateval f[icdf, fratio[9, 11]](0.975)$$

Nella fattispecie

$$0.49822695 \notin \{F \leq 0.2556188\} \cup \{F \geq 3.587898\}$$

per cui si deve concludere che i campioni disponibili non possono escludere che le varianze delle due popolazioni siano effettivamente uguali.

**L'ipotesi $H_0$ non può essere rigettata!**

## 8.4 Test per il confronto delle medie di due popolazioni normali indipendenti (varianze note)

Le stime campionarie delle medie

$$\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i \qquad \overline{z} = \frac{1}{q}\sum_{j=1}^{q} z_j$$

sono variabili casuali normali con media $\mu_1$ e $\mu_2$, mentre le rispettive varianze si scrivono

$$\frac{\sigma_1^2}{p} \qquad \text{e} \qquad \frac{\sigma_2^2}{q}$$

Di conseguenza, qualora sia $\mu_1 = \mu_2$ la variabile casuale

$$z = \frac{\overline{y} - \overline{z}}{\sqrt{\dfrac{\sigma_1^2}{p} + \dfrac{\sigma_2^2}{q}}}$$

è una variabile normale standard $N(0,1)$. Con un livello di significatività $\alpha \in (0,1)$, le regione di rigetto bilaterale dell'ipotesi nulla $H_0$: $\mu_1 = \mu_2$ contro l'alternativa $H_1$: $\mu_1 \neq \mu_2$ può essere espressa come

$$\{z \leq -z_\alpha\} \cup \{z_\alpha \leq z\}$$

il valore critico $z_\alpha > 0$ essendo definito dall'equazione

$$\frac{\alpha}{2} = \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_0^{z_\alpha} e^{-\xi^2/2} d\xi = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\Big(\frac{z_\alpha}{\sqrt{2}}\Big)$$

ossia da $1 - \alpha = \mathrm{erf}(z_\alpha/\sqrt{2})$

## 8.5 t-test disaccoppiato per il confronto delle medie di due popolazioni normali indipendenti (varianze incognite)

Date due popolazioni normali di medie $\mu_1$, $\mu_2$, si vuole testare l'ipotesi $H_0$: $\mu_1 = \mu_2$ contro l'ipotesi alternativa $H_1$: $\mu_1 \neq \mu_2$

Occorre distinguere il caso che le popolazioni abbiano la stessa varianza $\sigma_1^2 = \sigma_2^2$ e quello di varianze distinte $\sigma_1^2 \neq \sigma_2^2$

Il ricorrere della condizione $\sigma_1^2 = \sigma_2^2$ o di quella alternativa $\sigma_1^2 \neq \sigma_2^2$ può essere accertato mediante l'appropriato $F$-test, già esaminato (le varianze sono incognite!)

### ($i$) Caso di varianze uguali: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Se $\mu_1 = \mu_2$ la variabile casuale

$$t = \frac{\overline{y} - \overline{z}}{s\sqrt{\frac{1}{p} + \frac{1}{q}}}$$

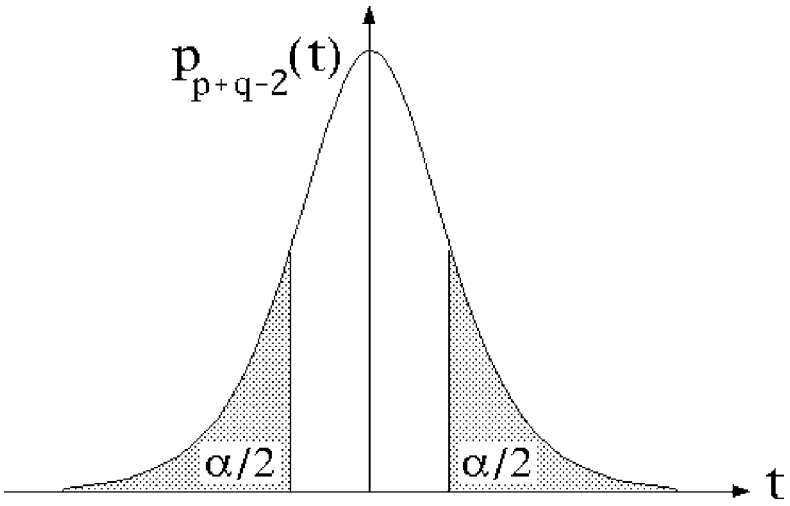
in cui si è introdotta la media ponderata delle varianze

$$s^2 = \frac{(p-1)s_y^2 + (q-1)s_z^2}{p+q-2}$$

segue una distribuzione di Student a $p+q-2$ gradi di libertà

La $t$ sarà quindi utilizzabile come variabile del test con una zona di rigetto costituita dall'unione di due intervalli, l'uno corrispondente a valori molto piccoli della variabile, l'altro a valori molto grandi della stessa $t$

$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](p+q-2)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](p+q-2)}\right\}$$




(*ii*) **Caso di varianze diverse** $\sigma_1^2 \neq \sigma_2^2$
Se $\mu_1 = \mu_2$ la variabile casuale

$$t = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{1}{p}s_y^2 + \frac{1}{q}s_z^2}}$$

segue **approssimativamente** una distribuzione di Student con un numero di gradi di libertà in generale **non intero**

$$n = \frac{\left(\frac{s_y^2}{p} + \frac{s_z^2}{q}\right)^2}{\frac{1}{p-1}\left(\frac{s_y^2}{p}\right)^2 + \frac{1}{q-1}\left(\frac{s_z^2}{q}\right)^2}$$

La distribuzione di Student **è comunque definita** e la regione di rigetto si scrive

$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](n)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n)}\right\}$$

Questo tipo di test è anche noto come **unpaired $t$-test**
Non si conoscono test esatti! (problema di Behrens-Fisher)

**Osservazione per il caso $\sigma_1^2 = \sigma_2^2$**
**Verifica che la variabile del test segue effettivamente una distribuzione di Student a $p+q-2$ gradi di libertà**

Basta osservare che per $\mu_1 = \mu_2$:

$(i)$ la variabile casuale

$$\xi \;=\; \frac{\overline{y} - \overline{z}}{\sigma\sqrt{\frac{1}{p} + \frac{1}{q}}} \;=\; \frac{\overline{y} - \overline{z}}{\sigma}\sqrt{\frac{pq}{p+q}}$$

segue una distribuzione gaussiana standard

$(ii)$ la variabile casuale

$$\mathcal{X}^2 \;=\; \frac{1}{\sigma^2}\Big[\sum_{i=1}^{p}(y_i - \overline{y})^2 + \sum_{j=1}^{q}(z_j - \overline{z})^2\Big]$$

è una variabile di $\mathcal{X}^2$ a $p-1+q-1 = p+q-2$ gradi di libertà

$(iii)$ le variabili $\xi$ e $\mathcal{X}^2$ sono stocasticamente indipendenti, come deducibile dal teorema di Craig

Di conseguenza, la variabile del test

$$\begin{aligned} t \;&=\; \sqrt{p+q-2}\frac{\xi}{\sqrt{\mathcal{X}^2}} \;= \\ &=\; \frac{\overline{y} - \overline{z}}{\sigma\sqrt{\frac{1}{p} + \frac{1}{q}}}\,\frac{\sqrt{p+q-2}}{\sigma^{-1}\sqrt{(p-1)s_y^2 + (q-1)s_z^2}} \;=\; \frac{\overline{y} - \overline{z}}{s\sqrt{\frac{1}{p} + \frac{1}{q}}} \end{aligned}$$

segue una distribuzione di Student a $p+q-2$ gradi di libertà

Una giustificazione del risultato si ottiene osservando che, posto

$$(x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}) \;=\; (y_1, \ldots, y_p, z_1, \ldots, z_q)$$

e

$$\overline{x} \;=\; \frac{1}{p+q}\sum_{k=1}^{p+q} x_k \;=\; \frac{1}{p+q}\Big[\sum_{i=1}^{p} y_i + \sum_{j=1}^{q} z_j\Big]$$

vale l'identità

$$\sum_{k=1}^{p+q}(x_k - \overline{x})^2 \;=\; \sum_{i=1}^{p}(y_i - \overline{y})^2 + \sum_{j=1}^{q}(z_j - \overline{z})^2 + \frac{pq}{p+q}(\overline{y} - \overline{z})^2$$

in cui:

$\dfrac{1}{\sigma^2}\displaystyle\sum_{k=1}^{p+q}(x_k - \overline{x})^2$ è una variabile di $\mathcal{X}^2$ a $p+q-1$ g.d.l.

$\dfrac{1}{\sigma^2}\Big[\displaystyle\sum_{i=1}^{p}(y_i - \overline{y})^2 + \sum_{j=1}^{q}(z_j - \overline{z})^2\Big]$ è una variabile di $\mathcal{X}^2$ a $p+q-2$ g.d.l.

$\dfrac{1}{\sigma^2}\dfrac{pq}{p+q}(\overline{y} - \overline{z})^2$ è una variabile di $\mathcal{X}^2$ a 1 g.d.l.

Le proprietà precedenti implicano l'indipendenza stocastica delle variabili casuali

$$\frac{1}{\sigma^2}\Big[\sum_{i=1}^{p}(y_i - \overline{y})^2 + \sum_{j=1}^{q}(z_j - \overline{z})^2\Big] \qquad \text{e} \qquad \frac{1}{\sigma^2}\frac{pq}{p+q}(\overline{y} - \overline{z})^2$$

e quindi, equivalentemente, di

$$\frac{1}{\sigma^2}\Big[\sum_{i=1}^{p}(y_i - \overline{y})^2 + \sum_{j=1}^{q}(z_j - \overline{z})^2\Big] \qquad \text{e} \qquad \frac{1}{\sigma}\sqrt{\frac{pq}{p+q}}(\overline{y} - \overline{z})$$

**Esempio: t-test disaccoppiato per il confronto di due medie (caso di varianze uguali incognite)**

Due processi sono caratterizzati dai valori elencati nella tabella seguente:

| Processo 1 | Processo 2 |
|---|---|
| 9 | 14 |
| 4 | 9 |
| 10 | 13 |
| 7 | 12 |
| 9 | 13 |
| 10 | 8 |
| | 10 |

Si vuole verificare, con un livello di significatività del 5%, l'ipotesi che le medie $\mu_1$ e $\mu_2$ dei due processi siano le stesse, assumendo popolazioni normali e varianze uguali ($\sigma_1^2 = \sigma_2^2$).

**Soluzione**
I due campioni consistono rispettivamente di $p = 6$ e $q = 7$ elementi. La media e la varianza del primo campione possono essere scritte nella forma:

$$\overline{x} = \frac{1}{p}\sum_{i=1}^{p} x_i = 8.17 \qquad s_x^2 = \frac{1}{p-1}\sum_{i=1}^{p}(x_i - \overline{x})^2 = 5.37$$

mentre quelle del secondo campione valgono:

$$\overline{y} = \frac{1}{q}\sum_{j=1}^{q} y_j = 11.29 \qquad s_y^2 = \frac{1}{q-1}\sum_{j=1}^{q}(y_j - \overline{y})^2 = 5.24$$

La varianza complessiva (pooled variance) dei due campioni viene allora calcolata nella forma:

$$s^2 = \frac{(p-1)s_x^2 + (q-1)s_y^2}{p+q-2} = \frac{5 \cdot 5.37 + 6 \cdot 5.24}{11} = 5.299$$

Come variabile del test, per verificare l'ipotesi $H_0 : \mu_1 = \mu_2$ si può usare il quoziente:

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{8.17 - 11.29}{\sqrt{5.299\left(\frac{1}{6} + \frac{1}{7}\right)}} = -2.436$$

Se l'ipotesi nulla è soddisfatta, la variabile del test segue una distribuzione di Student con $p+q-2$ gradi di libertà, in modo che la regione di accettazione bilaterale deve essere definita come:

$$|t| \leq t_{[1-\frac{\alpha}{2}](p+q-2)}$$

Nel caso specifico si ha $\alpha = 0.05$, $p = 6$, $q = 7$ e la regione di accettazione diventa

$$|t| \leq t_{[0.975](11)} = 2.201$$

Poichè $|t| = 2.436 > 2.201 = t_{[0.975](11)}$, i nostri campioni suggeriscono che l'ipotesi nulla $\mu_1 = \mu_2$ deve essere rigettata: **concludiamo che la differenza fra le medie dei due processi appare significativa.**

**Esempio**: **Confronto delle medie di due popolazioni normali indipendenti con varianze incognite ma presumibilmente uguali (unpaired $t$-test)**

Presso un centro di ricerche mediche, un gruppo di 22 volontari viene esposto agli stessi ceppi di virus influenzali e tenuto sotto stretto controllo medico. Una dosa stabilita e costante di vitamina C viene somministrata ad un campione casuale di $p = 10$ volontari. Un placebo, indistinguibile dal medicinale ma completamente inattivo, viene invece somministrato ai rimanenti $q = 12$ volontari. Per ciascun volontario viene registrata la durata (in giorni) della malattia. Ne risulta la lista seguente:

| vitamina C | placebo |
|---|---|
| 5.5 | 6.5 |
| 6.0 | 6.0 |
| 7.0 | 8.5 |
| 6.0 | 7.0 |
| 7.5 | 6.5 |
| 6.0 | 8.0 |
| 7.5 | 7.5 |
| 5.5 | 6.5 |
| 7.0 | 7.5 |
| 6.5 | 6.0 |
| | 8.5 |
| | 7.0 |

Le popolazioni si possono assumere indipendenti, normali e presumibilmente con la stessa varianza (incognita). Se vuole verificare, con un livello di significatività del 5%, l'ipotesi che la durata media della malattia sia la stessa per i volontari trattati

con vitamina C ($\mu_C$) e per quelli trattati con il placebo ($\mu_p$), contro l'ipotesi alternativa che $\mu_C < \mu_p$ — ossia che la somministrazione di vitamina C sia efficace nel ridurre la durata della malattia e dunque nel favorire la guarigione del paziente.

**Soluzione**
I due campioni consistono di $p = 10$ e $q = 12$ elementi, rispettivamente. Poichè l'ipotesi alternativa è unilaterale, si può applicare una versione unilaterale del $t$-test disaccoppiato per il confronto delle medie di due popolazioni normali indipendenti con la stessa varianza incognita, per controllare

$$H_0 \ : \ \mu_C \ = \ \mu_p \qquad \text{contro} \qquad H_1 \ : \ \mu_C \ < \ \mu_p$$

Indicate con $x_i$ le durate della malattia per il campione trattato con la vitamina e con $y_j$ le durate per i soggetti trattati con placebo, la media e la varianza del primo campione possono scriversi come:

$$\overline{x} \ = \ \frac{1}{p}\sum_{i=1}^{p} x_i \ = \ 6.4500 \qquad s_x^2 \ = \ \frac{1}{p-1}\sum_{i=1}^{p}(x_i - \overline{x})^2 \ = \ 0.5806$$

mentre quelle del secondo campione valgono:

$$\overline{y} \ = \ \frac{1}{q}\sum_{j=1}^{q} y_j \ = \ 7.1250 \qquad s_y^2 \ = \ \frac{1}{q-1}\sum_{j=1}^{q}(y_j - \overline{y})^2 \ = \ 0.7784$$

La varianza complessiva (pooled variance) dei due campioni viene allora calcolata come:

$$\begin{aligned} s^2 \ &= \ \frac{(p-1)s_x^2 + (q-1)s_y^2}{p+q-2} \ = \\ &= \ \frac{9 \ \cdot \ 0.5806 + 11 \ \cdot \ 0.7784}{20} \ = \ 0.6894 \end{aligned}$$

Per testare l'ipotesi $H_0 : \mu_C = \mu_p$, si introduce la variabile:

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{6.4500 - 7.1250}{\sqrt{0.6894\left(\frac{1}{10} + \frac{1}{12}\right)}} = -1.8987$$

Nel caso che l'ipotesi nulla sia verificata, tale variabile casuale è una $t$ di Student con $p + q - 2$ g.d.l., per cui la regione di rigetto unilaterale deve essere definita nel modo seguente:

$$t \leq t_{[\alpha](p+q-2)} = t_{[0.05](20)} = -t_{[0.95](20)} = -1.725$$

in quanto

$$t_{[\alpha](p+q-2)} = -t_{[1-\alpha](p+q-2)}$$

e

$$p = 10\,, \qquad q = 12\,, \qquad \alpha = 0.05\,.$$

Nella fattispecie si ha

$$t = -1.8987 < -1.725 = -t_{[0.95](20)}$$

I nostri campioni suggeriscono pertanto che l'ipotesi nulla $\mu_C = \mu_p$ debba essere rigettata a favore dell'ipotesi alternativa $\mu_C < \mu_p$: **si conclude che la differenza fra le medie delle due popolazioni appare significativa** e che presumibilmente la durata media $\mu_C$ della malattia sotto trattamento con vitamina C sia più breve di quella ($\mu_p$) registrata somministrando il placebo.

**Esempio**: **unpaired t-test su popolazioni normali con varianze incognite e presumibilmente diverse**

12 campioni di un certo materiale sono sottoposti ad un trattamento termico alla temperatura di 700 K. Altri 15 campioni dello stesso materiale vengono sottoposti ad un trattamento di eguale durata, ma alla temperatura di 1000 K. Si misura quindi la conducibilità elettrica di tutti i campioni, ottenendo i risultati in tabella (i dati sono in $10^3\,\mathrm{ohm}^{-1}\mathrm{cm}^{-1}$):

| $T = 700$ K | $T = 1000$ K |
|---|---|
| 6.00 | 6.25 |
| 7.02 | 8.48 |
| 6.52 | 7.33 |
| 7.30 | 6.89 |
| 6.80 | 8.70 |
| 6.95 | 7.15 |
| 7.45 | 8.66 |
| 5.38 | 6.44 |
| 5.77 | 6.95 |
| 6.25 | 8.10 |
| 7.13 | 7.47 |
| 6.68 | 6.05 |
| | 7.10 |
| | 6.24 |
| | 8.41 |

Si vuole stabilire, con un livello di significatività del 5%, se la conducibilità media dei due set di campioni sia la stessa o meno, ovvero se la temperatura del trattamente termico abbia un qualche effetto significativo sulla conducibilità elettrica del

materiale trattato. Le due popolazioni di dati si possono assumere normali, ma non si hanno elementi sufficienti per poter affermare che le rispettive varianze siano uguali.

**Soluzione**

Si indicano con $\mu_1$ e $\mu_2$ le medie dei due campioni, ossia i valori veri di conducibilità elettrica per il primo e per il secondo trattamento rispettivamente. Sia $p = 12$ il numero di dati $y_1, \ldots, y_p$ del primo campione e $q = 15$ quello dei dati $z_1, \ldots, z_q$ del secondo campione.

Si vuole testare l'ipotesi $H_0 : \mu_1 = \mu_2$ (i due trattamenti non modificano significativamente la conducibilità elettrica del materiale) contro l'ipotesi alternativa $H_1 : \mu_1 \neq \mu_2$ (i due trattamenti conducono a materiali con una diversa conducibilità elettrica).

Poichè per ipotesi le popolazioni possono assumersi normali ma le varianze non sono necessariamente uguali, la variabile del test è data da

$$t = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{1}{p}s_y^2 + \frac{1}{q}s_z^2}}$$

che per $H_0$ vera segue **approssimativamente** una distribuzione di Student con un numero di gradi di libertà pari a

$$n = \frac{\left(\frac{s_y^2}{p} + \frac{s_z^2}{q}\right)^2}{\frac{1}{p-1}\left(\frac{s_y^2}{p}\right)^2 + \frac{1}{q-1}\left(\frac{s_z^2}{q}\right)^2}$$

Nella fattispecie risulta

$$\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i = 6.6042 \qquad \overline{z} = \frac{1}{q}\sum_{j=1}^{q} z_j = 7.3480$$

$$s_y^2 = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \overline{y})^2 = 0.409517$$

$$s_z^2 = \frac{1}{q-1}\sum_{j=1}^{q}(z_j - \overline{z})^2 = 0.853617$$

per cui il numero di g.d.l. della variabile del test, qualora $H_0$ sia vera, risulta

$$n = \frac{\left(\frac{0.409517}{12} + \frac{0.853617}{15}\right)^2}{\frac{1}{11}\left(\frac{0.409517}{12}\right)^2 + \frac{1}{14}\left(\frac{0.853617}{15}\right)^2} = 24.576956$$

La regione di rigetto si scrive

$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](n)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n)}\right\}$$

con $\alpha = 0.05$ e $n = 24.576956$, per cui occorre calcolare il $t$-value

$$t_{[1-\frac{\alpha}{2}](n)} = t_{[0.975](24.576956)}$$

che ovviamente non è tabulato (il numero di g.d.l. non è intero).

Dalla tabella si ha però

$$t_{[0.975](24)} = 2.064 \qquad t_{[0.975](25)} = 2.060$$

ed è quindi possibile applicare uno schema di interpolazione lineare

$$\begin{array}{cc} 24 & 2.064 \\ 24.576956 & t_{[0.975](24.576956)} \\ 25 & 2.060 \end{array}$$

che porge la relazione

$$\frac{24.576956 - 24}{25 - 24} = \frac{t_{[0.975](24.576956)} - 2.064}{2.060 - 2.064}$$

dalla quale segue infine

$$t_{[0.975](24.576956)} = 2.0617 \ .$$

La regione di rigetto di $H_0$ diventa così

$$\{t \leq -2.0617\} \cup \{t \geq 2.0617\}$$

D'altra parte, la variabile del test assume il valore

$$t = \frac{6.6042 - 7.3480}{\sqrt{\frac{1}{12} \cdot 0.409517 + \frac{1}{15} \cdot 0.853617}} = -2.4653$$

che **è ricompreso nella regione di rigetto di** $H_0$. I valori di conducibilità elettrica ottenibili con le due temperature di trattamento sono **significativamente diversi**, con livello di significatività del 5%.

## 8.6 t-test accoppiato per il confronto delle medie di due popolazioni normali (varianze incognite)

Trova impiego quando i campioni estratti dalle due popolazioni normali, di medie $\mu_1$ e $\mu_2$ e varianze qualsiasi, non possono considerarsi indipendenti perchè i dati sono **organizzati per coppie di valori corrispondenti** $(p = q = n)$

$$(y_1, z_1)\ , \quad (y_2, z_2)\ , \quad \dots\ , \quad (y_n, z_n)$$

L'ipotesi nulla da testare è sempre $H_0 : \mu_1 = \mu_2$ contro l'ipotesi alternativa $H_1 : \mu_1 \neq \mu_2$

Le differenze fra i dati

$$d_i \ = \ y_i - z_i \ , \quad i = 1, \dots, n$$

sono variabili gaussiane i.i.d. di media $\mu_1 - \mu_2$ e varianza $\sigma_1^2 + \sigma_2^2$

Se $H_0$ è vera le $d_i$ costituiscono un campione di una variabile gaussiana $d$ di media nulla e varianza $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$, sicchè

$$t \ = \ \frac{\overline{d}}{s_d/\sqrt{n}} \ = \ \frac{\overline{y} - \overline{z}}{\dfrac{1}{\sqrt{n}}\sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2}}$$

è una $t$ di Student a $n - 1$ gradi di libertà

La regione di rigetto, con livello di significatività $\alpha$, si scrive

$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](n-1)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n-1)}\right\}$$

Il test è anche noto come **paired $t$-test**

### Esempio: t-test accoppiato per il confronto delle medie di due popolazioni normali

Un certo insieme di provini viene sottoposto ad un particolare trattamento chimico-fisico. Una certa grandezza fisica viene misurata **per ciascun provino** a monte e a valle del trattamento. I risultati sono riassunti nella tabella seguente

| prima del trattamento | dopo il trattamento |
|---|---|
| 9.5 | 10.4 |
| 4.2 | 9.5 |
| 10.1 | 11.6 |
| 7.3 | 8.5 |
| 9.6 | 11.0 |
| 7.9 | 10.9 |
| 10.1 | 9.3 |
| 8.7 | 10.3 |

Si vuole verificare, con un livello di significatività del 5%, l'ipotesi che le medie $\mu_1$ e $\mu_2$ della grandezza misurata siano le stesse prima e dopo il trattamento, assumendo popolazioni normali.

**Soluzione**

In questo caso è naturale applicare un $t$-test accoppiato per il confronto delle medie, dal momento che la grandezza viene misurata prima e dopo il trattamento **su ciascun provino**. Si accoppieranno perciò i valori relativi allo stesso provino:

$$(y_i, z_i) \qquad i = 1, \ldots, n$$

indicando con $y_i$ i valori misurati a monte del trattamento e con $z_i$ i corrispondenti valori misurati a valle del trattamento, sul totale degli $n = 8$ provini.

Si verifica l'ipotesi $H_0 : \mu_1 = \mu_2$, che il trattamento non abbia alcun effetto sul valor medio (e quindi vero) della grandezza misurata, contro l'ipotesi alternativa $H_1 : \mu_1 \neq \mu_2$. La variabile del test è

$$t = \sqrt{n} \frac{\overline{y} - \overline{z}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - z_i - \overline{y} + \overline{z})^2}}$$

che per $H_0$ vera segue una distribuzione di Student a $n-1 = 7$ gradi di libertà.

La regione di rigetto, con livello di significatività $\alpha$, è della forma

$$\left\{ t \leq -t_{[1-\frac{\alpha}{2}](n-1)} \right\} \cup \left\{ t \geq t_{[1-\frac{\alpha}{2}](n-1)} \right\}$$

e con $n = 8$, $\alpha = 0.05$ diventa

$$\left\{ t \leq -2.365 \right\} \cup \left\{ t \geq 2.365 \right\}$$

in quanto

$$t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.975](7)} = 2.365$$

Nella fattispecie, si ha:

$$\overline{y} = \frac{1}{8} \sum_{i=1}^{8} y_i = 8.425$$

$$\overline{z} = \frac{1}{8} \sum_{i=1}^{8} z_i = 10.1875$$

mentre

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2 \;=\; \frac{1}{7}\cdot 21.89875 \;=\; 3.12839286$$

e quindi

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2} \;=\; \sqrt{3.12839286} \;=\; 1.76872634$$

per cui la variabile del test assume il valore

$$t \;=\; \sqrt{8}\cdot\frac{8.425 - 10.1875}{1.76872634} \;=\; -2.8184704$$

Il valore calcolato appartiene alla coda inferiore della regione di rigetto

$$-2.8184704 \;<\; -2.365$$

e **si deve pertanto escludere**, con livello di significatività del 5%, che il valor medio della grandezza prima e dopo il trattamento **sia lo stesso**.

**L'ipotesi nulla viene rigettata!**

## 8.7 Test dei segni sulla mediana di una popolazione

Test **non parametrico**: non richiede alcuna ipotesi circa la forma della distribuzione di probabilità della popolazione

Si considera una variabile casuale $x$, discreta o continua, di distribuzione arbitraria $p(x)$

**Ipotesi da testare**:

$H_0$: la mediana della distribuzione è $m$

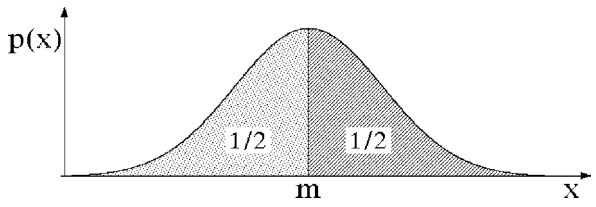
$H_1$: la mediana della distribuzione è diversa da $m$

Se la mediana della popolazione vale $m$, allora:

$x - m > 0$ con probabilità $1/2$

$x - m \leq 0$ con probabilità $1/2$

In caso contrario si avrà una maggiore probabilità di ottenere scarti positivi o negativi




**Variabile del test**

Per un set $x_1, x_2, \ldots, x_n$ di realizzazioni della $x$ si pone

$$s_i = \begin{cases} 1 & \text{se } x_i - m > 0 \\ 0 & \text{se } x_i - m \leq 0 \end{cases} \qquad \forall\, i = 1, \ldots, n$$

e si introduce come variabile del test il numero di differenze positive rispetto alla mediana

$$z = \sum_{i=1}^{n} s_i$$

## Regione di rigetto

Si è indotti a rigettare l'ipotesi $H_0$ qualora il numero di differenze positive rispetto alla mediana risulti eccessivamente grande o eccessivamente piccolo

La regione di rigetto sarà di tipo bilaterale simmetrico

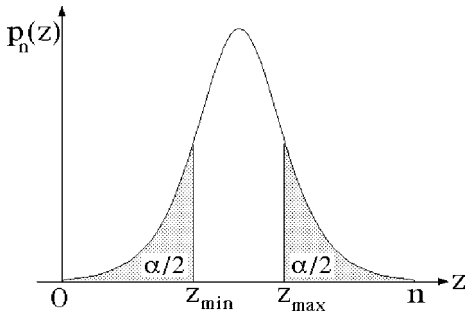
$$\{z \leq z_{\min}\} \cup \{z \geq z_{\max}\}$$

Se l'ipotesi $H_0$ è vera, la variabile del test $z$ segue una distribuzione binomiale $(1/2, 1/2)$

$$p_n(z) = \binom{n}{z} \frac{1}{2^n} \qquad \forall\, z = 0, 1, \ldots, n$$

Per un assegnato livello $\alpha$ di significatività del test, i limiti $z_{\min}$ e $z_{\max}$ della regione di rigetto sono perciò definiti da

$$\sum_{z=0}^{z_{\min}} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2} \qquad \sum_{z=z_{\max}}^{n} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2}$$



## 8.8 Test dei segni per verificare se due campioni accoppiati appartengono alla stessa popolazione

Il test dei segni è molto utile per verificare se due campioni accoppiati sono estratti dalla stessa popolazione statistica,

ossia se due campioni accoppiati seguono la stessa distribuzione di probabilità

Il test si basa sull'osservazione che date due variabili casuali continue indipendenti $X$ e $Y$ con uguale densità di probabilità $p(x)$ — o $p(y)$ — la variabile casuale **differenza**

$$Z \,=\, X \,-\, Y$$

segue una distribuzione di probabilità **pari**

$$p_Z(z) \,=\, p_Z(-z) \qquad\qquad \forall\, z \in \mathbb{R}$$

**qualunque sia la comune distribuzione** $p(x)$.

Si ha infatti

$$1 \,=\, \int\limits_{\mathbb{R}} dy \int\limits_{\mathbb{R}} dx\; p(x)\, p(y) \,=\, \int\limits_{\mathbb{R}} dy \int\limits_{\mathbb{R}} dz\; p(y+z)\, p(y) \,=$$
$$= \int\limits_{\mathbb{R}} dz \int\limits_{\mathbb{R}} dy\; p(y+z)\, p(y)$$

per cui la distribuzione di probabilità di $Z$ risulta

$$p_Z(z) \,=\, \int\limits_{\mathbb{R}} p(y+z)\, p(y)\, dy \qquad\qquad \forall\, z \in \mathbb{R}\,.$$

Ne deriva che

$$p_Z(-z) \;=\; \int_{\mathbb{R}} p(y-z)\, p(y)\, dy$$

e con il cambiamento di variabile $y = \xi + z$ l'integrale diventa

$$p_Z(-z) \;=\; \int_{\mathbb{R}} p(\xi)\, p(\xi+z)\, d\xi \;=\; \int_{\mathbb{R}} p(\xi+z)\, p(\xi)\, d\xi \;=\; p_Z(z)\,.$$

Le differenze positive e quelle negative o nulle hanno così la stessa probabilità di verificarsi:

$$p(z > 0) \;=\; \frac{1}{2} \qquad\qquad p(z \leq 0) \;=\; \frac{1}{2}\,.$$

Si può allora testare l'ipotesi nulla

$H_0$ : i due campioni sono estratti dalla stessa popolazione

contro l'ipotesi alternativa

$H_1$ : i due campioni sono estratti da popolazioni diverse

usando come **variabile del test** il **numero $z$ delle differenze positive** registrate sulle varie coppie di dati del campione

Si accetterà $H_0$ se $z$ è prossima al numero medio delle coppie di dati del campione, mentre si preferirà $H_1$ qualora $z$ sia sufficientemente basso o troppo alto (regione di rigetto bilaterale)

La regione di rigetto assume pertanto la forma a due code

$$\{z \leq z_{\min}\} \cup \{z \geq z_{\max}\}$$

Se $H_0$ è corretta, la variabile $z$ segue una distribuzione binomiale $(1/2, 1/2)$ e la probabilità che $z$ assuma un valore nella regione di rigetto deve identificarsi con la probabilità di errore di prima specie $\alpha$

Al livello di significatività $\alpha$, i limiti $z_{\min}$ e $z_{\max}$ della regione di rigetto sono perciò definiti da

$$\sum_{z=0}^{z_{\min}} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2} \qquad \sum_{z=z_{\max}}^{n} \binom{n}{z} \frac{1}{2^n} = \frac{\alpha}{2}$$

**Esempio**: **test dei segni**
Un'azienda dichiara che aggiungendo un additivo di propria produzione nel serbatoio di un'auto a benzina il rendimento del motore aumenta. Per verificare questa affermazione vengono scelte 16 automobili e si misura la distanza media da esse percorsa con un litro di carburante, con e senza additivo.

| additivo | no additivo | additivo | no additivo |
|---|---|---|---|
| 17.35 | 15.70 | 13.65 | 13.10 |
| 14.15 | 13.60 | 16.05 | 15.70 |
| 9.80 | 10.20 | 14.80 | 14.40 |
| 12.55 | 12.30 | 11.20 | 11.55 |
| 7.85 | 7.45 | 12.65 | 12.00 |
| 12.25 | 11.15 | 14.05 | 13.65 |
| 14.35 | 13.40 | 12.15 | 11.45 |
| 11.75 | 12.05 | 11.95 | 12.40 |

Assumendo che le condizioni di guida siano le stesse, si vuole verificare se vi è una qualche differenza nelle prestazioni dei motori a seguito dell'aggiunta dell'additivo, con livello di significatività 5% e 1%.

**Soluzione**
Si deve testare l'ipotesi nulla $H_0$ che non vi siano differenze di prestazioni fra le auto alimentate con carburante additivato e le stesse automobili rifornite con carburante normale, contro l'ipotesi alternativa $H_1$ che le auto trattate con l'additivo forniscano prestazioni migliori.

L'ipotesi $H_0$ viene abbandonata in favore di $H_1$ soltanto se il numero di differenze positive è sufficientemente grande: il test è unilaterale.

Le differenze dei chilometraggi per litro fra le auto trattate e quelle non trattate mostrano **12 segni positivi** e **4 segni negativi**.

Se $H_0$ è vera, la probabilità di ottenere almeno 12 segni positivi su un totale di 16 coppie di dati si scrive

$$\sum_{z=12}^{16} \binom{16}{z} \frac{1}{2^{16}} =$$

$$= \left[ \binom{16}{12} + \binom{16}{13} + \binom{16}{14} + \binom{16}{15} + \binom{16}{16} \right] \frac{1}{2^{16}} = 0.0384$$

per cui un numero di segni positivi pari a 12 è sicuramente contenuto nella regione di rigetto di $H_0$ se il livello di significatività viene fissato al 5%. Qualora sia invece $\alpha = 1\%$, il valore $z = 12$ di segni positivi non appartiene alla regione di rigetto di $H_0$, che quindi non può essere rifiutata.

## 8.9 Test sulla media di una popolazione arbitraria ma di varianza nota

Il teorema del limite centrale assicura che per una successione $(x_n)_{n\in\mathbb{N}}$ di variabili casuali indipendenti e identicamente distribuite, la cui comune distribuzione di probabilità abbia media $\mu$ e varianza $\sigma^2$ finite, la variabile casuale

$$z_n \;=\; \frac{1}{\sigma/\sqrt{n}}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\Big) \;=\; \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$$

segue una distribuzione normale standard nel limite $n \to +\infty$

Il teorema vale **qualunque sia la comune distribuzione di probabilità**, purchè con $\mu$ e $\sigma$ finite

In pratica, già per valori di $n$ dell'ordine di 5 la distribuzione della variabile $z_n$ può ritenersi normale standard, con ottima approssimazione

L'approssimazione migliora al crescere del numero $n$ di dati

Si può perciò ritenere che per una qualsiasi popolazione statistica di media e varianza finite, dato un campione ridotto di $n$ dati indipendenti $x_1, x_2, \ldots, x_n$, la variabile casuale

$$z_n \;=\; \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$$

segua una **distribuzione normale standard**

**Se $\sigma$ è nota**, si può allora verificare l'ipotesi nulla che la media della popolazione sia uguale ad un valore assegnato $\mu_0$, contro l'ipotesi alternativa che la media assuma un valore diverso:

$$H_0 \ : \ \mu = \mu_0 \qquad\qquad H_1 \ : \ \mu \neq \mu_0$$

usando la variabile

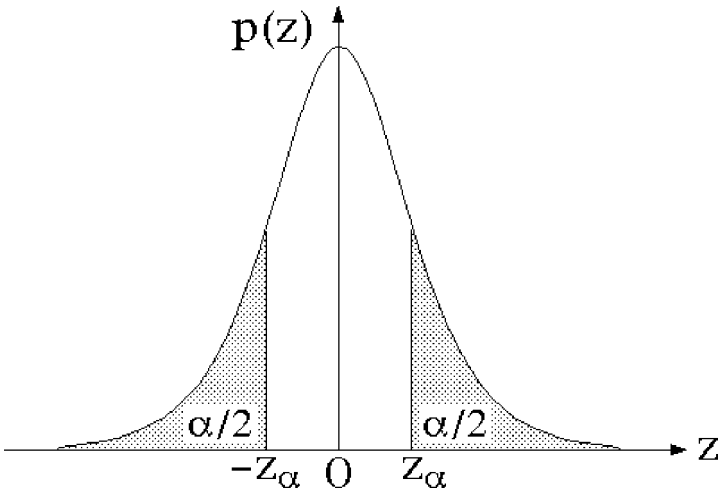
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

che per $\mu_0 = \mu$ coincide con la variabile normale standard $z_n$

Se $H_0$ è corretta, ci si aspetta che tipicamente risulti $\overline{x} \sim \mu = \mu_0$ e che pertanto sia $z \sim 0$. Viceversa, un valore di $z$ lontano da zero segnalerà che presumibilmente $\mu \neq \mu_0$

Si introduce perciò la regione di rigetto bilaterale simmetrica

$$\{z \leq -z_\alpha\} \cup \{z_\alpha \leq z\}$$




in cui il valore critico $z_\alpha > 0$ è individuato dall'equazione

$$\frac{\alpha}{2} = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^{z_\alpha} e^{-\xi^2/2} d\xi = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\Big(\frac{z_\alpha}{\sqrt{2}}\Big)$$

ossia da $1 - \alpha = \mathrm{erf}(z_\alpha/\sqrt{2})$

**Esempio**: **test sulla media di una popolazione di varianza nota**

La vita media di un campione di 100 lampadine a fluirescenza prodotte da una ditta è stata calcolata pari a 1570 ore, con una deviazione standard di 120 ore. Se $\mu$ è la vita media delle lampadine, si sottoponga a test l'ipotesi $\mu = 1600$ ore contro l'ipotesi alternativa che $\mu \neq 1600$ ore con un livello di significatività del 5%.

**Soluzione**

Si deve decidere fra le due ipotesi

$$H_0 : \mu = \mu_0 = 1600\,\text{ore} \qquad H_1 : \mu \neq \mu_0 = 1600\,\text{ore}$$

Poichè il campione è molto grande si può senz'altro ritenere che per $H_0$ vera la variabile casuale

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

segua una distribuzione normale standard. Si noti che per un campione così grande è lecito identificare la deviazione standard con la relativa stima campionaria:

$$\sigma \simeq s = 120$$

La media campionaria sugli $n = 100$ dati è pari a

$$\overline{x} = 1570$$

per cui la variabile del test assume il valore

$$z = \frac{1570 - 1600}{120/\sqrt{100}} = -2.50$$

Nella fattispecie si ha tuttavia

$$z_\alpha = z_{0.05} = 1.96$$

come si ricava risolvendo l'equazione

$$\frac{1}{\sqrt{2\pi}} \int_0^{z_{0.05}} e^{-\xi^2/2} d\xi = \frac{1-\alpha}{2} = \frac{1-0.05}{2} = 0.475$$

con l'ausilio della tabella della distribuzione cumulativa della normale standard.

La regione di rigetto di $H_0$ risulta pertanto

$$\{z \leq -z_\alpha\} \cup \{z_\alpha \leq z\}$$

ossia

$$\{z \leq -1.96\} \cup \{1.96 \leq z\}$$

e poichè questa contiene il valore calcolato $z = -2.50$ della variabile del test, **l'ipotesi $H_0$ deve essere rigettata**.

Si può dunque affermare che **la vita media delle lampadine è diversa dal valore dichiarato di 1600 ore**, con un livello di significatività del 5%.

## 8.10 $\mathcal{X}^2$-test di adattamento di un campione ad una distribuzione preassegnata

Serve a stabilire se una popolazione statistica segue una **distribuzione di probabilità assegnata** $p(x)$

Gli individui del campione ridotto vengono **istogrammati** su un certo numero di canali, di ampiezza conveniente ("binned" distribution). Sia $f_i$ il numero di individui nell'$i$-esimo canale ("frequenza osservata")

Il numero medio $n_i$ di individui atteso nell'$i$-esimo canale è calcolato per mezzo della distribuzione $p(x)$ ("frequenza teorica")

Se la distribuzione di probabilità ipotizzata $p(x)$ è corretta, la variabile casuale

$$\mathcal{X}^2 = \sum_i \frac{(f_i - n_i)^2}{n_i}$$

in cui la somma si intende su tutti i "canali" dell'istogramma, **segue approssimativamente una distribuzione di $\mathcal{X}^2$** a $k - 1 - c$ gradi di libertà, essendo:

$k$ il numero di canali;

$c$ il numero di parametri della distribuzione che **non sono preassegnati**, ma **calcolati sulla base dello stesso campione ridotto** (c.d. "vincoli")

L'approssimazione è buona se le frequenze osservate $f_i$ non sono troppo piccole ($\geq 3$ circa)

Si può ragionevolmente ritenere accettabile l'ipotesi nulla

$H_0$ : *$p(x)$ è la distribuzione di probabilità della popolazione*

qualora $\mathcal{X}^2$ **risulti non troppo grande**, e di doverla rigettare in caso contrario

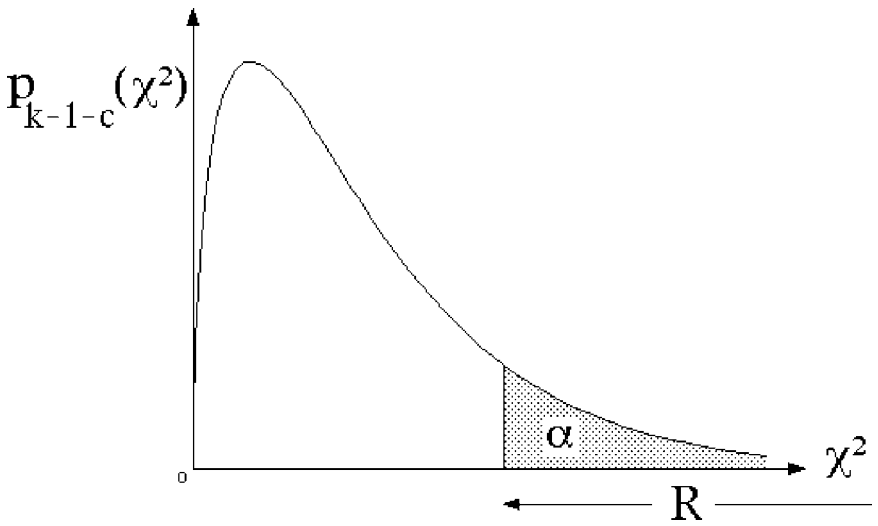
La regione di rigetto è perciò definita da

$$R = \left\{ \mathcal{X}^2 \geq \mathcal{X}^2{}_{[1-\alpha](k-1-c)} \right\}$$

Se l'ipotesi $H_0$ è vera, la probabilità che $\mathcal{X}^2$ assuma comunque un valore compreso nell'intervallo di rigetto $R$ vale

$$1 - (1 - \alpha) = \alpha$$

che rappresenta dunque la **probabilità di errore del I tipo**, nonché il **livello di significatività del test**



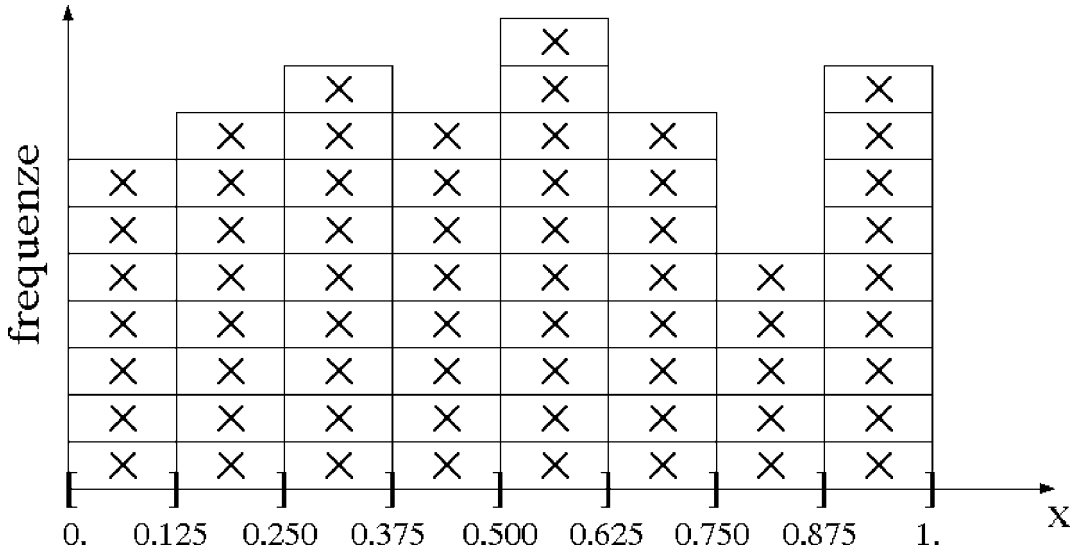
**Esempio**: **test del $\mathcal{X}^2$ per una distribuzione uniforme**
Misure ripetute di una certa grandezza $x$ hanno condotto alla seguente tabella di risultati

| $x_i$ | $x_i$ | $x_i$ | $x_i$ |
|---|---|---|---|
| 0.101 | 0.020 | 0.053 | 0.120 |
| 0.125 | 0.075 | 0.033 | 0.130 |
| 0.210 | 0.180 | 0.151 | 0.145 |
| 0.245 | 0.172 | 0.199 | 0.252 |
| 0.370 | 0.310 | 0.290 | 0.349 |
| 0.268 | 0.333 | 0.265 | 0.287 |
| 0.377 | 0.410 | 0.467 | 0.398 |
| 0.455 | 0.390 | 0.444 | 0.497 |
| 0.620 | 0.505 | 0.602 | 0.544 |
| 0.526 | 0.598 | 0.577 | 0.556 |
| 0.610 | 0.568 | 0.630 | 0.749 |
| 0.702 | 0.657 | 0.698 | 0.731 |
| 0.642 | 0.681 | 0.869 | 0.761 |
| 0.802 | 0.822 | 0.778 | 0.998 |
| 0.881 | 0.950 | 0.922 | 0.975 |
| 0.899 | 0.968 | 0.945 | 0.966 |

per un totale di 64 dati. Si vuole accertare, con un livello di significatività del 5%, se la popolazione da cui il campione è stato estratto possa avere una distribuzione uniforme in $[0, 1]$.

**Soluzione**
I 64 dati vengono posti in ordine crescente e istogrammati su 8 canali di uguale ampiezza nell'intervallo $[0, 1]$, chiusi a destra e aperti a sinistra (salvo il primo che è chiuso ad ambo gli estremi).

Si ottiene così il seguente istogramma di frequenza




che appare soddisfacente per l'applicazione del test del $\mathcal{X}^2$, dal momento che le frequenze empiriche in ogni canale non sono mai inferiori a $3 \div 5$.

Se la distribuzione uniforme in $[0, 1]$ è corretta, la frequenza attesa in ciascun canale è la stessa:

$$\text{numero dati del campione} \times \text{probabilità del canale} =$$
$$= 64 \times \frac{1}{8} = 8$$

Il $\mathcal{X}^2$ del campione per la distribuzione ipotizzata vale allora

$$\mathcal{X}^2 = \frac{(7-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(8-8)^2}{8} +$$
$$+ \frac{(10-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(9-8)^2}{8} = 2$$

Il numero di gradi di libertà è semplicemente il numero di canali ridotto di uno

$$k - 1 - c \;=\; 8 - 1 - 0 \;=\; 7$$

in quanto nessun parametro della distribuzione deve essere stimato usando gli stessi dati del campione ($c = 0$).

Con livello di significatività $\alpha = 0.05$, l'ipotesi che la distribuzione uniforme sia corretta viene rigettata se

$$\mathcal{X}^2 \;\geq\; \mathcal{X}^2{}_{[1-\alpha](k-1-c)} \;=\; \mathcal{X}^2{}_{[1-0.05](7)} \;=\; \mathcal{X}^2{}_{[0.95](7)}$$

Nella fattispecie, la tabella della distribuzione cumulativa di $\mathcal{X}^2$ fornisce

$$\mathcal{X}^2{}_{[0.95](7)} \;=\; 14.067$$

per cui il $\mathcal{X}^2$ campionario è minore di quello critico

$$\mathcal{X}^2 \;=\; 2 \;<\; 14.067 \;=\; \mathcal{X}^2{}_{[0.95](7)}$$

Si conclude pertanto che con il livello di significatività del 5% **non è possibile escludere che la popolazione abbia distribuzione uniforme in** $[0, 1]$. L'ipotesi nulla deve essere accettata, o comunque non rigettata.

**Esempio**: $\mathcal{X}^2$**-test per una distribuzione normale**

Un antropologo è interessato alle altezze dei nativi di una certa isola. Egli sospetta che le altezze dei maschi adulti dovrebbero essere normalmente distribuite, e misura le altezze di un campione di 200 uomini. Utilizzando queste misure, egli calcola la media $\overline{x}$ e la deviazione standard $s$ campionarie, e usa questi numeri come stima per il valor medio $\mu$ e la deviazione standard $\sigma$ della distribuzione normale attesa. Sceglie poi otto intervalli uguali in cui raggruppa i risultati delle sue osservazioni (frequenze empiriche). Ne deduce la tabella seguente:

| i | intervallo di altezze | frequenza empirica | frequenza attesa |
|---|---|---|---|
| 1 | $x < \overline{x} - 1.5s$ | 14 | 13.362 |
| 2 | $\overline{x} - 1.5s \leq x < \overline{x} - s$ | 29 | 18.370 |
| 3 | $\overline{x} - s \leq x < \overline{x} - 0.5s$ | 30 | 29.976 |
| 4 | $\overline{x} - 0.5s \leq x < \overline{x}$ | 27 | 38.292 |
| 5 | $\overline{x} \leq x < \overline{x} + 0.5s$ | 28 | 38.292 |
| 6 | $\overline{x} + 0.5s \leq x < \overline{x} + s$ | 31 | 29.976 |
| 7 | $\overline{x} + s \leq x < \overline{x} + 1.5s$ | 28 | 18.370 |
| 8 | $\overline{x} + 1.5s \leq x$ | 13 | 13.362 |

Si vuole verificare l'ipotesi della distribuzione normale con un livello di significatività ($a$) del 5% e ($b$) dell'1%.

**Soluzione**

Le frequenze empiriche $f_i$ sono piuttosto alte ($f_i \gg 5$), per cui è lecito verificare l'ipotesi applicando il test del $\mathcal{X}^2$.

Le frequenze attese $n_i$ si calcolano moltiplicando per il numero totale di individui — 200 — gli integrali della distribuzione normale standard su ciascun intervallo. Per gli intervalli $i =$

5, 6, 7, 8 la tavola degli integrali fra 0 e $z$ della normale standard porge infatti:

$$P(\overline{x} \leq x < \overline{x} + 0.5s) = \int_0^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.19146$$

$$P(\overline{x} + 0.5s \leq x < \overline{x} + s) = \int_{0.5}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= \int_0^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= 0.34134 - 0.19146 = 0.14674$$

$$P(\overline{x} + s \leq x < \overline{x} + 1.5s) = \int_{1}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= \int_0^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= 0.43319 - 0.34134 = 0.09185$$

$$P(\overline{x} + 1.5s \leq x) = \int_{1.5}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz =$$

$$= 0.5 - 0.43319 = 0.06681$$

mentre le probabilità degli intervalli $i = 1, 2, 3, 4$ sono simmetricamente uguali alle precedenti (data la simmetria della distribuzione normale standard rispetto all'origine).

Il $\mathcal{X}^2$ del campione è quindi dato da:

$$\mathcal{X}^2 = \sum_{i=1}^{8} \frac{(f_i - n_i)^2}{n_i} = 17.370884$$

Se la distribuzione di probabilità proposta è corretta, la variabile del test segue una distribuzione di $\mathcal{X}^2$ con

$$8 - 1 - 2 = 5$$

g.d.l. in quanto $k = 8$ sono i canali utilizzati e $c = 2$ i parametri della distribuzione ($\mu$ e $\sigma$) che vengono stimati usando gli stessi dati del campione.

La tabella delle distribuzioni cumulative di $\mathcal{X}^2$ fornisce i valori critici:

$$\mathcal{X}^2_{[1-\alpha](5)} = \mathcal{X}^2_{[0.95](5)} = 11.070 \qquad \text{per } \alpha = 0.05$$

$$\mathcal{X}^2_{[1-\alpha](5)} = \mathcal{X}^2_{[0.99](5)} = 15.086 \qquad \text{per } \alpha = 0.01$$

In ambo i casi il $\mathcal{X}^2$ calcolato sul campione è superiore: si conclude pertanto che, con ambo i livelli di significatività del 5 e dell'1%, **l'ipotesi deve essere rigettata**. I dati del campione suggeriscono che **la distribuzione delle altezze non sia normale** per la popolazione dei nativi.

**Esempio**: $\mathcal{X}^2$**-test per una distribuzione discreta**
Una coppia di dadi viene lanciata 360 volte e per ciascun lancio viene registrato il punteggio realizzato. I punteggi possibili sono, ovviamente, $2, 3, 4, \ldots, 11, 12$ e ciascuno di essi è stato realizzato il numero di volte indicato nella tabella seguente

| Punteggio | Numero risultati | Punteggio | Numero risultati |
|---|---|---|---|
| 2 | 6 | 8 | 44 |
| 3 | 14 | 9 | 49 |
| 4 | 23 | 10 | 39 |
| 5 | 35 | 11 | 27 |
| 6 | 57 | 12 | 16 |
| 7 | 50 | | |

Verificare l'ipotesi che i dati non siano truccati con un livello di significatività $(a)$ del 5% e $(b)$ del 1%.

**Soluzione**
Qualora i dadi non siano truccati, la probabilità dei singoli punteggi è calcolabile a priori, senza utilizzare i dati del campione:

| Punteggio | Probabilità | Punteggio | Probabilità |
|---|---|---|---|
| 2 | 1/36 | 8 | 5/36 |
| 3 | 2/36 | 9 | 4/36 |
| 4 | 3/36 | 10 | 3/36 |
| 5 | 4/36 | 11 | 2/36 |
| 6 | 5/36 | 12 | 1/36 |
| 7 | 6/36 | | |

ricordando che ciascuna faccia ha probabilità 1/6 di uscire.

Le frequenze attese $n_i$ dei vari punteggi $i = 2, \ldots, 12$ sono quindi quelle elencate nella terza colonna della tabella seguente:

| Punteggio | Frequenza empirica | Frequenza attesa |
|---|---|---|
| 2 | 6 | 10 |
| 3 | 14 | 20 |
| 4 | 23 | 30 |
| 5 | 35 | 40 |
| 6 | 57 | 50 |
| 7 | 50 | 60 |
| 8 | 44 | 50 |
| 9 | 49 | 40 |
| 10 | 39 | 30 |
| 11 | 27 | 20 |
| 12 | 16 | 10 |

Le frequenze empiriche $f_i$ sono tutte sufficientemente grandi ($> 5$) da consentire il ricorso al $\mathcal{X}^2$-test per verificare l'ipotesi **senza dover accorpare i punteggi con frequenze empiriche troppo basse** (un punteggio $\Leftrightarrow$ un intervallo).

Il $\mathcal{X}^2$ dei dati si scrive:

$$\mathcal{X}^2 = \sum_{i=2}^{12} \frac{(f_i - n_i)^2}{n_i} = 19.800$$

Se le probabilità proposte sono quelle corrette (dadi non truccati), la variabile del test seguirà una distribuzione di $\mathcal{X}^2$ a

$$11 - 1 = 10$$

g.d.l., attribuibili alle $n = 11$ frequenze osservate e all'unico

vincolo rappresentato dal numero totale di risultati registrati.

Dalla tabella delle distribuzioni cumulative di $\mathcal{X}^2$ si deducono i seguenti valori critici:

$$\mathcal{X}^2{}_{[1-\alpha](10)} = \mathcal{X}^2{}_{[0.95](10)} = 18.307 \qquad \text{per } \alpha = 0.05$$

$$\mathcal{X}^2{}_{[1-\alpha](10)} = \mathcal{X}^2{}_{[0.99](10)} = 23.209 \qquad \text{per } \alpha = 0.01$$

Si trova allora che

$$\mathcal{X}^2 = 19.800 > 18.307 = \mathcal{X}^2{}_{[0.95](10)}$$

per cui **al livello di significatività** $\alpha = 5\%$ **l'ipotesi che i due dadi siano regolari deve essere rigettata**.

Per contro, si ha evidentemente

$$\mathcal{X}^2 = 19.800 < 23.209 = \mathcal{X}^2{}_{[0.99](10)}$$

e di conseguenza **la stessa ipotesi non può essere confutata al livello di significatività** $\alpha = 1\%$.

Si osservi che la probabilità di ottenere un valore di $\mathcal{X}^2$ maggiore o uguale a quello effettivamente ricavato dal campione è pari a

$$\int_{19.800}^{+\infty} p_{10}(\mathcal{X}^2)\, d\mathcal{X}^2 = 0.033 = 3.3\%$$

essendo $p_{10}(\mathcal{X}^2)$ la densità di probabilità di una variabile di $\mathcal{X}^2$ a 10 g.d.l. Il valore non è tabulato, ma può essere ricavato facilmente per interpolazione lineare dai valori critici tabulati:

$$\mathcal{X}^2{}_{[0.95](10)} = 18.307 \qquad \mathcal{X}^2{}_{[0.975](10)} = 20.483\ .$$

**Esempio**: **$\mathcal{X}^2$-test per una distribuzione di Poisson**
Le specifiche di un certo campione radioattivo dichiarano che esso dà luogo a una media di due decadimenti al minuto. Per verificare l'affermazione si contano i decadimenti prodotti in 40 intervalli di tempo separati, tutti della durata di un minuto. I risultati sono riportati in tabella:

| Numero di decadimenti al minuto | Frequenza empirica |
|---|---|
| 0 | 11 |
| 1 | 12 |
| 2 | 11 |
| 3 | 4 |
| 4 | 2 |
| 5 o più | 0 |

Si vuole verificare, con livello di significatività del 5%, se il numero di decadimenti in un minuto segue:

(*a*) una distribuzione di Poisson con media $\mu = 2$;

(*b*) una distribuzione di Poisson con media stimata sulla base del campione.

**Soluzione**
Le frequenze empiriche delle osservazioni con almeno 3 decadimenti in un minuto — 4, 2 e 0, nella tabella — sono tutte troppo basse per consentire l'applicazione diretta del $\mathcal{X}^2$-test.

Conviene quindi accorpare le osservazioni con almeno 3 decadimenti in un unico canale, al quale si attribuirà una frequenza empirica complessiva pari a $4 + 2 + 0 = 6$.

($a$) Se la distribuzione corretta è una poissoniana di media $\mu = 2$, le probabilità di 0, 1, 2 e almeno 3 decadimenti in un minuto sono date da:

$$P(0) = e^{-2} = 0.135335 \qquad P(1) = \frac{2}{1!}\, e^{-2} = 0.270671$$

$$P(2) = \frac{2^2}{2!}\, e^{-2} = 0.270671$$

$$P(\geq 3) = 1 - e^{-2} \sum_{i=0}^{2} \frac{2^i}{i!} =$$

$$= 1 - 0.135335 - 0.270671 - 0.270671 = 0.323324$$

e le corrispondenti frequenze attese $n_i$ su 40 prove risultano perciò quelle elencate nella terza colonna della tabella seguente

| Numero di decadimenti al minuto | Frequenza empirica | Frequenza attesa |
|---|---|---|
| 0 | 11 | 5.413411 |
| 1 | 12 | 10.826823 |
| 2 | 11 | 10.826823 |
| 3 o più | 6 | 12.932943 |

Il $\mathcal{X}^2$ dei dati vale allora

$$\mathcal{X}^2 = \sum_{i=0}^{3} \frac{(f_i - n_i)^2}{n_i} = 9.611732$$

Per $4 - 1 = 3$ g.d.l. e livello di significatività $\alpha = 0.05$, il $\mathcal{X}^2$ critico risulta

$$\mathcal{X}^2{}_{[0.95](3)} = 7.815$$

e porta a **rigettare l'ipotesi**.

($b$) Il numero medio di decadimenti al minuto è stimato dalla media campionaria

$$\overline{x} = \frac{0 \cdot 11 + 1 \cdot 12 + 2 \cdot 11 + 3 \cdot 4 + 4 \cdot 2}{40} = \frac{54}{40} = 1.3500$$

Le probabilità di 0, 1, 2 e almeno 3 decadimenti in un minuto sono così

$$\begin{aligned} P(0) &= e^{-1.35} = 0.259240 \\ P(1) &= \frac{1.35}{1!}\, e^{-1.35} = 0.349974 \\ P(2) &= \frac{1.35^2}{2!}\, e^{-1.35} = 0.236233 \\ P(\geq 3) &= 1 - e^{-1.35} \sum_{i=0}^{2} \frac{1.35^i}{i!} = \\ &= 1 - 0.259240 - 0.349974 - 0.236233 = 0.154553 \end{aligned}$$

e le frequenze attese diventano

| Numero di decadimenti al minuto | Frequenza empirica | Frequenza attesa |
|---|---|---|
| 0 | 11 | 10.369610 |
| 1 | 12 | 13.998974 |
| 2 | 11 | 9.449308 |
| 3 o più | 6 | 6.182108 |

Il calcolo del $\mathcal{X}^2$ porge quindi

$$\mathcal{X}^2 \;=\; \sum_{i=0}^{3} \frac{(f_i - n_i)^2}{n_i} \;=\; 0.583608$$

D'altra parte, l'unico parametro della distribuzione di Poisson è stato stimato ricorrendo allo stesso campione, per cui il numero di g.d.l. risulta

$$4 \;-\; 1 \;-\; 1 \;=\; 2$$

e per $\alpha = 0.05$ individua il $\mathcal{X}^2$ critico

$$\mathcal{X}^2{}_{[0.95](2)} \;=\; 5.991 \;.$$

Il $\mathcal{X}^2$ calcolato è sensibilmente inferiore al valore critico.

Si conclude che sulla base del campione, e con livello di significatività del 5%, **è lecito accettare l'ipotesi che il numero di decadimenti al minuto del radionuclide possa essere descritto mediante una distribuzione di Poisson di media** $\mu = 1.35$:

$$P(i) \;=\; \frac{1.35^i}{i!}\, e^{-1.35} \qquad i \;=\; 0, 1, 2, \ldots \;.$$

Si vede dunque come una stima accurata del parametro $\mu$ sia cruciale nel determinare l'esito del test.

## 8.11 Test di Kolmogorov-Smirnov

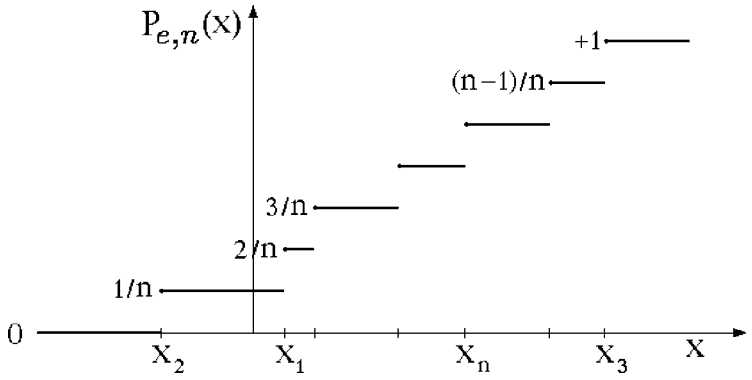
Metodo **non parametrico** per testare se un set di dati indipendenti segue una distribuzione di probabilità preassegnata

Per una variabile casuale continua $x \in \mathbb{R}$ con distribuzione $p(x)$, si considerano $n$ realizzazioni $x_1, x_2, \ldots, x_n$

Equivalentemente, $x_1, x_2, \ldots, x_n$ si assumono variabili casuali continue i.i.d. di distribuzione $p(x)$

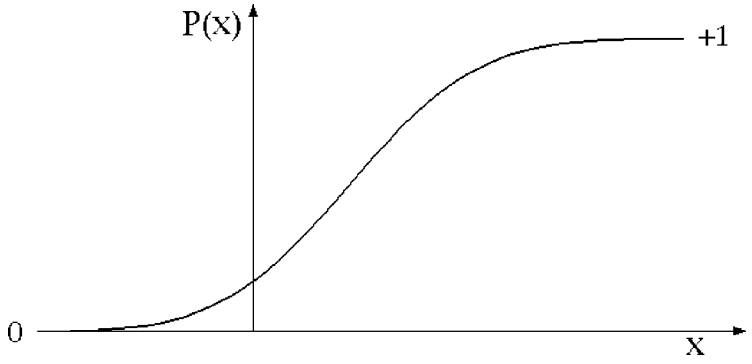
La **distribuzione empirica cumulativa** di frequenza a $n$ dati della variabile $x$ si definisce come

$$P_{e,n}(x) \;=\; \#\{x_i\,,\;\; i = 1, \ldots, n,\;\; x_i \;\leq\; x\} \,/\, n$$




e viene confrontata con la distribuzione cumulativa della $x$

$$P(x) \;=\; \int_{-\infty}^{x} p(\xi)\, d\xi$$



### Variabile del test

Si considera lo scarto massimo fra le distribuzioni cumulative teorica ed empirica

$$D_n \;=\; \sup_{x\in\mathbb{R}} \Big| P(x) - P_{e,n}(x) \Big|$$

$D_n$ è una funzione delle variabili casuali $x_1, x_2, \ldots, x_n$ e costituisce a sua volta una variabile casuale — di **Kolmogorov-Smirnov**

Il test di Kolmogorov-Smirnov si basa sul seguente teorema limite

### Teorema

Se le variabili casuali $x_1, x_2, \ldots$ sono i.i.d. di distribuzione cumulativa continua $P(x)$, $\forall\, \lambda > 0$ fissato vale

$$\text{Probabilità}\Big\{ D_n \;>\; \frac{\lambda}{\sqrt{n}} \Big\} \;\xrightarrow[n\to+\infty]{}\; Q(\lambda)$$

dove $Q(\lambda)$ è la funzione monotòna decrescente definita in $\mathbb{R}^+$ da

$$Q(\lambda) \;=\; 2\sum_{j=1}^{\infty} (-1)^{j-1}\, e^{-2j^2\lambda^2} \qquad \forall\, \lambda \;>\; 0$$

con

$$\lim_{\lambda\to 0+} Q(\lambda) \;=\; 1 \qquad\qquad \lim_{\lambda\to+\infty} Q(\lambda) \;=\; 0 \qquad\qquad \square$$

Il teorema assicura che per $n$ abbastanza grande vale

$$\text{Probabilità}\left\{ D_n > \frac{\lambda}{\sqrt{n}} \right\} \simeq Q(\lambda)$$

indipendentemente dalla distribuzione $P(x)$

Fissato il valore $\alpha \in (0,1)$ della precedente probabilità, si ha

$$\alpha \simeq Q(\lambda) \qquad \Longleftrightarrow \qquad \lambda = Q^{-1}(\alpha)$$

in modo che risulta

$$\text{Probabilità}\left\{ D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}} \right\} \simeq \alpha$$

**Ipotesi da testare**

$H_0$: la distribuzione cumulativa dei dati $x_1, x_2, \ldots, x_n$ è $P(x)$

$H_1$: la distribuzione cumulativa non è $P(x)$

**Regione di rigetto**

Un valore grande della variabile $D_n$ deve essere considerato indice di una cattiva rispondenza fra la distribuzione empirica cumulativa $P_{e,n}(x)$ e la distribuzione teorica $P(x)$ ipotizzata

L'ipotesi nulla sarà perciò rigettata, con livello di significatività $\alpha$, se

$$D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}}$$

In particolare

$$Q^{-1}(0.10) = 1.22384 \qquad Q^{-1}(0.05) = 1.35809$$

$$Q^{-1}(0.01) = 1.62762 \qquad Q^{-1}(0.001) = 1.94947$$

**Esempio**: **test di Kolmogorov-Smirnov per una distribuzione uniforme**

Un generatore di numeri casuali ha fornito la seguente sequenza di valori numerici nell'intervallo $[0, 1]$:

| $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|
| 1 | 0.750 | 17 | 0.473 |
| 2 | 0.102 | 18 | 0.779 |
| 3 | 0.310 | 19 | 0.266 |
| 4 | 0.581 | 20 | 0.508 |
| 5 | 0.199 | 21 | 0.979 |
| 6 | 0.859 | 22 | 0.802 |
| 7 | 0.602 | 23 | 0.176 |
| 8 | 0.020 | 24 | 0.645 |
| 9 | 0.711 | 25 | 0.473 |
| 10 | 0.422 | 26 | 0.927 |
| 11 | 0.958 | 27 | 0.368 |
| 12 | 0.063 | 28 | 0.693 |
| 13 | 0.517 | 29 | 0.401 |
| 14 | 0.895 | 30 | 0.210 |
| 15 | 0.125 | 31 | 0.465 |
| 16 | 0.631 | 32 | 0.827 |

Usando il test di Kolmogorov-Smirnov, si vuole verificare con un livello di significatività del 5% se la successione di dati ottenuta è effettivamente compatibile con una distribuzione uniforme nell'intervallo $[0, 1]$.

**Soluzione**
Per applicare il test è necessario calcolare la distribuzione cumulativa empirica e confrontarla con la distribuzione cumulativa di una variabile casuale uniforme in $[0, 1]$.

La distribuzione cumulativa della variabile casuale uniforme in $[0, 1]$ si calcola in modo elementare

$$P(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Per calcolare la distribuzione cumulativa empirica occorre riordinare i dati del campione in ordine crescente:

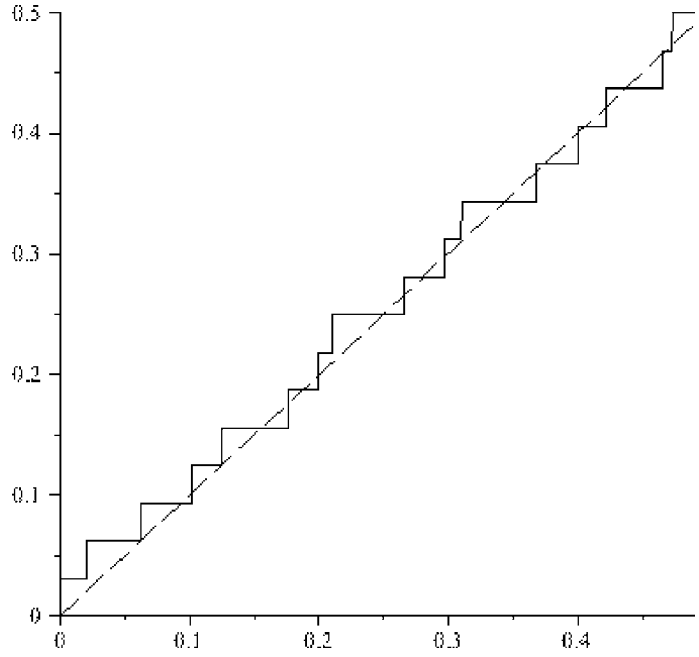
| $x_i$ | $x_i$ | $x_i$ | $x_i$ |
|---|---|---|---|
| 0.020 | 0.297 | 0.517 | 0.779 |
| 0.063 | 0.310 | 0.581 | 0.802 |
| 0.102 | 0.368 | 0.602 | 0.827 |
| 0.125 | 0.401 | 0.631 | 0.859 |
| 0.176 | 0.422 | 0.645 | 0.895 |
| 0.199 | 0.465 | 0.693 | 0.927 |
| 0.210 | 0.473 | 0.711 | 0.958 |
| 0.266 | 0.508 | 0.750 | 0.979 |

La distribuzione cumulativa empirica di frequenza del campione è quindi definita da

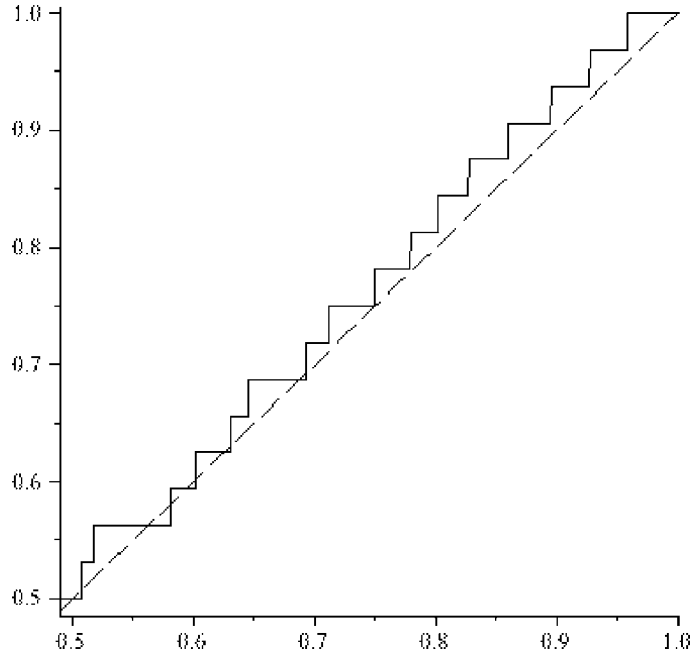
$$P_{e,n}(x) = \#\{x_i\,,\ i = 1, \ldots, n,\ \ x_i \leq x\} \,/\, n$$

in questo caso con $n = 32$. Ciascun "salto" della distribuzione cumulativa empirica ha dunque ampiezza $1/32 = 0.03125$.

Il confronto dei grafici delle due distribuzioni cumulative è illustrato nella figura seguente per l'intervallo $[0.000, 0.490]$




e in quella sottoriportata per l'intervallo $[0.490, 1.000]$



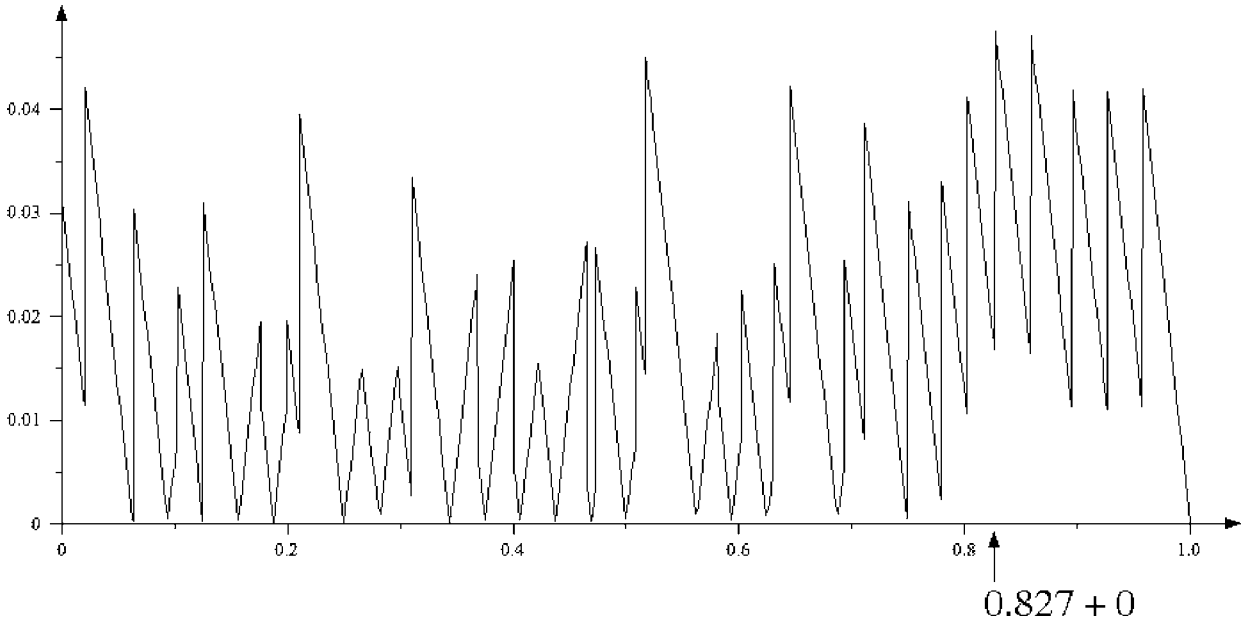
La variabile del test è quella di Kolmogorov-Smirnov

$$D_n \;=\; D_{32} \;=\; \sup_{x\in\mathbb{R}} \Big| P(x) - P_{e,32}(x) \Big|$$

che per questo campione assume il valore

$$D_{32} \;=\; 0.048$$

determinabile graficamente, esaminando con cura il grafico della funzione $\Big| P(x) - P_{e,32}(x) \Big|$ nell'intervallo $[0, 1]$:




oppure per via numerica (ad esempio con l'istruzione "maximize" di Maple applicata alla stessa funzione nello stesso intervallo).

Come evidenziato in figura, il massimo calcolato corrisponde al valore della funzione in $x = 0.827 + 0$ — il limite da destra della funzione (discontinua) in $x = 0.827$.

L'ipotesi nulla deve essere rigettata, con livello di significatività $\alpha$, se

$$D_n > \frac{Q^{-1}(\alpha)}{\sqrt{n}}$$

Nella fattispecie è $n = 32$ e $\alpha = 0.05$, per cui

$$Q^{-1}(0.05) = 1.35809$$

e la condizione di rigetto diventa

$$D_{32} > \frac{Q^{-1}(0.05)}{\sqrt{32}} = \frac{1.35809}{5.656854249} = 0.2400786621$$

Essendo però il valore calcolato della variabile di KS **minore** di quello critico

$$D_{32} = 0.048 < 0.2400786621$$

si conclude che **l'ipotesi nulla non può essere rigettata in base ai dati del campione**:

**si accetta l'ipotesi che i dati del campione siano estratti da una popolazione uniforme nell'intervallo [0,1]**

In altri termini, il generatore di numeri casuali in $[0, 1]$ **può considerarsi soddisfacente** in base al test di KS, secondo il livello di significatività preassegnato.

**Osservazione: calcolo della variabile di Kolmogorov-Smirnov**

Per determinare il valore della variabile $D_n$ in realtà non è necessario studiare il grafico di $|P(x) - P_{e,n}(x)|$ in $x \in \mathbb{R}$.

In primo luogo occorre ordinare i valori empirici $x_1, \ldots, x_n$ in senso crescente:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

essendosi posto $\forall\, j = 1, \ldots, n$

$$x_{(j)} = j\text{-esimo più piccolo valore tra } x_1, x_2, \ldots, x_n\,.$$

Poichè $P_{e,n}(x)$ risulta costante in ciascun intervallo di $x \in \mathbb{R}$ individuato dai punti empirici:

$$(-\infty, x_{(1)}) \qquad [x_{(1)}, x_{(2)}) \quad \ldots \quad [x_{(n-1)}, x_{(n)}) \quad [x_{(n)}, +\infty)$$

mentre la distribuzione cumulativa $P(x)$ è monotòna non decrescente, **gli scarti massimi fra le distribuzioni empirica e teorica possono verificarsi soltanto in corrispondenza dei punti** $x_1, \ldots, x_n$.

Basta perciò calcolare il massimo di un insieme **finito** di punti:

$$\max\left\{ \left| x_{(j)} - \frac{j-1}{n} \right|,\ \left| x_{(j)} - \frac{j}{n} \right|,\ j = 1, \ldots, n \right\}$$

**Osservazione: per ogni $\lambda > 0$ ed $n \in \mathbb{N}$ fissati, si ha che**

$$\text{Probabilità}\left\{ \sup_{x\in\mathbb{R}} \left| P(x) - P_{e,n}(x) \right| \; > \; \frac{\lambda}{\sqrt{n}} \right\}$$

**è indipendente dalla distribuzione $P(x)$**

Per la definizione di $P_{e,n}(x)$, tale probabilità vale infatti:

$$\text{Probabilità}\left\{ \sup_{x\in\mathbb{R}} \left| P(x) - \frac{\#\{i \,:\, x_i \le x\}}{n} \right| \; > \; \frac{\lambda}{\sqrt{n}} \right\}$$

ma siccome $P(x)$ è non decrescente si ha che:

$$x_i \;\le\; x \qquad \Longleftrightarrow \qquad P(x_i) \;\le\; P(x)$$

per cui la probabilità precedente diviene

$$\text{Probabilità}\left\{ \sup_{x\in\mathbb{R}} \left| P(x) - \frac{\#\{i \,:\, P(x_i) \le P(x)\}}{n} \right| \; > \; \frac{\lambda}{\sqrt{n}} \right\}.$$

La definizione di distribuzione cumulativa assicura inoltre che

$$P(x) \in [0,1] \qquad \forall\, x \in \mathbb{R}$$

e che le variabili casuali

$$u_i \;=\; P(x_i)\;, \quad i \;=\; 1, \ldots, n\,,$$

sono indipendenti e **uniformemente distribuite in** $[0,1]$. Ne deriva che, ponendo $u = P(x)$, la probabilità richiesta si può esprimere nella forma equivalente

$$\text{Probabilità}\left\{ \sup_{u\in[0,1]} \left| u - \frac{\#\{i \,:\, u_i \le u\}}{n} \right| \; > \; \frac{\lambda}{\sqrt{n}} \right\}.$$

in cui la sola dipendenza è da $n$ e da $\lambda$.

## 8.12 Criterio di Chauvenet per l'individuazione di outliers

Sia dato un campione ridotto di dati $x_1$, $x_2$, ... $x_n$

Si assuma che i dati siano estratti da una popolazione normale

Si supponga che un dato $x_k$ risulti molto lontano dalla media campionaria $\overline{x}$

Un dato che si colloca insolitamente lontano dal valore medio di una popolazione normale viene detto un **outlier** (punto che "giace fuori")

Un outlier $x_k$, se effettivamente appartiene alla popolazione normale, è un evento poco probabile, quindi sospetto

**Problema**: il dato $x_k$ appartiene effettivamente alla popolazione normale, o si tratta di un "intruso"?

**Una possibile strategia (criterio di Chauvenet):**

Si ponga

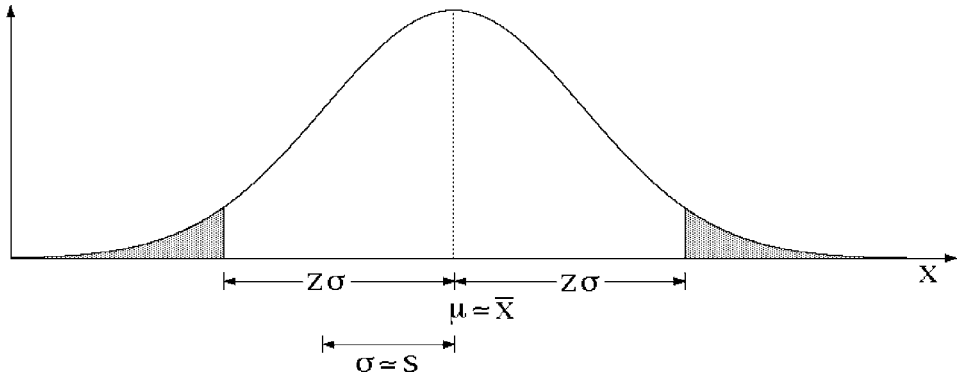
$$z = \frac{|x_k - \overline{x}|}{s} \simeq \frac{|x_k - \mu|}{\sigma}$$

dove $\mu$ e $\sigma$ sono sostituite dalle stime campionarie $\overline{x}$ e $s$ rispettivamente (**si tratta di una approssimazione!**)

$z$ è la distanza di $x_k$ dalla media campionaria $\overline{x}$ in unità di $s$

o, **approssimativamente**, dalla media $\mu$ in unità di $\sigma$

La probabilità $p(z)$ che un dato normale $x_k$ si collochi ad una distanza $\geq z$ dalla media è rappresentata dall'area tratteggiata nel seguente grafico della distribuzione $N(\mu, \sigma)$




La probabilità $p(z)$ può essere espressa come

$$p(z) \;=\; 1 - \int_{\mu - z\sigma}^{\mu + z\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

o, dopo un opportuno cambiamento di variabili, nelle forme equivalenti

$$p(z) \;=\; 1 - \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\xi^2/2} d\xi \;=\; 1 - \mathrm{erf}\Big(\frac{z}{\sqrt{2}}\Big)$$

dove erf indica la **funzione degli errori**:

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Il numero medio di dati ad una distanza $\geq z$ da $\mu$ in unità $\sigma$ **su** $n$ **misure** è

$$n\,p(z)$$

Si scelga $z$ in modo che il numero medio di dati indicato sopra sia significativamente minore di 1. Tipicamente si richiede che:

$$n\,p(z) \;=\; 1/2$$

ossia

$$1 - \mathrm{erf}\Big(\frac{z}{\sqrt{2}}\Big) \;=\; \frac{1}{2n} \qquad (1)$$

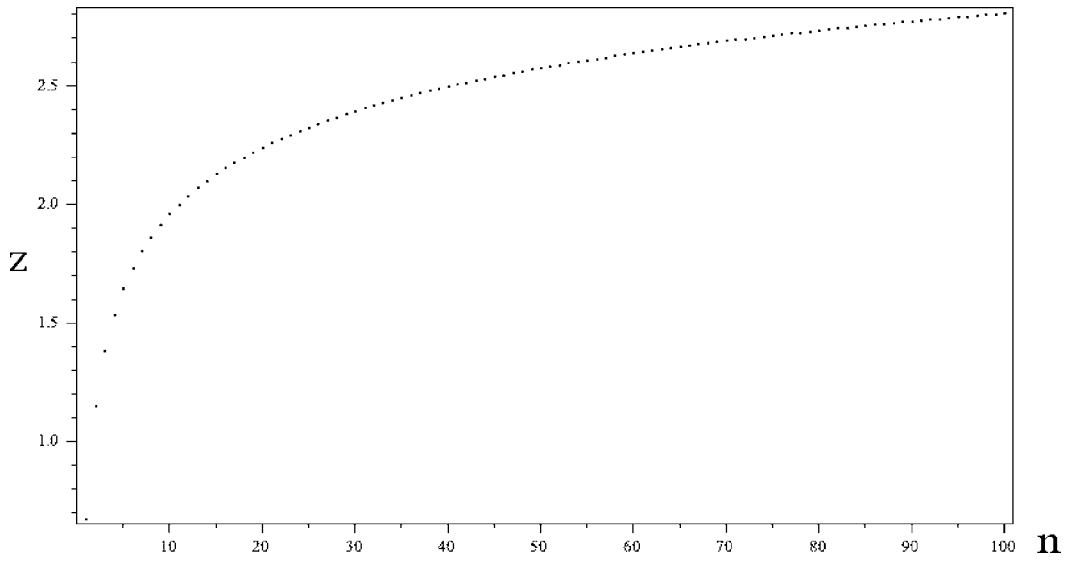
Usando lo $z$ precedente, se

$$\frac{|x_k - \overline{x}|}{s} \;\geq\; z$$

è improbabile che $x_k$ appartenga alla popolazione normale

In tal caso, **il dato sospetto dovrebbe essere rigettato** come non significativo (estraneo alla popolazione statistica)

Il grafico mostra il valore critico $z$ contro il numero $n$ dei dati



La tabella seguente riporta i valori critici $z$ in funzione del numero di dati, secondo la precedente equazione (1):

| $n$ dati | $z$ critico | $n$ dati | $z$ critico |
|---|---|---|---|
| 1 | 0.674489750 | 27 | 2.355084009 |
| 2 | 1.150349380 | 28 | 2.368567059 |
| 3 | 1.382994127 | 29 | 2.381519470 |
| 4 | 1.534120544 | 30 | 2.393979800 |
| 5 | 1.644853627 | 40 | 2.497705474 |
| 6 | 1.731664396 | 50 | 2.575829304 |
| 7 | 1.802743091 | 60 | 2.638257273 |
| 8 | 1.862731867 | 70 | 2.690109527 |
| 9 | 1.914505825 | 80 | 2.734368787 |
| 10 | 1.959963985 | 90 | 2.772921295 |
| 11 | 2.000423569 | 100 | 2.807033768 |
| 12 | 2.036834132 | 150 | 2.935199469 |
| 13 | 2.069901831 | 200 | 3.023341440 |
| 14 | 2.100165493 | 250 | 3.090232306 |
| 15 | 2.128045234 | 300 | 3.143980287 |
| 16 | 2.153874694 | 350 | 3.188815259 |
| 17 | 2.177923069 | 400 | 3.227218426 |
| 18 | 2.200410581 | 450 | 3.260767488 |
| 19 | 2.221519588 | 500 | 3.290526731 |
| 20 | 2.241402728 | 550 | 3.317247362 |
| 21 | 2.260188991 | 600 | 3.341478956 |
| 22 | 2.277988333 | 650 | 3.363635456 |
| 23 | 2.294895209 | 700 | 3.384036252 |
| 24 | 2.310991338 | 800 | 3.420526701 |
| 25 | 2.326347874 | 900 | 3.452432937 |
| 26 | 2.341027138 | 1000 | 3.480756404 |

**Esempio: criterio di Chauvenet**

Misure ripetute di una lunghezza $x$ hanno fornito i seguenti risultati (in mm):

$$46\,,\quad 48\,,\quad 44\,,\quad 38\,,\quad 45\,,\quad 47\,,\quad 58\,,\quad 44\,,\quad 45\,,\quad 43$$

che si possono assumere estratti da una popolazione normale. Applicando il criterio di Chauvenet, si verifichi se il risultato anomalo (outlier) $x_{\text{out}} = 58$ debba essere escluso o meno.

**Soluzione**
La media e la deviazione standard campionarie sono date da

$$\overline{x} \,=\, \frac{1}{10}\sum_{i=1}^{10} x_i \,=\, 45.8 \qquad\qquad s \,=\, \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2} \,=\, 5.1$$

La distanza del valore sospetto dalla media, in unità $s$, vale

$$\frac{x_{\text{out}} - \overline{x}}{s} \,=\, \frac{58 - 45.8}{5.1} \,=\, 2.4$$

La probabilità che un dato cada ad una distanza maggiore di 2.4 deviazioni standard dalla media può ricavarsi dalla tabella della distribuzione cumulativa per la variabile normale standard:

$$\begin{aligned} P(|x_{\text{out}} - \overline{x}| \geq 2.4s) \,&=\, 1 - P(|x_{\text{out}} - \overline{x}| < 2.4s) \,= \\ &=\, 1 - 2 \cdot P(\overline{x} \leq x_{\text{out}} < \overline{x} + 2.4s) \,= \\ &=\, 1 - 2 \cdot 0.49180 \,=\, 0.0164 \end{aligned}$$

Su 10 misure ci si aspettano, tipicamente, $10 \cdot 0.0164 = 0.164$ risultati "cattivi", ad una distanza di oltre $2.4s$ dalla media. **Poichè** $0.164 < 1/2$**, il criterio di Chauvenet suggerisce che l'outlier** $x_{\text{out}}$ **debba essere rigettato.**

# 9. COPPIE DI VARIABILI CASUALI
## 9.1 Coefficiente di correlazione lineare (r di Pearson)

Sia data una coppia di variabili casuali $(x, y)$

Problema: sono stocasticamente dipendenti o indipendenti?

Il caso estremo di dipendenza stocastica è quello in cui
una variabile è **funzione dell'altra**
y=f(x)

Caso notevole è quello lineare, per il quale si assume
$\mathbf{y = ax + b}$, con **a** e **b** costanti assegnate ($\mathbf{a \neq 0}$)

La distribuzione congiunta di probabilità di $(x, y)$
si scrive formalmente come
$$p(x, y) = p(x)\,\delta(ax + b - y)$$
in termini della funzione generalizzata **Delta di Dirac** $\delta$

Medie, varianze e covarianza di $(x, y)$ si scrivono allora:

$$\mu_x = \int_{\mathbb{R}^2} p(x)\,\delta(ax+b-y)x\,dydx = \int_{\mathbb{R}} p(x)x\,dx$$

$$\mu_y = \int_{\mathbb{R}^2} p(x)\,\delta(ax+b-y)(ax+b)\,dydx = a\mu_x + b$$

$$\mathrm{var}(x) = \int_{\mathbb{R}^2} p(x)\,\delta(ax+b-y)(x-\mu_x)^2 dydx = \int_{\mathbb{R}} p(x)(x-\mu_x)^2 dx$$

$$\mathrm{var}(y) = \int_{\mathbb{R}^2} p(x)\,\delta(ax+b-y)(y-\mu_y)^2 dydx =$$

$$= \int_{\mathbb{R}} p(x)(ax-a\mu_x)^2 dx = a^2 \int_{\mathbb{R}} p(x)(x-\mu_x)^2 dx$$

$$\mathrm{cov}(x,y) = \int_{\mathbb{R}^2} p(x)\,\delta(ax+b-y)(x-\mu_x)(y-\mu_y) dydx =$$

$$= \int_{\mathbb{R}} p(x)(ax-a\mu_x)(x-\mu_x) dx = a \int_{\mathbb{R}} p(x)(x-\mu_x)^2 dx$$

per cui la relativa correlazione diventa

$$\mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{a}{|a|} = \begin{cases} +1 & \text{se } a > 0 \\ -1 & \text{se } a < 0 \end{cases}$$

**Se $(x, y)$ sono indipendenti si ha invece** $\mathrm{corr}(x, y) = 0$

Qualora la $y$ sia una funzione $f(x)$ **non lineare** della variabile casuale $x$, la correlazione fra $x$ e $y$

$$\text{corr}(x, y) = \frac{\int_{\mathbb{R}} p(x)(x - \mu_x)[f(x) - \mu_y]dx}{\sqrt{\int_{\mathbb{R}} p(x)(x - \mu_x)^2 dx} \sqrt{\int_{\mathbb{R}} p(x)[f(x) - \mu_y]^2 dx}}$$

con

$$\mu_y = \int_{\mathbb{R}} p(x) f(x)\, dx$$

**non ha alcun significato particolare**

Nondimeno, i valori di $\text{corr}(x, y)$ potranno essere ritenuti un indice di:

– **dipendenza lineare diretta** ($a > 0$), se prossimi a $+1$;

– **dipendenza lineare inversa** ($a < 0$), se prossimi a $-1$;

– **indipendenza stocastica**, se vicini a 0

Della coppia di variabili casuali $(x, y)$ sia disponibile un campione ridotto $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$ con valori medi $\overline{x}$ e $\overline{y}$ rispettivamente

La correlazione è allora stimata facendo uso del campione ridotto, per mezzo del **coefficiente di correlazione lineare** (o di Pearson)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

In effetti segue dalla diseguaglianza di Cauchy-Schwarz

$$\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2}\ \sqrt{\sum_{i=1}^{n} b_i^2} \qquad \forall a_i, b_i \in \mathbb{R} \quad i = 1, \ldots, n$$

che la condizione $r = +1$ o $r = -1$ equivale all'avere tutte le coppie $(x_i, y_i)$ del campione ridotto allineate lungo una stessa retta, di slope positiva o negativa rispettivamente

## 9.2 Statistica del coefficiente di correlazione lineare

Appurato che le variabili casuali $(x, y)$ sono stocasticamente dipendenti, il coefficiente di correlazione lineare esprime convenzionalmente la "intensità" di questa dipendenza

Ma il coefficiente $r$ **non è in generale un buon indice** per testare la dipendenza o indipendenza stocastica delle variabili $(x, y)$

In generale infatti, anche assumendo vera l'ipotesi nulla

$$H_0 : \text{ le variabili } x \text{ e } y \text{ sono indipendenti}$$

la distribuzione di probabilità di $r$ non risulta determinata, **non essendo note quelle di** $x$ **e** $y$

$$\Downarrow$$

Non è possibile costruire alcun test per saggiare la fondatezza dell'ipotesi $H_0$, con un appropriato **livello di significatività**

**Occorre riferirsi ad alcuni casi notevoli**, per i quali la distribuzione di $r$ sia almeno approssimativamente determinabile

### 9.2.1 Variabili $(x, y)$ indipendenti

Per una coppia di variabili casuali $(x, y)$ **indipendenti**, le cui distribuzioni di probabilità abbiano un numero sufficientemente elevato di **momenti convergenti**, se $n$ è abbastanza grande $(n > 20)$

$$\Downarrow$$

$r$ segue approssimativamente una distribuzione normale con media nulla e varianza $1/n$

$$\Downarrow$$

La regione di rigetto dell'ipotesi nulla sarà l'unione di intervalli

$$\{|r| \geq \eta\} = \{r < -\eta\} \cup \{r > \eta\}$$

dove $\eta > 0$ è determinato dall'equazione

$$\frac{\alpha}{2} = \sqrt{\frac{n}{2\pi}} \int_{\eta}^{+\infty} e^{-z^2 n/2} dz = \frac{1}{2} \mathrm{erfc}\left(\eta \sqrt{\frac{n}{2}}\right)$$

e erfc indica la **funzione degli errori complementare**

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-\xi^2} d\xi$$

Al solito, $\alpha \leq 1/2$ è la probabilità di errore del I tipo

### 9.2.2 Variabili $(x, y)$ gaussiane

Per variabili $(x, y)$ **gaussiane**, con distribuzione

$$p(x,y) = \frac{1}{2\pi}\sqrt{a_{11}a_{22} - a_{12}^2}\, e^{-\frac{1}{2}(a_{11}x^2 - 2a_{12}xy + a_{22}y^2)}$$

$$a_{11}a_{22} - a_{12}^2 > 0 \qquad a_{11} + a_{22} > 0$$

e quindi correlazione

$$\rho = \operatorname{corr}(x,y) = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} \in (-1,1)$$

la distribuzione teorica di frequenza di $r$ dipende da $\rho$ e dal numero $n$ di punti.

Definita per $r \in (-1,1)$, tale distribuzione si scrive:

$$p_{\rho,n}(r) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}a_{\rho,n}(r)$$

con

$$a_{\rho,n}(r) = \sum_{s=0}^{\infty}\Gamma^2\Big(\frac{n+s-1}{2}\Big)\frac{(2\rho r)^s}{s!}$$

La distribuzione è un po' troppo complessa per essere utilizzata direttamente

Di solito si fa uso di una **approssimazione** valida per **grandi campioni** $(n > 10)$

Oppure ci si riferisce al caso $\rho = 0$ di **variabili scorrelate** $\Longleftrightarrow$ indipendenti (quello che usualmente interessa)

#### 9.2.2.1 Approssimazione per grandi $n$ Trasformazione di Fisher

Per $n$ abbastanza grande ($n \geq 10$), la variabile $z$ di Fisher:

$$z \;=\; \operatorname{arctanh} r \;=\; \frac{1}{2} \ln \frac{1+r}{1-r}$$

segue una distribuzione approssimativamente normale di media

$$\overline{z} \;=\; \frac{1}{2}\Big[\ln \frac{1+\rho}{1-\rho} + \frac{\rho}{n-1}\Big]$$

e deviazione standard

$$\sigma(z) \;\simeq\; \frac{1}{\sqrt{n-3}}$$

#### 9.2.2.2 Variabili gaussiane indipendenti (o scorrelate)

Per $\rho = 0$ la distribuzione di $r$ si riduce a:

$$p_{0,n}(r) \;=\; \frac{\Gamma[(n-1)/2]}{\sqrt{\pi}\Gamma[(n-2)/2]} (1-r^2)^{(n-4)/2}$$

in modo che la variabile definita da:

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

si distribuisce secondo una $t$ di Student a $n-2$ gradi di libertà.

Nel caso di **variabili gaussiane** $(x, y)$ è quindi possibile testare l'ipotesi

$$H_0 \ : \ x \text{ e } y \text{ sono variabili indipendenti}$$

che potrà ritenersi accettabile per valori della variabile del test

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

abbastanza vicini a 0 (allorquando anche $r$ risulta circa nullo)

L'ipotesi nulla $H_0$ dovrà essere rigettata se all'opposto $t$ (e dunque $r$) si presenta sensibilmente diversa da zero

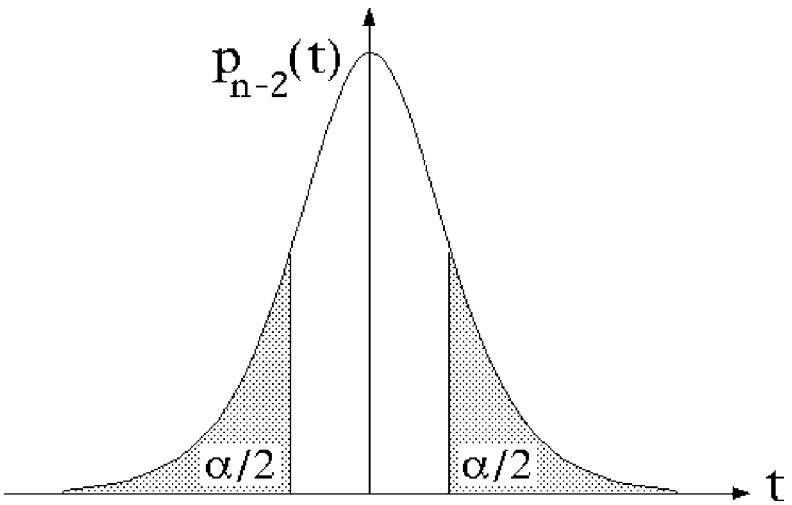
Se ne trae un tipico **test bilaterale**, con una regione di rigetto definita da

$$\left\{t \leq t_{[\frac{\alpha}{2}](n-2)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n-2)}\right\}$$

ossia

$$\left\{t \leq -t_{[1-\frac{\alpha}{2}](n-2)}\right\} \cup \left\{t \geq t_{[1-\frac{\alpha}{2}](n-2)}\right\}$$

con livello di significatività $\alpha \leq 1/2$



**Esempio**: **Coefficiente di correlazione lineare (Pearson)**
Si consideri la seguente tabella di dati, relativi ad alcune misure ripetute di due grandezze $x$ e $y$ (in unità arbitrarie), ottenute nelle stesse condizioni sperimentali. Il campione si assume normale.

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 1.8 | 1.1 |
| 2 | 3.2 | 1.6 |
| 3 | 4.3 | 1.1 |
| 4 | 5.9 | 3.5 |
| 5 | 8.1 | 3.8 |
| 6 | 9.9 | 6.2 |
| 7 | 11.4 | 5.6 |
| 8 | 12.1 | 4.0 |
| 9 | 13.5 | 7.5 |
| 10 | 17.9 | 7.1 |

Usare il coefficiente di correlazione lineare di Pearson per verificare se le grandezze $x$ e $y$ possano considerarsi stocasticamente indipendenti, con un livello di significatività del 5% e dell'1%.

**Soluzione**
Le medie campionarie delle due grandezze sono date da

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 8.8100 \qquad \overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 4.1500$$

e permettono di calcolare le seguenti somme di scarti

$$SS_{xy} = \sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) = 99.6050$$

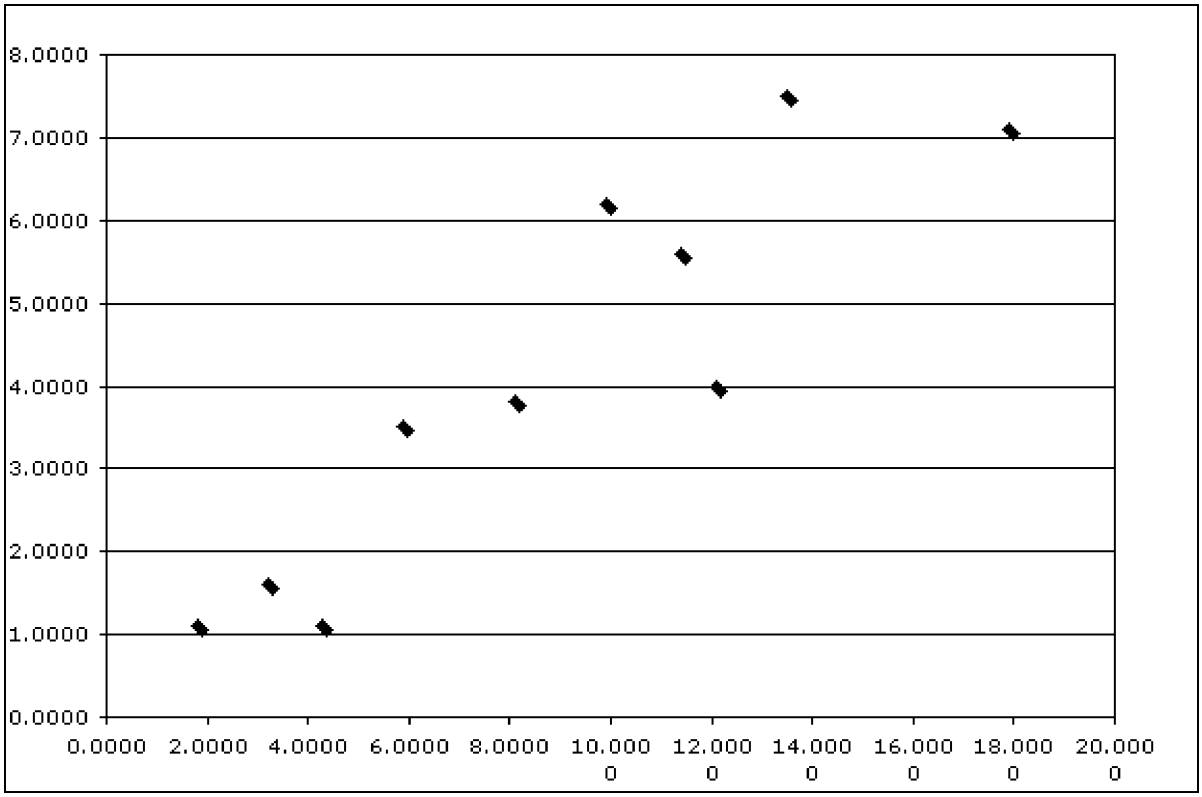
$$SS_{xx} = \sum_{i=1}^{10}(x_i - \overline{x})^2 = 233.2690$$

$$SS_{yy} = \sum_{i=1}^{10}(y_i - \overline{y})^2 = 51.9050$$

in modo che il coefficiente di correlazione lineare diventa

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}} = \frac{99.6050}{\sqrt{233.2690}\sqrt{51.9050}} = 0.9052$$

Il grafico dei dati suggerisce che $x$ e $y$ siano dipendenti:



Poichè $x$ e $y$ possono assumersi normali, si può testare l'ipotesi nulla

$$H_0 \ : \ x \text{ e } y \text{ sono stocasticamente independenti}$$

contro l'ipotesi alternativa

$$H_1 \ : \ x \text{ e } y \text{ sono stocasticamente dipendenti}$$

usando la variabile casuale

$$t \ = \ \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

che, per $H_0$ vera, segue una distribuzione di Student a $n-2$ g.d.l. Nella fattispecie si ha

$$t \ = \ \sqrt{10-2}\frac{0.9052}{\sqrt{1-0.9052^2}} \ = \ 6.0247$$

La regione di rigetto assume la forma

$$|t| \ > \ t_{[1-\frac{\alpha}{2}](n-2)} \ = \ t_{[0.975](8)} \ = \ 2.306$$

per un livello di significatività $\alpha = 5\%$, e la forma

$$|t| \ > \ t_{[1-\frac{\alpha}{2}](n-2)} \ = \ t_{[0.995](8)} \ = \ 3.355$$

qualora il livello di significatività richiesto sia $\alpha = 1\%$.

**In entrambi i casi $H_0$ deve essere rifiutata!**

## 9.3 Osservazione

Si è esaminato il caso che il coefficiente di correlazione lineare $r$ venga calcolato per un campione ridotto di **due** variabili casuali $(x, y)$

Sovente però il coefficiente è calcolato in una situazione del tutto diversa, quella in cui le coppie $(x_i, y_i)$:

- **non** possono essere considerate come individui estratti da **una stessa popolazione** statistica
- sono raggruppabili in un certo numero di **sottoinsiemi**, ciascuno dei quali appartiene **ad una propria popolazione** statistica

Esempio: gli esperimenti vòlti a evidenziare le variazioni subite da una grandezza $y$ al variare di una seconda grandezza $x$

I dati $(x_i, y_i)$ **non sono** le realizzazioni di **una** coppia di variabili casuali

$$\Downarrow$$

$r$ **non** presenta una distribuzione di probabilità nota e **non può** essere utilizzato per testare la dipendenza o indipendenza stocastica delle grandezze corrispondenti a $x$ ed $y$

$r$ fornisce una indicazione sul livello di allineamento dei dati $(x_i, y_i)$, ma non offre alcuna informazione di tipo statistico riguardo la significatività della dipendenza lineare proposta

Tale significatività deve essere verificata altrimenti: **problema della regressione lineare** e della **goodness of fit**

# 10. MODELLAZIONE DEI DATI

## 10.1 Impostazione generale del problema

Osservazioni

$\Downarrow$

Dati del tipo $(x_i, y_i)$, $i = 1, \ldots, n$

$\Downarrow$

Sintesi dei dati in un "modello"

"Modello" è una funzione matematica
che lega fra loro le grandezze osservate
e dipende da opportuni parametri "aggiustabili"

L'adattamento dei parametri aggiustabili ai dati è detto
**"fitting"**

La procedura di fitting consiste di regola:

nella definizione di una appropriata **funzione obiettivo** (merit function), che misura l'accordo fra i dati ed il modello per una data scelta dei parametri;

nel susseguente calcolo dei parametri del modello mediante la ottimizzazione della funzione obiettivo (best fit)

Di norma si conviene che

$$\begin{array}{ccc} \text{buon accordo} & & \text{valori piccoli} \\ \text{modello/dati} & \Longleftrightarrow & \text{della funzione obiettivo} \end{array}$$

per cui

$$\begin{array}{ccc} & & \text{ricerca del minimo} \\ \text{best fit} & \Longleftrightarrow & \text{della funzione obiettivo} \end{array}$$

Non si tratta però di un semplice problema di ottimizzazione (ricerca dei valori di best-fit dei parametri):

- i dati in genere sono affetti da errore e devono essere riguardati come le realizzazioni di appropriate variabili aleatorie;

- i parametri di best-fit del modello sono di conseguenza soggetti a errore e vanno assunti a loro volta come variabili casuali;

- dei parametri di best-fit si vogliono determinare gli errori ed eventualmente la distribuzione di probabilità;

- si richiedono opportuni test per verificare se il modello sia adeguato o meno a descrivere i dati (**goodness of fit**)

### Funzione obiettivo

La scelta della funzione obiettivo è affidata sovente al c.d. **metodo della massima verosimiglianza** (**maximum likelihood method**)

Si assume di determinare i parametri aggiustabili in modo che i dati effettivamente misurati costituiscano un evento di **probabilità massima**: il set di dati misurati è quello più probabile, il cui essersi verificato è più verosimile

Supposto che le variabili $x_i$ abbiano varianza trascurabile e possano considerarsi pressoché certe, e che le corrispondenti $y_i$ siano variabili normali indipendenti di varianza $\sigma_i^2$, la **maximum likelihood** conduce al **metodo del chi quadrato**

Qualora nelle stesse ipotesi le varianze delle variabili $y_i$ possano considerarsi uguali, il maximum likelihood method si traduce nel **least-squares fitting**

Il metodo di massima verosimiglianza può essere applicato pure nel caso di variabili **non gaussiane**

Talvolta le distribuzioni di probabilità delle $y_i$ non possono considerarsi gaussiane causa la presenza di realizzazioni sensibilmente distanti dai corrispondenti valori medi, il cui ricorrere sarebbe pressoché impossibile nel caso gaussiano (**outliers**, punti che distano dalla media più di 3 deviazioni standard)

In tal caso, le funzioni obiettivo dedotte per mezzo della maximum likelihood caratterizzano i c.d. metodi di fitting "robusti" (**robust fitting methods**)

## Relazione fra Maximum Likelihood e chi-quadrato

Probabilità di ottenere le coppie di dati $(x_i, y_i)$, $i = 1, \ldots, n$

$$P(x_1, y_1, \ldots, x_n, y_n) \sim \prod_{i=1}^{n} e^{-\frac{1}{2\sigma_i^2}[y_i - R(x_i)]^2} =$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}[y_i - R(x_i)]^2}$$

nell'ipotesi di variabili $x_i$ certe e di $y_i$ normali indipendenti, di media $R(x_i)$ e varianza $\sigma_i^2$

Il metodo di Maximum Likelihood consiste nel richiedere che

$$P(x_1, y_1, \ldots, x_n, y_n) \qquad \text{sia massima}$$

ossia che

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}[y_i - R(x_i)]^2 \qquad \text{sia minima}$$

cioè che il chi-quadrato dei dati $y_i$ rispetto alla funzione di regressione $R(x)$ sia minimo — **metodo del chi-quadrato**

## 10.2 Regressione lineare con il metodo del chi quadrato (chi-square fitting)

Sia

$$\left\{(x_i, y_i) \in \mathbb{R}^2\,,\ 1 \leq i \leq n\right\}$$

un insieme di dati soddisfacente le seguenti ipotesi:

($i$) per ogni $i = 1, \ldots, n$ il dato $y_i$ è la realizzazione di una variabile casuale gaussiana di varianza nota $\sigma_i^2$. Il valore medio di tale variabile casuale è a priori incognito, ma si assume dipendere esclusivamente dall'ascissa $x_i$ — di cui sarà funzione;

($ii$) l'insieme dei dati è descritto dal modello matematico

$$\zeta_i \,=\, \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \,+\, \varepsilon_i \quad , \qquad 1 \leq i \leq n \ , \tag{2}$$

nel quale le $\phi_j \ : \ \mathbb{R} \longrightarrow \mathbb{R}$, $1 \leq j \leq p$, sono $p < n$ funzioni preassegnate e gli $\alpha_j$, $1 \leq j \leq p$, dei parametri costanti da stimare convenientemente, mentre le $\varepsilon_i$ costituiscono un insieme di $n$ variabili gaussiane indipendenti, di media nulla e varianza $\sigma_i^2$

$$\mathbb{E}(\varepsilon_i) \,=\, 0 \qquad \mathbb{E}(\varepsilon_i \varepsilon_k) \,=\, \delta_{ik}\sigma_k^2 \quad 1 \leq i, k \leq n \ .$$

Il valore $y_i$ deve dunque intendersi come una realizzazione della variabile casuale $\zeta_i$, per ogni $i$.

Si definisce **regressione lineare** del set di dati, con il metodo del chi quadrato (**chi-square fitting**), l'espressione

$$R(x) = \sum_{j=1}^{p} a_j \phi_j(x) \quad , \qquad x \in \mathbb{R} \ ,$$

nella quale gli $a_j$ sono le stime dei parametri $\alpha_j$ ricavate per mezzo del campione $\{(x_i, y_i)\,,\ 1 \leq i \leq n\}$ minimizzando il funzionale quadratico $\mathcal{L} : \mathbb{R}^p \longrightarrow \mathbb{R}$ definito da

$$\mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[ -y_i + \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2 \qquad (3)$$

Qualora le varianze $\sigma_i^2$ delle singole variabili casuali $\varepsilon_i$ siano tutte uguali, il loro comune valore può essere ignorato nella ottimizzazione del funzionale (3), che si riduce pertanto a

$$\mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \left[ -y_i + \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2$$

In tal caso si parla di **regressione lineare con il metodo dei minimi quadrati** (**least-squares fitting**)

Il più generale metodo di regressione lineare del chi quadrato è anche noto come **metodo dei minimi quadrati pesato** (**weighted least-squares fitting**)

## Ottimizzazione. Equazione normale e sua soluzione

I punti critici del funzionale quadratico (3) sono tutte e sole le soluzioni $a$ dell'**equazione normale**

$$Fa \;=\; \Phi\sigma^{-1}y \;, \tag{4}$$

nella quale le matrici $F$ e $\Phi$ sono definite da

$$F_{jk} \;=\; \sum_{i=1}^{n} \frac{\phi_j(x_i)\phi_k(x_i)}{\sigma_i^2} \qquad 1 \le j,k \le p \tag{5}$$

e

$$\Phi_{ki} \;=\; \frac{\phi_k(x_i)}{\sigma_i} \quad 1 \le k \le p\,,\; 1 \le i \le n \tag{6}$$

mentre

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} \quad \sigma^{-1} = \begin{pmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \mathbb{O} \\ & & \ddots & \\ & \mathbb{O} & & 1/\sigma_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Vale in particolare che $F = \Phi\Phi^T$, matrice reale simmetrica semidefinita positiva. Nell'ipotesi che la matrice $p \times n$ $\Phi$ abbia rango massimo $p$, allora $F$ risulta definita positiva e l'unica soluzione dell'equazione normale (4) è data da

$$a \;=\; F^{-1}\Phi\sigma^{-1}y \;. \tag{7}$$

## I parametri di regressione lineare, con il metodo del chi quadro, come variabili casuali
## Media e matrice di covarianza

Se $\Phi$ ha rango massimo $p$, la stima (7) dei parametri di regressione costituisce un insieme di $p$ variabili gaussiane con valori medi

$$\mathbb{E}(a_j) \,=\, \alpha_j \;, \quad 1 \leq j \leq p \,,$$

e matrice di covarianza

$$C_a \,=\, \mathbb{E}\Big[\big[a - \mathbb{E}(a)\big]\big[a - \mathbb{E}(a)\big]^T\Big] \,=\, F^{-1} \tag{8}$$

ovvero matrice di struttura $F$

**N.B.**

$$a_1, \ldots, a_p \text{ indipendenti} \qquad \Longleftrightarrow \qquad F \text{ diagonale}$$

## Somma normalizzata dei quadrati degli scarti attorno alla regressione

Nell'ipotesi che $\Phi$ abbia rango massimo $p < n$, la somma normalizzata dei quadrati degli scarti attorno alla regressione

$$\text{NSSAR} \,=\, \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big[-y_i + \sum_{j=1}^{p} \phi_j(x_i) a_j\Big]^2 \tag{9}$$

costituisce una variabile di $\mathcal{X}^2$ a $n - p$ gradi di libertà.

## Indipendenza stocastica dai parametri stimati

La NSSAR è una variabile casuale stocasticamente indipendente da ciascuno dei parametri stimati $a_1, \ldots, a_p$ — definiti dalla (7).

## Goodness of fit

Valori troppo grandi di NSSAR indicano che presumibilmente il modello non è corretto e deve essere rigettato

L'ipotesi nulla

$$H_0 \ : \ \text{il modello proposto è corretto}$$

può dunque essere verificata con un test basato sulla variabile NSSAR, e la cui regione di rigetto abbia della forma

$$\left\{ \text{NSSAR} \geq \mathcal{X}^2{}_{[1-\alpha](n-p)} \right\}$$

essendo $\alpha$, al solito, la probabilità di errore del I tipo

Nella pratica comune sovente si preferisce calcolare la probabilità $Q$ che la variabile NSSAR assuma un valore pari a quello osservato o superiore

$$Q = \int_{\text{NSSAR}}^{+\infty} p_{n-p}(\mathcal{X}^2)\, d\mathcal{X}^2 =$$

$$= \int_{\text{NSSAR}}^{+\infty} \frac{1}{\Gamma[(n-p)/2]\, 2^{(n-p)/2}} e^{-\mathcal{X}^2/2} (\mathcal{X}^2)^{\frac{n-p}{2}-1}\, d\mathcal{X}^2$$

e usare $Q$ come indicatore della **goodness of fit** del modello

Valori **molto piccoli** di $Q$ inducono a **rigettare il modello** come inadeguato

**Che succede se $Q$ risulta molto piccolo?**

Si hanno tre possibilità:

($i$) il modello è errato, da rigettare

($ii$) le varianze delle variabili normali $y_i$ o $\varepsilon_i$ non sono state calcolate correttamente (il test le suppone note esattamente): NSSAR non segue una distribuzione di $\mathcal{X}^2$ o comunque nota

($iii$) le variabili casuali $y_i$ o $\varepsilon_i$ non seguono una distribuzione normale (o non tutte, almeno): NSSAR **non è** una variabile di $\mathcal{X}^2$ e la sua distribuzione di probabilità risulta sconosciuta

La circostanza ($iii$) è piuttosto frequente, come segnalato dalla presenza di **outliers**

**Che si fa?**
Si possono tollerare valori di $Q$ molto più piccoli di quelli calcolabili dalla distribuzione cumulativa di $\mathcal{X}^2$ **senza rigettare il modello**

Se le distribuzioni (**non gaussiane**) delle singole variabili $y_i$ o $\varepsilon_i$ sono note, si può applicare il **Metodo di Monte Carlo**:

si generano a caso i valori delle $y_i$;

si calcolano i valori di best-fit $a_j$ dei parametri $\alpha_j$ del modello

si calcola il valore di NSSAR;

ripetendo la procedura precedente un grande numero di volte, si ricavano per istogrammazione le distribuzioni di probabilità approssimate di NSSAR e dei coefficienti di best-fit $a_j$;

è così possibile **stimare** $Q$ (goodness of fit), nonché l'errore sui coefficienti di best-fit (**molto oneroso!**)

## Che succede se $Q$ è molto grande?

Di solito il problema non può imputarsi al fatto che le variabili $y_i$ non siano normali

Distribuzioni alternative a quella gaussiana in genere determinano una maggiore dispersione dei dati, un valore più grande della NSSAR calcolata e dunque un $Q$ più piccolo

Sovente il problema nasce da una **sopravvalutazione** delle deviazioni standard $\sigma_i$, per cui il valore calcolato della NSSAR risulta più piccolo del dovuto, e $Q$ più grande

## Come capire se $Q$ è troppo piccolo o troppo grande?

Basta ricordare che se il modello è corretto, le variabili $y_i$ gaussiane e le varianze $\sigma_i^2$ esatte, la NSSAR segue una distribuzione di $\mathcal{X}^2$ a $\nu = n - p$ gradi di libertà

La distribuzione di $\mathcal{X}^2$ a $\nu$ gradi di libertà ha media $\nu$ e varianza $2\nu$, e per grandi $\nu$ si approssima ad una distribuzione normale con media $\nu$ e varianza $2\nu$

Di conseguenza, valori tipici della variabile NSSAR sono compresi nell'intervallo

$$[\nu - 3\sqrt{2\nu}, \nu + 3\sqrt{2\nu}]$$

come regola generale

Se il valore calcolato di NSSAR cade fuori da questo intervallo, il risultato è sospetto

**Osservazione. Deviazioni standard $\sigma_i$ ignote**

Se le deviazioni standard $\sigma_i$ delle variabili $y_i$ sono sconosciute, la NSSAR non può essere determinata

**Se tuttavia**:

($i$) il modello è corretto;

($ii$) si può ritenere che tutte le varianze $\sigma_i$ siano uguali ad un comune valore $\sigma$;

**si può**:

– porre $\sigma = 1$;

– calcolare la soluzione di best-fit per i parametri del modello;

– determinare la relativa NSSAR $\implies$ NSSAR$|_{\sigma^2=1}$;

– stimare un valore verosimile di $\sigma$ per mezzo della relazione

$$\sigma^2 \;=\; \frac{1}{n-p}\mathrm{NSSAR}|_{\sigma^2=1} \;=\; \frac{1}{n-p}\sum_{i=1}^{n}\Big[-y_i + \sum_{j=1}^{p}\phi_j(x_i)a_j\Big]^2$$

in quanto NSSAR è $\mathcal{X}^2$ a $n-p$ gradi di libertà

$$\mathbb{E}[\mathrm{NSSAR}|_{\sigma^2=1}] \;=\; \sigma^2\mathbb{E}[\mathrm{NSSAR}] \;=\; \sigma^2(n-p)$$

**N.B.**: questo metodo non può provvedere alcuna indicazione sulla goodness of fit (il modello è assunto **a priori** corretto)!

## 10.3 $F$-test per la goodness of fit

Anziché per mezzo della NSSAR, variabile di $\mathcal{X}^2$ a $n-p$ gradi di libertà,

la goodness of fit può essere stimata mediante una opportuna variabile di Fisher (**$F$-test**)

• Si considerano modelli di regressione **con un parametro additivo**

$$\zeta_i \;=\; \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \qquad \forall\, i = 1, \ldots, n$$

e la funzione di regressione stimata viene indicata con

$$R(x) \;=\; a_1 + \sum_{j=2}^{p} a_j \phi_j(x)$$

• **Se il modello è corretto**, la variabile

$$\text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left[ y_i - a_1 - \sum_{j=2}^{p} a_j \phi_j(x_i) \right]^2 \;=\; {\mathcal{X}^2}_{n-p}$$

è una variabile di chi quadrato a $n-p$ gradi di libertà

Vero anche in mancanza di parametri additivi
e **quali che siano i valori di** $\alpha_1, \ldots, \alpha_p$
(per $\Phi$ di rango massimo $p$)

• È verificata l'**"equazione delle variazioni"**

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \overline{y})^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\big[R(x_i) - \overline{y}\big]^2 + \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\big[y_i - R(x_i)\big]^2$$

in cui compaiono le espressioni così definite:

$$\overline{y} = \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big)^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} y_i \quad \text{media pesata delle } y$$

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \overline{y})^2 \qquad \text{variazione totale dei dati}$$

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\big[R(x_i) - \overline{y}\big]^2 \quad \text{variazione "spiegata" dalla regressione}$$

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\big[y_i - R(x_i)\big]^2 \quad \text{variazione residua (NSSAR)}$$

Vale anche se il modello di regressione non risulta corretto, purché siano:

($i$) il modello lineare nei parametri di regressione

($ii$) i parametri stimati con il metodo del chi-quadrato

($iii$) **un parametro additivo!**

In mancanza di parametro additivo, l'equazione non è in generale soddisfatta

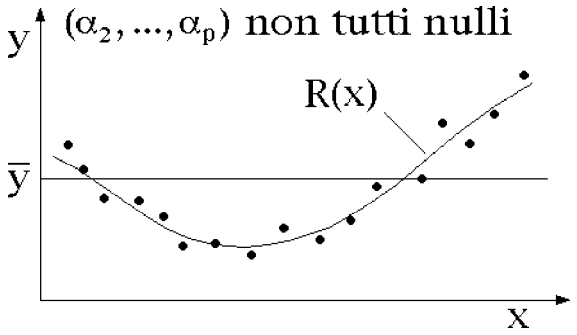
Se $y$ **dipende** effettivamente da $x$ ($y(x) \neq$ costante)
ci si aspetta un modello di regressione del tipo

$$\zeta_i \;=\; \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \qquad \forall\, i = 1, \ldots, n$$

con i parametri $\alpha_2, \ldots, \alpha_p$ **non tutti nulli**

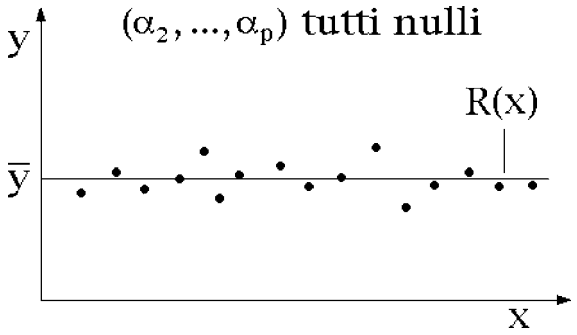
Se poi il modello è corretto la sovrapposizione fra i dati e la funzione di regressione sarà verosimilmente buona, per cui

$$\begin{array}{c}\text{variazione spiegata}\\ \text{dalla regressione}\end{array} \;\gg\; \text{variazione residua}$$




Se $y$ **non dipende** da $x$, allora $y(x) =$ costante e il modello di regressione sarà ancora quello precedente, ma con $\alpha_2, \ldots, \alpha_p = 0$. Perciò

$$0 \;=\; \begin{array}{c}\text{variazione spiegata}\\ \text{dalla regressione}\end{array} \;\leq\; \text{variazione residua}$$



• Per $y$ indipendente da $x$ (modello di regressione corretto con $\alpha_2, \ldots, \alpha_p = 0$) si hanno le proprietà seguenti:

la variazione totale è una variabile di $\mathcal{X}^2$ a $n-1$ gradi di libertà

$$\mathcal{X}^2{}_{n-1} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \overline{y})^2$$

la variazione spiegata dalla regressione è una variabile di $\mathcal{X}^2$ a $p-1$ gradi di libertà

$$\mathcal{X}^2{}_{p-1} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[R(x_i) - \overline{y}\right]^2$$

variazione spiegata e variazione residua NSSAR sono variabili di $\mathcal{X}^2$ stocasticamente indipendenti

$$\Downarrow$$

$$F = \frac{\text{varianza spiegata}}{\text{varianza residua}} = \frac{\text{variazione spiegata}/(p-1)}{\text{variazione residua}/(n-p)} =$$

$$= \frac{n-p}{p-1}\,\frac{\mathcal{X}^2{}_{p-1}}{\mathcal{X}^2{}_{n-p}} = \frac{n-p}{p-1}\,\frac{\mathcal{X}^2{}_{p-1}}{\text{NSSAR}}$$

segue una distribuzione di Fisher a $(p-1, n-p)$ gradi di libertà

**È necessario che sia presente il parametro additivo!**
In caso contrario le variabili $\mathcal{X}^2{}_{n-1}$ e $\mathcal{X}^2{}_{p-1}$ non seguono le distribuzioni indicate, né di conseguenza la $F$

• si testa l'**ipotesi nulla**

$$H_0 \ : \ \text{modello di regressione corretto con } \alpha_2 = \ldots = \alpha_p = 0 \qquad \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j$$

ossia

$$H_0 \ : \ y \text{ indipendente da } x$$

contro l'**ipotesi alternativa**

$$H_1 \ : \ \text{modello di regressione corretto con } \alpha_2, \ldots, \alpha_p \text{ non tutti nulli} \qquad \alpha_1 + \sum_{j=2}^{p} \alpha_j \phi_j$$

Variabile del test

$$F = \frac{\text{varianza spiegata}}{\text{varianza residua}} = \frac{n-p}{p-1} \frac{\mathcal{X}^2{}_{p-1}}{\text{NSSAR}}$$

$$F \text{ grande} \iff \text{fit buono con } \alpha_2, \ldots, \alpha_p \textbf{ non tutti nulli}$$

$$F \text{ piccolo} \iff \text{fit buono con } \alpha_2, \ldots, \alpha_p \textbf{ tutti nulli}$$

Ipotesi $H_0$ rigettata se

$$F > F_{[1-\alpha](p-1,n-p)}$$

con livello di significatività $\alpha \in (0, 1)$

## 10.4 $F$-test con osservazioni ripetute in $y$

Si suppone che per ogni $x_i$, $i = 1, \ldots, n$ siano disponibili più valori osservati della $y$ corrispondente

$$y_{ik} \qquad k = 1, \ldots, n_i$$

per un numero totale di osservazioni pari a

$$N = \sum_{i=1}^{n} n_i$$

Questi valori si intendono come le realizzazioni di variabili casuali gaussiane indipendenti di media $\mu_i$ e varianza $\sigma_i^2$

Le ipotesi della regressione sono verificate: in più si assume che le $n_i$ osservazioni di $y$ a $x = x_i$ assegnato siano **indipendenti**

Il modello di regressione (da testare) è quello consueto

$$\zeta_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) + \varepsilon_i \qquad \forall\, i = 1, \ldots, n$$

con parametri di regressione $\alpha_1, \ldots, \alpha_p$ e $\varepsilon_i$ variabile $N(0, \sigma_i)$

Se il modello è corretto, il valor medio di $\zeta_i$ vale

$$\mu_i = \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \qquad \forall\, i = 1, \ldots, n$$

Si distinguono due casi:

varianze $\sigma_i^2$ note

varianze $\sigma_i^2$ sconosciute ma eguali ($\sigma_i^2 = \sigma^2\ \forall\, i$)

### 10.4.1 Varianze $\sigma_i^2$ note

Si calcolano due somme normalizzate di quadrati di scarti

$$\mathrm{NSS}_{\mathrm{m.i.}} = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \frac{1}{\sigma_i^2} (y_{ik} - \overline{y}_i)^2$$

$$\mathrm{NSS}_{\mathrm{m.d.}} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} n_i \left[ \overline{y}_i - \sum_{j=1}^{p} a_j \phi_j(x_i) \right]^2$$

dove $\overline{y}_i$ è la media delle $y_{ik}$ a $x = x_i$ assegnato

$$\overline{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$$

e le $a_1, \ldots, a_p$ sono le stime di chi-quadrato dei parametri $\alpha_1$, $\ldots, \alpha_p$, ricavate minimizzando il funzionale

$$\mathcal{L}(\alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} n_i \left[ \overline{y}_i - \sum_{j=1}^{p} \alpha_j \phi_j(x_i) \right]^2$$

$\mathrm{NSS}_{\mathrm{m.i.}}$ descrive la variabilità dei dati in modo **indipendente dal modello di regressione**

$\mathrm{NSS}_{\mathrm{m.d.}}$ descrive variabilità dei dati **attorno al modello di regressione**: esprime la variabilità **"non spiegata"** dalla regressione

Euristicamente

$$\text{modello di regressione non corretto} \iff \mathrm{NSS}_{\mathrm{m.d.}} \gg \mathrm{NSS}_{\mathrm{m.i.}}$$

Più precisamente:

$\mathrm{NSS}_{\mathrm{m.i.}}$ è una variabile di chi-quadrato a $N-n$ gradi di libertà, **a prescindere dalla correttezza del modello di regressione**

$\mathrm{NSS}_{\mathrm{m.d.}}$ risulta una variabile di chi-quadrato a $n-p$ gradi di libertà **se il modello di regressione è corretto**

in tal caso $\mathrm{NSS}_{\mathrm{m.i.}}$ e $\mathrm{NSS}_{\mathrm{m.d.}}$ sono **stocasticamente indipendenti**

Se il modello è corretto, il rapporto delle variabili di chi-quadrato ridotte

$$\frac{\mathrm{NSS}_{\mathrm{m.d.}}/(n-p)}{\mathrm{NSS}_{\mathrm{m.i.}}/(N-n)} = \frac{N-n}{n-p}\,\frac{\mathrm{NSS}_{\mathrm{m.d.}}}{\mathrm{NSS}_{\mathrm{m.i.}}}$$

segue una distribuzione di Fisher a $(n-p, N-n)$ gradi di libertà

L'ipotesi

$$H_0 \;:\; \text{il modello di regressione è corretto}$$

verrà dunque rigettata per

$$\frac{N-n}{n-p}\,\frac{\mathrm{NSS}_{\mathrm{m.d.}}}{\mathrm{NSS}_{\mathrm{m.i.}}} > F_{[1-\alpha](n-p,N-n)}$$

con livello di significatività $\alpha \in (0,1)$

### 10.4.2 Varianze eguali (ancorché sconosciute)

Se le $\sigma_i^2$ assumono un comune valore $\sigma^2$, la discussione precedente rimane comunque valida per cui **si rigetta** il modello di regressione se

$$\frac{N-n}{n-p}\frac{\mathrm{NSS}_{\mathrm{m.d.}}}{\mathrm{NSS}_{\mathrm{m.i.}}} > F_{[1-\alpha](n-p,N-n)}$$

Il questo caso il quoziente delle variabili di chi-quadrato ridotte è **indipendente** da $\sigma^2$, che potrebbe anche essere **sconosciuto**

Basta scrivere

$$\frac{N-n}{n-p}\frac{\mathrm{NSS}_{\mathrm{m.d.}}}{\mathrm{NSS}_{\mathrm{m.i.}}} = \frac{N-n}{n-p}\frac{\mathrm{SS}_{\mathrm{m.d.}}}{\mathrm{SS}_{\mathrm{m.i.}}}$$

in termini delle somme

$$\mathrm{SS}_{\mathrm{m.i.}} = \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-\overline{y}_i)^2$$

$$\mathrm{SS}_{\mathrm{m.d.}} = \sum_{i=1}^{n} n_i\left[\overline{y}_i - \sum_{j=1}^{p} a_j\phi_j(x_i)\right]^2$$

**che non contengono** $\sigma^2$

Una stima di $\sigma^2$ è data da

$$\frac{\mathrm{SS}_{\mathrm{m.i.}}}{N-n} = \frac{1}{N-n}\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-\overline{y}_i)^2$$

ovvero, se il modello di regressione è corretto, da

$$\frac{\mathrm{SS}_{\mathrm{m.d.}}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} n_i\left[\overline{y}_i - \sum_{j=1}^{p} a_j\phi_j(x_i)\right]^2$$

## 10.5 $F$-test sul parametro addizionale (incremental $F$-test)

Per un assegnato set $\{(x_i, y_i), 1 \leq i \leq n\}$ di dati, si considerino **due** modelli di regressione lineare che differiscano l'uno dall'altro per un parametro aggiuntivo

$$\sum_{j=1}^{p} \alpha_j \phi_j(x) \qquad\qquad \sum_{j=1}^{p+1} \alpha_j \phi_j(x)$$

Se il primo modello è corretto, tale si può considerare anche il secondo (basta assumere $\alpha_{p+1} = 0$)

In tal caso la procedura di best-fit sulle somme pesate dei quadrati degli scarti conduce alle variabili di $\mathcal{X}^2$

$$\mathrm{NSSAR}_p \qquad \text{e} \qquad \mathrm{NSSAR}_{p+1}$$

rispettivamente a $n - p$ e $n - p - 1$ gradi di libertà

È evidente che comunque siano assegnati i dati $(x_i, y_i)$ risulta

$$\mathrm{NSSAR}_p \geq \mathrm{NSSAR}_{p+1}$$

Se ne può dedurre che la differenza

$$\mathrm{NSSAR}_p - \mathrm{NSSAR}_{p+1}$$

costituisce una variabile di $\mathcal{X}^2$ a 1 grado di libertà, **stocasticamente indipendente** da $\mathrm{NSSAR}_{p+1}$

Ciò consente l'introduzione della variabile di Fisher

$$F_A = \frac{\mathrm{NSSAR}_p - \mathrm{NSSAR}_{p+1}}{\mathrm{NSSAR}_{p+1}/(n - p - 1)}$$

a 1, $n - p - 1$ gradi di libertà

Il quoziente $F_A$ misura quanto il termine addizionale nel modello ha migliorato il chi quadrato ridotto del modello stesso

$F_A$ sarà **piccolo** se il modello a $p+1$ parametri non migliora sensibilmente il fitting offerto dal modello a $p$ parametri

Viceversa, per valori **grandi** di $F_A$ si dovrà ritenere che l'introduzione del parametro addizionale abbia migliorato in modo significativo il fitting e che quindi il modello a $p+1$ parametri **debba essere preferito** a quello con $p$ soli parametri

Perciò:

$$F_A \text{ piccolo} \iff \text{modello a } p \text{ parametri preferibile}$$

$$F_A \text{ grande} \iff \text{modello a } p+1 \text{ parametri più adeguato}$$

In definitiva:

se $F_A \leq F_{[1-\alpha](1,n-p-1)}$ **si accetta** l'ipotesi $H_0$ che il modello a $p$ parametri sia corretto;

se $F_A > F_{[1-\alpha](1,n-p-1)}$ **si rigetta** tale ipotesi, accogliendo il modello a $p+1$ parametri con $\alpha_{p+1}$ significativamente diverso da zero

## 10.6 Caso notevole. Retta di regressione

Il modello lineare è a due soli parametri e si riduce a

$$\zeta_i \;=\; \alpha + \beta x_i + \varepsilon_i$$

con parametri di regressione $\alpha$ e $\beta$

Le corrispondenti stime di chi quadrato $a$ e $b$ minimizzano la

$$\text{NSSAR} \;=\; \sum_{i=1}^{n} \Big( \frac{a + b x_i - y_i}{\sigma_i} \Big)^2$$

variabile di $\mathcal{X}^2$ a $n-2$ gradi di libertà, indipendente da $a$ e $b$

Le equazioni normali

$$\begin{cases} aS + bS_x \;=\; S_y \\ aS_x + bS_{xx} \;=\; S_{xy} \end{cases}$$

hanno la soluzione

$$a \;=\; \frac{S_{xx} S_y - S_x S_{xy}}{\Delta} \qquad b \;=\; \frac{S S_{xy} - S_x S_y}{\Delta}$$

con matrice di covarianza e coefficiente di correlazione

$$C \;=\; \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix} \qquad \text{corr}(a,b) \;=\; -\frac{S_x}{\sqrt{S S_{xx}}}$$

dove:

$$S \;=\; \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \qquad S_x \;=\; \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \qquad S_y \;=\; \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \;=\; \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} \;=\; \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} \qquad \Delta \;=\; S S_{xx} - S_x^2 > 0$$

**N.B.** Le variabili casuali $a$ e $b$ **non sono indipendenti**!

**Retta di regressione. Modello alternativo**

Si considera il modello a due parametri

$$\zeta_i \;=\; \mu + \kappa(x_i - \overline{x}) + \varepsilon_i$$

in cui

$$\overline{x} \;=\; \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big)^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i$$

mentre $\mu$ e $\kappa$ sono i parametri di regressione

Le relative stime di chi quadrato $m$ e $q$ minimizzano la

$$\text{NSSAR} \;=\; \sum_{i=1}^{n} \Big( \frac{m + q(x_i - \overline{x}) - y_i}{\sigma_i} \Big)^2$$

variabile di $\mathcal{X}^2$ a $n-2$ gradi di libertà indipendente da $m$ e $q$

Le equazioni normali porgono la soluzione

$$m \;=\; \frac{\displaystyle\sum_{i=1}^{n} \frac{1}{\sigma_i^2} y_i}{\displaystyle\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \qquad q \;=\; \frac{\displaystyle\sum_{i=1}^{n} \frac{(x_i - \overline{x}) y_i}{\sigma_i^2}}{\displaystyle\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_i^2}}$$

con matrice di covarianza diagonale

$$C \;=\; \begin{pmatrix} \Big(\displaystyle\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big)^{-1} & 0 \\ 0 & \Big(\displaystyle\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_i^2}\Big)^{-1} \end{pmatrix}$$

e dunque coefficiente di correlazione **nullo**.

**Le variabili casuali** $m$ **e** $q$ **sono indipendenti**! ($a$ e $b$ no)

**Retta di regressione**
**Intervalli di confidenza per i parametri $\alpha$ e $\beta$**

Dato il modello di regressione lineare a due parametri

$$\zeta_i \;=\; \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

gli intervalli di confidenza per le stime $a$ e $b$ dei parametri $\alpha$ e $\beta$ sono dati da

$$\alpha \;=\; a \pm t_{[1-\frac{\eta}{2}](n-2)} \sqrt{\frac{S_{xx}}{\Delta} \frac{\mathrm{NSSAR}}{n-2}}$$

$$\beta \;=\; b \pm t_{[1-\frac{\eta}{2}](n-2)} \sqrt{\frac{S}{\Delta} \frac{\mathrm{NSSAR}}{n-2}}$$

con livello di confidenza $1 - \eta \in (0, 1)$.

Vale sempre

$$a \;=\; \frac{S_{xx} S_y - S_x S_{xy}}{\Delta} \qquad b \;=\; \frac{S S_{xy} - S_x S_y}{\Delta}$$

Nel caso di eguali varianze $\sigma_i^2 = \sigma^2$ gli intervalli di confidenza sono indipendenti da $\sigma$.

**N.B.**
La probabilità che **simultaneamente** $\alpha$ e $\beta$ appartengano ai relativi intervalli di confidenza **non è** $(1-\eta)^2 = (1-\eta)(1-\eta)$
Le variabili $a$ e $b$ **non sono indipendenti!**

## Retta di regressione
## Intervalli di confidenza per i parametri $\mu$ e $\kappa$

Dato il modello di regressione lineare a due parametri

$$\zeta_i \;=\; \mu + \kappa(x_i - \overline{x}) + \varepsilon_i \qquad i = 1, \ldots, n$$

gli intervalli di confidenza per le stime $m$ e $q$ dei parametri $\mu$ e $\kappa$ si scrivono

$$\mu \;=\; m \pm t_{[1-\frac{\eta}{2}](n-2)} \sqrt{\Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big)^{-1} \frac{\mathrm{NSSAR}}{n-2}}$$

$$\kappa \;=\; q \pm t_{[1-\frac{\eta}{2}](n-2)} \sqrt{\Big(\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_i^2}\Big)^{-1} \frac{\mathrm{NSSAR}}{n-2}}$$

con livello di confidenza $1 - \eta \in (0, 1)$.

Nel caso di eguali varianze $\sigma_i^2 = \sigma^2$ gli intervalli di confidenza sono indipendenti da $\sigma$.

**N.B.**
La probabilità che **simultaneamente** $\mu$ e $\kappa$ appartengano ai relativi intervalli di confidenza **è** $(1-\eta)^2 = (1-\eta)(1-\eta)$
Le variabili $m$ e $q$ **sono indipendenti!**

## Retta di regressione
## Intervallo di confidenza per le previsioni

Si considera il modello di regressione lineare

$$\zeta_i \;=\; \mu + \kappa(x_i - \overline{x}) + \varepsilon_i$$

Si assume che il valore della variabile $y$ corrispondente ad una ascissa $x = x_0$ sia descritto dalla variabile casuale

$$\zeta_0 \;=\; \mu + \kappa(x_0 - \overline{x}) + \varepsilon_0$$

con $\varepsilon_0$ variabile gaussiana di media nulla e varianza $\sigma_0^2$, indipendente dalle $\varepsilon_i$.

Si definisce la **previsione** del valore di $y$ in $x_0$ in base alla regressione lineare come

$$y_0 \;=\; m + q(x_0 - \overline{x})$$

La previsione $y_0$ è gaussiana di media $\zeta_0$ e varianza

$$\mathbb{E}[(y_0 - \zeta_0)^2] \;=\; \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\Big)^{-1} + \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(x_i - \overline{x})^2\Big)^{-1} (x_0 - \overline{x})^2 + \sigma_0^2$$

L'intervallo di confidenza di $\zeta_0$ assume la forma

$$\zeta_0 \;=\; y_0 \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\mathbb{E}[(y_0 - \zeta_0)^2]} \sqrt{\frac{\mathrm{NSSAR}}{n-2}}$$

con livello di confidenza $1 - \alpha \in (0, 1)$

Se la varianza è indipendente dall'ascissa

$$\sigma_i = \sigma \qquad \forall\, i = 1, \ldots, n \qquad \text{e} \qquad \sigma_0 = \sigma$$

il precedente intervallo di confidenza è pure indipendente da $\sigma$

$$\zeta_0 \;=\; m + q(x_0 - \overline{x}) \pm t_{[1-\frac{\alpha}{2}],(n-2)}\, \sqrt{V}\, \sqrt{\frac{SS}{n-2}}$$

con

$$\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad SS \;=\; \sum_{i=1}^{n} \big[-y_i + m + q(x_i - \overline{x})\big]^2$$

$$V \;=\; \frac{1}{n} + \frac{1}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}(x_0 - \overline{x})^2 + 1\,.$$

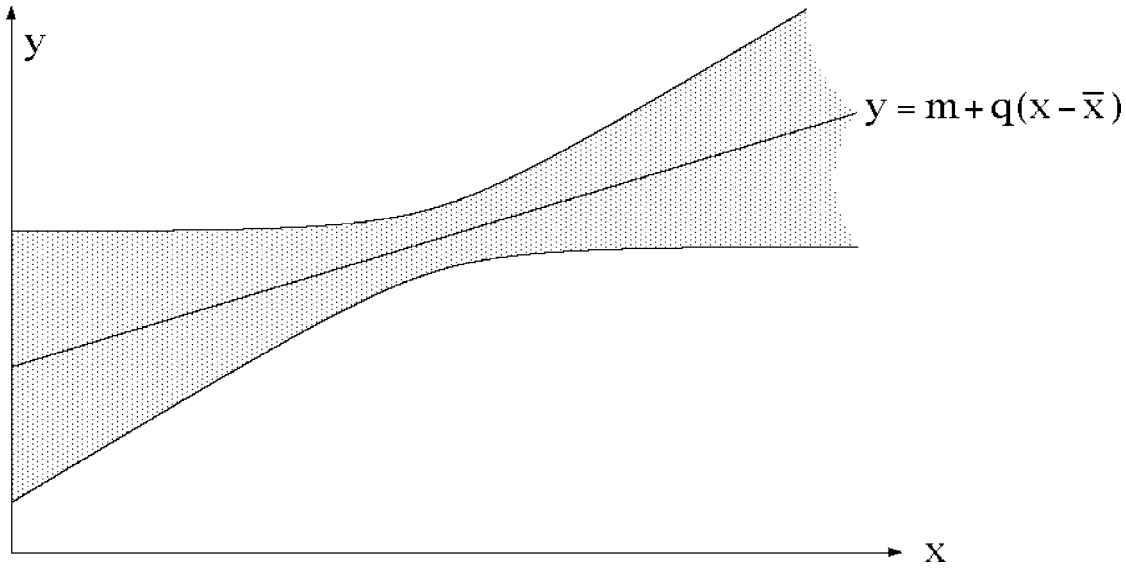
**Osservazione**
Il coefficiente $V$ diminuisce al crescere della dispersione delle $x_i$ attorno al loro valore medio $\overline{x}$

**Osservazione**
Al variare di $x_0 \in \mathbb{R}$ l'intervallo di confidenza descrive una regione del piano $(x, y)$ del tipo

$$\big\{(x,y) \in \mathbb{R}^2 \;:\; |m + q(x - \overline{x}) - y| \le c\sqrt{1 + g(x - \overline{x})^2}\big\}$$

con $c$ e $g$ costanti positive opportune.

Questa **regione di confidenza** costituisce un intorno della retta di regressione e il suo aspetto qualitativo è illustrato col tratteggio nella figura seguente




Per valori di $x_i$ molto dispersi attorno alla media $\overline{x}$ la regione di confidenza si restringe

**Esempio**: **Retta di regressione**

Alcune misure sperimentali di una grandezza $y$ contro un'altra grandezza $x$ sono raccolte nella tabella seguente

| $k$ | $x_k$ | $y_{k1}$ | $y_{k2}$ | $y_{k3}$ | $y_{k4}$ |
|---|---|---|---|---|---|
| 1 | 0.6 | 0.55 | 0.80 | 0.75 | $\times$ |
| 2 | 1.0 | 1.10 | 0.95 | 1.05 | $\times$ |
| 3 | 1.5 | 1.40 | 1.30 | 1.50 | $\times$ |
| 4 | 2.0 | 1.60 | 1.90 | 1.85 | 1.95 |
| 5 | 2.5 | 2.25 | 2.30 | 2.40 | 2.45 |
| 6 | 2.9 | 2.65 | 2.70 | 2.75 | $\times$ |
| 7 | 3.0 | 2.65 | 2.70 | 2.80 | 2.85 |

L'errore casuale sulle ascisse $x_k$ è trascurabile, mentre i dati $y_k$ sono variabili casuali indipendenti con la stessa varianza $\sigma^2$ (sistema omoscedastico).

Determinare:

($i$) la retta di regressione, secondo il metodo dei minimi quadrati, nella forma

$$y \,=\, \mu + \kappa(x - \overline{x})\,;$$

($ii$) gli intervalli di confidenza al 95% del parametro $\mu$ e del coefficiente angolare $\kappa$;

($iii$) la regione di confidenza al 95% per le previsioni;

($iv$) l'intervallo di confidenza al 95% per il valore di $y$ predetto a $x = 2.3$;

($v$) la goodness of fit del modello di regressione qualora sia $\sigma = 0.08$ la comune deviazione standard di tutti i dati $y$.

**Soluzione**

($i$) Poichè le deviazioni standard sono uguali, il fitting di chi-quadrato si riduce all'usuale fitting con i minimi quadrati e le stime di best-fit dei parametri $m$ e $q$ possono scriversi:

$$m = \frac{1}{n}\sum_{i=1}^{n} y_i = 1.8833 \quad q = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.87046$$

con $n = 24$ e $\overline{x} = 2.000$. La retta di regressione, calcolata con il metodo dei minimi quadrati, diventa pertanto:

$$\begin{aligned} y = m + q(x - \overline{x}) &= 1.8833 + 0.87046(x - 2.000) = \\ &= 0.1424 + 0.87046\, x \end{aligned}$$

($ii$) La somma dei quadrati attorno alla regressione vale:

$$\mathrm{SSAR} = \sum_{i=1}^{n}[m + q(x_i - \overline{x}) - y_i]^2 = 0.22704376$$

Al livello di confidenza $1 - \alpha \in (0, 1)$ l'intervallo di confidenza del parametro $\mu$ e quello del coefficiente angolare $\kappa$ assumono la forma:

$$\mu = m \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\frac{1}{n}\frac{\mathrm{SSAR}}{n-2}}$$

$$\kappa = q \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{-1}\frac{\mathrm{SSAR}}{n-2}}$$

Nel caso in esame si ha $\alpha = 0.05$, $n = 24$ e gli intervalli di confidenza diventano quindi:

$$\mu = m \pm t_{[0.975](22)} \sqrt{\frac{1}{24} \frac{\text{SSAR}}{22}}$$

$$\kappa = q \pm t_{[0.975](22)} \sqrt{\left[\sum_{i=1}^{24} (x_i - \overline{x})^2\right]^{-1} \frac{\text{SSAR}}{22}}$$

con:

$$\begin{aligned} m &= 1.8833 \\ q &= 0.87046 \\ \text{SSAR} &= 0.22704376 \\ \sum_{i=1}^{24} (x_i - \overline{x})^2 &= 17.060000 \\ t_{[0.975](22)} &= 2.074 \end{aligned}$$

Di conseguenza:

- l'intervallo di confidenza al 95% del parametro $\mu$ è

$$1.8833 \pm 0.0430 = [1.840, 1.926]$$

- il CI al 95% per il coefficiente angolare $\kappa$ vale

$$0.87046 \pm 0.05101 = [0.819, 0.921]$$

(*iii*) Poichè il modello si assume omoscedastico, il CI (al livello di confidenza $1-\alpha$) per la previsione di $y = y_0$ at una ascissa $x = x_0$ assegnata può essere calcolato mediante la formula generale:

$$\mathbb{E}(y_0) \;=\; m + q(x_0 - \overline{x}) \pm t_{[1-\frac{\alpha}{2}](n-2)}\, \sqrt{V}\, \sqrt{\frac{\mathrm{SSAR}}{n-2}}$$

dove:

$$\begin{aligned}
\overline{x} \;&=\; \frac{1}{n}\sum_{i=1}^{n} x_i \\
\mathrm{SSAR} \;&=\; \sum_{i=1}^{n} \big[-y_i + m + q(x_i - \overline{x})\big]^2 \\
V \;&=\; 1 + \frac{1}{n} + \frac{1}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}(x_0 - \overline{x})^2 \,.
\end{aligned}$$

Nella fattispecie si ha $\overline{x} = 2.000$, per cui:

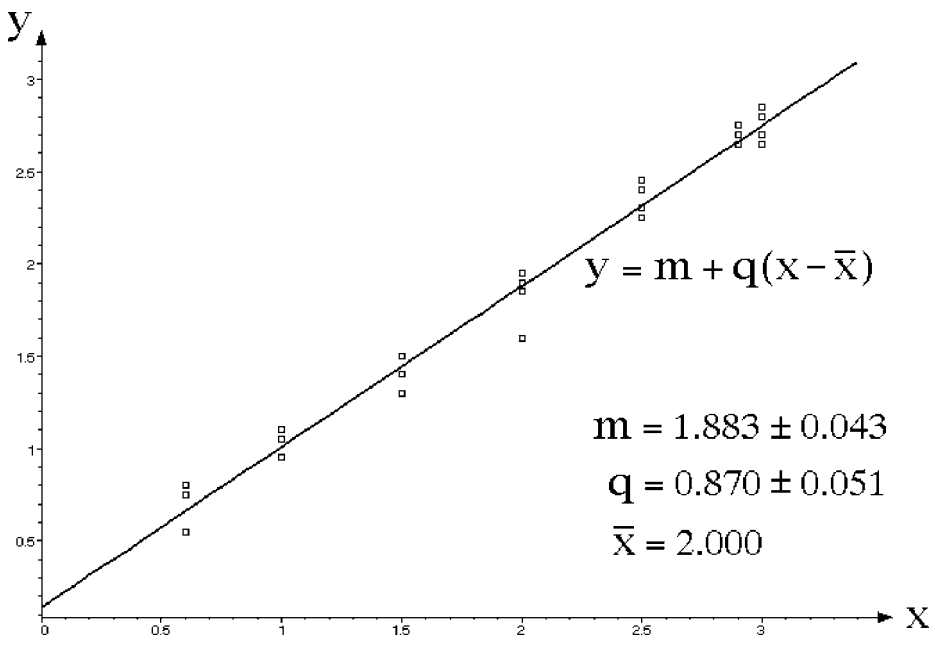
$$y_0 \;=\; 1.8833 + 0.87046(x_0 - 2.000) \pm t_{[0.975](22)} \cdot$$

$$\cdot \sqrt{1 + \frac{1}{24} + \frac{1}{17.060000}(x_0 - 2.000)^2}\sqrt{\frac{0.22704376}{22}}$$

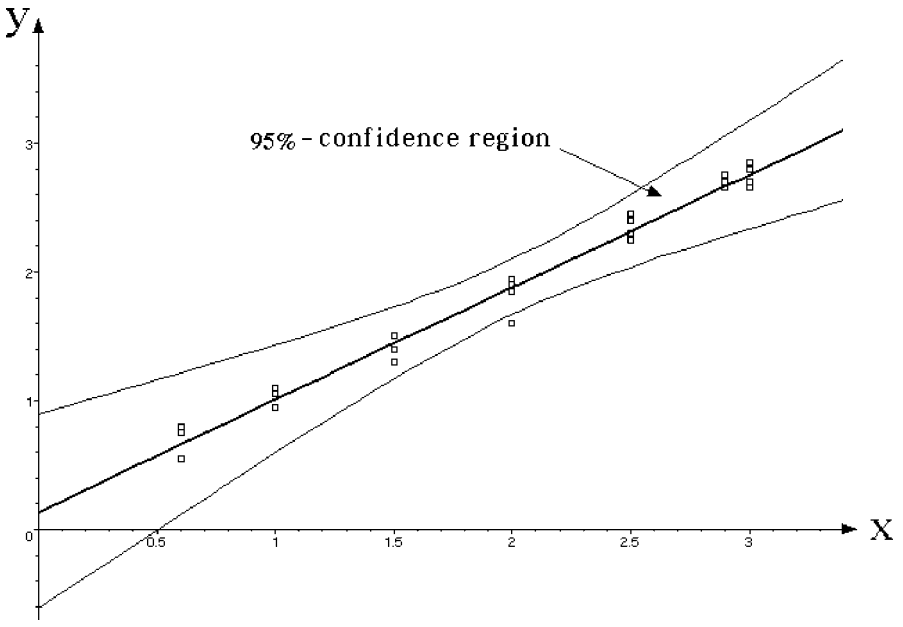
e l'intervallo di confidenza per la previsione di $y$ a $x = x_0$ si riduce a:

$$\begin{aligned}
y_0 \;&=\; 1.8833 + 0.87046(x_0 - 2.000) \pm \\
&\pm\; 0.21069\sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}
\end{aligned}$$

Nella figura seguente la retta di regressione è sovrapposta ai punti sperimentali (dots):




La regione di confidenza per le previsioni (al livello di confidenza del 95%) è evidenziato in figura (esagerando il fattore $V$ per maggiore chiarezza)



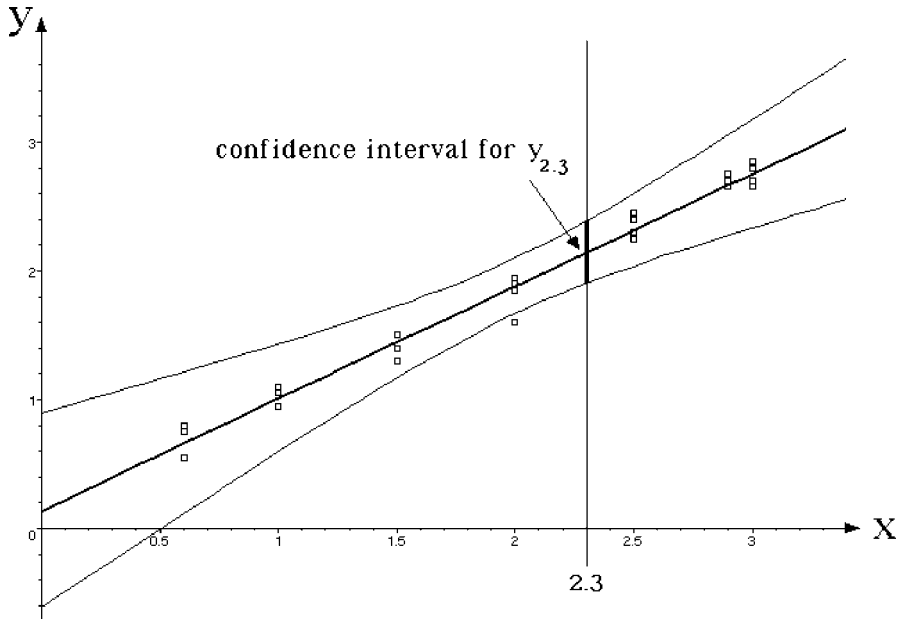
(*iv*) l'intervallo di confidenza al 95% per la previsione di $y$ in $x = 2.3$ può essere ottenuto sostituendo $x_0 = 2.3$ nella formula precedente

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm \\ \pm 0.21069\sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}$$

Si ha perciò:

$$y_{2.3} = [1.903, 2.385] = 2.144 \pm 0.241$$

Nel grafico seguente l'intervallo di confidenza al 95% è l'intersezione della regione di confidenza al 95% con la linea verticale di equazione $x = 2.3$:



(*v*) la goodness of fit $Q$ del modello di regressione è definita dalla relazione

$$Q = \int_{\text{NSSAR}}^{+\infty} \rho_{n-p}(\mathcal{X}^2)\, d\mathcal{X}^2$$

dove $\rho_{n-p}$ indica la distribuzione di $\mathcal{X}^2$ a $n-p$ g.d.l.

Questo perchè, le il modello di regressione è corretto, la somma normalizzata dei quadrati degli scarti attorno alla regressione

$$\text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma^2}[m + q(x_i - \overline{x}) - y_i]^2 = \frac{\text{SSAR}}{\sigma^2}$$

è una variabile di $\mathcal{X}^2$ a $n-p$ g.d.l.

Per eseguire i calcoli **è cruciale conoscere il comune valore della deviazione standard** $\sigma = 0.08$, poichè occorre determinare la NSSAR, e non semplicemente la SSAR.

Nel caso in esame si hanno $n = 24$ dati e il modello di regressione è basato su due parametri, $\mu$ e $\kappa$; di conseguenza, $p = 2$ e la NSSAR segue una distribuzione di $\mathcal{X}^2$ a $n - p = 22$ g.d.l.

Per il campione assegnato la somma normalizzata dei quadrati degli scarti vale

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{0.22704376}{0.08^2} = 35.47558$$

Dalla tavola dei valori critici superiori di $\mathcal{X}^2$ con $\nu = 22$ g.d.l. si trova

| Probabilità $\{\mathcal{X}^2 \geq 33.924\}$ | Probabilità $\{\mathcal{X}^2 \geq 36.781\}$ |
|---|---|
| 0.05 | 0.025 |

in modo che un semplice schema di interpolazione lineare:

| | |
|---|---|
| 33.924 | 0.05 |
| 35.476 | $Q$ |
| 36.781 | 0.025 |

$$\frac{35.476 - 33.924}{36.781 - 33.924} = \frac{Q - 0.05}{0.025 - 0.05}$$

fornisce la stima richiesta di $Q$:

$$Q = 0.05 + (0.025 - 0.05)\frac{35.476 - 33.924}{36.781 - 33.924} = 0.0364$$

Un valore più accurato di $Q$ segue da una integrazione numerica

$$Q = \text{Probabilità}\{\mathcal{X}^2 \geq 35.47558\} = \int_{35.47558}^{+\infty} p_{22}(\mathcal{X}^2)\, d\mathcal{X}^2$$

ad esempio usando l'istruzione Maple

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[22]](35.475587);$$

che conduce a $Q = 0.0345$.

Se il modello di regressione fosse rigettato, $Q$ esprimerebbe la probabilità di errore di prima specie — circa 3.5% nel caso considerato.

## 10.7 Chi-square fitting con la SVD (Singular Value Decomposition)

Il metodo del chi quadrato consiste nel determinare il minimo del funzionale quadratico (3)

$$\mathcal{L}(\alpha_1, \ldots, \alpha_p) \;=\; \sum_{i=1}^{n} \left[ -\frac{1}{\sigma_i} y_i + \sum_{j=1}^{p} \alpha_j \frac{1}{\sigma_i} \phi_j(x_i) \right]^2$$

che in termini delle matrici ausiliarie già introdotte

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} \quad \sigma^{-1} = \begin{pmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \mathbb{O} \\ & & \ddots & \\ & \mathbb{O} & & 1/\sigma_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

e

$$\Phi^T_{ih} \;=\; \Phi_{hi} \;=\; \frac{\phi_h(x_i)}{\sigma_i} \qquad 1 \le i \le n\,,\; 1 \le h \le p$$

assume la forma

$$\mathcal{L}(\alpha) \;=\; |\sigma^{-1} y - \Phi^T \alpha|_2^2$$

Si tratta perciò di individuare il vettore $a \in \mathbb{R}^n$ tale che

$$|\sigma^{-1} y - \Phi^T a|_2 \;=\; \min_{\alpha \in \mathbb{R}^n} |\sigma^{-1} y - \Phi^T \alpha|_2$$

ossia la soluzione nel senso dei minimi quadrati del sistema lineare

$$\sigma^{-1} y \;=\; \Phi^T a$$

Il problema può essere risolto, oltre che con l'equazione normale (7), mediante la **Singular Value Decomposition (SVD)** della matrice $\Phi^T$

**Singular Value Decomposition di una matrice** $m \times n$

La SVD di una matrice $A$ è una riscrittura di $A$ come prodotto

$$\begin{array}{ccccc} A & = & V & \begin{pmatrix} \Sigma & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} & U^T \\ m \times n & & m \times m & m \times n & n \times n \end{array}$$

dove:

- $U$ è una matrice $n \times n$ le cui colonne $u_1, \ldots, u_n$ costituiscono una base ortonormale di autovettori della matrice $n \times n$, simmetrica e semidefinita positiva, $A^T A$

$$U = (u_1 \ \ldots \ u_n) \qquad A^T A u_i = \sigma_i^2 u_i \quad \forall\, i = 1, \ldots, n$$
$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2 > \sigma_{r+1}^2 = \ldots = \sigma_n^2 = 0$$

- $\Sigma$ è una matrice diagonale $r \times r$, con $r \leq \min(m, n)$,

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{pmatrix}$$

  i cui elementi diagonali sono le radici quadrate aritmetiche degli autovalori positivi di $A^T A$ (c.d. **valori singolari** della matrice $A$)

- $V$ è una matrice $m \times m$ le cui prime $r$ colonne sono date dai vettori ortonormali

$$v_i = \frac{1}{|Au_i|_2} Au_i \qquad \forall i = 1, \ldots, r$$

  mentre le eventuali $m - r$ colonne residue si scelgono in modo da generare una base ortonormale di $\mathbb{R}^m$

**N.B.: Ogni matrice $m \times n$ ammette una SVD!**

La SVD di $A$

$$A = V \begin{pmatrix} \Sigma & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} U^T$$

consente di definire la c.d. **pseudo inversa** di $A$ (Moore-Pensore pseudo inverse)

$$A^+ = U \begin{pmatrix} \Sigma^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} V^T$$

**N.B.:** $U^{-1} = U^T$ e $V^{-1} = V^T$, le matrici $U$ e $V$ sono per costruzione ortogonali

Se $A$ è invertibile

$$A^+ = A^{-1}$$

In caso contrario, la $A^+$ risulta **comunque definita**

**Applicazione**

Dato il sistema lineare

$$Ax = b$$

il vettore

$$x^+ = A^+ b$$

ne costituisce la soluzione **nel senso dei minimi quadrati**

$$|Ax^+ - b|_2^2 = \min_{x \in \mathbb{R}^n} |Ax - b|_2^2$$

Per $A$ invertibile, $x^+$ è la soluzione in senso consueto

**SVD e matrice di covarianza dei parametri di best-fit**

Anche la matrice $\Phi_{ki} = \phi_k(x_i)/\sigma_i$, $k = 1, \ldots, p$, $i = 1, \ldots, n$, ammette una SVD

$$\begin{array}{ccccc} \Phi & = & V & \begin{pmatrix} \Sigma & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} & U^T \\ p \times n & & p \times p & p \times n & n \times n \end{array}$$

e la matrice di struttura dei parametri di best-fit — variabile normale multivariata — si scrive

$$F = \Phi\Phi^T = V \begin{pmatrix} \Sigma^2 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} V^T$$

dove le matrici a secondo membro sono tutte $p \times p$

Se $\Phi$ ha rango massimo — $p$ — vale

$$\begin{pmatrix} \Sigma^2 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} = \Sigma^2$$

con $\Sigma^2$ di elementi diagonali tutti positivi

Perciò la matrice di covarianza dei parametri di best-fit risulta

$$F^{-1} = V(\Sigma^2)^{-1}V^T$$

## 10.8 Modelli non lineari

Sono modelli in cui la dipendenza dai parametri di fitting è **non lineare**

Diversamente dai modelli lineari, la determinazione dei parametri di best-fit **non può essere ottenuta analiticamente**, mediante la soluzione di un sistema di **equazioni algebriche lineari** (equazioni normali)

I parametri di best-fit vanno calcolati mediante un appropriato algoritmo numerico di ottimizzazione (cioè di ricerca di minimi) L'algoritmo è di tipo iterativo

Algoritmi di ottimizzazione (ricerca di minimi) di uso corrente:

– steepest descent (metodo del gradiente)

– inverse Hessian method (metodo dell'hessiana inversa)

– metodo di Levenberg-Marquardt

– metodo del gradiente coniugato

– metodo di rilassamento (alla Barbasin-Krasovskii)

Per la distribuzione di probabilià della NSSAR e quella congiunta dei parametri di best-fit $a_1, \ldots, p$ valgono le stesse conclusioni tratte nel caso dei modelli lineari **a patto che le** $\sigma_i$ **siano piccole** e che di conseguenza i parametri di best-fit varino di poco

Se le distribuzioni di probabilità delle $y_i$ sono non gaussiane e ben conosciute, i risultati precedenti non sono più applicabili e conviene in genere ricorrere ai metodi di **Monte-Carlo**

Come già sottolineato, il passaggio traumatico non si verifica tanto fra i modelli lineari e quelli non lineari, quanto fra il caso delle variabili gaussiane e quello delle $y_i$ non normali

Questa constatazione conduce al concetto di **"robust fitting"**

### 10.8.1 Robust fitting

Genericamente, una stima di parametri o un fittaggio si dicono "robusti" se risultano poco sensibili a "piccole" deviazioni dalle ipotesi in base alle quali stima o fittaggio sono ottimizzati

Queste "piccole deviazioni" possono essere veramente "piccole" e riguardare **tutti** i dati

Oppure possono risultare anche "grandi" ma limitatamente ad una piccola frazione dei dati utilizzati per la stima o il fittaggio

Esistono varie tipologie di stime "robuste"

Le più importanti sono le c.d. $M$**-stime locali**, basate sul principio di massima verosimiglianza (maximum likelihood)

Queste costituiscono il metodo di valutazione dei parametri di best-fit nei "fittaggi robusti" (robust fitting)

### Stima dei parametri di best-fit di un modello mediante $M$-stime locali

Se in una serie di misure gli errori casuali non sono gaussiani, il metodo di massima verosimiglianza per la stima dei parametri $\alpha$ di un modello $y(x, \alpha)$ richiede che si massimizzi la quantità

$$P = \prod_{i=1}^{n} e^{-\rho[y_i, y(x_i, \alpha)]}$$

dove la funzione $\rho$ è il logaritmo naturale cambiato di segno della distribuzione di probabilità della variabile $y_i$, con media $y(x_i, \alpha)$

Ciò equivale a minimizzare la sommatoria

$$\sum_{i=1}^{n} \rho[y_i, y(x_i, \alpha)]$$

Spesso la funzione $\rho[y_i, y(x_i, \alpha)]$ dipende solo dalla differenza $y_i - y(x_i, \alpha)$, eventualmente scalata di un opportuno fattore $\sigma_i$ assegnabile ad ogni punto sperimentale.

In tal caso la $M$-stima si dice **locale**

La condizione di massima verosimiglianza diventa allora

$$\min_{\alpha \in \mathbb{R}^p} \sum_{i=1}^{n} \rho\Big(\frac{y_i - y(x_i, \alpha)}{\sigma_i}\Big)$$

Il minimo in $\alpha$ (stima di best-fit $a$ dei parametri $\alpha$) può essere ricercato direttamente (procedura più comune)

oppure può essere caratterizzato come punto critico, soluzione dell'equazione

$$\sum_{i=1}^{n} \frac{1}{\sigma_i} \psi\Big(\frac{y_i - y(x_i, \alpha)}{\sigma_i}\Big) \frac{\partial y(x_i, \alpha)}{\partial a_j} \;=\; 0 \qquad j = 1, \ldots, p$$

in cui si ponga

$$\psi(z) \;=\; \frac{d\rho(z)}{dz} \qquad \forall\, z \in \mathbb{R}$$

Per **variabili gaussiane** le funzioni coinvolte nella procedura di best-fit sono

$$\rho(z) \;=\; \frac{z^2}{2} \qquad \text{e} \qquad \psi(z) \;=\; z$$

Per **variabili a distribuzione gaussiana bilatera**

$$\text{Prob}[y_i - y(x_i, \alpha)] \sim \exp\Big(-\Big|\frac{y_i - y(x_i, \alpha)}{\sigma_i}\Big|\Big)$$

si ha invece

$$\rho(z) \;=\; |z| \qquad \text{e} \qquad \psi(z) \;=\; \text{sgn}(z)$$

Se le variabili seguono invece una **distribuzione lorentziana** (o di Cauchy)

$$\text{Prob}[y_i - y(x_i, \alpha)] \sim \frac{1}{1 + \frac{1}{2}\left(\frac{y_i - y(x_i, \alpha)}{\sigma_i}\right)^2}$$

le funzioni per la relativa stima di best-fit diventano

$$\rho(z) = \ln\left(1 + \frac{z^2}{2}\right) \qquad \text{e} \qquad \psi(z) = \frac{z}{1 + \frac{z^2}{2}}$$

Nell'analisi di sistemi di dati con outliers non drammatici può essere vantaggioso fare uso di funzioni peso $\psi$ che non corrispondono ad alcuna distribuzione di probabilità notevole

**Esempio I. Andrew's sine**

$$\psi(z) = \begin{cases} \sin(z/c) & |z| < c\pi \\ 0 & |z| \geq c\pi \end{cases}$$

**Esempio II. Tukey's biweight**

$$\psi(z) = \begin{cases} z(1 - z^2/c^2)^2 & |z| < c \\ 0 & z \geq c \end{cases}$$

**Se le distribuzioni risultano circa normali** i valori ottimali delle costanti $c$ si dimostrano essere $c = 2.1$ e $c = 6.0$, nei due casi

## 10.9 Modelli multidimensionali

I problemi di modellazione in più variabili, dipendenti o indipendenti, si trattano fondamentalmente come nel caso di una singola variabile indipendente e di una sola variabile dipendente

Per problemi a **più variabili indipendenti**, le $k$-uple di variabili indipendenti vengono considerate come variabili singole

Le sommatorie che definiscono le funzioni obiettivo sono estese a tutte le $k$-uple disponibili

Per problemi a **più variabili dipendenti**, il calcolo delle funzioni obiettivo viene eseguito trattando ogni singola $h$-upla di variabili dipendenti come un vettore di $\mathbb{R}^h$, di cui viene considerato il modulo (di solito la consueta norma euclidea)

### Caso notevole è quello delle grandi matrici di dati

Esso viene spesso studiato per mezzo della c.d. **Principal Component Analysis (PCA)**

### 10.9.1 Principal Component Analysis (PCA)

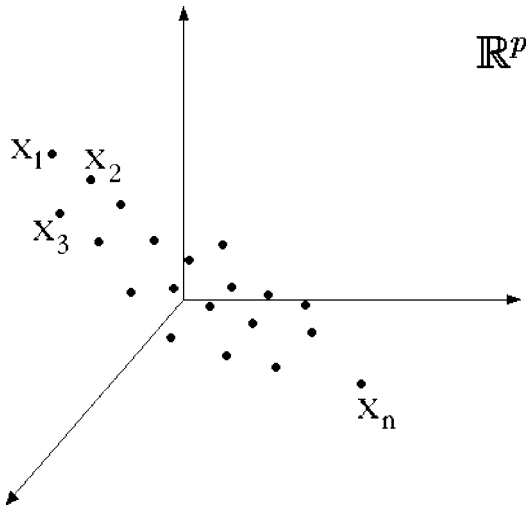
Sia data una matrice $X$ di dati, con $n$ righe e $p$ colonne

Si può assumere che ogni sua riga corrisponda ad un punto nello spazio $\mathbb{R}^p$ a $p$ componenti, una per ogni colonna

La matrice $X$ individua allora un set di $n$ punti in $\mathbb{R}^p$

$$X = \begin{pmatrix} x_1^T \\ \hline x_2^T \\ \hline \vdots \\ \hline x_n^T \end{pmatrix} \qquad x_i \in \mathbb{R}^p \quad \forall\, i = 1, \ldots, n$$

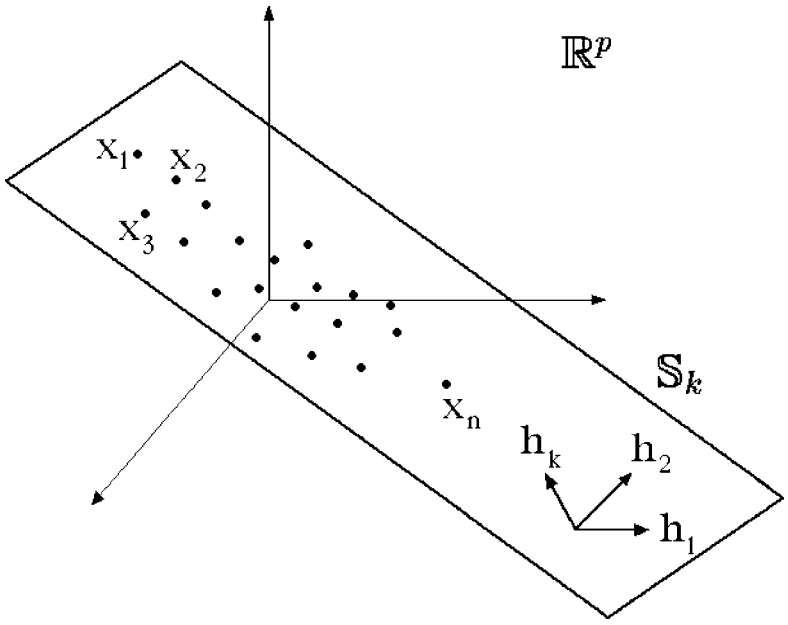
Può verificarsi che i punti $x_i$ non siano distribuiti a caso in $\mathbb{R}^p$, ma lungo un particolare sottoinsieme di $\mathbb{R}^p$



In particolare, si assuma che i dati siano descritti da uno spazio vettoriale $\mathbb{S}_k$, di dimensione $k \leq p$, generato da un sistema ortonormale di vettori di $\mathbb{R}^p$

$$\{h_1 \ , \ h_2 \ , \ \ldots \ , \ h_k\}$$

$$h_j^T h_q \ = \ \delta_{jq} \ \forall \, j, q = 1, \ldots, k$$




Tali vettori saranno le colonne di una matrice $p \times k$

$$H \ = \ \Big(h_1 \Big| h_2 \Big| \ldots \Big| h_k\Big)$$

Le proiezioni ortogonali dei vettori $x_i$ su $\mathbb{S}_k$ sono date dai vettori colonna

$$HH^T x_i \qquad \forall\, i = 1, \ldots, n$$

ovvero dai vettori riga

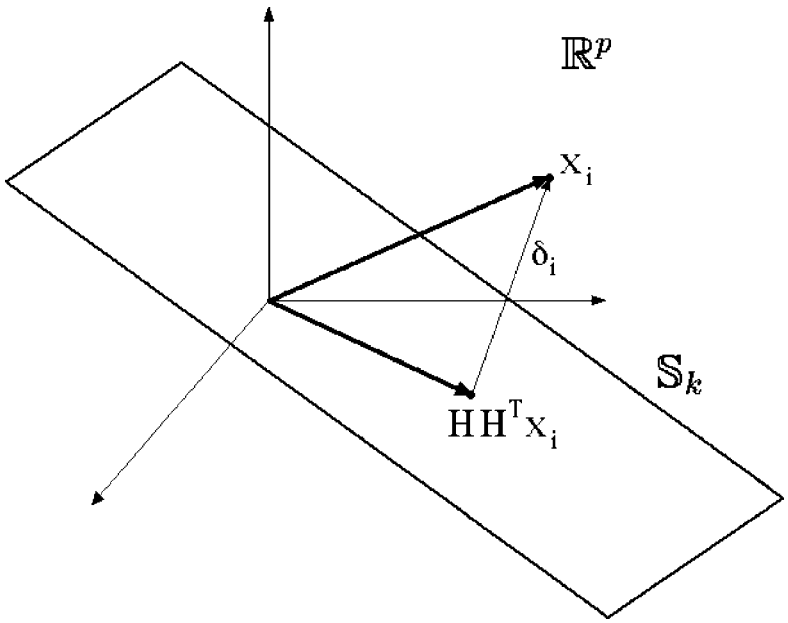
$$x_i^T HH^T \qquad \forall\, i = 1, \ldots, n$$

e la matrice delle proiezioni ortogonali vale dunque

$$XHH^T$$

per cui

proiezioni ortogonali su $\mathbb{S}_k$ degli $x_i$ $\iff$ righe di $XHH^T$



Le distanze dei punti $x_i$ da $\mathbb{S}_k$ — "scarti" — si scrivono

$$\delta_i \;=\; x_i - HH^T x_i \;=\; (\mathbb{I} - HH^T)x_i \qquad i = 1, \ldots, n$$

e la somma dei loro quadrati risulta

$$\begin{aligned} SS \;=\; \sum_{i=1}^{n} \delta_i^2 \;&= \mathrm{tr}[X(\mathbb{I} - HH^T)X^T] \;= \\ &= \mathrm{tr}[XX^T] - \mathrm{tr}[XHH^TX^T] \;= \\ &= \mathrm{tr}[XX^T] - \mathrm{tr}[H^TX^TXH] \end{aligned}$$

**Come determinare la matrice $H$ — ovvero $\mathbb{S}_k$?**

**Least squares:** si impone che la matrice $H$ — e quindi $\mathbb{S}_k$ — sia scelta in modo da rendere minima la somma dei quadrati degli scarti

$$\begin{array}{c} SS \quad \text{minima} \\ \Updownarrow \\ \mathrm{tr}[H^TX^TXH] \quad \text{massima} \end{array}$$

sotto la condizione che le colonne $h_1, \ldots, h_k$ di $H$ siano ortonormali

**Risultato**

Sia data una base ortonormale di autovettori

$$h_1 \ , \quad h_2 \ , \quad \dots \ , \quad h_p$$

della matrice reale simmetrica e semidefinita positiva

$$X^T X$$

con

$$X^T X h_i \ = \ \lambda_i h_i \quad \forall \, i = 1, \dots, p$$

e autovalori ordinati in senso decrescente

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

**Allora**

- si può porre

$$H \ = \ \left( h_1 \middle| h_2 \middle| \dots \middle| h_k \right)$$

- la corrispondente somma dei quadrati degli scarti vale

$$\mathrm{tr}[X(\mathbb{I} - HH^T)X^T] \ = \ \sum_{i=k+1}^{p} \lambda_i$$

Se si ritiene che i dati $x_i$ siano principalmente ubicati lungo $\mathbb{S}_k$, si sostituiscono i punti $x_i$ con le loro proiezioni ortogonali su $\mathbb{S}_k$

$$x_i \quad \longrightarrow \quad HH^T x_i$$

in modo che la matrice $X$ dei dati viene sostituita dal **modello a $k$ componenti principali**

$$XHH^T = \sum_{j=1}^{k} (Xh_j)h_j^T$$

(c.d. **calibrazione** del modello a $k$ componenti principali)

La somma dei quadrati dei dati ("devianza del campione")

$$\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 = \mathrm{tr}(X^T X) = \sum_{j=1}^{p} \lambda_j$$

si compone di una "devianza spiegata dal modello"

$$\mathrm{tr}[XHH^T X^T] = \mathrm{tr}[H^T X^T XH] = \sum_{j=1}^{k} \lambda_j$$

e di una "devianza non spiegata" o somma dei quadrati degli scarti

$$SS = \mathrm{tr}[X(\mathbb{I} - HH^T)X^T] = \sum_{j=k+1}^{p} \lambda_j$$

Se il modello a $k$ componenti principali è corretto

tutti i vettori $x \in \mathbb{R}^p$ di dati si rappresentano come combinazioni lineari degli autovettori ortonormali $h_1, \ldots, h_k$:

$$x = \sum_{j=1}^{k} \alpha_j h_j$$

con $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$

Ciò consente di esprimere $p-k$ componenti di $x$ come funzioni lineari delle restanti $k$

Ad esempio le prime $p-k$ in funzione delle ultime $k$:

$$\begin{aligned} x_1 &= f_1(x_{p-k+1}, \ldots, x_p) \\ x_2 &= f_2(x_{p-k+1}, \ldots, x_p) \\ \ldots &\quad \ldots \\ x_{p-k} &= f_{p-k}(x_{p-k+1}, \ldots, x_p) \end{aligned}$$

Il modello a $k$ componenti principali consente di prevedere i valori di $p-k$ componenti del dato $x \in \mathbb{R}^p$, noti che siano i valori delle restanti $k$!

(uso del modello a $k$ componenti principali per la **previsione**)

Per confronto, si verifica facilmente che il modello a $k$ componenti principali non è altro che il troncamento ai $k$ maggiori valori singolari della SVD di $X$

$$X = V \begin{pmatrix} \Sigma & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} U^T = \sum_{j=1}^{r} \sigma_j v_j u_j^T$$

con:

$$r = \text{Rank}(X) = \text{Rank}(X^T X) \leq \min(n, p)$$

$$U = \left( u_1 \middle| u_2 \middle| \ldots \middle| u_p \right)$$

$u_1, u_2, \ldots, u_p$ base ortonormale di autovettori di $X^T X$

$$X^T X u_i = \sigma_i^2 u_i \qquad i = 1, \ldots, p$$

secondo gli autovalori di $X^T X$ ordinati in senso decrescente

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_p^2$$

e

$$V = \left( v_1 \middle| v_2 \middle| \ldots \middle| v_n \right)$$

$v_1, v_2, \ldots, v_n$ base ortonormale di $\mathbb{R}^n$ tale che

$$v_i = \frac{1}{\sigma_i} X u_i \qquad \forall\, i = 1, \ldots, r$$

Pertanto, se $k \leq r$ — unico caso interessante

$$\sum_{j=1}^{k} \sigma_j v_j u_j^T = \sum_{j=1}^{k} \sigma_j \frac{1}{\sigma_j} (X u_j) u_j^T = \sum_{j=1}^{k} (X u_j) u_j^T$$

e per ottenere il modello a $k$ componenti principali basta porre $h_j = u_j \ \forall\, j = 1, \ldots, k$. Vale inoltre $\lambda_j = \sigma_j^2 \ \forall\, j = 1, \ldots, k$.

Il modello a $k$ componenti principali della matrice $X$ è perciò dato da

$$XHH^T = \sum_{j=1}^{k} (Xh_j)h_j^T$$

con:

- $h_i$ $i$-esima componente principale (**loading**)

e

- $Xh_i$ $i$-esimo **score**

**N.B.** La PCA non è altro che una SVD **troncata**

Gli algoritmi per la PCA sono perciò gli stessi della SVD!

### 10.9.2 PCA per dati non centrati (PCA generale)

Si può anche pensare di modellare la matrice $X$ dei dati

mediante una **varietà lineare di dimensione** $k$

che **non passa per l'origine** $O$

e dunque non è un sottospazio vettoriale di $\mathbb{R}^p$ come $\mathbb{S}_k$

Si prova che il risultato è ancora fornito dalla procedura precedente

a patto però di sottrarre a tutti i punti $x_i$ i loro valori medi

$$x_i - \overline{x} \;=\; x_i - \frac{1}{n}\sum_{q=1}^{n} x_q$$

ovvero di eseguire l'analisi sulla matrice dei dati "centrata"

$$X - NX \;=\; (\mathbb{I} - N)X$$

con la matrice $n \times n$ $N$ di elementi costanti

$$N_{ij} \;=\; \frac{1}{n} \qquad \forall\, i, j = 1, \ldots, n$$

(La prova segue da un argomento alla Huygens-Steiner)

# 11. ANALISI DELLA VARIANZA (ANOVA)

## Problema tipico

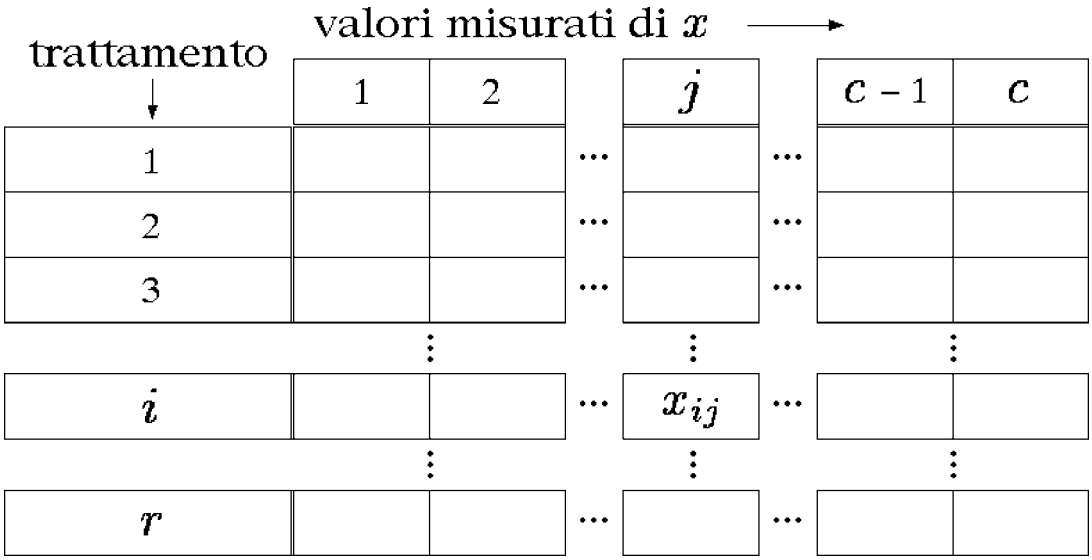
Un materiale è sottoposto a vari tipi di trattamento

$$\text{trattamenti} \qquad 1\,,\quad 2\,,\quad \ldots\,,\quad r$$

e per ogni materiale trattato si eseguono più misure di una proprietà $x$

| | | |
|---|---|---|
| valori misurati di $x$ sul materiale sottoposto al trattamento $i$ | $x_{ij}$ | $i = 1, 2, \ldots, r$<br>$j = 1, 2, \ldots, c$ |

(si suppone per semplicità che il numero di misure sia lo stesso per tutti i trattamenti)




## Domanda

I valori medi della grandezza $x$ sono significativamente diversi per i vari trattamenti?

Oppure sono sostanzialmente eguali, per cui i diversi trattamenti non influenzano la proprietà $x$ in modo significativo?

## Modello

Sistema di variabili casuali a due indici

$$x_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, c$$

$i$ individua la popolazione (trattamento)

$j$ specifica un elemento della $i$-esima popolazione
(valore misurato sul campione sottoposto al trattamento $i$)

Per ogni $i = 1, \ldots, r$ e $j = 1, \ldots, c$ la variabile casuale $x_{ij}$ è gaussiana di media $\mu_i$ e varianza $\sigma^2$

**N.B.** Tutte le popolazioni hanno la stessa varianza $\sigma^2$, ma potrebbero presentare **medie diverse** $\mu_i$

## Ipotesi da testare

$H_0 : \mu_i = \mu \; \forall \, i = 1, \ldots, r$ (medie tutte eguali)

$H_1 : \mu_h \neq \mu_k$ per almeno un $h$ e un $k$ in $1, \ldots, r$
(almeno due medie diverse)

## Metodo

**Analisi della varianza** a un fattore
(**AN**alysis **O**f **VA**riance - **ANOVA**)

## ANOVA

Per un sistema di dati organizzato in $r$ campioni di $c$ dati ciascuno, l'analisi della varianza a un fattore consiste nell'interpretare (o "spiegare") la dispersione dei dati distinguendo

- un contributo relativo alla scelta del campione (riga)
- un contributo intrinseco ai singoli campioni

I campioni sono diversificati sulla base di una opportuna condizione sperimentale, o **"fattore"**

Nell'esempio in esame, l'unico fattore che differenzia i vari campioni è il tipo di trattamento

L'ANOVA che si applica è a **un fattore**

Può darsi il caso di campioni che vengono differenziati da due o più condizioni sperimentali variabili in modo controllato

Ad esempio, il trattamento potrebbe consistere nell'attacco di agente chimico a temperature diverse

$$\text{campione} \iff \text{agente chimico} + \text{temperatura}$$

nel qual caso ogni campione sarà individuato da una **coppia di indici** e si utilizzerà una **ANOVA a due fattori**

Ancora, il campione potrebbe essere individuato dalla specifica dell'agente chimico utilizzato, dalla temperatura di trattamento e dalla durata del trattamento

$$\text{campione} \iff \begin{array}{l} \text{agente chimico} + \text{temperatura} \\ \qquad\qquad + \text{tempo} \end{array}$$

cioè da una **terna di indici**. L'ANOVA da applicare sarà a **tre fattori**

## Formulazione matematica generale dell'ANOVA

### ANOVA a un fattore

associato al **primo** indice

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \sum_i \alpha_i = 0 \qquad \varepsilon_{ij} \text{ i.i.d. } N(0,\sigma)$$

oppure associato al **secondo** indice

$$x_{ij} = \mu + \beta_j + \varepsilon_{ij} \qquad \sum_j \beta_j = 0 \qquad \varepsilon_{ij} \text{ i.i.d. } N(0,\sigma)$$

### ANOVA a due fattori senza interazione

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0$$

$$\varepsilon_{ij} \text{ i.i.d. } N(0,\sigma)$$

### ANOVA a due fattori con interazione

$$x_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij} \qquad \sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0$$

$$\sum_i (\alpha\beta)_{ij} = 0 \quad \forall\, j \quad \sum_j (\alpha\beta)_{ij} = 0 \quad \forall\, i \quad \varepsilon_{ij} \text{ i.i.d. } N(0,\sigma)$$

### ANOVA a tre fattori senza interazione

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \text{ i.i.d. } N(0,\sigma)$$

$$\sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_k \gamma_k = 0$$

**ecc. ecc.**

## 11.1 ANOVA a un fattore

L'ANOVA a un fattore fa uso di alcune definizioni fondamentali

**Media dei dati**

$$\overline{\overline{x}} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} = \frac{1}{rc}\sum_{i,j} x_{ij}$$

**Media di un campione**
la media dei dati di una riga, ossia entro un campione

$$\overline{x}_{i\cdot} = \frac{1}{c}\sum_{j=1}^{c} x_{ij}$$

**Scarto**
differenza fra il valore di una variabile e la media dei dati

$$x_{ij} - \overline{\overline{x}}$$

**Scarto "spiegato" dal fattore associato all'indice** $i$
la parte dello scarto imputabile al diverso valor medio dei vari campioni (righe), stimato per mezzo della relativa media campionaria

$$\overline{x}_{i\cdot} - \overline{\overline{x}}$$

**Scarto "non spiegato"**
componente residua dello scarto, indice della variabilità dei dati internamente al campione (entro la riga)

$$x_{ij} - \overline{\overline{x}} - (\overline{x}_{i\cdot} - \overline{\overline{x}}) = x_{ij} - \overline{x}_{i\cdot}$$

**Variazione totale**
la somma dei quadrati degli scarti

$$SS_t = \sum_{i,j} (x_{ij} - \overline{\overline{x}})^2$$

**Variazione spiegata dal fattore associato all'indice** $i$
la somma dei quadrati degli scarti spiegati dal fattore

$$SS_1 = \sum_{i,j} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2 = c \sum_{i=1}^{r} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2$$

Esprime la variazione dei dati imputabile al fatto che i singoli campioni (righe) presentano medie diverse

**Variazione non spiegata**
la somma dei quadrati degli scarti non spiegati

$$SS_u = \sum_{i,j} (x_{ij} - \overline{x}_{i\cdot})^2$$

Descrive la variazione dei dati entro ogni singolo campione (riga), variazione imputabile a fenomeni puramente casuali e non all'appartenenza ad un dato campione piuttosto che a un altro

Indipendente dal fatto che i singoli campioni abbiano o meno la stessa media

## ANOVA a un fattore

Modello:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, c$$

con:

$$\sum_i \alpha_i = 0$$

e

$$\varepsilon_{ij} \text{ variabili casuali i.i.d. } N(0, \sigma) \; \forall \, i, j$$

Si fa l'ipotesi forte che $\sigma$ sia la stessa per tutte le variabili!

L'identità algebrica

$$\begin{aligned} \sum_{i,j} (x_{ij} - \overline{\overline{x}})^2 = & \quad \text{variazione totale } SS_t \\ = & \sum_{i,j} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2 \quad \text{variazione spiegata } SS_1 \\ + & \sum_{i,j} (x_{ij} - \overline{x}_{i\cdot})^2 \quad \text{variazione non spiegata } SS_u \end{aligned}$$

vale **indipendentemente dalla correttezza del modello**

**Se il modello è corretto** si ha inoltre che

$$\frac{1}{\sigma^2}\sum_{i,j}(x_{ij}-\overline{\overline{x}}-\alpha_i)^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij}) \middle| (\mathbb{I}-P_1P_2)\frac{1}{\sigma}(\varepsilon_{ij}) \right\rangle = {\mathcal{X}^2}_{rc-1}$$
$$\frac{1}{\sigma^2}\sum_{i,j}(\overline{x}_{i\cdot}-\overline{\overline{x}}-\alpha_i)^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij}) \middle| P_2(\mathbb{I}-P_1)\frac{1}{\sigma}(\varepsilon_{ij}) \right\rangle = {\mathcal{X}^2}_{r-1}$$
$$\frac{1}{\sigma^2}\sum_{i,j}(x_{ij}-\overline{x}_{i\cdot})^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij}) \middle| (\mathbb{I}-P_2)\frac{1}{\sigma}(\varepsilon_{ij}) \right\rangle = {\mathcal{X}^2}_{r(c-1)}$$

con ${\mathcal{X}^2}_{r-1}$ e ${\mathcal{X}^2}_{r(c-1)}$ **stocasticamente indipendenti**

**Notazioni:**

- $(v_{ij}) = (v_{11}, v_{12}, \ldots, v_{rc}) \in \mathbb{R}^{rc}$ (generico vettore di $\mathbb{R}^{rc}$)
- $P_1(v_{ij}) = (\overline{v}_{\cdot j})$ (operatore di media sul $1^o$ indice)
- $P_2(v_{ij}) = (\overline{v}_{i\cdot})$ (operatore di media sul $2^o$ indice)
- $\alpha(v_{ij}) = (\alpha v_{ij}) \quad \forall\, \alpha \in \mathbb{R}, \forall\, (v_{ij}) \in \mathbb{R}^{rc}$
- $\langle (v_{ij}) | (u_{ij}) \rangle = \sum_{ij} v_{ij}u_{ij} \quad \forall\, (v_{ij}), (u_{ij}) \in \mathbb{R}^{rc}$
  (prodotto scalare standard in $\mathbb{R}^{rc}$)

**Teorema fondamentale**

Per ogni insieme di dati $x_{ij}$ la variazione totale è uguale alla somma della variazione spiegata dal fattore associato a $i$ e di quella non spiegata

$$SS_t \;=\; SS_1 + SS_u$$

Se poi le variabili $x_{ij}$ sono indipendenti e identicamente distribuite, con distribuzione gaussiana di media $\mu$ e varianza $\sigma^2$, si ha che

$SS_t/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $rc-1$ gradi di libertà (gdl)

$SS_1/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $r-1$ gdl

$SS_u/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $rc-r$ gdl

$SS_1/\sigma^2$ e $SS_u/\sigma^2$ sono stocasticamente indipendenti

**Osservazioni**

• si ha un bilancio del numero di gradi di libertà

$$rc-1 \;=\; (r-1) \;+\; (rc-r)$$

• le somme dei quadrati degli scarti normalizzate al rispettivo numero di gradi di libertà sono definite come **varianze**

$SS_t/(rc-1)$ varianza totale

$SS_1/(r-1)$ varianza spiegata dal fattore considerato

$SS_u/(rc-r)$ varianza non spiegata

• il rapporto della varianza spiegata su quella non spiegata

$$F \;=\; \frac{SS_1}{r-1}\,\frac{r(c-1)}{SS_u} \;=\; \frac{r(c-1)}{r-1}\,\frac{SS_1}{SS_u}$$

segue una distribuzione di Fisher a $(r-1, rc-r)$ gdl

Tornando al modello, se è verificata l'ipotesi nulla

$$H_0 \ : \quad \mu_i \ = \ \mu \qquad \forall\, i = 1, \ldots, r$$

il quoziente $F$ segue una distribuzione di Fisher a $(r-1, rc-r)$ gdl

la varianza spiegata $SS_1/(r-1)$ risulta molto vicina a zero

la varianza non spiegata — $SS_u/r(c-1)$ — si può ritenere sia dell'ordine di $\sigma^2$

Viceversa, se $H_0$ è falsa (cioè almeno una delle medie $\mu_i$ risulta diversa dalle altre) ci si aspetta che

$$\frac{1}{r-1} SS_1 \ \gtrsim \sigma^2$$

mentre la varianza non spiegata è comunque dell'ordine di $\sigma^2$

$$\Downarrow$$

$F$ si assume come variabile del test

e la regione di rigetto di $H_0$ è del tipo

$$F \geq F_{\text{critico}} \ = \ F_{[1-\alpha](r-1, r(c-1))}$$

con livello di significatività $\alpha$
(si rigetta $H_0$ se $F$ risulta abbastanza grande)

**Osservazione**

Nelle ipotesi del modello — stessa varianza $\sigma^2$ e medie $\mu_i$ per i vari campioni — è spesso importante determinare gli intervalli di confidenza delle differenze delle medie

Si può utilizzare il seguente risultato

**Teorema di Scheffé**
Per ogni $r$-upla di coefficienti reali $(C_1, C_2, \ldots, C_r)$ a media nulla

$$\sum_{i=1}^{r} C_i \;=\; 0$$

e per ogni $\alpha \in (0,1)$ fissato, la diseguaglianza

$$\left|\sum_{i=1}^{r} C_i(\overline{x}_{i\cdot} - \mu_i)\right|^2 \leq F_{[\alpha](r-1,rc-r)} \frac{SS_u}{r(c-1)} \, \frac{r-1}{c} \, \sum_{k=1}^{r} C_k^2$$

è verificata con probabilità $\alpha$

Ne segue l'intervallo di confidenza, con livello di confidenza $\alpha$, della combinazione lineare $\sum_{i=1}^{r} C_i \mu_i$

$$\sum_{i=1}^{r} C_i \overline{x}_{i\cdot} \pm \sqrt{F_{[\alpha](r-1,rc-r)}} \sqrt{\frac{SS_u}{r(c-1)}} \sqrt{\frac{r-1}{c} \, \sum_{k=1}^{r} C_k^2}$$

In particolare, per la differenza fra due medie $\mu_i$ e $\mu_h$ qualsiasi l'intervallo di confidenza è

$$\overline{x}_{i\cdot} - \overline{x}_{h\cdot} \pm \sqrt{F_{[\alpha](r-1,rc-r)}} \sqrt{\frac{SS_u}{r(c-1)}} \sqrt{2\frac{r-1}{c}}$$

sempre con livello di confidenza $\alpha$

## 11.2 ANOVA a due fattori

Il sistema di variabili casuali a due indici

$$x_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, c$$

descrive i valori misurati della grandezza $x$ relativa ad un materiale trattato con certo agente chimico ad una data temperatura

$i$ individua l'agente chimico trattante

$j$ specifica la temperatura del trattamento

### Domanda

L'agente chimico o la temperatura del trattamento hanno effetti sulla proprietà $x$ del materiale?

### Modello

Per ogni $i = 1, \ldots, r$ e $j = 1, \ldots, c$ la variabile casuale $x_{ij}$ è gaussiana di media $\mu_{ij}$ e varianza $\sigma^2$

**N.B.** La media della variabile $x_{ij}$ può dipendere tanto da $i$ **quanto da** $j$ (sia dall'agente chimico che dalla temperatura), mentre la varianza $\sigma^2$ è la stessa.

### Ipotesi da testare

$H_0 : \mu_{ij} = \mu \; \forall\, i = 1, \ldots, r,\; j = 1, \ldots, c$ (medie tutte eguali)

$H_1 :$ $\mu_{ij} \neq \mu_{hj}$ per almeno un $j$ e una coppia $i, h$ $(i \neq h)$
oppure
$\mu_{ij} \neq \mu_{ik}$ per almeno un $i$ e una coppia $j, k$ $(j \neq k)$

## Definizioni

### Media dei dati

$$\overline{\overline{x}} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} = \frac{1}{rc}\sum_{i,j} x_{ij}$$

### Media sull'indice $j$

$$\overline{x}_{i\cdot} = \frac{1}{c}\sum_{j=1}^{c} x_{ij}$$

### Media sull'indice $i$

$$\overline{x}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{c} x_{ij}$$

### Scarto

$$x_{ij} - \overline{\overline{x}}$$

### Scarto "spiegato" dal fattore associato all'indice $i$

$$\overline{x}_{i\cdot} - \overline{\overline{x}}$$

### Scarto "spiegato" dal fattore associato all'indice $j$

$$\overline{x}_{\cdot j} - \overline{\overline{x}}$$

### Scarto "non spiegato"

$$x_{ij} - \overline{\overline{x}} - (\overline{x}_{i\cdot} - \overline{\overline{x}}) - (\overline{x}_{\cdot j} - \overline{\overline{x}}) = x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{\overline{x}}$$

**Variazione totale**

$$SS_t = \sum_{i,j} (x_{ij} - \overline{\overline{x}})^2$$

**Variazione spiegata dal fattore associato all'indice $i$**

$$SS_1 = \sum_{i,j} (\overline{x}_{i.} - \overline{\overline{x}})^2 = c \sum_{i=1}^{r} (\overline{x}_{i.} - \overline{\overline{x}})^2$$

**Variazione spiegata dal fattore associato all'indice $j$**

$$SS_2 = \sum_{i,j} (\overline{x}_{.j} - \overline{\overline{x}})^2 = r \sum_{j=1}^{c} (\overline{x}_{.j} - \overline{\overline{x}})^2$$

**Variazione non spiegata**

$$SS_u = \sum_{i,j} (x_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{\overline{x}})^2$$

### 11.2.1 ANOVA a due fattori senza interazione

Modello:

$$x_{ij} \;=\; \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, c$$

con:

$$\sum_i \alpha_i \;=\; 0 \qquad\qquad \sum_j \beta_j \;=\; 0$$

e

$$\varepsilon_{ij} \text{ variabili casuali i.i.d. } N(0, \sigma)\ \forall\, i, j$$

$\sigma$ si assume ancora la stessa per tutte le variabili!

L'identità algebrica

$$\begin{aligned}
&\sum_{i,j} (x_{ij} - \overline{\overline{x}})^2 && \text{variazione totale } SS_t \\
&= \sum_{i,j} (\overline{x}_{i\cdot} - \overline{\overline{x}})^2 && \text{var.ne spiegata dal } 1^o \text{ fattore, } SS_1 \\
&+ \sum_{i,j} (\overline{x}_{\cdot j} - \overline{\overline{x}})^2 && \text{var.ne spiegata dal } 2^o \text{ fattore, } SS_2 \\
&+ \sum_{i,j} (x_{ij} - \overline{x}_{\cdot j} - \overline{x}_{i\cdot} + \overline{\overline{x}})^2 && \text{var.ne non spiegata } SS_u
\end{aligned}$$

è **indipendente dalla correttezza del modello**

**Se il modello è corretto** risulta inoltre

$$\frac{1}{\sigma^2}\sum_{i,j}(x_{ij}-\overline{\overline{x}}-\alpha_i-\beta_j)^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij})\middle|(\mathbb{I}-P_1P_2)\frac{1}{\sigma}(\varepsilon_{ij})\right\rangle = {\mathcal{X}^2}_{rc-1}$$

$$\frac{1}{\sigma^2}\sum_{i,j}(\overline{x}_{i\cdot}-\overline{\overline{x}}-\alpha_i)^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij})\middle|P_2(\mathbb{I}-P_1)\frac{1}{\sigma}(\varepsilon_{ij})\right\rangle = {\mathcal{X}^2}_{r-1}$$

$$\frac{1}{\sigma^2}\sum_{i,j}(\overline{x}_{\cdot j}-\overline{\overline{x}}-\beta_j)^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij})\middle|(\mathbb{I}-P_2)P_1\frac{1}{\sigma}(\varepsilon_{ij})\right\rangle = {\mathcal{X}^2}_{c-1}$$

$$\frac{1}{\sigma^2}\sum_{i,j}(x_{ij}-\overline{x}_{i\cdot}-\overline{x}_{\cdot j}+\overline{\overline{x}})^2 =$$
$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ij})\middle|(\mathbb{I}-P_2)(\mathbb{I}-P_1)\frac{1}{\sigma}(\varepsilon_{ij})\right\rangle = {\mathcal{X}^2}_{(r-1)(c-1)}$$

con

$${\mathcal{X}^2}_{r-1} \qquad {\mathcal{X}^2}_{c-1} \qquad {\mathcal{X}^2}_{(r-1)(c-1)}$$

**stocasticamente indipendenti**

## Teorema fondamentale

Per ogni insieme $x_{ij}$ la variazione totale è pari alla somma della variazione spiegata dal fattore associato a $i$, della variazione spiegata dal fattore associato a $j$ e di quella non spiegata

$$SS_t = SS_1 + SS_2 + SS_u$$

Se poi le variabili $x_{ij}$ sono indipendenti e identicamente distribuite, gaussiane di media $\mu$ e varianza $\sigma^2$, si ha che

$SS_t/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $rc-1$ gradi di libertà (gdl)

$SS_1/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $r-1$ gdl

$SS_2/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $c-1$ gdl

$SS_u/\sigma^2$ è una variabile di $\mathcal{X}^2$ a $(r-1)(c-1)$ gdl

$SS_1/\sigma^2$, $SS_2/\sigma^2$ e $SS_u/\sigma^2$ sono stocasticamente indipendenti

**Osservazioni**

• il numero di gradi di libertà si bilancia

$$rc - 1 \;=\; (r-1) \;+\; (c-1) \;+\; (r-1)(c-1)$$

• si definiscono le **varianze**

$SS_t/(rc-1)$ varianza totale

$SS_1/(r-1)$ varianza spiegata dal fattore associato a $i$

$SS_2/(c-1)$ varianza spiegata dal fattore associato a $j$

$SS_u/(r-1)(c-1)$ varianza non spiegata

• il rapporto della varianza spiegata dal fattore associato a $i$ su quella non spiegata

$$F \;=\; \frac{SS_1}{r-1}\,\frac{(r-1)(c-1)}{SS_u} \;=\; (c-1)\,\frac{SS_1}{SS_u}$$

segue una distribuzione di Fisher a $(r-1,(r-1)(c-1))$ gdl

• il rapporto della varianza spiegata dal fattore associato a $j$ su quella non spiegata

$$F \;=\; \frac{SS_2}{c-1}\,\frac{(r-1)(c-1)}{SS_u} \;=\; (r-1)\,\frac{SS_2}{SS_u}$$

segue una distribuzione di Fisher a $(c-1,(r-1)(c-1))$ gdl

**Test per la differenza fra le medie rispetto all'indice $i$**

**Variabile del test**

$$F \;=\; \frac{SS_1}{r-1}\,\frac{(r-1)(c-1)}{SS_u} \;=\; (c-1)\,\frac{SS_1}{SS_u}$$

Se le medie $\mu_{ij}$ non dipendono sensibilmente dall'indice $i$ vale

$$\frac{SS_1}{r-1} \ll \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \quad \Longleftrightarrow \quad F \text{ piccolo}$$

mentre in caso contrario ci si aspetta che

$$\frac{SS_1}{r-1} \gtrsim \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \quad \Longleftrightarrow \quad F \text{ grande}$$

Si definisce allora un valore critico di $F$

$$F_{\text{critico}} \;=\; F_{[1-\alpha](r-1,(r-1)(c-1))}$$

e una regione di accettazione di $H_0$ — medie indipendenti dall'indice $i$ —

$$F \leq F_{[1-\alpha](r-1,(r-1)(c-1))}$$

nonché una corrispondente regione di rigetto — medie $\mu_{ij}$ dipendenti in modo sensibile dall'indice $i$ —

$$F > F_{[1-\alpha](r-1,(r-1)(c-1))}$$

entrambe con livello di significatività $\alpha \in (0,1)$ prefissato

**Test per la differenza fra le medie rispetto all'indice $j$**

**Variabile del test**

$$F \;=\; \frac{SS_2}{c-1}\,\frac{(r-1)(c-1)}{SS_u} \;=\; (r-1)\,\frac{SS_2}{SS_u}$$

Se le medie $\mu_{ij}$ non dipendono sensibilmente dall'indice $j$ si ha

$$\frac{SS_2}{c-1} \ll \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \quad \Longleftrightarrow \quad F \text{ piccolo}$$

mentre in caso contrario è verosimile che si abbia

$$\frac{SS_2}{c-1} \gtrsim \frac{SS_u}{(r-1)(c-1)} \simeq \sigma^2 \quad \Longleftrightarrow \quad F \text{ grande}$$

Si definisce allora un valore critico di $F$

$$F_{\text{critico}} \;=\; F_{[1-\alpha](c-1,(r-1)(c-1))}$$

e una regione di accettazione di $H_0$ — medie indipendenti dall'indice $j$ —

$$F \leq F_{[1-\alpha](c-1,(r-1)(c-1))}$$

nonché una corrispondente regione di rigetto — medie $\mu_{ij}$ dipendenti in modo sensibile dall'indice $j$ —

$$F > F_{[1-\alpha](c-1,(r-1)(c-1))}$$

entrambe con livello di significatività $\alpha \in (0,1)$ prefissato

### 11.2.2 ANOVA a due fattori con interazione

Modello:

$$\begin{aligned} x_{ijk} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ & \quad i = 1, \ldots, r \quad j = 1, \ldots, c \quad k = 1, \ldots, s \end{aligned}$$

con:

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0$$

$$\sum_i \alpha\beta_{ij} = 0 \quad \forall\, j \qquad \sum_j \alpha\beta_{ij} = 0 \quad \forall\, i$$

e

$$\varepsilon_{ijk} \text{ variabili casuali i.i.d. } N(0, \sigma)\ \forall\, i, j, k$$

L'identità algebrica

$$\begin{aligned} & \sum_{ijk} (x_{ijk} - \overline{\overline{\overline{x}}})^2 && \text{variazione totale } SS_t \\ = & \sum_{ijk} (\overline{\overline{x}}_{i\cdot\cdot} - \overline{\overline{\overline{x}}})^2 && \text{var.ne spiegata dal } 1^o \text{ fattore, } SS_1 \\ + & \sum_{ijk} (\overline{\overline{x}}_{\cdot j\cdot} - \overline{\overline{\overline{x}}})^2 && \text{var.ne spiegata dal } 2^o \text{ fattore, } SS_2 \\ + & \sum_{ijk} (\overline{x}_{ij\cdot} - \overline{\overline{x}}_{\cdot j\cdot} - \overline{\overline{x}}_{i\cdot\cdot} + \overline{\overline{\overline{x}}})^2 && \text{var.ne spiegata dall'interazione fra i fattori 1 e 2, } S_{12} \\ + & \sum_{ijk} (x_{ijk} - \overline{x}_{ij\cdot})^2 && \text{var.ne non spiegata } SS_u \end{aligned}$$

vale **anche se il modello non è corretto**

**Se il modello è corretto** si ha inoltre che

$$\frac{1}{\sigma^2}\sum_{ijk}(x_{ijk}-\overline{\overline{\overline{x}}}-\alpha_i-\beta_j-\alpha\beta_{ij})^2 =$$

$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ijk}) \middle| (\mathbb{I}-P_3P_2P_1)\frac{1}{\sigma}(\varepsilon_{ijk}) \right\rangle = {\mathcal{X}^2}_{rcs-1}$$

$$\frac{1}{\sigma^2}\sum_{ijk}(\overline{\overline{x}}_{i..}-\overline{\overline{\overline{x}}}-\alpha_i)^2 =$$

$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ijk}) \middle| P_3P_2(\mathbb{I}-P_1)\frac{1}{\sigma}(\varepsilon_{ijk}) \right\rangle = {\mathcal{X}^2}_{r-1}$$

$$\frac{1}{\sigma^2}\sum_{ijk}(\overline{\overline{x}}_{.j.}-\overline{\overline{\overline{x}}}-\beta_j)^2 =$$

$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ijk}) \middle| P_3(\mathbb{I}-P_2)P_1\frac{1}{\sigma}(\varepsilon_{ijk}) \right\rangle = {\mathcal{X}^2}_{c-1}$$

$$\frac{1}{\sigma^2}\sum_{ijk}(\overline{x}_{ij.}-\overline{\overline{x}}_{.j.}-\overline{\overline{x}}_{i..}+\overline{\overline{\overline{x}}}-\alpha\beta_{ij})^2 =$$

$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ijk}) \middle| P_3(\mathbb{I}-P_2)(\mathbb{I}-P_1)\frac{1}{\sigma}(\varepsilon_{ijk}) \right\rangle = {\mathcal{X}^2}_{(r-1)(c-1)}$$

$$\frac{1}{\sigma^2}\sum_{ijk}(x_{ijk}-\overline{x}_{ij.})^2 =$$

$$= \left\langle \frac{1}{\sigma}(\varepsilon_{ijk}) \middle| (\mathbb{I}-P_3)\frac{1}{\sigma}(\varepsilon_{ijk}) \right\rangle = {\mathcal{X}^2}_{rc(s-1)}$$

con

$${\mathcal{X}^2}_{r-1} \qquad {\mathcal{X}^2}_{c-1} \qquad {\mathcal{X}^2}_{(r-1)(c-1)} \qquad {\mathcal{X}^2}_{rc(s-1)}$$

**stocasticamente indipendenti**