Statistica

(M.S. Bernabei)

Per ulteriori approfondimenti:

- **R. Johnson & G.K. Bhattacharyya, Statistics, Principles and methods, Wiley**

Per le applicazioni:

- **M. Middleton, Analisi statistica con Excel, Apogeo**

Introduzione

La parola statistica deriva dal latino "status" e significa stato. Per lungo tempo la statistica è stata identificata soltanto con la rappresentazione di dati e relativi grafici relativi al mondo dell'economia, politico, ecc.. Durante il ventesimo secolo la statistica è cresciuta notevolmente come disciplina scientifica.

La Statistica va ben oltre la semplice rappresentazione dei dati, infatti essa si occupa

- della raccolta dei dati;
- della descrizione e sintesi dei dati,
- dell' interpretazione dei dati in modo da trarre conclusioni sul fenomeno in esame.

La Statistica si applica a tutti i fenomeni che coinvolgono la raccolta e l'analisi dei dati: sondaggi di opinione (indagini di tipo sociale, economico o sulla salute), esperimenti nel campo dell'agricoltura (su nuovi semi, pesticidi), studi clinici di vaccini, inseminazione di una nuvola per la produzione di pioggia artificiale, economico (disoccupazione, costo della vita, soddisfazione per un certo prodotto), politico (elezioni, miglioramento dei servizi dello stato), industria (qualità e miglioramento della produzione), ecc..

I principi e le metodologie della statistica sono utili per rispondere a domande del tipo:

- Che tipo e quanti dati è necessario raccogliere?
- Come dovremmo organizzare e interpretare i dati?
- Come possiamo analizzare i dati e dedurre le conclusioni?
- Come stimare la potenza delle conclusioni e giudicare la loro incertezza?

Dai dati e tabelle distribuite dai mass media, dai reports, la statistica inferenziale fornisce criteri per determinare quali conclusioni sono supportate dai dati e quelle che non lo sono. La credibilità delle conclusioni dipende fortemente dai metodi statistici utilizzati.

Lo statistico

- stabilisce gli obiettivi di una indagine,
- raccoglie dati o informazioni sul fenomeno,
- analizza i dati, traccia conclusioni
- decide ulteriori linee di azione.

Esempi

Programma di formazione. I programmi di formazione sono specifici per un certo tipo di utenti (studenti, operai, persone con handicap fisico, bambini ritardati, ecc.) e vengono continuamente monitorati, valutati e modificati per migliorare la loro utilità verso la società. Per confrontare l' efficacia fra i diversi tipi di programmi, è essenziale raccogliere dati sul raggiungimento o miglioramento delle abilità dei tirocinanti alla fine di ogni programma.

Richieste di controllo pubblicitario. Il pubblico è costantemente bombardato dalla pubblicità che dichiara la superiorità di un prodotto di una certa marca rispetto ad altri. Quando tali confronti sono basati su una valida dimostrazione sperimentale, servono ad educare ed orientare il consumatore. Non di rado, comunque, dei messaggi pubblicitari fuorvianti sono basati su una insufficiente sperimentazione, su analisi dei dati errate, o anche su vistose manipolazioni dei risultati sperimentali. Le agenzie di governo e i gruppi di consumatori devono essere preparati a verificare la relativa qualità dei prodotti usando una adeguata procedura dei dati raccolti e appropriati metodi di analisi statistica.

Riproduzione delle piante. Per far crescere la produzione di cibo, gli scienziati nel campo dell' agricoltura sviluppano nuovi ibridi dall' incrocio di diverse specie di piante. Le nuove specie di piante devono essere confrontate con quelle già esistenti. La loro relativa produttività è stimata piantando in certo numero di posti ciascuna di ogni specie. I prodotti sono conservati e quindi analizzati per comprendere se ci sono sostanziali differenze. Le specie potrebbero essere anche confrontate sulla base di resistenza alle malattie o capacità di riprodursi.

Per ogni analisi statistica è molto importante avere informazioni reali. La branca della statistica chiamata **disegno sperimentale** potrebbe guidare colui che investiga nel progettare in che modo e in che quantità raccogliere i dati. Dopo che i dati sono stati raccolti, ci sono dei metodi statistici che sintetizzano e descrivono le caratteristiche principali dei dati, essi sono comunemente conosciuti come **statistica descrittiva**. Oggi, fornisce una maggiore conoscenza del fenomeno in esame la stima dell' informazione presente nei dati. Tale area è chiamata **statistica inferenziale**.

Bisogna capire che spesso la ricerca scientifica è tipicamente un processo di prove che potrebbero contenere errori. Raramente un fenomeno potrebbe essere compreso completamente o una teoria perfezionata per mezzo di un singolo e unico esperimento, E' troppo aspettarsi di ottenere tutto in una sola prova. I dati ottenuti da un esperimento fornisce nuova conoscenza. Tale conoscenza spesso suggerisce una revisione di una esistente teoria, e ciò potrebbe richiedere ulteriori ricerche attraverso altri esperimenti e analisi dei dati.

Anche se gli esempi precedenti provengono da campi molto lontani tuttavia ci sono delle caratteristiche che li accomuna:

- per acquisire nuova conoscenza si ha bisogno di raccogliere dati rilevanti per il fenomeno in esame.
- anche se spesso le osservazioni siano fatte nelle "stesse condizioni" tuttavia è impossibile eliminare la variabilità dei dati. Ad esempio il trattamento di una allergia potrebbe creare un lungo sollievo per qualche individuo, mentre potrebbe portare solo un momentaneo sollievo o nessuno per tutti gli altri pazienti.
- L'accesso all'intero insieme di dati è praticamente impossibile o da un punto di vista pratico non facile per le limitazioni di tempo, delle risorse e delle facilitazioni, così che dobbiamo lavorare con una informazione incompleta, cioè con i dati che abbiamo a disposizione nel corso di una studio sperimentale.

Dobbiamo distinguere l'insieme dei dati ottenuti attraverso le osservazioni sperimentali (**campione**) e tutte le potenziali osservazioni che possono essere fatte in un certo contesto (**popolazione**). Ogni misura in un insieme di dati è originata da una distinta sorgente (**unità**) che può essere un paziente, una famiglia, un albero, una fattoria,.

Unità: una singola unità è di solito una persona o un oggetto le cui caratteristiche o variabilità sono di interesse per l'indagine.
Popolazione: è l'insieme di tutte le unità su cui si svolge l'indagine statistica. Essa può essere finita o infinita.
Campione: sottoinsieme finito di unità della popolazione che sono state raccolte nel corso dell'indagine. Il campione dovrebbe essere il più possibile "rappresentativo" per la popolazione. Per esempio se vogliamo conoscere le preferenze musicali di una certa città e consideriamo come campione l'insieme degli ascoltatori che chiamano alla radio per esprimere il loro gusti non è rappresentativo perché dipende dal tipo di musica trasmessa da quella radio. Inoltre gli ascoltatori che chiamano ed aspettano per prendere la linea sono delle persone determinate nelle loro opinioni, per cui il risultato dell'indagine sarà falsato dal tipo di campione scelto. Un campione rappresentativo potrebbe essere scelto formando dei numeri di telefono scegliendo a caso tra le cifre 0,1,2,…, 9. Il computer potrebbe simulare tale esperimento (generazione di numeri casuali). E' importante l' ampiezza del campione?
Certamente le abitudini di acquisto di una persona potrebbe non rappresentare quella dell'intera popolazione. Comunque, nonostante le diverse abitudini di acquisto degli individui, potremmo ottenere un'informazione accurata sulle abitudini dell'intera popolazione se prendiamo un campione "sufficientemente" ampio.
Raccogliere i dati per rispondere ad una particolare domanda, in modo da fornire le basi per decidere una azione o migliorare un processo. Bisognerebbe a tal fine stabilire di un obiettivo che sia specifico e non ambiguo.

Esempio. Ogni giorno una città deve testare l'acqua di un lago per determinare se l'acqua è sicura per nuotare. La difficoltà più grande è la crescita delle alghe e il limite di sicurezza è stato stabilito in termini di limpidezza dell' acqua.
L'obiettivo in questo caso è determinare se la limpidezza dell'acqua sulla spiaggia è o no al di sotto del limite di sicurezza.

Obiettivi della Statistica

- fare inferenza sulla popolazione dall'analisi delle informazioni che sono contenute nel campione dei dati. Questo include una stima dell'incertezza contenuta in tale inferenza,
- delineare il processo e l'ampiezza del campione in modo che le osservazioni formino una base per fare valide inferenze.

Statistica descrittiva

La Statistica Descrittiva si occupa di illustrare e sintetizzare i dati osservati.
Fasi:

- Raccolta dei dati
- Organizzazione dei dati in tabelle e grafici.
- Sintesi dei dati mediante gli indici sintetici in modo da descrivere le caratteristiche essenziali.

Variabile: X grandezza che varia all'interno di una popolazione. Essa può essere **numerica** se i valori che essa può assumere sono numeri, in particolare essa si dice **discreta** se l'insieme dei valori è finito o numerabile, p.e. numero di chiamate ad un centralino, e **continua** se è continuo, p.e. altezze di una popolazione. Se la variabile non è numerica si dice **categorica** o **qualitativa**, per esempio gruppi sanguinei.

Frequenza assoluta: f_i è il numero di volte in cui la variabile X assume il valore quantitativo o qualitativo x_i Se la variabile X assume i valori $x_1, x_2, \ldots, x_k,$ con frequenze rispettivamente $f_1, f_2, \ldots, f_k$ allora

$$\mathbf{f_1 + f_2 + \ldots + f_k = n}$$

dove n è il numero totale degli oggetti del campione.

Frequenza relativa: è il rapporto tra la frequenza assoluta e il numero degli elementi del campione, cioè

$$\mathbf{p_i = f_i / n}$$

$$\mathbf{p_1 + p_2 + \ldots + p_n = 1.}$$

Frequenza cumulata: p_{ci} è la somma delle frequenze relative fino alla p_i, cioè

$$\mathbf{p_{ci} = p_1 + p_2 + \ldots + p_i.}$$

Se la variabile è categorica o numerica discreta con un numero di valori non molto alto allora si potrebbe costruire una tabella delle frequenze in cui a nella prima colonna ci sono i valori assunti dalla variabile nella seconda colonna le corrispondenti frequenze assolute e nella terza le frequenze relative. Nella quarta la frequenza cumulata.

Esempio 1
(Variabile categorica): gruppi sanguigni di un campione estratto da una certa popolazione:

gruppo	Frequenza assoluta
A	60
B	16
AB	7
O	66

TOTALE 149

Esempio 2

(variabile numerica discreta): Conteggio del numero di errori di stampa per pagina (variabile discreta) riscontrati su un testo di 45 pagine:

5 6 3 4 7 2 3 2 3 2 6 4 3 9 3
2 0 3 3 4 6 5 4 2 3 6 7 3 4 2
5 1 3 4 3 7 0 2 1 3 1 5 0 4 5

Distribuzione di frequenze assolute e relative (arrotondate) degli errori delle 45 pagine:

classe(x_i)	freq.ass.f_i	freq.rel.p_i	freq.cumul. p_{ci}
0	3	0,0667	0,0667
1	3	0,0667	0,1333
2	7	0,1556	0,2889
3	12	0,2667	0,5556
4	7	0,1556	0,7111
5	5	0,1111	0,8222
6	4	0,0889	0,9111
7	3	0,0667	0,9778
8	0	0,0000	0,9778
9	1	0,0222	1,0000
	45	1	

Distribuzione in classi per dati continui

Consideriamo la distribuzione di frequenza per dati che contengono misure su una virtuale scala continua. Naturalmente le misure sono sempre arrotondate. A differenza del caso discreto l'insime di misure su una scala continua potrebbe contenere molti valori distinti. In tal caso una tavola con tutti i valori distinti con le relative frequenze non darà una descrizione sintetica dei dati. In tal caso è più conveniente raggruppare le osservazioni in classi calcolando per ogni classe la relativa frequenza. A differenza della distribuzione discreta in cui i dati vengono raggruppati per ogni valore a cui viene assegnata la relativa frequenza, nel caso continuo si considera intervalli o classi di valori.
I passi da fare per costruire una distribuzione di frequenza sono i seguenti:

- individuare il valore minimo e quello massimo nell'insieme dei dati;
- scegliere intervalli di uguale lunghezza che ricoprono tutti i dati dal minimo al massimo senza sovrapposizione. Tali intervalli sono chiamati **classi.** Per far ciò si potrebbe calcolare l' intervallo di variazione, cioè la differenza tra il massimo e il minimo. Per convenienza si potrebbe allargare tale intervallo di variazione, tenendo conto che i suoi estremi non dovrebbero discostarsi troppo dai valori minimo e massimo. Il numero di classi è stabilito sulla base di $\sqrt{n}$ *(naturalmente si prende la parte intera)*. Per ottenere l' ampiezza di una classe si divide l'intervallo di variazione per il numero di classi.
- Si calcola il numero di dati che appartengono a ciascuna classe. Tale numero è la **frequenza della classe.**
- Con la tabella distribuzione di frequenza si perde l'informazione di come sono distribuiti i dati all'interno di ciascuna classe. Per questo si prende come punto di riferimento il valore centrale di ciascuna classe.

N.B.: le distribuzioni delle frequenze relative o percentuali sono indispensabili quando si confrontano due o più gruppi di misure, che presentano un diverso numero di osservazioni.

Osservazione: Raggruppando i dati si perde informazioni riguardo la distribuzione dei dati all'interno di ogni intervallo. Scegliendo un numero di classi troppo basso si rischia di sintetizzare troppo i dati perdendo informazioni sui dati. D'altra parte con un numero di classi troppo elevato rispetto al numero dei dati, le frequenze potrebbero variare in modo caotico e non sarebbe possibile riconoscere un certo andamento della distribuzione, a causa della eccessiva dispersione dei dati.

Osservazione: Quando il numero di dati è molto alto conviene calcolare la distribuzione di frequenza utilizzando il computer.

Esempio 3

(variabili continue): Raggruppamento in classi di una variabile continua: altezza in centimetri di 40 piante 111-119-130-170-143-156-126-113-127-107-83-100-128-143-127-117-125-64-119-130-120-108-95-192-124-129-143-198-131-163-152-104-119-161 178-135-146-158-176-98

Procedura:

- individuare il valore minimo (64) e massimo (198)
- stabilire l' intervallo di variazione, cioè un intervallo che comprenda tutti i dati, i cui estremi non si discostino troppo dai valori minimo e massimo, per esempio (60, 200), la cui ampiezza è 200-60=140.
- sulla base di $\sqrt{n}$ (circa 7) si decide il numero di classi (con ampiezza 20=140/7).

class	val.centx$_i$	freq.as-sol.f$_i$	freq.rel. p$_i$	freq.cumulata
60-80	70	1	0,025	0,025
80-100	90	3	0,075	0,1
100-120	110	10	0,25	0,35
120-140	130	12	0,3	0,65
140-160	150	7	0,175	0,825
160-180	170	5	0,125	0,95
180-200	190	2	0,05	1
		40	1	

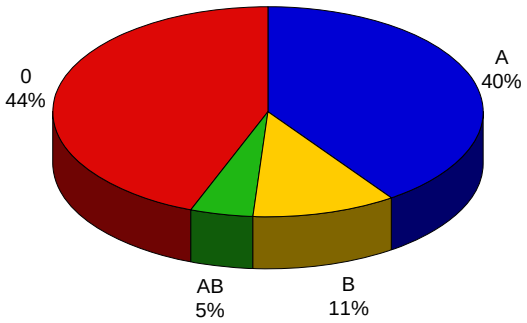
Grafici di distribuzioni di frequenze

Le distribuzioni contenute nella distribuzione di frequenza possono essere rappresentate graficamente. Ciò permette di sintetizzare i dati.

Diagrammi per dati qualitativi

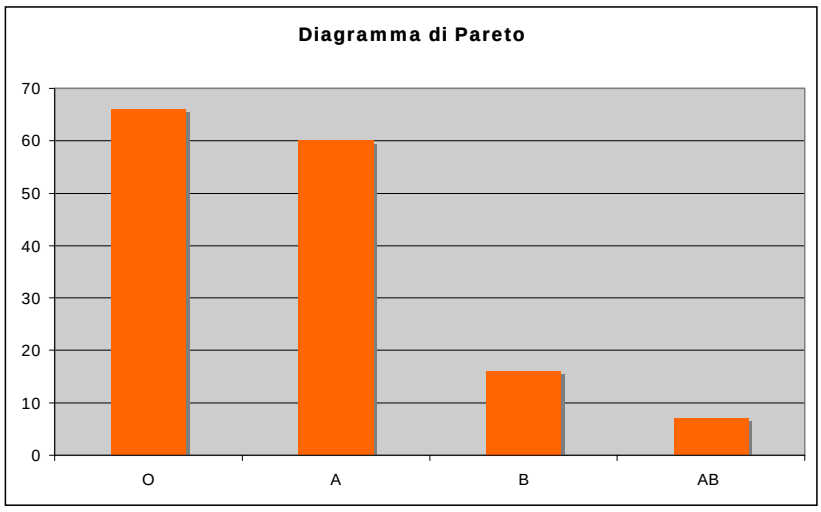
Aerogrammi a torta

Serve per rappresentare soprattutto le variabili categoriche in cui ogni "fetta" o settore di cerchio rappresenta la frequenza relativa della categoria. L'area di ciascun settore è proporzionale alla frequenza:
a_i:360°= f_i : n



Diagrammi di Pareto

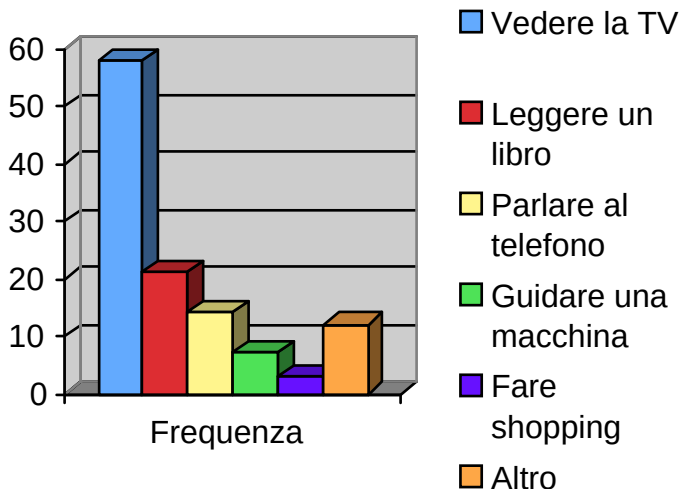
Secondo la legge empirica di Pareto fu formulata da Joseph M. Juran, ma che è nota anche con il nome di principio di Pareto, e che è sintetizzabile nell'affermazione: la maggior parte degli effetti è dovuta ad un numero ristretto di cause. Nei diagrammi di Pareto si ordinate categorie da quella con frequenza più alta e via via fino a quella più bassa. Sull' asse delle ascisse ci sono le caratteristiche della variabile qualitativa mentre sull' asse delle ordinate ci sono le frequenze assolute delle varie categorie. Nel caso dei gruppi sanguinei si ha:



Dal grafico si vede che 66 individui hanno il gruppo O e 66+60=126 su 149 il gruppo O e A, cioè l'84%. (frequenza cumulata).

Esempio. Un gruppo di studenti che frequentano un corso di psichiatria sono stati interrogati sulla abitudine che dovrebbe essere migliorata. Per ridurre l' effetto di tale abitudine dovrebbero raccogliere dati sulla frequenza e la circostanza in cui si manifesta. Uno studente ha raccolto le seguenti frequenze relativamente al vizio di mangiarsi le unghie in un periodo di due settimane:

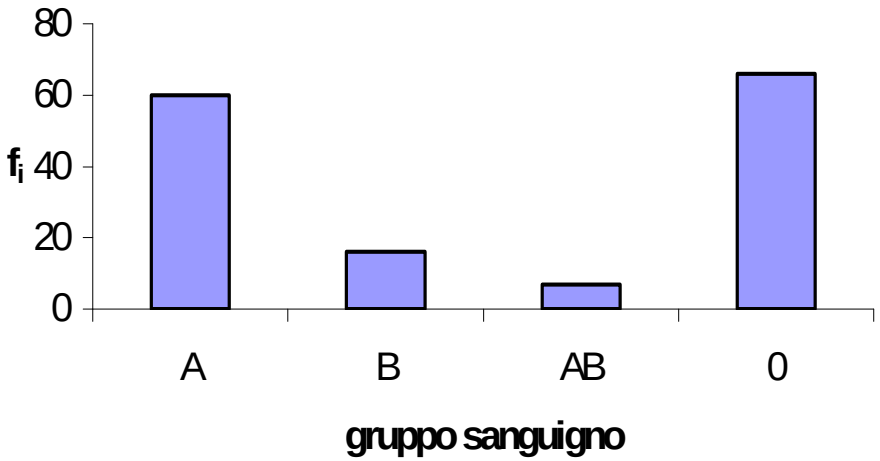
Attività	Frequenza
Vedere la TV	58
Leggere un libro	21
Parlare al telefono	14
Guidare una macchina	7
Fare shopping	3
Altro	12



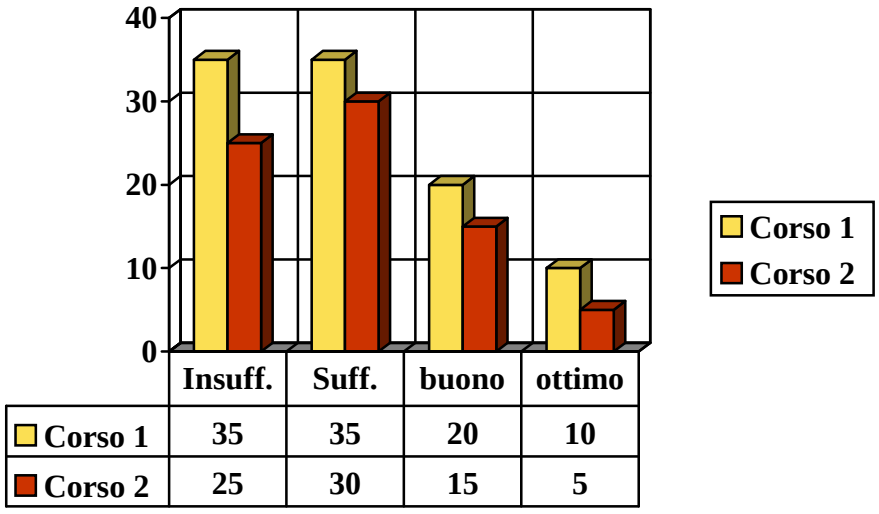
Il diagramma di Pareto mostra la relazione tra le varie attività e l mangiarsi le unghie. Guardare la TV rappresenta il 50,4% del totale.

Diagrammi a barre e istogrammi

I diagrammi a barre (o a linee) vengono utilizzati sia per rappresentare distribuzioni di variabili categoriche, sia variabili numeriche discrete o numeriche con pochi valori. Ad ogni classe corrisponde una barra (o linea) la cui ampiezza della base (per tutte uguali) non ha significato, mentre l'altezza rappresenta la frequenza (assoluta o relativa) della classe. Se le barre sono adiacenti allora si ha un' istogramma, e l'ordine delle barre ha significato nel caso in cui i valori della variabile si possono ordinare. Consideriamo l'esempio 1. Il relativo diagramma a barre è il seguente:



Il seguente diagramma rappresenta i risultati di un esame per due corsi distinti:



	Insuff.	Suff.	buono	ottimo
Corso 1	35	35	20	10
Corso 2	25	30	15	5

Oppure si potrebbe rappresentare nel seguente modo:

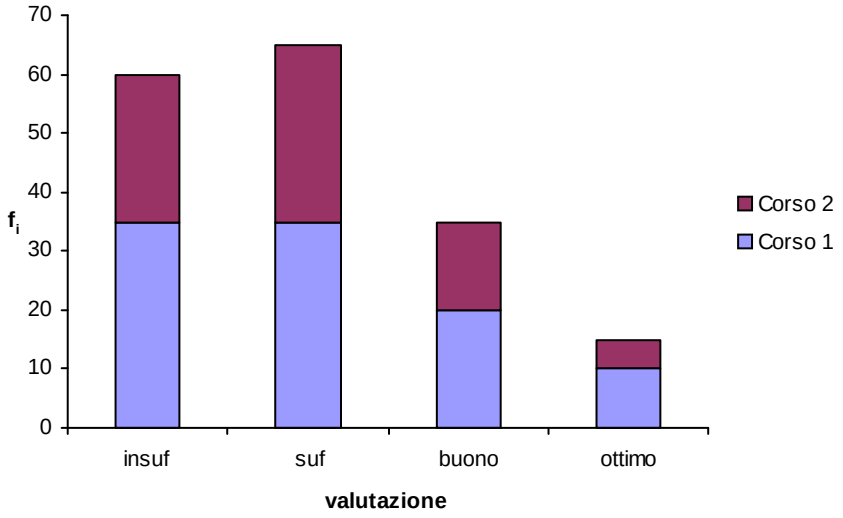
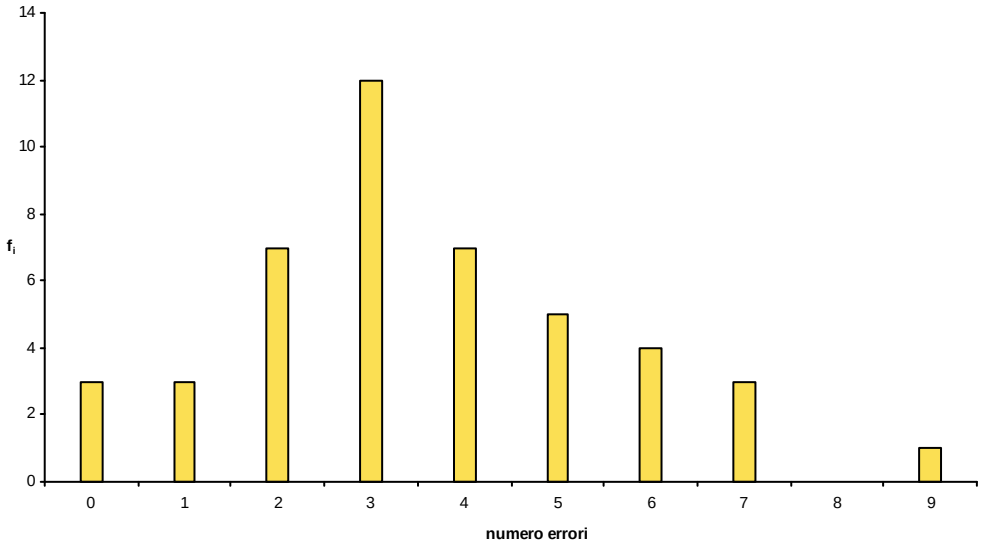
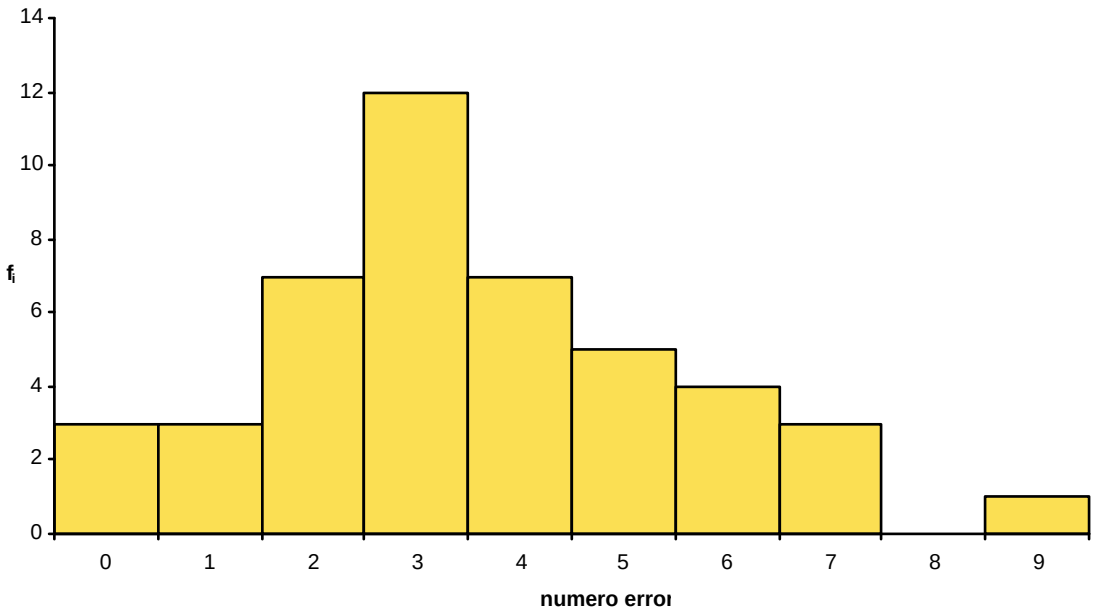


Diagramma relativo all'esempio 2:



oppure si potrebbe rappresentare come istogramma:

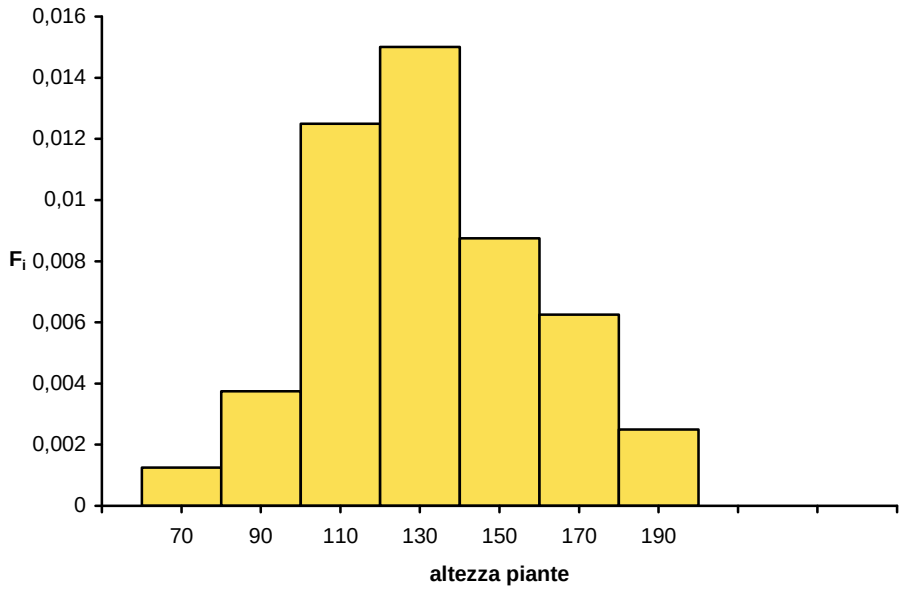


Istogrammi

L'istogramma è costituito mediante rettangoli adiacenti, le cui basi sono gli intervalli che definiscono le classi, e le altezze rappresentano le densità di frequenza, cioè

- F_i = frequenza relativa della classe i / ampiezza classe = p_i / ampiezza
- Le classi sono supposte di uguale ampiezza.

L’istogramma relativo all’esempio 3 è:



L’area totale dell’istogramma è 1. Infatti l’area di ogni rettangolo è proprio la frequenza relativa della classe e l’area dell’istogramma è la somma delle aree di ciascun rettangolo che lo costituisce, quindi la somma delle frequenze relative, che è uguale a 1.

Nel caso considerato tutte le classi hanno la stessa ampiezza per cui la densità di frequenza risulta proporzionale alla frequenza. In tal caso si potrebbe considerare la frequenza al posto della densità di frequenza.

Indici Sintetici dei dati

Essi danno delle informazioni quantitative sull'ordine di grandezza delle osservazioni (misure di posizione), sulla variabilità delle osservazioni (misure di dispersione, misure di forma). Indichiamo i dati osservati con i seguenti simboli: $x_1, x_2, \ldots, x_n$.
La misura di posizione localizza il valore centrale di una distribuzione di frequenza. Le più comuni sono la media, la mediana e la moda.

Nel caso di dati continui si prende come valore x_i, i=1,2,…,n, il valore centrale della classe i-esima.

Misure di posizione

Media

- per dati semplici:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

- per dati ponderati:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} x_i = \frac{1}{n}(x_1 f_1 + x_2 f_2 + \ldots + x_r f_r)$$

Osservazione: la media aritmetica sarà indicata con μ quando si riferisce alla popolazione, quando si riferisce al campione sarà indicata con $\bar{x}$

Nella Tabella 2 la media è:

$$\bar{x} = \frac{1}{45}(0 \cdot 3 + 1 \cdot 3 + 2 \cdot 7 + 3 \cdot 12 + 4 \cdot 7 + 5 \cdot 5 + 6 \cdot 4 + 7 \cdot 3 + 8 \cdot 0 + 9 \cdot 1) = 3.6$$

Nella Tabella 3 la media è:

$$\bar{x} = \frac{1}{40}(1 \cdot 70 + 3 \cdot 90 + 10 \cdot 110 + 12 \cdot 130 + 7 \cdot 150 + 5 \cdot 170 + 2 \cdot 190) = 132$$

Mediana

- E' il valore che occupa la posizione centrale in un insieme ordinato di dati.
- Non è influenzata dai valori estremi.
- Si usa per attenuare l'effetto dei valori estremi molto alti o molto bassi
- Per calcolare la mediana bisogna ordinare i valori.
- Se il campione ha un numero dispari di valori la mediana è il valore che occupa la posizione centrale.

Per esempio la mediana di 1, 4, 6, 7, 8 è 6. Se il campione ha un numero pari di valori, la mediana è il valor medio dei due valori che occupano la posizione centrale. Per esempio la mediana di 1, 4, 6, 7, 8, 9 è (6+7)/2=6.5.

Confronto tra media e mediana

Esempio: Il numero di giorni di sopravvivenza per i primi sei pazienti che hanno avuto un trapianto di cuore a Stanford sono stati: 15, 3, 46, 623, 126, 64.

Il valor medio è

$$\bar{x} = \frac{15+3+46+623+126+64}{6} = \frac{877}{6} = 146{,}2$$

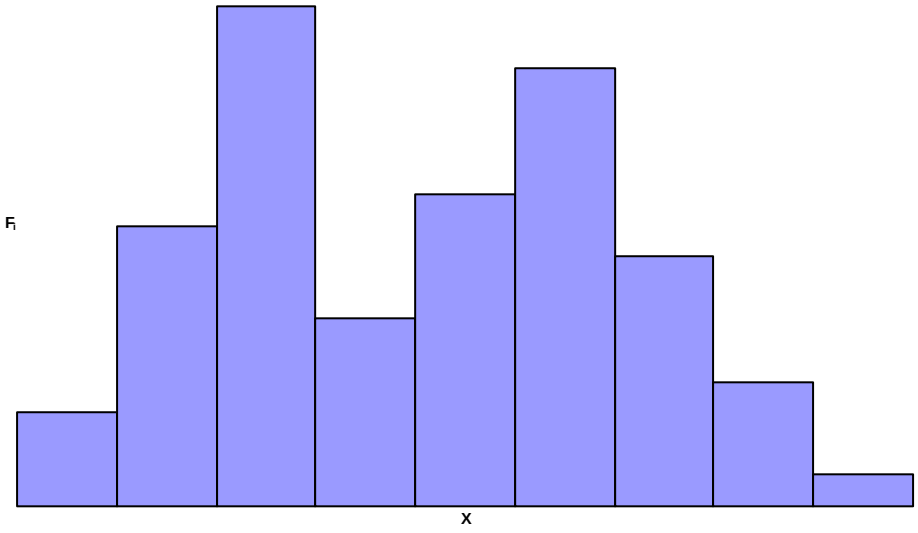
La mediana e' invece (46+64)/2 =55. Infatti i valori centrali della distribuzione ordinata sono 46 e 64 il cui valore medio è 55:

3, 15, 46, 64, 123, 623
Notiamo che il valore molto alto 623 influenza la media, che risulta significativamente più grande della mediana. In tal caso la mediana è un indicatore di posizione migliore della media.

L'esempio sopra dimostra che la mediana non è affetta da poche osservazioni molto basse o molto alte, mentre la presenza ti tali estremi potrebbe avere un effetto considerevole sulla media. Per distribuzioni molto asimmetriche la mediana risulta essere una misura di posizione più sensibile misura del centro della distribuzione rispetto che la media.

Moda

- E' il valore più frequente di una distribuzione
- Nelle distribuzioni di frequenza per dati raggruppati essa è molto sensibile alla modalità di costruzione delle classi.
- Le distribuzioni di frequenza unimodali hanno un'unica moda, quelle bimodali hanno anche mode secondarie.
- Nell'esempio 2 la moda è 3; nell'esempio 3 è il valore centrale 130.



Misure di dispersione

Due insiemi di dati che hanno dei valori centrali confrontabili potrebbero avere una variabilità molto diversa tra loro. Per studiare una distribuzione di dati le misure di posizione non bastano, ma è necessario anche analizzare come i dati variano rispetto al valore centrale.

Per definire la varianza introduciamo prima la deviazione dalla media:
$D_i = x_i - \bar{x},$
$i = 1,2,...,n$

Naturalmente la somma delle deviazioni dalla media è nulla: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Varianza

Per ottenere una misura di variabilità dei dati dovremmo considerare la deviazione senza segno perchè le parti negative potrebbero compensare quelle negative, e facciamo il quadrato. Otteniamo in tal modo la varianza:

Altro modo per calcolare la varianza
Per i dati semplici:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Per i dati ponderati:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot x_i^2 - \bar{x}^2$$

Esempio: Calcolare la varianza nei due modi descritti dei valori 5, 6, 7, 7, 8, 10

$$\bar{x} = \frac{5+6+7+7+8+10}{6} = \frac{43}{6} = 7.17$$

Allora si ha:

I metodo: $s^2 = \frac{1}{6}\left[(5-7.17)^2 + (6-7.17)^2 + (7-7.17)^2 + (7-7.17)^2 + (8-7.17)^2 + (10-7.17)^2\right] = 2.47$

II metodo: $s^2 = \frac{1}{6}\left[5^2 + 6^2 + 7^2 + 7^2 + 8^2 + 10^2\right] - 7.17^2 = 2.42$

Nell' esempio 2 la varianza calcolata con il secondo metodo è:

$$s^2 = \frac{1}{45}\left(0^2 \cdot 3 + 1^2 \cdot 3 + 2^2 \cdot 7 + 3^2 \cdot 12 + 4^2 \cdot 7 + 5^2 \cdot 5 + 6^2 \cdot 4 + 7^2 \cdot 3 + 8^2 \cdot 0 + 9^2 \cdot 1\right) - 3{,}6^2 = 3{,}7$$

Nell' esempio 3 la varianza calcolata con il primo metodo è:

$$s^2 = \frac{1}{40}\left[(70-132)^2 \cdot 1 + (90-132)^2 \cdot 3 + (110-132)^2 \cdot 10 + (130-132)^2 \cdot 12 + (150-132)^2 \cdot 7 + (170-132)^2 \cdot 5 + (190-132)^2 \cdot 2\right] = 756$$

Osservazione: nella statistica inferenziale, cioè quando si utilizzano i dati del campione per stimare le caratteristiche di una popolazione, si usa sempre la VARIANZA CAMPIONARIA con la correzione di Student:

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2$$

n-1: n° di osservazioni indipendenti, è chiamato GRADI DI LIBERTÀ : poiché la somma degli scarti dalla media è uguale a zero, l'ultimo valore è fissato a priori e non è libero di assumere qualsiasi valore

Deviazione standard (o Scarto quadratico medio) del campione

E' la radice quadrata della varianza

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

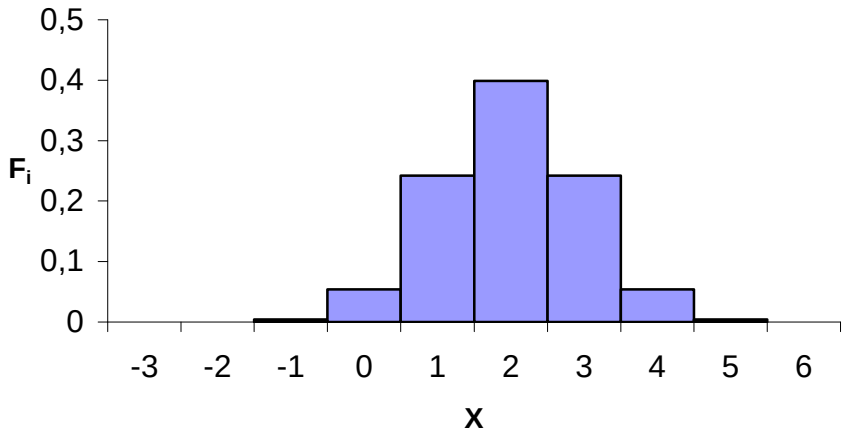
Analogamente per la **deviazione standard campionaria con la correzione di Student** si ha

$$s' = \sqrt{s'^2}$$

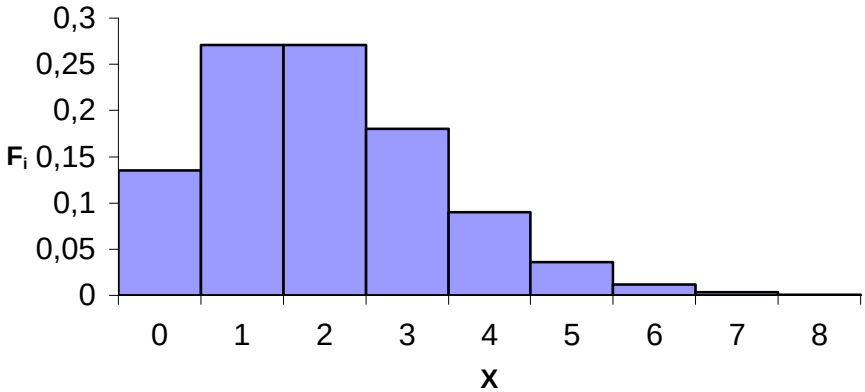
Misure di Forma

Servono per quantizzare due caratteristiche di una distribuzione di frequenza: **Simmetria**.

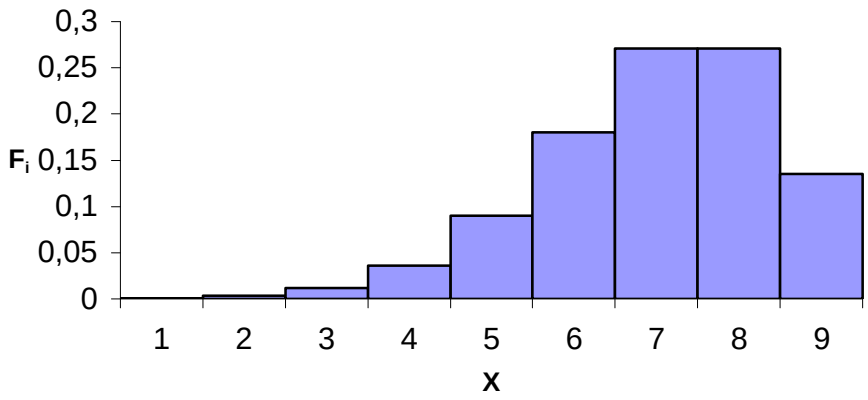
Si ha **Simmetria** se in una distribuzione di frequenza i valori equidistanti dalla media (o mediana) presentano la stessa frequenza. Si ha **Asimmetria Positiva** quando la distribuzione ha una coda verso destra e **Asimmetria Negativa** quando ha una coda verso sinistra.



Esempio di distribuzione simmetrica



Esempio di distribuzione con asimmetria positiva



Esempio di distribuzione con asimmetria negativa

Indice di asimmetria (skewness)

$$\frac{1}{ns^3}\sum_{i=1}^{k}(x_i-\bar{x})^3\cdot f_i$$

- > 0 coda a destra (simmetrica positivamente)
- < 0 coda a sinistra (simmetrica negativamente)
- **= 0 (quasi)** simmetrica

(s è la deviazione standard)

Applicazioni con Excel

Analisi dei dati

Analisi dei dati: è un'applicazione di Excel che permette di fare un'analisi statistica dei dati forniti in input insieme a dei parametri, attraverso delle funzioni macro statistiche. I risultati sono visualizzati sottoforma di tabelle o grafici.
Come aprire l'applicazione: scegliere Analisi dei dati dal menu Strumenti. Se non compare selezionare Analisi dei dati dall'opzione Componenti aggiuntive.

Statistica descrittiva. Scegliendo dalla finestra di dialogo l'opzione Statistica Descrittiva viene fatta una analisi statistica dei dati presi da una colonna di una tabella di Excel (vedi istruzioni per costruire un foglio elettronico).
Opzioni:

- Intervallo di input: dare le coordinate dei dati da analizzare, per esempio A1:A15, cioè tutti i dati contenuti nelle celle della colonna A dalla prima alla quindicesima riga.
- Intervallo di output: dare le coordinate della cella superiore sinistra della tabella di output.

Riepilogo statistiche: genera una tabella di output con le statistiche: Media, Errore standard (della media), Mediana, Moda, Deviazione Standard, Varianza, Curtosi, Simmetria, Intervallo, Min, Max, Somma, Conteggio

Istogramma

L'opzione Istogramma raggruppa un insieme di dati in classi
e calcola le frequenze assolute, e quelle cumulate.
Opzioni dalla finestra di dialogo:

- Intervallo di input: dare le coordinate dei dati da analizzare.
- Intervallo di classe (facoltativo): digitare gli indirizzi delle celle che contengono i valori che separano gli intervalli delle classi (per esempio B1:B5). Questi punti di separazione degli intervalli (o classi) devono essere in ordine crescente. Nel calcolo della frequenza viene compreso l'estremo destro dell'intervallo. Se non viene dato le classi vengono create automaticamente. Se la variabile è discreta allora per disegnare il grafico bisogna dare intervalli di ampiezza 1.
- Intervallo di output: dare le coordinate della cella superiore sinistra della tabella di output.
- Etichetta: tale opzione viene scelta quando la prima riga della colonna data è stata utilizzata per un titolo della colonna.
- Pareto (istogramma ordinato): scegliendo questa casella gli intervalli saranno ordinati in funzione delle frequenze prima di generare il diagramma.
- Percentuale cumulativa: fornisce il grafico della frequenza relativa cumulativa.
- Grafico in output: fornisce il grafico delle frequenze.

Per attaccare i rettangoli dell'istogramma fare doppio clic su una delle barre del grafico; nella scheda Opzioni della finestra di dialogo Formato serie dati impostare a 0 la Distanza tra le barre. Oppure si potrebbe cambiare il colore o scrivere i valori su ogni rettangolo, o la tabella dei dati. Fare clic su OK.

Calcolo delle probabilità

Introduzione

Sebbene la conoscenza completa sulla popolazione statistica rimanga lo scopo principale dell' analisi statistica, in genere abbiamo a disposizione solo la informazione parziale contenuta nel campione. Con la Statistica descrittiva abbiamo potuto descrivere i dati a disposizione attraverso il raggruppamento degli stessi, la loro rappresentazione grafica e il calcolo di indici sintetici quali la media, la mediana, la moda, la varianza e la deviazione standard. La nostra meta è di ottenere generalizzazioni o inferenza sulla popolazione sulla base delle informazioni contenute nel campione. Per capire il ragionamento che conduce a tale generalizzazione è essenziale introdurre il Calcolo della Probabilità.
Nel linguaggio comune sentiamo spesso delle frasi del tipo

- molto "probabilmente" il prossimo fine settimana il tempo sarà buono
- è "abbastanza improbabile" che riesca a vincere alla lotteria
- ho "50% di probabilità" di ottenere quel posto

Le parole "molto probabile", "abbastanza improbabile", indicano qualitativamente la possibilità che un evento accada. Il Calcolo delle Probabilità è una branca della matematica fornisce un metodo per quantizzare l'incertezza. In generale la probabilità di un evento è un valore numerico che misura quanto è verosimile che un evento accada. Assegniamo alla probabilità una scala che va da 0 a 1, dove un valore molto basso indica estremamente improbabile, mentre un valore prossimo a 1 molto probabile. Nella parte finale sulla inferenza statistica vedremo come il Calcolo delle Probabilità rappresenti uno strumento essenziale.

Il termine **esperimento** è usato in questo contesto non solo per studi condotti nei laboratori, ma più in generale per osservazioni di un fenomeno che presenta variabilità nei suoi possibili risultati.
Lo **Spazio campionario** è l'insieme di tutti i possibili esiti di un esperimento. Esso può essere **discreto** (Es.: Lancio di due dadi) o **continuo** (Es.: Peso). Ciascun punto o risultato di S sarà chiamato **evento elementare** e sarà denotato con la lettera *e*.
Un **evento** è un sottoinsieme di S cioè è un esperimento che a priori può avere diversi esiti, non prevedibili con certezza. Un evento accade quando uno dei suoi possibili risultati si verifica.
In particolare *S* è detto evento certo e Ø (insieme vuoto) è detto evento impossibile .

Definizione operativa di probabilità

Probabilità classica:
Dato uno spazio campionario S = $\{e_1, e_{2,\ldots,} e_N\}$ e supponiamo che tutti i risultati dell' esperimento siano simmetrici, cioè

$$P(e_1)=P(e_2)=\ldots=P(e_N)=1/N$$

in modo che

$$P(e_1)+P(e_2)+\ldots+P(e_N)=1$$

In generale, dato un evento A

$$P(A)=\frac{|A|}{N}$$

dove $|A|$ è il numero di elementi di A .
Nella definizione classica ogni evento è simmetrico, cioè non c'è nessuna ragione per supporre che un evento sia più probabile di un altro. Per esempio nel lancio di un dado, o in altri giochi d'azzardo si può costruire un modello simile. La probabilità classica è definita come il rapporto tra i casi favorevoli e tutti i possibili casi.

Esempio: nel lancio di una moneta "onesta" i possibili risultati sono "Testa", T e "Croce", C , cioè $S = \{T, C\}$ e $P(T)=P(C) = \frac{1}{2}$.

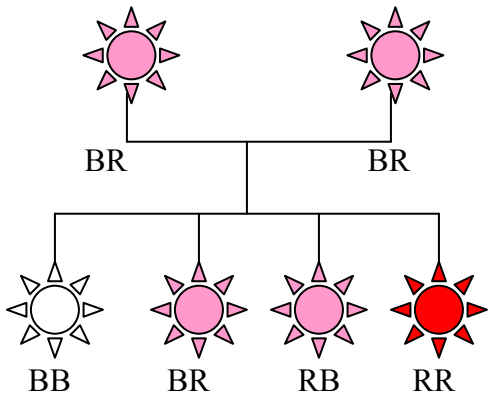
Esempio: nel lancio di un dado i possibili risultati sono $S = \{1,2,3,4,5,6\}$ e $P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=\frac{1}{6}$.

Sia A l'evento "Il risultato del lancio è un numero pari", allora $P(A)=\frac{1}{2}$.
Per tale definizione ci sono dei limiti: non tutti i fenomeni possono essere descritti da tale modello simmetrico.

Principio di enumerazione. Dati due possibili esperimenti di cui il primo ha n possibili risultati e il secondo m possibili risultati, allora se si considerano entrambi gli esperimenti contemporaneamente complessivamente ci sono $m \cdot n$ possibili esiti.

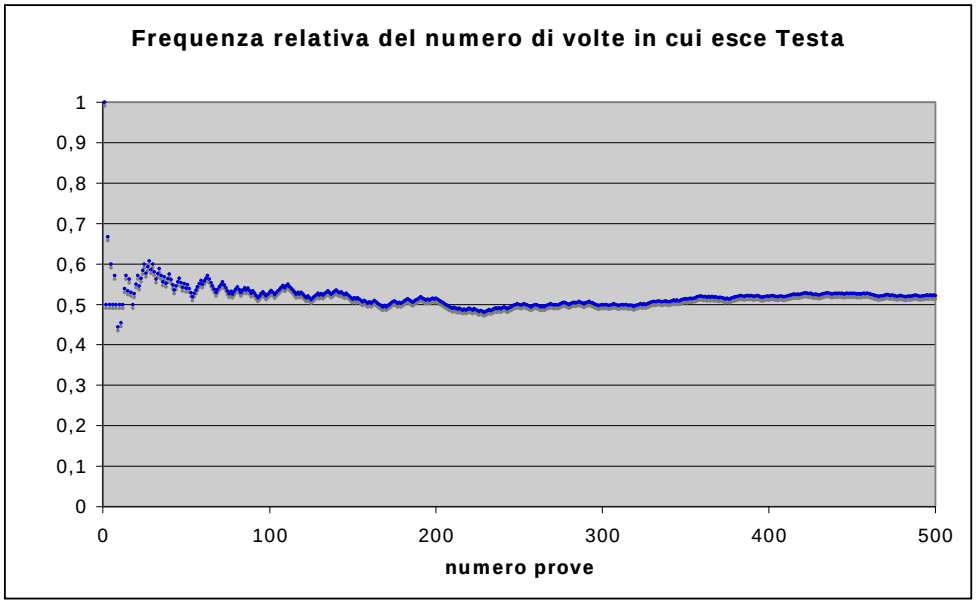
Esempio: in due lanci di un dado ci sono 6x6=36 possibili risultati. La probabilità che esca la coppia (2,5) è 1/36.

Esempio 5: Gorge Mendel, genetista pionieristico, ha scoperto un modo di caratterizzare le generazioni di piante di pisello e ideato una teoria di ereditarietà che spiega tale meccanismo. Secondo la legge di Mendel i caratteri ereditari vengono trasmessi da una generazione all' altra tramite i geni. I geni sono a coppie composte da un gene di ogni genitore. Un semplice modello uniforme sta alla base della spiegazione del meccanismo di selezione di Mendel. Un esperimento che illustra la teoria di Mendel consiste nell'incrocio tra una specie pura di fiori rossi (R) con una specie pura di fiori bianchi (B). Ciò produce ibridi aventi un gene di ogni tipo e sono fiori rosa. Incrociando tali ibridi si può ottenere una delle quattro possibili coppie di geni. Secondo la legge di Mendel, queste quattro coppie hanno la stessa probabilità. Quindi P(Rosa)= ½ e P(Rosso)=P(Bianco)= ¼



Un esperimento fatto da uno studente di Mendel, Correns mostra che le frequenze dei fiori bianchi, rosa e rossi sono rispettivamente 141, 291 e 132, che sono vicine alle proporzioni 1:2:1.

II. Probabilità frequentistica: si suppone di ripetere un esperimento alle stesse condizione, al crescere del numero delle prove secondo una legge empirica, la frequenza relativa in cui compare un certo evento tende a "stabilizzarsi" intorno ad un valore che definiamo come probabilità dell'evento. Per esempio il numero di volte in cui esce testa in n successivi lanci di una moneta. Anche tale definizione è limitative: per esempio non tutti i fenomeni sono ripetibili.



III. Probabilità soggettivistica: una altra definizione è quella soggettivistica secondo cui la probabilità di un evento è pari al grado di fiducia che l'osservatore ha nel verificarsi o meno di esso. Per esempio nelle scommesse si applica tale definizione.

IV. Probabilità assiomatica: La probabilità di un evento è un numero reale compreso tra 0 e 1 avente le seguenti proprietà (assiomi):

1. La probabilità dell'evento certo è 1 : $P(S) = 1$
2. La probabilità di un qualunque evento è compresa tra 0 e 1 : $P(A) \geq 0$ per ogni $A \subseteq S$.
3. La probabilità dell'unione di due eventi incompatibili A_1, A_2 è uguale alla somma delle probabilità di ciascun evento:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2),$$
$$A_1 \cap A_2 = \varnothing$$

Esempio 1. Nel lancio di un dado, $S = \{1,2,3,4,5,6\}$, un evento elementare è un punto di S, un evento è per esempio A = "Il risultato del lancio è un numero pari", cioè $A = \{2,4,6\}$.

Eventi composti

La negazione o complementare di un evento A ($A \subseteq S$) è l'evento che si verifica quando non si verifica A. Esso si indicherà con A^C . Allora A^C = S - A .

L' **intersezione** di due eventi A_1, A_2, è l'evento che si verifica quando si verificano contemporaneamente A_1 e A_2. Essa si indicherà con $A_1 \cap A_2$.

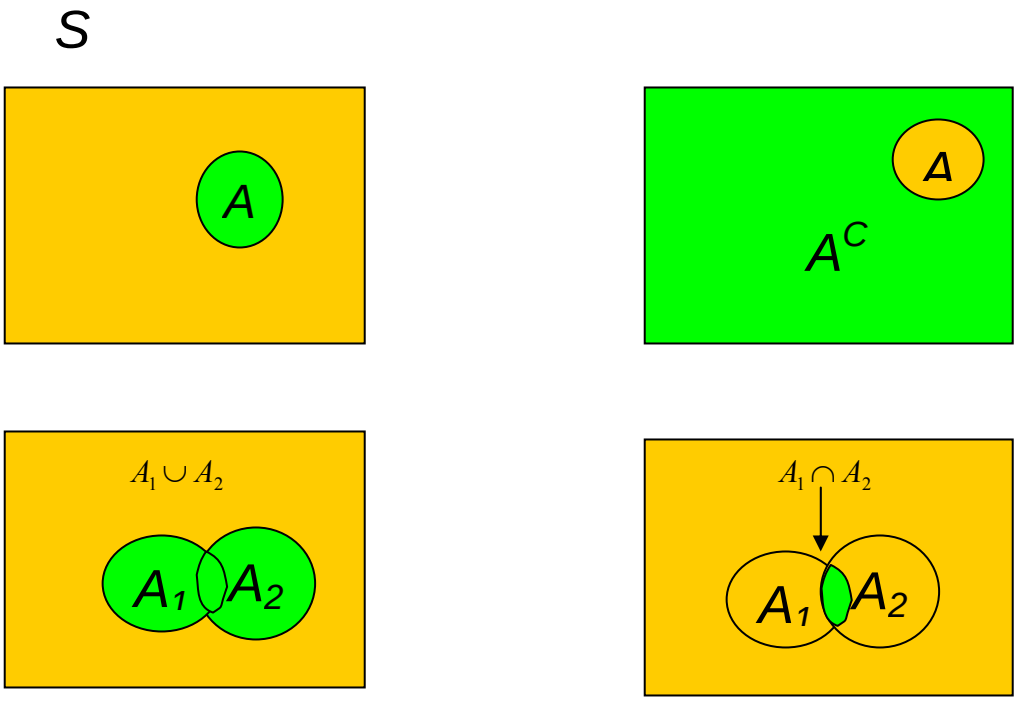
L' **unione** di due eventi A_1, A_2, è l'evento che si verifica quando si verifica A_1 o A_2. Essa si indicherà con $A_1 \cup A_2$.

Esempio 2. Sia A_1 l'evento "Il Signor Rossi prende il primo treno" e A_2 l'evento "Il Signor Bianchi prende il primo treno". Allora:

- l'evento A_1^C = "Il Signor Rossi perde il primo treno";
- l'evento $A_1 \cap A_2$ = "Il Signori Rossi e Bianchi prendono il primo treno",
- l'evento $A_1 \cup A_2$ = "Il Signor Rossi o il Signor Bianchi prendono il primo treno";
- l'evento $A_1^C \cap A_2^C$ = "Né il Signor Rossi né il Signor Bianchi prendono il primo treno";

Spesso per rappresentare gli eventi composti e lo spazio campionario vengono utilizzati i diagrammi di Venn.

Diagrammi di Venn



Proprietà della Probabilità

- $P(A) = \sum_{e \in A} P(e)$
- $P(A^C) = 1-P(A)$ per ogni $A \subseteq S$.
- $P(\varnothing) = 0$;
- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2), \qquad A_1, A_2 \subseteq S$

Definizione: Due eventi A_1 e A_2 si dicono **incompatibili** se $A_1 \cap A_2 = \varnothing$. In tal caso si ha che $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ per ogni evento A_1 e A_2 (III assioma).

Probabilità condizionata

Dato un evento A, il verificarsi di un ulteriore evento B potrebbe fornire maggiore conoscenza riguardo l'evento A e potrebbe quindi modificare la probabilità dell'evento A, P(A). Tale probabilità si chiama **probabilità condizionata** e si indica con P(A|B) (probabilità di A sapendo che si è verificato l'evento B). Ad esempio supponiamo di estrarre una carta da un mazzo di 40 (evento A). La probabilità di indovinare tale carta P(A)=1/40. Supponiamo che chi ha estratto la carta suggerisca che la carta estratta sia una carta di coppe. Allora, tenendo conto di tale suggerimento, la probabilità di indovinare la carta aumenta e diventa P(A|B)=1/10 perchè, sapendo che la carta estratta è di coppe, automaticamente scartiamo tutte le altre carte e il nostro spazio campionario si è ridotto alle sole carte di coppe, cioè 10, per cui la probabilità di estrarre una carta di coppe è diventata 1/10.

Probabilità condizionata: Dato uno spazio di probabilità *S* ed un evento B tale che P(B) >0, si definisce probabilità condizionata dell'evento A, sapendo che si è verificato l'evento B

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

Se il verificarsi dell' evento B non modifica la probabilità di A allora i due eventi si dicono **indipendenti,** altrimenti **dipendenti.**
Definizione: Dato uno spazio di probabilità, due eventi A e B di esso si dicono **indipendenti** se

$$P(A \mid B) = P(A) \qquad (P(B)>0)$$

o, equivalentemente

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

Esercizio. Supponiamo di avere un gruppo di dieci persone, delle quali 4 indossano un impermeabile e 5 hanno un ombrello. Siano 3 le persone che hanno entrambi (impermeabile ed ombrello).
Quale la probabilità che una persona con l'ombrello abbia anche l'impermeabile?

Svolgimento:
Invece di considerare il gruppo di dieci persone consideriamo il sottogruppo delle persone con l'ombrello, che è composto da 5 elementi, di cu 3 hanno anche l'ombrello. Se indichiamo con A l'evento "la persona ha l'impermeabile" e con B l'evento `"la persona ha l'ombrello", otteniamo:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{3}{5}$$

Gli eventi A e B dell'esempio sopra sono dipendenti. Infatti P(A)= 4/10 e P(B)= 5/10=1/2 e

$$P(A \cap B) = \frac{3}{10} \neq P(A) \cdot P(B) = \frac{2}{10}$$

Esempio. Un campione composto di ampiezza 200 è stato classificato secondo il peso corporeo e l' incidenza di ipertensione. I risultati sono stati riportati nella seguente tabella di contingenza

	Sovrappeso	Peso normale	Sottopeso	Totale
Ipertesi	20	16	4	40
Non ipertesi	30	90	40	160
Totale	50	116	44	200

a) Quale la probabilità che una persona scelta a caso da questo gruppo sia ipertesa?
b) Una persona scelta a caso da questo gruppo è stata trovata soprappeso. Qual è la probabilità che sia anche ipertesa?

Svolgimento:
Definiamo i seguenti eventi
A:La persona estratta è ipertesa
B:"La persona estratta è sovrappeso"
a) La probabilità di scegliere una persona ipertesa (evento A) è il totale della riga corrispondente agli ipertesi sul totale del campione: P(A)=(20+16+4)/200=0,2.
b) La probabilità richiesta si calcola considerando lo spazio campionario ristretto solo alle persone soprappeso del campione (50). La probabilità di trovare una persona ipertesa all'interno di questo sottogruppo è $P(A \mid B) = 20/50 = 0,4$.
Gli eventi A e B sono quindi dipendenti.
Infatti
P(A)= 0,2 e $P(A \mid B) = 0,4$
Quindi $P(A \mid B) \neq P(A)$
e gli eventi A e B sono dipendenti.

Esercizio. La probabilità che un gatto viva 12 anni è $\frac{1}{4}$, la probabilità che un cane viva 12 anni è $\frac{1}{3}$.

Calcolare la probabilità che il cane e il gatto appena nati:

a) il gatto non sia vivo fra 12 anni.
b) siano entrambi vivi fra 12 anni;
c) almeno uno sia vivo fra 12 anni;
d) nessuno dei due sia vivo fra 12 anni.

Svolgimento

Sia A l'evento "il gatto sia vivo tra 12 anni" e B l'evento "il cane sia vivo tra 12 anni".

a) $P(A^C) = 1\text{-}P(A) = 1 - \frac{1}{4} = \frac{3}{4}$.
b) Gli eventi A e B sono indipendenti. Quindi si può applicare la proprietà:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12}$$

c) L'evento ``almeno uno sia vivo fra 12 anni" è l'evento composto $A \cup B$, per Cui deriva che

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{4} + \frac{1}{3} - \frac{1}{12} = \frac{6}{12} = \frac{1}{2}$$

d) L'evento " nessuno dei due sia vivo fra 12 anni" è il complementare dell'evento $A \cup B$, quindi

$$P((A \cup B)^C) = 1 - P(A \cup B) = 1 - \frac{1}{2} = \frac{1}{2}$$

Formula di Bayes

Scambiando i ruoli di A e B e supponendo che $P(A)>0$ si ottiene la Formula di Bayes:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

- $P(B|A)$ si chiama probabilità a posteriori;
- $P(B)$ si chiama probabilità a priori;
- $P(A|B)$ si chiama verosimiglianza;

Il valore della verosimiglianza più o meno grande significa che la conoscenza dell'evento B rende più o meno grande la probabilità di A.
Si utilizza la formula di Bayes soprattutto quando B costituisce una causa nel senso di evento dal quale il verificarsi di A dipende in modo diretto e facilmente valutabile, mentre A rappresenta l' effetto.
La formula di Bayes si ottiene dalla definizione di probabilità condizionata. Infatti

$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A \mid B) \cdot P(B)$ e

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(B \mid A) \cdot P(A)$$

Tenendo conto che $P(A \cap B) = P(B \cap A)$, uguagliando le due equazioni sopra si ottiene

$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$ e quindi, dividendo primo e secondo membro per P(A) che è diverso da zero si ottiene la formula di Bayes.

Formula delle probabilità totali

Supponiamo di avere una partizione finita *{H_1, H_2, ..., H_n}* di eventi, con *$P(H_i)>0$* per ogni *$i=1, 2, ..., n$*. Dalla definizione di probabilità condizionata si ottiene la formula delle probabilità totali:

Formula di Bayes generalizzata

Data una partizione finita *{H_1, H_2, ..., H_n}* di eventi, con *$P(H_i)>0$* per ogni *$i=1, 2, ..., n$.* Dalla formula delle probabilità totali e dalla formula di Bayes si ottiene la formula di Bayes generalizzata:

$$P(H_i \mid A) = \frac{P(A \mid H_i) \cdot P(H_i)}{\sum_{j=1}^{n} P(A \mid H_j) \cdot P(H_j)}, i = 1,2,...,n$$

Supponiamo di avere una *n*-pla di eventi (ipotesi o cause esaustive) a due a due disgiunte, e di conoscere la loro probabilità a priori e supponiamo che si sia verificato un evento (effetto) del quale siano date le probabilità condizionate a ciascuna delle ipotesi. Allora la formula sopra consente di aggiornare la probabilità di ciascuna delle ipotesi condizionatamente all'evento osservato.

Esempio. In una data popolazione, una certa malattia si presenta con una frequenza di 1 persona ogni 1000 e la sua presenza può essere diagnosticata con un test che con una probabilità pari a 0,99 fornisce una risposta positiva qualora la persona sia effettivamente malata, ma che con una probabilità pari a 0,05 fornisce una risposta positiva qualora la persona sia invece sana. Valutare la probabilità che una persona che è risultata positiva al test sia effettivamente malata.

Risposta. *Se con M indichiamo la presenza della malattia, con S l'assenza di malattia e con il simbolo "+" il risultare positivo al test, si dispone delle seguenti informazioni: P(M)=0.001, P(+|M)=0.99, P(+|S)=0.05. Allora la probabilità richiesta può essere calcolata tramite il teorema di Bayes nel seguente modo:*

$$P(M|+) = \frac{0.001 * 0.99}{0.001 * 0.99 + 0.05 * 0.999} = 0.0194$$

Esempio. Si hanno tre scatole che contengono: la prima, 2 banconote da £100.000; la seconda, 1 banconota da £100.000 e 1 da £50.000; la terza, 2 banconote da £50.000. Si scelga a caso una delle tre

scatole (tra loro equiprobabili) e si estragga una banconota. Risulta estratta una banconota da £100.000; quale è la probabilità che la scatola dalla quale è stata estratta sia la prima?

Risposta. *Questo esercizio va risolto attraverso il teorema di Bayes. Avendo ipotizzato l'equiprobabilità di estrarre una delle tre scatole si ha che la probabilità a priori di estrarre una scatola qualsiasi è pari a 1/3. A questo punto se indichiamo con* S_1, S_2, S_3 *rispettivamente la prima, la seconda e la terza scatola e con C l'evento "estrazione di una banconota da £100.000" dobbiamo calcolare*

$$P(S_1|C) = \frac{P(S_1)P(C|S_1)}{P(S_1)P(C|S_1) + P(S_2)P(C|S_2) + P(S_3)P(C|S_3)} = \frac{\frac{1}{3} * 1}{\frac{1}{3} * 1 + \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * 0} = \frac{2}{3}$$

Campione casuale.
Da una popolazione finita di N unità vengono estratte n unità per osservare il valore di una certa caratteristica quantitativa X. Si assume che ciascuna unità una volta estratta venga riposta nella popolazione, cosicché ogni unità ha la stessa probabilità di essere estratta.
L' insieme delle unità estratte costituisce allora il **campione casuale**. Un modello probabilistico per tale procedura potrebbe essere un' urna con N palline dove su ciascuna pallina viene riportato il valore di X della caratteristica. Se l' estrazione viene fatta con reinserimento della pallina nell' urna, allora si tratta di un **campionamento con ripetizione**. Altrimenti, allora il campionamento é **senza ripetizione**. In tal caso la probabilità non é la stessa per ogni estrazione. Nel caso in cui N é molto grande e n / N molto piccolo, allora i due modelli sono equivalenti. In generale si assume che il modello si con ripetizione $N \geq 10\ n$.

Variabili aleatorie

Spesso i risultati di un esperimento sono valori numerici, per esempio il numero di clienti ad uno sportello di una banca, la paga oraria degli studenti che lavorano, ecc., ma molti esempi sono legati a caratteristiche qualitative (lancio di un dado o di una moneta). In questi casi è interessante concentrarsi su qualche aspetto numerico.
Dato uno spazio di probabilità una **Variabile aleatoria (v.a.) X** è una funzione che associa un valore numerico ad ogni risultato di un esperimento. L' aggettivo "aleatorio" ricorda che, a priori, non sappiamo né il risultato dell'esperimento né gli associati valori di X. Essa può essere discreta se l'insieme dei valori che può assumere è discreto, altrimenti è continua .

Data una variabile aleatoria discreta che può assumere i valori $x_1, ..., x_n$ possiamo associare ad ogni valore la probabilità che assuma quel valore, cioè

$p_i = P(X=x_i)$, $i=1, ..., n$.

L'insieme $\{p_1, ..., p_n\}$ è detto distribuzione di probabilità della variabile aleatoria.

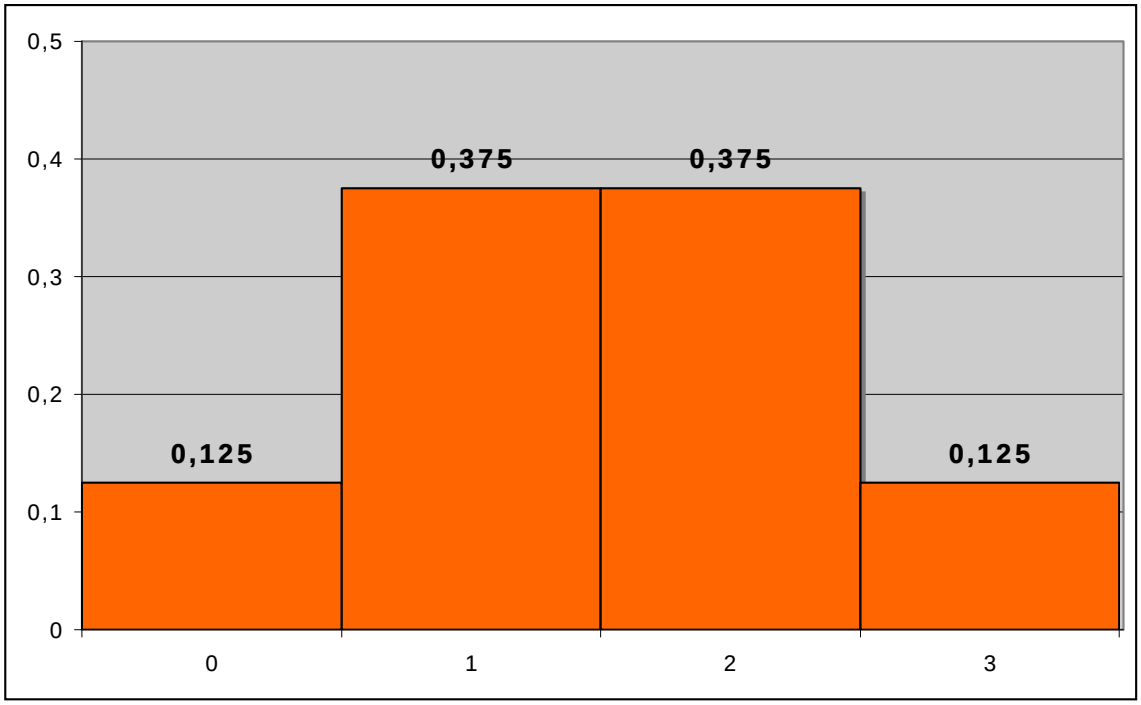
Esempio 7. Sia X la variabile aleatoria che misura il numero di "testa" in 3 lanci di una moneta non truccata.
Allora i possibili valori che X può assumere sono $\{0,1,2,3\}$, lo spazio degli eventi è

C	C	C	X=0
T	C	C	X=1
C	T	C	
C	C	T	
C	T	T	X=2
T	C	T	
T	T	C	
T	T	T	X=3

e la **distribuzione di probabilità** è la seguente:

x_i	p_i
0	1/8
1	3/8
2	3/8
3	1/8

Il relativo istogramma è



La **distribuzione di probabilità** di una variabile aleatoria discreta X è l'elenco dei distinti valori numerici che può assumere X con la rispettiva probabilità.
Le proprietà della distribuzione di probabilità di una v.a. discreta sono

- $p_i \geq 0$ per ogni i
- $p_1 + p_2 + \ldots + p_n = 1$

Valore atteso o media della v.a. X

$$\mu = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + ... + x_n p_n$$

Tale media è la media della popolazione.

Esempio. Una compagnia di assicurazione paga ad ogni cliente un premio di 1000 euro in caso di incidente o furto durante un viaggio di 5 giorni. Se il rischio di tale perdita è stimato 1 su 200, qual' è il prezzo onesto da pagare per tale polizza?
Soluzione. La probabilità che la compagnia paghi il cliente è 1/200=0,05, quindi se X è la somma pagata dalla compagnia allora la distribuzione di probabilità è

Pagamento	Probabilità
0 euro	0,995
1000 euro	0,05

In tal caso il valore atteso è

$\mu=0\cdot 0{,}995+1000\cdot 0{,}005=5$ euro

Il prezzo onesto da far pagare al cliente è quindi 5 euro.

Proprietà

$E(c_1 X_1 + c_2 X_2) = c_1 E(X_1) + c_2 E(X_2)$ dove c_1 c_2 sono costanti.

La **Varianza** *Var (X)* di una variabile aleatoria (variabile aleatoria) X con media E(X) è
$$Var(X) = \sum_{i=1}^{n} (x^2{}_i - \bar{x})^2 p_i = \sum_{i=1}^{n} x^2{}_i p_i - \bar{x}^2$$

Essa "misura" il grado di concentrazione della distribuzione attorno alla media. Essa è nulla se la variabile aleatoria è costante. *Var(X)* è lo spostamento quadratico medio o la deviazione standard. Essa è una misura di dispersione che ha la stessa unità di misura della variabile aleatoria.

Proprietà

1. $Var(c + X) = Var(X)$

2. $Var(c\,X)=c^2\,Var(X)$

3. $Var(c_1 X_1 + c_2 X_2 + ... + c_n X_n)= c_1^2\,Var(X_1) + c_2^2\,Var(X_2) + ... + c_n^2\,Var(X_n)$ per ogni n- pla di variabile aleatoria $X_1, X_2, ..., X_n$, indipendenti e costanti $c_1, c_2, ..., c_n$.

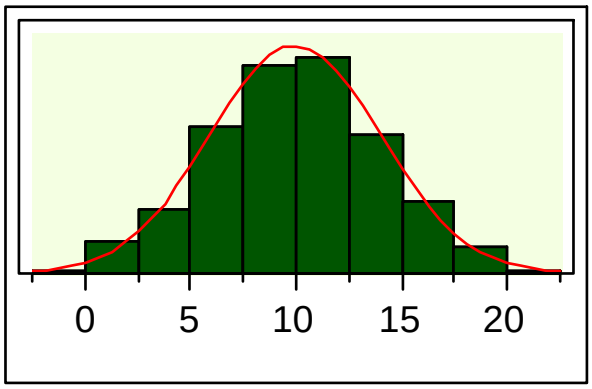
La **Deviazione standard o lo scarto quadratico medio** di una variabile aleatoria (variabile aleatoria) X o **Deviazione standard della popolazione** con media μ è la radice quadrata della varianza, $\sigma=\sqrt{Var(X)}$.

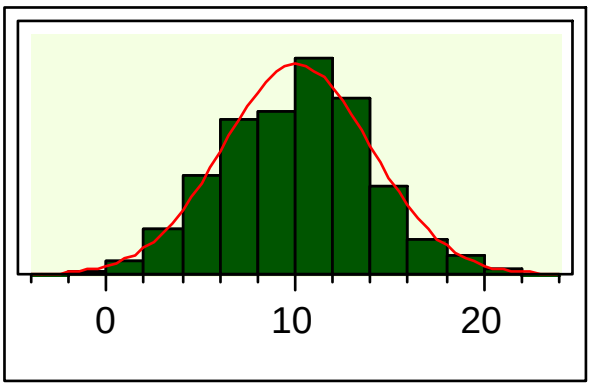
Se X è una variabile aleatoria con media $E(X)$ e varianza $Var(X)$, allora

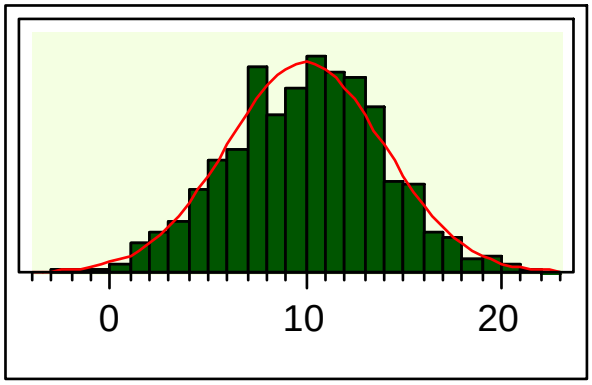
$$Y = \frac{X - E(X)}{\sqrt{Var(X)}}$$

è chiamata variabile aleatoria standardizzata , cioè $E(Y)=0$ e $Var\,(Y)=1$.

Distribuzione continue. Nel caso discreto abbiamo costruito per ogni campione e per una distribuzione discreta il relativo istogramma. In ogni caso l'area sottesa dall'istogramma e' uguale a 1. Se infittissimo le classi dell'istogramma aumentando i dati raccolti otterremmo degli istogrammi con più classi, e andando avanti con questa procedura otterremmo che l'istogramma si avvicina sempre di più ad una curva continua

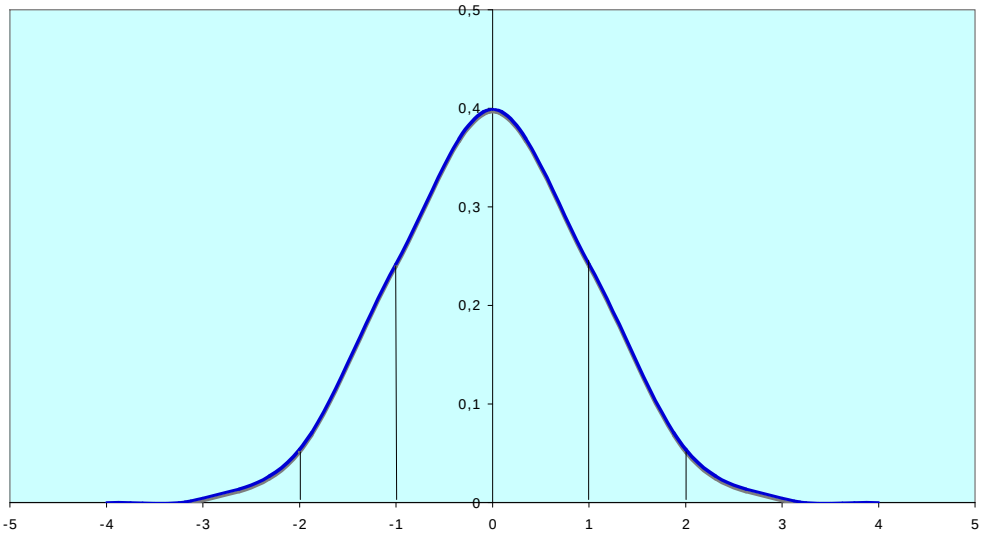






Distribuzione di Gauss

Fig.1 Distribuzione di Gauss standard



Proprietà:
- media μ
- varianza $\sigma 2$
- è simmetrica rispetto alla media
- ha media, moda e mediana coincidenti (e pari μ)
- non raggiunge mai lo zero per ogni valore di x.

Ad ogni distribuzione continua si associa la funzione densità di probabilità f(x) che corrisponde alla distribuzione di probabilità nel caso discreto. Soddisfa le seguenti proprietà

- f(x)≥0 per ogni x
- l'area totale della densità di probabilità è1
- l'area sottesa dalla curva normale nell'intervallo (a,b) è: $P(a \le X \le b) = \int_a^b f(x)dx$

Nel caso della distribuzione di Gauss $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

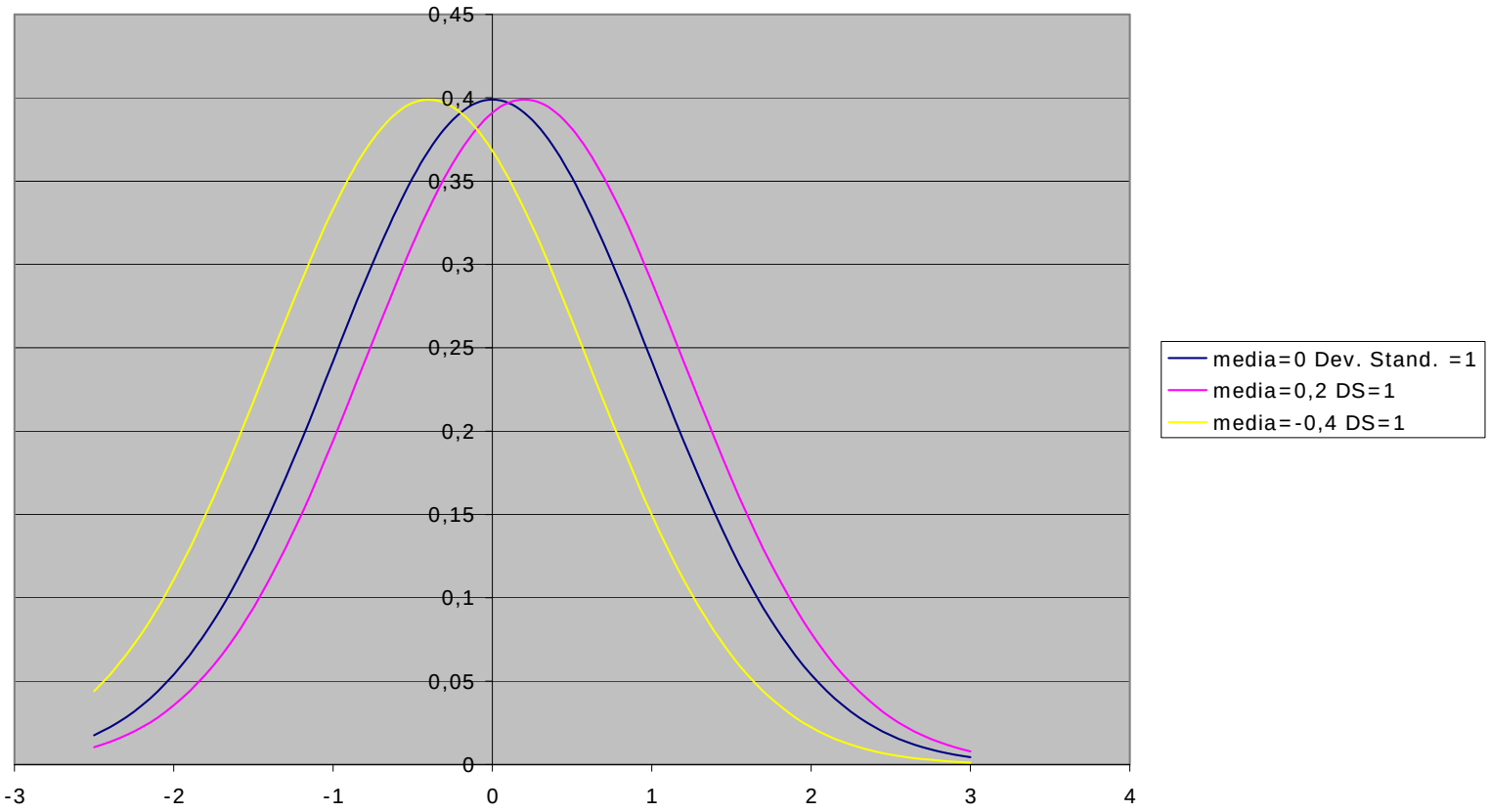
In particolare:

P(μ-σ,μ+σ) = 0.6827
P(μ-2σ,μ+2σ) = 0.9545
P(μ-3σ,μ+3σ) = 0.9973

La **funzione di distribuzione** di una variabile aleatoria è $F(x) \equiv P(X \le x) = \int_{-\infty}^{x} f(x)dx$

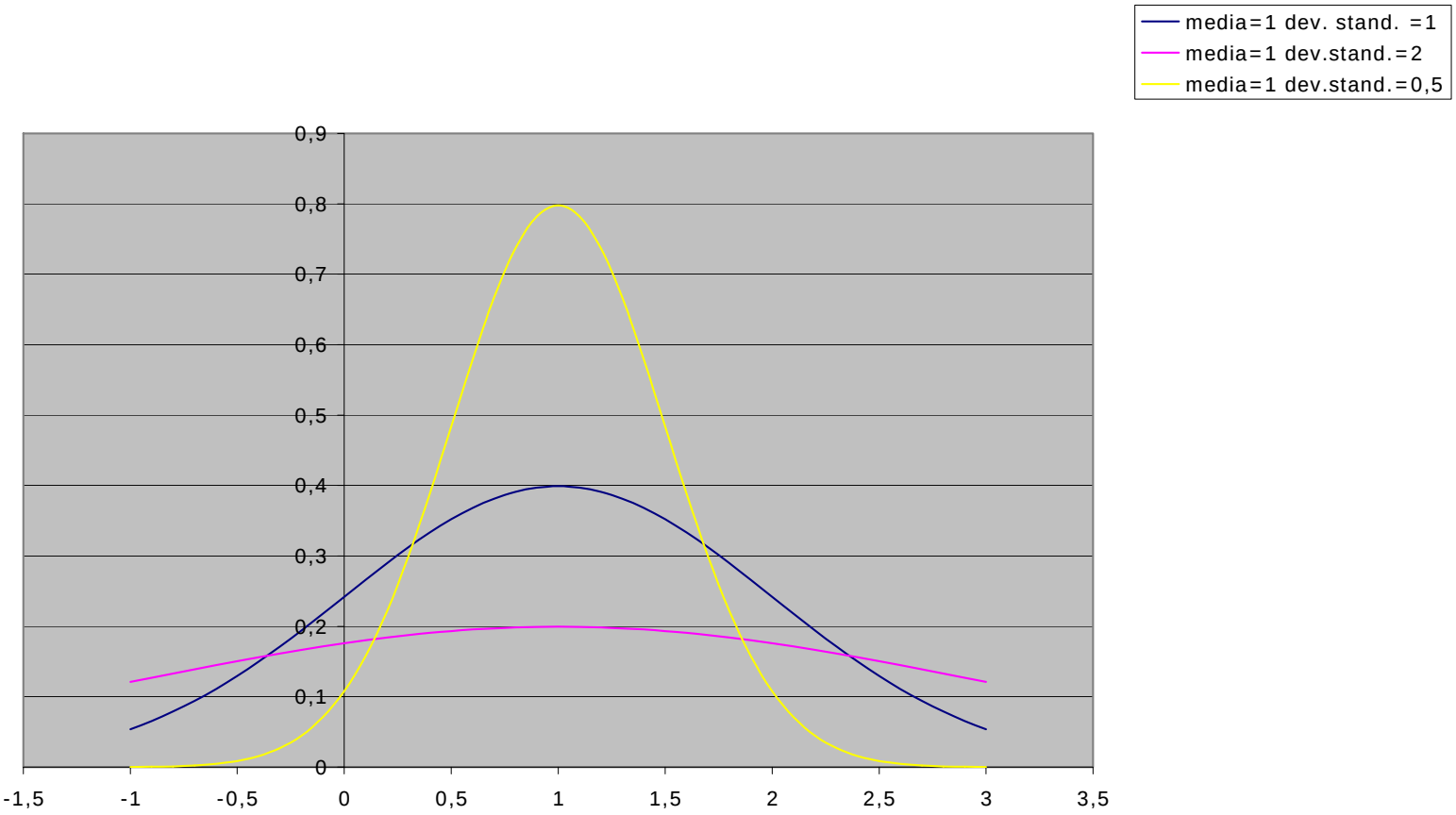
Osservazione. La **distribuzione di Gauss o curva degli errori** è la più importante distribuzione continua, è stata proposta da Gauss (1809) nell'ambito della teoria degli errori, è stata attribuita anche a Laplace (1812), che ne definì le proprietà principali in anticipo rispetto alla trattazione più completa di Gauss. Il nome deriva dalla convinzione che i fenomeni fisico-biologici solitamente si distribuiscono con frequenze più elevate nei valori centrali e frequenze progressivamente minori verso gli estremi, in quanto la distribuzione degli errori commessi nel misurare ripetutamente la stessa grandezza, è molto bene approssimata da questa curva .

Figura 1b: distr. normale con stessa Dev. Stand. =1



Cambiando la media la forma della configurazione non cambia ma si sposta il massimo.

Figura 1 a: distribuzione normale con stessa media =1



A differenza del caso precedente cambiando solo la deviazione standard la forma della curva cambia diventando più "appiattita" se essa aumenta, più "piccata" se diminuisce. La posizione del suo centro non cambia.

Dato una successione di variabili aleatorie indipendenti $\{X_n\}$ con la stessa distribuzione di probabilità, consideriamo la somma $S_n=X_1+X_2+\ldots+X_n$. Se $E(X_i)=\mu$ e $Var(X_i)=\sigma^2$, allora per le proprietà del valore aspettato e della varianza otteniamo $E(S_n)=n\mu$ e $Var(S_n)=n\sigma^2$. Normalizziamo la variabile

$$Z=\frac{S_n-n\mu}{\sigma\sqrt{n}}$$

Un risultato fondamentale nella Teoria della Probabilità noto come **Teorema del Limite Centrale** stabilisce che se l'ampiezza del campione è sufficientemente grande ($n\geq 30$) allora qualunque sia la distribuzione della successione la distribuzione della soma standardizzata Z campionaria può essere ben

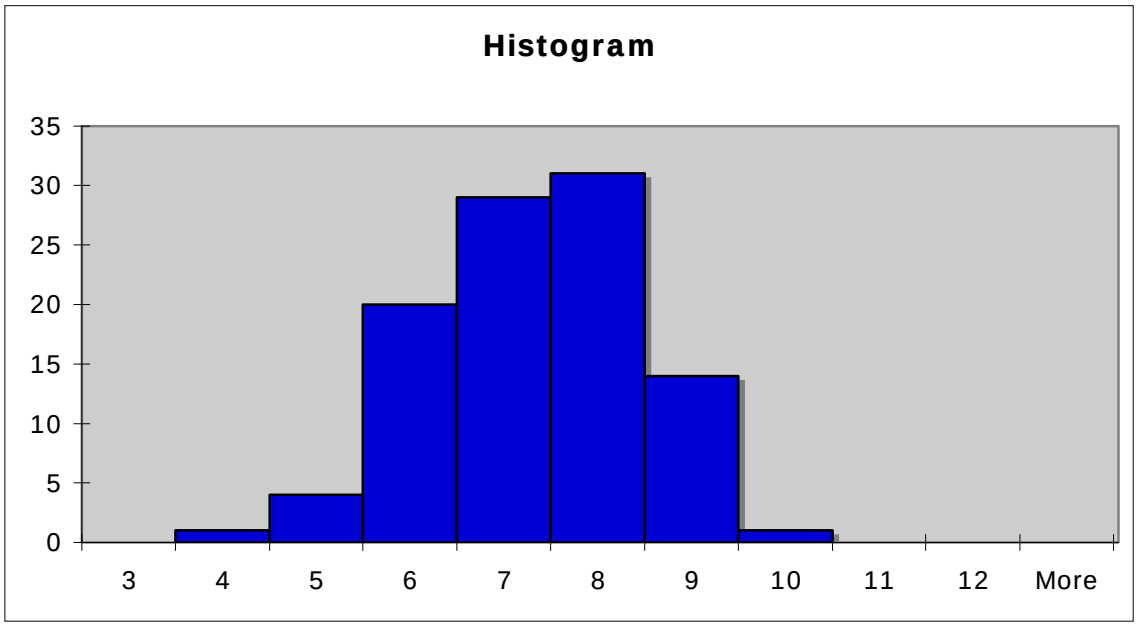
approssimata dalla distribuzione di Gauss standardizzata. La convergenza è in distribuzione nel senso che la funzione di distribuzione della media campionaria converge puntualmente a quella di Gauss.
La dimostrazione del Teorema del Limite Centrale non è semplice. Essa si base sul concetto di funzione caratteristica associata ad una variabile aleatoria. Diamo attraverso un esempio una conferma empirica.

Esempio: Si lanci due dadi e sia X la variabile aleatoria che misura la somma dei risultati. Supponiamo di lanciare 5 volte la coppia di dadi per 100 volte e otteniamo i seguenti risultati:

X_1	X_2	X_3	X_4	X_5	media
5	7	5	7	5	5,8
8	4	9	11	7	7,8
6	10	8	8	7	7,8
6	5	6	7	6	6
7	12	8	5	9	8,2
9	9	4	5	4	6,2
6	3	10	6	4	5,8
5	7	8	5	7	6,4
8	10	7	7	2	6,8
11	3	5	9	9	7,4
6	6	3	5	7	5,4
7	7	10	3	10	7,4
6	3	7	6	8	6
10	7	9	9	4	7,8
6	6	11	12	10	9
7	7	3	8	5	6
9	11	3	11	10	8,8
6	10	8	3	11	7,6
6	2	6	4	8	5,2
2	11	8	7	7	7
9	8	10	4	12	8,6
4	8	5	7	6	6
8	5	6	5	5	5,8
12	5	7	8	11	8,6
7	10	6	11	6	8
9	9	8	7	9	8,4
4	3	2	6	3	3,6
8	4	6	7	8	6,6
5	6	8	8	4	6,2
7	8	4	11	8	7,6
4	6	6	5	6	5,4
11	7	11	8	5	8,4
5	8	9	4	6	6,4

7	5	6	5	6	5,8
7	8	6	6	2	5,8
8	10	4	11	7	8
6	10	8	10	8	8,4
5	5	7	3	7	5,4
6	2	6	6	4	4,8
7	3	6	9	9	6,8
12	10	9	7	7	9
6	5	7	8	7	6,6
8	7	11	11	5	8,4
7	5	6	6	7	6,2
6	7	12	3	4	6,4
11	7	6	10	6	8
11	5	6	6	4	6,4
6	4	8	4	7	5,8
7	8	8	5	11	7,8
10	6	5	8	7	7,2
8	4	8	4	7	6,2
7	8	9	4	7	7
11	4	9	7	7	7,6
6	9	6	3	11	7
11	7	9	12	4	8,6
7	3	7	5	7	5,8
9	3	9	6	6	6,6
8	9	7	7	4	7
9	7	5	8	6	7
7	9	4	9	6	7
6	12	6	11	5	8
12	7	6	8	8	8,2
8	10	11	6	7	8,4
7	5	9	9	7	7,4
6	6	9	7	9	7,4
4	8	6	7	9	6,8
5	8	4	5	8	6
6	8	8	4	10	7,2
6	6	6	6	4	5,6
10	6	7	4	3	6
8	11	10	9	8	9,2
5	6	9	7	10	7,4
8	5	3	6	3	5
9	6	8	9	7	7,8
5	7	10	7	8	7,4

8	7	8	5	12	8
8	7	9	6	6	7,2
5	8	6	9	4	6,4
6	7	7	8	11	7,8
7	5	7	9	10	7,6
4	12	9	6	8	7,8
4	6	7	9	3	5,8
6	8	9	4	7	6,8
7	11	6	6	5	7
4	4	5	7	4	4,8
9	5	10	5	7	7,2
7	8	8	6	4	6,6
4	6	8	3	7	5,6
9	6	10	10	5	8
3	5	7	8	8	6,2
10	10	11	3	4	7,6
6	6	10	6	6	6,8
6	7	6	10	5	6,8
8	3	8	9	7	7
7	5	5	5	2	4,8
8	7	9	3	5	6,4
7	8	8	10	11	8,8
12	9	4	7	5	7,4
8	11	9	5	5	7,6
10	2	10	8	9	7,8



Statistica descrittiva	
Mean	6,96
Standard Error	0,11
Median	7,00
Mode	5,80
Standard Deviation	1,11
Sample Variance	1,24
Kurtosis	-0,25
Skewness	-0,25
Range	5,60
Minimum	3,60
Maximum	9,20
Sum	696,20
Count	100,00

Inferenza statistica

Introduzione

L' obiettivo dell' indagine statistica è ottenere un campione il più possibile rappresentativo della popolazione, cosicché le informazioni sulle caratteristiche della popolazione che da essa si traggono siano il più possibile accurate. Si cerca dalle osservazioni particolari di un campione di comprendere il caso generale, come la variabilità dei dati nella popolazione è trasmessa nel campione attraverso la sua media.
In generale siamo interessati a conoscere qualche grandezza numerica della popolazione, ad esempio la media o la deviazione standard. Tale grandezza si chiama **parametro.**
Il valore reale di un particolare parametro della popolazione è sconosciuto e può essere determinato solo dopo un' analisi su tutta la popolazione (ad esempio la media della popolazione). Se ciò è impossibile o non praticabile (per problemi di tempo, economici, ecc.) allora si utilizza l'inferenza statistica.
La **Stima dei parametri** è il procedimento con cui dal campione osservato si traggono informazioni per assegnare al parametro un valore (**stima puntuale**) o un insieme di valori (**stima per intervallo**).
La **Statistica** è una funzione numerica delle osservazioni del campione.
Ad esempio la **Media campionaria** definita da

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

è una statistica. perchè il suo valore numerico può essere calcolato dai dati del campione X_1, X_2, …, X_n, a disposizione. Altre statistiche sono la mediana del campione o la varianza.

Osservazioni.
1. Poiché il campione è una parte della popolazione il valore della statistica non può dare l'esatto valore del parametro
2. Il valore della statistica dipende dal particolare campione selezionato.
3. Esiste una variabilità nei valori della statistica su differenti modi di campionamento.

Il fatto che la media del campione vari con il campione suggerisce l'idea che la media campionaria sia una variabile aleatoria. La distribuzione di probabilità della statistica è chiamata **distribuzione campionaria**. La distribuzione campionaria è determinate da quella della popolazione di riferimento e dall' ampiezza del campione *n*.
Per approfondire il tipo di relazione partiamo da una popolazione con media μ e deviazione standard σ. Allora dato un campione estratto da essa X_1, …, X_n si ha

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + ... + X_n) = \frac{1}{n} n\mu = \mu$$

$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + ... + X_n) = \frac{1}{n^2} n\sigma^2 = \sigma^2 / n$$

$$Dev.Stand.(\overline{X}) = \sqrt{Var(\overline{X})} = \sigma / \sqrt{n}$$

La prima relazione ci dice che il valore aspettato della media campionaria è uguale a quello della popolazione, cioè μ. La terza relazione ci dice che la deviazione standard, e quindi la variabilità dei dati della media campionaria dipende dalla variabilità dei dati della popolazione, σ e dall'ampiezza del

campione, al crescere di σ è più difficile localizzare la media μ. D'altra parte si potrebbe diminuire tale incertezza aumentando l'ampiezza del campione. Infatti $\sigma/\sqrt{n}$ decresce al crescere di n.

Applicando il **Teorema del Limite Centrale** alla media standard

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

essa può essere ben approssimata dalla variabile normale standardizzata.

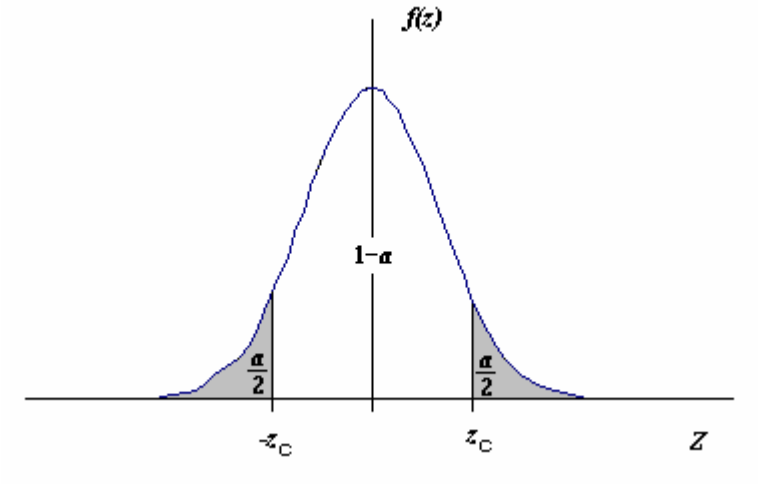
Inferenza su una media: intervallo di fiducia

Vorremo stimare la media della popolazione sconosciuta attraverso un campione, calcolando a partire dal campione un intervallo che potrebbe contenere la media cercata con un certo grado di credibilità.

Varianza nota. Sia $X_1, \ldots, X_n$ un campione estratto da una popolazion con media incognita μ e varianza nota σ^2 . Vogliamo trovare dei valori $(-z_c, z_c)$ tali che Z sia compresa nell' intervallo $(-z_c, z_c)$ con una alta probabilità a (ad esempio $1-a$=0,95), cioè

$$P(-z_c \le Z \le z_c) = 0{,}95$$

Dalle tavole di Gauss si vede che tale valore è esattamente z_c=1,96.



Sostituendo a Z l'espressione sopra si ottiene quindi

$$P\left(-1{,}96 \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le 1{,}96\right) = 0{,}95$$

Applicando le proprietà sulle disuguaglianze si ottiene

$$P\left(-1{,}96 \cdot \sigma/\sqrt{n} \le \overline{X} - \mu \le 1{,}96 \cdot \sigma/\sqrt{n}\right) = 0{,}95$$

e quindi

$$P\left(-\overline{X}-1{,}96\cdot \sigma/\sqrt{n} \le -\mu \le -\overline{X}+1{,}96\cdot \sigma/\sqrt{n}\right)=0{,}95$$

Applicando di nuovo le proprietà delle disuguaglianze otteniamo

$$P\left(\overline{X}-1{,}96\cdot \sigma/\sqrt{n} \le \mu \le \overline{X}+1{,}96\cdot \sigma/\sqrt{n}\right)=0{,}95$$

da cui si ottiene che l'intervallo di fiducia per la media con un grado di fiducia del 95% è

$$\left(\overline{X}-1{,}96\cdot \sigma/\sqrt{n}, \overline{X}+1{,}96\cdot \sigma/\sqrt{n}\right)$$

L'errore o livello di fiducia

$$E=1{,}96\cdot\frac{\sigma}{\sqrt{n}}$$

dipende dalla variabilità della popolazione e cresce con essa, dall'ampiezza del campione e decresce al crescere di n.

Fissando un grado di fiducia pari ad 1-*a* si ottiene dalle tavole un valore critico z_c tale che

$$P(-z_C \le Z \le z_C)=1-a$$

da cui si ottiene il corrispondente intervallo di fiducia

$$\left(\overline{X}-z_c\cdot \sigma/\sqrt{n}, \overline{X}+z_c\cdot \sigma/\sqrt{n}\right)$$

L'errore o livello di fiducia nel caso generale ha la seguente espressione

$$E=z_c\cdot\frac{\sigma}{\sqrt{n}}$$

Tale errore aumenta anche all' aumentare del grado di fiducia. Se volessimo fissare un errore a priori, conoscendo la varianza della popolazione, siamo in grado di trovare quanto deve essere ampio il campione in modo da avere un errore non più grande di quello fissato.

Esempio. Un urbanista é interessato alla superficie media μ delle abitazioni della propria città. Uno studio precedente indica che la deviazione standard della popolazione sia circa 8 m^2. In un campione di 50 appartamenti si osserva la media del campione è pari a 120 m^2 . Sulla base di questi dati l' intervallo di confidenza per μ con un grado di fiducia del 95% è

$$\left(120-1{,}96\cdot 8/\sqrt{50}, 120+1{,}96\cdot 8/\sqrt{50}\right)\approx(117{,}78;122{,}22)$$

Se volessimo trovare quanto deve essere grande il campione in modo da avere un errore non più grande di 1, per risolvere il problema basterebbe considerare la disequazione

$$|E| \leq 1 \Rightarrow z_c \cdot \frac{\sigma}{\sqrt{n}} \leq 1 \Rightarrow$$

$$1{,}96 \cdot \frac{8}{\sqrt{n}} \leq 1 \Rightarrow$$

$$\sqrt{n} \geq 8 \cdot 1{,}96 \Rightarrow$$

$$n \geq 15{,}68^2 \Rightarrow$$

$$n \geq [245{,}86]$$

Quindi basterebbe scegliere n=246 per ottenere un errore non più grande di 1.

Varianza non nota. Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione con media incognita μ e varianza non nota σ^2 . Allora in questo caso il parametro σ potrebbe essere approssimato da s', dove s' è la deviazione standard del campione con la correzione di Student, cioè

$$s' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Si ottiene in questo modo al posto di Z una nuova variabile

$$T = \frac{\bar{X} - \mu}{s' / \sqrt{n}}$$

tale distribuzione si chiama di **Student.** E' simile alla distribuzione di Gauss e per *n* molto grande potrebbe essere ben approssimata da essa. Poiché vale la seguente relazione

$$\frac{s'}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

potremmo anche considerare la variabile

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n-1}}$$

Fissando un grado di fiducia *1-a* si cercano dei valori critici $(-t_c, t_c)$ dalle tavole di Student tali che

$$P(-t_c \leq T \leq t_c) = 1-a$$

A differenza delle tavole di Gauss, quelle di Student dipendono dai **gradi di libertà** del campione cioè *n-1*. Ripetendo tutti i passaggi precedenti si ottiene come intervallo di fiducia per la media con varianza non nota

$$\left(\bar{X} - t \cdot s'/\sqrt{n}, \bar{X} + t \cdot s'/\sqrt{n}\right)$$

oppure

$$\left(\bar{X} - t \cdot s/\sqrt{n-1}, \bar{X} + t \cdot s/\sqrt{n-1}\right)$$

Esempio. Il produttore di una certa marca di sigarette desidera controllare il quantitativo di catrame in esse contenuto. A questo scopo si osserva un campione di 30 sigarette in cui la media è 10.92 mg e la deviazione standard 0.50 mg . Sulla base di questi dati l' intervallo di fiducia per la media pari al 99%
$\left(10{,}92 - 2{,}756 \cdot 0{,}51/\sqrt{30}, 10{,}92 + 2{,}756 \cdot 0{,}51/\sqrt{30}\right) \approx (10{,}66; 11{,}18)$
Il valore t_c è stato trovato considerando la 29- esima riga (gradi di libertà) e $2\gamma = 0{,}01$ (somma delle due code).

Intervallo di fiducia per una media con Analisi dei dati di Excel
L'intervallo di fiducia con Excel si può trovare solo nel caso in cui non si conosce niente della popolazione e vengono utilizzate per questo le tavole di Student.

- Scegliere Statistica descrittiva da Analisi dei dati.
- Scegliere come opzione livello di fiducia per la media.
- Se non si indica viene calcolato come default il livello 0,95.
- Come output apparirà il livello di confidenza cioè

$$E = t_c \cdot \frac{s'}{\sqrt{n}}$$

L' intervallo di fiducia è allora $(\overline{X} - E, \overline{X} + E)$.

Test di ipotesi

In contrasto con il problema della stima degli intervalli, in cui si cerca di valutare un parametro incognito, ci sono parecchie situazioni in cui si é costretti a scegliere tra due possibilità. Per esempio dopo la scoperta del vaccino Salk, molti test sono stati fatti per vedere se esso avrebbe potuto realmente prevenire il polio. Le due possibili azioni sono raccomandare o scoraggiare l'uso del vaccino. Il problema più semplice sorge é determinare quale fra due distribuzioni é preferibile come modello dei dati osservati.

Definizione. L' **ipotesi statistica** è una affermazione o una congettura intorno ad un parametro incognito ρ della popolazione, formulata sulla base di considerazioni teoriche o risultati sperimentali che riguarda un aspetto della popolazione studiata. L'ipotesi sottoposta a verifica viene in genere indicata con H_o e viene chiamata **ipotesi nulla**. Si chiama **test** il procedimento con cui si decide se, attraverso i dati del campione, accettare o rifiutare H_o. L'ipotesi H_o verrà rifiutata se dal confronto dei dati osservati con essa emergerà una "discrepanza considerevole". Il problema è "decidere" quanto deve essere grande la distanza tra il valore teorizzato e quello vero perché venga rifiutata l'ipotesi.

Esempio 1: L'industria A sostiene che le batterie elettriche da essa prodotte hanno una durata media di 36 mesi con una deviazione standard di 3 mesi. Un' industria automobilistica B è interessata al prodotto, ma prima di qualsiasi decisione di acquisto, intende controllare l' affermazione di A attraverso l'osservazione di un campione di batterie dalla popolazione composta dal numero indefinito di batterie che l'industria può potenzialmente produrre. L'affermazione di A, cioé μ=36, è un'ipotesi sulla media della popolazione, che si può approssimare con una distribuzione di Gauss se l'ampiezza del campione é sufficientemente ampia.

Data l'ipotesi nulla $H_0 : \rho = \rho_0$ allora l' **ipotesi alternativa** detta H_1 può essere
H_1 :

$\rho \neq \rho_0$

$\rho < \rho_0$

$\rho > \rho_0$

Relativamente all' Esempio precedente poiché B dubita dell'ipotesi H_0, penso che il valore medio dichiarato sia troppo alto allora si pone
H_1: $\mu < 36$.

Il test può condurre a decisioni errate, infatti esso è eseguito su base probabilistica utilizzando i risultati ottenuti con un campione

Si confrontano quindi due possibili ipotesi:
- H_0: Ipotesi nulla
- H_1: Ipotesi alternativa

Il metodo per eseguire il test è analogo a quello usato per calcolare gli intervalli di confidenza. In generale nell'ipotesi nulla H_0 si suppone che un parametro ρ della popolazione abbia valore pari a ρ_0. Non essendo certi di tale supposizione si esegue un test di tale ipotesi mediante un'indagine campionaria. Consideriamo come parametro incognito la media μ di una popolazione.

Test a due code per una media

Si formulano le ipotesi:

- H_0: $\mu = \mu_0$ (ipotesi nulla: il valore del parametro μ è pari a μ_0)
- H_1: $\mu \neq \mu_0$ (ipotesi alternativa: il valore del parametro μ è diverso da μ_0)

Varianza nota: sia σ^2 la varianza (nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 e varianza σ^2 / n (n dimensione dei campioni); quindi la variabile casuale

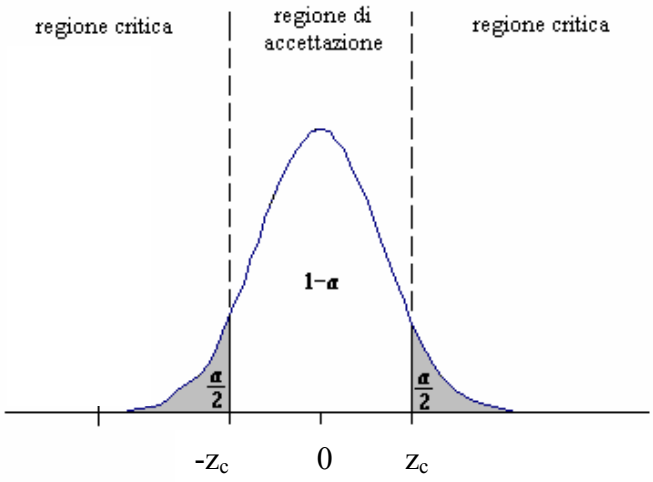
$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

ha distribuzione approssimativamente normale standard per n sufficientemente grande. Se l'ipotesi H_0 è vera, il valore Z è probabilmente vicino a *0*, anche se difficilmente sarà uguale. Se l'ipotesi H_0 è falsa, il valore Z è probabilmente abbastanza "lontano" da *0*. È opportuno quindi stabilire quanto deve discostarsi Z da 0 affinché H_0 possa essere ritenuta falsa. Tale decisione si prende in termini probabilistici ed essa è in funzione di un grado di fiducia *1-a* già introdotto nell' ambito dell' intervallo di fiducia per una media, a cui corrisponde un intervallo $(-z_c, z_c)$ di valori critici. Allora

- se $Z < -z_c$ oppure $Z > z_c$ (**regione critica**): si deve rifiutare l'ipotesi H_0, in quanto se essa fosse vera si avrebbe una probabilità pari a $a / 2$ di avere $Z < -z_c$ e si avrebbe una probabilità pari a $a / 2$ di avere $Z > z_c$; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se $-z_c \leq Z \leq z_c$ (**regione di accettazione**): non si può rifiutare l'ipotesi H_0, in quanto se essa fosse vera si avrebbe una probabilità pari a a di avere $-z_c \leq Z \leq z_c$; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Per mezzo delle tavole della distribuzione normale standard è possibile calcolare il valore critico z_c, tale che l'area della regione critica è pari ad $1-a$; se il valore della variabile casuale Z ottenuto dal campione cade nella regione di accettazione, H_0 si accetta; se il valore della variabile casuale Z ottenuto dal campione cade nella regione critica, H_1 si rifiuta.

Il test appena considerato è detto **Test a due code** perché l'ipotesi H_1 si accetta per valori Z sia maggiori di z_c che minori di $-z_c$. Questo tipo di test serve per verificare se Z è significativamente diverso da 0 o, equivalentemente, se il parametro μ è significativamente diverso da μ_0.



Test ad una coda per la media

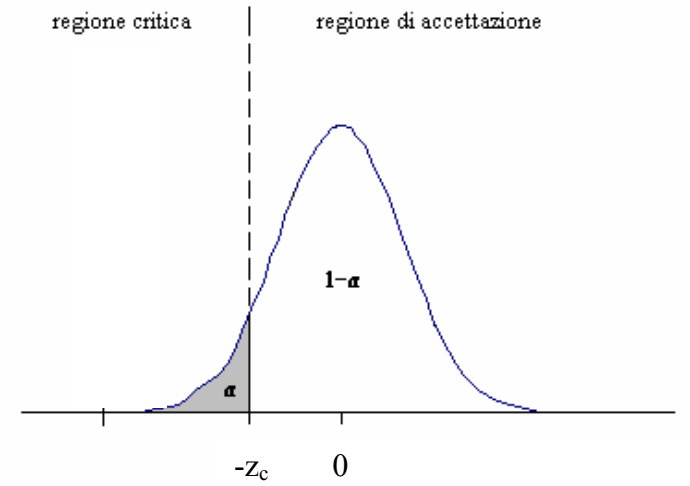
Varianza nota. A volte è necessario usare dei tests ad una coda in cui la regione critica è localizzata solo a destra di μ_0 o solo a sinistra di μ_0, o, se consideriamo la variabile standardizzata Z a destra di 0, o a sinistra di 0. Sia σ^2 la varianza (nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 (quella ipotizzata nell' ipotesi nulla) e varianza σ^2 / n (n dimensione dei campioni); quindi la variabile casuale

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

ha distribuzione approssimativamente normale standard per n sufficientemente grande.

I caso: **coda a sinistra**

- H_0: $\mu = \mu_0$
- H_1: $\mu < \mu_0$

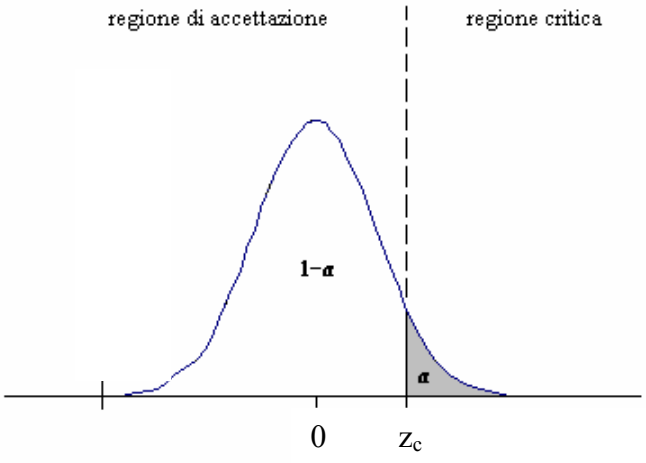


- se $Z < -z_c$ (**regione critica**) si deve rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a a di avere $Z < -z_c$; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se $Z \geq -z_c$ (**regione di accettazione**): non si può rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a a di avere $Z \geq -z_c$; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Il test appena considerato è detto **Test ad una coda** perché l'ipotesi H_1 si accetta per valori Z maggiori di $-z_c$.

II caso: **coda a destra**

- H_0: $\mu = \mu_0$
- H_1: $\mu > \mu_0$



- se $Z > z_c$ (**regione critica**) si deve rifiutare l'ipotesi H_0, in quanto se essa fosse vera si avrebbe una probabilità pari a a di avere $Z > z_c$; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se $Z \leq z_c$ (**regione di accettazione**): non si può rifiutare l'ipotesi H_0, in quanto se essa fosse vera si avrebbe una probabilità pari a 1-a di avere $Z \leq z_c$; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Il test appena considerato è detto **Test ad una coda** perché l'ipotesi H_1 si accetta per valori Z minori di z_c.

VARIANZA NON NOTA: sia σ^2 la varianza (non nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 e varianza σ^2 / n (n dimensione dei campioni), quindi la variabile casuale:

$$T = \frac{\overline{X} - \mu_0}{s' / \sqrt{n}}$$

ha distribuzione T di Student con $\nu = n-1$ gradi di libertà; quindi per mezzo delle tavole della distribuzione di Student è possibile calcolare t_c, detto **valore critico**, tale che l'area della regione critica è pari ad $1-a$; se il valore della variabile casuale T ottenuto dal campione cade nella regione di accettazione, H_0 si accetta; se il valore della variabile casuale T ottenuto dal campione cade nella regione critica, H_0 si rifiuta.

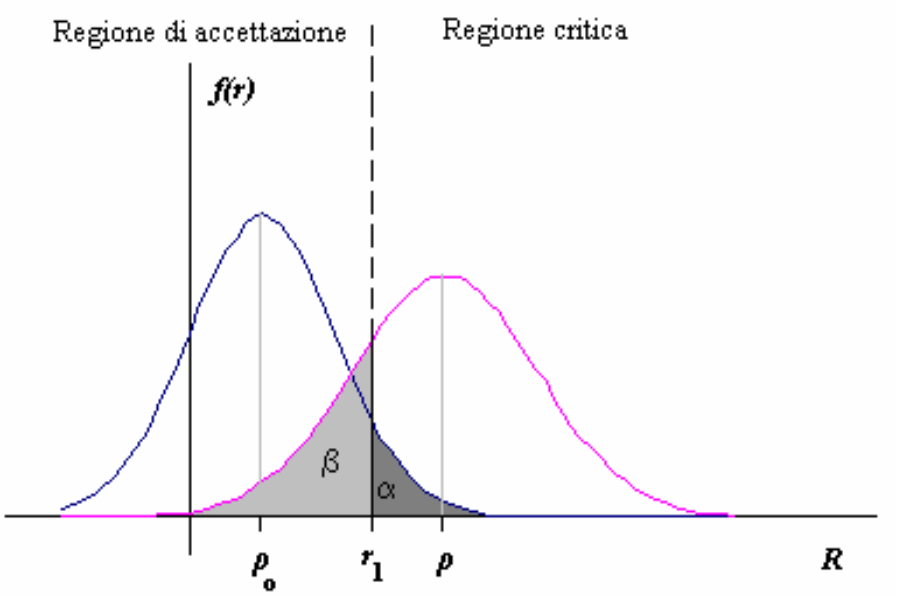
Osservazione: per grandi campioni ($n > 100$) la distribuzione T di Student con $\nu = n-1$ gradi di libertà è ben approssimata dalla distribuzione normale, quindi in tali ipotesi si può usare indifferentemente l'una o l'altra distribuzione; in particolare per grandi campioni ($n > 100$) la varianza della popolazione può essere considerata sempre nota perché s^2 fornisce una buona approssimazione di σ^2.

In tal caso vale il ragionamento fatto sopra per il test a due code e ad una coda, con T al posto di Z e t_c al posto di z_c.

Errori di prima e seconda specie

La decisione di rifiutare o accettare H_0 è sempre presa su base probabilistica considerando dei risultati campionari, è quindi possibile commettere degli errori nel prendere tali decisioni:

- Errori di prima specie: quando H_0 è vera, ma in base ai risultati campionari H_0 viene rifiutata
- Errori di seconda specie: quando H_0 è falsa, ma in base ai risultati campionari H_0 viene accettata



Si osserva che la probabilità di commettere un errore di prima specie è pari al livello di significatività *a*, mentre 1-*a* è la probabilità di accettare H_0 quando H_0 è vera. Sia β la probabilità di commettere un errore di seconda specie, mentre 1-β è la probabilità di accettare H_1 quando H_1 è vera. Il valore 1-β è generalmente detto POTENZA DEL TEST.

		Decisione	
		H_0	H_1
Realtà	H_0	DECISIONE ESATTA Prob.: $1 - \alpha$	ERRORI I SPECIE Prob.: α
	H_1	ERRORI II SPECIE Prob.: β	DECISIONE ESATTA Prob.: $1 - \beta$

Esempio: il contenuto di nicotina delle sigarette di un certo tipo risulta normalmente distribuito con deviazione standard di 4mg. Se, per rendere minimo il rischio di cancro ai polmoni, il contenuto medio di nicotina delle sigarette non deve superare 26mg e in un campione di 10 sigarette si sono ottenuti i seguenti valori di nicotina (in mg):

33 27 20 36 25 24 27 24 34 29

si può affermare, ad un livello di significatività pari a 0.05, che i consumatori di quel tipo di sigarette corrono un rischio minimo di cancro ai polmoni?
Risposta: si ha H_0: $\mu = 26$, H_1: $\mu > 26$ (test a una coda a destra). Dal livello di significatività $\alpha=0.05$ si ha $z_c=1.645$ (dalle tavole della distribuzione normale standard), inoltre dai dati campionari si ha $\bar{x}=27.9$ e quindi

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{27.9 - 26}{4/\sqrt{10}} = 1.502$$

cade nella regione di accettazione. Il campione non è statisticamente significativo si accetta l'ipotesi nulla H_0.
Osserviamo che quando il campione è troppo piccolo come in questo caso si deve fare l'ipotesi che la distribuzione della media campionaria è gaussiana, perchè non si può applicare il Teorema del Limite Centrale.

Esempio: in base all'esperienza degli anni precedenti risulta che le votazioni, ad un certo esame scritto, riportate dagli studenti di un certo corso di laurea sono distribuite in maniera approssimativamente normale con media di 23 trentesimi. Se un gruppo di 50 studenti dell'anno in corso riporta una votazione media di 25 trentesimi con deviazione standard di 4 trentesimi, si può accettare l'ipotesi che tali studenti non differiscano da quelli degli anni precedenti al livello di significatività di 0.02
Risposta: si ha H_0: $\mu = 23$, H_1: $\mu \neq 23$ (test a due code). Dal livello di significatività $a=0.02$ si ha $t_c=2.423$ (dalle tavole della distribuzione t di Student con 49 gradi di libertà), inoltre dai dati campionari si ha $\bar{x}=25$, $s'=4$, quindi

$$T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} = \frac{25-23}{4/\sqrt{50}} = 3.536$$

cade nella regione critica. Il campione è statisticamente significativo si rifiuta l'ipotesi nulla H_0.

Test sulle differenze di medie

Si considera due popolazioni. Sia μ_1 la media della prima popolazione e sia μ_2 la media della seconda popolazione; si formula la seguente ipotesi nulla:

H_0: $\mu_1 = \mu_2$

mentre le ipotesi alternative possono essere:

H_1: $\mu_1 \neq \mu_2$ (test a due code)
H_1: $\mu_1 < \mu_2$ (test a una coda a sinistra)
H_1: $\mu_1 > \mu_2$ (test a una coda a destra)

Campioni indipendenti: non esiste alcuna relazione tra i risultati dei due campioni, i risultati di un campione non influenzano quelli dell'altro.

Varianza nota: siano σ_1^2, σ_2^2 le varianze (note) delle popolazioni (generalmente si usa con $\sigma_1 = \sigma_2$); sia $\overline{X}_1$ la variabile casuale della media campionaria di campioni casuali di dimensione n_1 estratti dalla prima popolazione; sia $\overline{X}_2$ la variabile casuale della media campionaria di campioni casuali di dimensione n_2 estratti dalla seconda popolazione; allora la variabile casuale $\overline{X}_1 - \overline{X}_2$ ha distribuzione normale con media μ_1-μ_2 , che per H_0 è nulla, e varianza $\sigma_1^2/n_1 + \sigma_2^2/n_2$; quindi la variabile casuale:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ha distribuzione normale standard, quindi per mezzo delle tavole della distribuzione normale standard è possibile calcolare z_c, tale che l'area della regione critica è pari ad a; se il valore della variabile Z ottenuto dal campione cade nella regione di accettazione, si accetta H_0; se il valore della variabile casuale Z ottenuto dal campione cade nella regione critica, si rifiuta H_0.

Varianze non note: si suppone per semplicità che $\sigma_1 = \sigma_2$; siano $\bar{X}_1$ e S_1^2 rispettivamente la variabile casuale della media campionaria e la variabile casuale della varianza campionaria di campioni casuali di dimensione n_1 estratti dalla prima popolazione; siano $\bar{X}_2$ e S_2^2 rispettivamente la variabile casuale della media campionaria e la variabile casuale della varianza campionaria di campioni casuali di dimensione n_2 estratti dalla seconda popolazione; la stima di $\sigma = \sigma_1 = \sigma_2$ si può dare in termini di S_1^2 e S_2^2, cioè come media ponderata delle due varianze

$$\sigma^2 \approx \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \equiv S_p^2$$

inoltre la variabile casuale:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_p^2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_p^2}}$$

ha distribuzione T di Student con $\nu = n_1 + n_2 - 2$ gradi di libertà; quindi per mezzo delle tavole della distribuzione t di Student è possibile calcolare t_c, tale che l'area della regione critica è pari ad a; se il valore della variabile casuale Z ottenuto dal campione cade nella regione di accettazione, si accetta H_0; se il valore della variabile casuale Z ottenuto dal campione cade nella regione critica, si rifiuta H_0.

Esempio: un campione di 40 capsule di analgesico è stato fabbricato da una macchina A, il peso medio è $\bar{x}$=330 mg, la deviazione standard è s=7 mg; una macchina B ha prodotto 50 capsule con peso medio $\bar{x}$=320 mg e deviazione standard s=6.5 mg. Sottoporre a test l'ipotesi che le due macchine producano capsule di stesso con un livello di significatività pari a 0.05
Risposta: si ha H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ (test a due code). Si tratta di campioni indipendenti. Dal livello di significatività a=0.05 si ha t_c=1.98 (dalle tavole della distribuzione di Student), inoltre dai dati campionari si ha n_1=40, $\bar{x}_1$=330, s_1=7, n_2=50, $\bar{x}_2$=320, s_2=6.5 e quindi

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_p^2}} = \frac{330-320}{\sqrt{\left(\frac{1}{40}+\frac{1}{50}\right)\left(\frac{7^2(39)+6.5^2(49)}{88}\right)}} = 7{,}009$$

cade nella regione critica. Il campione è statisticamente significativo si rifiuta l'ipotesi nulla H_0.

Confronto tra due medie con uguale varianze con Analisi dei dati di Excel

Nella finestra di Dialogo bisogna inserire il range della prima variabile, l'intervallo della seconda, la differenza che si ipotizza (0 se si vuole verificare che le due medie coincidano), Alfa, cioè *a* dove *1-a* è il grado di fiducia e dove visualizzare la Tabella.

Apparirà una tabella di output dove sono descritti le medie, le varianze, le numerosità dei due campioni, la media pesata delle varianze, la differenza delle medie ipotizzata, i gradi di libertà.
T-stat è il valore di T ottenuto dai due campioni che va confrontato con il valore **t critico a due code** se il test è a due code oppure con **t critico ad una coda** se il test è ad una coda.
Il valore **P(T<=t) una coda** è la probabilità di ottenere una differenza di medie campionarie maggiore o uguale a quella osservata, quindi dovrebbe essere **P(T>=t) una coda.**
P(T<=t) due code è la probabilità di ottenere una differenza di medie campionarie in modulo maggiore o uguale a quella osservata.