EIDS – Teoria Statistica

Maria Teresa Giraudo

Table of Contents

[Capitolo 1 – Definizioni, tipi di esperimenti, campionamento. 1](#_Toc3627459)

[Introduzione 1](#_Toc3627460)

[Tipologie di dati 1](#_Toc3627461)

[Tipologie di esperimento 2](#_Toc3627462)

[Strategie di campionamento 3](#_Toc3627463)

[Capitolo 2 - Analisi esplorativa dei dati: Statistica Descrittiva 3](#_Toc3627464)

[Caratteristiche dei dati 3](#_Toc3627465)

[Frequenze 4](#_Toc3627466)

[Rappresentazione grafica dei dati 4](#_Toc3627467)

[Misure di centralità – Come valutare il centro 5](#_Toc3627468)

[Misure di dispersione – Come valutare l’ampiezza della distribuzione 6](#_Toc3627469)

[Misure di posizione – Dividere in parti la distribuzione di dati 7](#_Toc3627470)

[Analisi esplorativa dei dati – Trarre informazioni preliminari per uno studio 7](#_Toc3627471)

[Capitolo 3 – Probabilità 8](#_Toc3627472)

[Operazioni con le probabilità – Probabilità di due eventi 9](#_Toc3627473)

[Variabili Aleatorie 11](#_Toc3627474)

[Distribuzioni binomiali – Vero o Falso 11](#_Toc3627475)

[Distribuzione di Poisson – Eventi rari 12](#_Toc3627476)

[Distribuzione normale – Campana di Gauss 13](#_Toc3627477)

[Distribuzioni campionarie di una statistica 13](#_Toc3627478)

[Proporzioni -Distribuzione campionaria di una proporzione 14](#_Toc3627479)

[Il teorema del Limite Centrale 15](#_Toc3627480)

[Accertare la normalità di una distribuzione 16](#_Toc3627481)

[Capitolo 4 – Stime con un campione 16](#_Toc3627482)

[Stima delle proporzioni in una popolazione 16](#_Toc3627483)

[Gli intervalli di confidenza 17](#_Toc3627484)

[Stima della media di una popolazione 18](#_Toc3627485)

[La deviazione standard è nota 18](#_Toc3627486)

[La deviazione standard non è nota 19](#_Toc3627487)

[Stima della varianza di una popolazione 20](#_Toc3627488)

[Capitolo 5 – Verifica di un’ipotesi 21](#_Toc3627489)

[Prendere una decisione ed arrivare ad una conclusione 23](#_Toc3627490)

[Errori durante lo svolgimento di un test d’ipotesi 24](#_Toc3627491)

[Capitolo 6 – Inferenza per due campioni 25](#_Toc3627492)

[Inferenza su due proporzioni 26](#_Toc3627493)

[Inferenza su due medie: campioni indipendenti 26](#_Toc3627494)

[Inferenza su differenze per campioni appaiati 28](#_Toc3627495)

[Confronto della variabilità in due campioni 28](#_Toc3627496)

[Capitolo 7 – Correlazione e regressione 29](#_Toc3627497)

[La regressione 30](#_Toc3627498)

[Valori anomali e punti influenti. Minimi scarti quadrati 31](#_Toc3627499)

[Capitolo 8 – Analisi della varianza. Metodo ANOVA 31](#_Toc3627500)

[ANOVA ad una via 32](#_Toc3627501)

# Capitolo 1 – Definizioni, tipi di esperimenti, campionamento.

## Introduzione

La statistica studia gruppi ristretti di un gruppo più numeroso, cercando di scoprire qualcosa riguardo a quest’ultimo. Introduciamo i termini di dati, statistica, censimento, Campione e Popolazione:

* I **dati** sono le osservazioni, cioè le misure, raccolte.
* La **Statistica** consiste in metodologia per disegnare esperimenti, ricavare dati, e poi organizzarli e studiarli per ottenere e esporre risultati.
* La **popolazione** è l’insieme completo di tutti gli elementi in oggetto di studio.
* Un **censimento** è la raccolta di dati di ogni singolo elemento della popolazione.
* Un **campione** è un “sotto collezione” di membri della popolazione.

Vedremo che la selezione di una popolazione campione è estremamente importante per ottenere dei dati che, quando opportunamente manipolati, portano a conclusioni significative. Una cattiva selezione del campione porta molto probabilmente all’inutilità dei dati raccolti.

Possiamo avere due tipologie di dati, per differenziare se essi sono riguardanti alla popolazione, o ad un campione:

* Un **parametro** è una misura che descrive una caratteristica di un’intera popolazione.
* Una **statistica** è una misura che descrive una caratteristica di un campione.

Potremmo avere, ad esempio, sia il parametro “media delle età” che la statistica “media delle età”. Proprio per la definizione di popolazione, ottenere parametri è molto complesso se non impossibile. I metodi statistici si basano su, utilizzando statistiche, inferire sul possibile valore dei parametri.

## Tipologie di dati

Possiamo inoltre dividere i dati in base a di cosa sono costituiti:

* I dati quantitativi sono numeri, e questi numeri sono il risultato di conteggi o misure.
  + Ad esempio, Marco è alto 1.80 metri. Questo dato deriva da una misura diretta dell’altezza di Marco.
* I dati qualitativi sono codici numerici o di caratteri, e si riferiscono a categorie distinte in base a caratteristiche non numeriche.
  + Ad esempio, Marco è nato a Genova. Questo dato non arriva da una misura diretta, e ci dice soltanto che, qualitativamente, Marco è nato nella città “Genova”.

Da non confondere dati numerici come sempre quantitativi: ad esempio, il codice di immatricolazione “12345” non è un dato quantitativo, seppure numerico. Molti studi usano numeri per semplificare i dati qualitativi tabelle o raccolte di dati, ma questi numeri sono solo questo, codici.  
Ad esempio, uno studio sulla pressione potrebbe utilizzare “1” se il paziente è affetto da ipertensione, o “0” se non lo è. In questo caso, calcolare la media di questa statistica non avrebbe alcun senso.

Distinguiamo anche, tra i dati quantitativi, tra discreti e continui:

* I **dati discreti** corrispondono a casi in cui il possibile valore misurato può prendere solo un numero finito, o numerabile, di possibilità.
  + Ad esempio, “Uova deposte da galline in un giorno” è un dato discreto, dato che può prendere solo i valori della serie infinita, ma numerabile, 1, 2, 3, … Notare che, ovviamente, una gallina non potrà deporre “500” uova, ma neanche “1.3”.
* I **dati continui** corrispondono a quei casi in cui i valori possibili sono *infiniti* (o meglio, di più dell’infinità numerabile), e corrispondono ad una qualche scala continua, che copre un intervallo di valori senza interruzioni.
  + Ad esempio, “quantità di latte prodotta da una mucca” è un dato continuo, dato che può avere qualsiasi valore volumetrico positivo: 0, 0.00000…1, 0.00000…2, 1.2634…, etc…. Notare che, anche se in teoria il valore potrebbe avere un numero pressoché infinito di valori dopo la virgola, la sensibilità li limiterà sempre.

Possiamo inoltre dividere le variabili in quattro tipologie: nominali, ordinali, intervallari e rapportabili:

* Le **variabili nominali** sono variabili come il nome, una etichetta o una categoria. Questi dati non possono essere ordinati (se non in modo arbitrario).
  + Ad esempio, le risposte ad un sondaggio (Si, No, Non so), i colori di una pianta di pisello…
* Le **variabili ordinali** possono avere un ordine, ma le differenze tra i valori possono non essere calcolabili o non avere significato.
  + Ad esempio, i voti di un esame: A, B, C… sono ordinabili, cioè A è prima di C, ma le differenze tra di loro non sono calcolabili: A - B non può essere calcolato.
  + Oppure, la valutazione di aggressività come 1° più aggressivo, 2° più aggressivo… Possiamo ordinarli, ma 2 – 3 non ha significato: il cambiamento di aggressività tra il 1° e il 2° può essere diversa da quella del 2° e 3° e così via.
* Le **variabili intervallari** sono come quelle ordinali, ma qui le differenze di valori hanno un significato. In queste variabili, però, non esiste uno 0 di riferimento.
  + Temperature in celsius: tra 20 °C e 30 °C c’è una differenza, calcolabile, di 10°C. Tuttavia, 0 °C non è uno zero di riferimento, dato che è preso come convenzione. Dire, quindi, che 40 °C è *due volte più caldo* di 20 °C è sbagliato.
  + Gli anni: 2012, 2013, 2014. Possiamo calcolare le differenze (2012 – 2014 = 2 anni), ma l’anno “0” non significa “assenza di tempo”.
* Le **variabili rapportabili** sono come quelle intervallari, ma qui è presente uno zero di riferimento. Per queste, sia le differenze che i rapporti hanno un significato.
  + Il peso di un oggetto, l’età di una persona…

## Tipologie di esperimento

Possiamo raccogliere i dati in due modi distinti:

* Gli **studi osservativi** sono volti alla raccolta di dati da elementi, senza modificare nulla di quegli elementi.
  + Ad esempio, un sondaggio di opinione.
  + Uno studio osservativo si divide in:
    - Studio **trasversale**, dove i dati sono raccolti tutti nello stesso momento.
    - Studio **retrospettivo**, dove i dati sono riferiti al passato (e sono raccolti da archivi, sondaggi…)
    - Studio **di prospettiva**, dove i dati sono riferiti al futuro (raccolti seguendo nel tempo gruppi che condividono caratteristiche simili, dette **coorti**)
* Gli **studi pianificati**, invece, raccolgono dati da elementi dopo un *trattamento*.
  + Ad esempio uno studio clinico.

Gli studi pianificati possono essere affetti dal cosiddetto **confondimento**, cioè dalla mescolanza degli effetti di più variabili, che rendono i dati inutilizzabili. Ad esempio: somministriamo un vaccino, ma al tempo stesso variabili ambientali fanno scendere l’infettività della malattia. Il calo di incidenza è dovuto al vaccino o alla temperatura? In uno studio pianificato, quindi, è estremamente importante mantenere tutte le variabili costanti tranne quelle misurate.

Si sono adottati vari metodi per controllare le variabili e saper discernere tra gli effetti delle stesse:

* **Cieco**. Spesso usati negli studi clinici per rimuovere la variabile dell’effetto placebo. Tutti gli individui sono somministrati una pastiglia, ma solo nel gruppo trattato è realmente presente il principio attivo, e i pazienti non ne sono al corrente. Il doppio cieco c’è quando sia i pazienti che i medici che li controllano non sanno se i pazienti hanno preso o meno il principio attivo.
* **Blocchi**. Se si deve testare l’efficacia per uno o più trattamenti, e bene dividere il campione in blocchi con caratteristiche omogenee.
  + Ad esempio, dividiamo il campione in blocchi “maschi” e “femmine”, poi diamo il trattamento a metà maschi e metà femmine. Uno studio così è meglio preparato a capire le differenze di effetto del trattamento tra maschi e femmine.
* **Randomizzazione.** Al momento di assegnare i trattamenti ai pazienti, è di buona norma utilizzare una selezione casuale.
* **Dimensione campionaria.** Per minimizzare gli effetti che la naturale mutabilità dei singoli soggetti hanno nel mascherare il reale effetto del trattamento, è bene che il campione sia di grandi dimensioni.
* **Opportuna selezione del campione.** Ancora più importante della grandezza del campione è il modo in cui questo campione è selezionato. Le modalità più appropriate per selezionare un campione sono descritte più avanti.

## Strategie di campionamento

Se un campione non è opportunamente scelto, i dati prodotti possono rivelarsi completamente inutilizzabili. Vediamo i metodi di campionamento più comuni:

* Un **campione casuale**, gli individui di una popolazione sono scelti in modo che **ciascun singolo individuo ha la stessa probabilità di essere selezionato**. Un **campione casuale semplice** è in campione id taglia *n* in cui ogni gruppo di individui di taglia *n* possibile ha la stessa possibilità di essere scelto.
  + Il campionamento casuale semplice è da sempre il più adatto per ogni studio statistico, e porta ad avere risultati veramente inferibili sulla popolazione. È l’unico metodo utilizzato negli esempi da qui in avanti.
* Il **campionamento sistematico** prevede la selezione di un punto iniziale e la successiva selezione di ogni individuo in posizione k-esima dal punto iniziale.
* Nel **campionamento di convenienza** selezioniamo gli elementi per cui i dati sono facilmente raccoglibili.
* Nel **campionamento stratificato** si suddivide la popolazione in gruppi per una o più caratteristiche, e poi si seleziona un campione da ogni sottogruppo.
* Nel **campionamento a grappoli**, si suddivide la popolazione in sezioni, e le sezioni in gruppi, poi selezioniamo alcune sezioni, e poi alcuni gruppi in ciascuna sezione.
  + Un campionamento a grappolo con molti stage di selezione è anche detto **multistage**.

È facile confondere c. a grappoli e c. stratificato. In quello a grappoli, selezioniamo tutti gli individui da alcuni campioni di popolazione, in quello a strati selezioniamo un campione di popolazione da ciascuna categoria.

# Capitolo 2 - Analisi esplorativa dei dati: Statistica Descrittiva

In questo capitolo vedremo come analizzare i dati raccolti e ricavarne alcune informazioni. Questa è detta statistica descrittiva, e non tenta di inferire su parametri della popolazione, ma si limita a descrivere i dati raccolti.

## Caratteristiche dei dati

* Centro: è il valore rappresentativo o medio, che rappresenta il centro dei dati.
* Variazione: una misura di dispersione dei dati.
* Distribuzione: la forma della distribuzione dei dati (a campana, uniforme, asimmetrica, multimodale…)
* Outlier: dati che si discostano particolarmente dalla maggioranza delle osservazioni.
* Tempo: i cambiamenti dei dati nel tempo.

## Frequenze

In una **tabella delle frequenze** sono riportati i dati raccolti, singoli o, più comunemente, divisi in categorie, assieme alle frequenze i quali si sono presentati. Possiamo definire:

* I **limiti inferiori delle classi** sono i valori più piccoli appartenenti a ciascuna classe.
* I **limiti superiori delle classi** sono i valori più grandi appartenenti a ciascuna classe.
* I **separatori di classe** sono valori calcolabili che separano ciascuna classe. Si sottraggono i limiti superiori ed inferiori, si divide per due, e si sottrare il valore ottenuto a ciascun limite inferiore.
* I **punti medi** sono i valori centrali a ciascuna classe.
* L’**ampiezza delle classi** è la differenza tra limiti inferiori di due classi successive.

Da notare che il numero di classi è di solito arbitrario, ma un buon stimatore è , e la larghezza delle classi è

Le **Frequenze relative** sono le frequenze di un certo parametro diviso la taglia dei dati. La somma di tutte le frequenze relative deve dare 1, quindi bisogna fare attenzione ad eventuali arrotondamenti.

Le **Frequenze cumulate** sono le frequenze che si trovano “accumulando” via via le sequenze per ogni classe. La nuova tabella è, di solito, “Meno di X”, “Meno di Y” e così via. Si ottiene sommando le frequenze di una classe con quelle della classe prima.

L’analisi delle frequenze è di solito la prima analisi che si fa di qualche tipo di dato, e può produrre osservazioni interessanti (ad esempio su come i dati sono stati raccolti, come nell’esempio dei battiti al minuto, dove non ci sono valori dispari, e questo fa pensare ad una raccolta di dati che è stata poi moltiplicata per 2 o per 4).

## Rappresentazione grafica dei dati

Per studiare una distribuzione, è utile produrre un grafico. Vediamo i principali tipi di grafico che si usano per questo:

* Un **istogramma** è un grafico a barre che riporta sull’asse orizzontale delle ascisse (*x*) le classi in cui sono stati suddivisi i dati, e sull’asse verticale delle ordinate (*y*) le frequenze (o le frequenze relative). L’area (l’altezza) delle barre corrisponde al valore di frequenze, e le barre sono disegnate adiacenti l’una all’altra (per far vedere una continuità).
* Un **poligono delle frequenze** è una spezzata con in Y le frequenze, ed in x i punti medi delle classi. La linea spezzata è prolungata per toccare a sx e a dx l’asse delle ascisse.
* Il **Grafico delle frequenze cumulate** o **grafico ad Ogiva** è una spzzata in cui sull’asse x sono presenti i separatori superiori di classe (e il separatore inferiore della prima classe), e sull’asse y le frequenze cumulate.
* Un **dotplot** o **grafico a punti** è un grafico in cui ogni dato è rappresentato come un punto su di una scala graduata, e dati simili o uguali sono uno sopra l’altro.
* Un **diagramma ramo-foglia** o ***stem-leaf plot***è un grafico in cui i dati sono rappresentati elencando le cifre prima della virgola in ordine, e poi indicando, uno dopo l’altro, le cifre decimali di ogni dato in corrispondenza delle sue cifre prima della virgola.
* I **Diagrammi di Pareto** sono diagrammi a barre utilizzati per dati qualitativi, con in ascissa le classi, ed in ordinata le frequenze. Le barre sono sempre ordinate in modo decrescente da sinistra.
* I **Diagrammi a Torta** dividono l’area di un cerchio in spicchi, dove l’area di ogni spicchio è proporzionale alla frequenza (di solito la frequenza relativa).
* I **Diagrammi a Dispersione** o **Scatterplot** sono diagrammi in cui ad ogni punto (x, y) corrispondono ad una coppia di dati per cui vogliamo cercare una correlazione.

Esistono altri grafici, anche molto utili per alcune serie di dati, ma non sono affrontati qui. E’ utile ricordare che un grafico non è un risultato in sé, ma meglio un modo per descrivere, esaminare e confrontare tra loro serie di dati.

## Misure di centralità – Come valutare il centro

Una **misura di centralità** è un valore che si trova in qualche modo in mezzo alla distribuzione. Vediamo le misure di centralità più utilizzate:

* **Media**: La media aritmetica è una delle più importanti misure di centralità. Con il termine “media” si intende la media aritmetica. La media artimetica è quel valore che si ottiene sommando tutti i valori edividendo per il numero di valori:
  + **Definizione dei simboli**:  
     indica la somma di un insieme di dati.  
    x denota i singoli valori dell’insieme di dati.  
    n è il numero di valori in un campione  
    N è il numero di valori in una popolazione  
     è la media dei valori di un campione, o **media campionaria** è la media dei valori di una popolazione, anche chiamata solamente media.
* **Mediana**: La media è molto sensibile ai valori estremi della distribuzione. Per ovviare questo problema, si può trovare la mediana. La mediana di un insieme è il valore in posizione centrale (o la media dei valori in posizione centrale se n o N è dispari) dopo che i valori sono stati disposti in ordine crescente. È di solito indicata con .
  + La mediana è più rappresentativa per insiemi con pochi dati molto alti o molto bassi.
* **Moda**: La moda di un insieme di dati è il valore che si ripete più volte. È quindi il valore direttamente sotto il “picco” della distribuzione.
  + Se una distribuzione ha due valori con la stessa frequenza massima, prende il nome di **bimodale**.
  + Se una distribuzione ha più di due valori con la stessa frequenza massima, prende il nome di **multimodale**.
  + Se una distribuzione non ha valori ripetuti, si dice che la distribuzione non ha moda.
  + La moda non è molto usata per dati numerici, dato che spesso questi dati non hanno moda. È utile però per dati qualitativi come etichette o categorie.
* **Midrange**: Il midrange è definito come la media dei valori massimo e minimo.
  + È rarissimamente usato, proprio per la enorme sensibilità ai valori estremi. È da ricordare, però, perché è facile da calcolare, è utile per ricordare che ci sono molti modi per definire il centro di un insieme di dati, ed è spesso calcolato per sbaglio cercando la mediana.

Come regola generale di arrotondamento per questi calcoli, è utile riportare una cifra decimale in più di quelle nei dati. Così, la media di 2, 3 e 5 è 3.3.

Abbiamo anche altri modi per descrivere una distribuzione. Ne vediamo due che ci possono essere utili in seguito:

* **Media pesata** o **ponderata**: è usata quando i valori dei singoli dati possono avere diversa importanza, e volgiamo prendere in considerazione questo fatto nel calcolo della media.
  + Se x è un valore, e w è il peso corrispondente al valore x, allora la media ponderata è:
* **Asimmetria (Skewness):** una distribuzione è detta asimmetrica quando si estende in modo diverso dall’una e dall’altra parte. È facilmente visibile con un grafico, ma si può anche vedere osservando media, mediana e moda. Una distribuzione è asimmetrica a sinistra quando la media e la mediana sono a sinistra della moda. Una distribuzione è asimmetrica a destra quando la media e la mediana sono a destra della moda. Se media, moda e mediana coincidono, la distribuzione è simmetrica. Si può calcolare il valore di asimmetria con la formula (in cui sigma è la deviazione standard, definita più avanti):
* **Curtosi**: La curtosi è la misura di quanto la distribuzione è appiattita o “acuta”. La Normale gaussiana ha curtosi pari a zero. Essa si calcola con l’indice ρ (rho):

## Misure di dispersione – Come valutare l’ampiezza della distribuzione

La misura della dispersione di una serie di dati può essere molto utile per trarre conclusioni che non sono efficacemente deducibili dal semplice studio del centro. Vediamo alcuni indici utili per descrivere la dispersione:

* **Range**. Il range di una serie di dati è definito come il valore massimo meno il valore minimo. Come si nota, questo valore è poco significativo dato che è estremamente influenzato da valori estremi, e prende in considerazione solo due dati, magari in una collezione di molte osservazioni.
* **Deviazione Standard**. La deviazione standard è la misura di dispersione più utile ed importante. Definisce la variazione dei singoli valori rispetto alla media. È definita come:

Per un campione: oppure

Per una popolazione:

Ove x sono i dati, n è la numerosità del campione, N è la grandezza della popolazione, è la media campionaria e μ è la media della popolazione.

Notiamo che la deviazione standard non è mai negativa, ed è maggiore maggiore è la dispersione dei dati. L’u.d.m. della deviazione standard è identica a quella dei dati di partenza. La deviazione standard può anche essere chiamata con SD (Standard Deviation).

Possiamo anche interpretare la deviazione standard con una semplice regola empirica: Se la distribuzione di dati è a campana o simile ad una a campana, circa 68% dei dati cadono entro 1 SD dalla media; circa 95% dei dati cadono entro 2 SD dalla media, e il 99.7% dei dati cadono entro 3 SD dalla media.

Possiamo scomporre la formula della SD: innanzitutto, prende in considerazione la cosiddetta **deviazione** di ogni singolo dato “x”, . Poi combina questi valori in un unico numero, ma la semplice somma otterrebbe sempre 0. Per evitare che i dati negativi e positivi si annullino, essi sono portati a valori positivi tramite l’elevamento al quadrato. Nella SD, si applica la radice quadrata alla fine di tutti gli altri calcoli per compensare l’elevamento al quadrato. Semplicemente facendo la matrice della sommatoria della deviazione, e dividendo per n (), otteniamo la cosiddetta **deviazione assoluta media.** Dividiamo per per meglio descrivere la sommatoria al numeratore: solo n – 1 dati possono variare, dato che la media è fissata. Utilizzando n – 1, la varianza non sottostima sistematicamente la varianza di popolazione (come vediamo qui sotto).

* **Varianza**. La varianza di una serie di dati è definita come il quadrato della deviazione standard. Notare che la varianza campionaria è uno stimatore non distorto (cioè rappresenta bene, senza sovra o sotto-stimare) dellla varianza della popolazione.
  + La varianza campionaria è ;
  + La varianza della popolazione è

Dato che la varianza e la deviazione standard hanno la stessa unità di misura dei dati, può risultare difficile confrontare dati di origine e natura diversa. Per far questo, utilizziamo i **coefficienti di variazione**, o **CV**.

* Il **coefficiente di variazione** di una serie di dati è definito come la deviazione standard diviso la media in percentuale:

Per un campione: Per una popolazione:

## Misure di posizione – Dividere in parti la distribuzione di dati

Può essere utile dividere in parti la distribuzione per proseguire con la descrizione della distribuzione. Sono stati definiti alcuni valori utili in questo senso:

* **Quartili e percentili**. Come la mediana divide i valori esattamente a metà, i quartili la dividono in quarti. Definiamo tre quartili:
  + , o Primo Quartile: Separa il 25% inferiore dal 75% superiore dei dati. (25% ≤ ≤ 75%)
  + , o Secondo Quartile: Coincide con la mediana, separa il 50% inferiore dal superiore.
  + , o Terzo Quartile: Separa il 75% inferiore dal 25% superiore dei dati. (75% ≤ ≤ 25%)
  + , o Quarto Quartile, non è definito, e coinciderebbe con tutto l’insieme dei dati.

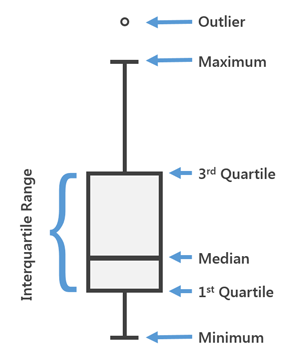
Per il calcolo dei quartili, c’è discostanza. Programmi diversi infatti producono risultati diversi.

* Come per i quartili, possiamo dividere la distribuzione in 99 **Percentili**, che dividono la serie di dati in 100 gruppi, ciascuno corrispondente all’ 1% della distribuzione. Possiamo calcolare il percentile associato ad un certo dato x come:
* I quartili coincidono con alcuni percentili, infatti , , (e ).
* Utilizzando quartili e percentili, possiamo introdurre altre statistiche:
  + Range interquartile:
  + Range semi-interquartile:
  + Midquartile:
  + Range Percentile 10-90:

## Analisi esplorativa dei dati – Trarre informazioni preliminari per uno studio

Riuscire a descrivere l’insieme di dati in un modo semplice ed efficace per vedere come i dati sono formati. Con il termine **analisi esplorativa dei dati** si intende l’uso di varie tecniche matematiche e grafiche per chiarire la struttur dei dati.

Preliminarmente, è importante cercare eventuali valori che sono ovviamente fuori dalla distribuzione di tutti gli altri dati. Questi valori sono detti **outlier**. Gli outlier possono influire pesantemente sulla media e sulla deviazione standard, o distorcere completamente un istogramma.

Alcuni Outlier sono frutto di errori: misurazione, battitura… e devono essere trovati e rimossi prima di continuare l’analisi. Altri outlier sono comunque valori corretti ma molto rari: in una qualsiasi distribuzione è possibile avere un valore ad esempio molto, molto grande, ma questo può essere molto improbabile. È buona pratica includere questi tipi di Outlier, ma è bene effettuare l’analisi sia utilizzandoli, sia escludendoli, per misurarne l’effetto.

Un grafico che bene esprime la forma dei dati è il **boxplot**, qui a sinistra. I Boxplot che indicano anche gli outlier sono chiamati **boxplot modificati**, mentre gli altri regolari.

È importante tenere a mente che i dati, oltre ad essere analizzati in questo modo, sono da considerare criticamente: Qual è l’origine dei dai? Il campione è scelto opportunamente? I dati sono stati raccolti opportunamente? Cosa voglio capire da questi dati? Il campione è abbastanza numeroso?  
Il pensiero critico è ugualmente importante al sapere maneggiare gli strumenti statistici più matematici.

Con “**riassunto in cinque numeri**”, si intende la panoramica che si ottiene calcolando Minimo, , Mediana, e Massimo. Per un’analisi preliminare, è utile calcolare il riassunto in cinque numeri assieme a Media e Deviazione standard.

# Capitolo 3 – Probabilità

Il calcolo delle probabilità è la base di tutti i metodi della statistica inferenziale. Il calcolo della probabilità interessa tutti gli eventi che sono casuali (aleatori), cioè non sono né certi né impossibili. La possibilità (o probabilità) che un evento succeda è un numero tra 0 e 1, dove 0 è l’impossibilità, e 1 è la certezza.

In statistica inferenziale, l’uso dello studio della probabilità può portare ad pensare che una certa assunzione (ad esempio un ipotesi, o spesso un evento complementare all’ipotesi) sia così improbabile da essere sbagliata, o, al contrario così probabile da essere esatta. Ovviamente, nella statistica una certezza perfetta non può verificarsi, ma, con un piccolo margine di tolleranza per gli errori, i metodi della statistica inferenziale sono alla basa di virtualmente qualsiasi studio scientifico.

Possiamo riassumere il contenuto del paragrafo precedente con la cosiddetta **legge degli eventi rari**: “*Se si è verificato un certo evento e, in base ad un’ipotesi, la probabilità che questo evento si verifichi è molto bassa, allora l’ipotesi è probabilmente errata*”.

Iniziamo a definire alcuni termini utili nello studio delle probabilità:

* Un **esperimento** è una azione, od un insieme di azioni, che portano alla produzione di dati.
  + Ad esempio, il lancio di un dado (a sei facce) è un esperimento che porta alla produzione di un risultato ogni volta che è ripetuto.
* Un **esito** è un singolo risultato di un esperimento.
  + Nel lancio del dado, il singolo tiro risulta essere un 6. Questo è un esito.
* Un **evento** è una raccolta di *esiti* di un esperimento.
  + Nel lancio del dado, “Esce un numero pari” è un evento, composto dagli esiti 2, 4, 6.
* Un **evento elementare** è un evento che non può essere suddiviso in eventi più semplici.
  + “Esce un 3” è un evento elementare, mentre “Esce un numero dispari” non lo è (può essere diviso in 3 eventi più semplici, cioè “Esce 1”, “Esce 3” ed “Esce 5”)
* Lo **spazio campionario** di un esperimento è l’insieme che contiene tutti i possibili esiti. Di solito è indicato con Ω.
  + Nel lancio del dado, Ω = [1, 2, 3, 4, 5, 6]
* **P** indica una probabilità. indica la probabilità dell’evento .
* (maiuscolo) indica una **variabile**. Una variabile è, in un esperimento, una caratteristica che è misurata.
* (minuscolo) indica l’esito (numerico) di una variabile dopo che l’esperimento è stato effettuato.

Ci sono vari modi per trovare la probabilità di un evento. I tre più importanti sono:

* **Probabilità soggettiva**. P(A) viene stimata conoscendo tutte le circostanze significative sull’avvenimento di A. Questa è la probabilità meno affidabile, ma per eventi molto poco probabili, o molto probabili, questo calcolo può non essere molto discostante dalla realtà.
  + Qual è la probabilità che un meteorite cada sulla mia macchina nel prossimo mese? Sapendo che i meteoriti cadono rarissimamente sulla terra, possiamo stimare che la P(x) sia molto, molto bassa, diciamo 1e-10.
* **Probabilità classica**. Se tutti i possibili eventi sono equiprobabili, e conosciamo tutto lo spazio campionario, allora la probabilità esatta può essere semplicemente calcolata utilizzandola formula:
* **Approssimazione della probabilità con la frequenza relativa**. Se non conosciamo lo spazio campionario, o gli eventi non sono equiprobabili, allora possiamo calcolare la probabilità di un certo evento come:

Notare che all’aumentare delle ripetizioni dell’esperimento, la probabilità con la frequenza relativa diventa sempre più uguale alla probabilità reale (effettiva). Questo è detto **legge dei grandi numeri**.

Se prendiamo in considerazione una variabile continua, essa può prendere un numero infinito di valori, e ciascun valore avrà una probabilità diversa di accadere. Possiamo quindi mostrare questa probabilità con una funzione, chiamata **funzione di probabilità** o **curva di densità di probabilità**, indicata con . Per trovare la probabilità che, ad una certa ripetizione di un esperimento, *x* sia tra i valori a e b, basta integrare la funzione di probabilità tra a e b:

è anche chiamata **funzione di distribuzione**. Notiamo due cose da questa definizione:

1. è sempre pari a 0, proprio per la definizione di integrale definito.
2. è sempre pari ad 1 (L’evento certo).

Per essere chiamata curva di densità, quindi, la funzione deve soddisfare due requisiti:

1. L’area totale sottesa dalla curva è pari a 1, quindi c’è una relazione area/probabilità; e
2. Ogni punto sulla curva deve essere ≥ a 0 (non esistono probabilità negative).

## Operazioni con le probabilità – Probabilità di due eventi

**1. Eventi complementari**. A volte ci è utile sapere qual è la probabilità che un evento A non accada. L’**evento complementare** di A, indicato con , è un insieme dello spazio campionario che contiene tutti gli esiti in cui A *non* si verifica. Possiamo calcolare semplicemente la probabilità di conoscendo :

**2. Probabilità sommate.** . Possiamo chiederci quale sia la probabilità che avvenga l’evento A, l’evento B, o entrambi. Questo è rappresentato dalle espressioni o . Per calcolarla, utilizziamo la **regola della somma** dopo aver calcolato sia che :

Dobbiamo rimuovere perché, se non lo facessimo, conteremo due volte gli eventi in cui A e B avvengono assieme. Nota che è anche indicato come .

**3. Probabilità prodotte.** . Per calcolare la probabilità che un evento A si verifichi contemporaneamente ad un evento B (e questi due eventi sono indipendenti tra loro, vedi punto 4), possiamo utilizzare la cosiddetta **regola del prodotto**:

Questa regola è facilmente dimostrabile empiricamente: Prendiamo un test formato da due domande, A e B. La prima (A) è una domanda Vero o Falso, e la risposta esatta è Falso. La seconda domanda (B) è chiusa, con le possibili risposte formate da a, b, c, d, e, in cui la risposta esatta è la “c”. Per calcolare la probabilità che, casualmente, uno studente risponda in maniera corretta ad entrambe le domande, possiamo utilizzare il metodo classico:

Notando che: , e ; salta subito all’occhio che .

**4. Probabilità condizionata. B in seguito ad A.** Se un certo evento B è condizionato dall’accadersi di un altro evento A, allora ci troviamo davanti ad una probabilità condizionata: “B dato A” o , quindi .

I due eventi, A e B, possono sia essere **dipendenti** tra di loro, cioè che l’esito di uno influenze l’esito dell’altro, sia **indipendenti**, cioè che l’accadere di uno non influenza l’altro.

Nel caso di due eventi **indipendenti**, , dato che l’avvenire di A non ha cambiato la probabilità dell’avvenire di B. L’esempio del paragrafo precedente è del tipo indipendente.  
Nel caso di due eventi **dipendenti**, allora , dato che, essendo già avvenuto A, la probabilità di B cambia. Ad esempio, estraendo due numeri da una scatola contenente tutti i numeri da 1 a 10 compresi, alla prima estrazione, la probabilità di estrarre un numero particolare “a” dalla scatola è , mentre alla seconda estrazione lo spazio campionario è cambiato, e la probabilità di estrarre un secondo numero “b” diverso da “a” è .

Nel calcolo di , se A e B sono **dipendenti** tra loro, non è più corretto utilizzare la regola del prodotto vista prima, ma è giusto usare la **regola formale del prodotto**:

La regola del prodotto è “estensibile” per più di due eventi, basta moltiplicare via via tra loro tutti gli eventi interessati. Se gli eventi sono tra loro dipendenti, allora la probabilità deve essere opportunamente ritoccata per tutti gli eventi oltre al primo.

Notiamo che, negli esperimenti che riguardano l’estrazione di più campioni, ci si potrebbe porre la domanda: “Estraendo più campioni contemporaneamente, il secondo e successivi campioni sono realmente scelti in modo casuale semplice?”. La probabilità che un soggetto del primo campione sia scelto per il secondo, proprio per il metodo di campionamento, è 0. Questo influisce sulla casualità del campionamento, come già detto essenziale per la significatività dei dati? In generale, **se un campione rappresenta meno del 5% della popolazione**, la probabilità che, potendo riselezionare gli individui, si selezioni di nuovo un individuo già selezionato è così bassa che questo problema non si pone, e le scelte successive possono essere considerate **indipendenti** tra loro.

Come regola generale, possiamo definire anche matematicamente con:

Questa formula è utile se non conosciamo esattamente gli effetti di A su B.

È importante sottolineare che le espressioni e **non** sono equivalenti.

**5. Regola del prodotto 2: “almeno uno”**. Potremmo volerci chiedere la probabilità che almeno in una ripetizione dell’esperimento, l’evento A avvenga. Per chiarirci, “Almeno uno” significa “uno o più”, e il complementare di “almeno uno” è “nessuno”. Allora:

Possiamo calcolare facilmente la probabilità di “nessuno” con la regola del prodotto. Ad esempio, la probabilità che su tre nascituri almeno uno sia femmina è:

**6. Correggere la probabilità alla luce di nuove informazioni. Il teorema di Bayes.** In un paese, il 51% delle persone sono uomini, e il 49% sono donne. Estraiamo a caso una persona che fuma il sigaro: Qual è la probabilità che essa sia un uomo? Dato che non sappiamo se ci sono relazioni tra fumare il sigaro e essere uomo, possiamo semplicemente dire che la probabilità sia il 51%, uguale all’estrarre casualmente una persona.

Successivamente, troviamo uno studio recente che dimostra che il 9.5% dei maschi sono fumatori di sigaro, mentre solo l’1,5% delle femmine lo sono. Ora potremmo voler cambiare la nostra probabilità alla luce di questo nuovo fatto: per far ciò, utilizziamo il **teorema di Bayes**:

Possiamo riassumere i nostri dati con: . Applicando il teorema di Bayes, otteniamo 0.853, cioè la probabilità che avendo selezionato un fumatore esso sia un uomo è 85,3%.

## Variabili Aleatorie

Iniziamo a definire cosa intendiamo con “variabile aleatoria”:

* Una **variabile aleatoria** è una variabile che assume un singolo valore numerico, determinato in base al caso, per ciascuno degli esiti dell’esperimento. Abbiamo già incontrato queste variabili nella sezione riguardante la probabilità.
  + Una variabile aleatoria **continua** assume un numero infinito ed innumerabile di valori, associabili ad una scala continua, senza salti o interruzioni.
  + Una variabile aleatoria **discreta** assume un numero finito o comunque numerabile di valori.
* Una **distribuzione delle probabilità** è un grafico, tabella o funzione che fornisce la probabilità di ciascun valore della variabile aleatoria. Come abbiamo già visto prima, solo una funzione riesce a rappresentare in maniera sensibile una variabile aleatoria continua.

Un esempio di Distribuzione di probabilità è rappresentato dall’**istogramma delle probabilità**, un istogramma in cui sull’asse delle ascisse (x) sono mostrate le classi, mentre sull’asse delle ordinate (y), sono mostrate le probabilità che il valore cada dentro ognuna classe. Notare che l’area totale dell’istogramma è pari ad 1, quindi , e che la probabilità di ogni possibile valore di x è sempre compresa tra 0 ed 1.

Possiamo calcolare centro, variazione, distribuzione ed outlier anche per questo tipo di dati:

* Media della d.d.p:
  + Deriva da: ma
* Varianza della d.d.p: oppure
  + Ottenuta in maniera simile alla media dalle formule di varianza e SD mostrate prima.
* SD della d.d.p:

{Ricordare: Potenze, poi moltiplicazioni, poi somme}

Vediamo di seguito alcune distribuzioni particolari che possono essere utili nello studio delle probabilità di variabili aleatorie:

## Distribuzioni binomiali – Vero o Falso

Quando trattiamo variabili con solo due possibili valori, ad esempio Vero o Falso, oppure Si o No, siamo di fronte ad una **distribuzione binomiale**.

* Con **distribuzione binomiale** intendiamo la distribuzione che i dati presentano se:
  1. L’esperimento a un numero fissato di prove;
  2. Le singole prove sono indipendenti tra loro;
  3. L’esito delle prove appartiene a solo due tipologie (dicotomico), ad esempio “risposta esatta” o “risposta sbagliata”;
  4. Le probabilità di ciascun risultato devono rimanere costanti tra prove diverse.

Alcuni simboli sono ricorrenti nelle distribuzioni binomiali:

* S = Success e F = Failure indicano i due possibili esiti.
* , e quindi
* indica il numero di prove.
* indica il numero di successi nelle prove, quindi ha valori interi tra 0 e inclusi.
* indica la probabilità di avere successi in prove.

Notare che l’indicazione *successo* può denotare anche una cosa negativa, in base a cosa stiamo considerando. Un errore comune è confondere e con due categorie diverse.

Come abbiamo già visto, il punto 2. dei requisiti per una distribuzione binomiale è rispettato anche in un campionamento senza reimmissione purché il campione rappresenti meno del 5% della popolazione.

Per una distribuzione binomiale, possiamo calcolare conoscendo utilizzando la **formula delle probabilità binomiali**:

Questa formula tiene conto di tutte le possibili permutazioni delle singole prove. Ad esempio, cerchiamo la probabilità di avere 3 successi e due fallimenti su 5 prove. Se utilizzassimo la semplice regola del prodotto, otterremmo le probabilità di avere, in ordine, 3 successi seguiti da 2 fallimenti. Utilizzando la formula delle probabilità binomiali di sopra, otteniamo la probabilità di avere una qualsiasi delle permutazioni (il numero delle quali è pari a ) che contengono almeno 3 successi e 2 fallimenti.

Possiamo riscrivere le formule viste sopra per una distribuzione generica, semplificandole conoscendo le definizioni dei simboli visti per le distribuzioni binomiali:



La formula per la media è intuitiva, mentre quelle per la SD e la varianza non lo sono altrettanto. Omettiamo i calcoli per arrivarci.

## Distribuzione di Poisson – Eventi rari

La distribuzione di Poisson è una distribuzione di probabilità discreta che è impiegata per descrivere la probabilità che un evento si verifichi in un certo intervallo specifico, cioè un tempo, un’area, una distanza… specifica. La probabilità che un evento avvenga volte in un intervallo è definita come:

Dove è il numero di Nepero, circa 2.71828 e μ è la media. La distribuzione di Poisson ha σ pari a .

Una variabile, per avere una distribuzione di Poisson, deve avere i seguenti requisiti:

* deve indicare il numero di volte che l’evento è soddisfatto in un certo intervallo.
* Gli eventi devono verificarsi in maniera indipendente l’uno dall’altro.
* Gli eventi devono verificarsi in modo casuale.

Dato che la distribuzione di Poisson può essere complessa da comprendere, utilizziamo un esempio esplicativo:

Nella Seconda guerra mondiale, la zona a sud di Londra è stata divisa in 576 regioni, ognuna di 0.25 Km quadrati. In questa zona, sono cadute un totale di 535 bombe. In questo caso, abbiamo tutti i requisiti per una distribuzione di Poisson:

* Abbiamo un numero per l’evento = bombe cadute.
* Ogni bomba è caduta indipendentemente da un’altra.
* Le bombe sono state sganciate in maniera pressoché casuale.

Possiamo calcolare μ:

Cerchiamo, ad esempio, la probabilità che siano cadute esattamente due bombe () in una regione qualsiasi:

Abbiamo una probabilità del 17% che, in una qualsiasi regione, cadano esattamente due bombe. Possiamo anche calcolare quante regioni (in teoria) siano bombardate due volte: . Il numero reale è 93.

## Distribuzione normale – Campana di Gauss

Se una variabile aleatoria continua ha una distribuzione che presenta un grafico simmetrico, con una forma riconducibile ad una campana e può essere descritta dalla formula:

Essa è detta avere una **distribuzione normale**, o semplicemente normale. Di solito, questa distribuzione prende il nome di .

Una **distribuzione normale standard** è una distribuzione normale con . La distribuzione normale standard è anche chiamata .

Dato che la funzione per la normale è definita solo da σ e μ, possiamo disegnarle molto facilmente dopo un’analisi preliminare dei dati. Come abbiamo già detto prima, c’è una correlazione diretta tra area a probabilità. Se una variabile è descrivibile da una normale particolare, possiamo calcolare la probabilità utilizzando la formula incontrata prima per le variabili continue:

La funzione che descrive la normale è però non elementarmente integrabile, cioè non possiamo calcolarne in maniera esatta la primitiva. Però, **I valori di probabilità per tutti i casi del tipo** , **sono stati pre-calcolati e tabulati nel caso della normale standard.**   
Possiamo calcolare anche i casi semplicemente con .

Per calcolare la probabilità, quindi, che una variabile aleatoria che segue una distribuzione normale standard cada tra i valori ed dove , cioè possiamo utilizzare le definizioni viste prima:

In pratica, però, una serie di dati può non seguire la normale standard, ma una normale con μ e σ diversi. Come proseguiamo per calcolare le probabilità in questo caso, dato che la funzione non è integrabile? Possiamo riportare i valori di della normale non standard a valori di della normale standard in modo tale che grazie alla formula seguente:

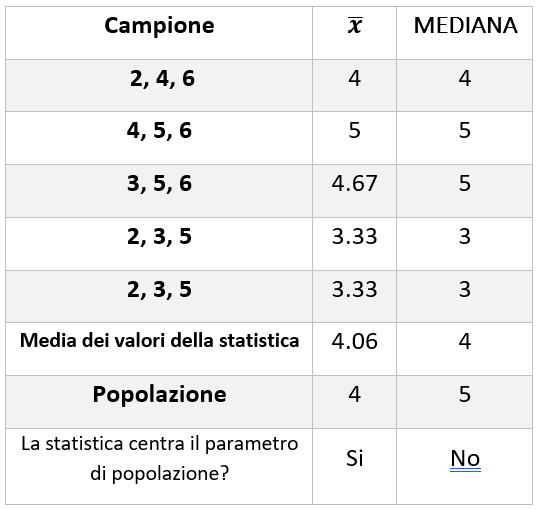
E possiamo “tornare indietro” (ad esempio, nel caso partiamo da una probabilità e vogliamo trovare il valore di corrispondente) dalla formula di conversione semplicemente facendo .

## Distribuzioni campionarie di una statistica

Un centro di analisi è rimasto aperto per cinque giorni prima di chiudere a causa del personale. In questi cinque giorni, sono stati visitati, in ordine per ogni giorno, 3, 2, 5, 4 e 6 pazienti. Dato che conosciamo l’intera **popolazione** “pazienti del centro”, possiamo calcolare μ e σ esattamente:

Di solito, però, non conosciamo ogni valore della popolazione, dato che trattiamo popolazioni estremamente grandi. In questi casi, prendiamo in considerazione solo un piccolo campione della popolazione, e, basandosi sulle caratteristiche di questo campione, cerchiamo di stimare alcune caratteristiche della popolazione.

Ad esempio, possiamo prendere dalla “popolazione di giorni” vista prima solo alcuni giorni, ad esempio 3, e calcolare la media di ciascuno. Dato che parliamo di campioni, dobbiamo parlare di *media campionaria*. Tabuliamo alcuni campioni (estratti con rimpiazzo), tutti scegliendo una taglia campionaria pari a 3, e calcoliamo alcuni parametri:

Calcolando la media di ciascun campione, possiamo notare che (ignorando la duplicazione dell’ultimo risultato, nata dal fatto che la popolazione è molto piccola) ogni media campionaria è diversa dalle altre. In generale, possiamo considerare ogni media campionaria diversa da tutte le altre medie di ogni altro campione.

Se calcolassimo tutte le media campionarie per ogni permutazione della popolazione, potremmo sapere ogni possibile valore che può prendere. Quindi, la probabilità che un campione casuale abbia una particolare media, è 1 sul numero di permutazioni. **Ogni media campionaria ha quindi la stessa probabilità di essere estratta.**

Immaginiamo ora di estrarre un numero di campioni elevato, tutti di taglia , da una popolazione di grandezza , per cui . In questo caso, la media campionaria di ogni campione (che sarà estratto senza reimmissione, ma questo non ci interessa dato che ) può coincidere in due campioni diversi. Ora immaginiamo di tabulare in una tabella di frequenze, divise in classi appropriate, tutte queste medie trovate. Possiamo calcolare la probabilità, utilizzando il metodo della frequenza relativa, che, estraendo un nuovo campione, la sua media cada in ognuna delle classi.  
Teniamo a mente questo concetto probabilistico, che utilizzeremo tra poco.

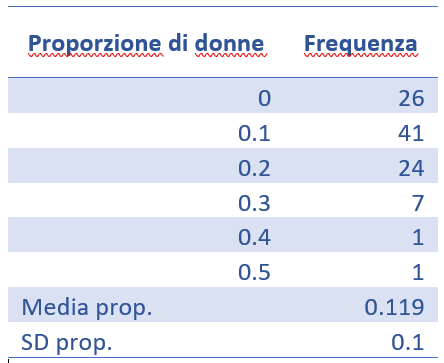
Dato che parliamo di probabilità, possiamo disegnare una **distribuzione campionaria della media**: essa è la distribuzione di probabilità per tutte le medie campionarie (o le classi di medie campionarie, come è più comune). Possiamo inoltre definire la **variabilità campionaria** di una statistica è la variazione di questa statistica a causa della casualità della variabile presa in questione.

Possiamo inoltre notare che la media delle medie campionarie è molto simile alla media della popolazione. Se conoscessimo tutte le possibili medie campionarie, e calcolassimo la media di queste, essa centrerebbe perfettamente la media della popolazione. In generale, possiamo dire che **la media di tutte le possibili medie campionare data una taglia è uguale alla media della popolazione**, e si deduce che **la media campionaria data una taglia diventa via via più vicina al valore reale man mano che aumentano il numero di campioni**.

## Proporzioni -Distribuzione campionaria di una proporzione

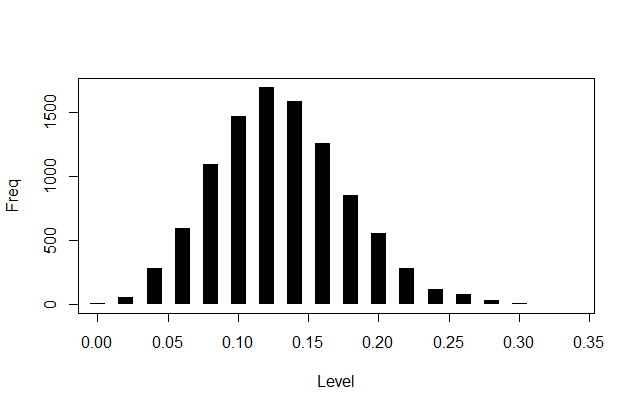
Prendiamo in considerazione un altro esempio. In un’aula di cento universitari (), solo tredici sono femmine. Possiamo calcolare la **proporzione** di donne nella classe con . Immaginando di non conoscere la popolazione, estraiamo a sorte, con rimpiazzo (dato che la taglia campionaria è > 5% della popolazione), cento campioni con taglia e calcoliamo la proporzione di donne sul totale del campione per ognuno dei campioni.

Le frequenze sono riportate in tabella:

Se permutassimo tutti i possibili campioni da 10 persone (1.73 possibili campioni) e facessimo la media delle proporzioni di donne otterremmo 0.13. In cento campioni, otteniamo una media delle proporzioni di 0.12, che è abbastanza vicina al valore reale.

In generale, **le proporzioni campionarie tendono a centrare il valore delle proporzioni di popolazione, diventando via via più precise maggiore è il numero di campioni raccolti.**

Come si potrebbe notare dalla tabella a fianco, le proporzioni campionarie hanno una classe di frequenza massima (0.1), e la proporzione scende sempre di più allontanandoci da questa classe.

Se ripetiamo lo stesso esperimento, ma prendendo 10'000 campioni con , otterremmo un grafico generalmente come quello a destra. Possiamo notare che la distribuzione delle probabilità sembra seguire molto bene una normale.

Possiamo dire che, in generale, **le distribuzioni delle proporzioni campionarie, sotto alcune condizioni, sono approssimabili da una distribuzione normale.**

Come abbiamo visto, alcune statistiche riescono a centrare bene i parametri della popolazione, mentre altri no. Le statistiche che centrano i parametri corrispondenti sono detti **stimatori non distorti**, mentre quelli che non li centrano sono detti **stimatori distorti**. Per riassumere:

* **Stimatori non distorti**: media, varianza, proporzione.
* **Stimatori distorti**: mediana, range, deviazione standard.

## Il teorema del Limite Centrale

Ricordando la definizione di *variabile aleatoria*, di *distribuzione delle probabilità* e *distribuzione della media campionaria* possiamo introdurre il teorema del limite centrale.

Il **teorema del limite centrale** afferma che, se la taglia del campione è abbastanza grande, la distribuzione delle medie campionarie può essere approssimata ad una **distribuzione normale**, anche se la popolazione non ha questa distribuzione.

Questo teorema ci permette di descrivere qualsiasi popolazione, ignorando la distribuzione della popolazione stessa, con una distribuzione normale se possiamo estrarre campioni sufficientemente grandi. Vediamo il teorema in dettaglio:

* **Ipotesi**
  + La variabile aleatoria ha una distribuzione (anche non normale), con media e deviazione standard .
  + Selezioniamo dalla popolazione campioni tutti di taglia in modo casuale semplice.
* **Conclusioni**
  + Se cresce, la distribuzione delle medie campionarie si avvicina alla distribuzione normale.
  + La media della distribuzione delle medie campionarie (la media delle medie) è la media della popolazione .
  + La deviazione standard della distribuzione delle medie campionarie è pari a .

Per parlare di media e SD della distribuzione delle medie campionarie occorrono nuovi simboli:

* Media delle medie campionarie:
* SD delle medie campionarie:
  + è anche chiamato **errore standard della media**.

Il teorema del limite centrale è ben visibile negli esempi fatti prima al riguardo delle medie campionarie. Anche se le possibili medie campionarie hanno tutte la stessa probabilità di essere estratte (), la distribuzione delle medie campionarie assomiglia ad una normale proprio per il fatto che tutte le medie hanno la stessa probabilità di estrazione.

Per utilizzare la distribuzione normale delle medie, però, probabilmente dovrà essere trasformata in una normale standard. Dobbiamo ricordare di utilizzare sia media che SD della distribuzione delle medie, sostituendola nella formula vista prima:

Questa formula è applicabile se conosciamo a priori della popolazione. Se non conosciamo , dobbiamo stimarla partendo dalle medie campionarie. Come abbiamo visto, la deviazione standard campionaria è uno stimatore distorto di , allora dobbiamo necessariamente calcolare la deviazione standard partendo dalla varianza (che non è distorto). Se è il numero di campioni, è la media di tutte le medie campionarie e è ogni singola media campionaria, allora:

## Accertare la normalità di una distribuzione

Possiamo controllare che una distribuzione è normale seguendo il processo generale:

1. **Istogramma**. Se la distribuzione non ha una forma a campana, rifiutiamo la normalità.
2. **Outlier**. Rifiutiamo la normalità se c’è più di un *outlier*.
3. **Grafico dei quantili normali**. Se i punti 1. E 2. sono soddisfatti, costruiamo il grafico dei quantili normali. In questo grafico, l’asse x mostra i valori originari della distribuzione, e l’asse y mostra i corrispondenti valori di dopo aver standardizzato . Se i punti non giacciono su una linea retta, o se seguono un disegno sistematico che non è una linea retta, rifiutiamo la normalità
4. Se tutti gli altri punti sono soddisfatti, allora non possiamo rifiutare la normalità della distribuzione.

# Capitolo 4 – Stime con un campione

Ora iniziamo a vedere come usare le statistiche di un campione per inferire sui parametri di popolazione. Questo è il vero nucleo della statistica inferenziale, e con questi metodi possiamo sia stimare il valore di un parametro partendo da una statistica, sia tastare alcune ipotesi circa la popolazione partendo da dei campioni.

## Stima delle proporzioni in una popolazione

Ora stimeremo la proporzione di una popolazione partendo da una proporzione campionaria. Per poter far questo, dobbiamo utilizzare dei dati con alcune caratteristiche:

* Il campione deve essere casuale semplice;
* La distribuzione deve essere di tipo binomiale: ci sono un numero finito di prove, ogni prova non influenza il risultato di un'altra prova, ci sono solo due possibili esiti (successo ed insuccesso) e la probabilità di successo rimane invariata tra le diverse prove.
* Prendiamo in considerazione solo le proporzioni per cui la distribuzione campionaria è assimilabile ad una normale. Considerando un evento binomiale, se è il numero di prove e è la probabilità di successo, e , la distribuzione delle proporzioni è normale. Dato che (e quindi anche ) sono incognite, dovremo utilizzare le proporzioni campionarie per stimarne i loro valori e accettarsi che queste condizioni sono soddisfatte.

Notiamo alcuni simboli utilizzati più avanti:

* è la proporzione di successi () della popolazione.
* è la proporzione campionaria di successi sul campione di prove di taglia .
* , cioè la proporzione di fallimenti su un campione di taglia .

Nota: possiamo considerare proporzioni anche probabilità e percentuali oltre che ad una generica proporzione , ma bisogna ricordarsi di trasformare da un valore percentuale ad un valore semplice queste due misure. Ad esempio, una probabilità del 22.3% è pari ad una probabilità di 0.223.

Ad esempio, se il 30% dei baccelli in un campione di 530 piante sono gialli, allora e .

Se vogliamo stimare la proporzione di popolazione con un singolo valore, la migliore stima sarà . In questo caso, è **stimatore puntuale** del parametro proporzione. In generale, uno stimatore puntuale è un singolo valore (un punto) usato per stimare un parametro di popolazione. Utilizziamo perché è uno stimatore non distorto, come abbiamo visto prima. Inoltre, la proporzione campionaria è lo stimatore più robusto tra quelli che possiamo utilizzare, cioè la deviazione standard di dal valore di popolazione, rispetto ad altri stimatori di , è minore.

## Gli intervalli di confidenza

Abbiamo detto che è il miglior stimatore di , ma abbiamo alcuna indicazione di *quanto* sia buona questa stima. Dato che la stima puntuale in sé non rivela la propria qualità, si sono messi a punto metodi per stimare il valore di popolazione non più con un punto, ma utilizzando un **range** di valori per poter avere una certa confidenza che il valore vero di cada in questo intervallo.

Un **intervallo di confidenza** (IC) è un range di valori usati per stimare il valore vero del parametro di popolazione.

Un intervallo di confidenza è calcolato partendo dal **livello di confidenza** (o **grado di fiducia**, o **coefficiente di confidenza**) scelto dall’operatore (ad esempio, 95%). Questo valore ci dà la percentuale di successo che, dopo aver calcolato l’intervallo di confidenza, esso contenga effettivamente . Quindi, se scegliamo 0.95, su un certo numeri di intervalli di confidenza calcolati per prove diverse, il 95% di questi conterranno . Bisogna sottolineare che il livello di confidenza **non è una probabilità**: quando calcoliamo l’intervallo per un certo set di dati, otterremo sempre lo stesso range. È quindi errato parlare di probabilità di successo, ma solo di confidenza di successo.

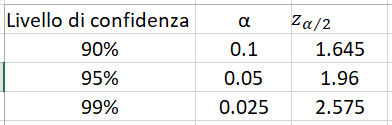
Di solito, il livello di confidenza è indicato come , dove è il complementare del livello di confidenza desiderato, ad esempio 0.05 per 0.95 di confidenza.

Il livello di confidenza determina pesantemente la forma dell’intervallo: per avere un 100% di confidenza che cada nell’intervallo, dovremmo prendere, in teoria, un intervallo di range infinito. Quindi, man mano che il grado di confidenza sale, l’intervallo cresce di range. Il contrario vale per un grado di confidenza decrescente.

La distribuzione delle proporzioni è una normale. Possiamo quindi trovare un valore di che divide le proporzioni che hanno più probabilità di essere esatte (verso il centro della distribuzione), da quelle che ne hanno di meno. Definiamo ora questo valore, considerazione per considerazione, chiamato **valore critico**:

* Sotto alcune condizioni, la distribuzione campionaria delle proporzioni è una normale.
* Le proporzioni che hanno probabilità più piccola di sono nelle “code” destra e sinistra della distribuzione normale. L’area delle code (entrambe sommate assieme) è quindi pari ad . L’area di una singola coda sarà .
* C’è una probabilità pari a che le proporzioni cadano nella parte più centrale della distribuzione.
* Il valore di che separa la coda destra (quindi , e l’area a destra partendo da è pari a ) è indicato come . è chiamato **valore critico**.

Per riassumere: Il **valore critico** è il valore positivo di in corrispondenza della linea verticale che separa un’area pari ad nella coda destra della distribuzione normale standard.

Per trovare il valore critico, iniziamo a definire . Se vogliamo un livello di confidenza del 95%, allora , quindi l’area sottesa dalla distribuzione dopo il valore critico sarà , cioè . La tabella che relaziona alle corrispondenti aree dà solo le aree da verso sinistra, quindi utilizziamo l’area di , che trova il valore di , cioè il nostro valore critico .

Altri valori critici, trovati nello stesso modo, sono ripotati nella tabella a fianco. Il livello di confidenza più usato è 95%, ma possiamo anche utilizzare 90% o 99%.

Ora definiamo un altro valore che è importante per calcolare gli intervalli di confidenza, il **margine di errore**. Il margine di errore, denominato , è la differenza massima tra la proporzione del campione e il vero valore della popolazione. Il margine di errore è calcolato moltiplicando la deviazione standard campionaria per il valore critico:

**Quindi, dato un , la proporzione campionaria sarà errata di una quantità non superiore ad con una probabilità di , e errata di una quantità superiore ad con una probabilità :** Questo si prende come viene?

Questo che abbiamo appena trovato è proprio l’**intervallo di confidenza** **per la distribuzione binomiale**. Spesso l’IC è segnato come .

Ricordiamo che l’intervallo di confidenza non è una probabilità: è una costante, quindi non ha senso dire “*abbiamo una probabilità del 95% che cada tra a e b*”, dato che il valore di non cambia, e non c’è nessuna probabilità coinvolta.

Quanto grande deve essere per ottenere un buon valore di ? Posiamo calcolare la taglia campionaria ideale per un certo esperimento in due modi, partendo sempre dalla formula inversa di . Se conosciamo già una stima di (magari da studi più vecchi), possiamo semplicemente applicare la formula del margine di errore in modo inverso:

Se non conosciamo una stima della proporzione campionaria, allora sostituiamo con il massimo valore che quella moltiplicazione può avere (se e ), cioè 0.25:

Ricordare di arrotondare sempre per eccesso, così da avere un abbastanza grande. Notare che non dipende da , cioè dalla taglia della popolazione.

## Stima della media di una popolazione

Per stimare la media occorre utilizzare ,la deviazione standard. La SD della popolazione potrebbe essere conosciuta (magari da altri studi) o essere incognita, e il metodo per calcolare la stima della media varia nei due casi. Il caso che la deviazione standard è conosciuta mentre la media di popolazione non lo è è raro, ma è utile considerarli entrambi per capire i metodi statistici al di sotto delle considearazioni:

### La deviazione standard è nota

Come prima, abbiamo dei requisiti:

1. Il campione è casuale semplice.
2. è noto.
3. Sono soddisfatti almeno uno di questi due requisiti di normalità per la distribuzione delle medie campionarie:
   1. La popolazione ha una distribuzione normale.
   2. . Questo rende, qualsiasi sia la distribuzione della popolazione, la distribuzione delle medie campionare una normale, per il **teorema del limite centrale**. Il valore di 30 basta dato che questo metodo è considerato **robusto**, cioè non è influenzato pesantemente dalla non perfetta normalità della distribuzione. Comunque, più cresce, migliori sono le stime.

Per iniziare, ricordiamo che la media campionaria è la migliore stima del parametro di popolazione. Questo è vero sperimentalmente: si è visto che la distribuzione delle medie campionarie è più consistente (cioè con SD minore) rispetto a altre statistiche come media o moda, e tende a centrare μ.

Anche qui calcoliamo un intervallo di confidenza, utilizzando la definizione di SD della media campionaria vista prima, sostituendola nella formula di :

Quindi il nostro intervallo di confidenza sarà , o .

Anche qui, possiamo stimare la taglia del campione per ottenere un intervallo di confidenza accettabile. Possiamo di nuovo proseguire come prima, facendo l’inverso della formula di :

Di nuovo, non conoscendo esattamente σ, potremmo cercare di stimarla da studi vecchi oppure da studi correlati: Se sappiamo che σ dei quozienti intellettivi della popolazione “italiani” è 15, allora, se prendiamo in considerazione come popolazione un sottogruppo di “italiani” possiamo provare ad utilizzare 15 come σ della nuova popolazione.

### La deviazione standard non è nota

Quando σ non è nota, abbiamo comunque il requisito di campionamento casuale semplice e che le medie campionarie seguano una distribuzione normale (quindi che la popolazione sia normale o che ), anche se il requisito di normalità è sempre non troppo severo per la robustezza del metodo.

Se σ è incognita, possiamo usare una distribuzione simile alla normale, chiamata ***distribuzione t di Student*** (sviluppata da William Gosset (1876-1987). Perché utilizziamo questa distribuzione? Non conoscendo σ, la stimiamo con la statistica s, la deviazione standard del campione. Questo però aggiunge una variabile di inaffidabilità: oltra a dover stimare la media con , dobbiamo anche stimare la SD con . Questa addizionale variazione è compensata dal range più ampio che la distribuzione di Student porta ad avere, dati i valori critici diversi calcolati, che in questo caso prendono il nome di .

Se la distribuzione della popolazione è essenzialmente normale, allora la distribuzione di in modo che:

È essenzialmente una distribuzione **t di Student,** o semplicemente **distribuzione t** per tutti i campioni di taglia .

I valori di sono tabulati, come per la normale standard. Esistono varie versioni della distribuzione t, in base ai **gradi di libertà** del campione. Questa nuova variabile è di nuovo per tenere conto del maggiore errore che l’introduzione della stima di porta nell’intervallo: minore è , maggiore il valore di s è incerto, quindi più ampio è l’intervallo che bisogna calcolare. I gradi di libertà sono, appunto, il numero di possibili valori che possono variare dopo aver imposto delle restrizioni. La distribuzione t è identica ad una normale standard se i gradi di libertà sono infiniti, cioè il campione ha taglia infinita (o molto, molto grande).

Ad esempio, se 10 studenti hanno preso dei voti con media pari a 80, la sommatoria di tutti i voti deve essere 800. Se i primi 9 voti sono aleatori, cioè possono prendere (con dovuti limiti) qualsiasi valore, l’ultimo no: deve essere per forza pari a . Quindi, per quello che interessa a noi, il valore dei gradi di libertà è pari a .

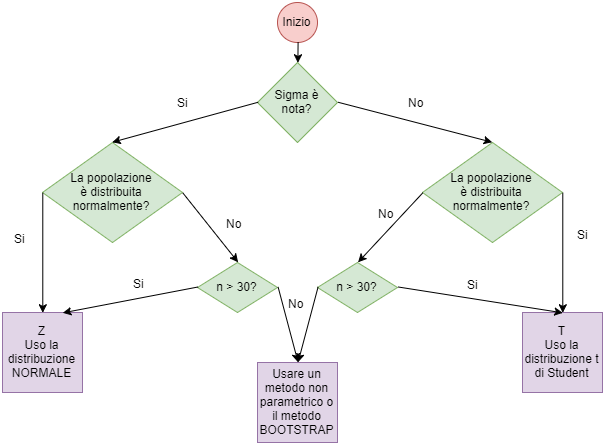
Possiamo ora definire l’intervallo di confidenza per la media campionaria , dato un , come:

con avente gradi di libertà.

Notiamo qui alcune caratteristiche della distribuzione t di Student:

* Essa cambia per taglie campionarie diverse (dato che cambia il valore di gradi di libertà);
* Essa ha comunque una forma a campana, ma con una maggiore variabilità (la curtosi cambia);
* Essa ha una media , come la normale standard;
* La deviazione standard cambia in base alla taglia campionaria, ma è sempre maggiore di 1 (con );
* All’aumentare della taglia campionaria, la distribuzione t di Student diventa sempre più simile ad una normale standard.
* L’area sottesa dalla funzione è sempre 1.

Quando usiamo la normale standard, e quando utilizziamo la t di Student? La tabella qui sotto può chiarirci le idee:

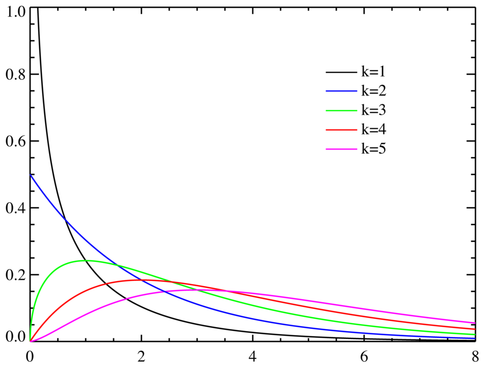


## Stima della varianza di una popolazione

Parallelamente a cosa abbiamo fatto per le altre statistiche, anche per la varianza sono necessari alcuni requisiti:

* I campioni sono casuali semplici;
* **La *popolazione* è distribuita normalmente, anche se è grande**. Questo requisito è tassativo, e la normalità deve essere appurata come descritto prima, con istogrammi e grafici dei quantili. Questo metodo non è robusto, e la mancata normalità della popolazione porta a grandi errori.

Anche qui dobbiamo utilizzare una nuova distribuzione, come per la media abbiamo usato la t di Student: la distribuzione *chi-quadro*, .

Data una popolazione con distribuzione normale, dopo aver estratto con il metodo casuale semplice una serie di campioni indipendenti di taglia , e calcoliamo la varianza campionaria per ogni campione, allora la statistica campionaria:

Dove è la varianza di popolazione, ha una distribuzione chiamata **distribuzione chi-quadro**.

Anche la distribuzione chi quadro è determinata di gradi di libertà del campione, determinati come . La distribuzione è mostrata a fianco, con . Descriviamo delle caratteristiche della funzione:

* Non è simmetrica. Essa diventa sempre più simmetrica all’aumentare di .
* I valori di una distribuzione chi quadro non possono essere negativi (tutto nella formula è necessariamente positivo)
* La distribuzione cambia per ogni valore di . Se , la funzione è identica ad una normale.
* L’area totale sottesa dalla funzione è sempre 1.

Dato che la distribuzione non è simmetrica, la notazione degli intervalli di confidenza non è corretta, dobbiamo infatti fare due calcoli di per il limite sinistro e destro, utilizzando i valori di e rispettivamente.

La varianza campionaria è uno stimatore non distorto della varianza di popolazione, e possiamo calcolarne l’intervallo di confidenza come:

Invece, la deviazione standard non è un buon stimatore, dato che è distorto, ma è comunque utilizzata. Possiamo calcolarne l’intervallo di confidenza come le radici degli estremi dell’intervallo per la varianza:

I calcoli per trovare la taglia campionaria non sono così semplici come gli altri due esempi. Di solito si utilizzano calcolatori come R o STATDISK. Comunque, a scopo di esempio, la taglia campionaria per avere un intervallo di confidenza del 95% è 3148, mentre per averlo del 99% è 77207. Da questi numeri capiamo quanto è inefficace il metodo per stimare la varianza e la deviazione standard.

# Capitolo 5 – Verifica di un’ipotesi

Spesso ci chiediamo quando una nostra ipotesi è confermata o meno da un esperimento. Ad esempio, gli esperimenti di Mendel portano ad avere un 26.7% di baccelli gialli, quando l’ipotesi mendeliana ne predice un 25%. Come possiamo affermare che l’ipotesi di Mendel è esatta? Come possiamo determinare se invece è falsa?

Definiamo alcuni termini:

* In statistica, un’**ipotesi** è una dichiarazione o affermazione riguardante una proprietà di una popolazione.
* Un **test di ipotesi** o **test di significatività** è una procedura standard per verificare un’ipotesi.

Questo capitolo fa uso pesantemente della **legge degli eventi rari** vista nel capitolo di probabilità: Se sotto una determinata ipotesi, la probabilità di un particolare evento osservato è eccezionalmente bassa, concludiamo che l’ipotesi è verosimilmente sbagliata.

Prendiamo in considerazione questo esempio: Una casa farmaceutica produce un prodotto che, a quanto affermato, aumenta la probabilità che venga concepita una femmina. Sappiamo che la probabilità normale che venga concepita una femmina è . In un esperimento, cento coppie che hanno utilizzato il prodotto hanno avuto un figlio ciascuna, e dei 100 nascituri 52 erano femmine. Notiamo che avere *x successi in n prove è un numero insolitamente alto di successi se è molto piccolo*, ad esempio più piccolo di 0.05. **Se è poco probabile che escano numeri più grandi di *,* allora è un numero molto grande in quella distribuzione.** Allo stesso modo, è un numero insolitamente piccolo se è piccola. Possiamo utilizzare il teorema del limite centrale per trattare i dati come una distribuzione normale, e calcolare la probabilità che siano nate 52 o più femmine su 100 nascituri: . Dato che la probabilità che nasca una femmina sono del 0.5, possiamo dire che il prodotto in realtà non funziona, dato che non abbiamo nessuna evidenza che ci fa pensare che la differenza tra le 50/100 femmine teoriche e le 52/100 femmine ottenute non sia dovuta al semplice caso. In altre parole, il 52% di femmine in un campione con non è insolito se la vera probabilità di avere una femmina è .

Per definire un test statistico, dobbiamo prima definire formalmente alcuni concetti che saranno utilizzati:

* L’ipotesi nulla, denotata con , afferma che il parametro di popolazione è uguale ad un certo valore.  
  Ad esempio,
* L’ipotesi alternativa, denotata con , afferma che il parametro della popolazione differisce da quanto affermato dall’ipotesi nulla. Per questo capitolo, l’ipotesi alternativa deve sempre contenere .
  + Se si vuole utilizzare un test statistico per sostenere le proprie affermazioni, bisogna esprimere tali affermazioni come ipotesi nulle, e questo richiede l’uso dei simboli . Non possiamo utilizzare un test di ipotesi per affermare che un parametro è uguale ad un valore.

**Statistica di test**

La statistica di test è un valore calcolato utilizzando i dati campionari, che è usato per decidere **se rifiutare** **l’ipotesi nulla**. Questo valore è determinato trasformando la statistica campionaria in un **punteggio immaginando che l’ipotesi nulla sia vera**:

Statistica di test per la **proporzione**:

Statistica di test per la **media**:

Statistica di test per la **deviazione standard**:

**Esempio esplicativo**:

Prendiamo in considerazione un test di genetica mendeliana: su 580 piante di pisello, il 26.2% () risulta avere baccelli gialli. Mendel dichiara che il valore teorico di baccelli gialli è 25%. Con una statistica di test cerchiamo di capire se i dati ottenuti si discostano molto dal valore teorico dichiarato da Mendel:

L’ipotesi nulla è , mentre l’ipotesi alternativa è . Dato che lavoriamo sotto l’ipotesi che sia vera, allora che , otteniamo la statistica di test (utilizzando la formula qui sopra per le proporzioni):

Dal capitolo precedente, possiamo dire che il valore di non è molto grande (sarebbe grande se fosse maggiore di 2 o minore di 2, per cui i dati campionari sono più distanti dalla media di 2 deviazioni standard): possiamo dire che in questo caso 26.2% non è *significativamente* diverso da 25%, quindi non possiamo dire che l’ipotesi è rifiutata.

* La **regione critica**, o **regione di rifiuto**, è l’insieme di tutti i valori della statistica di test che ci portano a rifiutare l’ipotesi nulla. Questi sono arbitrari, ma di solito sono le zone di “coda” della distribuzione normale.
* Il **livello di significatività** (α) è la probabilità che la statistica di test cada nella regione critica anche quando l’ipotesi nulla è vera; α, quindi è la probabilità di commettere un errore. Questa α è la stessa che abbiamo introdotto qualche capitolo fa al riguardo della creazione degli intervalli di confidenza, e quindi di solito prende il valore di 0.05, 0.01 o 0.10.
* Un **valore critico** è qualsiasi valore che separa la regione critica da quella che porta al non rifiuto dell’ipotesi nulla. Ad esempio, per α = 0.05 in un test a due code (vedi sotto), il valore critico è z = ± 1.96.

Notiamo che se la nostra ipotesi è del tipo “”, allora possiamo commettere due tipi di errori: sia sovrastimare la proporzione, e quindi cadere nella coda di destra della distribuzione, sia sottostimare la proporzione, e quindi cadere nella coda di sinistra. La nostra regione critica è quindi sia a sinistra che a destra. Questo tipo di test è detto **a due code**. I valori critici sono i valori di z che delimitano due aree, a sinistra e a destra, per cui ; quindi

Se la nostra ipotesi alternativa è del tipo “”, allora non sarà possibile sottostimare la proporzione, dato che stiamo solo considerando solo valori più grandi di . In questo caso, l’unico errore possibile è sovrastimare la proporzione, e l’area critica è solo a destra, e la sua area è pari a α, di solito 0.05. Questo test è chiamato **a coda destra,** e in questo caso il valore di , con è + 1.64.

Nello stesso modo, se l’ipotesi alternativa è del tipo “”, il test è **a coda sinistra**, e il valore di con è -1.64.

Possiamo ricordare quale tipo di test è associato ad un’ipotesi nulla semplicemente pensando a quali valori sono significativamente discostanti dall’ipotesi nulla, ed osservando il segno dell’ipotesi alternativa: .

Il **p-value** è la probabilità che il valore della statistica di test calcolato per la popolazione sia maggiore o uguale a quello calcolato con i dati campionari, se l’ipotesi nulla è vera. L’ipotesi nulla è rifiutata se il p-value è meno di α, ad esempio inferiore a 0.05. I P-value sono calcolati in questo modo:

* Per i test ad una coda, il p-value è il valore dell’area della distribuzione rispettivamente alla destra o alla sinistra della statistica di test per i test a coda destra e coda sinistra.
* Per i test a due code, il p-value è il **doppio** dell’area più piccola che si ottiene partendo dal valore di statistica di test e considerando l’area destra o sinistra.

### Prendere una decisione ed arrivare ad una conclusione

I test di ipotesi portano sempre ad avere una delle due conclusioni:

* L’ipotesi nulla è rifiutata
* L’ipotesi nulla non è rifiutata

La nostra ipotesi può essere sia l’ipotesi nulla che l’ipotesi alternativa, ma il risultato del test statistico è sempre lo stesso. Abbiamo quindi più modi per decidere se rifiutare o meno l’ipotesi nulla:

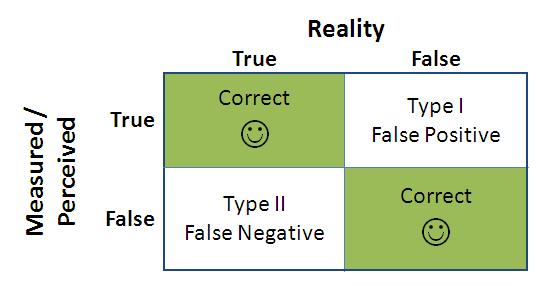
* **Metodo tradizionale**: rifiuto se la statistica di test cade nella regione critica. Non rifiuto se la statistica di test cade fuori dalla regione critica.
* **Metodo del p-value**: rifiuto se il valore del p-value è minore o uguale ad α. Non rifiuto se il p-value è maggiore di α.
  + **Opzione alternativa**: Non determino α, ma calcolo solo il p-value e lascio la scelta al lettore di determinare se è da rifiutare o meno.
* **Intervalli di confidenza**: Se il valore dell’ipotesi è fuori dall’intervallo di confidenza per quel parametro, rifiuto . Notare che se il test è ad una coda, la confidenza dell’intervallo deve essere .
  + L’uso dell’intervallo di confidenza aggiunge il possibile errore di stimare la SD della popolazione, pertanto non è molto utilizzato, e, se il metodo del p-value e quello tradizionale sono equivalenti tra loro, non lo sono con il metodo dell’intervallo di confidenza.

Se utilizziamo il metodo del p-value, è meglio determinare α prima dell’analisi, per non essere tentati di scegliere α in base al risultato.

Per esprimere I nostri risultati in forma più comprensibile per qualcuno che non ha seguito un corso di Statistica potremmo dire:

* Se l’ipotesi origine contiene un’uguaglianza (è quindi ):
  + è rifiutata: “*C’è un’evidenza statistica sufficiente per garantire il rifiuto dell’ipotesi per cui…*”
  + non è rifiutata: “*Non c’è un’evidenza statistica sufficiente per garantire il rifiuto dell’ipotesi per cui…*”
* Se l’ipotesi originale non contiene una condizione di uguaglianza (è quindi ):
  + è rifiutata: “*I dati campionari sostengono l’ipotesi per cui…*”
  + non è rifiutata: “*Non c’è un’evidenza campionaria sufficiente per sostenere l’ipotesi per cui…*”

### Errori durante lo svolgimento di un test d’ipotesi

Durante un test, anche se tutto è calcolato esattamente, rischiamo comunque di avere un errore. Questi errori sono di due nature: In realtà l’ipotesi è falsa, ma non l’abbiamo rifiutata (Errore di tipo 1, o **falso positivo**); in realtà l’ipotesi è vera, ma l’abbiamo rifiutata (Errore di tipo 2, o **falso negativo**).

Le probabilità di commettere questi errori sono definite come per l’errore di prima specie, e per l’errore di seconda specie.

Il migliore dei casi sarebbe quando sia alfa che beta fossero zero, ma questo non è possibile: dobbiamo trovare un buon compromesso tra i due tipi di errore. α è già selezionato prima, ad esempio 0.05. Si può dimostrare che α, β e (la taglia campionaria) sono in relazione matematica. Di solito si fissano due di queste variabili per ottenere la terza con un valore accettabile. Di solito si fissano α e , per ottenere un al di sotto di 0.2. Queste considerazioni pratiche sono vere:

* Per un valore di α fissato, un incremento di porta ad una diminuzione di β.
* Per un valore di fissato, α e β sono tra di loro inversamente proporzionali: all’aumento di uno l’altro diminuisce.
* L’incremento di porta ad una diminuzione sia di α che di β.

Dalla definizione di errore di seconda specie possiamo dire che è pari alla probabilità di rifiutare l’ipotesi nulla se essa è falsa. Ci si riferisce a questa probabilità come **potenza del test**. La probabilità β è influenzata, però, da α, , e dal valore che è considerato come alternativo nell’eseguire il test. Uno stesso test, quindi, può avere molteplici potenze. Spesso si cerca di determinare a priori considerando quale potenza vogliamo per il nostro test (minimo 0.80).

Riassumiamo qui i metodi per un test di ipotesi in ciascuno dei parametri visti:

#### Test di ipotesi per una **proporzione**

* **Requisiti**
  + Il campione è casuale semplice.
  + Le condizioni per una distribuzione binomiale sono soddisfatte:
    - Numero fisso di prove indipendenti, probabilità constanti per ogni prova, solo due esiti per prova, “successo” o “fallimento”
  + Le condizioni di approssimazione alla normalità sono tutte soddisfatte:
    - Se sono soddisfatte, allora e
* **Notazione**
  + è la taglia campionaria
  + è la proporzione campionaria
  + è la proporzione della popolazione (in )
* **Statistica di test per la verifica di un’ipotesi sulla proporzione:**
* Per il p-value e i valori critici, utilizziamo la **normale standard**.

#### Test di ipotesi per una **media**: σ nota

* **Requisiti**
  + Il campione è casuale semplice.
  + σ è noto.
  + Una o entrambe le condizioni di normalità sono soddisfatte:
    - La popolazione ha una distribuzione normale o .
* **Statistica di test per la verifica di un’ipotesi sulla media:**
* Per il p-value e i valori critici utilizziamo la **normale standard**.
* Per una media, il metodo dell’intervallo di confidenza è equivalente a quello del p-value e tradizionale.

#### Test di ipotesi per una **media**: σ non nota

* **Requisiti**
  + Il campione è casuale semplice.
  + σ non è noto.
  + Una o entrambe le condizioni di normalità sono soddisfatte:
    - La popolazione ha una distribuzione normale o .
* **Statistica di test per la verifica di un’ipotesi sulla media:**
* Per il p-value e i valori critici utilizziamo la **t di Student** con gradi di libertà. Il p-value è difficile da calcolare manualmente. Utilizzare un software o una tabella del p-value.
* Per una media, il metodo dell’intervallo di confidenza è equivalente a quello del p-value e tradizionale.

#### Test di ipotesi per la **deviazione standard** o la **varianza**

* **Requisiti**
  + Il campione è casuale semplice.
  + La popolazione è distribuita normalmente. (Vale la regola della rigidità della normalità vista prima, dato che i metodi per la deviazione standard non sono robusti).
* **Statistica di test per la verifica di un’ipotesi sulla SD:**
* Per il p-value e i valori critici utilizziamo il χ quadro. L’uso di software è raccomandato.

# Capitolo 6 – Inferenza per due campioni

Quando ci troviamo davanti a due campioni, e vogliamo confrontarli e decidere se le loro differenze sono statisticamente significative, siamo di fronte all’inferenza su due campioni. Questo tipo di statistica è comunemente effettuato ad esempio negli studi clinici, dove si vuole accertare sistematicamente che le differenze tra il gruppo placebo sono diverse da quello trattato.

## Inferenza su due proporzioni

Quando verifichiamo un’ipotesi con due campioni su una proporzione abbiamo i seguenti requisiti:

* I due campioni casuali semplici considerati non hanno alcun tipo di relazione, cioè non sono **appaiati** tra di loro.
* Per ciascuno dei due campioni, il numero di successi è almeno 5 e il numero di fallimenti è almeno 5.

Abbiamo queste notazioni per le proporzioni:

* è la proporzione della popolazione 1, da dove è estratto il campione 1
* è la taglia campionaria del campione 1
* è il numero di successi nel campione 1
* è la proporzione campionaria del campione 1

Notazioni del tutto analoghe sono per il campione 2. Notare che se un test non esplicita tutti questi valori, quasi sicuramente possiamo calcolarli con formule inverse dai dati forniti. Ricordiamo di arrotondare in modo sensibile (non possiamo avere 112.23 persone).

Vedremo se l’affermazione è rifiutabile o meno, ma ci occorrerà una stima aggregata del valore reale di :

Questo valore prende il nome di **stima cumulata**.

Possiamo ora vedere la statistica di test utilizzando come è :

Dato che consideriamo l’ipotesi nulla come vera, allora . Per il p-value e i valori critici utilizziamo un software o una tabella. Il procedimento per il test a due campioni è pressoché identico a quello ad un singolo valore.

Possiamo anche procedere con il metodo dell’intervallo di confidenza. In questo caso l’intervallo di confidenza è definito in questo modo:

## Inferenza su due medie: campioni indipendenti

Vediamo ora come verificare ipotesi per la differenza di due medie campionarie indipendenti. Definiamo intanto cosa si intende per campioni **dipendenti** ed **indipendenti**:

Due campioni sono **dipendenti** se gli elementi di un campione possono essere utilizzati per determinare in qualche modo l’altro. La dipendenza può anche essere temporale: stessi soggetti osservati prima del trattamento e dopo il trattamento. Due campioni sono **indipendenti** se non sono legati in alcun modo.

Per due campioni indipendenti abbiamo i seguenti requisiti:

* I due campioni sono **indipendenti**.
* Entrambi i campioni sono **casuali semplici**.
* Le condizioni di normalità sono rispettate per entrambi i campioni: la popolazione di partenza è normale o la taglia campionaria di entrambi i campioni è > 30. Questo test è comunque robusto.

Verifica di ipotesi per due medie di campioni indipendenti:

Per ricercare i gradi di libertà, manualmente possiamo usare il valore più grande tra e , ma un metodo più preciso è utilizzato dai calcolatori:

Il P-value e i valori critici sono entrambi calcolati con la funzione t di Student, e calcolati con tabelle o calcolatori. In questo caso, dato che gli intervalli di confidenza utilizzano la stessa distribuzione e lo stesso errore standard, l’utilizzo come test degli intervalli di confidenza è equivalente ai metodi tradizionali e del P-value.

L’intervallo di confidenza si calcola come segue:

E i gradi di libertà sono calcolati come descritto prima.

Possiamo utilizzare le espressioni viste prima dato che, se i requisiti sono soddisfatti, allora la distribuzione campionaria di è approssimabile ad una t di Student, con media pari a e deviazione standard uguale a , sfruttando la proprietà per cui la varianza della differenza di due variabili aleatorie indipendenti è uguale alla somma delle varianze delle due variabili. Esistono anche metodi per trovare la statistica di test se e sono note, o se, per qualche motivo, possiamo dire che . Questa uguaglianza è assai difficile da dimostrare, e quindi il metodo che la richiede è caldamente sconsigliato e non sarà mostrato qui.

**e sono note:** usiamo la normale standard al posto della t di Student.

Intervallo di confidenza:

## Inferenza su differenze per campioni appaiati

Ora vediamo come comportarci se i campioni sono in qualche modo correlato tra loro, cioè **dipendenti**. Possiamo utilizzare questi metodi per qualsiasi differenza, anche per la differenza di medie vista prima. I requisiti sono:

* Il campione consiste in coppie appaiate.
* I campioni sono casuali semplici.
* Le condizioni di normalità sono rispettate per entrambi i campioni: la popolazione di partenza è normale o la taglia campionaria di entrambi i campioni è > 30. Questo test non è particolarmente robusto: se i dati variano molto dalla distribuzione normale, è bene utilizzare un metodo non parametrico.

Annotiamo in questo modo i simboli:

* è la singola differenza dei valori di una coppia.
* è il valore medio delle differenze per la popolazione di tutte le coppie appaiate
* è il valore medio delle differenze per i campioni appaiati
* è la deviazione standard campionaria delle differenze .
* è il numero di coppie di dati

La statistica di test usa la t di Student:

Dove i gradi di libertà sono . Per il p-value e i valori critici usiamo la t di Student con una tabella o un software.

Per gli intervalli di confidenza delle coppie appaiate usiamo:

Per i valori critici di usiamo la tabella o un software.

## Confronto della variabilità in due campioni

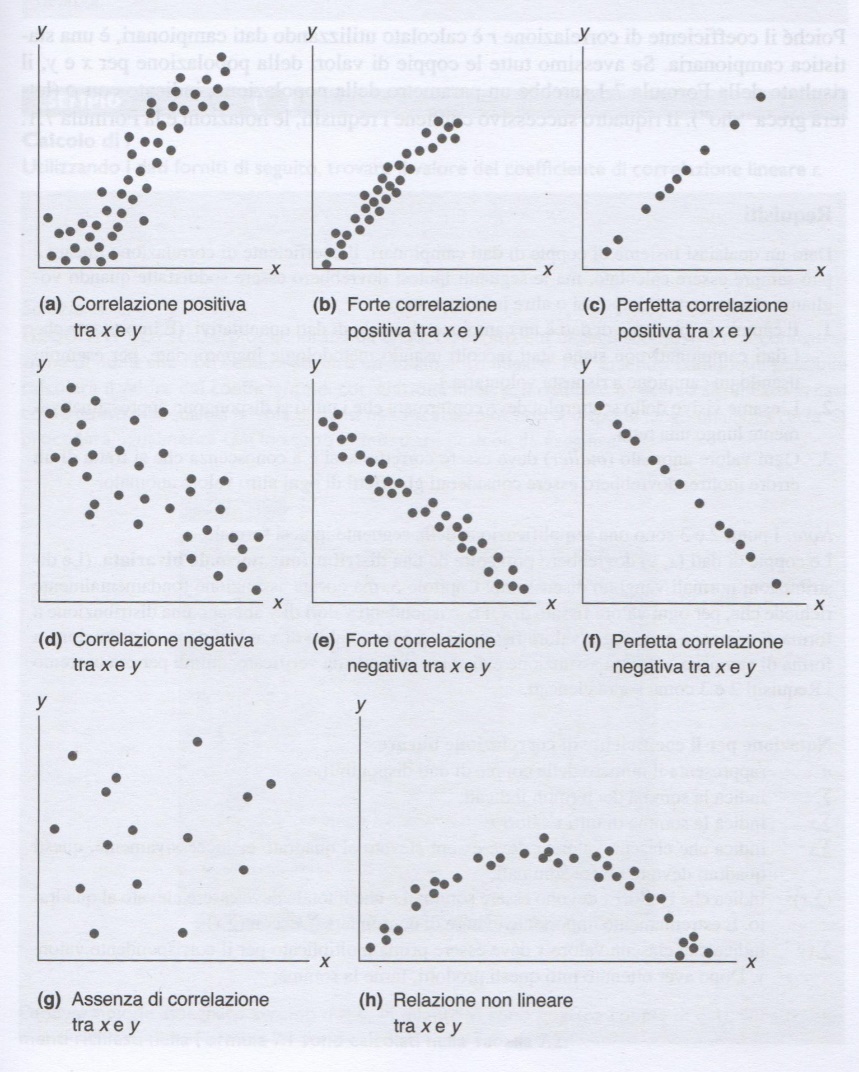
La procedura è formulata per le varianze, ma possiamo utilizzare questo metodo anche per le deviazioni standard (ricordiamo che le varianze sono il quadrato delle deviazioni standard). I simboli sono quelli già utilizzati.

* Requisiti
  + Le due popolazioni sono **indipendenti** tra loro.
  + Le due popolazioni sono **distribuite normalmente**. Questo metodo è estremamente fragile.
* Notazione particolare:
  + è la **maggiore** varianza campionaria
  + è la taglia del campione con varianza **maggiore**.
  + è la varianza della popolazione dalla quale il campione con varianza **maggiore** è stato estratto.
  + sono analoghe a riferite all’altro campione e all’altra popolazione.
* Statistica di test:
* Notiamo che si usa la distribuzione F, simile al , ma che dipende da due gradi di libertà, quello del numeratore () e del denominatore (). Con il grafico della distribuzione F, vediamo che come il anche essa è asimmetrica, non è mai negativa, e si avvicina alla normale mentre i due gradi di libertà si avvicinano all’infinito.
* L’utilizzo di un software elimina la difficile interpretazione della distribuzione F dandoci subito i valori critici o meglio il P-value.

# Capitolo 7 – Correlazione e regressione

In questo capitolo ci occupiamo di determinare se c’è una qualche associazione tra due variabili, e, se questa esiste, come descriverla attraverso un’equazione matematica. La correlazione è studiata attraverso un grafico, lo scatterplot, e attraverso un coefficiente detto **coefficiente di correlazione lineare**, una misura numerica del gradodi correlazione delle due variabili.

Lo scatterplot può darci l’idea di cosa intendiamo per correlazione, e anche una preliminare ipotesi sulla possibile correlazione che due variabili hanno prima di passare alle formule matematiche:

Se da un’analisi preliminare con uno scatterplot notiamo una qualche correlazione, possiamo calcolare il coefficiente di correlazione. Il **coefficiente di correlazione lineare di Pearson** è utilizzato per quantizzare quanto sono correlati, in modo lineare, due set di dati. Se viene calcolato con dati campionari, il coefficiente è denominato ; se è calcolato con tutti i dati della popolazione è denominato con la lettera greca (Rho).

Il coefficiente di correlazione lineare può essere sempre calcolato, ma può non aver significato se il campione è raccolto male o i dati non hanno significato; se lo scatterplot non fa notare una correlazione lineare o se ci sono outlier.

Più formalmente, potremmo dire che le coppie di dati devono provenire da una **distribuzione normale bivariata**, cioè ad ogni valore fissato di , i valori di seguono una distribuzione normale, e viceversa.

Elenchiamo alcuni simboli utilizzati nelle prossime formule:

* è il numero delle coppie di dati possibili,
* indica la sommatoria di tutti i valori di ,
* indica la sommatoria di tutti i valori di **dopo** averli elevati al quadrato,
* indica la sommatoria di tutti i valori di , **poi** li si eleva al quadrato. È estremamente importante non confondere le due scritture.
* indica la sommatoria di tutti i prodotti .
* è il coefficiente di relazione lineare riferito ad un campione,
* è il coefficiente di relazione lineare riferito alla popolazione.

Possiamo calcolare il coefficiente con la seguente formula:

Il valore di può essere interpretato con una tabella, oppure attraverso il p-value. Il p-value è ottenuto spesso da software, ed è interpretato in un modo identico a come abbiamo visto prima ( è rifiutato se p-value > , dove α è fissato *a priori*).

Il coefficiente di Pearson ha delle caratteristiche particolari:

* Il valore di è sempre compreso tra gli estremi, inclusi, di +1 e -1.
* Il valore di non cambia se tutti i valori di entrambe le variabili sono convertiti in una scala differente.
* Il valore di non dipende dall’ordine in cui vengono scelte ed : scambiando i valori tra le due variabili il valore di non cambia.
* misura l’intensità di un’associazione lineare, ma non è in grado di misurare l’intensità di associazioni non lineari.

Possiamo utilizzare un test di ipotesi per vedere se c’è o meno una correlazione lineare tra due popolazioni, cioè se o . Procediamo come abbiamo fatto per tutti gli altri test di ipotesi:

* **Metodo 1: la statistica di test è** 
  + Statistica di Test:
  + Valori critici e P-value sono calcolati con un numero di gradi di libertà pari ad .
  + Per trarre le nostre conclusioni possiamo dire che se , allora rifiutiamo e possiamo dire che c’è evidenza statistica per una correlazione lineare. Se invece , allora non possiamo rifiutare , e non c’è evidenza di correlazione lineare.
* **Metodo 2: utilizzo di come statistica di test**
  + Possiamo utilizzare direttamente i valori di per vedere se c’è correlazione o meno. Utilizziamo una tabella particolare dei valori critici di , e traiamo le nostre considerazioni in modo uguale a prima.

Per testare se c’è una correlazione lineare **positiva** o **negativa**, il test diventa ad una coda sola: dobbiamo modificare α in maniera opportuna, e modificare le nostre ipotesi per avere e per una correlazione negativa e per una correlazione positiva.

## La regressione

Ora vogliamo prendere due variabili in correlazione lineare e descrivere matematicamente una retta, chiamata **retta di regressione**, che descrive la relazione delle due variabili. L’equazione che definisce questa retta è chiamata **equazione di regressione**.

Assoceremo la variabile , chiamata **variabile indipendente**, con l’altra variabile , chiamata **variabile dipendente**. L’equazione tipica di una retta è . Noi utilizzeremo la stessa forma, scritta come . sono notazioni riferite al campione, mentre sono riferite alla popolazione. Possiamo stimare i valori della popolazione partendo dai valori del campione, come vedremo dopo.

Anche qui possiamo sempre creare la retta di regressione lineare, ma è bene considerare che questa può non avere senso se i seguenti requisiti non sono soddisfatti:

* Il campione è formato da dati quantitativi, e questi sono raccolti in modo corretto,
* L’esame visivo dello scatterplot mostra della linearità,
* Non ci sono outlier, o gli outlier sono considerati in maniera opportuna.

Il metodo è comunque robusto.

La **pendenza** è calcolata come:

E l’**intercetta** come:

La retta descritta dall’equazione è quella che meglio interpola i dati delle due variabili.

Possiamo utilizzare la retta di regressione per predire i valori di una variabile quando non sono noti i valori dell’altra variabile. Prendiamo ad esempio l’uso della legge di Lambert-Beer per predire le concentrazioni. Se non c’è relazione tra le due variabili, la migliore stima per predire un qualsiasi valore di è la sua media campionaria, . Se c’è correlazione con un'altra variabile , la migliore stima diventa il valore di che si trova sostituendo nell’equazione di regressione il valore di associato alla incognita.

Ricordiamo che la retta è utilizzabile solo se c’è una vera correlazione tra le due variabili, cioè se è significativo. Occorre quindi sempre fare un test di ipotesi per controllare se è significativo.

## Valori anomali e punti influenti. Minimi scarti quadrati

In uno scatterplot, possiamo definire i cosiddetti **punti influenti** come i punti che influiscono pesantemente sulla retta di regressione calcolata come prima. Possiamo inoltre avere **valori anomali**, cioè valori che si discostano tanto da tutti gli altri (e quindi dalla retta di regressione).

I punti influenti si trovano togliendo il punto pensato influente e ricalcolando la retta. I valori anomali sono facilmente visibili in uno scatterplot, e non sono molto diversi dagli **outlier**.

La retta che meglio descrive un set di dati è quella che porta ad avere meno **residui** tra di essa e i punti. I residui sono calcolati da ciascun punto con la differenza , dove è il valore predetto di dalla retta lineare. I **minimi quadrati** sono i quadrati più piccoli che si possono costruire con lato uguale a ciascuno dei residui. La retta che descrive meglio il set di dati è la retta che produce i minimi quadrati:

Il **grafico dei residui** è il grafico che otteniamo sostituendo le dello scatterplot con i residui corrispondenti. L’equazione della retta descrive bene i punti se il grafico degli scarti non presenta alcuna distribuzione se non quella completamente casuale.

Chiamiamo **deviazione totale** dalla media il valore .  
Chiamiamo **deviazione spiegata** il valore .  
Chiamiamo **deviazione non spiegata** il valore ().

Notiamo che il valore di correlazione lineare è anche definito come:

# Capitolo 8 – Analisi della varianza. Metodo ANOVA

Vorremmo ora introdurre un test statistico per verificare l’ipotesi che medie di tre o più popolazioni sono uguali. In questo caso, e .

Non possiamo effettuare un test per ogni coppia di variabili dato che la confidenza, in questo caso, scende esponenzialmente al pari passo dell’aumentare dei numeri di test che dobbiamo eseguire.

Il metodo **ANOVA** (**AN**alysis **O**f **VA**riance) è usato per testare questa ipotesi. Esso utilizza la distribuzione F, già stata introdotta. L’analisi si basa sul confronto delle varianze stimate per ciascuna popolazione, e queste stime saranno definite più avanti.

Parliamo di ANOVA ad *una via* se le nostre popolazioni sono diversificate da un solo fattore (o “via”). Ad esempio, in un esperimento potremmo dividere dei campi coltivati in “non trattati”, “pesticidi” e “coccinelle”, per il fattore “trattamento”: in questo caso si utilizza il metodo ANOVA ad una via.   
Parliamo di ANOVA a *due vie* se le nostre popolazioni sono diversificate per due fattori diversi, ad esempio il fattore “trattamento” come prima, e il fattore “irrigazione” che può essere “irrigato” o “non irrigato”.

Un fattore o via, quindi, è una proprietà, caratteristica o comunque una tipologia che ci permette di distinguere una popolazione dall’altra.

## ANOVA ad una via

I calcoli per il metodo ANOVA sono molto complessi, ma ricordiamo:

* e .
* Un p-value molto piccolo (minore di α), porta al rifiuto di . Il contrario porta al non rifiutare .

I requisiti per il metodo ANOVA sono abbastanza lassi, dato che il metodo è molto robusto, ma comunque sono elencabili:

* La popolazione è distribuita normalmente.
* Le popolazioni hanno la stessa varianza .
* I campioni sono campioni casuali semplici e i dati sono quantitativi.
* I campioni sono indipendenti gli uni dagli altri.

La procedura è necessariamente seguita utilizzando un software che produce un p-value, interpretato nella via consueta. Notare che il test ANOVA non ci da l’uguaglianza di tutte le medie ad un valore, ma solo che le medie sono uguali. Per ottenere quel numero, sono necessari altri test. Perché il test ANOVA funziona? Molto probabilmente è magico.