STATISTICA MEDICA

RACCOLTA DATI: raccogliere informazioni su come è la realtà in questo momento (es. questionario). Si raccolgono dati quando abbiamo un quesito a cui vogliamo dare una risposta. È un processo di pianificazione, raccolta, elaborazione, sintesi ed interpretazione dei dati raccolti.

DATI: numeri o caratteri alfanumerici (sequenze di numeri e lettere).

- Numeri con significato numerico (quantitativo).
- Numeri con significato categorico (non quantitativo).
- Caratteri alfanumerici (alfabetici e numerici).

Raccolta dati mediante questionario:

- Modalità di compilazione:
  - Questionario autosomministrato.
  - Compilazione assistita.
- Caratteristiche tecniche delle domande:
  - Domande chiuse: celta fra opzioni multiple prespecificate.
  - Domande aperte: inserimento di testo libero.

ARCHIVIAZIONE DATI

Per facilitarne la gestione, i dati, mediante l'utilizzo di un software, sono inseriti ed organizzati in un **database** che è un insieme strutturato di dati:

- Matrice dei dati: struttura rigida di righe e colonne.
- Riga: unità di osservazione (record).
- Colonna: caratteristica rilevata (variabile).

VARIABILI E SCALE DI MISURA

Variabili: caratteristiche rilevabili di un dato sistema oggetto di studio (es. peso, età, sesso).

- Variabile **quantitativa**: valore numerico
  - Continua: può assumere tutti i valori numerici. È sufficiente riportare tre cifre significative (40.593 → 40.6 ; 9.2368 → 9.24 ; 0.14521 → 0.145 ; 0.07391 → 0.0739).
  - Discreta: può assumere solo i numeri interi.
- Variabile **qualitativa**: modalità.

I valori numerici o le modalità assunti dalle variabili costituiscono i **dati**.

Scale:

- Scala **nominale** (a=b oppure a≠b): i valori possono essere riconosciuti esclusivamente come uguali o diversi.
  - Variabile **dicotomica**: due sole modalità possibili (sesso, lancio moneta).

  - Variabile **politomica**: più di due modalità possibili (gruppo sanguigno, colore, lancio dado).
- Scala **ordinale** (a > b oppure a < b): i valori possono essere disposti in un ordine naturale di grandezza, siamo in grado di associare una quantità a delle caratteristiche (es. scala dolore, stadio malattia, livello istruzione, taglia abiti, grado di soddisfazione).
  - VAS (Visual Analogue Scale): scala quantitativa che cerca di quantificare il dolore. possiamo considerarla una scala razionale.
- Scala **intervallare** (a-b): i valori possono essere disposti in ordine di grandezza, e la quantificazione della differenza fra due distinti valori ha un significato. Sono scale costruite con l'introduzione di uno *zero arbitrario* che rappresenta una convenzione e non l'assoluta assenza del fenomeno (es. temperature, date).
- Scala **razionale** (a/b) o **di rapporto**: i valori possono essere disposti in ordine di grandezza, e la quantificazione del rapporto fra due distinti valori ha un significato. È una scala caratterizzata da uno *zero assoluto*, che indica assenza del fenomeno, e da valori esclusivamente positivi (es. frequenza cardiaca, pressione arteriosa sistolica).

Tipologia	Caratteristica	Esempi	Peculiarità
NOMINALE	a=b a≠b	Sesso, status malattia, luogo di nascita, colori.	
ORDINALE	a>b a<b	Stadio malattia, titolo di studio, grado di soddisfazione.	Esiste un ordine naturale delle modalità.
INTERVALLARE	a-b	Temperature, date, score.	Zero arbitrario; ha un significato quantificare la differenza tra due distinti valori.
RAZIONALE	a/b	Statura, peso, età PAS, frequenza cardiaca.	Zero assoluto, valori solo positivi; hanno un significato sia la differenza sia il rapporto fra due valori.

ANALISI DATI

Controllo qualità dei dati: accertarsi che ogni variabile riportata nella matrice dei dati sia stata codificata correttamente.

Individuazione di dati non ammissibili o anomali: problemi di codifica, problemi di inserimento. La verifica dei dati avviene considerando le decisioni che abbiamo preso relativamente ai possibili valori che una variabile può assumere.

➔ Bisogna fare una **verifica all'origine**.

RACCOLTA E TRATTAMENTO DEI DATI

Quando l'unità statistica è una persona:

- Consenso informato: gli interessati devono essere informati e acconsentire.
- Comitato etico: l'indagine deve essere eticamente corretta.

— Trattamento dati nel rispetto della normativa in materia di privacy.

Codice in materia di protezione dei dati personali:

- art. 4 definizioni:
  — **Dato personale**: qualunque informazione relativa a persona fisica, persona giuridica, ente o associazione, identificati o identificabili, anche indirettamente mediante riferimento a qualsiasi altra informazione, ivi compreso un numero di identificazione personale.
  — **Dato sensibile**: i dati personali idonei a rilevare l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni di carattere religioso, filosofico, politico o sindacale, nonché i dati personali idonei a rilevare lo stato di salute e la vita sessuale.
- art. 11:
  1. I dati personali oggetto di trattamento sono:
     — Trattati in modo lecito e secondo correttezza.
     — Raccolti e registrati per scopi determinati, espliciti e legittimi, ed utilizzati in altre operazioni del trattamento in termini compatibili con tali scopi.
     — Esatti e, se necessario aggiornati.
     — Pertinenti, completi e non eccedenti rispetto alle finalità per le quali sono raccolti o successivamente trattati.
     — Conservati in una forma che consenta l'identificazione dell'interessato per un periodo di tempo non superiore a quello necessario agli scopi per i quali essi sono stati raccolti o successivamente trattati.
  2. I dati personali trattati in violazione della disciplina rilevante in materia di trattamento dei dati personali non possono essere utilizzati.
- art. 22 comma 6: I dati sensibili e giudiziari contenuti in elenchi, registri o banche di dati, tenuti con l'ausilio di strumenti elettronici, sono trattati con tecniche di cifratura o mediante l'utilizzazione di codici identificativi o di altre soluzioni che, considerato il numero e la natura dei dati trattati, li rendono temporaneamente inintelligibili anche a chi è autorizzato ad accedervi e permettono di identificare gli interessati solo in caso di necessità.
- I dati personali oggetto di trattamento sono custoditi e controllati, anche in relazione alle conoscenze acquisite in base al progresso tecnico, alla natura dei dati e alle specifiche caratteristiche del trattamento, in modo da ridurre al minimo, mediante l'adozione di idonee e preventive misure di sicurezza, i rischi di distruzione o perdita, anche accidentale, dei dati stessi, di accesso non autorizzato o di trattamento non consentito o non conforme alle finalità della raccolta.

Database clinici contengono dati sensibili.

Anonimizzazione: separazione fisica fra informazioni anagrafiche e dati sensibili con chiave di identificazione.

GLI STUDI SCIENTIFICI

- Ruolo del ricercatore:
  - **Osservazionali**: osservazione dei pazienti considerati senza eseguire nessun intervento.
  - **Sperimentali**: intervento sui pazienti considerati (es. terapeutico).
- Ruolo del tempo:
  - **Prospettici**: seguono il corso del tempo.
  - **Retrospettivi**: in senso contrario nel corso del tempo.
  - **Trasversali** (o cross-sectional): condotti in un solo istante.

Descrizione di casi (case reports e case series): casi clinici insoliti con qualche caratteristica clinica relativa a eziologia, presentazione o esiti, che si ritiene possa essere di qualche interesse. Suggeriscono ipotesi che possono essere successivamente indagate con studi metodologicamente più rigorosi.

STUDI **OSSERVAZIONALI**

1. Arruolamento: individuazione degli "esposti" e dei "non esposti".
2. Rilevazione dell'evento: seguo tutti i soggetti per un certo periodo di tempo e vedo quanti si ammalano nei due gruppi (follow up).
3. Studio **di coorte**: studio prospettico osservazionale: selezione di un gruppo di esposti e di un gruppo di non esposti. Si valuta *prospetticamente* la frequenza di eventi nel gruppo di esposti e nel gruppo di non esposti e si quantifica la forza dell'associazione.
   Studio **caso-controllo**: studio retrospettivo osservazionale: selezione di un gruppo di casi (malati) e di un gruppo di controlli (non malati). Si valuta *retrospettivamente* la frequenza di esposizione ad un dato fattore di rischio nel gruppo dei casi e in quello dei controlli e si quantifica la forza dell'associazione.
4. Risultati.

STUDI **SPERIMENTALI**

Il ricercatore non si limita ad osservare i soggetti inclusi nello studio ma interviene attivamente "obbligando" i partecipanti allo studio ad assumere particolari comportamenti, restando ovviamente nei confini di ciò che è eticamente accettabile.

Studio **trasversale** (cross-sectional): si valuta *nello stesso istante* la classificazione del paziente (malato o non malato) con i due metodi.

Studi **clinici randomizzati** (RCT): si valuta *prospetticamente* la frequenza di eventi nel gruppo di trattati e nel gruppo di non trattati e si quantifica la forza dell'associazione.

1. **Arruolamento** – selezione dei pazienti.
2. **Randomizzazione** – trattamenti: assegnazione casuale di ogni paziente ad uno dei due trattamenti. Garantisce una composizione media dei due gruppi simile per le caratteristiche individuali ritenute più importanti: unica differenza il trattamento.
3. Rilevazione dell'**esito** (guarigione – decesso).

4. **Risultati**.

Unione di più studi clinici: revisioni sistematiche /meta-analisi: rassegna sistematica di tutti i lavori presenti in letteratura aventi per oggetto un dato intervento, ad esempio la valutazione dell'efficacia di una data terapia per una certa patologia. Le revisioni non sono condotte con dati originali ma utilizzando dati già esistenti in letteratura e provenienti da altri studi. La meta-analisi è il metodo statistico che viene utilizzato per raggruppare i dati di diversi studi e fornire un risultato sintetico (obiettivo: cumulare i risultati di studi simili e cercare di dare una risposta definitiva sull'efficacia di un trattamento per una data patologia).

RAPPRESENTAZIONE DEI DATI

TABELLE

- **Proporzione**: uguaglianza di due rapporti, frazione del totale che possiede una certa caratteristica. Quattro numeri formano una proporzione se il rapporto fra il primo e il secondo è uguale al rapporto fra il terzo e il quarto.

  $$a : b = c : d$$

  — *a* e *c* → antecedenti
  — *b* e *d* → conseguenti
  — *a* e *d* → estremi
  — *b* e *c* → medi → se una proporzione ha i medi uguali si dice *continua* e il termine medio si dice *medio proporzionale* e l'ultimo termine si dice *terzo proporzionale*.
  — Proprietà delle proporzioni: il prodotto degli estremi è uguale al prodotto dei medi.
  — Si può esprimere in tre formati equivalenti: frazione (7/20), numero decimale (0.35) e percentuale (35%).
- **Incidenza** (tasso): numero di nuovi casi di malattia osservati in una popolazione in un dato intervallo di tempo.
  Incidenza cumulativa (proporzione): numero di nuovi casi di malattia in un dato periodo, rapportati al numero di individui sani, a rischio di ammalarsi, presenti all'inizio del periodo. Esprime il rischio di contrarre la malattia nel periodo considerato.
- **Prevalenza** (proporzione): numero di casi di malattia presenti in una popolazione in un dato istante.

VARIABILI QUALITATIVE

— Frequenza **assoluta**: numero di volte in cui compare, nella matrice dei dati, una specifica modalità (o valore, o classe di valori) di una variabile.
— Frequenza **relativa** (f_r = numero eventi di interesse / numero totale osservazioni): è data dal rapporto fra frequenza assoluta e numerosità totale. Può essere espressa sotto forma di proporzione o percentuale.
— Frequenza **cumulata**: solo per variabili su scala almeno ordinale (ordinale, intervallare, di rapporto). Le frequenze cumulate si ottengono riportando per ciascuna modalità (x) il numero di casi che hanno al massimo quella modalità ($f X \leq x$). Possono essere calcolate sia per le frequenze assolute che per le frequenze relative.

VARIABILI QUANTITATIVE DISCRETE: possono essere rappresentate in una tabella.

VARIABILI QUANTITATIVE CONTINUE

All'aumentare del numero dei valori riportati (numero delle righe) la tabella diventa difficilmente leggibile, per cui conviene raggruppare i dati in classi, riducendo così il numero di valori della variabile (righe) riportati in tabella.

Tabelle a doppia entrata: si considerano due variabili congiuntamente e si costruisce una tabella incrociando le due variabili. Se variabili dicotomiche: tabella 2 x 2.

ESEMPIO 1 – UTILIZZO DI INTERNET DISTINTO PER GENERE

	USANO INTERNET	**NON USANO INTERNET**	**TOTALE**
MASCHI	187	84	**271**
FEMMINE	174	115	**289**
TOTALE	**361**	**199**	**560**

Frequenze assolute

Frequenze marginali di colonna

Frequenze congiunte (di cella)

Frequenze marginali di riga

	USANO INTERNET	**NON USANO INTERNET**	**TOTALE**
MASCHI	33.4%	15.0%	**48.4%**
FEMMINE	31.1%	20.5%	**51.6%**
TOTALE	**64.5%**	**30.5%**	**100.0%**

Frequenze relative (percentuali) sul totale

Frazione di soggetti, sul totale complessivo, che possiede una caratteristica (congiunta).

	USANO INTERNET	**NON USANO INTERNET**	**TOTALE**
MASCHI	69.0%	31.0%	**100%**
FEMMINE	60.2%	39.8%	**100%**
TOTALE	**64.5%**	**35.5%**	**100.0%**

Frequenze relative (percentuali) di riga

Interpretabili come probabilità (condizionate)

Frazione di soggetti sul totale di riga, che possiede una data caratteristica.

	USANO INTERNET	NON USANO INTERNET	TOTALE
MASCHI	51.8%	42.2%	48.4%
FEMMINE	48.2%	57.8%	51.6%
TOTALE	100.0%	100.0%	100.0%

Frequenze relative (percentuali) di colonna.

Interpretabili come probabilità (condizionate)

Frazione di soggetti, sul totale di colonna, che possiede una data caratteristica.

ESEMPIO 2 – UTILIZZO DI INTERNET DISTINTO PER GENERE – SOLO UTILIZZATORI

	TUTTI I GIORNI	UNA O PIÙ VOLTE ALLA SETTIMANA	QUALCHE VOLTA AL MESE	QUALCHE VOLTA ALL'ANNO	TOTALE
MASCHI	142	40	3	2	**187**
FEMMINE	122	45	6	1	**174**
TOTALE	**264**	**85**	**9**	**3**	**361**

Tabella con più colonne, stessa interpretazione delle precedenti.

	TUTTI I GIORNI	UNA O PIÙ VOLTE ALLA SETTIMANA	QUALCHE VOLTA AL MESE	QUALCHE VOLTA ALL'ANNO	TOTALE
MASCHI	75.9%	21.4%	1.6%	1.1%	**100.0%**
FEMMINE	70.1%	25.9%	3.4%	0.6%	**100.0%**
TOTALE	**73.1%**	**23.5%**	**2.5%**	**0.8%**	**100.0%**

Percentuali di riga

	TUTTI I GIORNI	UNA O PIÙ VOLTE ALLA SETTIMANA	QUALCHE VOLTA AL MESE	QUALCHE VOLTA ALL'ANNO	TOTALE
MASCHI	53.8%	47.1%	33.3%	66.7%	**51.8%**
FEMMINE	46.2%	52.9%	66.7%	33.3%	**48.2%**
TOTALE	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Percentuali di colonna

GRAFICI

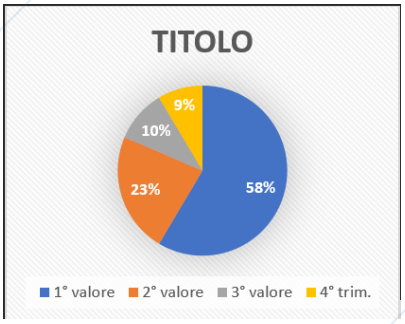
La rappresentazione grafica dei dati consente al lettore di cogliere con immediatezza le principali caratteristiche delle variabili che si stanno analizzando.

Componenti essenziali di un grafico (tutto il resto è superfluo):

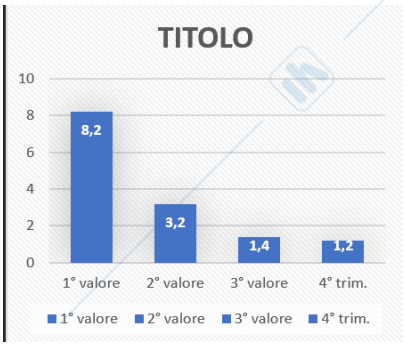
- Il titolo: legenda ed eventuali note esplicative.
- I dati: barre, aree, punti, linee.
- Gli assi: scala, con unità di misura ed etichette.

Tipi di grafici (variabili qualitative):

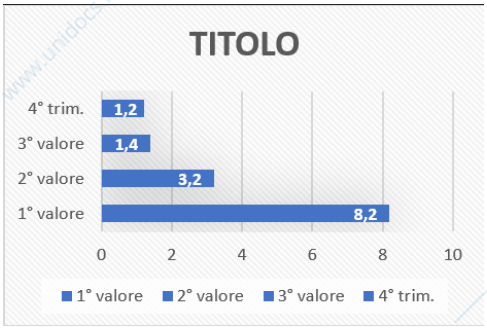
- **Areogramma** (diagramma a torta / ciambella): valutazione aree (aree proporzionali alla frequenza di unità statistiche aventi ognuna delle modalità indicate).



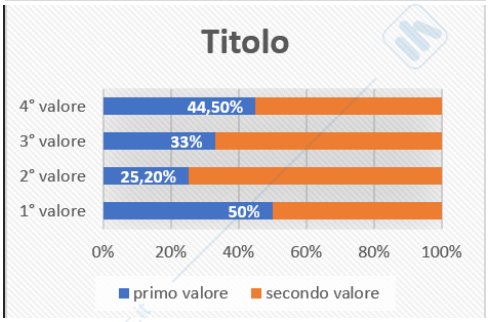
- **Diagramma a barre** (verticali): valutazione altezza barre (altezza delle barre proporzionale alla frequenza di unità statistiche aventi la caratteristica rappresentata dalla barra). Variabile (categorie) sull'asse orizzontale; conteggi (frequenze assolute o relative) sull'asse verticale.



- Diagramma a barre (orizzontali): valutazione lunghezza barre (lunghezza delle barre proporzionale alla frequenza di unità statistiche aventi la caratteristica rappresentata dalla barra. Variabile (categorie) sull'asse verticale; conteggi (frequenze assolute o relative) sull'asse orizzontale.

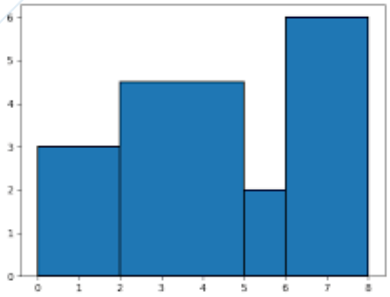


- Diagramma a **barre in pila**.

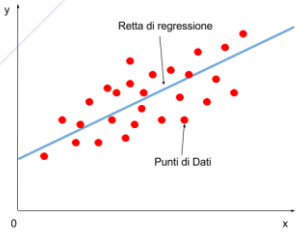


Tipi di grafici (variabili quantitative):

- Diagramma a torta.
- Diagramma a barre (verticali): valutazione altezza barre (altezza delle barre proporzionale alla frequenza di unità statistiche aventi quel particolare valore della variabile). Variabile sull'asse orizzontale, conteggi sull'asse verticale. Se la variabile è continua, le barre devono essere adiacenti; se la variabile è discreta, le barre devono essere separate.
    - Variabile *quantitativa continua* raggruppata in classi: se e solo se classi di uguale ampiezza, altezza delle barre proporzionale al numero di unità statistiche aventi quel particolare valore della variabile. Le barre successive devono essere una adiacente all'altra (variabile continua).
    - **Istogramma**: se si scelgono classi di ampiezza diversa, allora si deve ricorrere all'istogramma, in cui l'area della barra è proporzionale alla frequenza.



- Diagramma a barre in pila (verticale): valutazione composizione barre (altezza delle porzioni di barra proporzionale alla frequenza di unità statistiche aventi la caratteristica rappresentata dalla barra. Rappresentazione di differenti distribuzioni di frequenza. Variabile di confronto di interesse sull'asse orizzontale; conteggi (frequenze relative) suddivisi per le differenti modalità sull'asse verticale.
- **Grafico a dispersione**: andamento congiunto di due variabili quantitative continue. Ogni punto nel grafico rappresenta un soggetto; le coordinate del punto rappresentano i valori delle due variabili (x e y) in quel soggetto. Permettono di cogliere visivamente eventuali relazioni fra le due variabili.



LINEE GUIDA PER LA COSTRUZIONE DI GRAFICI:

- Scegliere il grafico più appropriato per ogni tipologia di variabile (torta, barra, dispersione).
- Riportare sempre un titolo, legenda ed eventuali note, che permettano di capire immediatamente il grafico.
- Scegliere le scale più opportune per gli assi, in modo da evitare fraintendimenti o rappresentazioni distorte.
- Evitare di riportare un numero eccessivo di categorie, soprattutto per i diagrammi a torta. Raggruppare se necessario.
- Evitare effetti grafici che potrebbero rendere meno immediata la lettura del grafico.
- Tenere presente la tipologia della variabile (qualitativa, quantitativa discreta o continua) e la scala di misura.

INDICI DI POSIZIONE E DI DISPERSIONE

- **Moda**: modalità della variabile in cui si registra il maggior numero di casi. È il valore che si verifica più spesso in una distribuzione di frequenza. Se il grafico è perfettamente simmetrico, la moda è anche la mediana.
- **Quantili**: indicatori di posizione di una distribuzione di frequenza che suddividono una serie ordinata di dati. (mediana, quartili, decili, centili).
  - **Mediana** (Me): valore che suddivide una serie ordinata di dati in 2 parti di uguale numerosità.
    - 50% delle osservazioni > Me; 50% delle osservazione < Me.
    - Se n dispari: (n+1)/2
    - Se n pari: fra (n/2) ed (n/2) + 1. Per il calcolo della mediana si considera il valore medio fra i valori corrispondenti alle due posizioni individuate.
  - **Quartili** (Q1, Q2, Q3): 3 valori che suddividono una serie ordinata di dati in 4 parti di uguale numerosità (Q2 = Me).
    - 1° quartile (Q1 = (n+1)/4): valore che divide una serie ordinata di dati in due sottoinsiemi. 25% delle osservazioni < Q1; 75% delle osservazioni > Q1.
    - 2° quartile (Q2, mediana = 2*(n+1)/4): valore che divide una serie ordinata di dati in due sottoinsiemi. 50% delle osservazioni < Q2; 50% delle osservazioni > Q2.
    - 3° quartile (Q3 = 3*(n+1)/4): valore che divide una serie ordinata di dati in due sottoinsiemi. 75% delle osservazioni < Q3; 25% delle osservazioni > Q3.
  - **Decili** (D1, D2, …, D9): valori che suddividono una serie ordinata di dati in 10 parti di uguale numerosità (D5 = Q2 = Me).
  - **Centili** (P1, P2, …, P99): 99 valori che suddividono una serie ordinata di dati in 100 parti di uguale numerosità. In ambito biomedico si utilizzano per le curve di crescita (P50 = D5 = Q2 = Me).
    - Utilizzati nelle curve di crescita (staturali e ponderali).

ALTRI INDICI

Media aritmetica = $\bar{x} = \frac{x_1 + x_1 + \ldots + x_n}{n} = \frac{\sum_1^n x_i}{n}$

Media aritmetica ponderata = $\bar{x} = \frac{\sum_1^n x_i * w_i}{\sum_1^n w_i}$ (w_i sono i pesi).

La media serve per sintetizzare i dati, per descrivere in modo sintetico le caratteristiche del mio campione. Però, lo stesso valore calcolato può rappresentare situazioni molto diverse tra loro. A volte utilizzare la mediana è meno fuorviante della media.

Se i dati del campione hanno distribuzione più o meno simmetrica è preferibile usare la media, altrimenti la mediana. Infatti, la media risente dei valori estremi, soprattutto in caso di numerosità non elevate. Entrambe, comunque, quantificano un valore tipico.

Usando esclusivamente la media, non siamo in grado di cogliere differenze tra campioni che hanno la stessa media, ma che sono molto diversi → **variabilità**: affiancata alla media, dà un'idea utile di come è fatto il campione.

MISURE DI VARIABILITÀ

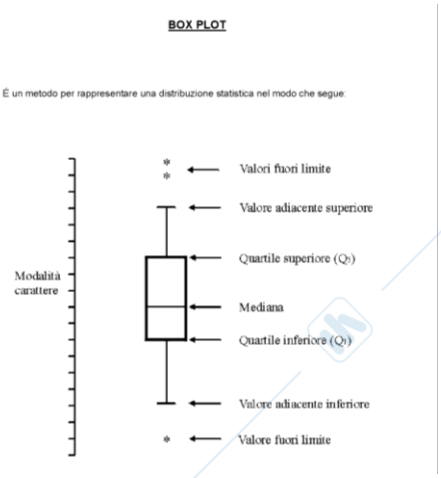
- **Range** o intervallo di variazione: differenza tra il minimo e il massimo. È un intervallo che contiene il 100% delle osservazioni.
- Range interquartile (IQR): intervallo tra il primo e il terzo quartile. È un intervallo che contiene il 50% dei valori centrali.

La variabilità si misura a partire dal concetto geometrico di distanza dalla media:

- **Varianza**: distanza media di ciascuna osservazione dalla media.
- **Deviazione standard**: è l'indice statistico che misura la dispersione delle singole osservazioni di un fenomeno intorno ad un valore atteso o alla sua stima.

RAPPRESENTAZIONE GRAFICA – VARIABILI QUANTITATIVE

- **Box** (and whiskers) **plot** (diagramma scatola e baffi): è la rappresentazione di mediana, range interquartile (IQR) e barre verticali (baffi) ed outliers.



LE PROBABILITÀ

— La probabilità è un numero non negativo, compreso tra 0 e 1.
— Se la probabilità è 0 l'evento è impossibile, se è uno è certo.
— S è lo spazio degli eventi. Certezza che si verifichi almeno uno fra gli eventi possibili.
— Se A e B sono eventi mutualmente esclusivi, cioè non possono verificarsi contemporaneamente.

- **Classica** (matematica): rapporto fra numero di eventi favorevoli e numero di eventi possibili, assumendo che tutti gli eventi siano egualmente possibili. È valida se tutti gli eventi sono egualmente possibili. $p = \frac{\text{n° eventi favorevoli}}{n° \text{ eventi possibili}}$
- **Frequentista**: frequenza relativa con cui un evento si realizza, al crescere del numero di ripetizioni dell'esperimento. È calcolata come frequenza relativa di successi osservati in un numero (tendente a) infinito di prove di un esperimento ripetibile. $p = \lim_{n\to\infty} \frac{Successi}{n}$
- **Soggettiva**: propensione soggettiva a credere nel realizzarsi di un evento. Preferibilmente basata su qualche base oggettiva. Opinioni informate.

MISURA DI ACCURATEZZA DIAGNOSTICA

— Sensibilità.
— Specificità.
— Valore predittivo positivo (VPP): probabilità di essere malato se il test ha dato esito positivo.
— Valore predittivo negativo (VPN): probabilità di essere non malato se il test ha dato esito negativo.

RISCHI E MISURE DI ASSOCIAZIONE FRA VARIABILI QUALITATIVE:

Associazione fra due variabili quantitative continue: assegnazione casuale, da parte del ricercatore, dei soggetti ad uno dei due trattamenti; confronto fra i due sottogruppi del numero (proporzione) di deceduti; valutazione efficacia.

- Confronto fra rischi di decesso: differenza fra rischi (RD) = rischio trattati – rischio controlli. (Prestare sempre attenzione a come è calcolato l'RD, e quindi a qual è il rischio sottratto all'altro. Cambiando il segno cambia l'interpretazione – se è fattore di rischio o protettivo).

— RD = 0 → i due rischi sono uguali. Non c'è differenza di rischio.
— RD > 0 → il rischio dei trattati è maggiore del rischio dei controlli (il trattamento è fattore di rischio.
— RD < 0 → il rischio dei trattati è minore del rischio dei controlli (il trattamento è fattore protettivo).

- Confronto (rapporto) tra rischi di decesso: rischio relativo (RR) = $\frac{\text{rischio trattati}}{\text{rischio controlli}}$

— RR = 1 → non c'è differenza di rischio.
— RR > 1 → il rischio dei trattati è maggiore del rischio dei controlli.

- RR < 1 → il rischio dei trattati è minore del rischio dei controlli.
- Interpretazione misure di confronto fra rischi:

RR	INTERPRETAZIONE		ESEMPIO
0.10	Rischio ridotto del 90%	1-RR: (1-0.10=0.90)	R_T=2% R_C=20%
0.25	Rischio ridotto del 75%	1-RR: (1-0.25=0.75)	R_T=5% R_C=20%
0.50	Rischio dimezzato	1-RR: (1-0.50=0.50)	R_T=10% R_C=20%
0.66	Rischio ridotto del 34%	1-RR: (1-0.66=0.34)	R_T=13.3% R_C=20%
0.75	Rischio ridotto del 25%	1-RR: (1-0.75=0.25)	R_T=15% R_C=20%
1	Assenza di effetto – rischio identico nei due gruppi		R_T=20% R_C=20%
1.25	Rischio aumentato del 25%		R_T=25% R_C=20%
1.5	Rischio aumentato del 50%		R_T=30% R_C=20%
2.0	Rischio aumentato del 100% (raddoppiato)		R_T=40% R_C=20%
4.0	Rischio quadruplicato		R_T=80% R_C=20%
10.0	Rischio decuplicato		R_T=40% R_C=4%

In uno studio caso-controllo non possiamo calcolare il rischio relativo, perché non possiamo calcolare i rischi di evento nei due campioni.
In sostituzione della % di esposti (esposti/totale) calcoliamo separatamente per casi e controlli il rapporto esposti / non esposti (odds).
Facendo il rapporto fra gli odds dei casi e gli odds dei controlli si ottiene una misura di associazione simile al rischio relativo.
Odds = p / 1-p = rapporto fra la probabilità che un evento si verfichi e la probabilità che l'evento non si verifichi.
È dimostrabile matematicamente che il rapporto fra odds è una misura interpretabile come un rischio relativo. Il rapporto fra questi due rapporti prende il nome di **odds ratio**

(OR) = $\frac{(n°\ casi\ esposti) * (n°\ controlli\ non\ esposti)}{(n°\ controlli\ esposti) * (n°\ casi\ non\ esposti)}$

L'interpretazione è simile a quella di un rischio relativo: misura della forza di associazione fra esposizione e malattia.

Odds ratio = $\frac{odds\ decesso\ esposti\ (o\ trattati)}{odds\ decesso\ controlli}$

- OR = 1 → non c'è differenza di odds (rischio) fra i due gruppi.
- OR > 1 → odds (rischio) degli esposti o trattati è maggiore del rischio dei controlli.
- OR < 1 → odds (rischio) degli esposti o trattati è minore del rischio dei controlli.

		EVENTO		
		SI	**NO**	
TRATTAMENTO (ESPOSIZIONE)	**SI**	a	b	**a + b**
	NO	c	d	**c + d**
		a + c	**b + d**	

RR = $\frac{a/(a+b)}{c/(c+d)} = \frac{rischo\ di\ evento\ fra\ i\ trattati\ (o\ esposti)}{rischio\ di\ evento\ fra\ i\ non\ trattati\ (o\ non\ esposti)}$

OR = $\frac{a*d}{c*b} = \frac{odds\ di\ evento\ fra\ i\ trattati\ (o\ esposti)}{odds\ di\ evento\ fra\ i\ non\ trattati\ (o\ non\ esposti)}$

➔ La scelta di calcolare RR o OR deve essere effettuata sulla base del disegno dello studio.

Studio prospettico (coorte, trial clinico) → utilizzare il rischio relativo (o differenza di rischi)

Studio retrospettivo (caso-controllo) → utilizzare l'odds ratio.

SINTESI

		EVENTO		
		SI	**NO**	
TRATTAMENTO (ESPOSIZIONE)	**SI**	a	b	**a + b**
	NO	c	d	**c + d**
		a + c	**b + d**	

Rischio trattati $= \frac{a}{a+b}$

Rischio non trattati $= \frac{c}{c+d}$

- **Differenza di rischio** $(\mathrm{RD}) = \frac{a}{a+b} - \frac{c}{c+d}$
  - RD = 0 → assenza di effetto.
  - RD > 0 → fattore di rischio (protettivo).
  - RD < 0 → fattore protettivo (di rischio).
- **Rischio relativo** $(\mathrm{RR}) = \frac{a/(a+b)}{c/(c+d)}$
  - RR = 1 → assenza di effetto.
  - RR > 1 → fattore di rischio.
  - 0 < RR < 1 → fattore protettivo.
- **Odds ratio** $(\mathrm{OR}) = \frac{a*d}{b*c}$
  - OR = 1 → assenza di effetto.
  - OR > 1 → fattore di rischio.
  - 0 < OR < 1 → fattore protettivo.

ASSOCIAZIONE FRA DUE VARIABILI QUANTITATIVE CONTINUE

Correlazione lineare → grafico a dispersione: rappresenta il valore congiuntamente di due variabili quantitative. Ci permette di vedere se esiste e di che tipo è un'associazione tra le due variabili.

Coefficiente di correlazione lineare (Pearson): indice che misura l'associazione lineare fra due variabili; misura il grado in cui due variabili si "muovono assieme" in maniera lineare (variabili quantitative continue).

- r = -1 → correlazione negativa. La relazione di proporzionalità inversa perfetta. Al crescere di x, diminuisce y (e viceversa).
- r = 0 → assenza di correlazione. Non esiste nessuna relazione tra le variabili, i valori di una variabile variano indipendentemente da quelli dell'altra.
- r = +1 → correlazione positiva. Tutti i punti cadono su una linea retta, la relazione tra le due variabili è perfetta. Al crescere di x, cresce anche y.

La valutazione della forza della correlazione è soggettiva e dipende dal contesto. Non importa quale sia la pendenza della retta: il coefficiente ci dice solo se è positiva o negativa.

Regressione lineare: quantificare come variazioni di valore una variabile causano variazioni di valore nell'altra variabile (possibile relazione di causa effetto): individuare variabile dipendente e indipendente e poi rappresentare i dati su un grafico a dispersione e si individua la retta ($y= \alpha+\beta * x$[1]). La relazione trovata è valida solo nel range di valori trovati.

L'intercetta (α) indica il valore medio della variabile dipendente in corrispondenza del valore 0 della variabile indipendente. Non sempre ha un senso la sua interpretazione.

Il coefficiente angolare (β, coefficiente di regressione) indica l'intensità della relazione lineare fra le variabili. Interpretazione geometrica: pendenza della retta. Mediamente, ad un incremento unitario della variabile indipendente corrisponde un incremento di intensità β della variabile dipendente.

Criterio dei minimi quadrati: retta che minimizza la somma del quadrato degli scarti (residui) dei dati osservati dalla retta.

Relazione fra variabili qualitative dicotomiche (che assumono due valori).

Tabella a doppia entrata (2 variabili).

[1] Formula della retta in statistica (equazione di regressione permette di studiare la relazione esistente fra due variabili (x e y) quantitative: y è la variabile dipendente; x è la variabile indipendente; α è l'intercetta; β è il coefficiente angolare (coefficiente di regressione).

ESERCIZI

STUDI – esempi

1. FUMARE FA MALE?
   - Obiettivo: valutare l'associazione fra esposizione a fumo di sigaretta e mortalità.
   - Studio: osserviamo un gruppo di fumatori ed un gruppo di non fumatori, e vediamo se la mortalità è differente fra i due gruppi.
   - Arruolamento: selezione di soggetti fumatori o ex fumatori (esposti); selezione di soggetti non fumatori, ma per il resto molto simili agli esposti (non esposti).
   - Rilevazione dell'evento: follow up – seguo tutti i soggetti per un certo periodo di tempo e vedo quanti si ammalano nei due gruppi.
   - Studio di coorte.
   - Risultati.
2. LAVORARE FA MALE?
   - Obiettivo: valutare l'associazione fra quantità di ore di lavoro e infarto miocardico acuto.
   - Arruolamento: selezione di soggetti che hanno avuto un infarto (casi) e di soggetti che non hanno avuto un infarto, ma per il resto molto simili ai casi.
   - Rilevazione dell'esposizione al fattore di rischio: questionario – raccolta informazione quantità di lavoro nel mese precedente.
   - Studio caso-controllo.
   - Risultati: chi ha avuto l'infarto aveva lavorato di più nell'ultimo mese?
3. TAC E CORONOGRAFIA
   - Si valuta se un nuovo test (meno invasivo) è sufficientemente "buono" nell'individuare coloro che hanno la malattia coronarica significativa (malati), confrontandolo con un metodo ritenuto ottimale.
   - Studio trasversale.
4. STREPTOMICINA
   - Obiettivo: valutare l'effetto della streptomicina nel trattamento della tubercolosi polmonare.
   - Arruolamento: selezione di soggetti affetti da tubercolosi e assegnazione casuale dei soggetti ad uno dei due trattamenti (trattamento e controllo).
   - Rilevazione dell'esito (guarigione o decesso a 6 mesi).
   - Risultati: chi ha preso streptomicina ha avuto un esito migliore rispetto a chi ha preso altro?
   - Studio clinico randomizzato.

FREQUENZE – ESERCIZI

1.

Titolo di studio	Frequenza semplice Assoluta	Relativa	Frequenza cumulata Assoluta	Relativa
Licenza Elementare	45	15.8%	45	15.8%
Licenza Media	101	35.2%	146	51.0%
Diploma 2-3 anni	18	6.0%	164	57.0%
Diploma 4-5 anni	89	31.1%	253	88.1%
Laurea	34	11.9%	287	100.0%
Totale	*287*	*100.0%*		

Possiamo affermare che…?

- Il 57% dei soggetti ha titolo di studio "diploma 2-3 anni → falso.
- Il 49% dei soggetti ha titolo di studio superiore alla "Licenza media" → vero.
- L'88.1% dei soggetti ha titolo di studio non superiore al "diploma 4-5 anni" → vero.
- Il 57% dei soggetti ha titolo di studio inferiore al "diploma 4-5 anni" → vero.
- L'11.9% dei soggetti ha titolo di studio superiore o uguale al "diploma 4-5 anni" → falso.

VARIABILI – ESEMPI

1. Variabili quantitative discrete: voto ottenuto all'esame di statistica da 170 studenti.

Voto	frequenza assoluta	frequenza relativa	frequenza cumulata assoluta	frequenza cumulata relativa
18	2	0.012	2	0.012
19	4	0.024	6	0.035
20	5	0.029	11	0.065
21	9	0.053	20	0.118
22	14	0.082	34	0.200
23	3	0.018	37	0.218
24	12	0.071	49	0.288
25	20	0.118	69	0.406
26	29	0.171	98	0.576
27	32	0.188	130	0.765
28	15	0.088	145	0.853
29	16	0.094	161	0.947
30	9	0.053	170	1.000
Totale	170			

2. Variabili quantitative continue: pressione sistolica (PAS) misurata in 12 adulti:

PAS	frequenza assoluta	frequenza relativa	frequenza cumulata assoluta	frequenza cumulata relativa
81	1	8.3%	1	8.3%
87	1	8.3%	2	16.7%
93	3	25.0%	5	41.7%
96	1	8.3%	6	50.0%
105	2	16.7%	8	66.7%
123	1	8.3%	9	75.0%
148	1	8.3%	10	83.3%
169	1	8.3%	11	91.7%
187	1	8.3%	12	100.0%
Totale	12	100.0%		

ESERCIZIO LETTURA TABELLE:

1. Studio per valutare se esiste un'associazione fra utilizzo di internet e genere.
   L'utilizzo di internet è maggiormente diffuso nei maschi o nelle femmine?

	Usano Internet	*Non usano Internet*	totale
maschi	187	84	**271**
femmine	174	115	**289**
totale	**361**	**199**	**560**

	Usano Internet	*Non usano Internet*	totale
maschi	33.4%	15.0%	**48.4%**
femmine	31.1%	20.5%	**51.6%**
totale	**64.5%**	**35.5%**	**100.0%**

	Usano Internet	*Non usano Internet*	totale
maschi	69.0%	31.0%	**100.0%**
femmine	60.2%	39.8%	**100.0%**
totale	**64.5%**	**35.5%**	**100.0%**

	Usano Internet	*Non usano Internet*	totale
maschi	51.8%	42.2%	**48.4%**
femmine	48.2%	57.8%	**51.6%**
totale	**100.0%**	**100.0%**	**100.0%**

L'utilizzo di internet è maggiormente diffuso fra i maschi o fra le femmine?

I maschi utilizzano maggiormente internet: il 69% dei maschi usa internet contro il 60.2% delle femmine (tabella 3).

2. INDAGINE SULLA SODDISFAZIONE DELLE CONDIZIONI DI SALUTE. RISULTATI PER GENERE.

Indagine sulla soddisfazione delle condizioni di salute. Risultati per genere.

	Soddisfatto		totale
	Si	*No*	
maschi	212	39	**251**
femmine	250	66	**316**
totale	**462**	**105**	**567**

	Soddisfatto		totale
	Si	*No*	
maschi	45.9%	37.1%	**44.3%**
femmine	54.1%	62.9%	**55.7%**
totale	**100.0%**	**100.0%**	**100.0%**

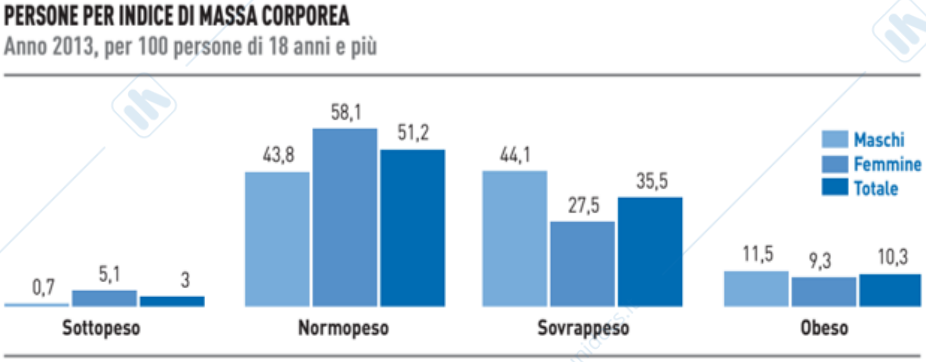
	Soddisfatto		totale
	Si	*No*	
maschi	84.5%	15.5%	**100.0%**
femmine	79.1%	20.9%	**100.0%**
totale	**81.5%**	**18.5%**	**100.0%**

Possiamo affermare che:

1. ✗ *Le femmine presentano un maggior livello di soddisfazione rispetto ai maschi*
2. ✗ *Il 37.1% dei maschi si dichiarano non soddisfatti*
3. ✓ *Il 79.1% delle femmine si dichiarano soddisfatte*
4. ✓ *Il 62.9% dei non soddisfatti sono femmine*

GRAFICI – esercizi

1. PERSONE PER INDICE DI MASSA CORPOREA:



Leggete ed interpretate il grafico

Possiamo affermare che:

✗ Fra le persone normopeso, la maggioranza (58.1%) sono femmine.

✓ Fra i maschi, la maggior parte dei soggetti (44.1%) sono sovrappeso.

2. Presso l'università di Paperopoli esistono solo due corsi di laurea, economia e lettere. Il 60% degli studenti segue il corso di laurea in economia. Nessuno abbandona gli studi. Un'indagine recente mostra che il 65% dei laureati presso l'università di paperopoli trova lavoro entro un anno dalla laurea. La stessa indagine inoltre riporta che, fra i giovani laureati in economia, il 25% non è riuscito a trovare un lavoro ad un anno dalla laurea. A quanto ammonta la proporzione di laureati in lettere fra coloro che hanno trovato lavoro entro un anno dalla laurea?

		Laurea		
		Economia	Lettere	Totale
Lavora	SI	45	20	65
	NO	15	20	35
	Totale	60	40	100

Proporzione "Lettere" fra "Lavora = SI": 20/65=30.8%

MEDIANA – ESERCIZI

Abbiamo rilevato le giornate di degenza di 41 pazienti ricoverati in un reparto ospedaliero. Calcolare la mediana della durata di degenza.

giorni degenza	1	2	3	4	5	6	7	8	9	11	14	15	18
n° di pazienti	1	3	8	12	7	2	0	2	1	1	2	1	1

La durata di degenza media è di 4 giorni.

PROBABILITÀ – ESERCIZI

1. A quanto ammonta la probabilità che si verifichi un incidente aereo?
   Ogni giorno ci sono circa 95000 voli di linea nel mondo. Ciò significa circa 34.675.000 di voli all'anno.
   Probabilità di incidente aereo mortale: consideriamo 10 incidenti mortali all'anno.
   Probabilità incidente mortale: 10/34675000 = 0.28/1000000
2. STUDIO DI ACCURATEZZA DIAGNOSTICA
   E' stato condotto uno studio per valutare l'accuratezza di uno strumento, l'Autism Diagnostic Observation Schedule (ADOS), nella diagnosi di disturbo dello spettro autistico in bambini in età prescolare (< 6 anni) con sospetto di disturbo autistico. Sono stati arruolati 180 bambini (con sospetto di disturbo) che sono stati sottoposti all'ADOS e ad una valutazione multidisciplinare (che richiede tempo e richiede un giudizio clinico) effettuata da un'equipe di esperti (reference standard). Le due valutazioni (ADOS e reference standard) sono state effettuate, per ogni bambino, ad una distanza massima di 3 mesi l'una dall'altra.
   Qual è il disegno dello studio? → Trasversale (cross-sectional).
   - Obiettivo: ADOS è uno strumento utile per la diagnosi di disturbo dello spettro autistico?
   - Index test: nuovo test da valutare (ADOS)
   - Malattia: disturbo dello spettro autistico.
   - Reference standard: valutazione multidisciplinare fatta da equipe di esperti.
   - ADOS: valutazione che si conclude con l'attribuzione di un punteggio numerico. Se punteggio superiore ad una soglia numerica (es. >7) allora test positivo. Assumiamo che l'esito fornito dalla valutazione multidisciplinare di esperti

(reference standard) permetta di fare diagnosi di disturbo autistico senza possibilità di errore.

— Come fare per valutare, quantitativamente, l'utilità dell'ADOS nella diagnosi di disturbo dello spettro autistico?

		Disturbo Autistico: Si	Disturbo Autistico: No	tot
ADOS	+	VP 63	22 FP	85
	-	FN 7	88 VN	95
tot		70	110	180

Veri positivi (VP): pazienti positivi al test ed affetti dalla malattia

Falsi positivi (FP): pazienti positivi al test ma non affetti dalla malattia

Falsi negativi (FN): pazienti negativi al test ma affetti dalla malattia

Veri negativi (VN): pazienti negativi al test e non affetti dalla malattia

—

ESERCIZIO 2

In uno studio di accuratezza diagnostica, per valutare il test T, abbiamo arruolato 500 pazienti sospettati di essere affetti dalla malattia M. fra i 50 pazienti che effettivamente sono risultati affetti da M, 28 hanno avuto un test positivo. Inoltre, il valore predittivo del test T è risultato essere 0.50.

Calcolare la probabilità di risultare positivo al test T per un paziente davvero affetto dalla malattia M → P (+ / M) = 28/50=0.560 (sensibilità).

Calcolare la probabilità di avere la malattia M per un paziente risultato negativo al test T. → P(M/-)=22/444=0.049

		M	non M	tot
Test	+	28	28	
	-			
tot		50	450	500

		M	non M	tot
Test	+	28	28	56
	-	22	422	444
tot		50	450	500

ESERCIZIO 3

Il 14% della popolazione di età > 85 anni soffre