

Corso di Laurea in **Medicina e Chirurgia** - 1° anno
C.I. Metodologia Medico Scientifica e
Metodi Quantitativi in Biomedicina

A.A.
2017/18

STATISTICA MEDICA

Prof. STEFANIA ROSSI
Dip. Medicina Molecolare e dello Sviluppo
stefania.rossi@unisi.it

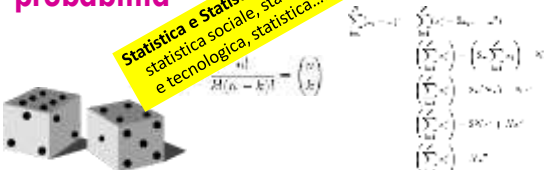


Presentazione del Modulo

STATISTICA

È una scienza di
derivazione **matematica**
con basi di **calcolo**
probabilità

Statistica e Statistica applicata: statistica economica,
statistica sociale, statistica per la ricerca sperimentale
e tecnologica, statistica.....medica



Statistica Medica

ovvero.....

L'USO della **Statistica**
nella **Ricerca Medica**



«ricerca scientifica»

- **GENERALIZZABILE** - siamo in grado di applicare i risultati non solo ai pazienti nello studio
- **EMPIRICAMENTE VERIFICABILE, RIPETIBILE** - ci viene detto **come**, **dove** e **quando** è stato fatto, così che noi possiamo verificare ciò che è stato effettivamente fatto e ripeterlo per verificare di poter ottenere gli stessi risultati
- **TRASPARENZA DEL PROCESSO DECISIONALE** - sappiamo **cosa** è stato fatto e **perché**.
- **COSTRUITO A PARTIRE DAI RISULTATI OTTENUTI DA ALTRI** (**accrescimento delle conoscenze**)-
- **GENERA, A SUA VOLTA, NUOVE IDEE DA TESTARE**

Libri consigliati

Rispondono alla necessità di conoscenze adeguate della Metodologia della Ricerca indispensabile per una progettazione efficiente e l'uso corretto della statistica –**ESERCIZI con soluzioni**

CAMPBELL MJ – MACHIN D, Statistica Medica. Un approccio Evidence-Based, Ediz. CSE Torino 2005


Trattazione concettuale, pratica e conciso

Studenti di Medicina e Infermieristica
Medici e operatori delle Professioni Sanitarie

FOWLER J. - JARVIS P. - CHEVANNES M, Statistica per Le Professioni Sanitarie, Ediz. Edises Napoli 2006

Trattazione scolastica con formule, **parte da zero** senza dare per scontata alcuna conoscenza precedente


Studenti e operatori delle Professioni Sanitarie




Def. «rivista» (1996)

EBM – Evidence Based Medicine

È un movimento culturale, che «**costituisce un approccio alla pratica clinica dove le decisioni cliniche risultano dall'integrazione tra l'esperienza del medico e l'utilizzo coscienzioso, esplicito e giudizioso delle migliori evidence scientifiche disponibili, mediate dalle preferenze del paziente**»

<http://www.gimbe.org> 

Perché insegnare la Statistica Medica?

Migliorare le conoscenze futuri medici
«**Consumatori/CO-produttori**» di ricerca
perché:

- **Poter essere autonomi** nell'interpretazione ottimale dei risultati della ricerca
- **Collaborare con i colleghi, compresi lo statistico e l'epidemiologo** per migliorare la qualità della progettazione degli studi

Letteratura Medica



Tutela del lettore da pubblicazioni fuorvianti?

- Riviste prestigiose: filtro=referee specialisti della disciplina, ma anche statistici
- Filtro manca nel caso di pubblicazioni generiche e nella maggior parte delle pubblicazioni promozionali sponsorizzate da aziende con interessi specifici

GARANZIA COMPLETA = impossibile

SOLUZIONE = saper interpretare correttamente i risultati della ricerca scientifica

25

“CO-PRODUTTORI “DI RICERCA MEDICA



Progettazione della Ricerca

Aspetti fondamentali di un Protocollo di RICERCA

1. **Il quesito di ricerca** (⇒ ipotesi di ricerca)
2. **Il razionale della ricerca** (rilevanza e novità)
3. **Il disegno dello studio** (in funzione dell'ipotesi)
4. **Popolazione di riferimento e Campione**
5. **I metodi di valutazione/misurazione delle variabili**
6. **La pianificazione dell'analisi statistica** (in funzione dell'ipotesi e, implicitamente, del tipo di variabili coinvolte)

Punto di partenza di un PROGETTO DI RICERCA

QUESITO DI RICERCA ben definito
a cui lo studio deve fornire una risposta

Quesito di ricerca

ESEMPIO:

- Nella cura prenatale di routine all'inizio del secondo trimestre di gravidanza, incorporando un semplice programma di screening e trattamento per le infezioni vaginali*subcliniche si può ridurre il tasso di parti pretermine spontaneo del 50%?

La produzione della ricerca

è caratterizzata da alcuni **step fondamentali**



LE FASI DELLA PROGETTAZIONE



Momento in cui una o più persone ipotizzano di realizzare un progetto, oppure decidono di accogliere una richiesta esterna (decodifica della domanda)

FONTI:

- Esperienza (clinica, ...) quotidiana
- Opportunità di finanziamento (Proposal or Foundation announcements)
- Confronto con colleghi (Congressi, ...)
-

LE FASI DELLA PROGETTAZIONE



Una volta avviata una prima ipotesi di progetto bisogna cercare di verificare quali sono le **risorse** (umane, strumentali, temporali, finanziarie ...) disponibili, identificare meglio ruoli e funzioni propri e degli altri soggetti coinvolti, identificare **rilevanza ed originalità/novità del quesito**.....



Elaborazione di un progetto cartaceo (**protocollo***) con identificazione e programmazione delle fasi di indagine

ACCURATA PROGETTAZIONE

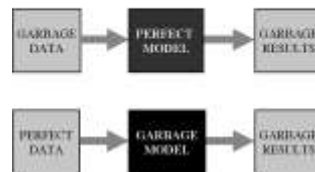
implica

PRECISA ESECUZIONE

- **GIGO** (garbage in, garbage out) is a concise expression of a concept common to computer science and mathematics: the quality of output is determined by the ...



MODEL CALCULATIONS
"Garbage In-garbage Out" Paradigm



LA TERMINOLOGIA

La Statistica ha una sua terminologia. Molti termini sono familiari; alcuni sono usati nel linguaggio ordinario sia pure in accezioni leggermente diverse

- Universo e Campione
- Unità di analisi
- Variabili e Dati
- Parametri e Stime
- Simboli alfabeto greco e latino



- **UNIVERSO o POPOLAZIONE** = concetto teorico per indicare un insieme di **UNITÀ di ANALISI**
- **CAMPIONI** = un sotto-insieme di **UNITÀ di ANALISI** estratte da una popolazione per ottenere **stime** dei parametri della popolazione stessa

Esempio:

- **Popolazione:** insieme di tutti gli studenti iscritti al 1° anno del CdL in *Medicina e Chirurgia* nell'a.a 2016/17 (Siena, N=240)
- **Campione:** n= 100 studenti estratti dal totale dei 240 studenti.

UNITÀ DI ANALISI: studentE iscritto al 1° anno del CdL in *Medicina e Chirurgia* nell'a.a 2016/17 (a Siena)

Sono UNITÀ DI ANALISI.....

- In uno studio sull'*Obesità* in età pediatrica?
- **bambini e adolescenti da 0 a 14 anni**
- In uno studio sull'*inquinamento indoor nelle scuole*?
- **le aule**
- In uno studio sulla % di disoccupati nelle regioni italiane?
- **le regioni**

Variabili & Dati

Si dice **variabile** qualsiasi **caratteristica soggetta a variazione**, che sia **misurabile/rilevabile** su un insieme di unità di analisi

I **dati** sono il risultato della misurazione

=

sono i valori numerici o le modalità assunti dalle variabili

Esempi:

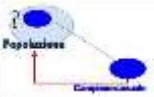
- **Valori numerici...**
 - 1,2,3 , 100 *anni*, per la variabile "Età"
 - 2, 10, 20, 50,*chili*, per il "Peso corporeo"
 - 80, 90, 110,.....*mmHg*, per la "Pressione arteriosa sistolica"
- **Modalità...**
 - *maschio* o *femmina* per la variabile "Sesso"
 - *Licenza elementare-Diploma media inferiore-Diploma maturità-Laurea*, per la variabile "Titolo di studio"

...assunti dalle variabili

PARAMETRI: statistiche usate per **descrivere** le caratteristiche di una popolazione - $\mu = \text{media}$



STIMATORE: ogni statistica, calcolata sui dati **campionari**, che fornisce una **stima di un parametro** della popolazione - $m = \text{media}$



$$\bar{x}$$

PARAMETRI di una Popolazione

Esempi:

- L'**altezza media** degli studenti iscritti al 1° anno del CdL in Medicina e Chirurgia nell'a.a 2012/13 – Siena (μ)
- La **percentuale** di studenti provenienti dal Liceo Classico tra i 235 studenti iscritti al 1° anno del CdL in Medicina e Chirurgia nell'a.a 2012/13 Siena (π)

Aspetti fondamentali di un Protocollo di RICERCA*



N.B. Protocollo di ricerca # **Protocollo Clinico:** LINEE GUIDA che descrivono la «buona pratica» nelle diverse situazioni cliniche.
Per es. **PERCORSO DIAGNOSTICO-TERAPEUTICO-** assistenziale nel reparto di MEU, in caso di sospetto infarto

Aspetti STATISTICI di un Protocollo di RICERCA



1. **Il quesito di ricerca** (\Rightarrow ipotesi di ricerca)
2. **Il razionale della ricerca** (rilevanza e novità)
3. **Il disegno dello studio** (in funzione dell'ipotesi)
4. **Popolazione di riferimento e Campione**
5. **I metodi di valutazione/misurazione delle variabili**
6. **La pianificazione dell'analisi statistica** (in funzione dell'ipotesi e, implicitamente, del tipo di variabili coinvolte)

Protocollo di RICERCA

1. **Il quesito di ricerca** (\Rightarrow ipotesi di ricerca)
2. **Il razionale della ricerca** (rilevanza e novità)
3. **Il disegno dello studio** (in funzione dell'ipotesi)
4. **Popolazione di riferimento e Campione**
5. **I metodi di valutazione/misurazione delle variabili**
6. **La pianificazione dell'analisi statistica** (in funzione dell'ipotesi e, implicitamente, del tipo di variabili coinvolte)

CARATTERISTICHE DI UN BUON QUESITO DI RICERCA:

FINER

2. Il razionale della ricerca (rilevanza e novità)

1314 RESEARCH QUESTION Research Fundamentals II: Choosing and Defining a Research Question
THOMAS RWYSEKAWSKI, MD, ROBERT REVERMAN, MD

F...Feasible. There are many issues that must be addressed when initiating a study that will determine whether it can be pursued **successfully**. It is essential to determine whether the research question can be **answered in a specific environment** (time, resources, and personnel) and within the next important question asked. **Feasibility** is a function of the **adequate time and resources**.

I...Interesting. When initiating a project, the investigator must have a **strong interest** in the question being asked.

R...Relevant. [Tomkowiak JM, Gunderson AJ. To IRB or Not to IRB? Acad Med. 2004;79:628-32.]

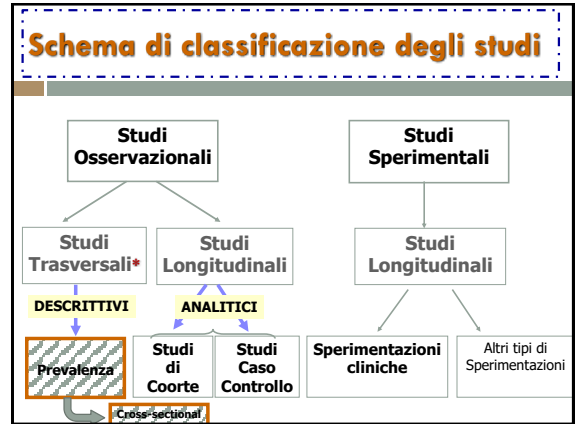
R...Relevant. As previously defined, **Institutional Review Board** research is to discover new information that will **improve overall (health status)**.

www.gimbe.univ.it

PILLOLE DI METODOLOGIA DELLA RICERCA

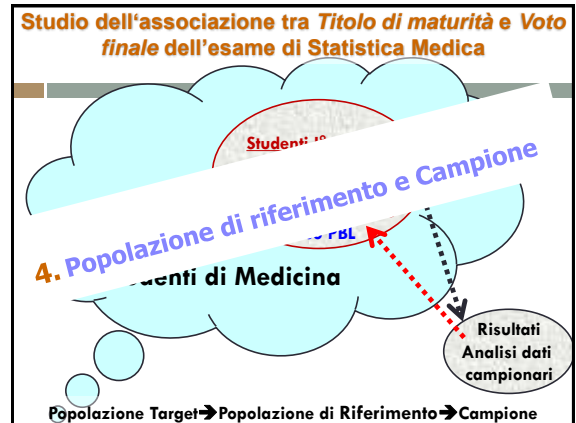
Architettura della ricerca clinica
Come scegliere il disegno di studio appropriato?

- Classificare i **disegni di studio** che appartengono alla ricerca primaria **QUANTITATIVA**
- Definire per quali **quesiti di ricerca** possono essere utilizzati i vari disegni di studio



3. Il disegno dello studio (in funzione del quesito)

CATEGORIA	Quesito clinico <i>Esempio</i>
Eziologia/Rischio Caso-controllo Coorte con corte parallela	Qual è la responsabilità eziologica del fattore di rischio X nell'insorgenza della malattia Y? <i>L'utilizzo del telefono cellulare aumenta il rischio di neoplasia cerebrale?</i>
Diagnosi Cross-sectional	Qual è l'accuratezza del test diagnostico X (rispetto al gold-standard Y) nella diagnosi della malattia Z? <i>Quale è l'accuratezza diagnostica della risonanza magnetica nucleare nei pazienti con sospetta lesione del menisco?</i>
Prognosi Coorte senza corte parallela	Qual è la storia naturale della malattia X e la potenza dei fattori prognostici? <i>In un paziente con neurite ottica, quali sono i fattori prognostici (favorevoli e sfavorevoli) che condizionano l'evoluzione in sclerosi multipla?</i>
Terapia Sperimentazione clinica controllata e randomizzata	Qual è l' EFFICACIA del trattamento X (preventivo, terapeutico o riabilitativo), rispetto al trattamento Y, nella malattia Z? <i>In un paziente con osteoartrite in trattamento cronico con FANS (malattia/condizione), l'omeprazolo (intervento), rispetto al misoprostolo (confronto) è in grado di prevenire l'ulcera da FANS sintomatica e le sue complicanze (evento)?</i>



Criteri di eleggibilità

Definiscono le caratteristiche dei soggetti da includere in uno studio. Sono distinti in criteri di **inclusione** e criteri di **esclusione**:

Criteri di Inclusione = si utilizzano per definire il tipo di paziente che corrisponde all'obiettivo di ricerca: **caratteristiche demografiche relative al paziente** (ad esempio sesso, età), alla **malattia** (Es. Gravità, Tipo istologico, ...) oppure alle modalità diagnostiche che identificano la popolazione oggetto dello studio, ma anche **temporali** e **geografiche**

Caratteristiche cliniche, ESEMPIO: soggetti affetti da asma lieve, non possono essere buoni candidati per uno studio sugli effetti di un nuovo farmaco sul tasso di ospedalizzazione per attacchi d'asma!

Criteri di Esclusione: restrizioni applicate alla popolazione includibile in uno studio per ragioni di **sicurezza** o di **fattibilità**

- Le Variabili confondenti sono usate spesso come criteri di esclusione
- Non parlare italiano o... /non saper leggere e/o scrivere
- Soggetti carcerati, per evitare potenziali violazioni etiche, poiché potrebbero non sentirsi liberi di rifiutare di partecipare allo studio

4.1 Popolazione di riferimento e Campione

- Quali sono criteri di inclusione ed esclusione dei soggetti per definire la popolazione di riferimento?
- Se lo studio è sperimentale, come si intende **assegnare** i soggetti a gruppi sperimentali?

Aspetti STATISTICI di un

Protocollo di RICERCA

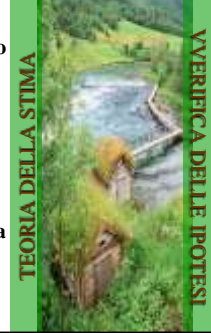
1. Il quesito di ricerca (⇒ ipotesi di ricerca)
2. Il razionale della ricerca (rilevanza e novità)
3. Il disegno dello studio (in funzione dell'ipotesi)
4. Popolazione di riferimento e Campione
5. I metodi di valutazione/ misurazione delle variabili

La pianificazione dell'analisi statistica (in funzione dell'ipotesi e, implicitamente, del tipo di variabili coinvolte)

Metodo del campionamento statistico

Procedimento logico in base al quale l'insieme delle norme per la formazione del campione **si salda** alla teoria della stima e della verifica delle ipotesi

TEORIA DELLA STIMA



VERIFICA DELLE IPOTESI

Campionamento probabilistico = l'insieme delle regole seguite nella formazione di un campione ci consentono di conoscere a priori le probabilità di inclusione di ciascuna unità di analisi nel campione

Campionamento Probabilistico

(Ponte tra analisi Descrittiva e Inferenziale)

Descrittiva



Inferenziale

Confronto tra campionamento probabilistico e non probabilistico	
Campionamento probabilistico	Campionamento non probabilistico
Obiettivo	Conoscenza
Procedimento	Arbitrario
Popolazione	Definita
Selezione	Arbitraria
Validità	Non valutata
Risultato	Non valutato

Tipo di campionamento

- Campionamento probabilistico: Campionamento NON –probabilistico:

Campione casuale semplice
Campione casuale stratificato
Campione casuale sistematico
Campione a grappolo (Cluster)

Campione per quote
Campione di convenienza
Campione a palla di neve

N.B. Sebbene le tecniche non-probabilistiche abbiano un valore scientifico minore, rappresentano il tipo di campionamento più utilizzato nella ricerca biomedica.

Campionamento NON –probabilistico

Campione di convenienza = soggetti che sono convenienti per il ricercatore

Campione a palla di neve (*snowball*) = si identifica il primo soggetto, che recluta altri soggetti, che reclutano altri....

Campione per quote = i soggetti vengono scelti in proporzione al loro numero nella popolazione di interesse (utilizzato prevalentemente in ricerche di market)

Campione di volontari (*judgmental*)

Che cosa fare se non è possibile utilizzare la selezione casuale?

- Cercare di ridurre al minimo il bias di selezione utilizzando chiari ed espliciti criteri di inclusione ed esclusione
- Campione consecutivo di casi..... ★



**BMJ 1999 Oct 16;319(7216):1027-30:
Stressful life events and difficulties and onset of breast cancer: case-control study.**

Protheroe D, Turvey K, Horgan K, Benson E, Bowers D, House A.
Department of Liaison Psychiatry, Leeds General Infirmary, Leeds LS1 3EX.

- **CAMPIONI** = sono estratti dalla popolazione per ottenere *stime* della popolazione stessa



- Nella maggior parte dei casi la ricerca biomedica cerca di risalire alle caratteristiche di una popolazione partendo da quelle di un **campione rappresentativo**, ovvero un numero limitato di individui che nel loro complesso rappresentano le caratteristiche della popolazione
- Tutte le volte che è possibile si cerca di utilizzare un **campione casuale semplice**

DISEGNO DI CAMPIONAMENTO

- Definizioni: casuale e probabilistico
- Tipi di campionamento probabilistico
- Il Campione casuale semplice

Campione: insieme ridotto di elementi estratto dalla popolazione e considerato rappresentativo della stessa.

La selezione del campione (**campionamento**) deve essere operata mediante rigorosi criteri di casualità e di rappresentatività, impiegando specifiche metodologie statistiche di estrazione.

Esistono diverse modalità di campionamento, tra le quali le più frequentemente impiegate sono:

campionamento casuale semplice;

campionamento sistematico;

campionamento stratificato proporzionale;

Campionamento a grappolo

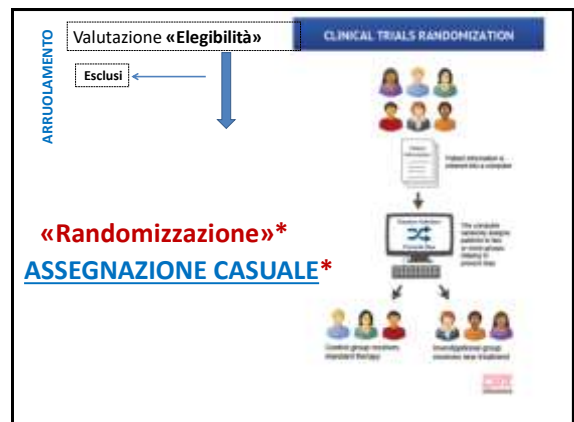
Il **campionamento casuale semplice** è la tecnica più semplice di selezione di un campione (simile allo schema di estrazione da un'urna).

La metodologia che sta alla base della scelta del campione è la **selezione «random»***

In un campione casuale semplice ogni individuo della popolazione ha la stessa probabilità di essere scelto ed inoltre campioni della stessa dimensione hanno tutti la stessa probabilità di essere selezionati.

In pratica per scegliere il campione si può usare una tecnica basata sulle **tavole dei numeri casuali**.

«randomizzazione»*



NOTA:**Sperimentazioni Cliniche controllate e randomizzate**

In una sperimentazione è fondamentale ottenere dei gruppi di studio confrontabili non solo per le variabili note, ma anche per i fattori non noti che potrebbero influenzare il risultato.

La soluzione ideale sarebbe quella di creare due gruppi di pazienti identici, però....

almeno in termini di prognosi si ignorano i fattori che permettono di determinare con precisione la prognosi di ogni paziente

inoltre esiste una enorme variabilità fra i pazienti affetti da una stessa malattia nella distribuzione delle varie caratteristiche

RANDOMIZZAZIONE:

Il problema è stato superato capovolgendo l'approccio al problema: INVECE DI CERCARE DI RENDERE CONFRONTABILI I DUE GRUPPI IN TERMINI DI PROGNOSI, SI AMMETTE CHE I DUE GRUPPI POSSANO ESSERE DIVERSI TRA DI LORO A CONDIZIONE CHE LE DIFFERENZE SIANO ESCLUSIVAMENTE DETERMINATE DAL CASO

**Ancora a proposito di Selezione Random e Randomizzazione
Validità esterna e Validità interna dei risultati di uno studio**

- La **selezione casuale** (*random*) si riferisce al **campionamento**, ovvero al modo in cui vengono selezionate le unità campionarie da includere nel campione ed è riferibile alla «**validità ESTERNA**» (o **generalizzabilità**) **dei risultati della ricerca**. Infatti, utilizziamo la selezione casuale proprio perché il nostro campione rappresenti nel modo migliore possibile la popolazione da cui è estratto.
- La **randomizzazione**, si riferisce invece alle **caratteristiche del disegno** di studio, ovvero, quando noi assegniamo i soggetti partecipanti ai due gruppi di trattamento, noi abbiamo scelto, per definizione, uno **studio sperimentale**. L'assegnazione casuale è dunque relativa più alla «**validità INTERNA**» **dei risultati della ricerca**. Infatti, utilizziamo l'assegnazione casuale dei soggetti, proprio per assicurarci che i «**gruppi a confronto siano simili prima del trattamento**».....

**Tipi di campioni**

- Campioni non probabilistici
- Campioni probabilistici

Campioni non probabilistici

- **Volontari**
- **Campioni di convenienza**
- **A scelta ragionata** (per es. campioni di unità tipiche)

Probabilità di essere selezionati: sconosciute!!!!

- mancano di **accuratezza** (→) a causa della selezione distorta
- impossibile **generalizzare** i risultati

Rappresentatività:

Campione = immagine ridotta (quantitativamente) fedele dal punto di vista delle caratteristiche (qualitativamente)

'Garantirsi' la rappresentatività prima, confermarla dopo!!!!!!

CAMPIONAMENTO CASUALE SEMPLICE■ **Vantaggi**

- Semplice
- Si calcola facilmente l'errore campionario

■ **Svantaggi**

- C'è bisogno della lista completa delle unità
- Non sempre raggiunge la **miglior** rappresentatività

- **Si effettuano estrazioni** (per sorteggio, mediante lancio di dadi o monete, usando la **Tabella dei numeri casuali**) fino al raggiungimento della **numerosità campionaria prefissata**

Campionamento sistematico

- $N = 1200$ $n = 60$
 ⇒ **passo di campionamento** = $1200/60 = 20$
- Lista unità da 1 a 1200
- Selezione casuale di un numero tra 1 e 20 (es. 8)
 ⇒ 1^a persona selezionata = 8^a nella lista
 ⇒ 2^a persona = $8 + 20 = 28^a$
 ecc.

Campionamento sistematico

1	Albert D.	25	Monique G.
2	Richard D.	26	Régine D.
3	Bella H.	27	Lucille L.
4	Raymond L.	28	Lucy B.
5	Stéphane B.	29	Michelle B.
6	Albert T.	30	Gilles D.
7	Jean William V.	31	Renaud S.
8	Clara D.	32	Pierre K.
9	Jeremy W.	33	Étienne M.
10	Anthony Q.	34	Marie M.
11	James B.	35	Gaston Z.
12	Denis G.	36	Fidèle D.
13	Amanda L.	37	Maria P.
14	Jennifer L.	38	Anne-Marie G.
15	Philippe K.	39	Michel K.
16	Eva F.	40	Gaston C.
17	Frisolia O.	41	Alan M.
18	Robert D.	42	Olivier P.
19	Brian F.	43	Geneviève M.
20	Hélène H.	44	Berthe D.
21	Isabelle R.	45	Jean Pierre P.
22	Jean T.	46	Jacques S.
23	Samantha D.	47	François R.
24	Berthe L.	48	Dominique M.
		49	Antoine S.

"Stratificato"

- Stratificato è un campione tratto da una popolazione stratificata
- Ogni strato è una popolazione in senso proprio e *il campione che si seleziona da uno strato è idoneo a rappresentarla*
- Il campione ottenuto dall'unione dei campioni dei singoli strati rappresenta l'intera popolazione

CAMPIONAMENTO A GRAPPOLO (CLUSTER)

Utile quando la popolazione di interesse si presenta **naturalmente suddivisa** in **sottogruppi (cluster)**, **eterogenei** (piccole miniature della popolazione)

- Viene estratto un campione casuale di cluster
- Vengono selezionate tutte le persone o un campione casuale di persone in ogni cluster

ES. **Pazienti ricoverati** in differenti reparti di un ospedale, oppure **pazienti assistiti** da diversi medici di Medicina generale

Vantaggioso = quando i grappoli costituiscono una **naturale aggregazione** delle unità finali di analisi, **per le quali non si possiede una lista**)

Efficiente = grappoli molto eterogenei al loro interno e piuttosto omogenei tra loro

Aspetti STATISTICI di un

Protocollo di RICERCA

1. Il quesito di ricerca (⇒ ipotesi di ricerca)
2. Il razionale della ricerca (rilevanza e novità)
3. Il disegno dello studio (in funzione dell'ipotesi)
4. **Popolazione di riferimento e Campione**
5. **I metodi di valutazione/misurazione delle variabili**
6. **La pianificazione dell'analisi statistica (in funzione dell'ipotesi e, implicitamente, del tipo di variabili coinvolte)**

5. I metodi di valutazione/misurazione delle variabili (a)

ovvero, delle caratteristiche/variabili dei pazienti che interessano per la ricerca:

- ❖ **strumenti meccanici/elettronici** utilizzati e loro caratteristiche (Ecografo, ECG, test di laboratorio, metro e bilancia,.....)
- ❖ **questionari**

PREMESSA: Per svolgere **attività di ricerca** seguendo il **metodo scientifico** è necessario **adottare metodi che garantiscano**

- **validità alle misure** dei fenomeni oggetto di studio
- possibilità di **comparare risultati ottenuti in contesti diversi e in modi diversi**

[TRATTO DA: Corbetta P.G., *La ricerca sociale: metodologia e tecniche: II. Le tecniche quantitative*. Ediz. Il Mulino, Bologna 2004]

Affidabilità e Validità di una "misura" rispetto ad un concetto

RELIABLE da **RELIABILITY = AFFIDABILITÀ**
 Grado in cui, in diverse condizioni di osservazione, si ottengono gli stessi risultati, sottoponendo a rilevazione gli stessi individui (unità di analisi)

VALID da **VALIDITY = VALIDITÀ**
 Grado in cui uno strumento di misura rileva ciò che si intende effettivamente misurare

EVERITT B. *Dizionario Cambridge di Statistica Medica*. Il Pensiero Scientifico Editore, ROMA 1998

TIPi di Variabili (Dati)

Qualitative o categoriali

- **Nominali Dicotomiche**
2 categorie, che indicano presenza/assenza di attributi [0/1; SI/NO; Binari; Vero/Falso]
- **Nominali Policotomiche**
più di 2 categorie
- **Ordinali**

Quantitative o numeriche

- **Discrete**
- **Continue**

VARIABILE QUANTITATIVA CONTINUA

- Si dice **continua** una variabile che può **virtualmente** assumere un qualsiasi valore reale, in un certo ambito
- **Esempio:** Achille può essere alto esattamente 1,7724538509... metri
- **In pratica**, tuttavia, le misure di una variabile quantitativa **possono assumere solo certi valori**, in relazione al potere di risoluzione dello strumento di misura
- **Esempio:** l'Altezza di Achille è 1.77 m, se misurata con un metro da sarto; è invece 1.772 m se misurata con lo stadio-metro Harpenden

GERARCHIA DELLE VARIABILI

Una **variabile quantitativa** può essere ridotta a e ancora a

Quantitative
 Quantitative/ordinali
 Qualitative nominali

Come si fa?

1) L'**Età** dei pazienti può essere espressa a **livello ordinale?**
 Bambini (0-10) /Adolescente (11-18) /Adulti (19-64)/Anziani (65+)

2a) L'**IMC** (*Indice di Massa Corporea, kg/m²*) può essere espressa a **livello ordinale?**
 Sottopeso (<18)/ Normo (18-25)/ Sovra (26-29)/ Obeso (≥30)

2b) e a **livello dicotomico?**
 Obeso (≥30) / Non Obeso (<30)

Indicare il tipo per ognuna delle seguenti variabili (1):

- | | |
|--|-----------------------|
| <input type="checkbox"/> Consumo giornaliero di sale | Quantitativa continua |
| <input type="checkbox"/> Valutazione della struttura corporea sulla base del BMI: normale, sovrappeso, obeso | Qualitativa ordinale |
| <input type="checkbox"/> Eventi avversi di un farmaco | Qualitativa nominale |
| <input type="checkbox"/> Gravità di una patologia in una scala da I° a V° grado | Qualitativa ordinale |
| <input type="checkbox"/> Gruppo sanguigno | Qualitativa nominale |
| <input type="checkbox"/> Quantità di lipidi in un alimento | Quantitativa continua |
| <input type="checkbox"/> Forma di un eritrocita | Qualitativa nominale |
| <input type="checkbox"/> N° di visite mediche richieste da un paziente in un anno | Quantitativa discreta |
| <input type="checkbox"/> N° di comorbidità | Quantitativa discreta |

RUOLO delle variabili nella formulazione del quesito/ipotesi

- Esposizione** / Indipendente
- Esito outcome** (endpoint)/ dipendente

In Ginecologia, c'è differenza nel numero di **complicazioni** legate all'anestesia tra ospedali che fanno riferimento ad **anestesisti-infermieri** (CRNA) rispetto a quelli che fanno riferimento ad **anestesisti-medici**?

In the United States, a Certified Registered Nurse Anesthetist (CRNA) is an advanced practice registered nurse (APRN) who has acquired graduate-level education and board certification in anesthesia.

Indagine conoscitiva sulle Caratteristiche degli studenti del CDL 1°anno - 2015/16

TIPO di Quesito/IPOTESI

(descrittive/di associazione/di differenza*)

- Come sono distribuiti in base al **Sesso**?
- Quale è la proporzione di quelli che seguono **corsi e-learning**?
- Essere in regola con gli esami del 1° anno** è **associato** all'**Avere/non avere una borsa di studio**? Ipotesi «plausibile»
- Il **Numero di esami registrati alla fine della prima sessione di esami del 1° anno** è **diverso** in base al **Tipo di diploma di scuola secondaria (Liceo scient /Liceo clas/ Liceo tecnol/Altro)**?
- Il Voto della prova scritta di Biologia è **diverso** fra chi ha la Maturità Scientifica e tutti gli altri?



Tre Tipi di Ipotesi «di base»

Generalmente, l'ipotesi consiste in una affermazione che riguarda la relazione tra 2 o più variabili.

I quesiti di ricerca sono uguali alle ipotesi, ma sono poste in forma interrogativa.

Distinguiamo tre tipi di quesito: *quesito di differenza*, *quesito di associazione* e *quesito descrittivo*.

- **Per i quesiti di differenza e di associazione**, la specifica «di base», sta a indicare che nella loro formulazione ci sono una variabile dipendente (esito, outcome) e una variabile indipendente (esposizione/variabile esplicativa/ fattore di rischio....determinante) **Più precisamente**, una variabile indipendente qualitativa e una variabile dipendente quantitativa nel caso di ipotesi di differenza, mentre nelle ipotesi di associazione, entrambe le variabili sono qualitative.
- **Per le ipotesi descrittive**, invece, «di base» significa che c'è una sola variabile

* Morgan GA, Harmon RJ. Research questions and hypotheses. J Am Acad Child Adolesc Psychiatry. 2000 Feb;39(2):261-3.

L'Analisi statistica

Analisi statistica DESCRITTIVA: consiste nell'uso di **tecniche** statistiche sviluppate per **sintetizzare e organizzare i dati**, sia di **popolazione** sia **campionari**



Analisi statistica INFERENZIALE: procedure per lo studio dei campioni che consentono di valutare la possibilità di **generalizzazione dei risultati** ottenuti dall'elaborazione dei **dati campionari** alle popolazioni da cui sono stati estratti



Analisi DESCRITTIVA - 1



Consiste nell'uso di **tecniche** statistiche sviluppate per **sintetizzare e organizzare i dati**, sia di **popolazione** sia **campionari**, attraverso:

- Misure di **centralità** (misure di **posizione**)
- Misure di **variabilità** (o **dispersione**)
- **Tabelle**
- **Grafici**



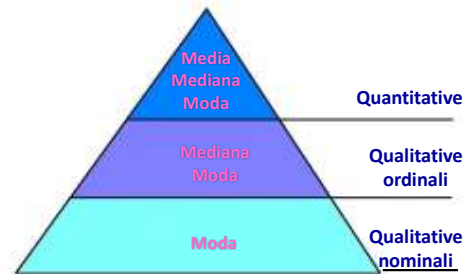
MISURE di SINTESI

La tendenza centrale



- ❖ è una misura statistica che consente di riassumere un insieme di dati in un solo numero
- ❖ è un tentativo di identificare gli aspetti "tipici", "medi" di una distribuzione

- **Misure di Tendenza centrale:** Media aritmetica, ... (m)
Mediana (Me)
Moda
- **Misure di Variabilità:**
Deviazione Standard (DS)
Range (min-max)
Differenza Interquartile (DI)



L'unica misura di sintesi statistica che possiamo calcolare per variabili qualitative nominali è la moda. Possono esistere variabili con più di una moda (al max tante mode quanti sono le classi, caso di equidistribuzione).

Quasi tutte le distribuzioni che si incontrano in statistica medica sono unimodali



MISURE di VARIABILITÀ

Deviazione Standard
Varianza
Intervallo interquartile
Coefficiente di Variazione

Il valore medio delle **deviazioni** elevate al quadrato si chiama **Varianza** e la sua formula è:

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

- Il numeratore è noto come «*somma dei quadrati degli scarti dalla media*»
- Si misura in unità quadratiche (Es. se x è l'altezza in cm la varianza sarà espressa in cm^2)
- Per questo conviene avere una misura espressa nelle unità originarie di x , e ciò si realizza calcolando la radice quadrata della varianza: questa grandezza è appunto nota come **Deviazione Standard**

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

INTERVALLO di variazione $R = x_{(1)} - x_{(n)}$
INTERVALLO interquartile: $RI = Q_1; Q_3$

La distanza tra il valore più piccolo e il valore più grande è indicata come **Campo di variazione (Range)**

$x_{(1)} \longleftarrow \text{Range} \longrightarrow x_{(n)}$

Se utilizziamo la **mediana** come indice del centro della distribuzione, dividendo la distribuzione in due parti, possiamo usare la stessa idea **per misurare la dispersione**

La distanza tra Q_1 e Q_3 è una misura di dispersione detta **Range interquartile**

$Q_1 \longleftarrow M_e \longrightarrow Q_3$

L'Analisi DESCRITTIVA - 1

Consiste nell'uso di **tecniche** statistiche sviluppate **per organizzare e sintetizzare i dati**, sia di popolazione sia campionari, attraverso:

- ✓ Misure di **centralità** e misure di **posizione**
- ✓ Misure di **variabilità** (o dispersione)

✓ **Tabelle**
 • **Grafici**



CODIFICA

Agecat = Classe di Età: 1, <25aa
 2, 25-34
 3, 35-44
 4, 45 o+

Gender = Genere: 1, Donna / 0, Uomo

Marital = Coniugato: 1, SI / 0, NO

Active = Adeguata Attività Fisica: 1, SI / 0, NO

Bfast = Breakfast: 1, Colazione al bar
 2, Colazione con Oatmeal
 3, Colazione con Cereali

Gender	Frequenza	%
Male	424	48,2
Female	456	51,8
Total	880	100,0

LifeStyle	Frequenza	%
Inactive	474	53,9
Active	406	46,1
Total	880	100,0

Marital status	Frequenza	%
Unmarried	303	34,4
Married	577	65,6
Total	880	100,0

Preferred breakfast	Frequenza	%
Breakfast Bar	231	26,2
Oatmeal	310	35,2
Cereal	339	38,5
Total	880	100,0

Organizzare: DISTRIBUZIONE di FREQUENZA

- Primo passo per descrivere i dati raccolti è la costruzione di una.....

"Tabella che indica il numero di osservazioni per ogni valore/modalità della variabile"

GRUPPO SANGUIGNO	Frequenza
A	32
B	8
AB	6
O	9
Totale	55

N Tot. popolazione
 n Tot. campione

Frequenze assolute, relative, percentuali

- ❖ Il conteggio del numero delle volte che una specifica modalità, oppure un singolo valore, della variabile compare in un determinato campione/popolazione, viene definita **frequenza assoluta**
- ❖ L'informazione derivabile da una frequenza è più completa quando la **frequenza assoluta viene messa in relazione alla totalità dei soggetti su cui la misurazione viene effettuata (n)**. In sostanza si pone in relazione una parte con il tutto ottenendo un dato relativo, **definito frequenza relativa (f/n)**
- ❖ Una indicazione ancora più efficace si può ottenere **moltiplicando per 100 la frequenza relativa, calcolando così la frequenza percentuale**, o più semplicemente **percentuale (%)**

"Gruppo sanguigno"
di un campione di n=55 donne con trombo-embolia

$$\frac{32}{55} = 0,58$$

Gruppo sanguigno	Frequenza assoluta	Frequenza relativa	%
A	32	0,58	58
B	8	0,15	15
AB	6	0,11	11
O	9	0,16	16
Totale	55	1	100

Nota: una frequenza relativa è il rapporto tra la frequenza assoluta con cui si manifesta una modalità/valore e la numerosità totale del campione/popolazione

Proporzioni & percentuali

Valore	f	p=(f/N)	%=p(100)
25	3	0,3	30,00%
26	2	0,2	20,00%
27	2	0,2	20,00%
28	0	0	0,00%
29	2	0,2	20,00%
30	1	0,1	10,00%

$$p = \frac{f_i}{n}$$

$$\% = p \times 100$$

L'Analisi DESCRITTIVA II°



Consiste nell'uso di **tecniche** statistiche svil **sintetizzare e organizzare** i dati, sia di popolazione sia campionari, attraverso:

- Misure di **centralità** (misure di **posizione**)
- Misure di **variabilità** (o **dispersione**)

- Tabelle**
- Grafici**

Per variabili quantitative continue (1)

❖ conviene **raggruppare le osservazioni**, suddividendo l'intervallo di variabilità, in **classi**

Variabile: Volume Espiratorio Forzato al secondo (FEV1)

Dati di FEV1 in un campione di n=57 studenti maschi di Medicina

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

Per variabili quantitative continue (2)

❖ è preferibile che le classi abbiano un **intervallo costante** (ad es. soggetti con età tra 36 e 45; 46 e 55; 56 e 65 anni)

❖ le classi devono essere **mutualmente esclusive**, in modo che ciascuna misura sia riferibile senza ombra di dubbio a una singola classe (ad es. **evitare classi quali 50-55, 55-60, 60-65 anni, ecc**)

VARIABILE QUANTITATIVA CONTINUA

Perché le classi non si sovrappongano bisogna decidere quale, tra due intervalli contigui, debba contenere il valore soglia, al fine di evitare conteggi duplici

PER CONVENZIONE, si è soliti **includere l'estremo inferiore nell'intervallo**, e attribuire invece l'estremo superiore all'intervallo successivo

FEV1	Frequenza	Frequenza relativa %
2.0 - 2.49	0	0.0
2.5 - 2.99	3	5.3
3.0 - 3.49	9	15.8
3.5 -	14	24.6
4.0 -	15	26.3
4.5 -	10	17.5
5.0 -	6	10.5
5.5 -	0	0
Totale	57	100.0

2.0-2.5; 2.5-3.0; 3.0-3.5;???????

L'Analisi DESCRITTIVA



Consiste nell'uso di **tecniche** statistiche sviluppate per **organizzare e sintetizzare i dati**, sia di popolazione sia campionari, attraverso:

- ✓ Tabelle
- ✓ Misure di centralità e misure di posizione
- ✓ Misure di variabilità (o dispersione)
- ✓ **Grafici**

- Torta (Diagramma a settori circolari)
- Diagramma a barre
- Grafico Box-Whisker
-
- ❖ Istogramma (Diagramma a canne d'organo)
-
- Grafico a punti
- *Diagramma a dispersione (*Scatter plot*)

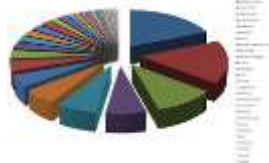


Diagrammi a Torta

In un diagramma a torta le **frequenze relative** vengono rappresentate dividendo un cerchio in settori, in modo che ogni settore sottenda un **angolo proporzionale** alla frequenza relativa della categoria corrispondente:

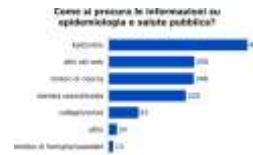
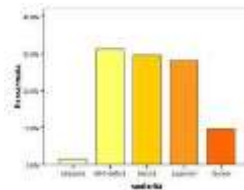


Le **torte** sono utilizzabili preferibilmente se il numero delle categorie è limitato



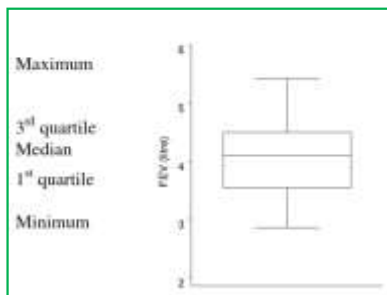
Diagrammi a barre (o a nastri)

- Servono per rappresentare le distribuzioni di frequenze di variabili qualitative
- "Forme" rettangolari **DISTANZIATE**



Box-Whisker:

Nei grafici a **Scatole e Baffi** vengono visualizzati la **mediana**, la **distanza inter-quartilica** e i **valori anomali ed estremi** di una variabile quantitativa



Br Med J (Clin Res Ed). 7 Jul 25;295(6592):231-4

Milson S, Ibbertson K, Hannan S, Shaw D, Pybus J. (1987). *Simple test of intestinal calcium absorption.*

Rappresentazione di due variabili: una quantitativa (**concentrazione di stronzio nel liquido extracellulare**) e una qualitativa (condizione patologica/non)

Riporta i **valori dei singoli** individui e le **differenze** tra gruppi sono facilmente apprezzabili

Dati anomali facilmente individuabili

Adatto a pochi dati

Grafico a punti

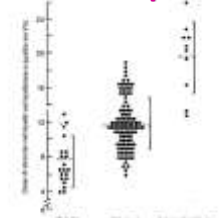


Figura 11. Concentrazioni di stronzio in plasma con la radiocinetica, negli affetti da malattia di Crohn (gruppi) e nei controlli (gruppo). (Milson et al., 1987)

Particolarmente **utili per misure ripetute** in più di una **occasione**: in questo caso i **dati appaiati** sono collegati da un **segmento**

ESEMPIO da letteratura:
 Gli autori **mettono in relazione** la **Velocità di filtrazione glomerulare** in **n=7 diabetici insulinodipendenti nutriti con Dieta normale o ipoproteica**

Figura: mostra livelli di poco, ma significativamente inferiori in tutti i pazienti durante la dieta ipoproteica, evidenziata grazie ai collegamenti individuali

Grafico a punti

Figure 4.7. Velocità di filtrazione glomerulare (ml/min per 1,73 m²) in 7 pazienti diabetici insulinodipendenti nutriti con Dieta normale o ipoproteica (da Collier et al., 1987)

Diagramma a dispersione

- Rappresentazione grafica di due variabili quantitative
- Ciascun punto rappresenta un **sogetto***

Figure 1. Scatter diagram showing muscle strength and height for 41 male alcoholics

- Torta (Diagramma a settori circolari)
- Diagramma a barre
- ❖ Grafico a punti
- ❖ Diagramma a dispersione (*Scatter plot*)
- Grafico Box-Whisker

- **Istogramma** (Diagramma a canne d'organo)

Istogramma

Peso (kg)	freq assoluta	freq relativa	freq cumulata
40-49,9	2	0,09	0,09
50-59,9	5	0,23	0,32
60-69,9	8	0,36	0,68
70-79,9	6	0,27	0,95
80-89,9	0	0,00	0,95
90+	1	0,05	1,00
Tot.	22	1,00	

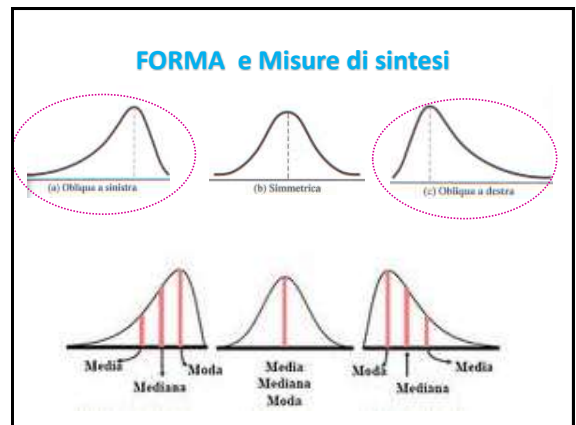
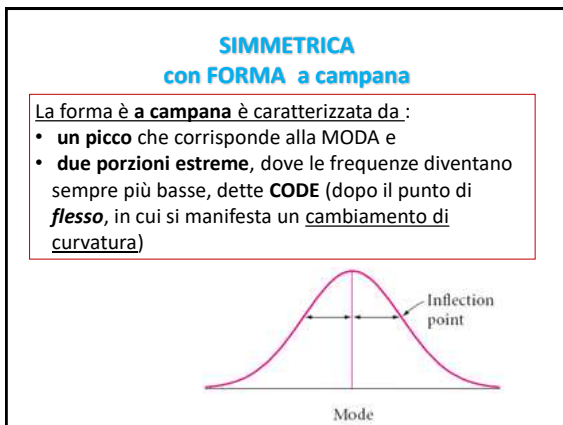
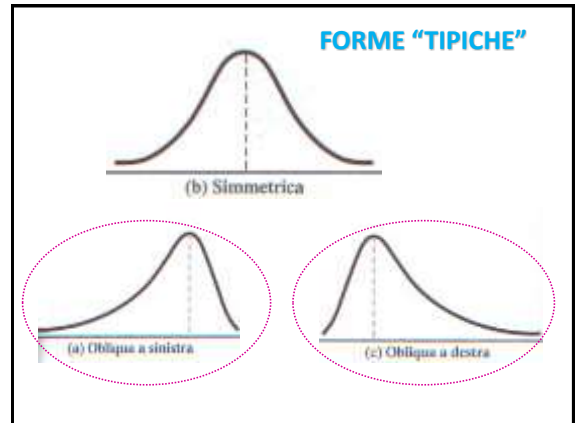
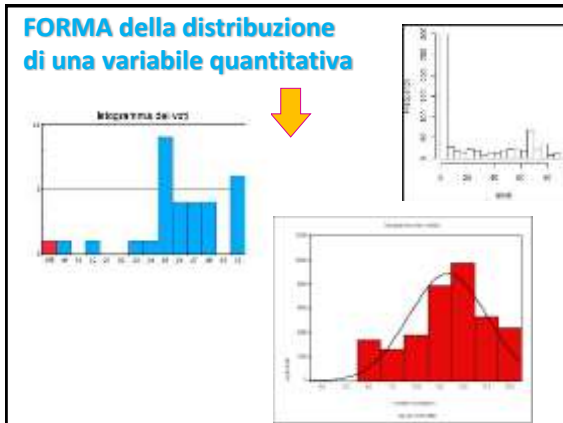
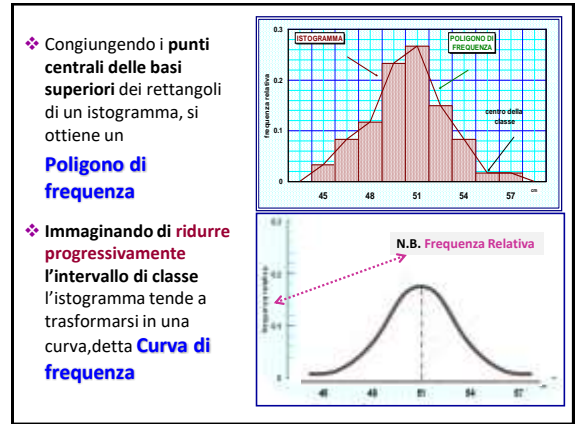
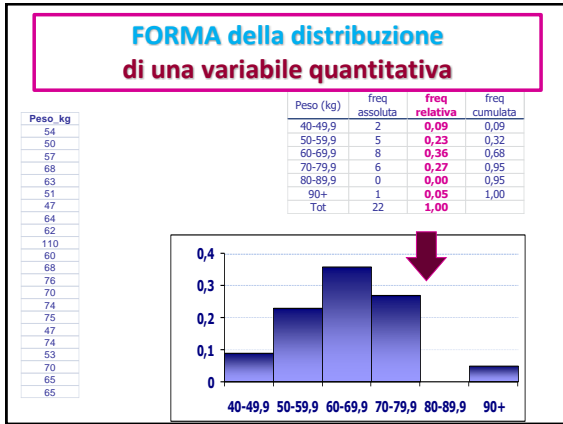
Le distribuzioni di frequenza di variabili quantitative continue possono essere rappresentate visivamente tramite ISTOGRAMMI

- **Asse X:** estremi degli intervalli che rappresentano **le classi** di suddivisione dei dati
- **Asse Y:** **sulle basi** rappresentate dalle classi si costruiscono i **rettangoli**, **le cui altezze risultano uguali alladella corrispondente alla classe**

- Per classi di ampiezza UGUALE $h = \text{frequenza}$
- Per classi di ampiezza DIVERSA $\text{Area} = b \times h = \text{frequenza}$

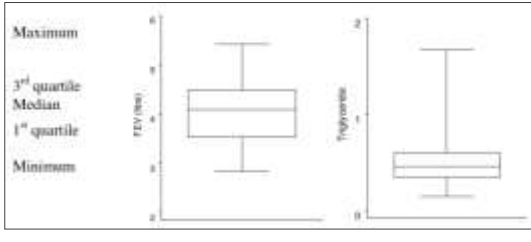
DENSITÀ DI FREQUENZA

$$h = \frac{\text{frequenza}}{b}$$



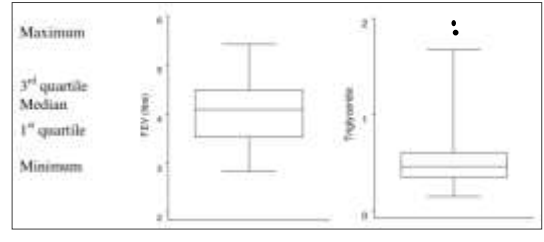
Box-Whisker:

- Il grafico della distribuzione del FEV mostra una distribuzione **simmetrica**, dove i **baffi sono di lunghezza simile**
- Il grafico del **Siero Trigliceride** (dal cordone ombelicale) presenta invece una **asimmetria positiva**, evidenziata dal **baffo superiore molto più lungo** di quello inferiore



Box-Whisker:

- **NOTA:** un dato la cui distanza dai lati orizzontali della scatola (ossia i quartili) sia maggiore di una volta e mezzo l'altezza della scatola (ossia il Range interquartile) può essere definito «**outlier**» - possono essere rappresentati per mezzo di punti isolati



Dal campione alla popolazione

Analisi statistica INFERENZIALE: procedure per lo studio dei dati **campionari** che consentono di valutare la possibilità di **generalizzazione dei risultati** descrittivi, ottenuti dall'elaborazione dei dati campionari, alle popolazioni da cui sono stati estratti



TEST per la VERIFICA delle IPOTESI sui parametri della popolazione

rappresentano la **PROCEDURA** che utilizziamo **per decidere**

I test statistici più utilizzati in letteratura



- **Test t-Student:**
- **ANOVA (ANalysis Of VAriance)**
- **Test Chi-quadrato (χ^2)**
- **Test di Mann-Whitney**
-

Indagine conoscitiva sulle Caratteristiche degli studenti di xxxxxxxx 2° anno-2011/12

Ho bisogno di sapere:

- Essere in regola con gli esami del 1° anno è associato all'Avere/non avere una borsa di studio?
- Il Numero di esami registrati alla fine della prima sessione di esami del 2° anno è diverso in base al Tipo di diploma di scuola secondaria (Liceo scient /Liceo clas/ Liceo tecnol/Altro)
- Il Voto della prova di Metodi Statistici è diverso fra chi ha la Maturità Scientifica e tutti gli altri

• Ricercatore progetta una RICERCA

- Ha in mente una teoria/ipotesi..... dimostrare

Esiste una differenza tra il voto medio della prova di Metodi Statistici di chi ha la Maturità Scientifica e il voto medio di chi non ha la Maturità scientifica

IPOTESI di studio/ricerca

STEP 1. Si formula l'ipotesi al negativo:

NON esiste una differenza tra il **voto medio della prova di Metodi Statistici** di **chi ha la Maturità Scientifica** e il **voto medio** di **chi non ha la Maturità scientifica**

L'Ipotesi nulla (H_0)

L'Ipotesi nulla (H_0)

"nulla"

perché espressa al negativo

$$H_0 : \bar{x}_{LiceoScientifico} = \bar{x}_{AltraScuola}$$

H_A

è l'ipotesi alternativa all'ipotesi nulla e coincide con l'ipotesi di ricerca

$$H_0 : \bar{x}_{LS} = \bar{x}_{noLS}$$

$$H_A : \bar{x}_{LS} \neq \bar{x}_{NoLS}$$

2 Ipotesi contrapposte: H_0 e H_A

H_0 o H_A ????

STEP 2.

Si sceglie il test in funzione al tipo di variabile....e del tipo di ipotesi

- **Test t-Student** per la verifica dell'ipotesi riferite alla media di uno o due gruppi
- **ANOVA** (medie di tre o più gruppi): *a una via oppure a due vie*
- **Test Chi-quadrato** per la verifica dell'ipotesi di associazione fra 2 variabili qualitative*

Test t-Student



Dati INDIPENDENTI

Es. serie di dati ottenuti dalla misurazione dell'altezza in un gruppo di bambini ricoverati in un ospedale

Dati DIPENDENTI

Es. serie di dati ottenuti dalla misurazione dell'altezza sullo stesso campione di bambini, in un periodo di 5 anni (monitoraggio)

N.B. Dati DIPENDENTI casi particolari

Es. serie di dati ottenuti dalla misurazione dell'altezza sullo stesso campione di bambini, in un periodo di 5 anni (monitoraggio)

Il test si basa sul calcolo della probabilità (valore p) che si avrebbe di osservare, per puro caso*, i nostri risultati campionari se fosse vera l'ipotesi nulla

QUANDO la probabilità associata a un certo risultato del test diventa così bassa* da ritenersi 'poco probabile' che sia un effetto solo del caso:
RIFIUTIAMO L'IPOTESI NULLA!!!

***QUANDO $p < 0,05$**

Assunzioni per tecniche/test parametrici

Test non parametrici

Si è detto sopra che, per le variabili continue, si parte dal presupposto di una sottostante distribuzione normale. Se questo non è il caso, si può passare a procedure alternative che non richiedono la normalità dei dati. In questi casi, il t -test per dati appaiati è sostituito dal test di Wilcoxon, e dal test per dati non appaiati di Wilcoxon descritto alternativamente anche come test U di Mann-Whitney.

Questi test sono spesso chiamati indipendenti dalla distribuzione o non parametrici.

Scegliere il Test per la verifica delle ipotesi (2)

	TEST PARAMETRICI		TEST NON PARAMETRICI
QUANTITATIVE	frequenza cardiaca, pressione arteriosa	Variabili	dolore, Gilgoug (nota scorse)
	Test t per dati non appaiati	2 campioni indipendenti	Test di Wilcoxon-Mann-Whitney
	ANOVA a 1 criterio	K campioni indipendenti	Test di Kruskal-Wallis
	Test t per dati appaiati	2 campioni dipendenti	Test di Wilcoxon
	ANOVA per misure ripetute	K campioni dipendenti	Test di Friedman
	Correlazione e regressione	Associazione tra due variabili	Coeff. di correlazione ordinale di Spearman
			Equivalente non-parametrico del Coeff. di correlazione di Pearson

Scegliere il Test per la verifica delle ipotesi (3)
APPROSSIMAZIONE dei dati alla Distribuzione Normale

Tests of Normality^a

Accurate	Kolmogorov-Smirnov ^b			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Interval No	.180	119	.000	.901	119	.000
Yes	.188	150	.000	.861	150	.000

a. Lilliefors Significance Correction
 b. There are no valid cases for Interval when Accurate = .000. Statistics cannot be computed for this level.

$H_0 : \approx \text{dati} \approx N$
 $H_A : \approx \text{dati} \neq N$

Livelli di Significatività e Tipi di errore. RIEPILOGO (2)

Quanto «piccolo» è piccolo?

- P= 0,05.** Significa che, se l'ipotesi nulla È VERA, noi potremo prendere una decisione sbagliata 1 volta su 20
- Supponiamo di aver eseguito un test per la verifica delle ipotesi e di aver ottenuto una probabilità $p < 0,01$ che rappresenta una prova ragionevole contro l'ipotesi nulla. SE l'ipotesi nulla È VERA, rifiutandola noi potremmo aver preso la decisione sbagliata meno di 1 volta su 100.

STEP 4. DECISIONE e conclusioni H_0 o H_A ???

QUANDO $p < 0,05$

Rifiutiamo H_0 e Accettiamo H_A
 Possiamo generalizzare i risultati osservati nel campione affermando che:
 Esiste (non solo nel campione!) una differenza/associazione *statisticamente significativa*
INCERTEZZA = $p < 0.05$

QUANDO $p \geq 0,05$

NON possiamo rifiutare H_0
 La nostra conoscenza rimane limitata al campione
 OVVERO:
 Non possiamo trarre conclusioni relativamente alla popolazione

Non raggiungere la significatività statistica ($p > 0,05$) NON rifiuto l'ipotesi nulla

non è l'equivalente di "effetto assente" ma consente solo di concludere che lo studio non ha permesso di mettere in evidenza alcuna associazione/differenza

- Questa è una particolare funzione della **STATISTICA**:
 permettere al ricercatore di fare una **valutazione numerica dell'incertezza delle sue conclusioni**

Snedecor (1950)

N.B. Differenza tra:

Misure di Variabilità delle stime campionarie

SE= Standard Error

Variabilità dei dati

DS= Deviazione Standard

Concetto di Standard Error (SE*)
 Variabilità delle STIME

ES. **L'errore standard DELLA MEDIA** è **indice dell'Imprecisione** con la quale la media campionaria **stima** la media vera della popolazione

Si determina dividendo la deviazione standard dei dati campionari (σ) per la numerosità campionaria

$$\frac{\sigma}{\sqrt{n}}$$

N.B. se **n** aumenta: **SE** diminuisce: **•** aumenta la precisione

* Indicato anche con SEM = Standard Error della Media

Formula per il calcolo dell'Intervallo di confidenza della **MEDIA***
 se il campione è grande, $n > 100$

...95%

Stima - 1.96 x SE(stima) < parametro < Stima + 1.96 x SE(stima)

...90%

Stima - 1.64 x SE(stima) < parametro < Stima + 1.64 x SE(stima)

...99%

Stima - 2,58 x SE(stima) < parametro < Stima + 2,58 x SE(stima)

Non solo SE della media campionaria.....

$$\frac{\sigma}{\sqrt{n}}$$

- Media
- Proporzione (Es. Prevalenza di malattia)
- Rischio Relativo
- OR,
-

Intervalli di confidenza al 95% ... di una proporzione p

- La proporzione p può essere un rischio, una sensibilità, una specificità, ...

Esempio: valutazione della stima della sensibilità di uno strumento diagnostico (stima campionaria puntuale, $n=17$ sensibilità=64,7%)

$$p - 1,96 \times \sqrt{\frac{p \times (1-p)}{n}} < \pi < p + 1,96 \times \sqrt{\frac{p \times (1-p)}{n}}$$

$$0,647 - 1,96 \times \sqrt{\frac{0,647 \times 0,353}{17}} = 0,42$$

$$0,647 + 1,96 \times \sqrt{\frac{0,647 \times 0,353}{17}} = 0,874$$

IC 95% : 42% < Valore vero della sensibilità < 87.4%

Seconda parte



Significatività statistica e clinica

Relazione dei **valori-p** con gli **intervalli di confidenza**

APPROFONDIMENTO:



Bollettino di informazione sui farmaci:

<http://www.agenziafarmaco.gov.it/content/bollettino-di-informazione-sui-farmaci>

BIF Nov-Dic 2001 - N. 6, pp.257-260

Come viene riportata l'importanza clinica dei risultati degli studi?

La **differenza minima clinicamente importante (MCID)**, dall'inglese *minimal clinically important difference* fra una terapia sperimentale e la rispettiva terapia di controllo è **definita** come «*quella differenza che sarebbe sufficiente a giustificare un cambiamento nelle decisioni terapeutiche, tenendo anche conto del rischio di eventi avversi, inconvenienti e costi*».

La MCID è un **concetto chiave** sia nel disegno che nell'interpretazione dei risultati degli RCT.

Nel disegno, infatti, la numerosità del campione (sample size) del trial dovrebbe riflettere la MCID che si vuole evidenziare fra il trattamento in sperimentazione e quello di controllo (**valore delta**).

Nell'interpretazione, il raggiungimento o meno di una MCID è il criterio principale di cui tener conto per la prospettiva di applicare i risultati del trial alle decisioni terapeutiche

Uso scorretto delle P ...

... ottenute in due studi diversi per confrontare l'efficacia dei trattamenti

Nel primo studio l'efficacia è > che nel secondo ???

STUDIO – Farmaco A/placebo

$n = 2000$ pz

End-point = guarigione

Trattati = 25% guarigioni

Controlli = 20% guarigioni

$P < 0,001$

STUDIO – Farmaco B/placebo

$n = 100$ pz

End-point = guarigione

Trattati = 35% guarigioni

Controlli = 20% guarigioni

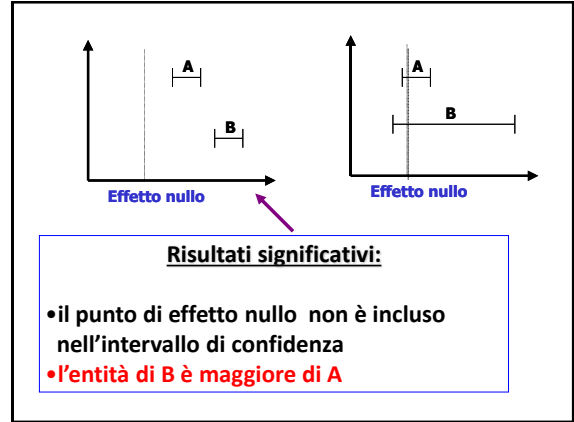
$P = 0,07$

Farmaco A = siamo ragionevolmente sicuri che sia efficace e che questa efficacia sia modesta

Farmaco B = non siamo sicuri che sia efficace ma questa efficacia potrebbe essere notevole

P + Intervallo confidenza (2)

- L'IC consente anche di stimare quanto realmente è grande un effetto:
maggiore lontananza dal punto dell'effetto nullo = maggiore entità dell'effetto



Analisi delle contingenze il Test Chi-quadrato

	Tosse	No tosse	TOTALE
Non fumatori	248 (20,4%)	1077 (79,6%)	1325 (100%)
Occasionali	388 (26,4%)	1077 (73,6%)	1465 (100%)
Regolari	80 (46,0%)	92 (53,9%)	172 (100%)

Applicazione al nostro database sugli STILI DI VITA

agecat	gender	marital	active	lifest
1	0	1	1	3
2	0	0	1	1
3	0	1	0	2
4	1	1	1	2
5	0	1	0	2
6	0	1	0	3
7	1	1	0	1
8	1	0	0	2
9	1	1	1	2
10	1	1	0	1
11	0	1	0	3
12	1	1	0	1
13	1	1	0	1
14	1	1	1	2
15	1	1	1	2

Distribuzioni di frequenza UNI-VARIABILE

Gender	Frequenza	%
Male	424	48,2
Female	456	51,8
Total	880	100,0

LifeStyle	Frequenza		%
	Frequenza	%	
Inactive	474	53,9	
Active	406	46,1	
Total	880	100,0	

Marital status	Frequenza	%
Unmarried	303	34,4
Married	577	65,6
Total	880	100,0

Preferred breakfast	Frequenza		%
	Frequenza	%	
Breakfast Bar	231	26,2	
Oatmeal	310	35,2	
Cereal	339	38,5	
Total	880	100,0	

Distribuzioni di frequenza BI-VARIABILE

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male	104	155	165	424
		24,5%	36,6%	38,9%	100,0%
	Female	127	155	174	456
		27,9%	34,0%	38,2%	100,0%
Totale		231	310	339	880

- **Organizzazione** in forma tabellare dei dati riferiti all'osservazione di 2 variabili contemporaneamente
- **Impatto visivo** che rende più evidente la **RELAZIONE** tra 2 variabili qualitative

Esempio: Genere & Tipo di colazione???

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male				
	Female				
Totale		231	310	339	880

Gender	Frequenza
Male	404
Female	456
Total	880

Preferred breakfast	Frequenza
Breakfast Bar	231
Oatmeal	310
Cereal	339
Total	880

Distribuzioni di frequenza BI-VARIABILE

- Tabella di contingenza
- Tabella a doppia entrata
- **Tabella r x c** – Esempio: Tabella **2 x 3** ?

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male	104	155	165	424
		24,5%	36,6%	38,9%	100,0%
Female	Female	127	155	174	456
		27,9%	34,0%	38,2%	100,0%
Totale		231	310	339	880

QUESITO: L'abitudine scorretta di «Fare colazione al bar» è più frequente tra i maschi o tra le femmine?

RISPOSTA dell' Analisi Descrittiva:

SI
 NO

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male	104	155	165	424
		24,5%	36,6%	38,9%	100,0%
Female	Female	127	155	174	456
		27,9%	34,0%	38,2%	100,0%
Totale		231	310	339	880

NI SI, nel nostro campione!

- Esiste (nella nostra popolazione di riferimento) un'associazione tra **Genere** e **Tipo di colazione** ?
- L'associazione tra **Genere** e **Tipo di colazione** è statisticamente significativa?

RISPOSTA dell'Analisi INFERENZIALE:

SI
 NO

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male	104	155	165	424
		24,5%	36,6%	38,9%	100,0%
Female	Female	127	155	174	456
		27,9%	34,0%	38,2%	100,0%
Totale		231	310	339	880

Applicazione del test chi² ad una tabella di contingenza. Esempio 1

		Preferred breakfast			Totale
		Breakfast Bar	Oatmeal	Cereal	
Gender	Male	104	155	165	424
		24,5%	36,6%	38,9%	100,0%
Female	Female	127	155	174	456
		27,9%	34,0%	38,2%	100,0%
Totale		231	310	339	880

Chi-quadrato			
	Valore	df	Sig. stat. (2 vie)
Chi-quadrato	1,367*	2	,505
N. di casi validi	880		

* 0 celle (.0%) hanno un conteggio atteso inferiore a 5. Il conteggio atteso minimo è 111,30

persone guarite nei due gruppi

Trattamento	Guariti	Non guariti	Totale
Tipo A:	23 (a)	7 (c)	30 (m)
Tipo B:	18 (b)	13 (d)	31 (n)
Totale	41 (r)	20 (s)	61 (N)

VALORI ATTESI??? Frequenza attesa di cella = (Totale Riga * Totale Colonna)/n

persone guarite nei due gruppi			
Trattamento	Guariti	Non guariti	Totale
Tipo A:	(30x41)/61 (a)	(30x20)/61 (c)	30 (m)
Tipo B:	(41x31)/61 (b)	(31x20)/61 (d)	31 (n)

N.B. χ^2 per tabelle 2x2

- **Correzione (per la continuità) di Yates**
- **Test esatto di Fisher**

Per la lettura dei risultati del Chi-quadrato, la regola generale dice che: **il test esatto di Fisher deve essere utilizzato se qualunque dei valori attesi è meno di 5**

NOTA IMPORTANTE

Non disponiamo di una procedura alternativa (tipo Fischer per 2x2) per tabelle di contingenza con più di 2 righe e colonne e numeri piccoli [CAMPBELL 2005 vers. italiana, p.105]

TUTTAVIA: Il test tradizionalmente applicato per la validità dell'approssimazione del chi-quadrato è che almeno l'80% dei valori attesi sia maggiore di 5 e che tutti siano maggiori di 1 [ARMITAGE e BERRY, 1994]

Scegliere il Test per la verifica delle ipotesi (3) - ESEMPIO

Associazione tra due variabili QUALITATIVE, ordinale e nominale	Test Chi-quadrato per il trend
Associazione tra due variabili QUALITATIVE nominali	Test Chi-quadrato

*EXP ORDINALE con più di due categorie

	Tosse	No tosse	TOTALE
Non fumatori	266 (20.4%)	1037 (79.6%)	1303 (100%)
Occasionali	395 (28.8%)	977 (71.2%)	1372 (100%)
Regolari	80 (46.5%)	92 (53.5%)	172 (100%)

p < 0,0001

↓

Misura di concordanza (agreement)*

K di Cohen

Es. associazione = misura del legame tra due variabili
Concordanza = il grado in cui i valori di due set di dati concordano

Concordanza : grado di **accordo** che si osserva fra due test clinici, oppure fra due (o più) operatori che interpretano l'esito di *uno stesso test* (es. radiografie, elettrocardiogramma, auscultazione cardiaca ecc), oppure fra due letture effettuate da uno stesso operatore **in tempi diversi**.

N.B.
 Non si vuole stabilire quale classificazione sia più corretta, bensì stabilire se i criteri utilizzati per l'interpretazione del test siano efficienti, e se la **classificazione sia riproducibile**

K-di Cohen

Forza della concordanza- Range: 0-1

- <=0.20 scarsa
- 0.21-0.40** indifferente
- 0.41-0.60** moderata
- 0.61-0.80** buona
- 0.81-1.00** molto buona

Quaderno di epidemiologia veterinaria
www.quadernodiepidemiologia.it
 Foglio di lavoro complementare all'unità didattica:
http://www.quadernodiepidemiologia.it/screen/ind_con.htm
 Prof. Ezio Bottarelli

	+	-	
+	10	11	21
-	8	59	67
	18	70	88

Concordanza complessiva: 0,158
 Rapporto di Cohen: 0,102

5) Interpretazione di un test diagnostico

Legge: concordanza

3 D: nulla

3 D: -C: scarsa

3 D: -C: C: moderata

3 D: -C: C: buona

3 D: -C: C: molto buona

Variabili e Dati...particolari

Statistics Notes

Time to event (survival) data

BMJ VOLUME 317 15 AUGUST 1998 www.bmj.com

In many medical studies an outcome of interest is the time to an event. Such events may be **death** or recurrence of a tumour, **postoperative** such as operation or discharge from hospital, or **recovery** such as cessation of breast feeding. It is conventional to talk about survival data and survival analysis, regardless of the nature of the event. Similar data also arise when measuring the time to complete a task, such as walking 100 metres.

La caratteristica che contraddistingue i **dati di sopravvivenza** è che alla fine del periodo di **Follow-up**, è probabile che **l'evento non si sia verificato per tutti i pazienti**.

Per questi pazienti si dice che: **il tempo di sopravvivenza è CENSURATO**, ovvero che, per alcuni pazienti, il periodo di osservazione è terminato prima che l'evento si verificasse.

Non possiamo sapere **se e quando** l'evento si verificherà per quel paziente.

ES: Tempo di sopravvivenza a due anni dalla diagnosi di cancro al polmone (Dati: 11 mesi, 11, 13,, 20, 22, 24, 24, censurato (> 24, ancora vivo dopo 2 anni),

Tipo di studio epidemiologico?

Time to event

Survival analysis

Si applica ai dati di studi longitudinali di coorte

Criticità:

- non tutti i soggetti reclutati potrebbero essere entrati nello studio allo stesso tempo (=tempi di osservazione diversi)
- Alcuni soggetti possono decidere di uscire dallo studio prima della conclusione, oppure essere persi al follow-up* (=censurati)

* Continuazione della valutazione dei pazienti anche dopo la fine di un trattamento

Survival ANALYSIS

In entrambi i casi:

- tempi di osservazione diversi
- dati censurati

risulta difficile «sintetizzare» i dati , per determinare il tempo medio di sopravvivenza

Tempo MEDIANO di sopravvivenza

Variabili e Dati...particolari

- Life table*
- Survival curve
- Kaplan-Meier curve
- Logrank test
- Cox regression model*

Statistics Notes

Time to event (survival) data
Douglas G. Altman, J Martin Flaherty

In many medical studies an outcome of interest is the time to an event. Such events may be adverse, such as death or occurrence of a tumour, positive, such as recovery, or discharge from hospital, or neutral, such as cessation of breast feeding. It is conventional to talk about survival data and survival analysis, regardless of the nature of the event. Survival data also arise when measuring the time to complete a task, such as walking 50 metres.

The distinguishing feature of survival data is that at the end of the follow-up period the event will probably not have occurred for all patients. For these patients the survival data is said to be censored, indicating that the observation period was cut off before the event occurred. We do not know when (or, indeed, whether) the patient will experience the event, only that he or she has not done so by the end of the observational period.

* L'analisi multivariabile è uno strumento statistico che si utilizza per determinare il singolo contributo di vari fattori alla determinazione di un outcome.

Variabili e Dati...particolari

HR= Hazard Ratio:

Esprime il rapporto tra i rischi nell'analisi di sopravvivenza. In pratica esprime , in un periodo di follow-up definito, il rapporto tra i rischi di un evento (morbilità* oppure mortalità) in un gruppo di soggetti trattati , rispetto ad un gruppo di non trattati, **per l'intera durata dello studio**
Se HR>1, indica aumento del rischio nel gruppo sperimentale
Se HR<1, indica diminuzione del rischio nel gruppo sperimentale
Se HR=1 = indica uguale rischio nei due gruppi di trattamento

*MORBILITA'= Rapporto esistente tra il numero di malati e la popolazione

RELAZIONE tra 2 variabili.....



Sui termini: relazione o.....

- associazione o correlazione o covarianza
- Termini analoghi nel loro significato sostanziale
- Quando osserviamo che 2 variabili presentano VARIAZIONE CONCOMITANTI



- 2 variabili..... qualitative = **Associazione**
- 2 variabili..... quantitative= **Correlazione**

Diagramma a dispersione

- Rappresentazione grafica di due variabili quantitative
 - Ciascun punto rappresenta un **soggetto***

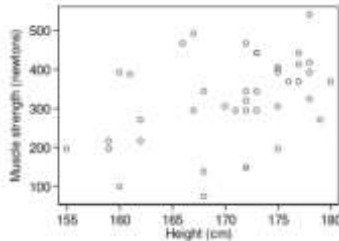
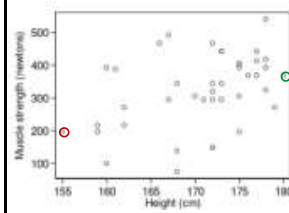


Figure 1. Scatter diagram showing muscle strength and height for 41 male alcoholics

Diagramma a dispersione

- Rappresentazione grafica di due variabili quantitative
 - Ciascun punto rappresenta un **soggetto***

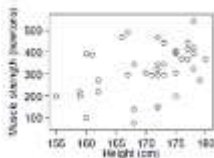


Nella figura è facile vedere che gli uomini più alti **tendono** ad essere più forti degli uomini più bassi

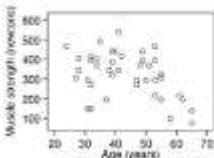
Figure 1. Scatter diagram showing muscle strength and height for 41 male alcoholics

Coefficiente di Correlazione lineare , r

Valore massimo $r = 1,00$
Valore minimo $r = -1,00$



Correlazione tra
«**Altezza e Forza muscolare**» = +0,42
Correlazione **positiva** piuttosto bassa



Correlazione tra
«**Anni di età e Forza muscolare**» = -0,42
Correlazione **negativa** piuttosto bassa

Coefficiente di Correlazione lineare , r

ci consente di misurare la FORZA della relazione tra **variabili quantitative continue di cui almeno una distribuita normalmente (r di Pearson)**

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

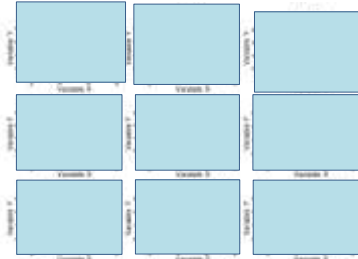


Figure 5. Simulated data from populations with different relationships between the two variables, and the population correlation coefficient

Coefficiente di Correlazione lineare, r di Spearman

ci consente di misurare la FORZA della relazione tra **variabili quantitative continue con distribuzione NON Normale**

Generalizzabilità di un *coefficiente di correlazione lineare* campionario

- **Regola pratica:** in mancanza del relativo intervallo di confidenza o del p-value:

Se $r > 2/\sqrt{n}$

possiamo considerare questo risultato **statisticamente significativo***

*N.B. Per avere un risultato VALIDO del test di significatività: ALMENO UNA delle due variabili quantitative deve seguire una Distribuzione Normale

Mentre, per avere un Intervallo di Confidenza VALIDO, è necessario che ENTRAMBE le variabili quantitative seguano una Distribuzione Normale

Generalizzabilità di un *coefficiente di correlazione lineare* campionario

- **Regola pratica:** in mancanza del relativo intervallo di confidenza o del p-value:

$$\text{Se } r > 2/\sqrt{n}$$

possiamo considerare questo risultato **statisticamente significativo***

*N.B. Per avere un risultato VALIDO del test di significatività: ALMENO UNA delle due variabili quantitative deve seguire una Distribuzione Normale

Mentre, per avere un Intervallo di Confidenza VALIDO, è necessario che ENTRAMBE le variabili quantitative seguano una Distribuzione Normale

Introduzione ai MODELLI di Analisi Multivariabile

- Relazione tra una variabile **dipendente** e **più** variabili **indipendenti**
- **Obiettivo:** determinare quali variabili indipendenti, separatamente e insieme, contribuiscono a **predire** l'outcome

Analisi Multivariabile

ESEMPIO:

Sappiamo che esistono numerosi fattori (variabili) associate con lo sviluppo delle malattie cardiache, p.es. il fumo, l'obesità, lo stile di vita sedentario, diabete, valori elevati di colesterolo, l'ipertensione.

L'analisi multivariabile è uno strumento statistico che *si utilizza per determinare il singolo contributo di vari fattori alla determinazione di un outcome.*

L'analisi multivariabile ci consente di determinare il CONTRIBUTO INDIPENDENTE di ciascuno di questi fattori di rischio allo sviluppo di malattie cardiache

Katz MH. *Multivariable analysis: a primer for readers of medical research.* Ann Intern Med. 2003 Apr 15;138(8):644-50.

Il tipo di variabile **Outcome** determina la scelta del **modello**

Outcome	Modello
Continua	Linear regression
Conteggio	Poisson regression
Survival	Cox regression
Binomiale	Logistic regression



MODELLI di Regressione: Semplice e Multipla

Modelli di regressione:

- Lineare
- Poisson
- COX
- Logistica

• Utilizzo per:

- Stimare la forza del legame tra un outcome e più covariate
- Controllare il confondimento
- Costruire modelli causali
- Predire il rischio

Modelli di Regressione: Terminologia

y	=	$\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$
Variabile predetta		Predittori
Variabile risposta		Variabili esplicative
Variabile "Outcome"		Covariate/Fattori
Variabile Dipendente		Variabili Indipendenti