

CDL in INFERMIERISTICA & OSTETRICIA

Statistica Medica

gbarbati@units.it

A.A. 2024-25



UNITÀ DI BIOSTATISTICA

Dipartimento Universitario Clinico di
Scienze Mediche Chirurgiche e della Salute

CALENDARIO DELLE LEZIONI:

09/4: Lezione introduttiva + statistica descrittiva I

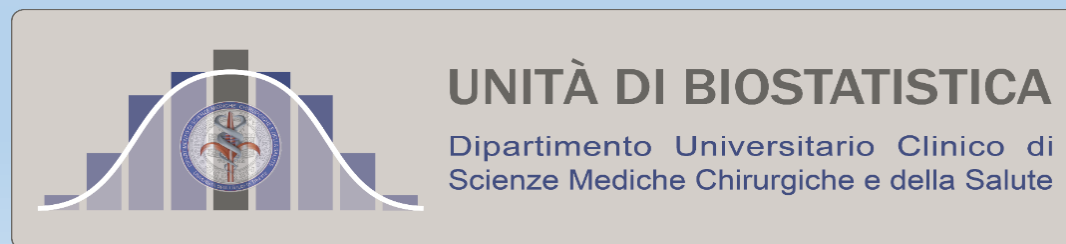
16/4: Statistica descrittiva II

30/4: Introduzione alla statistica inferenziale

07/5: Lezione di statistica inferenziale II

14/05: Test diagnostici, Dimensione Campionaria, misure di rischio

28/05: Modelli di regressione



Scaricare il materiale didattico da Moodle:

<https://moodle2.units.it/>

Dipartimento Universitario Clinico di Scienze mediche, chirurgi... / Laurea triennale (DM270) / ME10 - INFERMIERI

702ME-1 - STATISTICA MEDICA 2025

Corso

Impostazioni

Partecipanti

Valutazioni

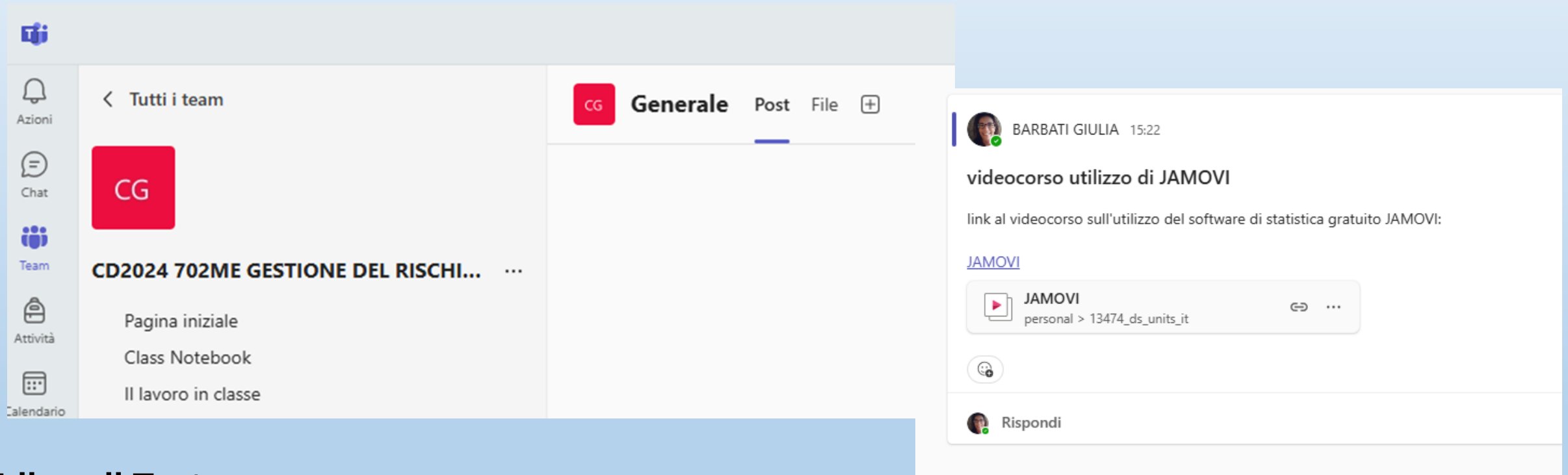
Report

Altro ▾

Materiale didattico:

Lezioni (slides in .pdf), esercizi

Video Corso per l'utilizzo di JAMОВI su Microsoft Stream:



Libro di Testo:

W.W. Daniel, C.L. Cross, «*Biostatistica. Concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria*», EdiSES, 2019.

Modalità di esame

La modalità di esame scritto sarà su **Moodle**, attraverso l'utilizzo di un **quiz** online.

Ci sarà poi una prova orale, con il Prof. Sanson, basata sulla lettura e interpretazione critica di un articolo scientifico, da organizzare a gruppi di studenti.

La parte di corso con JAMOV **NON** sarà oggetto dell'esame, ma questa parte è utile per poi rendervi autonomi nella elaborazione dei dati relativi alla vostra tesi di laurea

Ricevimento studenti: al termine della lezione o tramite videochiamata in MS teams o in Università Centrale, Edificio D, studio 4.2 [quarto piano]

Previo avviso tramite e-mail: gbarbati@units.it



SOMMARIO (I):

- Introduzione alla Statistica
- Concetti generali e terminologia
- Studi sperimentali e osservazionali

Che cos'è la statistica?

La statistica è la tecnica che ha come scopo la conoscenza **quantitativa** dei **fenomeni collettivi**.

Es: Quante persone ci sono in questa classe?

L'operazione che si deve fare è forse la più importante, ma certo la più semplice, operazione della statistica: il **conteggio**.
In base ad esso si perviene alla conoscenza del numero delle persone che costituiscono il gruppo.

I fenomeni che la nostra mente non può conoscere con una sola osservazione, ma che invece apprende tramite la sintesi delle osservazioni di fenomeni più semplici, si possono definire *fenomeni di massa* o *fenomeni collettivi*.

Essi sono composti da una collettività di osservazioni di fenomeni più semplici che sono detti quindi, *fenomeni individuali* o *singoli*.

Tramite la statistica è possibile sostituire ad un'impressione *qualitativa* di un fenomeno la sua misura *quantitativa*.

Che cosa si intende in statistica per 'fenomeni collettivi'?

- Collettività di casi singoli:

Caratteristiche di una popolazione osservando gli individui che compongono.
(Es: nascite, matrimoni, morti...)

- Ripetizioni dell'osservazione di un unico fenomeno singolo:

La lunghezza di un segmento, la massa di un corpo, la temperatura di un materiale ecc...: un'unica misurazione non può fornire l'esatta misura del fenomeno, poiché eseguendo misure della stessa grandezza si ottengono generalmente misure diverse (anche se di poco). Se le differenze sono effetti di cause che sfuggono ad ogni controllo, si può avere una misura verosimilmente esatta **ripetendo** l'osservazione un certo numero di volte e sintetizzando in un unico numero i risultati ottenuti. Alla misura si perviene quindi mediante una **collettività di misurazioni**.

Chi usa la statistica?

- Le scienze che studiano un aspetto di una collettività di individui o enti:
per es: demografia, economia, sociologia, *epidemiologia* (cioè quel settore della medicina che studia il tipo, la frequenza, la modalità di distribuzione e il ruolo di fattori causali di un determinato evento nelle popolazioni)...ecc...
- Le scienze cosiddette 'sperimentali' quelle cioè che devono fornire ed utilizzare dati misurati con molta precisione: la fisica, la chimica, l'ingegneria, la biologia, la ricerca biomedica...ecc...

Qual è il ruolo della statistica nella metodologia della ricerca?

Lo studio dei fenomeni collettivi avviene attraverso diverse fasi:

FASE 1: Schematizzazione, consiste nella **definizione del fenomeno**, nell'individuazione della **collettività** in cui esso si realizza e nella scelta delle **caratteristiche** della collettività che interessano.

Per esempio: se si intende studiare la *propensione al consumo di alcolici* (= il fenomeno) nella *popolazione europea* per la fascia di età compresa tra i 14 e i 30 anni (= collettività) si potrebbe decidere di rilevare *la spesa media per alcolici in una settimana* (= caratteristica di interesse).

(A) TUTTA LA POPOLAZIONE VIENE STUDIATA:

Se la collettività di individui alla quale è legato il fenomeno in esame è *interamente da osservare* ed allora alla fase di schematizzazione segue la:

FASE 2A: Osservazione: raccolta, nell'ordinamento e nella classificazione del materiale di osservazione.

Al termine di queste due fasi, alla collettività di individui si sostituisce un ***insieme di dati*** (=es. dati sulla spesa media settimanale per alcolici per individuo) su cui si può operare con procedimenti per lo più propri della matematica per la:

FASE 3A: Descrizione.

FASE 1: Schematizzazione, consiste nella **definizione del fenomeno**, nell'individuazione della **collettività** in cui esso si realizza e nella scelta delle **caratteristiche** della collettività che interessano.

(B) VIENE STUDIATA UNA PARTE DELLA POPOLAZIONE:

Se della collettività di individui alla quale è legato il fenomeno conviene osservare solo una parte (un *campione*), allora alla fase di *schematizzazione* segue quella di:

FASE 2B: Formulazione di ipotesi sulla collettività più vasta, o di enunciazione di una teoria o di ricorso ad un modello;

FASE 3B: Osservazione del campione, che è una parte della collettività, e a cui poi si sostituisce un insieme di dati, che viene poi

FASE 4B: Descrizione con procedimenti matematici dei dati rilevati sul campione.

A tutte queste fasi segue quella di:

FASE 5B: Induzione o di **Inferenza** che fa risalire dalla descrizione del campione a quella dell'insieme più ampio e che è anche una verifica delle ipotesi formulate nella fase 2b.

(A): TUTTA LA POPOLAZIONE:

Nel primo caso la metodologia impiegata è puramente descrittiva, da cui deriva il termine di “**statistica descrittiva**” dato all’insieme di tecniche relative a questo tipo di indagini. Occorre comunque notare che **anche nella statistica descrittiva si può fare riferimento a dei campioni di popolazione**; l’intera collettività viene generalmente osservata solo in occasioni di indagini particolari, come per esempio il **censimento**.

(Statistica descrittiva).

(B) UN CAMPIONE DELLA POPOLAZIONE:

Nel secondo caso, la metodologia impiegata può essere detta inferenziale ed il relativo campo della statistica viene definito “**statistica inferenziale**”. Ovviamente nella statistica inferenziale si devono compiere le operazioni di schematizzazione e di descrizione e quindi necessitano tutti gli strumenti che sono propri della statistica descrittiva; ad essi poi vanno aggiunti i procedimenti puramente inferenziali.

(Statistica inferenziale).

La statistica ci aiuta nel delineare il quesito scientifico e le diverse fasi del progetto di ricerca:

quesito scientifico -> ipotesi -> disegno dello studio -> analisi dei dati

Raccolta e gestione dei dati & analisi statistica sono delle **componenti cruciali** per la maggioranza degli studi clinici.

Quesiti di ricerca:

- Quali gruppi si vogliono **descrivere/confrontare**?
- Quali parametri si vogliono **misurare** ?
- **Quando**/in che occasione/i si vogliono misurare ?
- Quali parametri si ha bisogno di “**controllare**” per confrontare i gruppi ?
- ...

OBIETTIVI DELLA RICERCA

Descrivere associazioni/*patterns*.

- Moderatamente semplice...
- Occorre compilare un *report* delle analisi statistiche fatte.

————→ Statistica descrittiva

Inferenza: generalizzare dal campione alla popolazione.

- Moderatamente difficile...
- Richiede un piano di analisi *prima* di cominciare.

————→ Statistica
inferenziale

Prevedere cosa accadrà ai soggetti della popolazione di interesse.

- Descrittiva + Inferenza...(usando i modelli adeguati...)
- Piano di analisi pre/post per pianificare/documentare.

└————→ Modelli di regressione
[...forse faremo alcuni
cenni...]

TERMINOLOGIA:

Outcome : il risultato di interesse per i soggetti (unità statistiche) dello studio.

- ◀ es : tempo ad un evento (morte/recidiva);
- ◀ es : un parametro di funzionalità fisiologica o uno stato di disabilità;

Predittore/i di interesse : la/le variabile/i che si vogliono studiare come possibili **cause** dell'outcome.

- ◀ es : genotipo/i
- ◀ es : trattamento

Potenziale Confondente: variabile che può essere “**confusa**” con un predittore di interesse

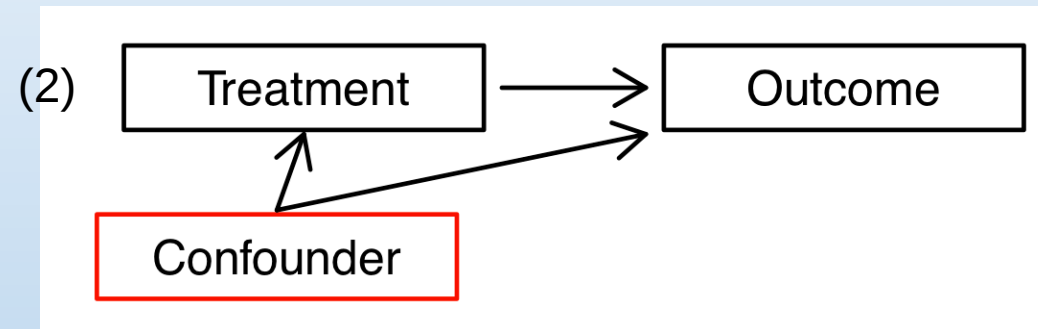
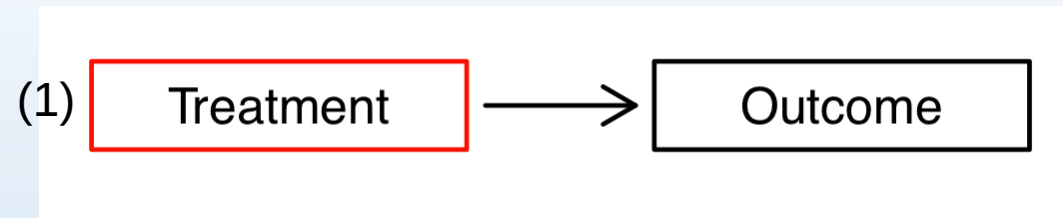
(associata sia all'outcome che al predittore).

- ◀ es: indicazioni cliniche ad un certo trattamento;
- ◀ es: luogo di arruolamento;

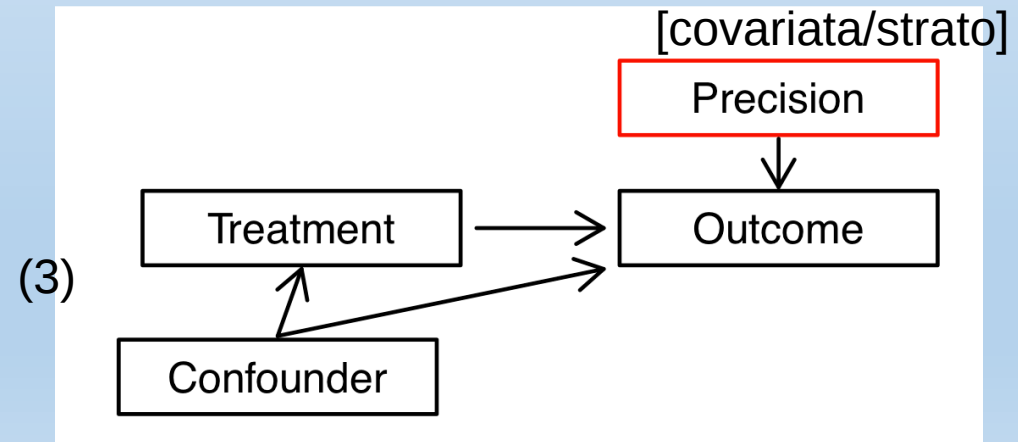
‘Precision’/[covariata/strato]: spiega parte della **variabilità** dell'outcome (non è associata al predittore di interesse).

- ◀ es: baseline (pre-randomization) stato di salute;
- ◀ es: età/sexo...

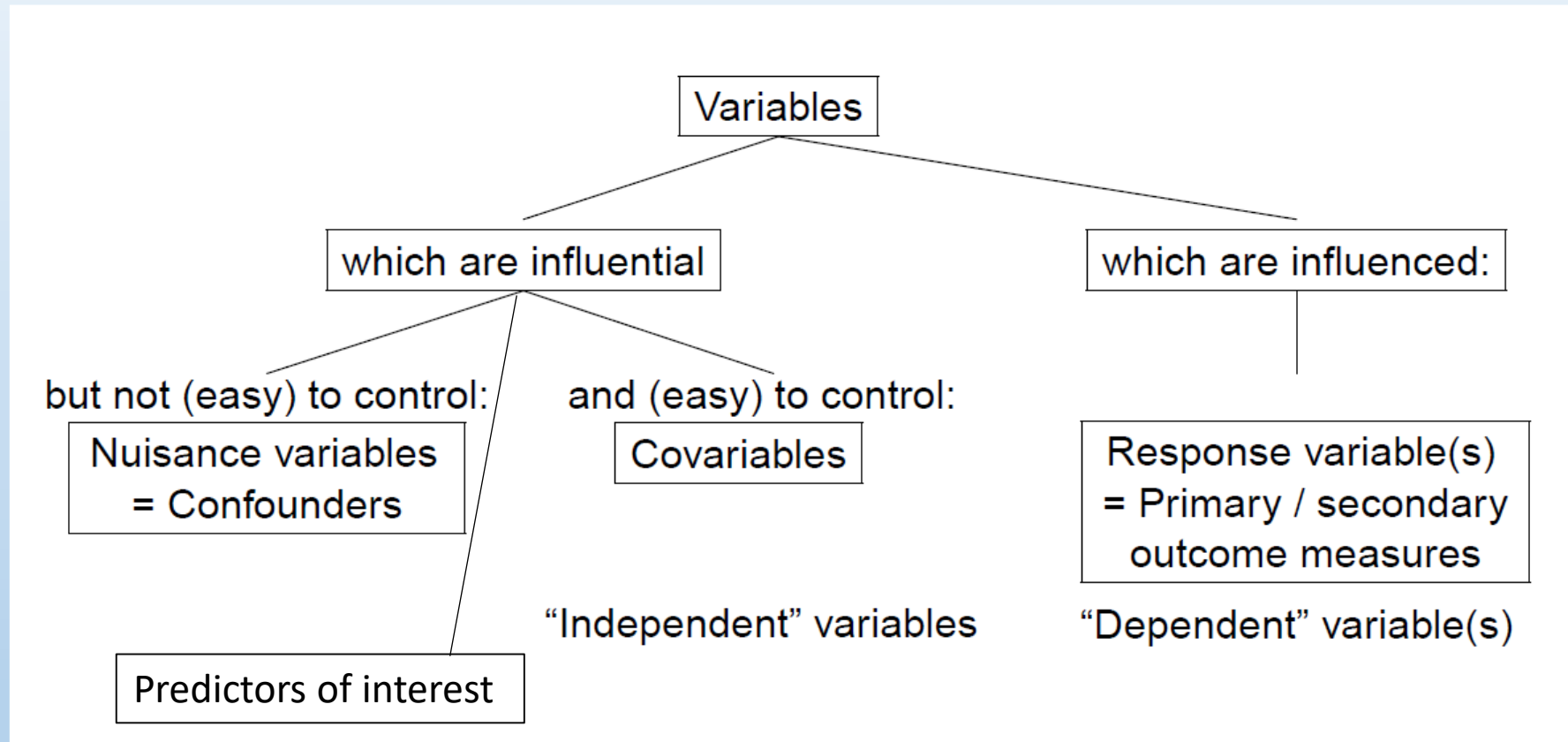




Livello crescente di «complessità»
del disegno dello studio
da (1) a (3)



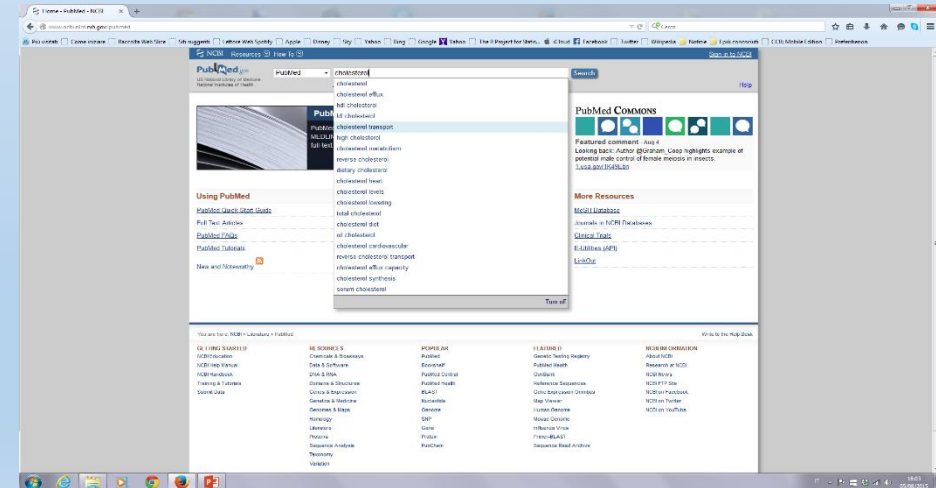
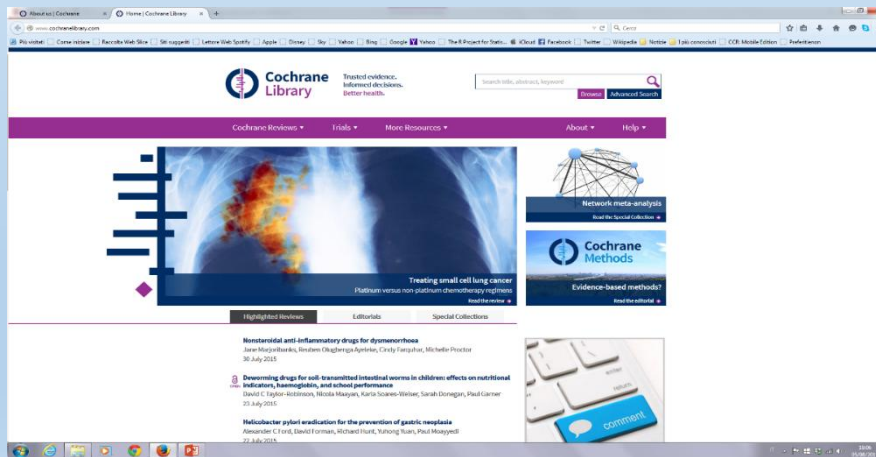
Le 'Variabili' in un Disegno di Studio:



Prima di cominciare: revisione sistematica della letteratura !!

Revisione **aggiornata** della letteratura sull'argomento da studiare:

- Preparare un report (=background) sullo stato dell'arte della ricerca sul fenomeno, (referenze -> **bibliografia**) ; discutere **in modo critico** risultati e metodi dei papers riportati.
- Documentare dai risultati di altri studi la efficacia del trattamento/farmaco/intervento oggetto di studio o sul potenziale fattore di rischio, per avere delle **stime quantitative** del suo effetto... (-> serve per determinare la dimensione campionaria opportuna).



Nel pianificare uno studio due aspetti sono particolarmente rilevanti:

-> **precisione** dei risultati

-> **capacità di rispondere alle domande** che ci si era posti

Questioni più frequenti da risolvere:

- (i) come **selezionare** gli individui da osservare;
- (ii) come decidere il **numero** di osservazioni da fare su uno o più gruppi di soggetti;
- (iii) come **ripartire** le osservazioni tra differenti possibili gruppi
 - per es. come suddividere i soggetti che devono ricevere differenti trattamenti-

Il campionamento

Alla base di uno studio statistico è il tipo di *campionamento* che si è condotto; a parte infatti alcuni casi particolari che richiedono la completa enumerazione dei soggetti studiati (ex: il *censimento* della popolazione), la maggior parte degli studi richiede la formazione di un *campione*.

Il ***campionamento semplice casuale*** è il modello basilare per la maggior parte delle indagini:

- si definisce una sorta di *lista* che identifichi gli individui della popolazione da cui estrarre il campione;
(ex: in un'indagine ospedaliera i codici che identificano in modo anonimo i pazienti ricoverati)
- si estraggono *un certo numero* di codici *a caso* utilizzando dei 'generatori di numeri casuali' che andranno ad identificare i pazienti selezionati nel campione.

Altri tipi di campionamento:

-campionamento stratificato: la popolazione viene dapprima suddivisa in gruppi (*strati*) (ex: l'Italia suddivisa nelle sue regioni) e si estrae un campione per ogni strato (ex: 100 cittadini per ogni regione);

-campionamento a più stadi: la popolazione viene suddivisa in unità appartenenti ad un primo *stadio di campionamento* (ex: si estraggono *casualmente* 10 scuole in una certa zona) e poi in unità appartenenti ad un *secondo stadio di campionamento* (ex: per ogni scuola estratta nella prima fase si estraggono 2 classi).

Una prima distinzione tra due tipi di studi statistici:

(1) **Esperimenti**

(2) **Indagini** (o studi osservazionali).

Esperimenti:	interferenza pianificata con il corso naturale degli eventi cosicché <u>l'effetto sperimentale</u> possa essere osservato.
Indagine:	il ricercatore è un <u>osservatore</u> più passivo ed interferisce il meno possibile con il fenomeno che deve essere studiato.

Esempi:

- Indagine sui tipi di veicoli a motore che passano per un certo casello stradale durante un certo intervallo di tempo;
(indagine)
- Sondaggio di opinione; (indagine)
- Studio sulla funzione respiratoria (misurata attraverso vari test) dei lavoratori in una certa industria; (indagine)
- Osservazione sui tempi di sopravvivenza di topi di diverse razze dopo avere inoculato loro una certa dose di una sostanza tossica; (esperimento)
- Indagine clinica per confrontare i vantaggi tra intervento chirurgico o trattamento farmacologico per pazienti in una certa condizione, avendo assegnato in modo casuale i soggetti ai due trattamenti. (esperimento)

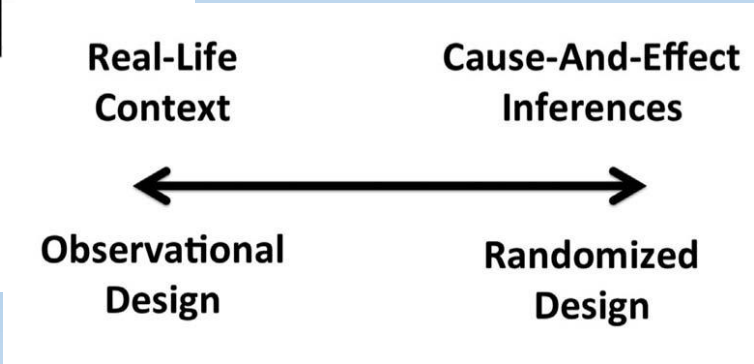
```
graph LR; SD[Study Designs] --> O["Observational  
(the researcher studies,  
but does not  
alter, what occurs)"]; SD --> E["Experimental  
(the researcher intervenes  
to change reality, then  
observes what happens)"];
```

Study Designs

- Observational**
(the researcher studies, but does not alter, what occurs)
- Experimental**
(the researcher intervenes to change reality, then observes what happens)

Real-World Context

← **Observational Design**



Gli esperimenti

Studi che vengono pianificati in generale con lo scopo di confrontare gli effetti di diversi trattamenti sullo stesso tipo di unità sperimentali.

In particolare in medicina tali esperimenti si definiscono con il termine inglese: '*clinical trials*'. Il trattamento da applicare a ciascuna unità viene deciso dal ricercatore.

esempi:

- confronto sull'effetto in diversi animali di diverse dosi una sostanza chimica;
unità=gli animali, il trattamento= le diverse dosi della sostanza.
- studio di profilassi per confrontare l'efficacia sui bambini di diversi vaccini contro il morbillo.
unità=bambino, il trattamento=vaccino.
- confronto in un paziente che soffre di attacchi ricorrenti di una malattia cronica di diversi metodi per alleviare il dolore.
unità=le occasioni successive degli attacchi, il trattamento= metodi anti-dolore.
- studio sui vantaggi e svantaggi relativi a diversi programmi di educazione alla salute.
unità=aree, trattamento= programmi di educazione

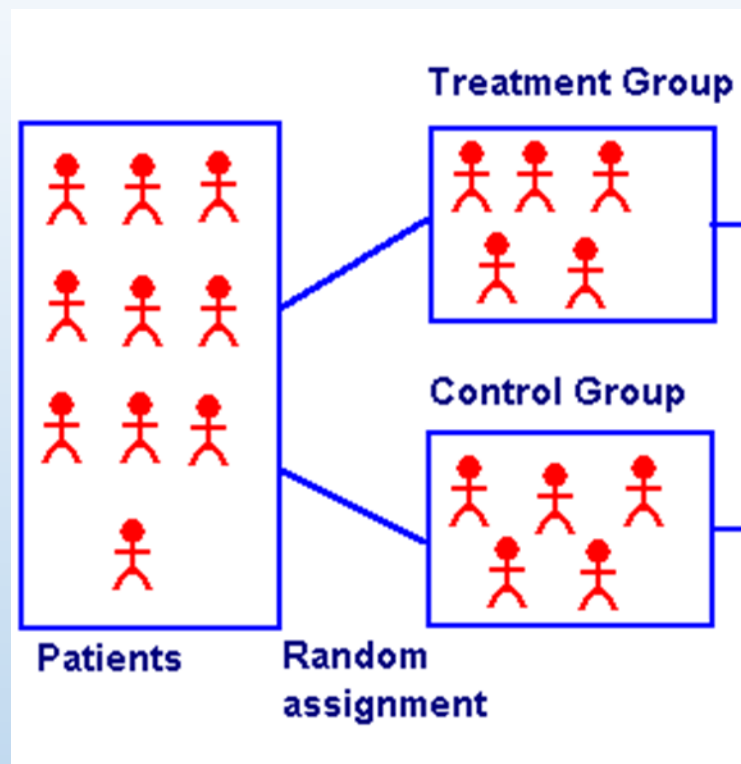
...come *assegnare* i trattamenti alle unità disponibili?

I gruppi di unità sperimentali a cui vengono assegnati i diversi trattamenti devono essere il più *simili* possibile, almeno negli aspetti più rilevanti per la sperimentazione.

Nonostante l'attenzione del ricercatore a che i *fattori* ritenuti importanti siano *bilanciati*, non si può mai essere sicuri che i gruppi non differiscano in modo rilevante per qualche fattore anch'esso importante ma che è stato ignorato durante l'attribuzione ai trattamenti.

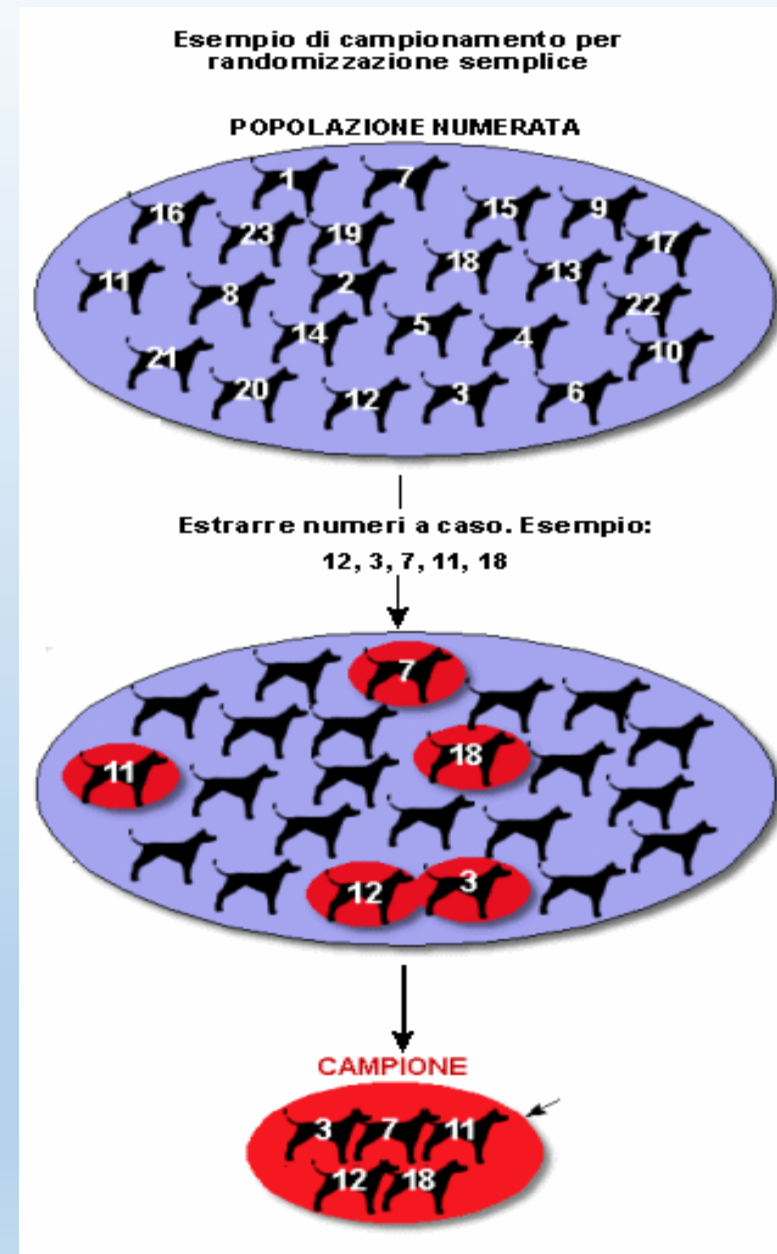
Soluzione (semplice) = **randomizzazione**

= fissato il numero n di unità da assegnare ad ogni gruppo, si estraggono dei campioni casuali di n unità dal gruppo complessivo di N unità sperimentali.



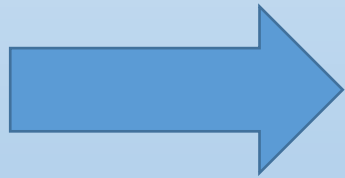
I gruppi di unità sperimentali a cui vengono assegnati i diversi trattamenti devono essere il più *simili* possibile, almeno negli aspetti più rilevanti per la sperimentazione...

Soluzione (semplice) = **randomizzazione**



...*Quante* unità assegnare ad ogni trattamento?

...in generale occorre infatti la *replicazione* delle osservazioni, cioè l'uso di più di una unità sperimentale per ogni trattamento.



Problema della **dimensione campionaria** «ottimale»....

Perché ripetere le osservazioni:

- assicura che i confronti tra i trattamenti siano sufficientemente precisi;

il cosiddetto 'errore di campionamento' della stima diminuisce all'aumentare delle osservazioni

- le osservazioni ripetute tengono conto della *variabilità* tra le unità sperimentali;

nel confrontare per es. due procedure chirurgiche, è utile un esperimento 'cooperativo' in cui le metodologie da confrontare vengano applicate in vari ospedali, cosicchè gli effetti di variazione dovuti a diversi medici e alle diverse strutture possa essere tenuto in conto..

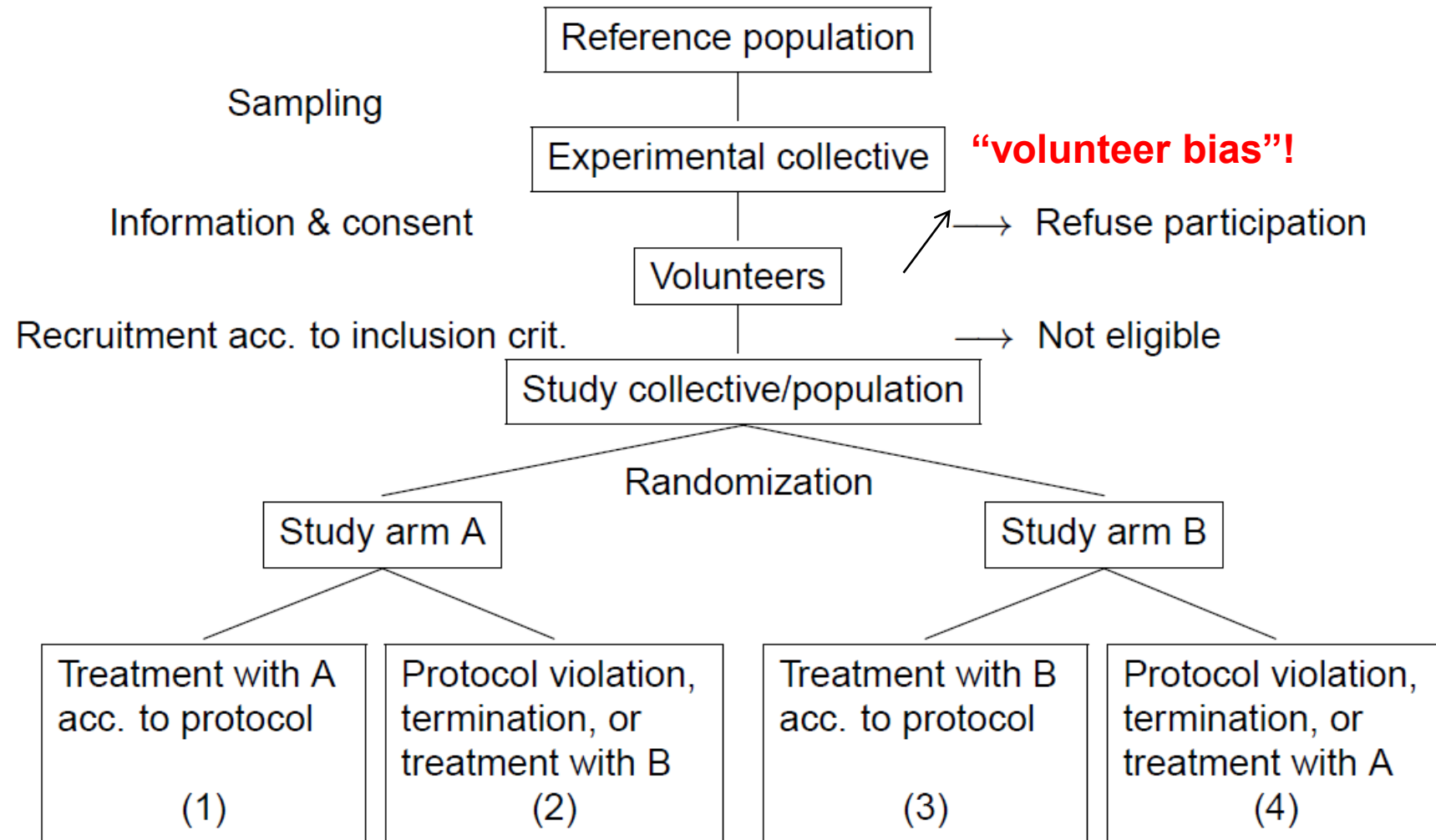
'errore di stima'
(per la media, per esempio): $\longrightarrow \frac{\sigma}{\sqrt{n}}$

σ = misura della **variabilità** tra le unità sperimentali rispetto al fenomeno oggetto di studio
 n = **dimensione** del campione di osservazioni, ovvero il *numero* delle unità sperimentali.

Per diminuire l'errore di stima si hanno due possibilità:

- si aumenta il denominatore, cioè il numero delle unità sperimentali n ;
- si riduce il numeratore σ : si cerca di prendere le unità sperimentali in modo più *omogeneo possibile* rispetto al trattamento da sottoporre a verifica

Studi Clinici Sperimentali (Clinical Trials): SCHEMA



Studi Clinici Sperimentali (Clinical Trials): BLINDING (MASKING)

Per minimizzare la soggettività di chi valuta i pts o analizza i dati.

- **SINGLE BLIND:** i soggetti dello studio non sanno quale trattamento ricevono
- **DOUBLE BLIND:** i soggetti e i ricercatori (medici/biologi) non sanno
- **TRIPLE BLIND:** i soggetti, i ricercatori, gli statistici non sanno



Figure 1: The authors: double blinded versus single blinded
Schulz & Grimes. Lancet 2002

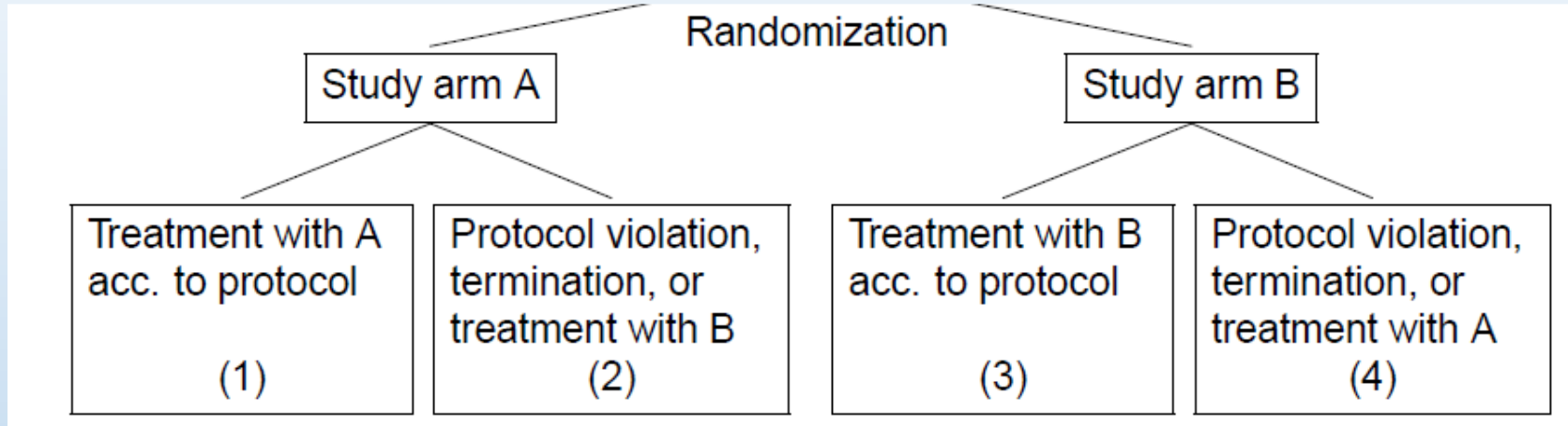
Se non viene effettuata: **OPEN TRIAL**

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002 Feb 16;359(9306):614-8



Deciphering the allocation concealment scheme

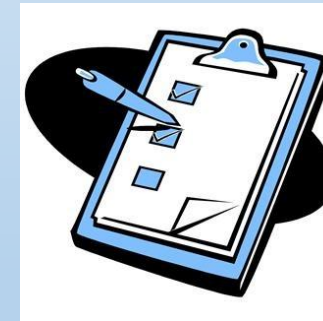
Studi Clinici Sperimentali (Clinical Trials): Analisi dei dati



ITT: intention to treat
(1)+(2) vs (3)+(4)

PP: per protocol (1) vs (3)

AT: as treated (1)+(4) vs (2)+(3)
(ma escludendo violazioni del protocollo o chiusure precoci)



Studi Clinici Sperimentali (Clinical Trials): limiti

- **Mis-classificazione/ 'information bias'** sull'outcome o sull'esposizione;
- **Co-interventi**: se un braccio riceve un intervento addizionale più frequentemente rispetto all'altro;
- **Contaminazione**: quando una parte di soggetti di «*controllo*» ricevono il trattamento al di fuori dello studio

Inoltre:

«**Volunteer bias**»: partecipano solo soggetti volontari (eligibilità allo studio, *compliance*, vicinanza geografica, stato socio-economico, stato di salute...)

Popolazione omogenea selezionata -> validità interna (bassa variabilità)
Ma: limite per la validazione esterna (generalizzabilità) dei risultati.

EFFICACY ≠ EFFECTIVENESS



Successo di un trattamento in un contesto
«artificiale»



Successo in un contesto
«real world»

Sommario (II):

- Concetti generali della Statistica Descrittiva
- Esempi di rappresentazione grafiche dei dati

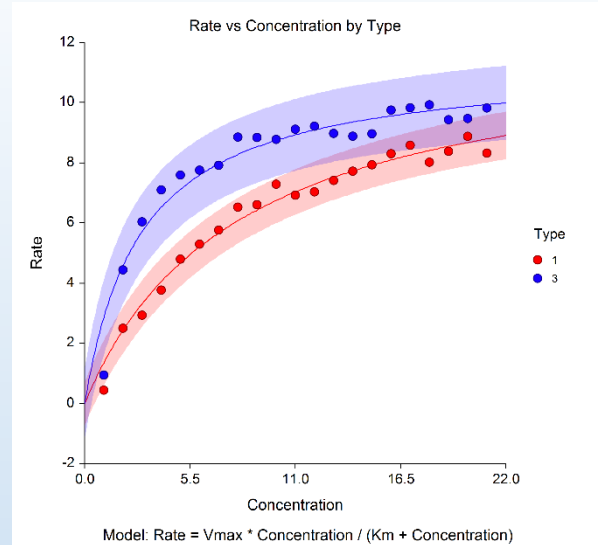
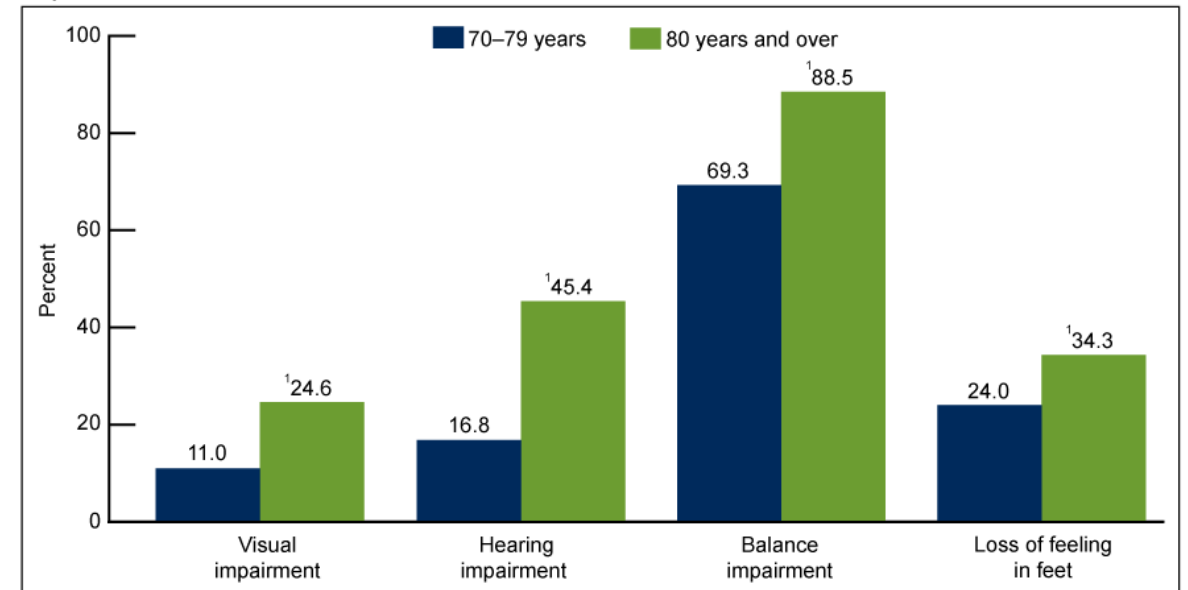


Figure 3. The prevalence of sensory impairments among persons aged 70–79 years compared with persons aged 80 years and over: United States, 1999–2006



¹Significantly different from the 70–79 age group.

SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey.

Che cos'è la statistica?

La statistica è la tecnica che ha come scopo la conoscenza **quantitativa** dei **fenomeni collettivi**.

Es: Quante persone ci sono in questa classe?

L'operazione che si deve fare è forse la più importante, ma certo la più semplice, operazione della statistica:
il **conteggio**.

In base ad esso si perviene alla conoscenza del numero delle persone che costituiscono il gruppo.

I fenomeni che la nostra mente non può conoscere con una sola osservazione, ma che apprende tramite la sintesi delle osservazioni di fenomeni più semplici, si possono definire

fenomeni di massa o fenomeni collettivi

Essi sono composti da una collettività di osservazioni di fenomeni più semplici

fenomeni individuali o singoli

Terminologia

L'oggetto dell'osservazione di ogni *fenomeno individuale* che costituisce il *fenomeno collettivo* in studio è detto **'unità statistica'**

Ex: un individuo, una A.S.L., un ospedale...

Ciascuna *unità statistica* presenta delle caratteristiche definite **'caratteri/variabili'**

Ex: il colore degli occhi in un individuo, il numero di medici che lavorano in una A.S.L., il numero di posti letto in un ospedale....

Ciascun *carattere* è presente in ogni *unità* con una determinata **'modalità'**

Ex: colore blu degli occhi, 10 medici in una certa ASL, 300 posti letto in un dato ospedale,...

La classificazione dei caratteri

I caratteri possono essere classificati in:

-Caratteri **qualitativi** (colore degli occhi, tipo di diagnosi...), a loro volta distinti in:

- ***ordinabili***: è possibile *ordinare* le modalità del carattere in senso crescente o decrescente (es: titolo di studio, livello di gravità della diagnosi...);
- ***sconnessi***: non c'è alcun ordinamento intrinseco tra le modalità (es: colore degli occhi, sesso...);

-Caratteri **quantitativi** (es: età, peso, numero di medici,...) distinguibili in:

- ***discreti***: le modalità del carattere sono numeri interi (es: numero di medici...)
- ***continui***: le modalità del carattere sono misurate su una scala continua (es: peso, altezza...).

Alla base di tale classificazione dei caratteri vi è la **'scala di misura'** con cui sono espresse le modalità: se attraverso dei **numeri** o delle **'etichette'**.

Nominale
(qualitativi
sconnessi)

Ordinale
(qualitativi
ordinabili)



«factor data»



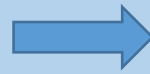
SCALA DI MISURA

«date/time data»



Discreta
(quantitativo)

Continua
(quantitativo)



«numeric data»



Sintesi scale di misura:

Nominale:

Variabili categoriali/qualitative non ordinate. Queste possono essere **binarie** (due sole categorie, come il genere: maschio o femmina) o **multinomiali** (più di due categorie, come lo stato civile: sposato, divorziato, mai sposato, vedovo, separato). La cosa fondamentale qui è che **non esiste un ordine logico** per le categorie.

Ordinale:

Categorie ordinate. Sempre variabili categoriali/qualitative, ma con un ordine. Gli elementi delle scale «Likert*» con risposte come: "Mai, a volte, spesso, sempre" sono ordinali.

Le variabili numeriche/quantitative possono essere ulteriormente suddivise in due tipologie: discrete e continue. Le variabili discrete, come i conteggi, possono assumere solo numeri interi: numero di bambini in una famiglia, numero di giorni persi dal lavoro. Le variabili continue possono assumere qualsiasi numero, anche oltre la virgola decimale.

Non è sempre ovvio che questi livelli di misurazione non riguardino solo il significato intrinseco della variabile stessa. Altrettanto importanti sono il significato della variabile **all'interno del contesto della ricerca** e **come è stata misurata**.

*Scale likert: costituite da un certo numero di affermazioni - definiti item - che esprimono un atteggiamento positivo e negativo rispetto ad uno specifico tema. La somma di tali giudizi tenderà a delineare l'atteggiamento del soggetto nei confronti del tema. Per ogni item si presenta una scala di accordo/disaccordo, generalmente a 5 o 7 modalità

Un esempio: l'età

L'età è, tecnicamente, una variabile numerica continua. L'età di una persona è continua se la si misura con sufficiente precisione. È significativo dire che qualcuno (o qualcosa) ha 7.28 anni. Detto questo, potremmo non essere in grado di trattarla come continua. Dipende da come è stata misurata e se ci sono implicazioni qualitative dell'età nel contesto di ricerca

Età come ordinale: non è raro dare alle persone categorie di età come possibili risposte a un sondaggio. Le ragioni comuni sono che le persone non vogliono rivelare la loro età effettiva o perché non ricordano l'età effettiva in cui si è verificato un evento.

L'età come discreta: età misurata in numero di giorni in cui i semi germinati di una specie iniziano a germogliare. La maggior parte lo farà entro pochi giorni e il valore potrebbe variare da 2 a 9 giorni. In questo contesto, l'età è sicuramente un conteggio discreto: il numero di giorni.

Età come multinomiale: a volte le variabili numeriche sono rese categoriche a causa della mancanza di valori. In uno studio, si misuravano le età di alcuni testimoni in un processo. Tecnicamente le età sono continue, ma in questo studio c'erano solo quattro valori: 49, 59, 70 e 80.

Età come binaria: in uno studio che confrontava l'equilibrio tra lavoro e vita privata di uomini e donne, la variabile di interesse era il numero di ore lavorate a settimana. Un predittore chiave per le donne, ma non per gli uomini, era l'età del figlio più piccolo. C'è una differenza qualitativa tra un bambino di 5 anni, che può essere idoneo solo per la scuola materna part-time e un bambino di 6 anni, che è abbastanza grande per andare a scuola a tempo pieno. Questa differenza qualitativa esiste in questo contesto tra 5 e 6 anni ma non esiste in altre differenze di età di un anno. Questa differenza qualitativa è infatti la caratteristica più importante dell'età del bambino più piccolo. Trattare l'età come continua in realtà ignora questa importante differenza qualitativa.

Esempio di matrice dei dati:

Abbiamo chiesto ad un campione di studenti se avessero visto l'ultimo Festival di Sanremo e se lo avessero gradito. Inoltre, abbiamo chiesto il genere (maschio/femmina) e il numero di ore giornaliere dedicate al sonno e ad attività di studio.

variabile

↓

Sappiamo definire la scala di misura di queste variabili?

Stu.	sex	Sanremo	...	sleep	study
1	maschio	Non l'ho visto	...	8	2
2	femmina	L'ho visto e mi è piaciuto	...	6	30
3	maschio	Non l'ho visto	...	9	5
4	femmina	Non l'ho visto	...	8	25
⋮	⋮	⋮	⋮	⋮	⋮
52	femmina	Non l'ho visto	...	8	20

←

unità statistica

Classificazione dei caratteri e scala di misura

CARATTERE		SCALA
qualitativo	Sconnesso	Nominale
	Ordinabile	Ordinale
quantitativo		Ad intervalli (scala numerica discreta o continua)

Operazioni che è possibile fare sui caratteri in base alla loro scala di misura:

Operazioni sulle modalità del carattere	Carattere		
	qualitativi		Quantitativi
	<i>sconnessi</i>	<i>ordinabili</i>	<i>(discreti/continui)</i>
= ; ≠	si	si	si
> ; <	no	si	si
+ ; -	no	no	si

Distribuzioni

Quando si rilevano le **modalità** con cui uno (o più) **caratteri** si presentano in ciascuna **unità** di una popolazione (o di un campione) si ottiene una **'distribuzione'** della popolazione/campione secondo il carattere considerato.

Ex: distribuzione 'unitaria' del peso in una classe:

Studente	Peso (kg)
Marco	62
Chiara	59
Luca	56
...	...

distribuzione «unitaria» = ad ogni unità del gruppo è associato il suo peso, cioè la corrispondente modalità del carattere considerato.

Possiamo suddividere in **'classi'** la popolazione secondo il carattere considerato, allora le modalità del carattere vengono raggruppate in classi ed otteniamo una **distribuzione di 'frequenze'** -> *frequenza della classe* = numero di unità statistiche che appartengono alla classe.

Distribuzioni di frequenze

Tipo di scuola			
		Frequenza	Percentuale
Validi	Altro	11	34,4
	Istituto Professionale	4	12,5
	Liceo Classico	3	9,4
	Liceo Scientifico	14	43,8
	Totale	32	100,0

Il numero di unità che appartengono ad una classe= **frequenza assoluta** della classe.

E' possibile calcolare anche le distribuzioni di **frequenze relative** e di **frequenze percentuali**

Alcuni simboli:

Abbiamo k classi di un dato carattere (nell'ex: 6 classi) cioè $k=6$; le possiamo indicare tramite un indice i che varia da 1 a 6: $i=1,2,3,4,5,6$. Abbiamo rilevato le modalità del carattere su N unità (nell'ex: $N=30$).

- Frequenza **assoluta** della classe i $\rightarrow n_i$
- Frequenza **relativa** della classe i $\rightarrow f_i = n_i / N$
- Frequenza **percentuale** della classe i $\rightarrow p_i = f_i * 100 = (n_i / N) * 100$

Esiste poi un simbolo matematico indica l'operazione di **somma** su un gruppo di oggetti ed è chiamato '**sommatoria**' :

$$\sum_{i=1}^k n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 30$$

$$\sum_{i=1}^k n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$$

$$\sum_{i=1}^k n_i = N ; \quad \sum_{i=1}^k f_i = 1 ; \quad \sum_{i=1}^k p_i = 100$$

(per $k=6$)

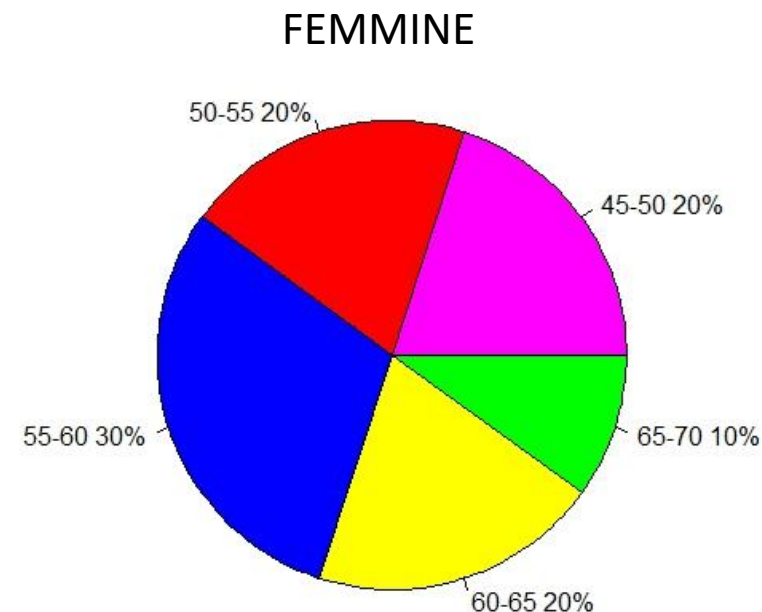
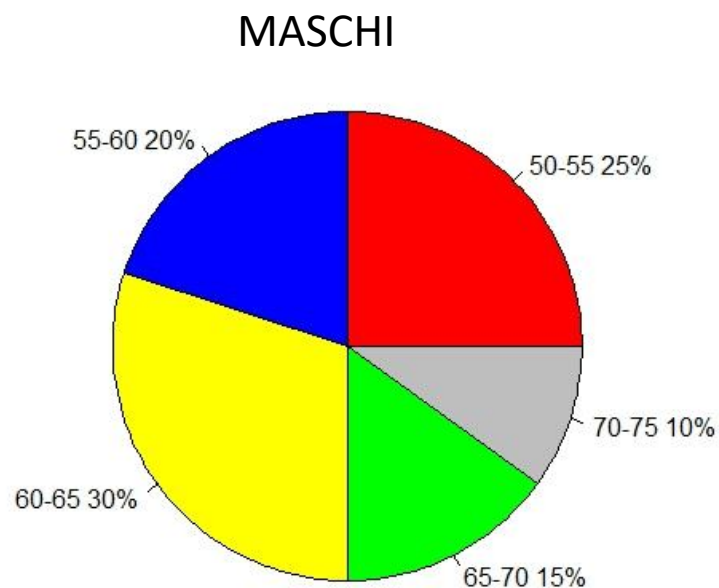
Distribuzioni doppie (dati categorici/in classi)

Se si rilevano **due** caratteri in una popolazione si ottiene una **distribuzione doppia**; è uso comune rappresentare le distribuzioni doppie delle modalità rilevate per due caratteri (**categorici o espressi in classi**) tramite la '**tabella di contingenza**' o '**tabella a doppia entrata**':

CLASSI DI PESO	M	F	TOT
45-50	0	2	2
50-55	5	2	7
55-60	4	3	7
60-65	6	2	8
65-70	3	1	4
70-75	2	0	2
TOT	20	10	30

Per confrontare correttamente la distribuzione delle classi di peso rispetto al sesso occorre calcolare le **frequenze relative** all'interno della tabella di contingenza:

CLASSI DI PESO	M	F
45-50	0	20
50-55	25	20
55-60	20	30
60-65	30	20
65-70	15	10
70-75	10	0
TOT	100	100



La tabella di contingenza permette inoltre di studiare **l'associazione** fra i caratteri analizzati
→ **test di ipotesi** che vedremo in seguito → **Statistica Inferenziale**

Tabella di Contingenza (a doppia entrata):

n_{aibj} = numero di soggetti che presentano *contemporaneamente* la modalità i del carattere A e la modalità j del carattere B.

		B				Totale
		B ₁	B ₂	...	B _m	
A	A ₁	na_1b_1	na_1b_2	...	na_1b_m	NA ₁
	A ₂	na_2b_1	na_2b_2	...	na_2b_m	NA ₂
	...	...	...	na_ib_j	...	...
	...	...	...	...	...	...
	A _k	na_kb_1	na_kb_2	...	na_kb_m	NA _k
Totale		NB ₁	NB ₂	...	NB _m	N

← ← ←
Marginali di Riga

↖ ↗
Marginali di colonna

Distribuzioni di frequenza per caratteri su scala quantitativa

Modalità	Frequenza assoluta
160	2
164	2
165	1
166	1
168	1
170	5
172	1
173	4
174	1
175	4
176	4
177	1
178	2
180	6
181	2
182	2
183	3
184	2
185	3
186	1
187	2
188	1
190	2

Abbiamo raccolto dei dati sulla distribuzione delle altezze in un campione di 53 studenti.

Vogliamo suddividere questa distribuzione in classi di frequenza.

Con che criterio scegliamo le classi?

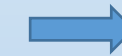
Modalità	Frequenza assoluta
[160,165]	5
(165,170]	7
(170,175]	10
(175,180]	13
(180,185]	12
(185,190]	6

Possiamo creare ad esempio

6 classi di frequenza:

[: la classe comprende il valore;

(: la classe NON comprende il valore;



Classi [circa] di uguale ampiezza

Modalità	Frequenza assoluta
[160,185]	47
(185,190]	6

Ma possiamo creare ad esempio solo 2 classi di frequenza:

[: la classe comprende il valore;

(: la classe NON comprende il valore;

La scelta delle classi è arbitraria, ma va fatta in maniera ragionevole....

Può capitare, o per scelta (si vuole fornire informazioni più dettagliate su una parte della distribuzione) o per necessità (i dati ci sono stati forniti già raggruppati in classi e non disponiamo della distribuzione unitaria) di costruire delle classi utilizzando *intervalli di lunghezza differente*.

In questo caso è conveniente definire il concetto di **densità** di frequenza della classe:

$$\left(\begin{array}{c} \text{densità} \\ \text{di una classe} \end{array} \right) = \frac{\text{frequenza assoluta di } Y \text{ sull'intervallo}}{\text{lunghezza dell'intervallo}}$$

La densità ci dice il ***numero atteso di unità statistiche*** per ogni unità di misura della variabile.

Modalità	Freq. ass.
0	1
11	1
79	1
100	1
112	1
119	1
130	2
140	1
150	1
162	1
169	1
176	1
200	2
231	1
254	1
257	1
277	1
300	1
349	1
350	1
356	2
370	1
400	1
438	1
439	1
450	1
463	1
469	1
470	1
500	2
520	1
543	2

Esempio: Numero di amici su Facebook, indagine fatta su un campione di 52 studenti.

Modalità	Frequenza assoluta	Ampiezza Classe	Densità	“Altezza” della densità (grafico) Freq rel/ampiezza
[0,100]	4	100	0,04	$(4/52)/100=0.0008$
(100,200]	11	100	0,11	0.0021
(200,300]	5	100	0,05	0.0010
(300,400]	6	100	0,06	0.0012
(400,500]	8	100	0,08	0.0015
(500,1000]	11	500	0,022	0.0004
(1000, 3500]	7	2500	0,0028	0.0001

La densità ci dice il *numero atteso di unità statistiche* per ogni unità di misura della variabile.

Nella classe [0,100] ci aspettiamo di osservare 4 persone in un intervallo di 100 unità.

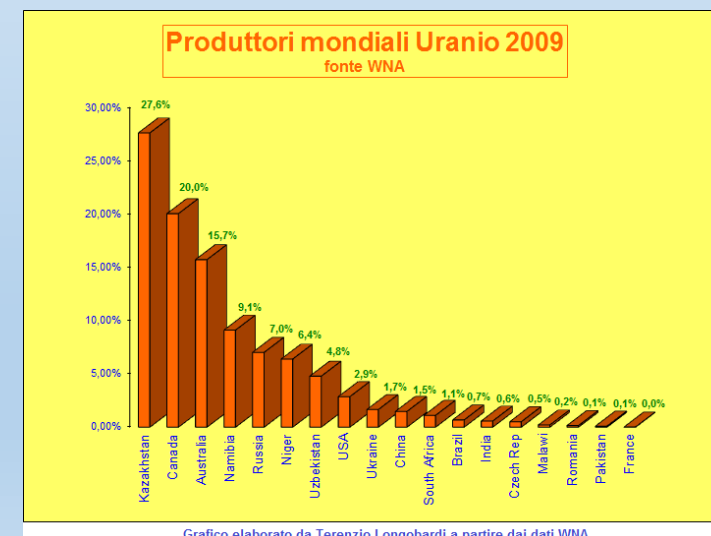
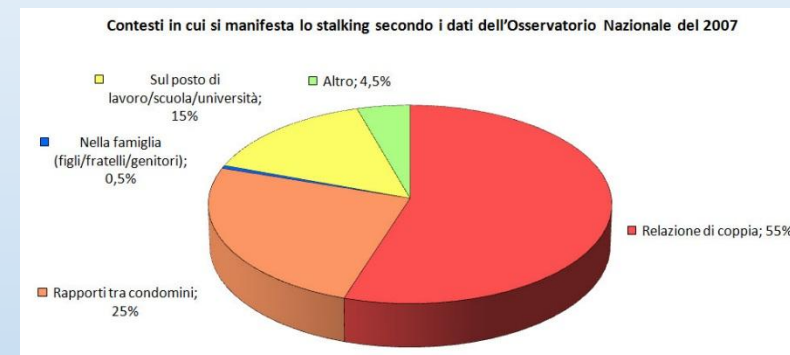
Nella classe (500,1000] ci aspettiamo di vedere 2.2 persone ogni 100 unità (cioè 2.2 persone tra 500 e 600, altrettante tra 600 e 700, ...etc).

Rappresentazioni grafiche dei dati

Una parte molto importante della statistica descrittiva è la rappresentazione grafica dei dati

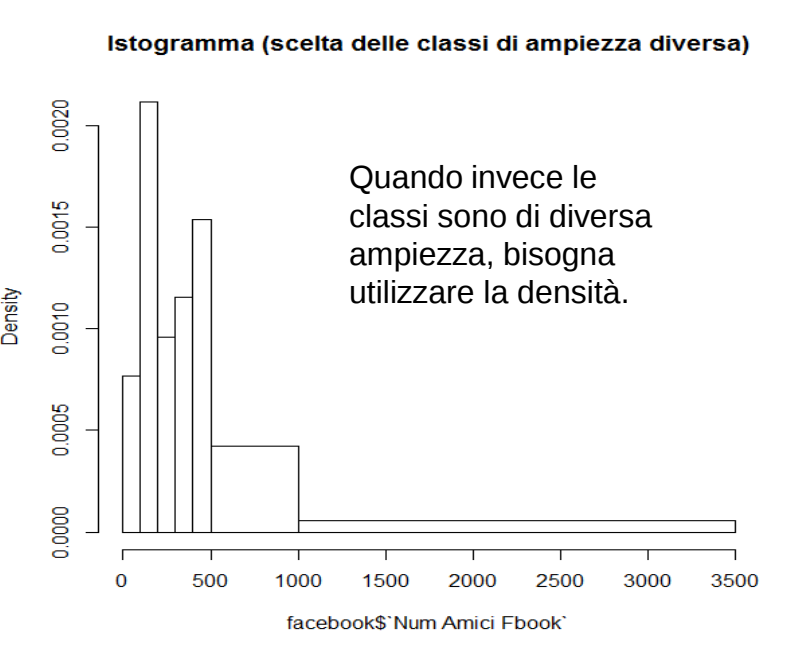
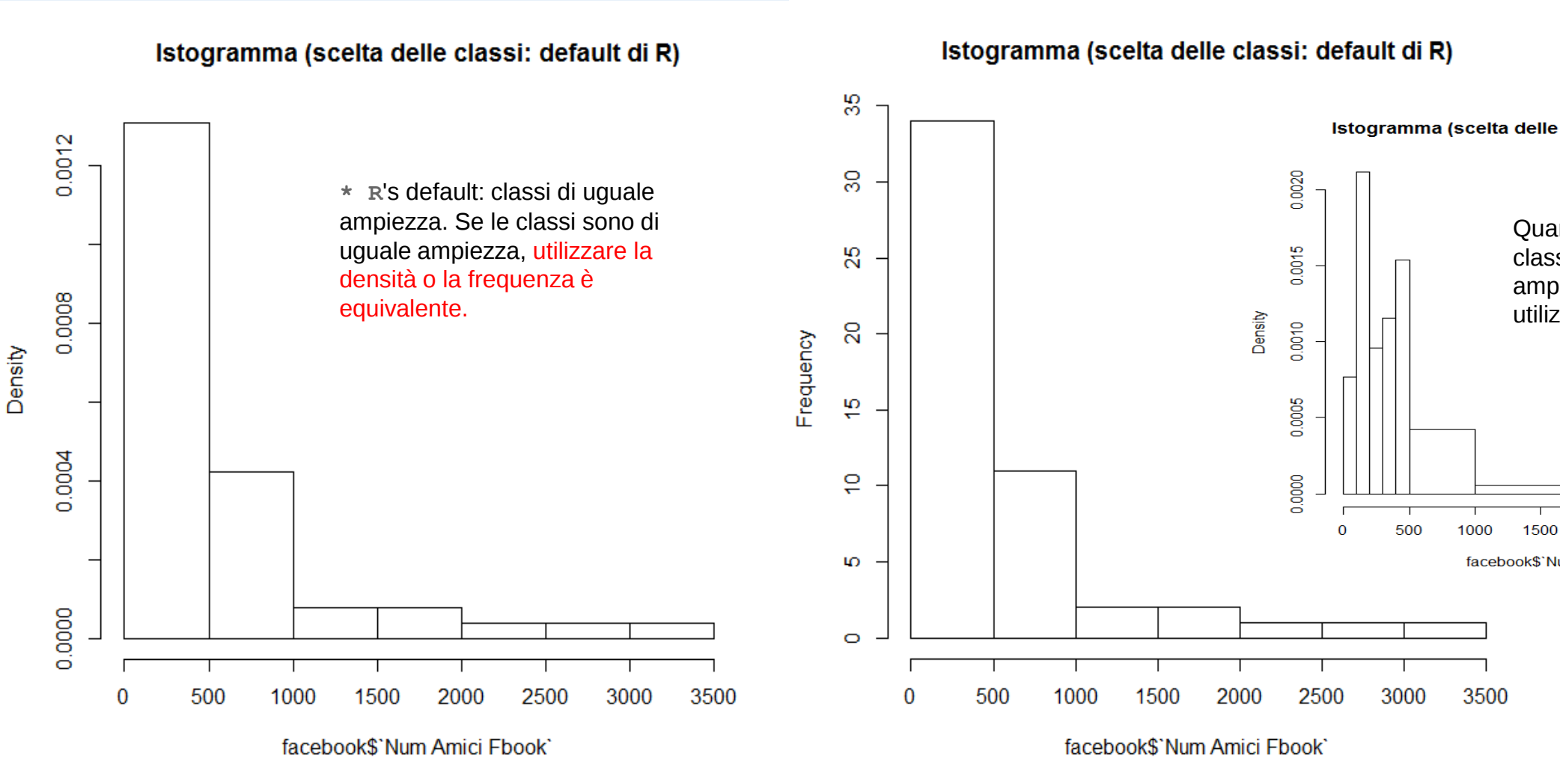


1. Scoprire la **struttura** dei dati;
2. Identificare i **valori più rilevanti** assunti dal fenomeno;
3. Identificare eventuali **valori anomali** (*outliers*)
4. **Comunicare** i risultati dello studio in modo efficace



Caratteri su scala numerica:

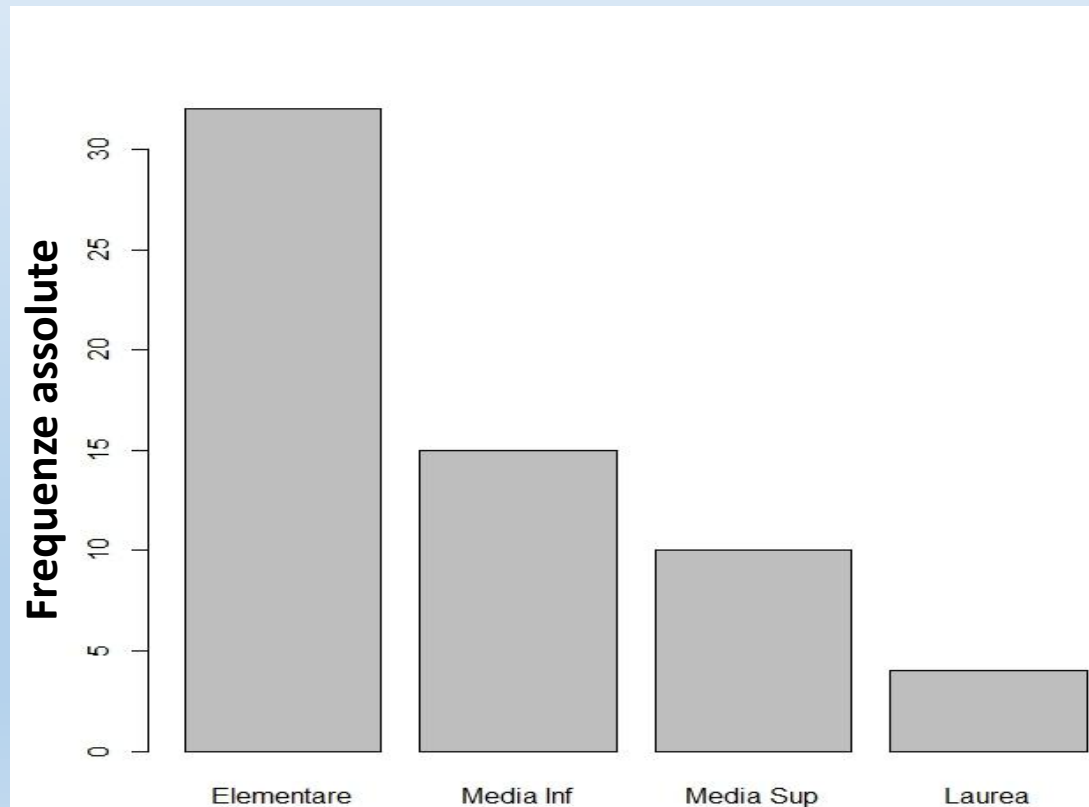
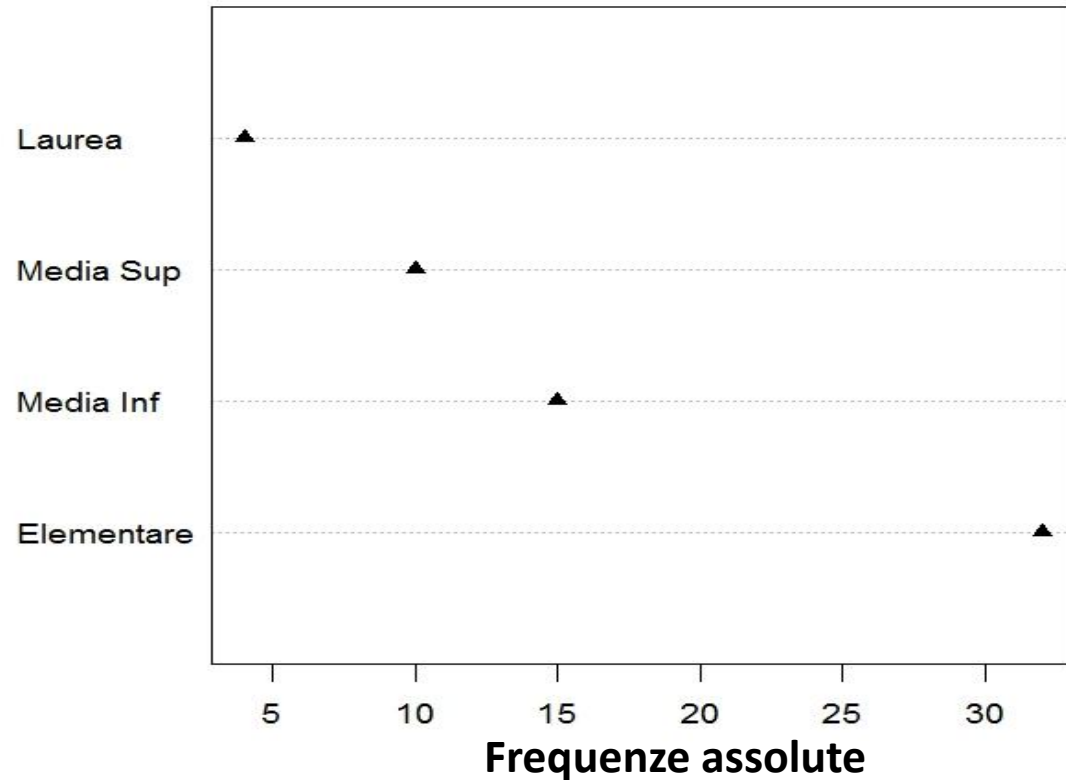
Per rappresentare graficamente una distribuzione secondo un carattere di tipo quantitativo possiamo utilizzare l'istogramma:



Caratteri su scala nominale/ordinale (I):

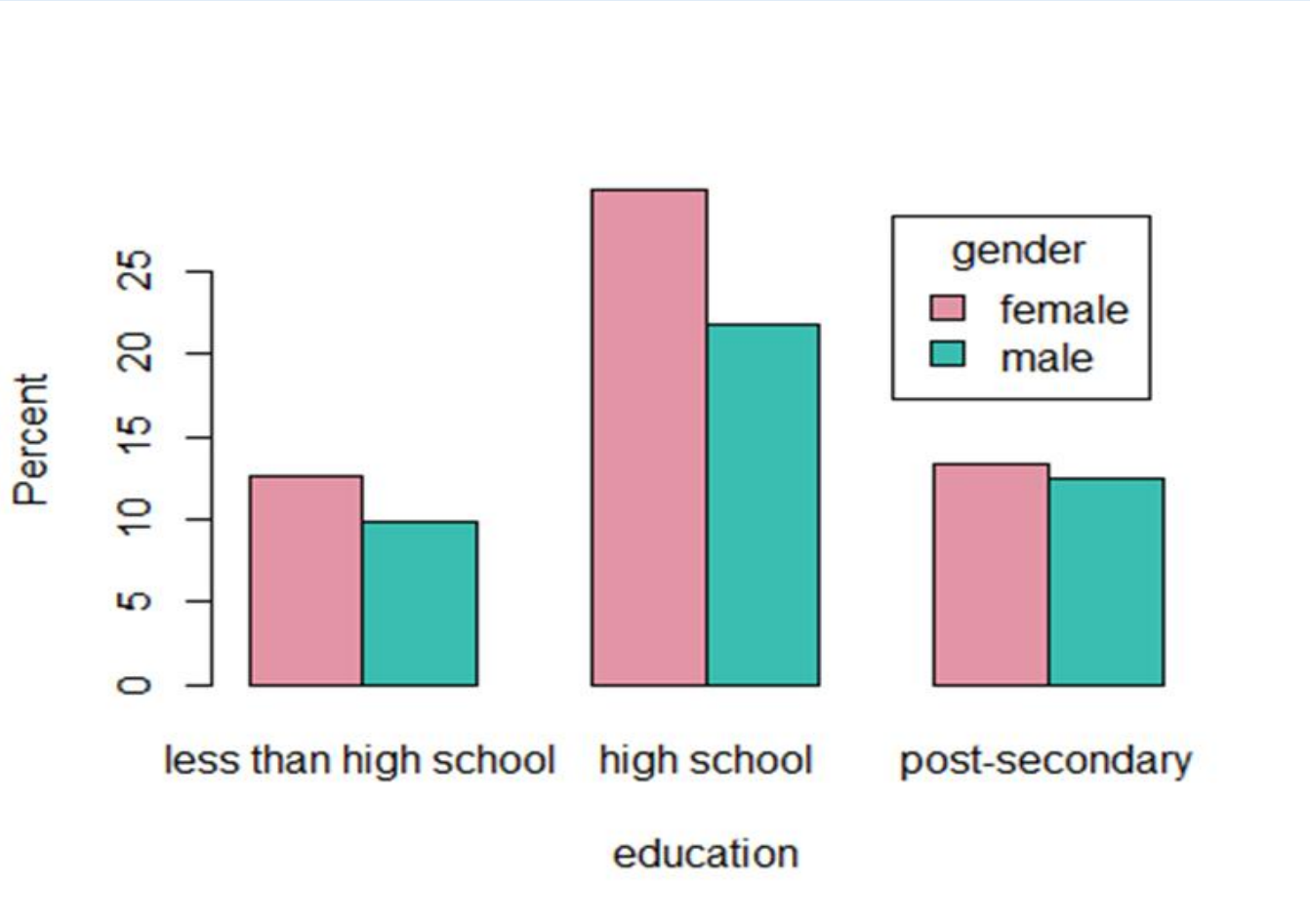
Per rappresentare graficamente una distribuzione secondo un carattere di tipo nominale o ordinale possiamo utilizzare il grafico «dotchart» oppure il «barplot»:

Elementare	Media Inf	Media Sup	Laurea
32	15	10	4



Caratteri su scala nominale/ordinale (II):

Per rappresentare graficamente una distribuzione doppia di due caratteri nominali/ordinali possiamo utilizzare il grafico «barplot» [meglio utilizzare le frequenze percentuali]:



Frequency **table**:

education	gender	
	female	male
high school	9987	7264
less than high school	4228	3280
post-secondary	4436	4159

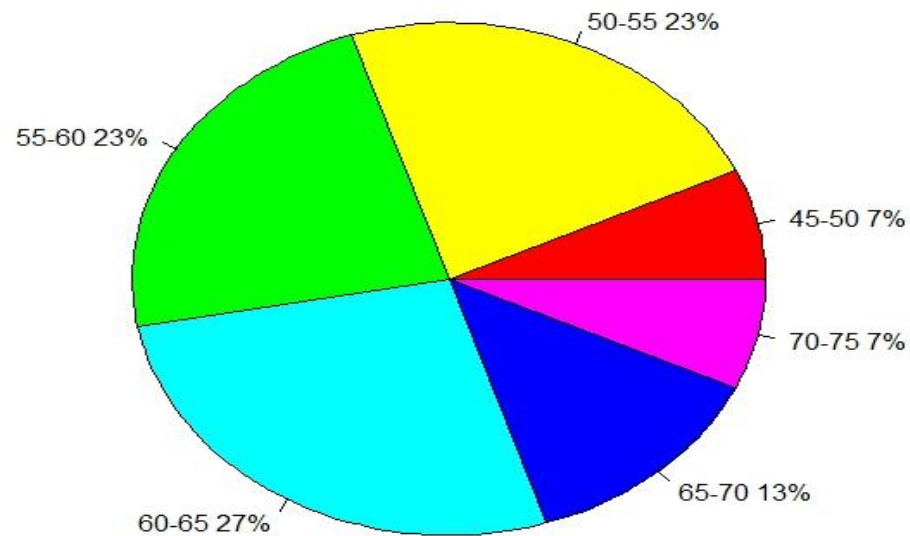
Total percentages:

	female	male	Total
high school	29.9	21.8	51.7
less than high school	12.7	9.8	22.5
post-secondary	13.3	12.5	25.8
Total	55.9	44.1	100.0

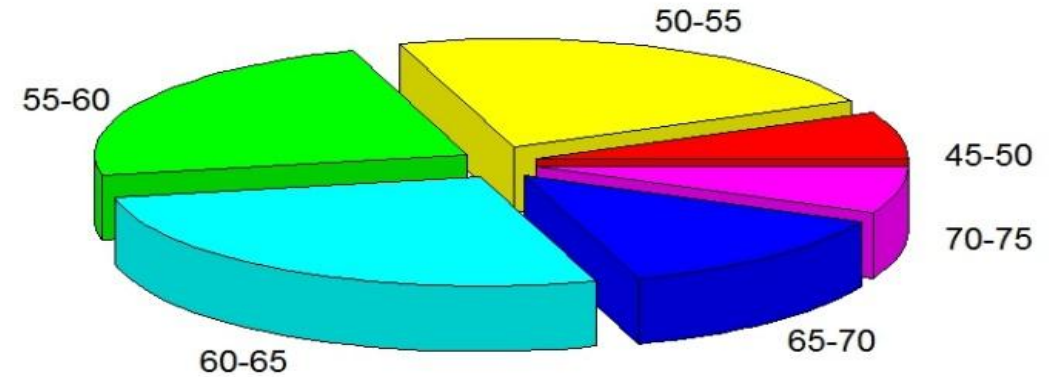
Rappresentazione grafica di distribuzioni in %

Per rappresentare una distribuzione in percentuale secondo un carattere di tipo categorico/in classi possiamo anche utilizzare il grafico a torta:

Pie Chart of Peso



Pie Chart of Peso



Distribuzioni doppie (dati su scala numerica)

Se si rilevano **due** caratteri in una popolazione su scala numerica e si è interessati a studiare l'associazione tra i due caratteri, la rappresentazione grafica usuale è lo **scatter plot (diagramma cartesiano)**:

Distribuzione su un campione di 252 maschi delle dimensioni del polso e del collo: si ipotizza che l'ampiezza del collo sia circa due volte quella del polso.

Vedremo in seguito come sintetizzare da un punto di vista statistico queste relazioni
(**correlazione/regressione lineare...**)

