LEZIONE 1

Esistono due definizione che definiscono al meglio la statistica:

1) **è una scienza che si occupa di sintetizzare e interpretare dati numerici** (Bland),
2) **è la disciplina che utilizza dati numerici che derivano da gruppi di unità statistiche** (Armitage).

Ad accumunare le due definizioni è il concetto di **dato** (informazione che deve essere analizzata e interpretata) e il concetto di **unità statistica.**

Sono quattro le categorie a cui bisogna prestare attenzione:

1. **Unità statistiche: l'elemento su cui si va a rilevare le informazioni**, ossia l'oggetto di studio, che non deve essere confusa con l'unità biologica, che tratta persone, oggetti, territori, tempi ecc…
2. **Popolazione**: **insieme di unità statistiche omogenee su cui si vuole condurre la rilevazione**, spesso **non definita nel tempo e nello spazio** (es. compresse prodotte in un processo produttivo), ed ecco perché può riferirsi ad infinite unità statistiche.
3. La valutazione può esser fatta riferendosi a tutte le unità della popolazione: **esaustiva** *(*es. censimento) o ad una parte della popolazione, **campionaria,** – poiché non è definita nel tempo e nello spazio (es. principio attivo di un medicinale);
4. **Variabile**: **il fenomeno oggetto di studio rilevato su unità statistiche** e **può assumere valori differenti** a seconda dell'unità statistica su cui si applica (genere, peso…);
5. **Modalità: il valore assunto da ciascuna variabile**, modo con cui una variabile si manifesta in un'unità statistica.

La statistica descrive ed analizza dei dati all'interno delle unità statistiche, in questo ambito è quasi sempre necessario il lavoro di sintesi e quindi **descrivere il fenomeno nel suo complesso**. Pertanto, **è necessario analizzare dei dati** e capire qual è la sintesi più opportuna da utilizzare per descrivere le informazioni.

Le variabili si dividono in qualitative e quantitative.

- Le variabili **qualitative non derivano da un numero ma dall'osservazione di una certa caratteristica**. Si tratta di etichette linguistiche (es. genere, intensità del dolore, grado d'istruzione…) e sebbene queste variabili in alcuni casi vengano rappresentate da numeri, non bisogna trattarle come variabili quantitative. All'interno delle variabili qualitative si distinguono le:
    a. variabili **qualitative nominali** in cui si stabilisce un uguaglianza o una differenza (es. genere);
    b. variabili **qualitative ordinali** che possono avere un criterio di ordinamento ed un ordinamento gerarchico (es. grado d'istruzione).
- Le variabili **quantitative** derivano da valori numerici e da un'operazione di misurazione o da un'operazione di conteggio.

Pertanto, **le variabili qualitative nominali** sono quelle che **daranno meno informazioni**, **le variabili qualitative ordinali daranno un numero maggiore di informazioni**.

Le variabili quantitative permetteranno di effettuare delle valutazioni più precise, prendendo in considerazione anche delle misurazioni.

Rappresentazione dei dati

Distribuzioni di Frequenza

I dati sono raggruppati in base alle diverse modalità osservate ed ad ognuna di queste è associata la corrispondente frequenza. Si utilizza generalmente per variabili qualitative e quantitative discrete.

Esempio

Unità	Valore
1	Modalità 1
2	Modalità 1
3	Modalità 2
4	Modalità 1
5	Modalità 2
6	Modalità 3
7	Modalità 3

È possibile stabilire una scala gerarchica tra le variabili e raggrupparle in una tabella di dati di frequenza per osservare le modalità.

Per poter eseguire un lavoro di sintesi si possono rappresentare in forma tabellare o grafica i dati posseduti. Il primo passo è utilizzare una **distribuzione di frequenza in classi**: raggruppare in intervalli consecutivi e disgiunti i valori osservati.

Rappresentazioni Tabellari

Subject	Gender	Age	Adverse event (AE)	Day to AE
1	Male	36	Anemia	22
2	Female	40	Hypokalaemia	6
3	Male	36	Abdominal Distension	7
4	Male	38	Abdominal Pain	16
5	Female	38	Vomiting	8
6	Male	35	Dysphagia	7
7	Female	39	Fatigue	16
8	Female	40	Vomiting	46
9	Female	40	Fatigue	41
10	Male	39	Fatigue	40
11	Male	35	Vomiting	40
12	Male	39	Vomiting	30
13	Female	37	Fatigue	32
14	Male	37	Vomiting	47
15	Female	36	Fatigue	9
16	Male	36	Vomiting	38
17	Female	36	Fatigue	35
18	Female	36	Fatigue	47
19	Female	38	Fatigue	1

In questa tabella sono presenti 4 variabili, di cui 2 qualitative nominali e 2 quantitative, in cui sono stati tabellati i campioni, permettendo di confrontarli, sebbene in questo caso non è stato fatto un lavoro di sintesi.

Rappresentazione Tabellari

Spesso, nel caso di variabili quantitative continue, l'impiego di distribuzioni di frequenza non è utile a causa dell'elevato numero di differenti modalità che possono essere osservate. In questo caso si ricorre ad una distribuzione di frequenza in classi. Essa si ottiene raggruppando in intervalli consecutivi e disgiunti (classi) i valori osservati e associando a ciascuna di questi la corrispondente frequenza

Day to AE	Frequenza
1	1
6	1
7	2
8	1
...	...
40	2
41	1
46	1
47	2
Totale	19

Day to AE	Frequenza
0 -| 15	6
15 -| 30	4
30 -| 60	9
Totale	19

La scelta delle classi può avvenire utilizzando criteri semi-automatici (classi equi-ampie, classi equi-frequenti) o ricorrendo a conoscenza esperta

Nel caso in cui si abbia a che fare con **variabili quantitative continue**, **l'impiego di distribuzione di frequenza non è utile** a **causa dell'elevato numero di differenti modalità** che possono essere osservate. In questo caso, si ricorre ad una **distribuzione di frequenza in classi**. Essa **si ottiene raggruppando in intervalli consecutivi e disgiunti** (classi) **i valori osservati e associando a ciascuna di queste la corrispondente frequenza**. La scelta delle classi può avvenire utilizzando criteri semi-automatici o ricorrendo a conoscenza esperta.

Per descrivere il comportamento congiunto di più variabili si utilizzeranno le **tabelle di contingenza (**tabella dei dati di frequenza) che consente di incrociare le distribuzioni di frequenza relative a due o più variabili ed evidenziarne relazioni e associazioni, e può essere utilizzata con un numero indefinito di righe e colonne. Le frequenze all'interno sono quelle **congiunte**, mentre i dati posti all'esterno si definiscono **marginali**.

Rappresentazioni Tabellari

Tabelle di contingenza

La tabelle di contingenza consentono di incrociare le distribuzioni di frequenza relative a due o più variabili allo scopo di evidenziare la presenza di relazioni e associazioni

Unità	Variabile X	Variabile Y
1	x_1	y_1
2	x_1	y_2
3	x_2	y_2
4	x_1	y_2
5	x_2	y_2
6	x_2	y_1
7	x_2	y_1

	x_1	x_2	Totale
y_1	1	2	3
y_2	2	2	4
Totale	3	4	7

I totali di riga e di colonna sono invece le frequenze marginali. Esse indicano il numero di soggetti che presentano una specifica modalità di una variabile indipendentemente da quanto accade per l'altra variabile considerata

Rappresentazioni tabellari

Qualora si debbano confrontare collettivi di numerosità differenti è opportuno ricorrere a frequenze relative o percentuali.

Maschi

Nr. di eventi avversi	Freq. Ass.
0	74
1	123
2	85
3	32
>3	18
Totale	332

Femmine

Nr. di eventi avversi	Freq. Ass.
0	26
1	62
2	23
3	12
>3	4
Totale	127

I maschi sono meno soggetti a eventi avversi???

Maschi

Nr. di eventi avversi	Freq. Rel.
0	0.20
1	0.49
2	0.18
3	0.10
>3	0.03
Totale	1

Femmine

Nr. di eventi avversi	Freq. Rel.
0	0.22
1	0.37
2	0.26
3	0.10
>3	0.05
Totale	1

Le frequenze relative (percentuali), ottenute dividendo le frequenze assolute per la numerosità del collettivo esaminato (moltiplicando poi il risultato per 100), consentono di annullare l'effetto della diversa numerosità poiché esprimono ciascuna frequenza come quota di uno stesso totale (1 nel caso di frequenze relative e 100 nel caso di frequenze percentuali)

Le frequenze prese finora in considerazione sono quelle assolute che, però, possono dare un'immagine distorta dell'evento. Pertanto, è necessario ricorrere a **frequenze relative o percentuali,** qualora si debbano confrontare collettivi di numerosità differenti. Ad esempio, analizzando una frequenza assoluta (prima riga) si avrebbe un dato distorto poiché il numero degli eventi avversi maschili (74) risultano essere maggiori dei femminili (26): bisognerà analizzare, per questo, le frequenze relative (frequenze assolute per la numerosità del collettivo esaminato), avendo così una visione reale del fenomeno.

Proporzione

Nel caso di variabili qualitative, l'unica operazione possibile è il conteggio *n* delle occorrenze di una o più specifiche modalità in un collettivo di *N* unità statistiche. Questo conteggio (frequenza assoluta) non consente però confronti utili quando i collettivi hanno numerosità differenti.

Il secondo passo per raccogliere dati è usare gli **indicatori**. Si tratta di operazioni sintetiche sulle modalità individuali delle variabili che le sintetizzano in un unico valore numerico (ancora più sintetici delle tabelle) e si avrà, così, la massima sintesi del fenomeno poiché indicato da un unico valore (es. media aritmetica).

L'indicatore più utilizzato per descrivere dati categorici è la Proporzione (π) data dal rapporto tra il valore di n e la dimensione del collettivo N.

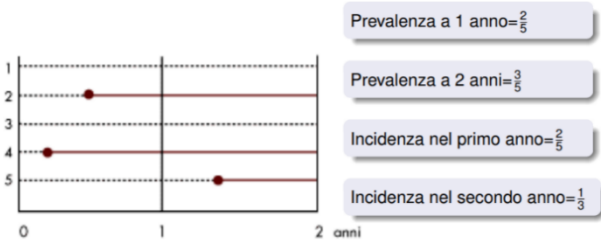
$$\pi = \frac{n}{N}$$

La Proporzione è un *rapporto di parte al tutto* ed è quindi, per costruzione, un numero compreso tra 0 e 1

(valore collettivo N = quante volte la caratteristica è stata contata rapportata al totale).

Un esempio di indicatore sintetico è la **prevalenza** di una patologia = numero di soggetti con la patologia al tempo t / la popolazione totale al tempo t. (es. soggetti diabetici 31 dic)
Un altro esempio di indicatore sintetico è **l'incidenza cumulativa** = numero di malattie in un periodo di tempo / popolazione a rischio al tempo zero (non fa riferimento ad un preciso istante temporale ma ad un orizzonte temporale). (es. soggetti ad una malattia nell'anno 2020).
(Sono entrambe proporzioni poiché date dal rapporto della parte per il tutto).

Esempi di proporzione



Analizzando questo esempio possiamo concludere di non poter mai ottenere un valore maggiore di 1, poiché al massimo il numeratore ha un valore uguale al denominatore.

STUDI STATISTICI

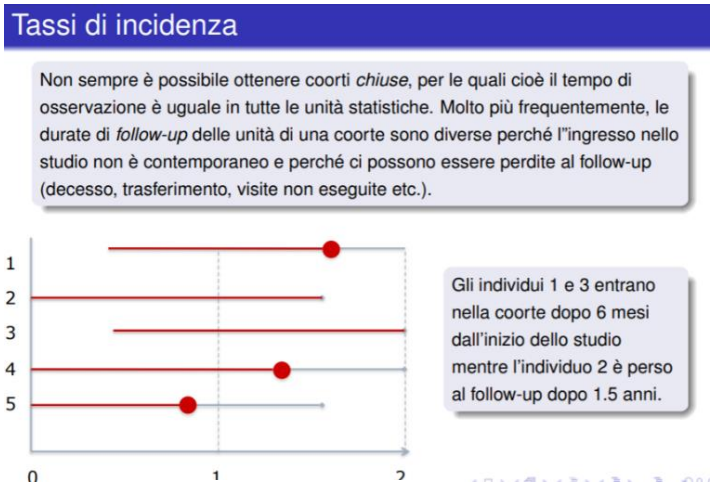
Gli studi che si conducono in ambito statistico si definiscono **studi di corte** (e sono studi epidemiologici) in cui **vengono arruolati gruppo di soggetti che si seguono prospetticamente nel tempo per vedere in quanti di loro sorge quella data patologia**. Questi studi di corte possono essere suddivisi in:

- **studi di corte prospettici**: si analizza la corte in un intervallo di tempo (si analizzano al momento e si studiano i cambiamenti che hanno nel futuro);
- **studi di corte retrospettivi:** si analizza la situazione che è già avvenuta nel passato e tendono a studiare l'incidenza cumulativa.

Un'altra tipologia di studi sono gli studi **cross sectional** (trasversali), in cui si valuta in un istante temporale chi possiede e chi non possiede una data caratteristica e tendono a studiare la prevalenza di un fenomeno.

Un altro indicatore in un fenomeno qualitativo è il **TASSO D'INCIDENZA COMULATIVA** applicabile quando i soggetti a rischio sono stati seguiti nello stesso intervallo di tempo.

Qualora in cui **l'individuo riesca ad entrare o uscire dallo studio** si parla di **corti aperte**, mentre si parla di **corti chiuse** quando **l'osservazione è uguale in tutte le unità statistiche.**



Nel caso di corti chiuse si considera il **TASSO D'INCIDENZA,** che tiene conto della partecipazione della persona allo studio.

Tasso d'incidenza = n di nuove diagnosi / tempo persona

NB. tempo persona (person time) = tempo trascorso da ciascuna persona nella popolazione.

In ambito diagnostico si utilizzano: la **sensibilità** (soggetti che presentano l'infezione / pazienti sottoposti al test diagnostico) (Es. uno strumento che dà come risultato 80/100 significa che 80 persone sono soggette a malattia e 20 persone sono falsi negativi) e la **specificità**: procedura diagnostica capace di far emergere l'assenza della malattia. (Es se abbiamo un rapporto 95/100 avremo 95 positivi e 5 falsi positivi, ovvero, individui non affetti da malattia ma considerati come malati).

LEZIONE 2.

Le **rappresentazioni grafiche rappresentano uno strumento importante nella descrizione di un'informazione**, ma, nonostante ciò, devono essere usati con cautela perché si rischia di confondere il lettore piuttosto che aiutarlo a comprendere il fenomeno che si sta descrivendo.
Questo vuol dire che nonostante le rappresentazioni grafiche siano molto numerose, in realtà quelle che si utilizzano in un contesto scientifico sono molto limitate: **nel caso di variabili qualitative**, lo strumento che generalmente si costruisce prevede la **costruzione di una rappresentazioni a torta o, preferibilmente, a barra**.

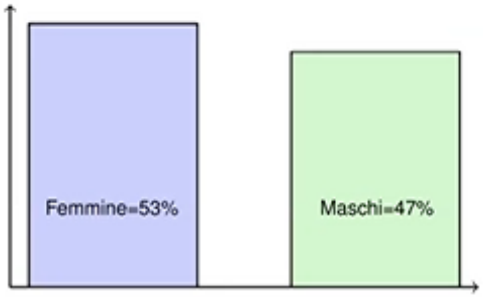


Diagramma a *Barre*

Le **rappresentazioni a barra** sono rappresentazioni in cui le diverse **modalità di una variabile** (variabili qualitative in cui le modalità sono etichette linguistiche) **sono descritte da barre**, da delle colonne che poggiano su un asse che può essere quello delle ascisse (nel caso in cui le barre siano orizzontali) o delle ordinate (nel caso in cui le barre siano verticali). Il numero di barre è pari al numero di modalità che caratterizza quella variabile e l'altezza è pari alla frequenza relativa con cui ciascuna modalità è stata osservata.

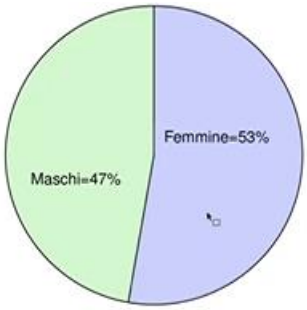


Diagramma a *Torta*

L'alternativa, anche se non molto utilizzata a livello scientifico, è un **diagramma a torta in cui una superfice circolare viene suddivisa in tanti spicchi quanti sono le modalità della variabile** e ciascuno spicchio ha una superfice che è pari alla frequenza relativa di quella particolare modalità.

Nel caso del diagramma a barre è chiaro che l'ordine con cui si rappresentano le modalità è assolutamente arbitrario, poiché non vi è informazione metrica. L'unica cosa da dover rispettare è l'altezza, perché deve essere proprio pari alla frequenza relativa dell'evento.

Per quanto riguarda, invece, **le variabili numeriche**, il discorso è un po' particolare: premesso che nelle variabili numeriche le modalità della variabile sono quantitative, gli aspetti che bisogna descrivere per poter caratterizzare il fenomeno sono molteplici e quindi sono molteplici le classi e le tipologie di indicatori che si andranno a costruire.
Nello specifico, sono **tre gli aspetti** che interessano per poter descrivere una variabile numerica:

1. **La posizione**
2. **La variabilità**
3. **La forma**

Studiare la **posizione**, o eventualmente l'intensità di una variabile vuol dire **descrivere e capire quale è**, attraverso un indicatore sintetico**, l'ordine di grandezza con cui quel fenomeno si è presentato nel collettivo di unità statistiche che si è arruolato**. Esistono diversi indicatori di posizione e tra quelli più utilizzati bisogna ricordare: **la media aritmetica e la mediana.**

La **media aritmetica è la somma di tutte le modalità diviso il numero di termini sommati**, ed è perciò una sommatoria **con indice che va da un range con Xi** (x con i), dove in Xi si identifica la generica manifestazione della variabile X nella ***iesima unità statistica*** (per evitare di doversi riferire ad un'unità statistica specifica), tutto diviso n. La media viene indicata col simbolo dell'alfabeto greco "mi".
È importante sottolineare che si tratta di media aritmetica, perché il concetto di media è un concetto più ampio che tocca anche strade diverse, come la media geometrica.