Lezione 1, 6/10/2020 - Introduzione al linguaggio statistico

Introduzione

Per poter comprende il concetto di statistica, i suoi utilizzi ed aspetti fondamentali ci possiamo rifare a due principali autori, che diedero una definizione del termine "statistica". Le due definizioni di Bland ed Armitage sottolineano infatti alcuni aspetti fondamentali della statistica.

- **Bland:** la statistica è quella scienza che consiste nella sintesi ed interpretazione dei dati numerici.
- **Armitage:** la statistica è una disciplina che riguarda il trattamento dei dati numerici, che derivano da gruppi di unità statistica.

Quello che si comprende da queste definizioni è che la statistica procede per *sintesi*. Nella statistica gli individui che vengono arruolati sono fondamentali, perché rappresentano strumento attraverso il quale possiamo dare una definizione di collettivi. Questo è il punto fondamentale: caratterizzare i collettivi, i gruppi di individui. Perché le caratteristiche individuali sono caratterizzate da una variabilità che non deve far parte del processo di sintesi, in quanto legato troppo allo specifico individuo. Invece è fondamentale attraverso gli studi clinici giungere a una conclusione. Per esempio se si deve confrontare l'efficacia di due differenti trattamenti andare a vedere le singole reazioni degli individui significherebbe andare a perdere il punto di vista generale. Potrebbero esserci alcuni individui che reagiscono meglio a un farmaco X e altri individui che reagiscono meglio ad un altro tipo di farmaco. L'unico modo per giungere ad una conclusione è ragionare dunque sui collettivi e non sulle singole osservazioni.

Vocabolario essenziale

<u>Unità statistica</u>

Si definisce tale l'unità elementare su cui vengono osservati i **caratteri oggetti di studio** (persone o esseri viventi, oggetti, territori, tempi...). Quindi sono i pazienti che entrano a far parte del nostro studio. Questo però non è valido per tutti i campi, ad esempio in odontoiatria l'unità statistica è il dente. Possono essere anche qualcosa di **astratto**. Nelle indagini sociali per esempio, il censimento dell'Istat, l'unità statistica è la famiglia. Dunque, *l'unità statistica è l'elemento su cui si andrà a misurare e a rilevare le caratteristiche che sono oggetto dell'indagine.*

<u>Popolazione</u>

È l'insieme di tutte le unità statistiche omogenee su cui si vuole condurre la rilevazione. La popolazione rappresenta dunque l'obiettivo dello studio.
Il concetto di popolazione ci introduce un primo problema. Si pensi a uno studio che vuole valutare l'efficacia di un nuovo trattamento per ridurre il colesterolo, è chiaro che la popolazione è rappresentata da soggetti affetti da ipercolesterolemia. Questa popolazione è indefinita nel tempo e nello spazio, non è individuabile, in quanto non esistono i confini temporali e territoriali. Quindi molto spesso le popolazioni in ambito medico non sono popolazioni chiuse, ma indefinite che in linea di principio contengono infinite unità statistiche.
Nelle scienze sociali è diverso. Se si vuole fare l'indagine sul possesso o meno per il possesso di una residenza, la popolazione è definita. Tutti i nuclei familiare in un determinato istante sono residenti in un certo territorio. In ambito medico questo non si può attuare.

L'obiettivo è descrivere una popolazione, ma in linea di principio è impossibile rilevare le unità statistiche. quindi nella maggior parte dei casi non si potranno raccogliere i dati di tutta la popolazione, ma si raccoglierà solo una quota (un campione). Questo significa che l'analisi statistica deve essere scomposta in due momenti diversi: un **momento descrittivo** ed un **momento inferenziale.**

Il **momento descrittivo** si occupa di descrivere, utilizzando gli strumenti più opportuni, le caratteristiche del collettivo di unità statistiche che sono oggetto della rilevazione. Forniscono una fotografia di quello che ho davanti agli occhi. È un momento importante, ma non conclusivo. Questo perché se, ad esempio, l'obiettivo di uno studio è valutare l'efficacia di un farmaco per ridurre il colesterolo (ritorniamo all'esempio precedente) e negli individui arruolati nella sperimentazione questo farmaco è efficace, non si può fare una generalizzazione sulla popolazione. Bisogna comprendere se l'efficacia del trattamento osservata in una quota limitata di unità statistica, può essere generalizzata all'intera popolazione, da cui sono state selezionate le unità statistiche. Quest'ultimo è il **momento inferenziale**.

L'inferenza statistica si occupa di *trasferire l'evidenza raccolta su un campione sull'intera popolazione*. È un momento fondamentale in cui le unità statistiche analizzate non esauriscono l'intera popolazione. Il **censimento** rappresenta il momento in cui il momento inferenziale è assente. È l'unica indagine di popolazione, l'obiettivo è descrivere la popolazione residente in Italia in un dato momento. L'inferenza quindi ci permette di comprendere se l'informazione raccolta su un campione, si può generalizzare sulla popolazione. *L'inferenza non può negare i dati raccolti sul campione.*

Variabile

Si intende qualunque fenomeno oggetto di studio che è rilevato sulle unità statistiche e che è suscettibile di assumere valori differenti nelle differenti unità statistiche. *Quindi i fenomeni nello studio non sono costanti. Il* motivo per cui certe caratteristiche sono diverse tra gli individui è dovuto in parte ad elementi che non fanno parte del processo di sintesi (variabilità inter ed intra individuale che non è l'obiettivo dello studio. Si pensi alla misurazione della pressione di un individuo, questa può variare nell'arco di dieci minuti.) ed in parte perché assumono valori differenti a secondo delle unità statistiche su cui vengono rilevate.

I valori assunti da ciascuna variabile nelle singole unità statistiche prendono il nome di **modalità**. Le variabili si possono suddividere in due classi: qualitative e quantitative.

<u>Variabili qualitative</u>

Le modalità non derivano da un'operazione di misurazione e nemmeno da un'operazione di conteggio. Sono qualità di cui si rileva la presenza o assenza.
Esempi:

- genere (maschile o femminile)
- intensità del dolore
- grado di istruzione

Le opzioni di analisi dati saranno molto limitate rispetto a quelle che si avranno a disposizione per le informazioni quantitative. Le variabili qualitative a loro volta possono essere distinte in **nominali e ordinali**. Sono differenti per il contenuto informativo. Le **variabili nominali** come il genere o l'insorgenza di un evento avverso, nel momento in cui si confronta la variabile su due unità statistiche è possibile dire se le due unità sono uguali o diverse rispetto a questa variabile. Quindi uno è maschio e l'altro è femmina, uno ha avuto un evento avverso e l'altro no. Quindi il confronto si svolge in *termini di uguaglianza e disuguaglianza.*

Differentemente nella **variabili ordinali**, i dati raccolti possono essere ordinati. Ad esempio, se si sta analizzando la variabile "*intensità del dolore*", oltre a poter dire se due unità statistiche sono uguali o diverse, si possono ordinare.
<u>In conclusione:</u>
Una variabile è nominale se l'unica relazione che si può instaurare tra due unità statistiche è di uguaglianza o disuguaglianza. Una variabile è ordinale se tra due unità statistiche è possibile porre un ordine, pur restando sempre qualitative.

Variabili quantitative

Le variabili derivano da un'operazione di misurazione o di conteggio. Le variabili assumono come modalità valori numerici con preciso significato e unità di misura.

Esempi:

- altezza
- potenza di un vaccino
- BMI
- Contenuto del principio attivo di un farmaco

Alle variabili (qualitative e quantitative) è possibile associare dei valori numerici in un foglio di lavoro, ad esempio è possibile associare lo 0 al genere maschile e l'1 al genere femminile, convertendo una variabile qualitativa in quantitativa.

Rappresentazione dei dati

La prima cosa da fare, una volta che si sono raccolte una serie di informazioni sulle unità statistiche e si vuole comprendere cosa le caratterizza, è contare il numero di volte in cui le diverse modalità di ciascuna delle variabili analizzate sono state osservate.

Questo è possibili tramiate la costruzione di rappresentazioni tabellari di frequenza.

Rappresentazioni tabellari: distribuzioni di frequenza

Una tabella di frequenza è una struttura attraverso la quale si individuano le distinte modalità osservate e ad ognuna di queste è associata la corrispondente frequenza (cioè il numero di volte che è stata osservata). Si utilizza generalmente per le variabili qualitative e quantitative discrete.

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

Modalità	Frequenza
Modalità 1	3
Modalità 2	2
Modalità 3	2
Totale	7

Quindi questo è l'elenco che presenta in colonna le unità statistiche e associata a ciascuna unità vi è la modalità. La rappresentazione della frequenza (tabella a destra) è una forma compatta delle modalità (non si riportano tutte le unità) e accanto si indica il numero di volte con cui si presenta quella modalità.

Esempio

Sono state raccolte informazioni su 19 soggetti che hanno partecipato ad una sperimentazione, di cui sono state rilevate alcune caratteristiche: genere, età, insorgenza evento avverso ed i giorni dall'inizio del trattamento. Quindi ci sono 4 variabili, 19 unità statistiche. Delle 4 variabili, 2 sono quantitative (età e tempo dall'evento avverso) e qualitative nominali (genere e tipologia dell'evento avverso).

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

Vi è una tabella che riporta la frequenza delle variabili in funzione del **genere**. Si può fare anche per le altre. In questo modo la sintesi inizia a perdersi perché distinte modalità diventano numerose. Diventano ancora più numerose nel momento in cui si utilizza la prima forma di sintesi (la distribuzione di frequenza) in cui le modalità sono differenti fra individui.

Genere	Frequenza
Male	9
Female	10
Totale	19

In queste situazioni non si ha un vantaggio di sintesi, quindi si procede con una distribuzione di frequenza in classi. Sono riportate le tabelle relative alla frequenza per **tipologia di evento avverso** e per **età**.

Day to AE	Frequenza
1	1
6	1
7	2
8	1
9	1
16	2
22	1
30	1
32	1
35	1
38	1
40	2
41	1
46	1
47	2
Totale	19

Età	Frequenza
35	2
36	6
37	2
38	3
39	3
40	3
Totale	19

Distribuzione di frequenza in classi

Si suddivide il campo di variazione (detto anche *range.* Si intende l'intervallo di valori che va dal più piccolo al più grande dei valori osservati) in intervalli disgiunti e consecutivi, ad ognuno di questi si associa il numero di unità statistiche la cui modalità rientra all'interno di questo intervallo.

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

In questo caso sono stati utilizzate tre classi con ampiezza differente e ad ognuna delle classi sono associate il numero di unità statistiche. Il fatto di usare classi di ampiezza differente è molto frequente. Valori molto grandi in una variabile caratterizzano poche unità statistiche, questo accade in molti fenomeni economici e medici. Ad esempio: si pensi al BMI, valori molto elevati in una popolazione sono poco frequenti. Fare classi che mantengono un'ampiezza iniziale piccola in modo da evitare un'eccessiva perdita di informazione, ma conservare quest'ampiezza fino alla fine non ha senso, perché se ci sono classi di BMI di 2 o 3 unità di BMI quando ci si sposta su BMI elevate non ci sarà nessuno in quelle classi. Quindi conviene allargare l'ampiezza delle classi in modo tale da inglobare nelle classi più estreme a destra un numero sufficiente di unità statistiche.

Una volta che abbiamo contato il numero di volte in cui una diversa modalità è stata osservata, incominciamo ad interessarci allo studio delle relazioni tra due variabili. Importante nell'ambito medico, come l'insorgenza di una patologia in un individuo. Quindi si lavorerà su tabelle in cui sono presenti coppie variabili: **tabelle di contingenza.**

Le 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.

In queste tabelle le modalità della variabile y sono rappresentate per riga e le modalità della variabile x sono rappresentate per colonna e in corrispondenza di ogni incrocio fra riga e colonna, si va a contare il numero di unità statistiche con quella particolare combinazione di modalità.

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

La frequenze interne sono dette frequenze congiunte dal momento che indicano il numero di unità statistiche che congiuntamente presentano quelle specifiche modalità di riga e di colonna

Quindi se vediamo questa struttura dati riportata a sinistra e la sua rappresentazione sotto forma di tabella di contingenza (destra), si nota che vi è un solo individuo che presenta la modalità x1 della variabile x e la modalità y1 della variabile y.

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

Ai margini vi sono le **distribuzioni di frequenza marginali.** Il numero 3 rappresenta che ci sono, complessivamente ed indipendentemente da quello che si è osservato per questa variabile, 3 individui con la prima modalità della variabile y.

Quindi all'interno di una tabella di contingenza si individueranno le **frequenze congiunte**, che valutano congiuntamente la contemporanea presenza delle modalità a cui si sta facendo riferimento. Analizzare due o più caratteristiche in contemporanea permette di iniziare a stabilire se vi è una relazione tra queste.

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

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

I maschi sono meno soggetti a eventi avversi???

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)

È chiaro che sia possibile costruire delle distribuzioni di frequenza congiunte anche nel caso in cui una delle due variabili sia stata precedentemente divisa in classi. In questo caso viene analizzata la relazione genere/tempo di insorgenza di eventi avversi, quest'ultimo suddiviso in 3 classi. Nei primi 15 giorni dunque 6 individui hanno presentato AE di cui 4 maschi.