Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 2

Piano della presentazione

1. Statistica descrittiva e inferenziale
2. Terminologia
3. Parametri e statistiche
4. Statistica Univariata, bivariata e multivariata
5. Variabili e modalità
6. Scale di modalità
7. Esperimenti e osservazioni
8. Conclusioni

Statistica descrittiva e inferenziale

- La Statistica riguarda i **metodi scientifici per raccogliere, ordinare, riassumere e analizzare** informazioni riguardanti i fenomeni collettivi, ed anche per trarre valide conclusioni e prendere **ragionevoli decisioni** sulla base di tali analisi.
- Generalmente i **fenomeni collettivi** sono tali perché sono relativi ad una collettività di casi singoli (le cosiddette unità statistiche). Quindi sono fenomeni collettivi quelli che riguardano una popolazione.

Statistica descrittiva e inferenziale

- In tutti i problemi di statistica ci si trova di fronte ad una massa (grande o piccola) di dati che sono raccolti, classificati ed elaborati in vista di certi scopi.
- Questi possono essere diversi da problema a problema e si distinguono nella teoria della statistica due rami principali che rispondono ai nomi di:
  1. statistica descrittiva
  2. statistica inferenziale

Statistica descrittiva

- La statistica descrittiva fornisce gli strumenti per rappresentare, sintetizzare ed interpretare il modo in cui un fenomeno di interesse si è manifestato nel collettivo osservato.

  . . . quanti sono, in uno specifico campione casuale di studenti di Psicologia, gli studenti provenienti da un liceo scientifico, quelli provenienti da un liceo classico, istituto tecnico e così via?

- Sarà sufficiente considerare studente per studente, classificarlo secondo la scuola di provenienza e riassumere i dati ottenuti in una tabella.

- Questo prospetto ha però un significato molto ristretto in quanto vale solamente per quel particolare campione.

In altre parole, i dati così raccolti servono solamente a descrivere certe caratteristiche degli studenti che sono stati osservati, non degli studenti di Psicologia in generale.

Statistica inferenziale

- La statistica inferenziale permette con le sue metodologie basate sul calcolo delle probabilità di estendere (mediante logica induttiva) all'intera popolazione di riferimento, i risultati e le interpretazioni effettuate sulla base di una rilevazione parziale.
- Da una certa popolazione, i cui caratteri sono ignoti, si estrae un campione casuale, per esempio, e in base ai dati che esso fornisce si vuol provare (cioè accettare o respingere) una certa ipotesi, che riguarda la popolazione.

nota terminologica

Popolazione	(*di riferimento*) insieme ben definito di oggetti di interesse per l'indagine;
reale	(es., i cittadini italiani di età compresa tra i 18 e i 24 anni)
ipotetica	(es., tutti gli individui che, sotto certe condizioni, potrebbero esibire degli stati d'ansia).
Campione	sottoinsieme della popolazione oggetto della rilevazione.
Unità	(*statistiche*) elementi della popolazione di riferimento.
Dati	risultato della rilevazione di caratteristiche delle unità statistiche.
Variabili	caratteristiche rilevate sulle unità statistiche.
Modalità	valori distinti assunti da una variabile.

Parametri e statistiche

- Un parametro è un numero che descrive un qualche aspetto della popolazione.
- Il reddito italiano medio nel Nord - Est μ è un parametro. Per esempio, $\mu = €20.800$.
- In qualsiasi situazione concreta, i parametri sono sconosciuti.

Parametri e statistiche

- Una statistica è un numero che può essere calcolato utilizzando i dati forniti da un campione, senza alcuna conoscenza dei parametri della popolazione.
- Supponendo che il reddito medio di un *campione casuale* di $n = 100$ famiglie del Nord - Est è pari ad €19.800,
- la media campionaria $\bar{x} =$ €19.800 è una statistica.

Statistica Univariata, bivariata e multivariata

- Si ha un problema di statistica descrittiva quando i dati raccolti si considerano come un ente a sè, isolati quindi da altri analoghi che non sono stati raccolti.
- ciascuna unità statistica può essere analizzata facendo riferimento
  - ad una sola variabile: si parlerà allora di statistica descrittiva univariata;
  - a due variabili: si parlerà allora di statistica descrittiva bivariata;
  - a tre o più variabili: si parlerà allora di statistica descrittiva multivariata.

Variabili e modalità

- Se si considera una sola variabile sarà necessario analizzare le manifestazioni di quella variabile nelle varie unità statistiche.
- Tali manifestazioni prendono il nome di modalità

Variabile

- **Variabile**: caratteristica delle unità statistiche che al variare delle unità (statistiche) può assumere almeno due valori.
- Notazione: $X; Y; Z$.
- Esempio. Per le unità statistiche corrispondenti agli studenti universitari, si possono analizzare variabili quali $X =$ genere, Y età, $X =$ voto all'esame, ...

Modalità

- **Modalità**: valori assumibili da una variabile (generalmente noti a priori). L'insieme di tali valori è detto insieme delle modalità della variabile.
- Notazione: $x; y; z$.
- Esempio: $x = \{m, f\}$, $y = \{19, ..., 50\}$, $z = \{ins., 18/30, 19/30, ..., 30/30\}$., ...

scale di modalità

Nell'analisi di una variabile è necessario predisporre quella che è chiamata scala delle modalità, ovvero l'insieme delle modalità che saranno adottate per descrivere le possibili manifestazioni della variabile considerata.

Scale di modalità

Le scale delle modalità possono essere:

nominali quando si articolano in modalità suscettibili di classificare le unità statistiche;

ordinali quando si articolano in modalità suscettibili di un ordinamento;

intervallari quando si articolano in modalità che si identificano con numeri aventi la prerogativa di quantificare l'intensità della manifestazione di una variabile;

di rapporto quando si articolano in modalità caratterizzate dall'avere una origine fissa che risulta in via naturale lo zero.

Fenomeni quantitativi o qualitativi

Se si effettua una ricerca di statistica descrittiva univariata considerando quindi solamente una variabile si avrà che:

- se la rilevazione della variabile viene effettuata su scala di modalità di tipo nominale o ordinale si parlerà di fenomeno di tipo qualitativo;
- se la rilevazione della variabile viene effettuata su scala di modalità di tipo intervallare o di rapporto si parlerà di fenomeno di tipo quantitativo.

Scala nominale

- Nella scala nominale esiste una sola relazione, quella di identità: le unità statistiche (u.s.) attribuite a classi diverse sono tra loro differenti, mentre tutte quelle della stessa classe sono tra loro equivalenti, rispetto alla proprietà utilizzata nella classificazione.
- L'unica operazione aritmetica ammessa è il conteggio delle u.s. presenti in ogni categoria. I quesiti statistici che possono essere posti correttamente riguardano le frequenze, sia assolute che relative.

Scala nominale

- Sono possibili confronti tra frequenze osservate.

  "Una classe è significativamente più numerosa dell'altra?"

  "Le varie classi hanno tutte lo stesso numero di u.s., escludendo le variazioni casuali?"

- oppure tra le frequenze osservate e le rispettive frequenze attese sulla base di leggi, ipotesi od altro.

  "I risultati ottenuti da un esperimento sono in accordo con la sua distribuzione teorica?"

Scala ordinale o per ranghi

- La scala ordinale o *per ranghi* assume modalità logicamente sequenziali.
- Alla proprietà precedente di equivalenza tra le u.s. della stessa classe si aggiunge una gradazione tra le classi o tra le u.s. di classi differenti

Scala ordinale o per ranghi

- Con la scala nominale, si ha la sola informazione che le u.s. appartenenti a classi differenti sono tra loro diverse, ma non è possibile stabilire un ordine. Con la scala per ranghi, le differenti classi possono essere ordinate sulla base dell'intensità del fenomeno.
- In una scala ordinale, però, non è possibile quantificare le differenze di intensità tra le osservazioni.

Scala ad intervalli

- La scala ad intervalli aggiunge la proprietà di misurare le distanze o differenze tra tutte le coppie di u.s. nei termini di una misura costante, detta unità di misura.
  - Esempio di scala ad intervalli è la temperatura misurata in gradi Celsius o Fahrenheit, ma non Kelvin.
  - Valori (*in gradi*) di temperatura, oltre a poter essere ordinati secondo l'intensità del fenomeno, godono della proprietà che le differenze tra loro sono direttamente confrontabili e quantificabili.

Scala ad intervalli

- Il limite della scala ad intervalli è quello di non consentire il calcolo del rapporto tra coppie di misure.

Ad esempio, una temperatura di 80 gradi Celsius non è il doppio di una di 40 gradi. Se infatti esprimiamo le stesse temperature nei termini della scala Fahrenheit, allora i due valori non saranno in rapporto di 1 a 2 tra loro

Scala ad intervalli

- In una scala ad intervalli solo le differenze tra i valori sono quantità isomorfiche al fenomeno considerato.
- Solo per le differenze sono permesse tutte le operazioni aritmetiche: possono essere tra loro sommate, elevate a potenza oppure divise, determinando le quantità che stanno alla base della statistica parametrica.

scala di rapporti

- La scala di rapporti è caratterizzata dal possedere un'origine non arbitraria: lo 0 significa intensità nulla della caratteristica.
  - Sono tipiche scale di rapporti l'altezza, la distanza, la velocità, l'età, il peso, il reddito, la temperatura in gradi Kelvin.
- Non solo le differenze, ma gli stessi valori della scala possono essere moltiplicati o divisi per quantità costanti senza alterare la relazione isomorfica esistente tra i valori della scala, le differenze tra i valori, i rapporti tra i valori, da una parte, e le intensità del fenomeno oggetto di indagine, dall'altra.

Relazioni tra u.s. nelle scale di misurazione

	Fenomeno			
	qualitativo		quantitativo	
	nominale	ordinale	intervallare	di rapporto
uguaglianza	sì	sì	sì	sì
ordine	no	sì	sì	sì
intervallo unitario	no	no	sì	sì
zero assoluto	no	no	no	sì

Operazioni sulle modalità

scale di modalità	operazioni aritmetiche
nominale	enumerare le classi di equivalenza e/o
	le frequenze per ciascuna classe di equivalenza
ordinale	enumerare le classi di equivalenza e/o
	le frequenze per ciascuna classe di equivalenza
intervalli unitario	differenze (rapporti tra differenze)
di rapporti	rapporti diretti tra le misure

Studi osservazioni

La validità delle inferenze statistiche dipende dall metodo che è stato utilizzato per raccogliere i dati.

- Negli studi osservazionali, il ricercatore raccoglie dei dati che sono 'naturalmente' prodotti dai fenomeni studiati, senza manipolare il valore di alcuna variabile.
  - Auto-selezione degli individui al trattamento.
  - Differenze sistematiche nelle risposte dovute al trattamento e/o altri fattori non controllati.

Studi sperimentali

- Negli studi sperimentali, il valore delle variabili "indipendenti" è direttamente manipolato dallo sperimentatore.
  - Assegnazione casuale delle u.s. al trattamento.
  - Fonti di variabilità sotto controllo.
  - Differenze sistematiche nelle risposte dovute al trattamento.

Inferenze

- Negli studi osservazionali le inferenze di natura causale sono problematiche.
  - La relazione tra una variabile esplicativa (indipendente) e una variabile risposta (dipendente) può essere prodotta da una variabile interveniente di cui il ricercatore non è consapevole.

Inferenze

- Negli esperimenti programmati le variabili indipendenti sono sotto il diretto controllo dello sperimentatore.
  - In un esperimento adeguatamente pianificato le inferenze causali sono molto meno problematiche che nel caso degli studi osservativi.
  - Sfortunatamente, molti fenomeni psicologici non possono essere studiati in laboratorio.

Domande degli studi osservazioni

- Il focus di uno studio osservazionale è solitamente quello di analizzare la distribuzione di una variabile oppure la relazione tra due o più variabili in una popolazione di individui. Per esempio,

  . . . quanti cittadini italiani esibiscono sintomi depressivi di gravità o durata sufficienti a giustificare una diagnosi di depressione?

  Che relazione c'è tra genere e depressione? – La depressione si manifesta con maggior frequenza tra i maschi o tra le femmine?

Domande degli studi osservazioni

- Per ragioni facilmente comprensibili, solitamente non è possibile raccogliere i dati relativi a tutti gli individui della popolazione.
  - Questo problema si risolve selezionando un numero relativamente piccolo di individui della popolazione.
  - Per esempio, un ricercatore potrebbe intervistare un numero relativamente piccolo di cittadini italiani (diciamo 1000) per valutare l'incidenza della depressione.

Raccolta dati

- Come devono essere raccolti i dati di uno studio osservazionale? Quale metodo dovrebbe essere usato per selezionare gli individui del campione tra gli individui della popolazione?
  - Un buon campione è quello che viene scelto a caso. Per capire perché le cose stiano così, consideriamo due comuni procedure di campionamento non casuale.
    1. Campione di individui che si prestano in maniera volontaria a partecipare allo studio (voluntary response sample).
    2. Campione di individui che sono facilmente raggiungibili (convenience sample).

Campionamento non casuale

- Il problema dei campioni non casuali è che gli individui di quei campioni si differenziano in maniera sistematica dalla popolazione generale.
- Di conseguenza, i risultati dello studio basato su questi dati saranno distorti.

Campionamento non casuale

- Per esempio, per determinare l'incidenza della depressione in Italia potremmo chiedere a dei volontari di partecipare all'indagine (*voluntary response sample*), oppure potremmo limitarci ad intervistare gli individui che possono essere reperiti con più facilità (*convenience sample*).
- In entrambi i casi finiremmo per sottostimare grandemente l'incidenza della depressione, dato che gli individui depressi sono meno disposti a farsi intervistare.

Campionamento casuale semplice

- Per evitare gli errori sistematici discussi in precedenza possiamo ricorrere al campionamento probabilistico che seleziona a caso gli individui del campione all'interno della popolazione.
- Il campionamento probabilistico più semplice è detto *campionamento casuale semplice*

Campionamento casuale semplice

Un campione casuale semplice (CCS) di grandezza n è costituito da n individui selezionati dalla popolazione attraverso una procedura che fa in modo che tutti i possibili insiemi di grandezza n all'interno di quella popolazione abbiano la stessa probabilità di essere selezionati.

Il campionamento casuale semplice può essere realizzato assegnando un codice di identificazione a ciascun individuo nella popolazione per poi selezionare a caso un insieme n di questi codici.

Campionamento casuale semplice

- Dato che gli individui sono selezionati in maniera completamente casuale, nel campione non vi può essere la tendenza di favorire un certo risultato piuttosto che un altro.
- Ci sarà comunque una variazione nei risultati osservati in campioni diversi (variabilità campionaria).
- Se il campione è sufficientemente grande, però, la caratteristica misurata (per esempio, la proporzione di individui depressi) non varierà molto da campione a campione.

Campionamento casuale semplice

- Nel campionamento casuale semplice un individuo della popolazione non può comparire più di una volta in un dato campione.
- In tali circostanze, diciamo che è stato utilizzato un campionamento senza rimessa.

Campionamento casuale indipendente

Nel campionamento casuale indipendente (CCI), dopo che un individuo è stato selezionato (ovvero, inserito in un dato campione), viene reinserito nella popolazione così che, in linea di principio, possa essere selezionato un'altra volta.

- Tale procedura di selezione del campione va sotto il nome di campionamento con rimessa.

- Se il campione è molto più piccolo della popolazione è molto improbabile che lo stesso individuo venga scelto più di una volta.
- Di conseguenza, in tali circostanze, il campionamento casuale semplice e il campionamento indipendente sono sostanzialmente equivalenti.

CCS e CCI: illustrazione con R

```
 x<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20)

# Campionamento (probabilistico) casuale semplice
sample(x,size=3,prob=rep(1/20,20))
 # [1] 10  9  5
sample(x,size=5,prob=rep(1/20,20))
 # [1] 17 13  7  5  4
sample(x,size=10,prob=rep(1/20,20))
 # [1]  9 14  5 13  3 12 16  7  8  4

# Campionamento (probabilistico) casuale indipendente
sample(x,size=3,replace=TRUE,prob=rep(1/20,20))
# [1]  2 18 18
sample(x,size=5,replace=TRUE,prob=rep(1/20,20))
# [1] 13 18 15 15 14
sample(x,size=10,replace=TRUE,prob=rep(1/20,20))
# [1] 15 13  1  5  7 15  5  4  2 19
```

Altri metodi di campionamento

- I tradizionali metodi di inferenza statistica assumono che i dati provengano da un campione casuale indipendente,
- ma possono anche essere applicati ai dati provenienti da un campione casuale semplice purché la grandezza n del campione non sia maggiore, diciamo, al 10% della grandezza N della popolazione

Altri metodi di campionamento

- La caratteristica fondamentale del campionamento statistico è la selezione casuale: ciascun individuo nella popolazione deve avere una probabilità nota diversa da zero di essere selezionato e inserito nel campione.
- Non è necessario però che ciascun campione abbia la stessa probabilità di essere scelto, o che ciascun individuo abbia la stessa probabilità di essere inserito nel campione.

Altri metodi di campionamento

- Campionamento casuale stratificato
- Campionamento casuale a grappoli
- Campionamento multistadio

In questo corso ci limiteremo al CCS che rappresenta la forma di campionamento più usata in psicologia.

Limiti delle interviste

Anche se un campione di individui selezionati in maniera opportuna fornisce, in linea di principio, dati rappresentativi della popolazione, diversi fattori possono introdurre degli errori sistematici nei dati raccolti mediante interviste.

Limiti delle interviste

- Alcuni gruppi di individui non sono adeguatamente rappresentati nel campione $\smile$ per esempio, gli individui senza fissa dimora, coloro che non hanno un telefono, ...
- Alcuni individui selezionati che non intendono partecipare $\smile$ se tali individui differiscono in maniera sistematica da coloro che partecipano all'indagine, allora i risultati dell'indagine risultano distorti.
- Alcuni individui potrebbero non fornire informazioni veritiere, oppure vengono influenzati nelle loro risposte dalle caratteristiche dell'intervistatore o da altri aspetti estranei.
- Le risposte al questionario sono influenzate dal modo in cui sono poste le domande.

Limiti delle interviste

- Alcuni gruppi di individui non sono adeguatamente rappresentati nel campione $\smile$ per esempio, gli individui senza fissa dimora, coloro che non hanno un telefono, ...
- Alcuni individui selezionati che non intendono partecipare $\smile$ se tali individui differiscono in maniera sistematica da coloro che partecipano all'indagine, allora i risultati dell'indagine risultano distorti.
- Alcuni individui potrebbero non fornire informazioni veritiere, oppure vengono influenzati nelle loro risposte dalle caratteristiche dell'intervistatore o da altri aspetti estranei.
- Le risposte al questionario sono influenzate dal modo in cui sono poste le domande.

Limiti delle interviste

- Alcuni gruppi di individui non sono adeguatamente rappresentati nel campione $\smile$ per esempio, gli individui senza fissa dimora, coloro che non hanno un telefono, ...
- Alcuni individui selezionati che non intendono partecipare $\smile$ se tali individui differiscono in maniera sistematica da coloro che partecipano all'indagine, allora i risultati dell'indagine risultano distorti.
- Alcuni individui potrebbero non fornire informazioni veritiere, oppure vengono influenzati nelle loro risposte dalle caratteristiche dell'intervistatore o da altri aspetti estranei.
- Le risposte al questionario sono influenzate dal modo in cui sono poste le domande.

Limiti delle interviste

- Alcuni gruppi di individui non sono adeguatamente rappresentati nel campione $\smile$ per esempio, gli individui senza fissa dimora, coloro che non hanno un telefono, ...
- Alcuni individui selezionati che non intendono partecipare $\smile$ se tali individui differiscono in maniera sistematica da coloro che partecipano all'indagine, allora i risultati dell'indagine risultano distorti.
- Alcuni individui potrebbero non fornire informazioni veritiere, oppure vengono influenzati nelle loro risposte dalle caratteristiche dell'intervistatore o da altri aspetti estranei.
- Le risposte al questionario sono influenzate dal modo in cui sono poste le domande.

Conclusioni

- La validità delle inferenze basate sui dati dipende dalla validità dei metodi che sono stati utilizzati per selezionare il campione e per misurare le variabili di interesse.
- Si noti che considerazioni di tipo statistico devono essere fatte prima di raccogliere i dati.
- Se per rispondere alla domanda che motiva la ricerca è necessario applicare un determinato test statistico, allora i dati dovranno esibire i requisiti necessari per l'applicazione del test richiesto.

Conclusioni

- È dunque necessario decidere come analizzare i dati prima di raccoglierli.
- Altrimenti, potremmo scoprire troppo tardi di non avere misurato una o più variabili che erano necessarie per l'analisi statistica, oppure di avere misurato le variabili rilevanti ad un livello di scala inadeguato.