7 - Ragionamenti statistici

Comunicazione e critical thinking a.a. 2023-2024

Michele Paolini Paoletti (Università di Macerata)
m.paolinipaoletti@unimc.it

pagina insegnamento: http://tiny.cc/criticalthinking2324

Gli argomenti di questo modulo

(a) gli argomenti statistici;
(b) valutare gli argomenti statistici: ampiezza del campione;
(c) valutare gli argomenti statistici: rappresentatività del campione;
(d) valutare gli argomenti statistici: fattori nascosti;
(e) valutare gli argomenti statistici: domande viziate;
(f) le generalizzazioni sul futuro;
(g) le generalizzazioni scientifiche;
(h) il sillogismo statistico;
(i) il problema della classe di riferimento.

Argomenti induttivi per enumerazione semplice

D_1, D_2, D_3, etc. (**tutti** i **dati** fin qui **disponibili**) inerenti un fenomeno F **concordano** su una qualche caratteristica **C**.

Dunque: **tutti** i **dati** inerenti **F** concordano su **C** / il **prossimo dato** inerente F concorderà su C.

Ulteriore passaggio dopo l'argomento induttivo per enumerazione semplice:

Dunque: **C** è la spiegazione di F.

Problema 1: il **numero** di dati fin qui disponibili potrebbe essere **ristretto** → nuovi dati potrebbero emergere e rendere falsa la conclusione.

Problema 2: i **dati** fin qui disponibili potrebbero essere inconsapevolmente affetti da qualche **bias** (da qualche pregiudizio) → es. potrei essermi limitato a raccogliere dati soltanto in una certa settimana, oppure soltanto in certe situazioni, potrei aver ignorato dati incompatibili con la mia ipotesi esplicativa, etc.

Problema 3 (nell'ulteriore passaggio): la caratteristica **C** su cui tutti i dati (disponibili o tout court) concordano potrebbe **non** essere la **spiegazione** di F → potrebbe essere, ad esempio, un **fenomeno G** che si **accompagna sempre** a **F** senza spiegarlo → es. G è la carenza di sonno che si accompagna sempre ai miei frequenti mal di testa, ma entrambi hanno una causa comune: lo stress.

Argomenti statistici (a)

Argomenti statistici: la **distribuzione** di una certa **proprietà** all'interno di un certo insieme (la **popolazione**) viene inferita (non-deduttivamente) dalla **distribuzione** di quella **proprietà** o di un'altra proprietà **rilevante** all'interno di un sottoinsieme proprio di quell'insieme (il **campione**).

L'**n**% del **campione** ha la proprietà **Q**.

Dunque: l'**n**% della **popolazione** ha la proprietà **P** (laddove P = Q o Q è rilevante per possedere P).

Il 27% del campione intervistato dichiara di votare il Partito Democratico.

Dunque: il 27% degli elettori italiani vota il Partito Democratico.

N.B. Dichiarare di votare il Partito Democratico e votare il Partito Democratico sono due proprietà diverse. Tuttavia, il possesso della prima proprietà è rilevante per il possesso della seconda.

Argomenti statistici (b)

L'**n**% del **campione** ha la proprietà **Q**.

Dunque: l'**n**% della **popolazione** ha la proprietà **P** (laddove P = Q o Q è rilevante per possedere P).

Proprietà presente nel **campione**: **proprietà misurata** (Q).

Proprietà presente nella **popolazione**: **proprietà target** (P).

Idealmente, Q dovrebbe essere **identica** a P o il possesso di Q dovrebbe **necessitare** il possesso di P.

La **connessione** tra Q e P, tuttavia, **non** è sempre così **forte**:

- **non** è possibile **misurare direttamente** la proprietà **target** nel campione (es. nel segreto dell'urna, non sappiamo quanti intervistati voteranno effettivamente il Partito Democratico, cioè avranno P);
- il possesso della proprietà **misurata non implica** sempre con buona probabilità il possesso della proprietà **target** (es. in un regime autoritario o in un contesto particolarmente conformista, molti membri del campione potrebbero dichiarare di votare un certo partito e poi non votarlo).

Argomenti statistici (c)

L’**n**% del **campione** ha la proprietà **Q**.

Dunque: l’**n**% della **popolazione** ha la proprietà **P** (laddove P = Q o Q è rilevante per possedere P).

Perché i **buoni** argomenti statistici **possono** essere (non-deduttivamente) **forti**?

Canale et al.: Dato un insieme S (la **popolazione**) e posto che una **percentuale** di membri di S abbia la **proprietà P**, la **maggior parte** dei **sottoinsiemi** di S che sono sufficientemente **ampi** (i **campioni**) hanno una **percentuale analoga** di membri che possiedono **P** (sulla base della proprietà misurata Q).

I campioni, tuttavia, come vedremo, devono essere anche **rappresentativi…**

Argomenti statistici (d)

L'**n**% del **campione** ha la proprietà **Q**.

Dunque: l'**n**% della **popolazione** ha la proprietà **P** (laddove P = Q o Q è rilevante per possedere P).

- **Non-deduttivo**: sulla base delle premesse, la verità della conclusione può essere molto **plausibile**, ma la verità della conclusione **non** segue **logicamente** dalla verità di tutte le premesse;
- **ampliativo**: nella **conclusione** vi sono informazioni **nuove** (es. si parla della popolazione, e non del campione, della proprietà target, e non della proprietà misurata);
- **fallibile**: a fronte di **nuove informazioni** diverse da quelle contenute nelle premesse, la **conclusione** potrebbe risultare **falsa** E anche a fronte delle **stesse informazioni** la conclusione potrebbe risultare **falsa** → ogni argomento statistico ha un **margine di errore**, cioè un margine di **divergenza** positiva o negativa tra la **percentuale** di Q nel **campione** e la **percentuale** di P nella **popolazione** (es. + o - 2%)
- **non-monotono**: l'**aggiunta** di nuove premesse **non** lascia **immutato** il grado di plausibilità della conclusione.

Valutare gli argomenti statistici: proprietà misurata e target

Un buon argomento statistico deve rispettare alcune condizioni:

(1) il possesso della **proprietà misurata** deve implicare con una **buona probabilità** il possesso della **proprietà target**:

Dichiarare di votare il Partito Democratico implica con una buona probabilità (in un contesto in cui è valorizzata la sincerità degli intervistati) votare effettivamente il Partito Democratico.

Comprare un SUV non implica con una buona probabilità (in un contesto nel quale sono presenti molti modelli di SUV economici) il possesso di grandi ricchezze.

Dichiarare di andare a messa tutte le domeniche implica con una buona probabilità la conoscenza del Catechismo della Chiesa Cattolica?

Valutare gli argomenti statistici: ampiezza del campione

(2) Il **campione** deve essere **sufficientemente ampio**.

L'80% delle 1000 recensioni di questo ristorante sono positive. Dunque, l'80% dei clienti di questo ristorante ha mangiato bene → campione sufficientemente ampio.

L'80% delle 5 recensioni di questo ristorante sono positive. Dunque, l'80% dei clienti di questo ristorante ha mangiato bene → campione troppo ristretto.

Il 30% dei miei amici dichiara di votare Fratelli d'Italia. Dunque, il 30% degli elettori italiani vota Fratelli d'Italia → campione troppo ristretto.

Il 30% di tutti i maceratesi dichiara di votare Fratelli d'Italia. Dunque, il 30% degli elettori italiani vota Fratelli d'Italia → campione sufficientemente ampio, ma potrebbero esserci altri problemi…

Valutare gli argomenti statistici: rappresentatività (a)

(3) Il **campione** deve essere **sufficientemente rappresentativo**.

Che cosa significa?

Assumiamo che F_1, F_2, F_3, etc. siano **fattori** che possono **incidere** sulla **presenza** o sull'**assenza** della proprietà **target** P nella **popolazione** (**fattori incidenti**).

Es. riguardo al votare effettivamente il Partito Democratico, i fattori che possono incidere sulla presenza o sull'assenza di tale proprietà nella popolazione sono: l'età, il reddito, la regione di residenza, la condizione lavorativa, il grado di istruzione, etc.

Riguardo a ciascun fattore incidente F_1, F_2, F_3, etc. i membri della **popolazione** e del **campione** possono essere caratterizzati da **diversi valori**.

Es. riguardo alla condizione lavorativa, i membri della popolazione e del campione possono essere: disoccupati, sottoccupati, occupati a tempo indeterminato, occupati a tempo determinato, pensionati, etc.

Valutare gli argomenti statistici: rappresentatività (b)

Un campione **sufficientemente rappresentativo** è un campione nel quale, per **ogni valore** di **ogni fattore** incidente, è **minima** o **ridotta** la **divergenza** tra la percentuale dei **membri** del **campione** che sono caratterizzati da quel **valore** e la percentuale dei **membri** della **popolazione** che sono caratterizzati da quel **valore**.

In un campione sufficientemente rappresentativo, se il 30% degli italiani sono pensionati, allora il 30% del campione dovrà essere costituito da pensionati.

Alcuni problemi:

(a) **quali** sono i **fattori** incidenti per la distribuzione di P nella popolazione?
(b) Rispetto a P, tali fattori hanno tutti lo stesso **peso**? O hanno pesi diversi? (es. l'età e la regione di residenza hanno lo stesso peso? Oppure no? E quale pesa di più?)
(c) Quali sono i **valori** di ciascun fattore incidente?
(d) I valori di ciascun fattore incidente sono stati determinati in modo **esclusivo** oppure no? (es. un pensionato non è anche un occupato a tempo pieno)

Valutare gli argomenti statistici: rappresentatività (c)

Alcuni problemi:

(e) Alcuni **fattori incidenti** potrebbero **non** essere stati **individuati** (es., in un sondaggio politico, le ore trascorse su Facebook).

(f) Lo **stesso membro** del campione è caratterizzato da **valori diversi** rispetto a **fattori** incidenti **diversi** (es. un pensionato di 70 anni dell'Alto Adige che guadagna più di 60.000 euro l'anno e ha una laurea quadriennale)→ è difficile ottenere una **piena rappresentatività** del campione rispetto a **tutti** i valori di tutti i fattori incidenti.

(g) **Non** sempre è possibile conoscere la **percentuale** di membri della **popolazione** che sono caratterizzati da **ciascun valore** di **ciascun fattore** incidente (es. il grado di fede religiosa, che è una proprietà target che non si può misurare direttamente).

(h) Potrebbero partecipare alla **rilevazione statistica** - ed essere membri del campione - **soltanto** i **membri** della **popolazione** caratterizzati da un **certo valore** di un certo fattore incidente (es., in un sondaggio politico condotto su Facebook, soltanto gli utenti di Facebook) → quel valore sarebbe **sovrarappresentato** nel campione.

Valutare gli argomenti statistici: rappresentatività (d)

Per risolvere questi problemi, è possibile pianificare attentamente la composizione del campione.

Oppure: è possibile **randomizzare**, cioè scegliere membri del campione **a caso** all'interno della popolazione. In tal caso, però, occorre:

- rispettare il criterio (2): il **numero** dei membri scelti deve essere sufficientemente **ampio**. Perché?

  In una popolazione sufficientemente ampia, i sottoinsiemi-campione **randomizzati** sono **più numerosi** di quelli non-randomizzati;

  i sottoinsiemi-campione sufficientemente **ampi** hanno una **percentuale** di membri che possiedono la proprietà target **P analoga** a quella della popolazione

  → dunque: i sottoinsiemi-campione **randomizzati** e sufficientemente **ampi** hanno una **maggiore probabilità** di avere una **percentuale** di membri che possiedono P **analoga** a quella della popolazione;

- rispettare il criterio (4) (v. slide successiva): **non** devono esserci **fattori incidenti nascosti** nella scelta casuale dei membri.

Valutare gli argomenti statistici: fattori nascosti (a)

(4) Lo studio del **campione** non deve essere influenzato da **fattori incidenti nascosti**, che possano influenzare la scelta dei **membri** e/o la percentuale di membri con la **proprietà misurata** Q e/o la **connessione** tra la proprietà **misurata** Q e la proprietà **target** P.

Canale et al.

Nell'ospedale A 900 pazienti su 1000 sono sopravvissuti ad un intervento all'anca.

Nell'ospedale B 800 pazienti su 1000 sono sopravvissuti allo stesso intervento.

Dunque, i pazienti hanno più probabilità di sopravvivere a questo intervento nell'ospedale A che nell'ospedale B.

Ma Giovanni vuole sapere quanti pazienti in condizioni di salute gravi sono sopravvissuti all'intervento. Qui le cose cambiano.

Nell'ospedale A ci sono state 100 persone in condizioni di salute gravi. Sono sopravvissute all'intervento all'anca in 30.

Nell'ospedale B ci sono state 400 persone in condizioni di salute gravi. Sono sopravvissute all'intervento all'anca in 200.

Dunque, per i pazienti in condizioni di salute gravi, c'è maggiore probabilità di sopravvivere nell'ospedale B.

Valutare gli argomenti statistici: fattori nascosti (b)

Problema 1: **connessione** tra la proprietà misurata **Q** (sopravvissuti all'intervento all'anca) e la proprietà target **P** (sopravvissuti all'intervento all'anca in gravi condizioni di salute). Inizialmente il campione è stato esaminato sulla base della prima proprietà. Pertanto, è stato ignorato il fattore incidente delle gravi condizioni di salute - decisivo per la proprietà target.

Problema 2: il campione è stato selezionato senza tenere in considerazione il fatto che nell'ospedale B vi sono state più persone in condizioni di salute gravi. In termini comparativi, i due campioni **non** avevano lo **stesso grado** di **rappresentatività**.

Valutare gli argomenti statistici: domande viziate

(5) Lo studio del **campione** deve essere effettuato **senza domande viziate** (o domande con **bias**), cioè senza domande poste in modo tale da poter influenzare la risposta.

Domanda non-viziata: voteresti il Partito Democratico?

Domanda viziata con bias (positivo): voteresti l'unico partito che ha sempre sostenuto il più autorevole governo dal Secondo Dopoguerra ad oggi?

Domande viziata con bias (negativo): voteresti il partito erede del Partito Comunista Italiano, cioè di un movimento che si è macchiato di innumerevoli crimini nel '900?

Le generalizzazioni sul futuro (a)

L'**n**% dei **casi passati** (≃**campione)** ha la proprietà **P**.

Dunque: l'**n**% di **tutti** i **casi** (passati, presenti e futuro) **o** dei casi **futuri** (≃**popolazione)** ha la proprietà **P**.

L'Inter ha vinto finora il 50% delle partite.

Dunque, l'Inter vincerà il 50% di tutte le partite della stagione.

Marco ha centrato finora 10 birilli su 100.

Dunque, Marco centrerà 10 birilli su 100 nel corso di tutta la partita.

Argomento **non-deduttivo**. **Forza**: sulla base dell'**uniformità** tra il passato, il presente e il futuro.

Maggiore è l'**uniformità** tra il passato, il presente e il futuro, **più forte** è la generalizzazione.

Minore è l'**uniformità** tra il passato, il presente e il futuro, **più debole** è la generalizzazione.

Le generalizzazioni sul futuro (b)

N.B.1 Per valutare la **forza** della generalizzazione, occorre **conoscere** il grado di **uniformità** tra passato, presente e futuro.

Ma, per **conoscere** il grado di **uniformità** tra passato, presente e futuro, occorre basarsi sul **successo** o sull'**insuccesso** di generalizzazioni **analoghe** e **passate**.

Per valutare la generalizzazione sull'Inter, dovrò basarmi su generalizzazioni riguardanti squadre di calcio simili all'Inter in contesti simili e sul loro successo/insuccesso.

N.B.2 Spesso la verità della **prima premessa** si fonda su un precedente **argomento statistico**, in cui si studia la distribuzione di una certa **proprietà misurata** Q (rilevante per P) all'interno di un **campione** più **ristretto**. A parità di condizioni, ciò **indebolisce** l'argomento.

Nel 2021 il 10% degli italiani ha letto almeno un libro al mese → conclusione tratta da un argomento statistico del tipo: il 10% del **campione** degli italiani intervistati **dichiara** di aver letto almeno un libro al mese nel 2021. Dunque, il 10% degli italiani ha letto almeno un libro al mese nel 2021.

Dunque, nel 2022 il 10% degli italiani leggerà almeno un libro al mese.

Le generalizzazioni scientifiche

Il **100**% dei **casi passati** (≈**campione)** ha la proprietà **P**.

Dunque: il **100%** di **tutti** i **casi** (passati, presenti e futuro) **o** dei casi **futuri** (≈**popolazione)** ha la proprietà **P**.

Tutti i giorni il Sole è sorto.

Dunque, tutti i giorni il Sole sorgerà.

La rivoluzione della Terra attorno al Sole negli ultimi millenni è durata circa 365 giorni.

Dunque, per qualche altro millennio, la rivoluzione della Terra attorno al Sole durerà circa 365 giorni.

Fondate su un **alto** grado di **uniformità** tra passato, presente e futuro.

Questo alto grado di uniformità, a sua volta, è fondato sulla presenza di determinate **leggi di natura**.

Il sillogismo statistico (a)

L’**n**% di una certa **popolazione E** ha la proprietà **P**.

L’entità ***a*** appartiene alla **popolazione E**.

(**prima versione**) Dunque: vi è un **n%** di **probabilità** che ***a*** abbia **P**.

(**seconda versione**, se **n** è sufficientemente **alto/basso**): Dunque: ***a*** ha/non ha **P**.

Il 70% dei percettori di reddito di cittadinanza ha votato il Movimento 5 Stelle.

Luca è un percettore di reddito di cittadinanza.

(prima versione) Dunque: vi è un 70% di probabilità che Luca abbia votato il Movimento 5 Stelle.

(seconda versione) Dunque: Luca ha votato il Movimento 5 Stelle.

Il sillogismo statistico (b)

In **entrambe** le **versioni**, il sillogismo è **non-deduttivo**.

Anche quando si conclude che la probabilità è dell'**n%** (come nella **prima versione**), la conclusione **non** segue necessariamente dalla verità delle **premesse**. Date le stesse premesse, **altri fattori** potrebbero aggiungersi, influenzare la probabilità rilevante e modificarla.

Il 70% dei percettori di reddito di cittadinanza ha votato il Movimento 5 Stelle.

Luca è un percettore di reddito di cittadinanza.

Ma: Luca odia Giuseppe Conte.

Dunque: vi è un 70% di probabilità che Luca abbia votato il Movimento 5 Stelle → più debole.

Il sillogismo statistico (c)

Il sillogismo è **più forte** nella **prima versione** che nella seconda versione.

La **conclusione** della **prima** versione (vi è un 70% di probabilità…) è **meno informativa** della **conclusione** della **seconda** versione (Luca ha votato il Movimento 5 Stelle).

Dunque, la **conclusione** della **seconda** versione ha **maggiore** probabilità di essere **falsificata**.

Il sillogismo statistico (d)

N.B.1 Spesso la verità della **prima premessa** si fonda su un precedente **argomento statistico**, in cui si studia la distribuzione di una certa **proprietà misurata** Q (rilevante per P) all'interno di un **campione** più **ristretto**. A parità di condizioni, ciò **indebolisce** l'argomento.

Il 70% dei percettori di reddito di cittadinanza ha votato il Movimento 5 Stelle → conclusione tratta da un argomento statistico del tipo: il 70% del **campione** dei percettori di reddito intervistati **dichiara** di aver votato il Movimento 5 Stelle. Dunque, il 70% dei percettori di reddito di cittadinanza ha votato il Movimento 5 Stelle.

N.B.2 La **probabilità** di cui si parla nella **prima versione** del sillogismo (vi è un 70% di probabilità che Luca abbia votato il Movimento 5 Stelle) è di natura **epistemica**. Essa consiste nel grado di **attendibilità** della credenza che Luca abbia votato il Movimento 5 Stelle sulla base dei dati conosciuti.

La **probabilità reale** che Luca abbia votato il Movimento 5 Stelle, invece, è **1** (se lo ha votato) o **0** (se non lo ha votato).

Il problema della classe di riferimento (a)

L’**n**% di una certa **popolazione E** ha la proprietà **P**.

L’entità ***a*** appartiene alla **popolazione E**.

(**prima versione**) Dunque: vi è un **n%** di **probabilità** che ***a*** abbia **P**.

(**seconda versione**, se **n** è sufficientemente **alto/basso**) Dunque: ***a*** ha/non ha **P**.

L’entità ***a*** potrebbe appartenere a **diverse popolazioni E**, **F**, **G**, etc.

Tali popolazioni potrebbero avere **percentuali diverse** n%, n’%, n’’%, etc., di **distribuzione** della proprietà **P**.

Prima versione: Qual’è la **probabilità** che ***a*** abbia **P**?

Seconda versione: Quanto è **forte** la **conclusione** che ***a*** ha/non ha **P**?

Il problema della classe di riferimento (b)

Il 70% dei percettori di reddito di cittadinanza ha votato il Movimento 5 Stelle.

Il 30% di chi ha votato Ingroia nel 2018 ha votato il Movimento 5 Stelle.

Il 10% di chi vive in Lombardia ha votato il Movimento 5 Stelle.

Luca è un percettore di reddito di Cittadinanza, ha votato Ingroia nel 2018 e vive in Lombardia.

Prima versione: qual'è la probabilità che Luca abbia votato il Movimento 5 Stelle?

Seconda versione: quanto è forte la conclusione che Luca ha votato il Movimento 5 Stelle?

Il problema della classe di riferimento (c)

Prima strategia: si cerca di determinare la **popolazione più rilevante** di tutte quando si tratta di valutare la distribuzione della proprietà **P**.

In tal caso, si considera **solo** la **percentuale** di distribuzione di **P** nella popolazione **più rilevante**.

Per aumentare la probabilità di aver votato il Movimento 5 Stelle, conta di più il percepire il reddito di cittadinanza? O l'aver votato Ingroia nel 2018? O il risiedere in una certa regione, come la Lombardia?

Seconda strategia: Se ci sono **più popolazioni** che paiono **egualmente rilevanti**, occorre determinare la **percentuale** di distribuzione della proprietà **P** nell'**intersezione** tra queste popolazioni, cioè nella popolazione che presenta **tutte** le **caratteristiche** delle prime.

Se le tre popolazioni precedenti sono egualmente rilevanti, occorre determinare la distribuzione della proprietà di aver votato il Movimento 5 Stelle nell'intersezione tra queste popolazioni, cioè nella popolazione dei percettori di reddito di cittadinanza che hanno votato Ingroia nel 2018 e vivono in Lombardia.

Nuova indagine? Media tra le tre percentuali delle singole popolazioni? Media ponderata sulla base del numero di membri delle singole popolazioni? Dipende…