

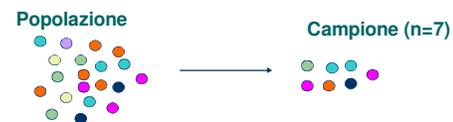
LA VERIFICA DELLE IPOTESI

Popolazioni, campioni, parametri ed indicatori

Popolazioni e campioni

- Viene definita **popolazione** o **universo** *l'insieme completo di tutti gli elementi che hanno in comune una determinata caratteristica.*
- Le popolazioni possono essere:
 - ✓ **finite**, cioè definibili in termini di un determinato numero di elementi (**es. tutti gli studenti iscritti quest'anno a Psicologia, tutte le possibili combinazioni delle facce di due dadi, ecc.**)
 - ✓ o **infinite**, ossia non definibili in termini di un numero di elementi che le compongono (**es. tutti i possibili temi svolti da studenti di 15 anni, ecc.**)

- Il **campione** è un *sottoinsieme di N (o n) elementi dell'insieme dato*, dove per "insieme dato" si intende la popolazione o universo che ci interessa.
- Il numero **n** definisce l'**ampiezza** del campione, cioè il numero di elementi di cui esso è costituito (possiamo avere campioni di ampiezza $n=2$, $n=3$, $n=50$, ecc. estratti dalla stessa popolazione)



- Il ricercatore è interessato ad acquisire conoscenze che riguardano la popolazione ma molto spesso non è "economico" o è del tutto impossibile rilevare i dati su tutti i soggetti della popolazione (Es. sondaggi sulle intenzioni di voto).
- Il ricercatore lavora quindi (quasi) sempre sui campioni estratti dalla popolazione.
- La procedura attraverso cui il campione è estratto dalla popolazione è detta **campionamento**.

- Un problema importante che si pone nella selezione del campione è la sua **rappresentatività** rispetto alla popolazione.
- Un **campione è rappresentativo** quando possiede le principali caratteristiche della popolazione da cui proviene.
- Nessun campione è mai perfettamente rappresentativo della popolazione, ma i campioni estratti con **campionamento casuale** (in cui tutti gli elementi della popolazione hanno la stessa probabilità di essere estratti) ci permettono di ritenere che le differenze tra campione e popolazione siano dovute al caso.

Parametri e indicatori

- Un **parametro** è una quantità numerica che misura un qualche aspetto di una **popolazione**. Ad esempio la media μ del quoziente intellettivo di tutti gli studenti di 15 anni. Simbolicamente i parametri riferiti alla popolazione sono indicati con le lettere greche (es. media = μ ; varianza = σ^2)
- Si chiama **indicatore o statistica** la stessa quantità riferita ad un **campione**. Ad esempio la media \bar{X} del quoziente intellettivo di un campione di 20 ragazzi estratto casualmente dalla popolazione degli studenti di 15 anni.

7

Le distribuzioni campionarie

8

Una **distribuzione campionaria** è una **distribuzione di frequenza** relativa ad una **statistica (o indicatore)**. A differenza delle distribuzioni teoriche, le distribuzioni campionarie sono **empiriche**, cioè possono essere calcolate a partire dai dati osservati.

Per ottenere la distribuzione campionaria di una statistica (es. media, deviazione standard ecc.):

1. si estraggono dalla popolazione **tutti i possibili campioni** di una certa numerosità;
2. si calcola la statistica (o indicatore) in questione su ciascuno di essi;
3. si assegna ad ogni valore la relativa frequenza. La probabilità associata ad ogni valore è quella che si ottiene in seguito ad estrazione casuale del campione.

9

- Esistono distribuzioni campionarie note di vari indicatori (o statistiche: es. proporzioni, media, varianza). Ciascuna di tali distribuzioni ha una sua **media**, una sua **varianza** e una sua deviazione standard (che nel caso di distribuzioni campionarie è chiamata **errore standard**).
- La più importante è la **distribuzione campionaria della media**, i cui elementi sono le medie calcolate su tutti i possibili campioni di una certa numerosità (N) estratti da una stessa popolazione.

10

Come si costruisce una distribuzione campionaria.

Esempio: la distribuzione campionaria della media

11

Consideriamo una **popolazione** finita che assume i seguenti valori: 10; 4; 0; 2; 6; 8 ($\mu = 5$; $\sigma = 3,41$; $\sigma^2 = 11,62$). Vogliamo costruire la distribuzione campionaria delle medie per campioni di numerosità **N = 2**

1. Estraiamo dalla popolazione tutti i possibili campioni di numerosità = 2 e calcoliamo su ciascuno la media: \bar{X}

Campione	media	Campione	media	Campione	media
10 e 4	7	4 e 0	2	0 e 6	3
10 e 0	5	4 e 2	3	0 e 8	4
10 e 2	6	4 e 6	5	2 e 6	4
10 e 6	8	4 e 8	6	2 e 8	5
10 e 8	9	0 e 2	1	6 e 8	7

12

Scalisi - Tecniche Psicometriche

2. Contiamo quante volte si ripete ogni media

Campione	media	Campione	media	Campione	media
10 e 4	7	4 e 0	2	0 e 6	3
10 e 0	5	4 e 2	3	0 e 8	4
10 e 2	6	4 e 6	5	2 e 6	4
10 e 6	8	4 e 8	6	2 e 8	5
10 e 8	9	0 e 2	1	6 e 8	7

La media 7 si ripete
2 volte

La media 5 si ripete
3 volte

13

3. Costruiamo la distribuzione di frequenza (f_i) delle medie e calcoliamo per ognuna la relativa probabilità (frequenza/N)

Medie (\bar{x})	1	2	3	4	5	6	7	8	9	Tot
f_i	1	1	2	2	3	2	2	1	1	15
prob. ($f_i/15$)	0,07	0,07	0,13	0,13	0,2	0,13	0,13	0,07	0,07	1,0

Quindi: la probabilità di estrarre dalla popolazione originaria un campione casuale di numerosità $N = 2$ e media (ad esempio) = 3, è 0,13 (13%). Per un campione con $N = 2$ e media = 9, invece, la probabilità è 0,07 (7%).

Se calcoliamo la media delle medie otteniamo 5, cioè esattamente il valore della media della popolazione (μ) da cui abbiamo estratto i campioni.

14

Proprietà della distribuzione campionaria delle medie

La media delle medie dei campioni coincide con la media della popolazione da cui i campioni sono stati estratti.

$$\mu_{\bar{x}} = \mu$$

Media delle medie dei campioni Media della popolazione

Man mano che l'ampiezza N del campione aumenta, la sua media diventa una stima sempre più "precisa" della media della popolazione, coincidendo con essa quando N è uguale alla numerosità della popolazione.

15

Proprietà della distribuzione campionaria delle medie

La deviazione standard calcolata sulla distribuzione campionaria delle medie si chiama **errore standard** ed è uguale alla deviazione standard della popolazione diviso la radice quadrata di N (ampiezza del campione).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

All'aumentare di N la variabilità della distribuzione campionaria delle medie diminuisce fino a tendere a 0 (**legge dei grandi numeri**).

16

Proprietà della distribuzione campionaria delle medie

Secondo il **teorema del limite centrale** per campioni di ampiezza abbastanza grande ($N > 30$) la distribuzione campionaria delle medie si approssima alla distribuzione normale, *qualunque sia la forma della distribuzione della popolazione*.

17

La verifica delle ipotesi

18

Scalisi - Tecniche Psicometriche

L'obiettivo della scienza non è quello di descrivere fenomeni che si manifestano all'interno di particolari campioni di persone o eventi, ma quello di sviluppare **teorie generali** che riguardano intere **popolazioni**.

Le **ipotesi** sono **predizioni** (se ... allora), espresse in termini operativi, che derivano logicamente da una teoria e devono essere vere se la teoria è vera.

La falsificazione di una ipotesi derivata da una teoria mette in discussione la teoria stessa, al contrario se una ipotesi è verificata ciò non implica necessariamente che la teoria sia vera, perché altre ipotesi derivate dalla stessa teoria potrebbero risultare non vere.

19

Esempio

Teoria: La capacità di inibire informazioni irrilevanti favorisce l'apprendimento del calcolo.

Ipotesi che ne derivano:

- 1) Bambini con discalculia hanno punteggi inferiori a test di "inibizione", rispetto a bambini che non presentano discalculia (confronto tra gruppi).
- 2) Esiste una correlazione positiva significativa, nella popolazione scolastica, tra punteggi a prove di inibizione e rendimento in aritmetica (ricerca correlazionale).
- 3) Il punteggio ad una prova di "inibizione" misurato alla fine della scuola dell'infanzia è in grado di predire il rendimento in aritmetica un anno dopo (ricerca longitudinale).

Se la teoria è vera, tutte e tre le ipotesi formulate devono essere vere.

20

La significatività statistica dei risultati

- La fase di valutazione della significatività statistica dei risultati fa parte della **statistica inferenziale**, cioè della **verifica probabilistica delle ipotesi**.
- Il ricercatore ha interesse ad affermare che il risultato della sua ricerca dipende dall'esistenza, nella popolazione, di una **relazione tra le variabili considerate**. Per fare ciò deve dimostrare che il suo risultato ha una **probabilità molto bassa** di essere dovuto al caso (cioè a fattori accidentali che riguardano lo specifico campione esaminato).
- Da questo dipende la **validità delle conclusioni statistiche** che vengono tratte dai risultati ottenuti.

21

Verifica della significatività statistica dei risultati: i passi da seguire

- 1) Formulazione dell'ipotesi nulla e dell'ipotesi alternativa:

Ipotesi nulla (H_0): Il risultato è ottenuto per effetto del caso

Ipotesi alternativa (H_1): Il risultato è ottenuto a causa della relazione esistente tra la variabile indipendente e la variabile dipendente (o della relazione esistente tra le diverse variabili considerate nel disegno sperimentale)

22

Verifica della significatività statistica dei risultati: i passi da seguire

Se la distribuzione teorica di probabilità che si utilizza è simmetrica (es. la curva normale), l'ipotesi alternativa può essere di tre tipi:

- $A > B$ monodirezionale destra
- $A < B$ monodirezionale sinistra
- $A \neq B$ bidirezionale

Numerosi esempi di ipotesi alternative monodirezionali destre o sinistre o bidirezionali e delle procedure di verifica della significatività statistica dei risultati a seconda dell'ipotesi formulata, sono riportati nei libri di Psicometria della triennale.

23

Verifica della significatività statistica dei risultati: i passi da seguire

- 2) Costruzione della distribuzione campionaria della statistica calcolata

La distribuzione campionaria ci consente di calcolare la probabilità di ottenere un certo valore di una statistica (o indicatore) **se l'ipotesi nulla è vera, ossia se il risultato è ottenuto per effetto del caso.**

24

Verifica della significatività statistica dei risultati: i passi da seguire

3) Uso della distribuzione teorica di probabilità della statistica calcolata

Non accade mai che il ricercatore debba costruire la distribuzione campionaria della statistica che ha calcolato per conoscere la probabilità di ottenere per caso un certo risultato, perché egli dispone in anticipo di un certo numero di informazioni (sotto forma di **teoremi**) che gli consentono di utilizzare la **distribuzione teorica di probabilità** che più si avvicina alla **distribuzione campionaria** della statistica da lui calcolata. Ad esempio sappiamo che la distribuzione campionaria della media, per campioni con $N > 30$, si approssima alla distribuzione normale, quindi in questo caso si usa la tavola di probabilità della distribuzione normale.

26

Verifica della significatività statistica dei risultati: i passi da seguire

4) Calcolo della statistica del test di significatività:

Il test di significatività che deve essere utilizzato dipende dall'indicatore sul quale si vuole verificare l'ipotesi. Continuando a prendere come esempio la verifica delle ipotesi sulla media, per $n > 30$, per poter utilizzare la distribuzione teorica di probabilità della curva normale, dobbiamo trasformare la media del campione in punto z e quindi utilizzare il test z (per un campione), con la formula:

$$z = \frac{\bar{X} - \mu_x}{\sigma_{\bar{x}}} \quad \begin{array}{l} \rightarrow \mu_{\bar{x}} = \mu \\ \rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \end{array}$$

26

Verifica della significatività statistica dei risultati: i passi da seguire

5) Scelta del livello di significatività:

Il livello di significatività α rappresenta la cosiddetta **zona di rifiuto** o **regione critica** dell'ipotesi nulla H_0 .

Fissando il livello di significatività stabiliamo la **regola decisionale** per accettare o respingere l'ipotesi nulla, cioè quanto deve essere bassa la probabilità di ottenere per caso un certo risultato per poter respingere H_0 .

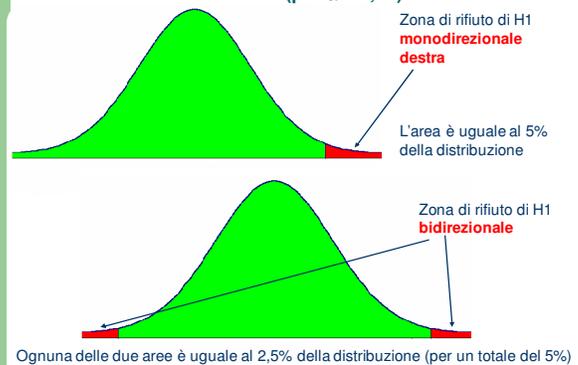
Il livello di significatività (α) per convenzione corrisponde a **0,05** oppure **0,01**.

Esso va stabilito **prima** di calcolare la statistica del test, e non può essere successivamente modificato in base al risultato ottenuto!

Se la distribuzione teorica di riferimento è **simmetrica** (come la curva normale) la zona di rifiuto varia in funzione dell'**ipotesi alternativa** (H_1).

27

Zona di rifiuto per H_1 monodirezionale o bidirezionale (per $\alpha = 0,05$)



Verifica della significatività statistica dei risultati: i passi da seguire

6) Conclusione:

Dopo aver calcolato la probabilità di ottenere per caso un certo risultato (utilizzando le tavole della distribuzione di probabilità più appropriata), confrontiamo tale probabilità con il livello di significatività scelto α :

Se la **probabilità** di ottenere per caso un certo risultato è **maggiore** di α ($p > \alpha$) \Rightarrow Non si può rifiutare H_0

Se la **probabilità** di ottenere per caso un certo risultato è **minore** di α ($p < \alpha$) \Rightarrow Si **rifiuta** H_0 con una probabilità di errore = α .

Ricordando che l'**ipotesi nulla** (H_0) afferma sempre che il **risultato è ottenuto per effetto del caso**.

29

La probabilità di errore nella decisione statistica

Sia che l'ipotesi nulla venga rifiutata oppure no, la decisione finale comporta sempre un certo **marginale di errore**. Si distinguono due tipi di errore:

Errore di I° tipo: rifiutare l'ipotesi nulla quando è vera.

Errore di II° tipo: accettare l'ipotesi nulla quando è falsa.

Quando si accetta l'ipotesi nulla non è possibile calcolare con esattezza la **probabilità dell'errore di II° tipo**, perché dipende da molti fattori (ampiezza del campione, errore sperimentale, sensibilità del disegno sperimentale ecc.).

L'**errore di I° tipo** dipende invece da un solo fattore: il **livello di significatività** α . Quindi, se si rifiuta l'ipotesi nulla con $\alpha = 0,05$, l'errore di I° tipo è $\alpha = 0,05 = 5\%$.

30

Verifica della significatività statistica dei risultati: il risultato è casuale? (un esempio, utilizzando il test z per un campione)

Uno psicologo scolastico vuole sperimentare un metodo di potenziamento delle abilità di calcolo e chiede al dirigente di una scuola primaria di estrarre a caso due classi di III primaria da sottoporre alla sperimentazione. Il dirigente segnala due classi, per un totale di 50 bambini.

Prima di iniziare la sperimentazione lo psicologo somministra ai 50 bambini del campione una prova di calcolo ed ottiene che la media del campione è $\bar{X} = 6,5$. Il manuale del test indica che il punteggio medio della popolazione dei bambini di III primaria è $\mu = 5$, con scarto quadratico medio $\sigma = 4,5$. La media del campione è più alta di quella della popolazione: lo psicologo ha quindi il dubbio che la scelta delle classi da parte del dirigente non sia stata casuale ma che questi abbia selezionato due classi con bambini particolarmente bravi nel calcolo. Ciò potrebbe compromettere il risultato della sua sperimentazione, che è destinata a bambini con capacità di calcolo "nella norma".

31

Ragioniamo su casualità e probabilità

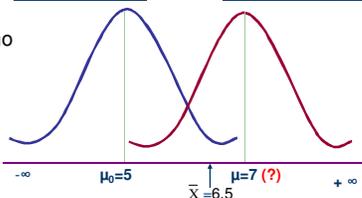
In sintesi, un campione di 50 bambini appartenenti ad una popolazione con media $\mu = 5$, ha media $\bar{X} = 6,5$. Questa differenza di 1,5 punti può essere dovuta al caso? Sì, i bambini del campione potrebbero essere "nella norma", ma, "per caso", si potrebbe trattare dei più bravi tra i "normali". Oppure, in alternativa, si potrebbe trattare di un campione appartenente ad una popolazione di bambini molto dotati nel calcolo, quindi con media μ più elevata di quella della popolazione cui si riferiscono le norme del test.

In quest'ultimo caso, la probabilità di estrarre casualmente dalla popolazione "normale" con media $\mu = 5$ un campione con media $\bar{X} = 6,5$, sarà **molto bassa**. Ma quanto deve essere bassa questa probabilità per poter affermare che il risultato non è casuale? **Per convenzione si considera come limite un valore di 0,05 o 0,01 (livello di significatività α).**

32

POPOLAZIONE BAMBINI "NORMALI" (dati normativi del test) **POPOLAZIONE BAMBINI "PIU' BRAVI" NEL CALCOLO (popolazione ipotetica)**

Esaminiamo meglio il problema



Il problema proposto prevede quindi due possibilità:

- il campione è un caso estremo della popolazione di riferimento i cui dati sono noti (curva blu): la media 6,5 del campione ha probabilità casuale bassa (coda di destra della curva blu)
- il campione appartiene ad una popolazione di bambini "più bravi" i cui dati sono ignoti (che, per esempio, potrebbe avere media 7 ed è rappresentata dalla curva rossa); in tal caso non possiamo conoscere la probabilità associata alla media del campione, perché i dati di questa popolazione non sono noti.

33

Formulazione di ipotesi nulla e ipotesi alternativa:

- (H_0): $\mu = \mu_0 = 5$** (La popolazione di provenienza del campione ha media uguale a quella della popolazione nota, ovvero l'estrazione è stata casuale)
- (H_1): $\mu > \mu_0$** (La popolazione di provenienza del campione ha media superiore a quella della popolazione nota, ovvero l'estrazione non è stata casuale)

N.B. L'ipotesi alternativa è monodirezionale destra, in quanto la differenza attesa è > 0 e quindi il risultato ricadrà nella parte destra della distribuzione normale.

34

Perché si ragiona solo sull'ipotesi nulla?

Noi non possiamo sapere quale sia la media della popolazione dei bambini «più bravi nel calcolo» cui il campione potrebbe appartenere (7 o un altro valore), quindi non possiamo sottoporre a verifica l'ipotesi alternativa. L'unica ipotesi che può essere sottoposta a verifica è l'ipotesi nulla (il risultato è casuale), perché le distribuzioni di probabilità ci forniscono solo valori relativi ad eventi casuali (v. esempio del dado nella lezione sulla probabilità).

35

Quindi l'obiettivo finale è ricavare dalla tabella della curva normale standardizzata la probabilità di ottenere per caso:

$\bar{X} = 6,5$ oppure un valore superiore *

e confrontare tale probabilità con il livello di significatività scelto.

* Perché?

Perché l'ipotesi alternativa prevede che il campione provenga da una popolazione con media superiore a quella della popolazione nota (ipotesi monodirezionale destra). Nel caso di ipotesi monodirezionale sinistra si calcola la probabilità di ottenere un valore uguale o inferiore al valore calcolato.

36

Scalisi - Tecniche Psicometriche

Scelta del livello di significatività: Prima di procedere con i passi successivi dobbiamo fissare il livello di significatività α .
 Convenzionalmente $\alpha = 0,05$ o $0,01$.
 Scegliamo $\alpha = 0,05$. Se la probabilità casuale del risultato ottenuto sarà inferiore al livello di significatività scelto, noi rifiuteremo l'ipotesi nulla, altrimenti la accetteremo. Fissando tale livello di significatività ci assumiamo un rischio di rifiutare H_0 quando è vera (ossia di commettere un errore di I tipo) pari al 5%.

37

Uso della distribuzione teorica di probabilità della statistica calcolata

Poiché abbiamo formulato un'ipotesi sulla media, facciamo riferimento alle proprietà della distribuzione campionaria delle medie. Sappiamo che quando $N > 30$ questa distribuzione si approssima alla curva normale, quindi possiamo utilizzare la tavola della distribuzione normale per trovare la probabilità di estrarre a caso, da una popolazione con $\mu = 5$ e $\sigma = 4,5$, un campione con $\bar{X} \geq 6,5$.

Calcolo della statistica del test di significatività

Per poter utilizzare la distribuzione teorica di probabilità della curva normale, dobbiamo trasformare la media del campione in punto z e quindi utilizzare il test z (per un campione), con le formule:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \quad \mu_{\bar{X}} = \mu \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = 4,5 / \sqrt{50} = 0,64$$

$$z = (6,5 - 5) / 0,64 = 1,5 / 0,64 = 2,34$$

38

Sulla tavola della curva normale standardizzata ("area al di là di z") osserviamo che la probabilità di ottenere un valore di z **superiore o uguale a 2,34** è **0,0096**.

Estratto dalla Tavola I in Areni & Scalisi "Esercizi di statistica per la ricerca psicologica", pag. 317:

(A) z	(B) area compresa tra la media e z	(C) area al di là di z	(A) z	(B) area compresa tra la media e z	(C) area al di là di z
.....
2,32	,4898	,0102	2,89	,4981	,0019
2,33	,4901	,0099	2,90	,4981	,0019
2,34	,4904	,0096	2,91	,4982	,0018
2,35	,4906	,0094	2,92	,4982	,0018
2,36	,4909	,0091	2,93	,4983	,0017
.....

39

Conclusione

Nel passo precedente abbiamo calcolato che estraendo a caso un campione di 50 bambini da una popolazione con media 5, la probabilità che la media del campione sia $\geq 6,5$ è 0,0096 (0,96%). Questa probabilità è **inferiore** al livello di significatività scelto ($\alpha = 0,05$ o 5%).

Possiamo **rifiutare l'ipotesi nulla** e concludere, **con una probabilità d'errore (di primo tipo) dello 0,05 (5%)** che il risultato ottenuto non è casuale, quindi è molto probabile che il dirigente non abbia scelto a caso le due classi, ma abbia selezionato due classi con bambini particolarmente bravi nel calcolo (lo psicologo dovrà quindi selezionare un altro campione per effettuare la sua sperimentazione).

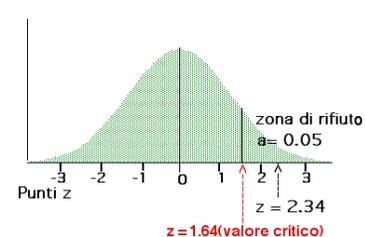
40

Un altro modo per verificare se un risultato è **significativo** (cioè se ci permette di rifiutare H_0) è quello di vedere se cade nella **zona di rifiuto** (o regione critica).

Questa zona rappresenta la "coda" estrema dei risultati che hanno probabilità totale pari al livello di significatività scelto.

Il valore della statistica utilizzata che delimita tale zona è chiamato **valore critico**. Nel nostro caso il valore che delimita la zona di rifiuto dello 0,05 è $z = 1,64$ (ricavabile dalla tavola della distribuzione normale). Rifiuteremo l'ipotesi nulla per tutti quei valori di z che cadono a destra di 1,64.

41



Poiché z calcolato ($z = 2,34$) $>$ z critico ($z = 1,64$) possiamo rifiutare H_0 .

42