Statistica inferenziale

Docente: Prof.ssa Paola Borrelli
paola.borrelli@unich.it

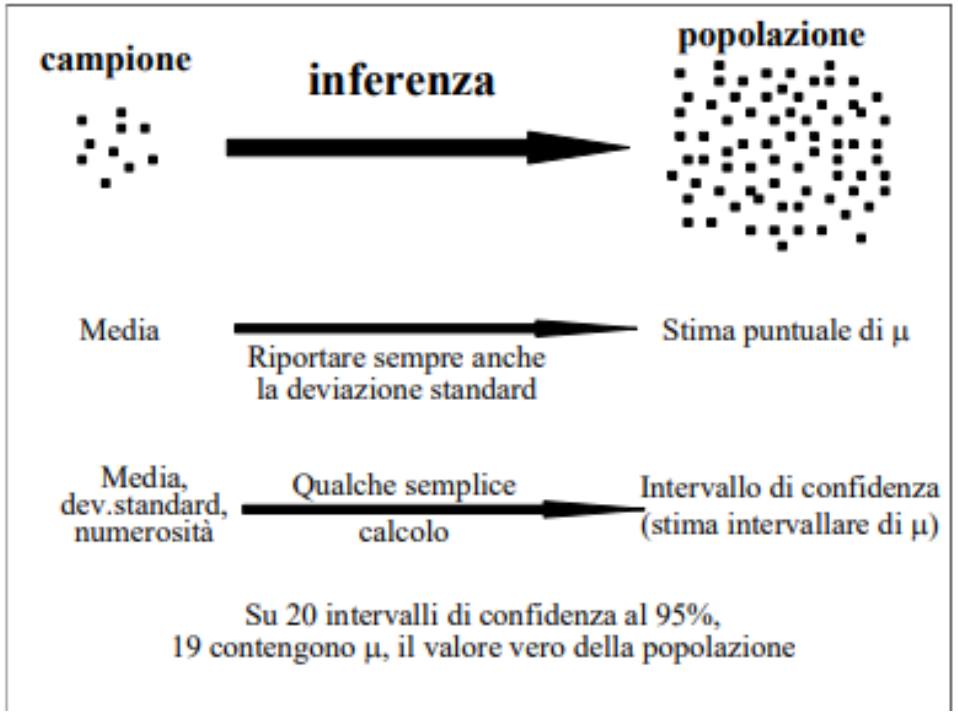
Statistica inferenziale

Insieme di procedimenti utili a trarre conclusioni sulla popolazione obiettivo

(stima dei parametri, verifica delle ipotesi)

Paola Borrelli, PhD

Stimare il parametro della popolazione e la precisione delle stime



Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.
Attraverso l'inferenza statistica si cerca di stimare e limitare la probabilità di commettere errori

Stima **puntuale**

La stima puntuale fornisce un singolo valore.
Tuttavia questo valore non coincide quasi mai con il valore vero (parametro) della popolazione

Stima **intervallare**

La stima intervallare fornisce un intervallo, che ha una predeterminata probabilità di contenere il valore vero della popolazione. Pertanto quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione

Di solito si individuano dei limiti all'interno dei quali è *verosimilmente* contenuto il **parametro** della popolazione

Media – E
Media + E

Si dice che il parametro giace all'interno dell'intervallo compreso tra tali limiti

Media – E <μ<Media +E

Stima **intervallare**

Media del FEV = 4.062

Deviazione standard FEV = 0.67

Soggetti adulti n= 57

Stima **intervallare**

Intervallo di confidenza al 95%

Media $\mp 1.96\ \sigma/\sqrt{n}$

3.89-4.24

Siamo sicuri al 95% che l'intervallo 3.89-4.24 contenga il vero valore di μ

La precisione della stima

si valuta attraverso **l'intervallo di confidenza**

è un intervallo costruito a partire dalla stima stessa che ha una certa probabilità di contenere il parametro della popolazione stimato

Verifica di ipotesi e test di significatività

Le ipotesi di ricerca formulate vengono verificate, mediante tecniche opportune, calcolando la probabilità di commettere un errore nel considerare validi per tutta la popolazione i risultati osservati in uno o più campioni

Sequenza logica di assiomi e decisioni nella verifica di ipotesi

- FORMULARE L'IPOTESI
- SCEGLIERE IL TEST STATISTICO PER VERIFICARLA
- IDENTIFICARE LA DISTRIBUZIONE TEORICA DEL TEST
- STABILIRE LA REGOLA DI DECISIONE
- PRENDERE UNA DECISIONE STATISTICA
  - Verificare o rifiutare l'ipotesi formulata

Le ipotesi

Confronto tra valori di una variabile osservati in due (o più) campioni ipoteticamente diversi o trattati in modo diverso

- Ipotesi nulla (H_0): ipotesi da verificare, non c'è differenza tra campioni
  - I valori della variabile sono uguali in quanto provengono dalla *stessa popolazione* di origine
  - La differenza osservata rientra nella casualità campionaria
  - E' sempre di uguaglianza, di indipendenza, di assenza di effetto
- Ipotesi alternativa (H_1): c'è differenza tra campioni
  - I campioni presentano valori della variabile diversi in quanto provenienti da *popolazioni diverse*

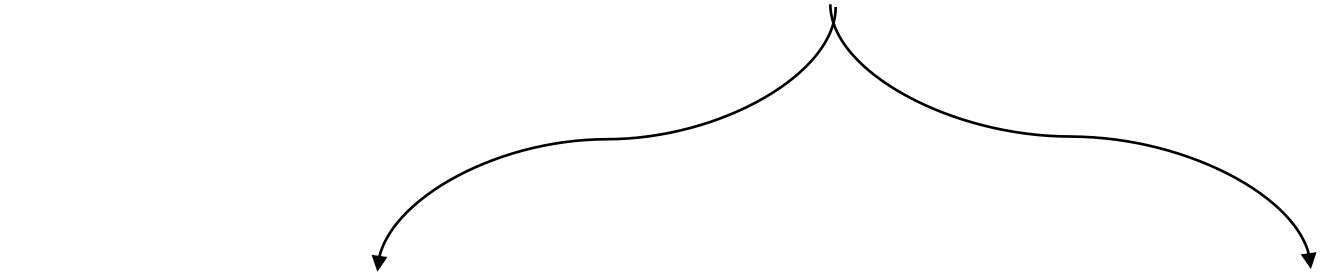
DETERMINARE IL TEST STATISTICO DA UTILIZZARE IN BASE A:

- tipo di variabile che si vuole studiare
- ASSUNZIONI:
  - distribuzione della variabile
  - ampiezza del campione

TEST STATISTICO

- Il valore della statistica test deve permettere al ricercatore di decidere se accettare/rifiutare l'ipotesi nulla
- La decisione si basa sulla probabilità associata al valore del test
  - è la probabilità teorica di rifiutare l'ipotesi nulla quando essa è vera
  - deriva dal calcolo del test statistico su infiniti campioni

Il risultato del test PUÒ portare al RIFIUTO o al NON RIFIUTO della ipotesi NULLA (H_0)



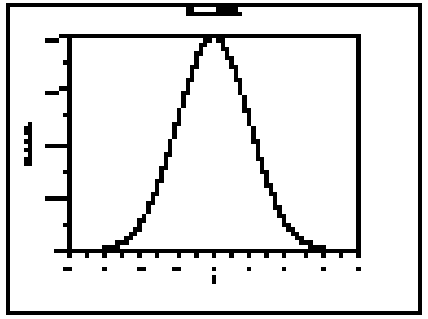
NON RIFIUTO

i dati non forniscono sufficiente evidenza per causare il rifiuto di H_0

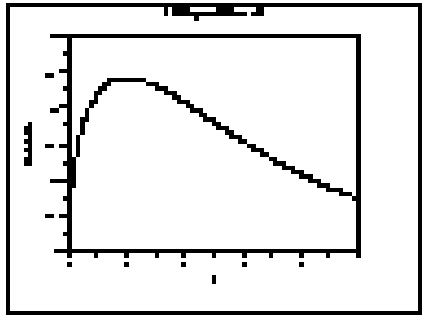
RIFIUTO

dati non sono compatibili con H_0 ma supportano un'altra ipotesi (H_A)...

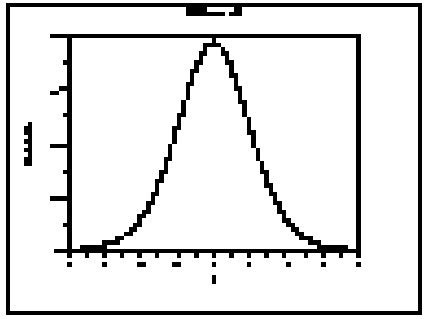
IDENTIFICARE LA DISTRIBUZIONE DEL TEST STATISTICO



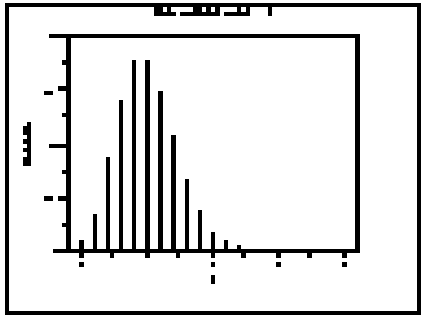
Normale



χ^2



t di Student



Poisson

DISTRIBUZIONI DI PROBABILITÀ TEORICA

STABILIRE LA REGOLA DI DECISIONE

Individuare nella distribuzione teorica di probabilità del test due distinte zone, o aree di valori, una di non rifiuto e una di rifiuto dell'ipotesi nulla

- **regione di accettazione (di non rifiuto)**: valori del test più probabili se $\mathbf{H_0}$ è vera
- **regione di rifiuto**: valori del test meno probabili se $\mathbf{H_0}$ è vera

STABILIRE LA REGOLA DI DECISIONE

- **regione di accettazione =** insieme di valori del test per i quali risultati diversi tra gruppi si possono considerare compatibili con la causalità
- **regione di rifiuto** = insieme di valori che permettono di rifiutare l'ipotesi nulla e quindi prendere in considerazione l'ipotesi alternativa

Parliamo di una distribuzione di probabilità, quindi la regione di rifiuto corrisponde alla probabilità di errore prescelta dal ricercatore

Livello di significatività

La regione di rifiuto viene stabilita sulla base del desiderato

livello di significatività α:

- rappresenta l'area sotto la curva della distribuzione del test delimitata dai valori (sull'asse x) che rappresentano la regione di rifiuto
- **È la probabilità di rifiutare l' H_0 quando essa è vera**
- poiché rifiutare un' H_0 quando è vera è un errore, è ragionevole rendere α piccolo (di solito 0.05)
- **α = ERRORE DI I TIPO**

PRENDERE UNA DECISIONE STATISTICA

Se il test cade nella regione di rifiuto ($P \leq 0.05$)

RIFIUTIAMO H_0

Se il test cade nella regione di accettazione ($P > 0.05$)

NON RIFIUTIAMO H_0

RILEVANZA CLINICA

La significatività statistica è una condizione preliminare, utile e indispensabile, per poter parlare dell'importanza che assume l'ampiezza della differenza riscontrata, cioè la sua **significatività o rilevanza clinica**

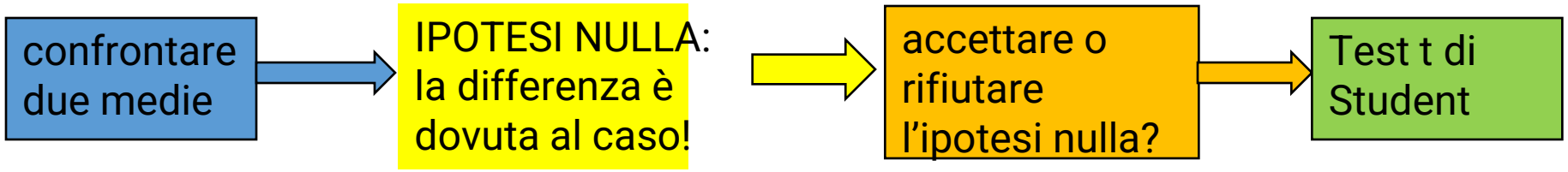
Test t di Student
Disegno a due popolazioni indipendenti

Serve per studiare la relazione tra una **variabile di risposta quantitativa** e una **variabile esplicativa qualitativa dicotomica**

Lo studio della relazione avviene attraverso **il confronto di medie** di due campioni dove **le osservazioni in un campione sono indipendenti dalle osservazioni nel secondo campione**

Infatti spesso il ricercatore ha l'esigenza di confrontare dati provenienti da **popolazioni diverse** allo scopo di evidenziare differenze e trarre conclusioni in merito a tali differenze.

test *t* Student per dati indipendenti



$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

H_0 NON ESISTE differenza tra l'età media dei soggetti malati e non

$$\mu_1 = \mu_2$$

H_A **ESISTE differenza** tra l'età media dei soggetti malati e non

$$\mu_1 \neq \mu_2$$

Assunzioni di applicabilità

- Variabile quantitativa
- 2 ‘gruppi’ diversi su cui è misurato la **stessa variabile**
- la variabile è distribuita in modo **normale** nelle 2 ‘popolazioni’ cui appartengono i 2 gruppi
- varianza della variabile è omogenea tra le due ‘popolazioni’

$$\bar{x}_1 \; e \; \bar{x}_2$$

$$\mu_1 \; e \; \mu_2$$

$$\sigma^2_1 = \sigma^2_2$$

$|t_{calcolato}| < |t_{tabulato}|$ → **Non Rifiuto H_0**

$|t_{calcolato}| > |t_{tabulato}|$ → **Rifiuto H_0**

ma questo è PASSATO!!!

Nella pratica, oggi, esistono numerosi software per effettuare i test e qualsiasi sia il software utilizzato, esso esplicita anche (o solamente, come Excel) un **valore di ‘p’ o p-value**

p-*value*

- stima quantitativa della **probabilità** che le differenze osservate siano dovute al caso
- È una probabilità, quindi può assumere solo valori compresi tra 0 e 1
- Un valore di p che si avvicina a 0 indica una **bassa probabilità** che la differenza osservata possa essere attribuita al caso

p-value

- Qualsiasi **programma** statistico darà come risultato una probabilità (p-value)
- Tale probabilità di errore deve poi essere valutata in base al **livello di significatività** scelto
- Il livello di significatività può essere **scelto** dallo sperimentatore
- Di solito si sceglie un livello di probabilità di **0.05 (5%)** o di 0.01 (1%)

Significatività

- Il livello di significatività 5% viene adottato molto frequentemente in quanto si ritiene che il rapporto **1/20** (cioè 0.05) sia sufficientemente piccolo da poter concludere che sia piuttosto improbabile che la differenza osservata sia dovuta al semplice caso
- In effetti, la differenza potrebbe essere dovuta al caso, e lo sarà 1 volta su 20
- Tuttavia, questo evento è **improbabile**
- Ovviamente, se si vuole escludere con maggiore probabilità l'effetto del caso, si adotterà un livello di significatività inferiore (es. 1%)

Significatività

- Quindi, se l'ipotesi zero viene respinta al livello di significatività 5%, allora abbiamo il 5% di probabilità di respingere un'ipotesi zero che era vera; se l'ipotesi zero viene respinta al livello di significatività 1%, allora abbiamo l'1% di probabilità di respingere un'ipotesi zero che era vera

- In generale, se l'ipotesi zero viene respinta al livello di significatività n%, allora abbiamo n% di probabilità di respingere un'ipotesi zero che era vera

p-value

- Ipotizziamo di aver scelto come livello di significatività p=0.05
- Se otteniamo dal test p=0.032, cioè un p **MINORE** di quello scelto, possiamo concludere che la differenza tra i gruppi che abbiamo sottoposto al test è **SIGNIFICATIVA**
- Se otteniamo dal test p=0.09, cioè un p **MAGGIORE** di quello scelto, possiamo concludere che la differenza tra i gruppi che abbiamo sottoposto al test **NON è SIGNIFICATIVA**

Statisticamente significativo

Quindi:

statisticamente significativo vuol dire che ciò che è stato osservato è difficilmente dovuto al caso

Test statistico

- Un test di significatività **non può provare con certezza** che una ipotesi zero è vera o falsa

- Rimane sempre un **margine d'errore** (direttamente proporzionale a p)

- Può fornire una **indicazione della forza** con cui i dati contrastano l'ipotesi zero

Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev
0	1584,0000	90116,0000	56,8914	212,1701	14,5661
1	120,0000	8266,0000	68,8833	121,0619	11,0028

	Minimum	25%	Median	75%	Maximum	Mode
0	21,0000	46,0000	59,0000	68,0000	92,0000	65,0000
1	29,0000	62,5000	71,0000	76,0000	90,0000	72,0000

T-Test

	Method	Mean	95% CL Mean		Std Dev
Diff (Group 1 - Group 2)	Pooled	-11,9919	-14,6541	-9,3297	14,3457
Diff (Group 1 - Group 2)	Satterthwaite	-11,9919	-14,1039	-9,8799	

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1702	-8,83	0,0000
Satterthwaite	Unequal	152.50	-11.22	0.0000

Età e Ricoveri per malattie cardiovascolari

IL TEST DEL CHI-QUADRATO

Verifica l'associazione tra due variabili qualitative

$$\chi^2 = \sum_{tuttecelle} \frac{(fr.osservate - fr.attese)^2}{fr.attese}$$

Il test si basa sul **confronto** tra le frequenze **O**sservate e quelle **A**ttese **(O-A).**

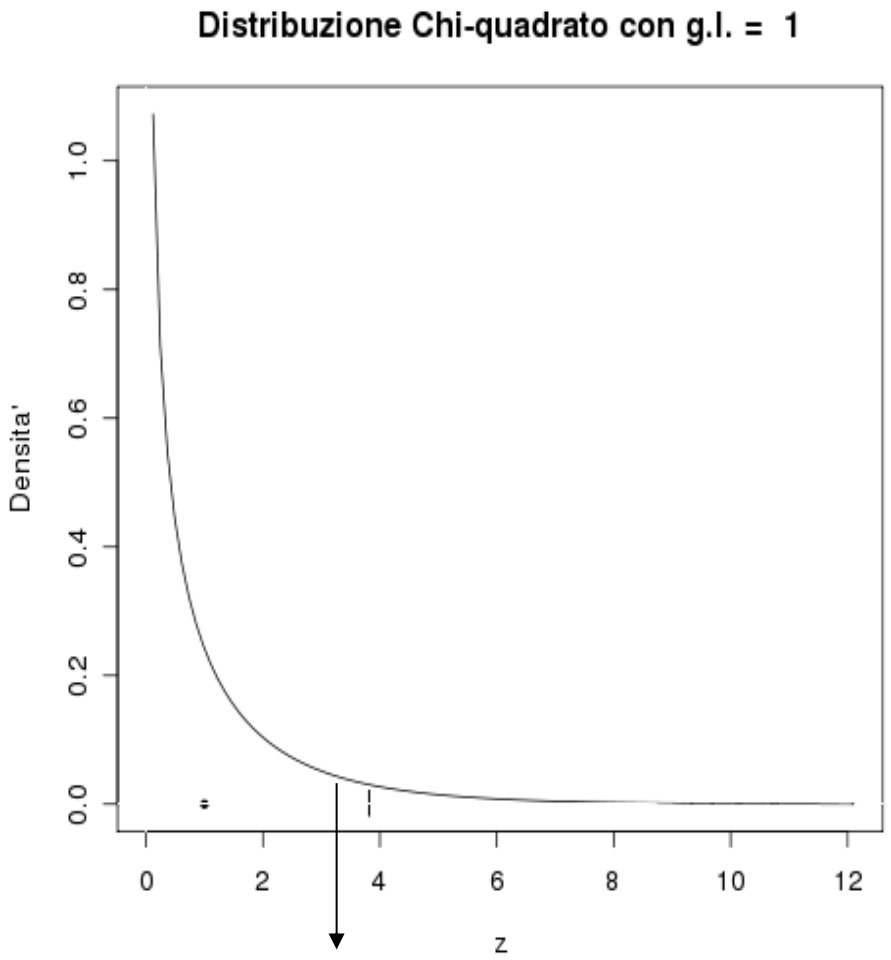
La statistica campionaria alla base del test è

$$\sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$

dove la sommatoria è estesa a tutte le k celle delle tabelle degli Osservati e degli Attesi

Se H_0 è vera e le variabili sono indipendenti la differenza tra gli osservati e gli attesi è effetto del caso.

Si procede dunque a calcolare la **probabilità** di ottenere un valore della statistica campionaria uguale a quella calcolata o ancora più grande per effetto del caso



3,84

α=0.05

- tante curve a seconda dei gradi di libertà
- il X^2 è sempre positivo
- varia tra 0 e $+\infty$

Ipertensione e Ricoveri per malattie cardiovascolari

Hpt	RIC_CVD 0	1	Total
0	1,218	59	1,277
	1,187.1	89.9	1,277.0
	76.89	49.17	74.94
1	366	61	427
	396.9	30.1	427.0
	23.11	50.83	25.06
Total	1,584	120	1,704
	1,584.0	120.0	1,704.0
	100.00	100.00	100.00

Pearson chi2(1) = 45.6669 Pr = 0.000

CORREZIONE DI YATES

- Quando, in tabelle 2x2 le frequenze sono **basse** (ma sempre >5) è consigliabile utilizzare un correttivo, detto **di Yates**. Anche in questo caso avremo una p…

$$\chi^2 = \sum_{\text{tuttecelle}} \frac{\left(\left|\text{fr. osservate} - \text{fr. attese}\right| - 0.5\right)^2}{\text{fr. attese}}$$

TEST ESATTO DI FISHER

- Quando la dimensione campionaria è piccola (celle con 0-5 elementi), è possibile elencare tutte le possibili combinazioni delle osservazioni e quindi calcolare le probabilità esatte associate a ogni possibile combinazione di dati

$$p = \frac{(a+b)!*(c+d)!*(a+c)!*(a+d)!}{a!*b!*c!*d!*(a+b+c+d)!}$$

TEST DI McNEMAR

- Quando è necessario confrontare risultati prima e dopo relativi agli stessi individui o risultati relativi a uno studio in cui i dati sono appaiati

$$\chi^2 = \frac{(|b-c|-1)^2}{(b+c)}$$