

Annotazioni su alcuni dei quesiti del compito del 24/02/2012

Due quesiti erano impostati su dati in una tabella come la seguente:

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>				
Gruppo	X	n	%	M	S	Casi con Y=1				
A	$x_0 - x_1$	nA	pA	MA	SA	rA				
B	$x_1 - x_2$	nB	pB	Mb	SB	rB				
		n								

La tabella mostra dati osservati in un campione di pazienti suddivisi in due gruppi: A e B; i due gruppi sono determinati sulla base di un carattere quantitativo continuo X, di cui cioè consideriamo 2 classi (colonna *b*). Le colonne *c* e *d* riportano rispettivamente le frequenze assolute e percentuali di soggetti in ciascun gruppo. Per ogni gruppo si conosce la media esatta M e la deviazione standard S del carattere X (colonne *e* ed *f*). Si ha inoltre (colonna *g*) il numero di soggetti per cui risulta che un secondo carattere, Y, dicotomico, risulta pari a 1.

1. Primo dei quesiti con i dati in tabella, si chiede di:

- Valutare la media di X complessiva → media ponderata

► infatti il metodo di usare un valore centrale per rappresentare i valori osservati di X nella classe è da considerarsi un'approssimazione, da utilizzarsi in mancanza di informazioni più precise; in questo caso, conoscendo Media M e frequenza n di ciascun gruppo, possiamo ricostruire l'ammontare esatto del carattere X nel gruppo – è proprio l'operazione che si fa calcolando la media ponderata. Notare che questo discorso è valido anche se si hanno più di 2 gruppi, e che la media ponderata si fa anche per calcolare la media complessiva fra $k > 2$ gruppi.

- Valutare la classe in cui cade la mediana di X complessiva → vedere l'ultima esercitazione; la mediana cade nella classe a cui corrisponde la prima frequenza cumulata maggiore di 50%
- Confrontare la variabilità di X nei due gruppi → tramite il calcolo del coefficiente di variazione di ciascun gruppo (come al solito!!!)
- Confrontare la densità di frequenza di X nei due gruppi → basta calcolarle
- Valutare la probabilità complessiva che $Y=1$ → ovviamente, $(rA+rB)/n$

2. Quesito sull'utilizzo di alcuni indici sintetici: fonte di errore è una mancanza di piena comprensione di cosa sia il "confronto fra gruppi" rispetto alla "associazione fra caratteri". Un confronto fra due o più gruppi rispetto a un carattere Y è una valutazione dell'associazione fra il carattere che definisce i gruppi (chiamiamolo X) e il carattere Y. Ma se due caratteri X e Y sono entrambi continui, NON si stanno confrontando gruppi.

► Quando si studia, è utile farsi degli esempi: confronto la pressione Y (carattere continuo) fra Maschi e Femmine = valuto se la pressione (Y) e il sesso (X) sono associati; valuto se la pressione Y (carattere continuo) e l'età X (carattere continuo) sono associati: NON sto confrontando gruppi; quando però l'età è ridotta in classi, ad esempio "giovani" "maturi" e "anziani" e voglio confrontare le medie della pressione Y nei 3 gruppi ho di nuovo un confronto fra gruppi (che si fa con l'Anova – vedere la dispensa di inferenza)

3. Secondo quesito con i dati in tabella, si chiede di:

- Dire di quanto la probabilità che $Y=1$ è maggiore nel gruppo A rispetto al gruppo B → calcolare il Risk Ratio
- Essere d'accordo o meno sul fatto che la differenza così valutata fosse "significativa"

► la significatività è un concetto statistico preciso e si valuta attraverso un test di ipotesi; quindi bisognava fare il test e NON esprimere un giudizio sulla RILEVANZA della differenza (vedere la distinzione fra rilevanza e significatività, affrontata a lezione). Bisognava fare il test anche per valutare le altre affermazioni. Il test è ovviamente il Chi-Quadrato, ed è stato fatto un ulteriore esempio di svolgimento nell'ultima esercitazione.

- Essere d'accordo o meno con affermazioni rispetto al risultato del test.

► Vediamo varie affermazioni corrette o errate in merito alle conclusioni di un test di livello di significatività α ; i concetti sono generali e dunque inseriamo qui anche altre affermazioni che erano proposte per un altro dei quesiti del compito – dove non c'era un confronto fra gruppi, ma un test su una ipotesi nulla basata su un valore di riferimento (la media μ di una popolazione pari a un valore μ_0).

- SITUAZIONE DI RIFIUTO DELL'IPOTESI NULLA: sono affermazioni corrette e coerenti:
 - L'ipotesi nulla H_0 è rifiutata al livello di significatività α
 - Il valore della statistica test calcolato cade nella regione di rifiuto
 - Il p-value risulta inferiore a α
 - I dati supportano l'ipotesi alternativa (nel caso del confronto fra gruppi ovvero della valutazione della presenza di una associazione, H_0 è sempre un'ipotesi di indipendenza, quindi: i dati supportano l'ipotesi che vi sia una relazione fra X e Y, o “indicano la presenza di una relazione”)
 - I dati sono *significativamente* diversi dal valore di riferimento / La differenza osservata fra i due gruppi è *significativa*
- SITUAZIONE DI ACCETTAZIONE DELL'IPOTESI NULLA: sono affermazioni corrette e coerenti:
 - L'ipotesi nulla H_0 è accettata al livello di significatività α
 - Il valore della statistica test calcolato cade nella regione di accettazione
 - Il p-value risulta superiore a α
 - I dati confermano l'ipotesi nulla (*)
 - I dati NON supportano l'ipotesi alternativa (es: che vi sia una relazione fra X e Y)
 - La differenza osservata fra i due gruppi NON è *significativa*
 - La differenza osservata fra i due gruppi è attribuibile al caso, non riproducibile in un altro esperimento simile

(*) Attenzione: NON E' CORRETTO affermare che:

- I dati dimostrano che non vi è relazione fra X e Y

► infatti nella logica del test statistico i dati vengono utilizzati per “falsificare” H_0 , cioè per dimostrare che H_0 è falsa; possono o riuscirci (rigo H_0 , dati significativi) oppure no (accetto H_0 , dati non significativi); ma l'accettazione di H_0 NON vuol dire che H_0 è dimostrata essere vera. Può sembrare una sottigliezza ma è un concetto preciso e importante, e coloro che in futuro parteciperanno a qualche progetto di ricerca vedranno bene questa importanza.

Raccomandazione: Studiare in modo corretto significa comprendere bene le relazioni fra tutti gli aspetti del test, compresa la relazione fra test e IC, mentre è del tutto inutile cercare di apprendere quanto sopra in maniera mnemonica!!

4. Proseguiamo quindi sull'inferenza. Altri due quesiti vertevano sulla relazione fra Intervallo di Confidenza e Test di ipotesi.

Questa relazione è stata ricordata in sede di esame:

Per un parametro θ di cui si fa la stima mediante IC e su cui si fa un test per rigettare $H_0: \theta = \theta_0$:

IC al livello $(1 - \alpha) \leftrightarrow$ test a due code al livello di significatività $\alpha \leftrightarrow$ test a una code al livello $\alpha/2$

IC al livello $(1 - 2\alpha) \leftrightarrow$ test a una coda al livello di significatività α

Se θ_0 è contenuto dentro l'IC il test accetta H_0

Se θ_0 non è contenuto dentro l'IC il test rifiuta H_0

e viceversa, ovviamente:

I valori θ_0 accettati dal test sono dentro l'IC

I valori θ_0 rifiutati dal test sono esterni all'IC

Quindi se θ_0 è dentro l'IC, il test accetta H_0 , e quindi vuol dire che la statistica test cade in zona di accettazione, che il p-value è $> \alpha$, etc.

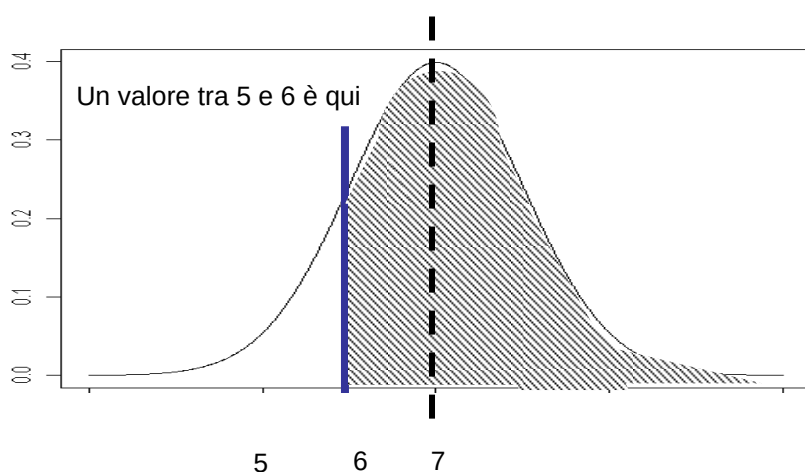
5. L'esercizio sulla curva Normale richiedeva solo di conoscere e utilizzare la nozione di primo e terzo Quartile, e per gli altri punti di risolvere diversi calcoli – si individuano sul grafico le aree da calcolare, si standardizzano i valori, si usano le tavole ... Ad esempio:

La Normale ha media 7 e deviazione standard 2. Si chiede se è vero o falso che:

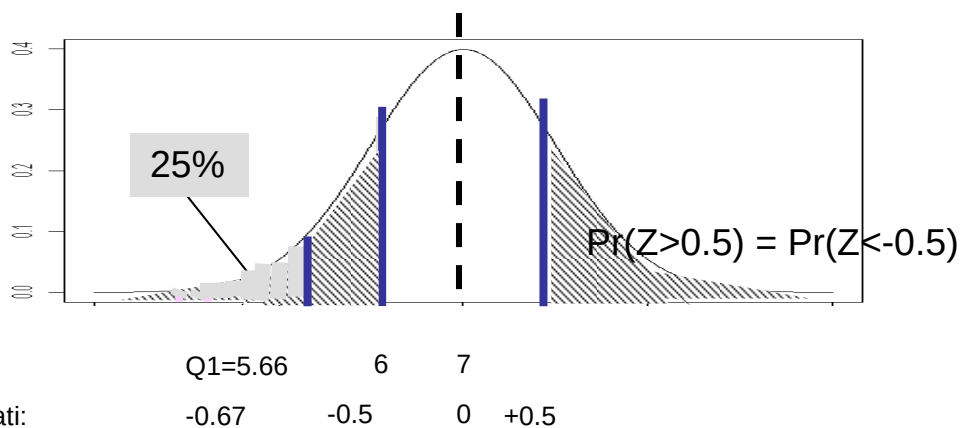
- 1 Un individuo su quattro ha un valore di X almeno pari a 5 o 6
- 2 Circa il 30% di individui ha un valore di X minore di 6

La prima affermazione ci porta subito a pensare ai quartili; attenzione però: si chiede quale è quel valore x tale che $\Pr(X > x) = 25\%$, e quindi dovremmo calcolare $x = Q_3$ – non Q_1 !! $Q_1 = 7 - 0.67 \cdot 2 = 5.66$ MA 3 individui su 4 hanno $X > 5.66$. quindi l'affermazione è FALSA.

► Notare che facendo un grafico della curva la risposta poteva essere data anche senza pensare ai quartili: bastava vedere che si stava chiedendo di valutare un'area maggiore della metà della curva (poiché i valori tra 5 e 6 stanno comunque prima dell'asse centrale), quindi sicuramente maggiore di 25%.



Per la seconda affermazione: notiamo intanto che NON è in contraddizione col fatto che $Q_1 = 5.66$. Infatti 6 è maggiore di 5.66, quindi può essere che fra 5.66 e 6 vi sia un'ulteriore 5% di frequenza. Facciamo quindi il calcolo: valore standardizzato corrispondente a $x=6$ è: $z = (6-7)/2 = -0.5$. Sulle tavole non trovo i valori negativi di z , quindi guardo l'area per $z=0.5$, che vale 0.691. Ricordiamo che essa è l'area FINO A z , a me serve l'area nella coda (nel grafico qui sotto, le code sono tratteggiate), quindi la calcolo come $1 - 0.691 =$ circa 30%: l'affermazione è VERA.



6. Esercizio di probabilità, impostazione: gli eventi: C =concepimento, A =aborto, M =diventare madre: $M = (C \cap \text{non}A)$ dove $\text{non}A$ indica il complementare di A ; non diventare madre: $\text{non}M = \Omega - M$ (e quindi $\Pr(\text{non}M) = 1 - \Pr(M)$ e anche, direttamente: $\text{non}M = (\text{non}C) \cup (C \cap A)$)