Inferenza Statistica I B

Prova A del 25 giugno 2002

1. Si è osservato il livello di un certo parametro biologico in 23 individui affetti da una certa patologia. E' risultato

$$\sum_{i=1}^{23} y_i = 5894{,}4 \text{ e } \sum_{i=1}^{23} y_i^2 = 1533753$$

dove con y_i, $i = 1, \ldots, 23$, è stato indicato il valore osservato sull'i-simo soggetto. Il valore medio del parametro biologico considerato in soggetti sani di pari età è 270. Supponendo che le osservazioni siano assimilabili ad un campione di osservazioni normali indipendenti ed identicamente distribuite, dire se sulla base dei risultati osservati si può affermare che la presenza della patologia in esame non altera il livello del parametro considerato (rispondere sia utilizzando un intervallo di confidenza che un test).

[**Schema di risposta**] Calcoliamo innanzitutto le stime della media e della varianza della distribuzione normale, indicate al solito nel seguito utilizzando rispettivamente μ e σ^2

$$\hat{\mu} = \overline{y} = \sum_{i=1}^{n} y_i/n = \frac{5894{,}4}{23} = 256{,}28$$

e

$$\hat{\sigma}^2 = s^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2/(n-1) = \frac{23}{22}\left(\frac{1533753}{23} - 256{,}28^2\right) = 1051{,}18$$

dove, per il calcolo di s^2 abbiamo utilizzato la formula

$$\sum_{i=1}^{n} (y_i - \overline{y})^2/n = \left(\sum_{i=1}^{n} y_i^2/n\right) - \overline{y}^2.$$

Un intervallo di confidenza per la media può essere calcolato come (si veda l'Unità E dei lucidi)

$$\overline{y} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

Ponendo $\alpha = 0{,}05$, troviamo $t_{22,0.975} = 2{,}07$ e quindi un intervallo al 95% per la media è

$$256{,}28 \pm 2{,}07\sqrt{\frac{1051{,}18}{23}} = [242{,}29 \text{ - } 270{,}27]$$

Il valore medio nei soggetti sani (270) appartiene ma solo "per un pelo" all'intervallo di confidenza calcolato. Quindi la storia che ci racconta il "calcolo" precedente è che è possibile che nei soggetti con la patologia considerata ci sia una alterazione del parametro biologico considerato (visto che la "maggior parte dell'intervallo di confidenza è "sotto" 270) ma che l'evidenza contenuta nei dati a favore di questa ipotesi non è molto forte e conclusiva (visto che, per quanto "per un pelo", 270 appartiene all'intervallo di confidenza).

Volendo procedere con un test, possiamo utilizzare i dati per verificare l'ipotesi nulla

$$H_0 : \mu = 270$$

verso l'ipotesi alternativa

$$H_1 : \mu \neq 270.$$

Una appropriata statistica test è

$$t_{oss} = \frac{\sqrt{n}(\overline{y} - 270)}{s} = \frac{\sqrt{23}(256{,}28 - 270)}{\sqrt{1051.18}} = -2{,}03$$

Quando è vera H_0 t_{oss} è una determinazione di una variabile casuale t di Student con 22 $(= n-1)$ gradi di libertà. Seguendo un approccio del tipo accetto/rifiuto con α (= probabilità di rifiutare H_0 quando H_0 è vera) posto uguale a 0,05 dovremmo accettare H_0 se $-t_{n-1,0.975} \leq t_{oss} \leq t_{n-1,0.975}$ cosa che in questo caso accade. In realtà vediamo che in questo caso $t_{oss} \approx -t_{22,0.975}$. Quindi non possiamo andare oltre ad una "dubbiosa" accettazione (o, che è lo stesso, ad un "dubbioso" rifiuto).

2. Per verificare se l'esposizione ad un certo agente inquinante altera la probabilità di sviluppare una certa patologia respiratoria sono stati raccolti i dati riassunti nella seguente tabella

soggetti esposti all'inquinante	patologia	
	assente	presente
NO	32	21
SI	24	36

Analizzare la tabella precedente con gli stumenti a voi noti.

[Schema di risposta] Il problema a cui si vuole rispondere con dei dati di questo tipo consiste nel cercare di capire se l'esposizione all'inquinante altera o non altera la probabilità di sviluppare la patologia considerata ovvero se nella popolazione da cui sono tratti gli individui per cui sono disponibili i dati esiste o non esiste indipendenza tra esposizione all'inquinante e presenza della malattia.

La statistica test classica per verificare questa ipotesi è l'X^2 di Pearson (si veda l'Unità D dei lucidi) che in questo caso vale 4,67. Sotto l'ipotesi di indipendenza tra i due fattori (esposizione e malattia) questo valore può essere visto come una determinazione di una variabile casuale χ^2 con un grado di libertà (infatti in generale i gradi di libertà sarebbero dati da

$$[(\text{num. righe}) - 1] \times [(\text{num. colonne}) - 1]$$

e nel nostro caso abbiamo una tabella con due righe e due colonne). In presenza di dipendenza ci aspettiamo viceversa di osservare valori "grandi" di X^2.

Il valore osservato della statistica test, 4,67, risulta più grande del 95% percentile di una variabile χ^2 con un grado di libertà che vale 3,84. Possiamo quindi concludere che, visto che il valore osservato cade in una regione in cui non ci aspetteremmo di trovarlo se ci fosse indipendenza, i dati suggeriscono la presenza di una qualche forma di relazione tra esposizione all'inquinante e presenza della patologia considerata.

Inferenza Statistica I B

Prova B del 25 giugno 2002

1. In un sondaggio condotto su 100 docenti dell'Ateneo di Padova all'inizio della scorsa settimana è stato rilevato che 58 avevano intenzione di partecipare all'elezione del Rettore mentre gli altri 42, per vari motivi, non si sarebbero recati a votare. Per la validità dell'elezione è necessario che almeno il 50% degli elettori si rechi alle urne. Sulla base del sondaggio è possibile affermare che: (a) è sicuro che il quorum verrà raggiunto; (b) è molto plausibile che il quorum verrà raggiunto; (c) è poco plausibile che il quorum non verrà raggiunto; (d) i dati non ci permettono di scegliere tra nessuna delle alternative precedenti.
   Rispondere sia utilizzando un intervallo di confidenza che un appropriato test.

   [**Schema di risposta**] Poniamo

   $$n = (\text{num. individui intervistati}) = 100$$

   e

   $$\vartheta = \left(\begin{array}{c} \text{percentuale di docenti che hanno intenzione di} \\ \text{andare a votare} \end{array} \right)$$

   Supponendo che sia possibile assumere che le risposte dei soggetti intervistati siano indipendenti ed identicamente distribuite il numero di intenzioni di voto espresse (58) può essere visto come una determinazione di una variabile casuale binomiale con probabilità di successo ϑ e numero di prove uguale a 100. La stima di ϑ vale

   $$\hat{\vartheta} = \frac{\text{num. intenzioni di voto}}{\text{num. intervistati}} = \frac{58}{100} = 0{,}58.$$

   Un intervallo di confidenza per ϑ può essere calcolato come (si veda l'Unità B dei lucidi)

   $$\hat{\vartheta} \pm z_{1-\alpha/2}\sqrt{\hat{\vartheta}(1-\hat{\vartheta})/n}$$

   Ponendo $\alpha = 0{,}05$ otteniamo $z_{0{,}975} = 1{,}96$ e quindi l'intervallo di confidenza diventa

   $$0{,}58 \pm 1{,}96\sqrt{0{,}58 \times 0{,}42/100} = [0{,}46 \text{ - } 0{,}68]$$

   Questo "calcolo" mostra che valori della percentuale di votanti inferiori al 50% non possono essere esclusi sulla base dei dati. Quindi il raggiungimento del quorum necessario per rendere le elezioni valide è incerto. Può d'altra parte essere osservato, che la "maggior parte" dell'intervallo di confidenza si estende su valori superiori al 50% dei votanti. Non possiamo quindi neanche escludere la possibilità che il quorum venga raggiunto.

   Volendo utilizzare un test potremmo considerare l'ipotesi

   $$H_0 : \vartheta \leq 0{,}5$$

   contro l'alternativa

   $$H_1 : \vartheta > 0{,}5$$

   e la relativa statistica test

   $$z = \frac{\hat{\vartheta} - 0.5}{\sqrt{0.5 \times 0.5/100}} = 1{,}6.$$

   Questo valore va confrontato con i valori che ci aspetteremmo da una normale standard sapendo che

(a) valori più bassi di quelli previsti da una $N(0,1)$ ce li aspettiamo se $\vartheta < 0{,}5$;

(b) valori "uguali" a quelli generati da una $N(0,1)$ ce li aspettiamo se $\vartheta = 0{,}5$;

(c) valori più alti di quelli previsti da una $N(0,1)$ ce li aspettiamo se $\vartheta > 0{,}5$.

Ovviamente i primi due casi sono a favore di H_0, l'ultimo caso a favore di H_1.

Ora, 1,6 è all'incirca il quantile 0,945 di una $N(0,1)$. Quindi il valore osservato è abbastanza "grande" ma non "enormemente" grande. In conclusione sembra ragionevole concludere a favore di una "dubbiosa" accettazione o, equivalentemente, di un "dubbioso" rifiuto di H_0. La conclusione non cambierebbe se avessimo formulato il problema come uno di verifica di ipotesi bidirezionale.

In conclusione, tra le ipotesi formulate nel testo del problema la (d) sembra essere la più "vicina" ai dati seguita dalla (b).

2. Per studiare l'effetto di 4 differenti tecniche di coltivazione di un certo ortaggio, un appezzamento di terreno è stato diviso in 24 sotto-appezzamenti tutti della stessa dimensione. I 24 appezzamenti sono poi stati divisi casualmente in 4 gruppi ciascuno composto da 6 sotto-appezzamenti differenti. I 6 sotto-appezzamenti del 1° gruppo sono stati coltivati con la prima tecnica, i 6 del 2° gruppo con la seconda tecnica e così via. Su ciascun sotto-appezzamento è poi stata rilevata la quantità di ortaggio prodotta. Supponendo che valgano le ipotesi alla base dell'analisi della varianza con un criterio di classificazione e sapendo che, con i dati rilevati, la decomposizione della varianza totale

$$\text{(varianza totale)} = \text{(varianza entro i gruppi)} + \text{(varianza tra i gruppi)}$$

è risultata essere uguale a

$$1327{,}12 = 976{,}08 + 351{,}04$$

dire se è possibile afferermare che le tecniche di coltivazione considerate hanno una efficienza differente.

[**Schema di risposta**] Per l'analisi della varianza con un criterio di classificazione si veda l'Unità G dei lucidi. La statistica test per verificare, all'interno delle assunzioni fatte, l'ipotesi che non ci siano differenze tra le medie delle quantità di ortaggio prodotto utilizzando le 4 differenti tecniche di coltivazione è

$$F_{oss} = \left(\frac{\text{varianza tra i gruppi}}{\text{varianza entro i gruppi}}\right)\left(\frac{n-k}{k-1}\right)$$

dove n indica il numero totale di osservazioni (24 in questo caso) e k il numero dei gruppi (4 nella situazione in esame). Quindi, con i dati considerati,

$$F_{oss} = \frac{351{,}04/3}{976{,}08/20} = 2{,}4.$$

Nell'ipotesi di nessuna differenza tra le tecniche di coltivazione questo valore può essere visto come una determinazione di una variabile casuale F di Snedecor con 3 e 20 gradi di libertà. Viceversa se ci sono differenze tra i gruppi ci aspettiamo di osservare valori più grandi di quelli previsti dalla variabile casuale menzionata.

In questo caso, possiamo osservare che F_{oss} è risultato molto vicino al valore del 90-simo percentile di una F di Snedecor con 3 e 20 gradi di libertà (il percentile infatti vale 2,38). Siamo

quindi nella coda destra della distribuzione ma ancora su valori "possibili". La conclusione è quindi di accettazione magari dubbiosa e da confermare con altri studi.

Volendo fare un test accetto/rifiuto garantendoci, ad esempio, che la probabilità di accettare H_0, l'ipotesi di non differenza tra l'efficienza delle quattro tecniche, quando H_0 è vera sia uguale al 95% dovremmo accettare H_0 quando F_{oss} risulta minore del percentile 95-simo di una F con 3 e 20 gradi di libertà e rifiutare altrimenti. In questo caso, la procedura ci porterebbe ad accettare visto che il percentile 95-simo di una F con 3 e 20 gradi di libertà vale 3,1.