1 Correzione pre-appello SF1 Statistica II parte 16/04/2022 - canale MZ

Es. 1.1 . Si stimi il parametro A pari alla cifra meno significativa del proprio numero di matricola. Esplicitare la risposta.

R: 1.1 La cifra meno significativa per un numero intero è la cifra più a destra che non sia nulla. Esempio:

$$123456 \to A = 6 \qquad 123450 \to A = 5$$

Nota: se si volesse rappresentare il caso in cui la cifra meno significativa sia davvero 0, si dovrebbe scrivere il numero in notazione decimale o scientifica:

$$1.23450 \qquad \text{oppure} \qquad 1.23450 \cdot 10^5$$

Es. 1.2 . Un fioraio controlla l'andamento delle vendite negli ultimi 5 mesi in particolare mette in un istogramma il numero di rose vendute (A e' il parametro del primo esercizio): Ipotizza che le vendite in ciascun mese abbiano una probabilità costante. Alternativamente ipotizza che le vendite siano equiprobabili tranne che a Febbraio, per un picco di vendite a San Valentino. Q1.1 Si effettuino due test del χ^2, uno per ognuna delle due ipotesi, e si commenti opportunamente il risultato indicando quale delle due ipotesi è più credibile (ci si limiti al confronto dei χ^2 e all'esito del test e non si confrontino le verosimiglianze) Si usi un livello di confidenza del 99% Q1.2 Quante rose in più si può stimare abbia venduto in febbraio? Per la risoluzione si usi la tabella del χ^2.

Mese	Rose Vendute (n_i)
Gennaio	25
Febbraio	$46 + A$
Marzo	32
Aprile	26
Maggio	23

R: 1.2 Q1.1 Prima ipotesi. In questo caso si intende che la probabilità é uniforme per tutti 5 i mesi[1]. Allora, per ogni mese $p = 0.2$ e considerando (e.g. $A = 9$) che il numero totale di rose vendute e' 171, allora $n^* = 171 \cdot 0.2 = 34.2$. Per applicare un test del χ^2 anche in forma approssimata, è necessario valutare che incertezza usare. Si puo' procedere o stimando la incertezza dai dati, oppure a partire dalla ipotesi. In questi casi, non avendo una vera e propria stima sperimentale delle incertezze delle misure, ma solo quella legata alla ipotesi, utilizziamo l'incertezza della ipotesi (metodo di Neyman). Da questo possiamo calcolare la incertezza su n^* che va assunta Binomiale:

$$s^2_{n*} = N\, n^*\, (1 - n^*) = 27.4$$

e da questa anche i termini del $\chi^2_i = \left(\frac{n_i - n^*}{s_{n^*}}\right)^2$. Si riportano i dati in Figura 1. La distribuzione di riferimento del χ^2 ha $k = N - \nu = 5 - 1 = 4$ gradi di libertà, in quanto per stimare n^* si e' usato il fatto che $N = 171$. Si noti come il dato di Febbraio sia in effetti il piu' anomalo, con un $\chi^2_i \gg 1$. Anche i valori degli altri dati non sono troppo regolari e quindi possiamo ipotizzare comunque che, al di là del test, la credibilità della ipotesi sia difficile da valutare. Il valore critico di riferimento corrispondente ad un livello di confidenza del 99% e': $\chi^2_0 = 13.28$. Siccome $\chi^2_m = 23.79 > 13.28$ allora l'ipotesi nulla risulta rigettata. Ad una conclusione simile si sarebbe arrivati attraverso altre due vie: utilizzando come stima di incertezza per la singola classe la

[1] per altro, senza correggere per il numero diverso di giorni tra i mesi, cosa che non era stata esplicitata e che per fortuna nessuna ha fatto

Mese	ni	pi	Npi	s^2	chi^2
gennaio	25	0,2	34,2	27,4	3,09
febbraio	55	0,2	34,2	27,4	15,81
marzo	32	0,2	34,2	27,4	0,18
aprile	36	0,2	34,2	27,4	0,12
maggio	23	0,2	34,2	27,4	4,58
	171	1	171		23,79

Figura 1: Tabella di riferimento Ipotesi 1

$s^2_{n*} = N\, n_i\, (1 - n_i)$ o alternativamente usando l'approssimazione Poissoniana delle binomiale, ovvero usando: $s^2_{n*} = n^* = 34.2$. Tuttavia in alcuni casi quest'ultima approssimazione ha portato all'accettazioen anche di questa ipotesi. In questo caso la scelta corretta era comunque quella relativa alla statistica binomiale, le altre, se giustificate sono state solo minimamente penalizzate.

Assumere vera l'ipotesi alternativa che dice che unicamente a Febbraio ci sia stata una fluttuazione consiste in escludere appunto Febbraio dalla serie di dati[2] In questo caso, procedendo come sopra si sarebbe ottenuto

Mese	ni	pi	Npi	sigma^2	chi^2
gennaio	25	0,25	29	21,8	0,74
marzo	32	0,25	29	21,8	0,41
aprile	36	0,25	29	21,8	2,25
maggio	23	0,25	29	21,8	1,66
0	116	1	116		5,06

Figura 2: Tabella di riferimento Ipotesi 2

In questo caso, i valori del test sono piu' regolari La distribuzione di riferimento del χ^2 ha $k = N - \nu = 4 - 1 = 3$ gradi di libertà, in quanto per stimare n^* si e' usato il fatto che $N = 171 - (46 + A)$. Il valore critico di riferimento corrispondente ad un livello di confidenza del 99% e': $\chi^2_0 = 11.34$. Siccome $\chi^2_m = 5.06 < 11.34$ allora l'ipotesi nulla risulta accettata. Ad una conclusione simile si sarebbe arrivati le altre vie menzionate. Possiamo quindi concludere che l'ipotesi piu' credibile é la seconda.

Q1.2 Per la seconda domanda, era sufficiente calcolare:

$$\delta = n_{Feb} - \frac{n_{Gen} + n_{Mar} + n_{Apr} + n_{Mag}}{4} = 26$$

non era chiesto di stimarne la incertezza, ma si sarebbe potuto procedere calcolando:

$$s_\delta = \sqrt{s^2_{n_2} + s^2_{\bar{n}}} = \sqrt{N p_{2Feb}(1 - p_{Feb}) + \frac{1}{4} s^2_{n*}} \simeq 6$$

. Qualcuna ha confrontato i valori del χ^2_m, ma non era questa la domanda.

[2]Assumere un valore specifico di rosa vendute o di probabilità in relazione agli altri mesi (ad es, il doppio) per Febbraio sarebbe stata una scelta in parte arbitraria (come teniamo in conto delle fluttuazioni?), anche se ragionevole. Nel primo caso tuttavia, il numero di misure indipendenti sarebbe passato a 4 visto che il conteggio era fissato. Nel secondo caso, rimaneva pari a 5 il numero di misure indipendenti.

Es. 1.3 . Uno studente misura l'andamento velocità-tempo in un piano inclinato per quattro intervalli di lunghezza, trovando i dati in tabella (A parametro del primo esercizio): Osservando i punti, allo studente essi sembrano piuttosto allineati, tuttavia si propone di effettuare un test di Student sull'indice di correlazione campionario per valutare quanto le linearità possano essere causata unicamente dalle fluttuazioni statistiche delle misure. Come fareste voi? Q2.1 Calcolate il coefficiente di correlazione lineare campionario r della serie, e la sua incertezza. Q2.2 Specificate a parole quale sia l'ipotesi nulla per questo test Q2.3 Scrivete la variabile di Student t e dite quale sia la PDF di Student di riferimento, ovvero specificate il numero di gradi di libertà (non serve la formula esplicita della PDF) Q2.4 Effettuate il test con un livello di confidenza dello 0.5% ad una coda: definite se l'ipotesi nulla e' accettata o rifiutata, e cosa implica questo a livello della domanda posta dallo studente.

Tempo (s)	Velocità (m/s)
1.1	3.5
2	2.6
2.9	$2 + 0.05$
4	0.8

R: 1.3 Q2.1 Si calcola il coefficiente di correlazione lineare campionario r e la sua incertezza s_r, secondo le formule

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}} \qquad s_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{N - 2}}$$

Usare la formula per la popolazione porta ad una stima con bias, che va corretta come per la varianza campionaria. Chi lo ha calcolato in questo modo ha ricevuto una piccola penalità. Si ricordi infine che essendo piuttosto probabile per bassa numerosità avere valori alti di r (al limite per 2 punti passa sempre una retta) era importante ritenere un elevato numero di cifre significative per poi svolgere il test.

Q2.2 La domanda che si pone lo studente e' se vi sia correlazione solo a seguito di fluttuazioni casuali, e quindi l'ipotesi nulla e' che non via siano correlazioni lineare, ovvero $r = 0$. La ragione di questa scelta rispetto a quella $r = 1$ é che il significato di margine di confidenza é chiaro per l'esclusione di una ipotesi ma non per la sua conferma.

Q2.3 La variabile di Student viene quindi costruita come

$$t = \frac{r - 0}{s_r} = r\sqrt{\frac{N - 2}{1 - r^2}}$$

La PDF di Student di riferimento ha $k = N - \nu = 4 - 2$ gradi di libertà, in quanto si usano i dati per calcolare $\overline{x}, \overline{y}$.

Q2.4 Il test puo' essere effettuato ad una o a due code, con significato leggermente diverso. Se fatto ad una coda, allora il valore critico per 2 gradi di libertà, ad una coda e $\alpha = 0.005$ é $t_0 = 9.925$. Per $A = 1, 2$ risulta $|t| < t_0$ e il test é passato, altrimenti no. Se il test passa, allora l'interpretazione é che le fluttuazioni statistiche sono tali che é probabile che il valore di r sia compatibile con zero, ovvero con non correlazione. Nell'altro caso, ovviamente non stiamo dimostrando che vi sia una relazione linerare tra i dati, ma stiamo affermando la probabile presenza di un grado di correlazione oltre le fluttuazioni.

Es. 1.4 . Una popolazione ha la probabilità del 10% di avere un certo gene denominato RC che influisce sulla concentrazione sotto stress. Osservando un campione di 50+A individui (A e' il parametro del primo esercizio): Q3.1 Qual è la probabilità che 5 individui abbiano il gene RC? Q3.2 Qual è la probabilità che, nel campione, 1 o più individui abbiano il gene?

Si supponga di eseguire un test per la presenza del gene RC. Il test e' certificato con un tasso di errori di prima specie del 15% e di errori di seconda specie del 7%. Q3.3 Si identifichino le possibili situazioni alternative che si possono verificare a seguito del test e si chiarisca quali rappresentano gli errori di prima e seconda specie. Q3.4 Si identifichino le probabilità di queste quattro situazioni (e per verifica si controlli che la somma delle probabilità collegate alle 4 eventualità sia 1). Q3.5 Si calcoli la probabilità che, sul campione precedentemente descritto, 2 persone siano correttamente dichiarate munite del gene RC?

R: 1.4 Q.3.1 Avere o no il gene è un evento di Bernoulli. Si applica la relativa statistica per $k = 5$ eventi positivi con probabilità $p = 0.1$ di accadere su $N = 50 + A$ tentativi (per A=6):

$$\mathcal{B}(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k} = 0.177$$

Q3.2 Si tratta dell'evento complementare a quello in cui nessun individuo del campione abbiano il gene:

$$1 - \mathcal{B}(0; N, p) = 0.997$$

Q3.3-4 Gli errori di prima specie corrispondono alla probabilità di rifiutare erroneamente l'ipotesi nulla. Gli errori di seconda specie corrispondono al rifiuto erroneo dell'ipotesi alternativa. Per rispondere alla domanda va quindi definita quale sia l'ipotesi nulla del test. Qui il testo non era chiaro e si basava sul fatto che, per convenzione sui test medici, l'ipotesi nulla è l'assenza della patologia o della caratteristica ricercata: $\mathcal{H}_0$ «L'individuo è sprovvisto di gene RC». L'ipotesi alternativa è univocamente definita: $\mathcal{H}_a$ "L'individuo ha il gene RC". Sulla base di questo si possono costruire i quattro casi e la tabella delle quattro eventualità:

(a) Vero negativo: il paziente non ha il gene e il test risulta negativo al gene. La probabilità e' quindi: $P(a) = 0.9 \cdot (1 - 0.15) = 0.765$
(b) Vero positivo: il paziente ha il gene e il test risulta positivo al gene. La probabilità e' quindi: $P(b) = 0.1 \cdot (1 - 0.07) = 0.093$
(c) Errore di tipo I: Falso negativo: il paziente non ha il gene ma il test risulta positivo al gene. La probabilità e' quindi: $P(c) = 0.9 \cdot (0.15) = 0.135$
(d) Errore di tipo II: Falso positivo: il paziente ha il gene ma il test risulta negativo al gene. La probabilità e' quindi: $P(d) = 0.1 \cdot (0.07) = 0.007$

In tabella risulta

	H_0 vera Assenza RC	H_0 falsa Presenza RC	
Test negativo	p_a : 0.765	p_d : 0.007 Err. Tipo II	
Test positivo	p_c : 0.135 Err. Tipo I	p_b : 0.093	
	0.9	0.1	1

Nel caso si fosse presa l'ipotesi alternativa come ipotesi nulla, ovvero come ipotesi nulla il paziente portatore di gene, la tabella sarebbe risultata analoga scambiando ovviamente i due valori per gli errori.

Q3.5 Nel campione di $N = 50 + A$ persone, avendo a disposizione la probabilità di essere dichiarato correttamente positivo $p(b) = 0.093$ allora, attraverso la statistical binomiale si puo' calcolare ($A = 6$):

$$\mathcal{B}(2; N, p(b)) = 0.068$$