1 Correzione Appello SF1 Statistica 16/06/2022 - canale MZ

Es. 1.1 . Si stimi il parametro A pari alla dispersione massima delle cifre che compongono il proprio numero di matricola. Esplicitare la risposta sull'elaborato scritto.

R: 1.1 La dispersione massima e' la differenza tra il valore massimo e il valore minimo. Esempio:

$$123456 \rightarrow A = 5 \qquad 123450 \rightarrow A = 5$$

Es. 1.2 .

Primo set	Secondo set
100	101
106	92+A
89	102
101	108
100	96
94	

Un ricercatore vuole controllare la larghezza X di un contatto metallico prodotto per litografia. Effettua un primo set di misure tramite un microscopio calibrato e un secondo set con un'altro nominalmente simile. Si risponda alle seguenti domande: 1.1) Si riporti la miglior stima dei valori di larghezza e delle loro incertezze nelle due serie; 1.2) Ci si chiede se i due strumenti siano calibrati in modo compatibile con un livello di confidenza del 95.5%, assumendo che le misure siano affette da incertezze gaussiane. Si descrivano in modo esaustivo i calcoli e le formule utili per rispondere al quesito; 1.3) Può il ricercatore combinare le due serie? Che valore finale di misura e incertezza riporta? Per i margini di confidenza si usino le tabelle di valori critici.

R: 1.2 Si tratta di valutare la compatibilità tra le due serie di misure. Essendo il numero di misure piccolo (inferiore a circa 20-30) la valutazione é meglio svolta attraverso un test di Student.

1.1) Cominciamo valutando le miglior stime delle due serie. Essendo stato specificato che si tratta di un microscopio calibrato, possiamo assumere che le incertezza sistematiche siano state annullate, e che rimangano solo le fluttuazioni statistiche (casuali) della misura, dovuta alla sensibilità dello strumento[1] Allora la miglior stima e' quella ottenuta assumendo una distribuzione normale, e quindi:

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

dove $N = 6$ per il primo campione e $N = 5$ per il secondo campione. Ad esempio per $A = 7$ $\overline{x_1} = 98.3 \pm 2.4)$ μm e $\overline{x_2} = 101.2 \pm 2.0)$ μm (si poteva anche portare una unica cifra significativa visto il numero esiguo di valori).

1.2) Come detto i due valori vanno confrontati attraverso un test di Student. Cercandosi pura compatibilità, si può procedere ad un test a due code, ovvero che non guarda il segno della differenza tra i due valori. La variabile t di Student va impostata secondo la formula **??** (si ponga M numerosità del II campione e $Y = X_2$):

$$|t_m| = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{1}{M} + \frac{1}{N}}\sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2 + \sum_{j=1}^{M}(y_j - \overline{y})^2}{N + M - 2}}}$$

[1] E' sufficiente che le incertezze statistiche non corrette siano inferiori a quelle casuali.

distribuita con una PDF di Student a $N+M-2=9$ gradi di libertà. Nel caso $A=7$ allora $t_m=0.89$ per ottenerlo bisognava costruire una colonna di valori $(x_i-\overline{x}),(y_i-\overline{y})$ e calcolarne le somme. Eseguendo un test a due code il valore critico per un livello di confidenza del 95.5% é pari a $t_0=2.32$. Essendo ora $t_m<t_0$ il test di Student risulta superato. La conclusione e' che con un margine di confidenza di almeno il 95.5% le due misure sono compatibili entro le fluttuazioni sperimentali.

1.3) Essendo le misure compatibili entro le fluttuazioni, si potrebbe riportare una misura combinata della due attraverso il calcolo della media pesata.

$$x_{MP}=\frac{\sum_{i=1}^{2}\overline{x_i}/s^2_{\overline{x_i}}}{\sum_{i=1}^{2}1/s^2_{\overline{x_i}}} \qquad s_{MP}=\sqrt{\frac{1}{\sum_{i=1}^{2}1/s^2_{\overline{x_i}}}}$$

Per $A=7$ si sarebbe ottenuto $x=(99.6\pm1.8)\ \mu$m. Si noti che nonostante le due incertezza di partenza fossero molto vicine, la procedura di media aritmetica, ancorché avesse dato un valore numerico simile, non sarebbe stata la più corretta formalmente.

Es. 1.3 . In un esperimento, un elettrone prossimo alla velocità della luce viene fatto interagire con un materiale. A seguito dell'attraversamento del materiale l'elettrone può aver subito o meno una deflessione dalla sua traiettoria (D deflessione D' non deflessione) e inoltre può aver causato o meno l'emissione di un fotone (E emissione, E' non emissione). Ogni elettrone ha il 90% di probabilità di essere deflesso e l' 85% di emettere un fotone. Si sa inoltre che gli elettroni che emettono il fotone hanno il 99% di essere deflessi.

Vengono mandati sul materiale $100+A$ elettroni e viene determinata la deflessione e l'eventuale emissione del fotone per ciascuno elettrone. 2.1) Se un elettrone risulta deflesso quale è la probabilità che emetta un fotone? 2.2) Quanti elettroni in media sono deflessi e non emettono un fotone? 2.3) Quale è la probabilità di misurare meno di 3 elettroni (0 o 1 o 2) deflessi senza l'emissione del fotone (del tipo precedente) ?

R: 1.3 Identifichiamo gli eventi aleatori come $E=$[elettrone Emette un fotone] e $D=$[elettrone viene Deflesso]. Allora i dati del problema sono $P(E)=0.85, P(D)=0.90, P(D|E)=0.99$.

2.1) La probabilità che un elettrone che viene deflesso emtta un fotone e' la probabilità condizionata $P(E|D)$ che si puo' in questo caso ottenere facilmente attraverso la formula di Bayes:

$$P(E|D)=\frac{P(E)P(D|E)}{P(D)}=0.935$$

2.2) La probabilità che un elettrone sia emesso e deflesso e' la probabilità congiunta

$$P(E'D)=P(E'|D)P(D)=(1-P(E|D))P(D))=0.0585$$

. Per $100+A$ elettroni allora il numero medio di eventi di questo tipo e' $N=0.059(100+A)$.

	E	E'	
D	0.8415	0.0585	0.9
D'	0.0085	0.0915	0.1
	0.85	0.15	1

2.3) La probabilità di osservare meno di 3 elettroni di questo tipo é quindi di tipo Bernoulliano:

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

dove $k=0,1,2$, $N=100+A$, $p=0.059$. La probabilità richiesta era $\mathcal{B}(0)+\mathcal{B}(1)+\mathcal{B}(2)$. Nel caso $A=7$ si trovava $P=0.002+0.011+0.035=0.047$. Essendo la probabilitá p piuttosto piccola e N piuttosto grande si poteva anche usare una approssimazione di tipo Poissoniano. Questa approssimazione avrebbe portato ad una probabilità di $P=0.051$, sovrastimata tuttavia di circa il 10%.

Es. 1.4 .

Il fattore di rischio ambientale F dovuto alla concentrazione di tre sostanze è definito da

$$F = 127\frac{C_1\, C_3}{C_2}$$

dove C_1, C_2, C_3 sono le concentrazioni (in ppm) delle tre sostanze, Si svolgono 5 prese dati in cui si misurano C_1, C_2, C_3 come riportato in tabella Stimare motivando le risposte: 3.1 La miglior stima di F; 3.2 L' incertezza di F in maniera diretta calcolando i diversi F_i 3.3 L'incertezza di F utilizzando il metodo della propagazione delle incertezze e trascurando il termini di covarianza tra le serie 3.4 Discutere le possibili ragioni delle differenze tra le stime 3.2 e 3.3 (3.5 facoltativo) ripetere la stima della incertezza 3.3 calcolando le covarianze tra le serie

C_1	C_2	C_3	F
5.007	19.4	1.56	
4.994	19.39	1.38	
5.005	19.63	1.68	
4.99	19.85	1.28	
4.999	19.78	1.33	

R: 1.4 In questo caso abbiamo 5 serie ripetute di misure attorno al valore centrale di ciascuna. E' un caso molto specifico riscontrato per altro nell'esercizio H2 del GUM. E' il caso in cui ogni serie di misure é 'completa' e quindi si possono calcolare i 5 valori $F_i = 127\frac{C_1^i\, C_3^i}{C_2^i} = (51.13, 45.14, 54.40, 40.87, 42.69)$

3.1 La miglior stima del valore centrale é quindi la media aritmetica dei 5 valori F_i: $\overline{F} = 46.845$. Si noti che la procedura di stima attraverso le medie $\overline{F} = 127\overline{C_1}\,\overline{C_3}/\overline{C_2}$ avrebbe portato ad una leggera sottostima (errore sistematico): $\overline{F} = 46.814$.

3.2 La miglior stima della incertezza é proprio la deviazione standard della media dei valori, che vale 2.563. Questa é anche la piu' esatta in questo caso e la scelta da usare.

3.3 Se si fosse proceduto per propagazione delle incertezze, si sarebbe usato la formula *approssimata* della propagazione degli errori (si veda as es **??**)

$$s^2_{z_0} \simeq \left(\frac{\partial F}{\partial C_1}\right)^2_{(***)} s^2_{C_1} + \left(\frac{\partial F}{\partial C_2}\right)^2_{(***)} s^2_{C_2} + \left(\frac{\partial F}{\partial C_3}\right)^2_{(***)} s^2_{C_3}$$
$$+2\left(\frac{\partial F}{\partial C_1}\frac{\partial F}{\partial C_2}\right)_{***} r_{C_1C_2}\, s_{C_1}\, s_{C_2} + 2\left(\frac{\partial F}{\partial C_1}\frac{\partial F}{\partial C_3}\right)_{***} r_{C_1C_3}\, s_{C_1}\, s_{C_3} + 2\left(\frac{\partial F}{\partial C_2}\frac{\partial F}{\partial C_2}\right)_{***} r_{C_2C_3}\, s_{C_2}\, s_{C_3}$$
$$+\mathcal{O}(2)$$

dove le derivate parziali prime

$$\frac{\partial F}{\partial C_1} = 127\frac{C_3}{C_2} = 9.4; \qquad \frac{\partial F}{\partial C_2} = -127\frac{C_1C_3}{C_2^2} - 2.4; \qquad \frac{\partial F}{\partial C_3} = 127\frac{C_1}{C_2} = 32.4;$$

sono valutate nei punti medi delle tre serie e quindi $(***) = (\overline{C_1}, \overline{C_2}, \overline{C_3})$. Le deviazioni standard sono di nuovo quelle delle medie, e quindi $s_{C_i} = s_{\overline{C_i}}$. Questa procedura avrebbero portato a

$$s_z = 2.445 \text{ no correlazioni, e} \qquad s_z = 2.571 \text{ con correlazioni}$$

Nel primo caso avremmo quindi sottostimato l'incertezza, mentre le secondo la avremmo, ma solo leggermente, sovrastimato, a causa appunto della natura approssimata della formula della propagazione. Per il calcolo completo si sarebbe dovuto procedere a calcolare tutte le covarianze campionarie.

Es. 1.5 .

Si consideri la seguente densità di probabilità non normalizzata a variabile continua definita per x appartenente ad $\mathcal{R}$ e dipendente da un parametro b generico :

$$f(x) = Ke^{-|x|/b}$$

. 4.1. Si stimi il fattore K di normalizzazione della densità per ogni b 4.2. Si calcoli la speranza matematica della densità di probabilità in funzione di b. 4.3. Si definisca la varianza della densità di probabilità in funzione di b (facoltativo: se ne calcoli il valore esplicito in funzione di b) 4.4. Si calcoli la funzione di distribuzione $F(x)$ associata alla densità di probabilità. *Suggerimento per i punti 4.1-4.4: attenzione che avendo /x/, conviene considerare gli integrali come somme degli integrali tra* $[-\infty; 0]$ *e tra* $[0; +\infty]$

Si continui considerando i seguenti eventi distribuiti in 6 classi di frequenza come in tabella e si ipotizzi che seguano la distribuzione di cui sopra con il parametro $b = 10$. Effettuare il test del χ^2 della suddetta ipotesi. In particolare: 4.5. Calcolare i conteggi attesi per ogni classe di frequenza sfruttando la funzione di distribuzione di probabilità calcolata al punto 4.4 4.6. Calcolare la varianza associata a questi conteggi 4.7 Calcolare il χ^2, i gradi di libertà e il valore di riferimento per l'accettazione dell'ipotesi al 95% di livello di confidenza. 4.8. Possiamo affermare che i dati seguono la distribuzione con il suddetto livello di confidenza?

Intervallo	Conteggi
$[-30, -20]$	$8+A$
$[-20, 10]$	18
$[-10, 0]$	66
$[0, 10]$	59
$[10, 20]$	16
$[20, 30]$	8

R: 1.5 4.1 Affinche' la funzione sia PDF deve essere sempre positiva (cio' avviene sempre) e normalizzata a 1 e quindi:

$$1 = \int (Ke^{-|x|/b})dx = 2\int_0^\infty (Ke^{-x/b})dx = 2K\left[be^{-x/b}\right]_0^\infty = 2Kb \to K = 1/2b$$

4.2 Essendo la funzione simmetrica rispetto all'origine, il valore di aspettazione é $\mu = \mathrm{E}[X] = 0$

4.3 La varianza si calcola secondo

$$\sigma_X^2 = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2 = \mathrm{E}[X^2] = 2b^2 (\text{per parti})$$

4.4 La CDF si calcola per segmenti, tra $[-\infty, 0]$ a partire da $F(x) = \int_\infty^0 f(x)dx = 0.5e^{x/b}$. Essendo $F(0) = 0.5$ allora per $[0, +\infty]$ si ha analogamente $F(x) = 0.5 - 0.5e^{-x/b}$

4.5 Avendo a disposizione una PDF di riferimento, il numero di conteggi attesi per ciascun bin diventa: $n_i^* N\, P_{bin,i}$ dove $P_{bin,i} = F_{max,i} - F_{min,i}$ ovvero

$$N_1 = N(F(-20) - F(-30) = 0.043\, N; N_2 = N(F(-10) - F(-20) = 0.116\, N; \ldots$$

Per $A = 7$ i conteggi di riferimento risultano $n^* = (7.78, 21.16, 57.52, 57.52, 21.16, 7.78)$.

4.6 La varianza si calcola secondo la binomiale: $s_i^2 = N\, P_i', (1-P_i) = (7.45, 18.70, 39.34, 39.34, 18.70, 7.45$

4.7 I singoli termini del χ^2 sono quindi $\chi_i^2 = (n_i - n_i^*)^2/s_i^2 = (6.99, 0.53, 1.83, 0.06, 1.42, 0.01)$ e quindi $\chi_m^2 = 10.83$

4.8 Confrontando con una distribuzione del χ^2 con $k = N - 1 = 5$ gradi di libertà, il valore critico per $\alpha = 0.05$ é $\chi_0^2 = 11.07$. Essendo $\chi_m^2 > \chi_0^2$ il test é passato: entro le fluttuazioni, i dati sono compatibili con la distribuzione di riferimento almeno al 95% di margine di confidenza.