Politecnico di Milano – Facoltà di Ingegneria Informatica

Anno Accademico 2008/2009

Corso di Statistica (2L) per INF e TEL

Docente: Antonio Pievatolo; esercitazioni: Raffaele Argiento

Esercitazione del 12/06/2009

Esercizio 1

Una macchina imbottigliatrice è impiegata per riempire flaconi di bagnoschiuma. A causa di fluttuazioni casuali, la quantità di bagnoschiuma per flacone è una variabile aleatoria X gaussiana di media e varianza entrambe incognite. Se la varianza del volume riempito supera $25ml^2$, una frazione non accettabile dei flaconi sar à sotto-riempita o sovra-riempita. Se la varianza non supera $25ml^2$, la macchina imbottigliatrice è considerata precisa. Per controllare la precisione della macchina imbottigliatrice, abbiamo misurato la quantità di bagnoschiuma presente in 46 flaconi (espressa in ml), e abbiamo ottenuto che la funzione di ripartizione empirica $\hat{F}_{46}$ è

x_i	242.1	246.5	248.5	251.0	253.5	255.6
$\hat{F}_{46}(x_i)$	8/46	21/46	27/46	33/46	43/46	1

1. Calcolate la media e la varianza campionarie.
2. Costruite un intervallo di confidenza al 95% per la varianza, unilatero del tipo $(c, 1)$.
3. Costruite un test sulla varianza tale che sia al più pari a 5% la probabilità di commettere lÂ¿errore di prima specie di ritenere imprecisa una macchina effettivamente precisa.
4. Calcolate la probabilità di errore di secondo tipo del test costruito al punto 3. se la varianza vale effettivamente $42ml^2$, o indicate un intervallo dove tale probabilità cade.

Soluzione

1. Osserviamo che la media campionaria $\bar{X}$ coincide con la media di una v.a. Y la cui funzione di ripartizione è $\hat{F}_{46}(x), x \geq 0$. Ponenso $x_0 = -\infty$ si ha

x_i	242.1	246.5	248.5	251.0	253.5	255.6
$\hat{F}_{46}(x_i) - \hat{F}_{46}(x_{i-1})$	8/46	13/46	6/46	6/46	10/46	3/46

   Quindi:

$$\bar{X} = \mathbb{E}(Y) = \sum_{i=1}^{6} x_i \left(\hat{F}_{46}(x_i) - \hat{F}_{46}(x_{i-1}) \right) = 248.6978,$$

   Allo stesso modo, il momento campionario secondo M_2 coincide col momento secondo di Y, quindi:

$$M_2 = \mathbb{E}(Y^2) = \sum_{i=1}^{6} x_i^2 \left(\hat{F}_{46}(x_i) - \hat{F}_{46}(x_{i-1}) \right) = 61868.36,$$

   Quindi la varianza campionaria,

$$S^2 = \frac{45}{46} \left(M_2 - \bar{X}^2 \right) \simeq 18.159$$

2. Sia σ^2 la varianza del volume di bagnoschiuma in un flacone. Un IC(σ^2) di confidenza al 95% della forma $(c, 1)$ è dato da

$$\sigma^2 > (n-1)\frac{S^2}{\chi^2_{n-1}}.$$

Con i dati a nostra disposizione abbiamo l'intervallo $(13.2535, \infty)$

3. Deduciamo dalla domanda di dover impostare un test di verifica delle ipotesi nulla H_0: "La macchina è precisa" contro l'alternativa H_1: "La macchina è imprecisa", che in termini di σ^2 traduciamo come

$$H_0 : \ \sigma^2 \leq 25 \text{ contro l'alternativa } H_1 : \ \sigma^2 > 25.$$

In base al punto precedente, con probabilità 0.95, $\sigma^2 \geq \frac{45S^2}{61.656} = 13.2535$; poichè $25 \geq 13.2535$, per la dualit à fra IC e test di ipotesi, riteniamo plausibile l'ipotesi nulla $H_0 : \ \sigma^2 \leq 25$ con significatività $\alpha = 1 - 0.95 = 0.05$.

4. Dobbiamo calcolare

$$\begin{aligned}
\Pr_{\sigma^2=42}(25 \in IC(\sigma)) &= \Pr_{\sigma^2=42}\left(25 > (n-1)\frac{S^2}{\chi^2_{n-1}(1-\alpha)}\right) = \Pr_{\sigma^2=42}\left((n-1)S^2 < 25\chi^2_{n-1}(1-\alpha)\right) \\
&= \Pr\left((n-1)\frac{S^2}{\sigma^2} < 25\frac{\chi^2_{n-1}(1-\alpha)}{\sigma^2}\right) = \Pr\left(U < \frac{25}{42}61.656\right) \\
&= \Pr\left(U < 36.7\right) \in (0.125, 0.2)
\end{aligned}$$

Dove U è una v.a. con distribuzione χ^2_{n-1}. Si osservi che il valore esatto della probabilità di errore di seconda specie è 0.194

∎

Esercizio 2

Abbiamo raccolto dei dati su 200 allievi del corso di laurea in xxx che hanno superato l'esame di LIN e SIN, entrambi obbligatori, e li abbiamo sintetizzati come segue.

LIN SIN	1	2	3 o pi'u	
1	60	10	10	
2	30	15	15	
3 o pi'u	10	20	30	

Tabella 1: Numero di appelli sostenuti per superare LIN e SIN

In Tabella 1 troviamo il numero di appelli sostenuti dagli allievi per superare gli esami LIN e SIN; per esempio, 20 è il numero di allievi che in SIN sono stati promossi al secondo appello sostenuto e che in LIN hanno dovuto faticare almeno 3 appelli. Poi, in (1) abbiamo le statistiche sui voti finali registrati, espressi in trentesimi:

$$\sum_{j=1}^{200} x_j = 4600, \sum_{j=1}^{200} x_j^2 = 107600, \sum_{j=1}^{200} y_j = 4900, \sum_{j=1}^{200} y_j^2 = 120850, \sum_{j=1}^{200} x_j y_j = 113704, \tag{1}$$

dove $\{x_j\}$ sono i voti di LIN e $\{yj\}$ quelli di SIN.

1. Usate un opportuno test di livello $\alpha = 5\%$ per stabilire se ci sia dipendenza fra il numero di appelli necessari per superare gli esami LIN e SIN.

2. Stabilite se il voto medio registrato di LIN sia più basso di quello medio registrato per SIN. A tal fine, costruite un opportuno test tale che sia al più pari ad $\alpha = 5\%$ la probabilità di commettere l'errore di prima specie di ritenere il voto in LIN minore di quello in SIN, quando effettivamente è maggiore o uguale. Ipotizzate la normalità dei voti.

3. Stabilite se i voti riportati in LIN e SIN siano o no indipendenti, a livello $\alpha = 5\%$. Ipotizzate la normalità dei voti.

Soluzione

1. Impostiamo un test chi-quadrato di indipendenza fra le variabili L ed S che contano rispettivamente il numero di appelli necessari per superare LIN e quello per superare SIN. Le ipotesi nulla e alternativa sono rispettivamente

$$\mathrm{H}_0\text{: “L e S sono indipendenti”, contro } \mathrm{H}_1\text{: “L e S non sono indipendenti.”}$$

Completiamo la Tabella 1 con le numerosità marginali:

LIN SIN	1	2	3 o pi'u	
1	60	10	10	80
2	30	15	15	60
3 o pi'u	10	20	30	60
	100	45	55	

La statistica di Pearson ha valore

$$Q = 200 \left(\sum_{i=1}^{3} \sum_{j=1}^{3} \frac{N_{ij}}{N_i N_j} - 1 \right) \simeq= 47.9 \ .$$

Asintoticamente (osserviamo che ciascuna cella della tabella 1 contiene almeno 5 elementi) Q ha f.d.r. $\chi^2_{(3-1)(3-1)} = \chi^2_4$. Poiché risulta $47.9 > \chi^2_4(1 - 0.05) = 9.488$, rifiutiamo l'ipotesi H_0 di indipendenza fra le variabili L e S. Inoltre, il p-value del test risulta $1 - F_{\chi^2_4}(47.9) \leq 1 - F_{\chi^2_4}(18.467) = 0.1\%$: concludiamo che cè una netta evidenza sperimentale contro l'ipotesi H_0.

2. Siano μ_X il voto medio in LIN e μ_Y quello in SIN. Dobbiamo impostare un t-test per dati gaussiani accoppiati di significatività $\alpha = 5\%$ per verificare

$$\mathrm{H}_0 : \ \mu_X - \mu_Y \geq 0 \text{ contro } \mathrm{H}_1 : \ \mu_X - \mu_Y < 0;$$

quindi rifiutiamo H_0 se

$$t := \frac{\bar{x} - \bar{y}}{\sqrt{s^2_{X-Y}/200}} \leq -t_{199}(95\%).$$

Con i dati forniti abbiamo

$$\bar{x} = 23, \bar{y} = 24.5, s^2_X = 9.05, s^2_Y = 4.02, \mathrm{cov}_{X,Y} = \frac{1}{199} \left(\sum_{j=1}^{200} x_j y_j - 200 \bar{x}\bar{y} \right) \simeq 5.05$$

$$s^2_{X-Y} = s^2_X - s^2_Y - 2\mathrm{cov}_{X,Y} = 9.05 + 4.02 - 2 \times 5.05 \simeq 2.97 \ .$$

Poiché $t \simeq -12.31 < -1.645$, allora rifiutiamo lÂ¿ipotesi nulla che il voto in LIN sia mediamente maggiore o uguale di quello in SIN.

3. Verifichiamo le ipotesi sul coefficiente di correlazione lineare date da

$$\mathrm{H}_0 : \ \rho = 0 \text{ contro } \mathrm{H}_1 : \ \rho \neq 0,$$

con il t-test che prescrive di rifiutare H_0 a livello 5% se

$$T = \left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| \geq t_{198}(97.5\%).$$

Il coefficiente di correlazione campionario

$$R = \frac{\mathrm{COV}_{X,Y}}{\sqrt{S_X^2 \times S^2 Y}}$$

ha realizzazione $r = 0.837$ e la statistica T ha realizzazione 21.523. Poiché $t_{198}(97.5\%) \simeq z_{97.5} = 1.96$, allora rifiutiamo H_0 a livello $\alpha = 5\%$. In realtà, la statistica test ha un valore così elevato che ne risulta un p-value prossimo allo zero, rifiutiamo quindi l'ipotesi di indipendenza a qualunque livello del test.

Esercizio3

L'azienda OISAC ha proposto una pila di ultima generazione yyy più durevole del tipo xxx di vecchia generazione. Per confrontare la durata delle due pile xxx e yyy abbiamo a disposizione i seguenti due campioni indipendenti di dati continui (espressi in ore):

$$x_i: \ 7.26, \ \ 2.04, \ \ 0.94, \ \ 1.76, \ \ 11.08, \ \ 0.60, \ \ 9.04$$

(=durate di m=7 pile xxx)

$$y_i: \ 0.80, \ \ 1.71, \ \ 4.10, \ \ 6.10, \ \ 7.89, \ \ 24.10$$

(=durate di n=6 pile yyy).

1. Proponete a OISAC un test per verificare se le pile yyy durano più di quelle xxx, che funzioni anche quando non si ha nessun'altra informazione sul modello statistico generatore dei dati. Sulla base dei dati forniti che decisione prendete a livello $\alpha = 10\%$

In realtà , successivamente, la nostra conoscenza sul modello statistico generatore dei dati è aumentata. Infatti, ora riteniamo plausibile modellare le durate delle pile, sia di nuova che di vecchia generazione, come variabili aleatorie gaussiane.

2. Avendo questa ulteriore informazione, la risposta alla domanda 1. cambia o no? Argomentate la risposta impostando una opportuna metodologia statistica.

[Per risparmiare tempo, nell'eventualità vi occorrano, vi abbiamo già calcolato qualche statistica per i due campioni: $\sum_{i=1}^{7} x_i = 32.72$, $\sum_{i=1}^{6} y_i = 44.7, \sum_{i=1}^{7} x_i^2 = 265.7, \sum_{i=1}^{6} y_i^2 = 700.65$].
SOLUZIONE

1. Sia X la v.a. che rappresenta il tempo di durata di una pila di tipo xxx e Y il tempo di durata di una pila yyy. Possiamo tradurre l'ipotesi che le pile yyy durano più di quelle xxx dicendo che Y domina stocasticamente X. Si vuole quindi studiare il problema

$$H_0: \ F_X(x) = G_Y(x) \ \forall x \in \mathbb{R} \text{ contro } H_1: F_X(x) \geq G_Y(x) \ \forall x \ \text{e} \exists x \text{ t.c. } F_X(x) > G_Y(x).$$

Dove $F_X(x)$ e $G_Y(x)$ sono le distribuzioni di X e Y rispettivamente. Distribuzioni che supporremo essere continue. Un problema di questo tipo è detto di Wilcoxon-Mann-Whitney. La statistica test è la somma dei ranghi $T_X \overset{H_0}{\sim} W_{m,n}$. Si ha che

$$\mathbb{E}(W_{m,n}) = \frac{m(m+n+1)}{2}$$
$$\mathrm{Var}(W_{m,n}) = \frac{mn(m+n+1)}{12}$$

La distribuzione di $W_{m,n}$ è simmetrica rispetto alla media, se $m, n \leq 20$ i percentili di $W_{m,n}$ sono tabulati altrimenti si può approssimare la distribuzione di $W_{m,n}$ con quella di una normale con la stessa media e la stessa varianza. Inoltre dalla proprietà di simmetria segue che se $w_{m,n}(\alpha)$ è il percentile di ordine α di W allora

$$w_{m,n}(\alpha) + w_{m,n}(\alpha) = m(m+n+1) \tag{2}$$

Per calcolare il valore di T_X riscriviamo i dati nella seguente forma tabellare

x_i	7.26	2.04	0.94	1.76	11.08	0.60	9.04
ranghi di X	9	6	3	5	12	1	11
y_i	0.80	1.71	4.10	6.10	7.89	24.10	
ranghi di Y	2	4	7	8	10	13	

Da cui si ottiene facilmente che $t_X = 47$. Dato che nell'ipotesi nulla suppongo che Y domina stocasticamente X rifiuterò tale ipotesi se T_X assume valori piccoli, la regione critica si scriverà dunque guardando alla coda sinistra della distribuzione di T_X sotto l'ipotesi H_0. la regione critica è quindi

$$C = \{dati : t_X \leq w_{m,n}(\alpha)\},$$

con $\alpha = 0.05$, $m = 7$ e $n = 6$ dalle tavole ottengo $w_{7,6}(0.05) = 37$ quindi $C = \{dati : t_X \leq 37\}$. Dato che $t_X = 47 \notin C$ NON Rifiuto H_0 al livello di significatività $\alpha = 0.05$.

In modo alternativo possiamo calcolare

$$\text{p-value} = \mathbb{P}(T_X < t_X) = \mathbb{P}(T_X < 47) > \mathbb{P}(T_X \leq 40) = 0.1$$

otteniamo dunque p-value > 0.1 che indica assenza di evidenza sperimentale per affermare che le pile yyy sono più durevoli delle xxx.

Osserviamo ora che lo stesso test si può effettuare considerando la statistica $T_Y \overset{H_0}{\sim} W_{n,m}$, in questo caso però rifiuto l'ipotesi nulla quando T_Y assume valori elevati. Per scrivere la regione di rifiuto bisogna quindi guardare alla coda destra della statistica test. Si ottiene la regione critica

$$C = \{dati : t_y \geq w_{n,m}(1-\alpha)\}.$$

Usando la formula (2) si ottiene

$$C = \{dati : t_y \geq n(n+m+1) - w_{n,m}(\alpha)\},$$

dalle tavole $C = \{dati : t_Y \geq 54\}$ ed in conclusione dato che $t_Y = 44 \notin C$ non rifiuto H_0 al livello $\alpha = 0.05$.

Per calcolare il p-value per il test basato sulla statistica T_Y si osservi che p-value $= \mathbb{P}(T_Y > t_y)$. Dalla proprietà di simmetria della distribuzione di Wilcoxon si ricava che p-value $= \mathbb{P}(T_Y < t'_Y)$, dove t'_Y è il simmetrico di t_Y rispetto al punto $n(m+n+1)/2$. Si ha dunque che $t'_Y = n(m+n+1) - t_Y = 84 - 44 = 40$, in conclusione p-value $= \mathbb{P}(T_Y < 44) > 0.1$.

2. Supponiamo ora che le variabili X e Y siano indipendenti e distribuite normalmente i.e.

$$X \sim \mathrm{N}(\mu_X, \sigma_X^2) \ \perp \ \mathrm{N}(\mu_Y, \sigma_Y^2)$$

Allora possiamo affrontare il problema dividendo il test in due passi

$$\mathrm{H}_0^{(1)} : \ \sigma_X^2 = \sigma_Y^2 \text{ contro } \mathrm{H}_1^{(1)} : \ \sigma_X^2 \neq \sigma_Y^2$$

Se in base all'osservazione campionaria non rifiutiamo l'ipotesi $\mathrm{H}_0^{(1)}$ allora studiamo il test

$$\mathrm{H}_0^{(2)}. \ \mu_X = \mu_Y \text{ contro } H_1^{(2)} \mu_X < \mu_Y$$

Per quanto riguarda il primo test la statistica di riferimento è

$$U := \frac{S_X^2}{S_Y^2} \overset{H_0}{\sim} F_{m-1,n-1},$$

dove S^2 è la varianza campionaria, e $F_{m,n}$ è la distribuzione di Fisher con m gradi di libertà al numeratore e n gradi di libertà al denominatore. È facile ricavare i valori osservati $s_X^2 = 18.84$ e $s_Y^2 = 73.53$, si ha dunque $u = 0.256$. Fissiamo ora il livello di significatività $\alpha_1 = 0.05$ e scriviamo la regione critica

$$C^{(1)} = \left\{ dati : u \leq F_{m-1,n-1}\left(\frac{\alpha}{2}\right); u \geq F_{m-1,n-1}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Dalle tavole si ricava $F_{6,5}(0.975) = 6.98$ e $F_{6,5}(0.025) = \frac{1}{F_{5,6}(0.975)} = \frac{1}{5.99} \simeq 0.67$. Si ha dunque $u = 0.256 \notin C^{(1)}$ e non si rifiuta l'ipotesi nulla al livello di significatività $\alpha_1 = 0.05$.

Per calcolare il p-value del test osservo che la mediana di una distribuzione di Fisher $F_{6,5}$ è 1.02, dato che $u = 0.256 < 1.02$ ho che $\frac{\text{p-value}}{2} = \mathbb{P}(U < 0.256)$. Ricordo ora che se $U \sim F_{m,n}$, allora $\frac{1}{U} \sim F_{n,m}$ quindi $\frac{\text{p-value}}{2} = \mathbb{P}\left(\frac{1}{U} > \frac{1}{0.256}\right)$ e dalle tavole ricavo che $0.1 < \text{p-value} < 0.2$.

Dato che non abbiamo rifiutato l'ipotesi nulla nel primo test, studiamo il secondo. In questo caso la statistica test è

$$T := \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \overset{H_0}{\sim} t_{m+n-2}$$

dove

$$S_p^2 = \frac{(m-1)\,S_X^2 + (n-1)\,S_Y^2}{m+n-2}$$

è lo stimatore raggruppato (*pooled*) della varianza. Dai dati si ricava $s_p^2 = 43.7$ e $t = -0.756$. Fissiamo il livello di significatività $\alpha_2 = 0.05$ e ricaviamo la regione critica

$$C^{(2)} = \{ dati : \ t < t_{m+n-2}(\alpha_2) \}.$$

dalle tavole si ricava $t_{11}(0.05) = -1.796$. Dato che $t = -0.756 \notin C^{(2)}$ non rifiuto l'ipotesi nulla $\mathrm{H}_0^{(2)}$ al livello di significatività $\alpha_2 = 0.05$

Posso concludere dicendo che non rifiuto l'ipotesi di omogeneità fra la distribuzione di X e quella di Y con un livello di significatività $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2) = 0.0975$

∎