Esame di Statistica del 14 dicembre 2007 (Corso di Laurea Triennale in Biotecnologie, Università degli Studi di Padova).

Cognome Nome Matricola

Es. 1	Es. 2	Es. 3	Es. 4	Somma	Voto finale

Attenzione: si consegnano SOLO i fogli di questo fascicolo.

Esercizio 1. La fertilizzazione dei fiori di una certa pianta è stata controllata in modo da ottenere nuove piante con fiore rispettivamente rosso, rosa e bianco con probabilità rispettivamente 1/4, 1/2, 1/4. n semi scelti a caso producono piante che sono tutte coltivate fino alla fioritura. Calcola la probabilità che:

1. se $n = 4$, si ottengano due piante con fiore rosa e due non rosa;
2. se $n = 4$, si ottengano almeno due piante con fiore rosa;
3. se n è generico, che le piante siano tutte bianche;
4. se n è generico, che ci sia almeno una pianta bianca.
5. Qual è il minimo n tale che la probabilità di ottenere almeno ottenere una pianta bianca sia almeno 0.999?

Esercizio 2. Una compagnia farmaceutica produce un nuovo farmaco contro le emicranie con un principio attivo molto rapido ad entrare in circolo. Per convincere l'ente preposto al controllo dei nuovi medicinali che il tempo medio che il farmaco impiega a raggiungere il sangue è inferiore ai 10 minuti, questa ditta raduna un campione di persone soggette ad emicranie e conduce un esperimento.

1. Quale tipo di test bisogna effettuare e come vanno scelte l'ipotesi e l'alternativa?
2. Supponiamo di prendere un gruppo da 30 persone e di avere le seguenti statistiche cumulative:

   $$\bar{X} = 9.267, \qquad s_X = 2.29$$

   Supponendo che l'ente preposto al controllo voglia una significatività dell'1%, si può affermare che il farmaco raggiunga il sangue in meno di 10 minuti?
3. Calcolare l'intervallo di confidenza al 99% del tempo medio. In base a questo intervallo, si può rispondere alla domanda del punto 2.?

Esercizio 3. Con lo scopo di determinare se le cause legali per negligenza siano più frequenti per certi tipi di interventi chirurgici, si sono studiati campioni casuali di tre tipi di interventi, ottenendo i dati seguenti.

tipo di intervento	casi campionati	cause intentate
Chirurgia cardiaca	400	26
Chirurgia cerebrale	300	19
Appendicectomia	300	7

1. Ci sono differenze tra i diversi tipi di intervento nel portare ad una causa giudiziaria? Effettuare il test con un livello $\alpha = 0.05$.
2. Stabilire se ci sono interventi in cui il tasso di cause sul numero di interventi si può ritenere uguale.

Esercizio 4. Si vuole investigare il legame tra l'età (x, in anni) e diametro massimo del tronco (y, in pollici) di alberi di Eucalyptus Robusta. Si raccolgono quindi i dati di 15 tronchi, con le seguenti statistiche cumulative:

$$\sum_{i=1}^{15} x_i = 511, \qquad \sum_{i=1}^{15} y_i = 268.7, \qquad \sum_{i=1}^{15} x_i^2 = 17555, \qquad \sum_{i=1}^{15} y_i^2 = 4835.29, \qquad \sum_{i=1}^{15} x_i y_i = 9169.3$$

1. Calcolare la retta di regressione che lega il diametro massimo del tronco all'età dell'albero.
2. Eseguire un test per verificare se ci può essere una relazione lineare; riportare limitazioni al valore P.
3. In un bosco di Eucalyptus Robusta si è deciso di tagliare gli alberi di 50 anni. Quale ci aspettiamo che sia in media il diametro massimo del tronco? E quale può essere un intervallo di confidenza al 95% per questa misura?

Soluzioni

Esercizio 1.

1. Definiamo le variabili aleatorie di Bernoulli $(X_i)_i$ e poniamo $X_i = 1$ se l'i-esima pianta ha il fiore rosa; allora le $(X_i)_i$ hanno legge $Be(1/2)$, e $S_n := \sum_{i=1}^n X_i$, che esprime il numero di piante con fiori rosa in n piante, ha legge $B(n, 1/2)$. Vogliamo calcolare

$$\mathbb{P}\{S_4 = 2\} = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^2 = 6\cdot\frac{1}{2^4} = \frac{3}{8}$$

2. Vogliamo calcolare

$$\begin{aligned}\mathbb{P}\{S_4 \geq 2\} &= \mathbb{P}\{S_4 = 2\} + \mathbb{P}\{S_4 = 3\} + \mathbb{P}\{S_4 = 4\} = \\ &= \binom{4}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^2 + \binom{4}{3}\left(\frac{1}{2}\right)^3\left(1-\frac{1}{2}\right)^1 + \binom{4}{4}\left(\frac{1}{2}\right)^4 = \\ &= (6+4+1)\cdot\frac{1}{2^4} = \frac{11}{16}\end{aligned}$$

3. Definiamo ora le variabili aleatorie di Bernoulli $(Y_i)_i$ e poniamo $Y_i = 1$ se l'i-esima pianta ha il fiore bianco; allora le $(Y_i)_i$ hanno legge $Be(1/4)$, e $T_n := \sum_{i=1}^n Y_i$, che esprime il numero di piante con fiori bianchi in n piante, ha legge $B(n, 1/4)$. Vogliamo calcolare

$$\mathbb{P}\{T_n = n\} = \binom{n}{n}\left(\frac{1}{4}\right)^n = \left(\frac{1}{4}\right)^n$$

4. Vogliamo calcolare

$$\mathbb{P}\{T_n \geq 1\} = 1 - \mathbb{P}\{T_n = 0\} = 1 - \binom{n}{0}\left(1-\frac{1}{4}\right)^n = 1 - \left(\frac{3}{4}\right)^n$$

5. Imponendo che $\mathbb{P}\{T_n \geq 1\} > 0.999$, si ha

$$1 - \left(\frac{3}{4}\right)^n > 0.999$$

cioè

$$\left(\frac{3}{4}\right)^n < 0.001$$

Passando ai logaritmi si ha $n \log\frac{3}{4} < \log 0.001$, e infine (ricordandosi che $\log\frac{3}{4} < 0$)

$$n \geq \frac{\log 0.001}{\log 3/4} \simeq 17$$

Esercizio 2.

1. Bisogna fare un test unilatero sulla media di ipotesi $H_0 : \mu = 10$ e alternativa $H_1 : \mu < 10$.

2. Calcoliamo

$$t = \frac{\bar{X} - 10}{s_{\bar{X}}} = \frac{9.267 - 10}{0.4181} = -1.753$$

dove $s_{\bar{X}} = s_X/\sqrt{n} = 2.29/\sqrt{30} = 0.4181$. Il valore critico è dato da $t_\alpha(\nu) = t_{0.01}(29) = -t_{0.99}(29) = -2.462$. Siccome $t > t_\alpha(\nu)$, accettiamo H_0: non ci sono quindi abbastanza elementi per affermare che il farmaco raggiunga il sangue in meno di 10 minuti.

3. Ci serve il quantile $t_{1-\alpha/2}(\nu) = t_{0.995}(29) = 2.756$. Allora gli estremi dell'intervallo di confidenza sono dati da $\bar{X} \pm s_{\bar{X}} t_{1-\alpha/2}(\nu) = 9.267 \pm 0.4181 \cdot 2.756$ e risulta uguale a $[8.1147; 10.4193]$.

   Con questo intervallo NON si può rispondere alla domanda del punto 2., poichè bisogna utilizzare un test unilatero. Se invece, per rispondere alla domanda, fosse servito un test bilatero di livello $\alpha = 0.01$, allora ci sarebbe stato un legame tra l'intervallo di confidenza e il punto 2.

Esercizio 3.

1. Vogliamo fare un test di ipotesi $H_0 : p_A = p_B = p_C$ contro l'alternativa $H_1 : \exists i, j$ tali che $p_i \neq p_j$. Costruiamo la tabella attesa:

ch.card. (A)	26	(20.8)	374	(379.2)	400
ch.cer. (B)	19	(15.6)	281	(284.4)	300
app. (C)	7	(15.6)	293	(284.4)	300
totali	52		948		1000

   Per costruire la tabella attesa, abbiamo usato $\hat{p} = 52/1000 = 0.052$. Siccome tutti i numeri della tabella attesa sono maggiori di 5, si puó usare il metodo del χ^2:

$$\chi^2 = 7.154$$

   Il valore critico è $\chi^2_{0.95}(2) = 5.991$; siccome $\chi^2 > \chi^2_{0.95}(2)$, rifiutiamo H_0 e accettiamo H_1.

2. Facciamo prima un confronto tra A e B, ponendo $H_0 : p_A = p_B$ e $H_1 : p_A \neq p_B$. Abbiamo

A	26	(25.7)	374	(374.3)	400
B	19	(19.3)	281	(280.8)	300
totali	45		655		700

   Per costruire la tabella attesa, abbiamo usato $\hat{p} = 45/700 = 0.064$. Calcoliamo

$$\chi^2 = 0.008$$

   Il valore critico è almeno $\chi^2_{1-\alpha'}(1) > \chi^2_{0.95}(1) = 3.841$; poichè $\alpha' = \alpha/k < \alpha = 0.05$, dove k sarà uguale a 2 o a 3, a seconda del numero di test multipli. Siccome $0.008 = \chi^2 < \chi^2_{0.95}(1)$, accettiamo H_0. Possiamo quindi unificare i gruppi A e B.

   Facciamo ora un confronto tra il nuovo gruppo A-B e il gruppo C, ponendo $H_0 : p_{AB} = p_C$ e $H_1 : p_{AB} \neq p_C$. Abbiamo

A + B	45	(36.4)	655	(663.6)	700
C	7	(15.6)	293	(284.4)	300
totali	52		948		1000

   Per costruire la tabella attesa, abbiamo ancora usato $\hat{p} = 0.052$. Calcoliamo

$$\chi^2 = 7.144$$

   Abbiamo effettuato $k = 2$ tests, per cui $\alpha' = \alpha/2 = 0.025$; il valore critico è quindi $\chi^2_{1-\alpha'}(1) = 5.024$. Siccome $7.144 = \chi^2 > \chi^2_{1-\alpha'}(1)$, rifiutiamo H_0 e accettiamo H_1.

   La conclusione sembra essere che $p_A = p_B \neq p_C$.

Esercizio 4. Partiamo calcolando le quantità:

$$
\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = 34.0667 \quad \text{(0.5 punti)}, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 17.9133 \quad \text{(0.5 punti)},\\
s_X^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = 10.4952 \quad \text{(0.5 punti)},\\
s_Y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right) = 1.5698 \quad \text{(0.5 punti)},\\
s_{XY} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}\right) = 1.1133 \quad \text{(0.5 punti)}
\end{aligned}
$$

1. Calcoliamo i coefficienti della retta di regressione:

$$
\begin{aligned}
b_1 &= \frac{s_{XY}}{s_X^2} = 0.1061\\
b_0 &= \bar{Y} - b_1\bar{X} = 14.2995
\end{aligned}
$$

   La retta di regressione è quindi $y = 0.1061x + 14.2995$.

2. Bisogna effettuare un test di ipotesi $H_0 : \beta_1 = 0$ e alternativa $H_1 : \beta_1 \neq 0$. Bisogna prima calcolare

$$
\begin{aligned}
s_{Y|X} &= \sqrt{\frac{n-1}{n-2}(s_Y^2 - b_1^2 s_X^2)} = 1.2504\\
s_{b_1} &= \frac{s_{Y|X}}{\sqrt{(n-1)\cdot s_X^2}} = 0.1032
\end{aligned}
$$

   Abbiamo allora

$$
t = \frac{b_1 - 0}{s_{b_1}} = 1.0284
$$

   Bisogna confrontare $|t|$ con una legge di Student a $\nu = 15 - 2 = 13$ gradi di libertà. Abbiamo che $t_{0.95}(13) = 1.770 > |t|$, quindi $P > 0.1$. Con una P così alta siamo portati ad accettare H_0, il che significa che sembra non esserci una relazione lineare.

3. Se la regressione lineare fosse un modello valido (cosa per cui non ci sono abbastanza elementi in base al punto 2.), il diametro medio che ci attendiamo per un'età di 50 anni è $y = b_1x + b_0 = 19.6035$, e l'errore standard di questa media è

$$
s := s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{(n-1)s_x^2}} = 1.6749
$$

   Dato che $t_{0.975}(13) = 2.160$, l'intervallo di confidenza cercato ha estremi $y \pm t_{1-\alpha/2}(\nu)s = 19.6035 \pm 2.160 \cdot 1.6749$, ed è quindi $[15.9857; 23.2213]$

Esame di Statistica del 14 dicembre 2007 (Corso di Laurea in Biotecnologie, Universitá degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

Bellardi Micael	$25 + 3^{+}$
Debiasi Stefano	$16.5 + 3^{-}$
Furlan Lucia	$22 + 3^{+}$
Niero Mattia	$24.5 + 3^{-}$
Peroni Edoardo	$15.5 + 3^{-}$

Visione compiti corretti, registrazione voto e/o orali: giovedì 20 dicembre ore 10.00 nel mio studio.
Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.