Esame di Statistica del 19 settembre 2006 (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. Supponiamo che il numero di raffreddori di una persona in un anno sia una variabile aleatoria di Poisson di parametro 3. Viene presentato un nuovo "miracoloso" farmaco che abbassa la media della Poisson a 2 nel 75% della popolazione. Nel restante 25% dei casi, il farmaco non ha effetto.

1. Supponiamo che per un dato individuo il farmaco sia efficace. Qual è la probabilità che si ammali esattamente 3 volte? E che non si ammali nemmeno una volta?
2. Rispondere alle stesse domande nel caso in cui il farmaco non sia efficace su quell'individuo.
3. Supponiamo ora di non sapere se il farmaco è efficace o meno su quell'individuo. Rispondere ancora alle domande del punto 1.
4. Se l'individuo prova il farmaco per due anni e in effetti non si ammala mai, qual è la probabilità che su di lui il farmaco sia efficace? E se invece si ammala esattamente 6 volte?

Esercizio 2. In un certo procedimento chimico, è di fondamentale importanza che il pH di uno dei reagenti sia esattamente 8.20. Supponiamo che 10 misurazioni indipendenti abbiano dato i seguenti valori:

$$8.18 \quad 8.16 \quad 8.17 \quad 8.22 \quad 8.19 \quad 8.17 \quad 8.15 \quad 8.21 \quad 8.16 \quad 8.18$$

1. Calcolare l'intervallo di confidenza al 95% del pH.
2. Effettuare un test per vedere se effettivamente il valore del pH si può considerare uguale a 8.20. Riportare limitazioni al valore P.

Supponiamo ora che il valore 8.20 venga da una misurazione di un altro esperimento, che ha dato anche il valore 0.02 per la deviazione standard. Vogliamo soddisfare le richieste seguenti: se il pH è realmente pari al valore dell'esperimento "di riferimento" (8.20), il test deve affermarlo con probabilità 95%; d'altra parte, se il pH vero differisce dal valore di riferimento di almeno 0.03, tale differenza deve essere evidenziata almeno nel 95% dei casi. Supponiamo anche che la vera deviazione standard sia circa uguale a quella stimata in precedenza.

3. Quanto numerosi dovranno essere i campioni?
4. Se con il nuovo campione avessimo $\bar{X} = 8.31$, che conclusione si trarrebbe?

Esercizio 3. Per studiare la parte microbiologica della fibrosi cistica, si sono recuperati vari batteri da pazienti all'inizio della malattia (¡ 2 anni), a distanza di tempo ("clonali", età ¿ 14 anni), e dall'ambiente, e con questi 3 tipi di batteri si sono infettati topi. Dopo 14 giorni di infezione si è analizzato il tasso di morte / sopravvivenza e quello di infezione / clearance. Lo scopo del lavoro era capire se ceppi clonali e no mostrano differenze significative tra di loro e in rapporto ai ceppi ambientali, che non essendo isolati dai pazienti non dovrebbero essere adattati e in grado di stabilire un'infezione cronica.

	morti	infetti	guariti	totale
ceppi iniziali	21	33	52	106
ceppi clonali	9	52	61	122
ceppi ambientali	26	17	22	65

Rispondere alle seguenti domande (in tutti i tests usare $\alpha = 0.05$).

1. Tra i ceppi clinici isolati alla prima infezione e i ceppi isolati dopo anni di infezione ci sono delle differenze?
2. I ceppi clinici (iniziali + clonali) sono diversi dai ceppi ambientali in termini di mortalità?
3. E in termini di infezione?

Esercizio 4. Un tale è convinto che la velocità di guida non influisca sui consumi della sua macchina. Per verificare questa supposizione, misura i consumi dell'auto a diverse velocità tra le 45 e le 70 miglia orarie. Le miglia percorse alle diverse velocità sono state le seguenti:

velocità (x)	72	80	88	96	104	112	120
km / l (y)	10.27	10.61	9.89	9.34	9.13	8.75	8.41

1. Trovare la retta di regressione del consumo di carburante rispetto alla velocità.
2. Eseguire un test per vedere se in effetti c'è una relazione lineare. Riportare limitazioni al valore P.
3. Supponiamo che questa persona faccia il prossimo viaggio su una strada con limite di 90 km/h. Che consumo medio si attende?
4. Fornire un intervallo di confidenza al 95% del consumo medio calcolato al punto 3.

Soluzioni

Esercizio 1. Se chiamiamo X la variabile aleatoria "numero di raffreddori" ed E l'evento "il farmaco ha effetto", allora X ha legge $Po(2)$ se si verifica E e legge $Po(3)$ se non si verifica E.

1. Dobbiamo calcolare

$$\mathbb{P}(\{X=3\}|E) = e^{-2}\frac{2^3}{3!} = 0.18045, \qquad \mathbb{P}(\{X=0\}|E) = e^{-2}\frac{2^0}{0!} = 0.13534$$

2. Dobbiamo calcolare

$$\mathbb{P}(\{X=3\}|E^c) = e^{-3}\frac{3^3}{3!} = 0.22404, \qquad \mathbb{P}(\{X=0\}|E^c) = e^{-3}\frac{3^0}{0!} = 0.04979$$

3. Utilizzando la formula della probabilità totale, abbiamo

$$\begin{aligned}\mathbb{P}\{X=3\} &= \mathbb{P}(\{X=3\}|E)\mathbb{P}(E) + \mathbb{P}(\{X=3\}|E^c)\mathbb{P}(E^c) = 0.18045\cdot 0.75 + 0.22404\cdot 0.25 = 0.19135,\\ \mathbb{P}\{X=0\} &= \mathbb{P}(\{X=0\}|E)\mathbb{P}(E) + \mathbb{P}(\{X=0\}|E^c)\mathbb{P}(E^c) = 0.13534\cdot 0.75 + 0.04979\cdot 0.25 = 0.11395\end{aligned}$$

4. Se chiamiamo X_i la variabile aleatoria "numero di raffreddori nell'i-esimo anno", allora ha senso supporre che, condizionatamente ad E e ad E^c, le variabili aleatorie siano indipendenti. Allora, per ogni i, X_i ha legge $Po(2)$ se si verifica E e legge $Po(3)$ se non si verifica E; questo significa che $Y := X_1 + X_2$ (numero di raffreddori in 2 anni) ha legge $Po(4)$ se si verifica E e legge $Po(6)$ se non si verifica E.

   Utilizzando la formula di Bayes, abbiamo

$$\begin{aligned}\mathbb{P}(E|\{Y=0\}) &= \frac{\mathbb{P}(\{Y=0\}|E)\mathbb{P}(E)}{\mathbb{P}(\{Y=0\}|E)\mathbb{P}(E) + \mathbb{P}(\{Y=0\}|E^c)\mathbb{P}(E^c)},\\ \mathbb{P}(E|\{Y=6\}) &= \frac{\mathbb{P}(\{Y=6\}|E)\mathbb{P}(E)}{\mathbb{P}(\{Y=6\}|E)\mathbb{P}(E) + \mathbb{P}(\{Y=6\}|E^c)\mathbb{P}(E^c)}\end{aligned}$$

   Bisogna quindi calcolare:

$$\begin{aligned}\mathbb{P}(\{Y=0\}|E) &= e^{-4}\cdot\frac{4^0}{0!} = e^{-4}, \qquad \mathbb{P}(\{Y=0\}|E^c) = e^{-6}\cdot\frac{6^0}{0!} = e^{-6},\\ \mathbb{P}(\{Y=6\}|E) &= e^{-4}\cdot\frac{4^6}{6!} = 0.1041, \qquad \mathbb{P}(\{Y=6\}|E^c) = e^{-6}\cdot\frac{6^6}{6!} = 0.1606\end{aligned}$$

   e i risultati finali sono quindi

$$\begin{aligned}\mathbb{P}(E|\{Y=0\}) &= \frac{e^{-4}\cdot 0.75}{e^{-4}\cdot 0.75 + e^{-6}\cdot 0.25} = 0.9568,\\ \mathbb{P}(E|\{Y=0\}) &= \frac{0.1041\cdot 0.75}{0.1041\cdot 0.75 + 0.1606\cdot 0.25} = 0.6605\end{aligned}$$

Esercizio 2.

1. Calcolando media e varianza campionarie otteniamo $\bar{X} = 8.179$ e $s_X = 0.0223$, quindi $s_{\bar{X}} = s_X/\sqrt{10} = 0.00706$. Abbiamo poi $t_{1-\alpha/2}(\nu) = t_{0.975}(9) = 2.262$. Allora l'intervallo di confidenza ha estremi $\bar{X} \pm t_{1-\alpha/2}(\nu) s_{\bar{X}} = 8.179 \pm 2.262 \cdot 0.00706$, e quindi risulta uguale a $[8.163; 8.195]$.

2. Facciamo un test t sulla media con ipotesi $H_0 : \mu = 8.20$ e alternativa $H_1 : \mu \neq 8.20$. Abbiamo

$$t = \frac{\bar{X} - 8.20}{s_{\bar{X}}} = \frac{8.179 - 8.20}{0.0706} = -2.973$$

I gradi di libertà sono $\nu = 10 - 1 = 9$. Siccome $t_{0.99}(9) = 2.821 < |t| < t_{0.995}(9) = 3.249$, possiamo riportare $0.01 < P < 0.02$. Questo significa che possiamo rifiutare H_0 e accettare H_1.

3. Dato che entrambe le stime della deviazione standard (0.0223 e 0.02) vengono da campioni di taglia 10, hanno egual peso, quindi volendo stimare la varianza usando entrambi questi valori si ha che

$$s_X^2 = \frac{0.0223^2 + 0.02^2}{2} = 0.00044865$$

e quindi $s_X = \sqrt{0.00044865} = 0.021$. Abbiamo poi $\alpha = 0.05$, $\beta = 0.05$, quindi il minimo numero di individui in ogni campione dovrà essere

$$\bar{n} = 2\left(\frac{\sigma}{\delta}\right)^2 (q_{1-\beta} + q_{1-\alpha/2})^2 = 12.94 \simeq 13$$

Bisogna quindi utilizzare almeno 13 misurazioni in ogni campione.

4. Rifacciamo il test t con questi nuovi valori: stavolta abbiamo $H_0 : \mu_X = \mu_Y$ e $H_1 : \mu_X \neq \mu_Y$. Supponiamo che la stima finale della deviazione standard sia 0.021:

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{8.31 - 8.20}{\sqrt{\frac{0.021^2}{13} + \frac{0.021^2}{13}}} = 13.35$$

I gradi di libertà stavolta sono $\nu = 26 - 2 = 24$. Siccome $t_{0.999}(24) = 3.467 < t$, possiamo riportare $P < 0.002$. Questo significa che possiamo rifiutare H_0 e accettare H_1.

Esercizio 3.

1. Dato che le variabili aleatorie nei 2 campioni possono assumere 3 valori, l'unico test che appare appropriato è il test χ^2, con ipotesi $H_0 : p_1^I = p_1^C, p_2^I = p_2^C, p_3^I = p_3^C$ e alternativa $H_1 : \exists i$ tale che $p_i^I \neq p_i^C$. Per poterlo eseguire, bisogna verificare che tutti i numeri nella tabella attesa siano maggiori di 5. Calcoliamo innanzitutto

$$\hat{p}_1 = \frac{21 + 9}{106 + 122} = 0.13, \qquad \hat{p}_2 = \frac{33 + 52}{106 + 122} = 0.37$$

e di conseguenza $\hat{p}_3 = 0.50$. Raccogliamo i dati osservati ed attesi nella seguente tabella:

	morti		infetti		guariti		totale
I	21	(14)	33	(39)	52	(53)	106
C	9	(16)	52	(46)	61	(61)	122
	30		85		113		228

Siccome tutti i numeri della tabella attesa sono maggiori di 5, si puó usare il metodo del χ^2:

$$\chi^2 = \frac{(21-14)^2}{14} + \frac{(9-16)^2}{16} + \frac{(33-39)^2}{39} + \frac{(52-46)^2}{46} + \frac{(52-53)^2}{53} + \frac{(61-61)^2}{61} = 8.68$$

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: ci sono differenze significative nei ceppi clinici prelevati a diverse fasi della malattia.

2. Per rispondere a questa domanda, è necessario accorpare i primi due campioni (riga 1 e 2) insieme, così come gli stati (colonne) 2 e 3. Calcoliamo $\hat{p} = \frac{30+26}{228+65} = 0.19$, e le tabelle osservata e attesa risultano quindi

	morti		vivi		totale
C	30	(44)	198	(184)	228
A	26	(12)	39	(53)	65
	56		237		293

Vogliamo fare un test di ipotesi $H_0 : p_M^C = p_M^A$ e alternativa $H_1 : p_M^C \neq p_M^A$. Siccome tutti i numeri della tabella attesa sono maggiori di 5, si puó usare il metodo del χ^2:

$$\chi^2 = \frac{(30-44)^2}{44} + \frac{(26-12)^2}{12} + \frac{(198-184)^2}{184} + \frac{(39-53)^2}{53} = 23.57$$

Il valore critico è $\chi^2_{0,95}(1) = 5.991$; siccome $\chi^2 > \chi^2_{0,95}(2)$, rifiutiamo H_0 e accettiamo H_1: ci sono differenze significative nella mortalità tra i ceppi clinici e i ceppi ambientali. Notiamo che, se invece di effettuare il test χ^2 avessimo effettuato il test Z, avremmo avuto lo stesso risultato.

3. Per rispondere a questa domanda, questa volta è necessario accorpare i primi due campioni (riga 1 e 2) insieme, così come gli stati (colonne) 1 e 2. Calcoliamo $\hat{p} = \frac{113+22}{228+65} = 0.46$, e le tabelle osservata e attesa risultano quindi

	infetti		guariti		totale
C	113	(105)	115	(123)	228
A	22	(30)	43	(35)	65
	135		158		293

Vogliamo fare un test di ipotesi $H_0 : p_G^C = p_G^A$ e alternativa $H_1 : p_G^C \neq p_G^A$. Siccome tutti i numeri della tabella attesa sono maggiori di 5, si puó usare il metodo del χ^2:

$$\chi^2 = \frac{(113-105)^2}{105} + \frac{(22-30)^2}{30} + \frac{(115-123)^2}{123} + \frac{(43-35)^2}{35} = 5.03$$

Il valore critico è $\chi^2_{0,95}(1) = 5.991$; siccome $\chi^2 > \chi^2_{0,95}(2)$, rifiutiamo H_0 e accettiamo H_1: ci sono differenze significative anche nella guarigione tra i ceppi clinici e i ceppi ambientali. Come per il punto 2., notiamo che, se invece di effettuare il test χ^2 avessimo effettuato il test Z, avremmo avuto lo stesso risultato.

Nota: se invece di fissare il livello $\alpha = 0.05$ avessimo fissato $\alpha = 0.01$, il primo e il terzo test avrebbero dato risultati differenti.

Esercizio 4. Facendo i calcoli si ottiene

$$\sum_{i=1}^{7} X_i = 672, \quad \sum_{i=1}^{7} Y_i = 66.4, \quad \sum_{i=1}^{7} X_iY_i = 6293.9$$

$$\sum_{i=1}^{6} X_i^2 = 66304, \quad \sum_{i=1}^{6} Y_i^2 = 633.74$$

Partiamo calcolando le quantità:

$$\begin{aligned}
\bar{X} &= 96 \quad \text{(0.5 punti)}, \qquad \bar{Y} = 9.48 \quad \text{(0.5 punti)}, \\
s_X^2 &= 116.66 \quad \text{(0.5 punti)}, \\
s_Y^2 &= 0.648 \quad \text{(0.5 punti)}, \\
s_{XY} &= -13.41 \quad \text{(0.5 punti)}
\end{aligned}$$

1. Calcoliamo i coefficienti della retta di regressione:

$$\begin{aligned}
b_1 &= \frac{s_{XY}}{s_X^2} = -0.044 \\
b_0 &= \bar{Y} - b_1\bar{X} = 13.79
\end{aligned}$$

La retta di regressione è quindi $y = -0.044x + 13.79$.

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

$$\begin{aligned}
s_{Y|X} &= \sqrt{\frac{n-1}{n-2}(s_T^2 - b_1^2 s_X^2)} = 0.234 \\
s_{b_1} &= \frac{s_{T|X}}{\sqrt{(n-1)\cdot s_X^2}} = 0.0055
\end{aligned}$$

Abbiamo allora

$$t = \frac{b_1 - 0}{s_{b_1}} = -8.11$$

Bisogna confrontare $|t|$ con una legge di Student a $\nu = 7 - 2 = 5$ gradi di libertà. L'ultimo valore in tavola è $t_{0.999}(5) = 5.893 < |t|$, quindi possiamo riportare $P < 0.002$. Con una P così bassa, possiamo rifiutare H_0 e accettare H_1, concludendo che in effetti c'è una relazione lineare significativa.

3. Il consumo atteso a $x = 90$ km/h è $y = b_1x + b_0 = 9.75$.

4. L'errore standard della stima della y è

$$s := s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_x^2}} = 0.252$$

e $t_{0.975}(5) = 2.570$, quindi l'intervallo di confidenza cercato ha estremi $y \pm t_{1-\alpha/2}(\nu)s = 9.75 \pm 2.570 \cdot 0.252$, ed è quindi $[9.10; 10.40]$

Esame di Statistica del 19 settembre 2006 (Corso di Laurea in Biotecnologie, Universitá degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

Cusinato Claudia	21
Giorio Armando	17.5
Zuccato Valentina	15.5 + 3

Visione compiti corretti, registrazione voto e/o orali: giovedì 21 settembre ore 14.30 aula 1B/50 torre Archimede.
Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.