Esame di Statistica del 9 gennaio 2008 (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. Supponi che vi sia un test per diagnosticare un certo tipo di tumore con affidabilità che è pari al 95% sia per le persone malate, sia per quelle sane, e supponiamo che lo 0.4% della popolazione soffra di questa forma di tumore.
Calcolare la probabilità:

1. di risultare positivi al test;
2. di essere veramente malati se si è risultati positivi al test.
3. Supponiamo ora che questo test venga migliorato e che la probabilità di risultare negativi al test se si è sani aumenti al 99%. Rispondere di nuovo alle prime 2 domande.

Esercizio 2. Vogliamo verificare se due campioni di soluzioni possono provenire dalla stessa sorgente. Vengono quindi fatte 10 misurazioni di pH per campione.

1. Supponendo che la deviazione standard sia uguale a 0.03 per entrambe le soluzioni e di voler effettuare un test di livello 5%, se il loro pH differisce per più di 0.05 vorremmo evidenziarlo con il 95% di probabilità. Quante misurazioni sono necessarie perchè succeda questo?
2. Vengono comunque fatte 10 misurazioni per campione, con i seguenti risultati:

$$\bar{X} = 6.267, \qquad \bar{Y} = 6.285, \qquad s_X = 0.0295, \qquad s_Y = 0.0327$$

   Si può affermare che le due soluzioni provengono da sorgenti diverse? Riportare limitazioni al valore P.

3. Calcolare l'intervallo di confidenza al 95% dei pH di entrambe le soluzioni.

Esercizio 3. La tabella seguente riporta la mortalità infantile in funzione del peso del neonato alla nascita, per 72730 nati vivi a New York nel 1974.

	vivi dopo un anno	deceduti entro un anno
Neonati fino a 2.5 kg	4597	618
Neonati oltre i 2.5 kg	67093	422

Vogliamo studiare la dipendenza tra mortalità e peso alla nascita.

1. Calcolare l'intervallo di confidenza al 99% dei tassi di mortalità dei due gruppi di peso.
2. Calcolare l'intervallo di confidenza al 99% della differenza tra i tassi di mortalità dei due gruppi di peso.
3. Eseguire un test per stabilire se i due gruppi di peso possono avere lo stesso tasso di mortalità, riportando limitazioni al valore P.
4. Confrontare il risultato del punto 3. con i risultati dei primi 2 punti.

Esercizio 4. È stato dimostrato che la probabilità che un quarantenne che abbia fumato per 10 anni si ammali di tumore ai polmoni entro i 20 anni successivi è una funzione del numero giornaliero di sigarette. Un modello approssimato consiste nel supporre che questa funzione sia lineare.
Supponiamo che uno studio estensivo (fatto sui topi ed estrapolato agli esseri umani) dia i seguenti risultati:

n. giornaliero sigarette	5	10	20	30	40	50	60	80
probabilità di ammalarsi	0.061	0.113	0.192	0.259	0.339	0.401	0.461	0.551

1. Calcolare la retta di regressione che lega la probabilità di contrarre un tumore con il numero giornaliero di sigarette.
2. Eseguire un test per verificare se ci può essere una relazione lineare; riportare limitazioni al valore P.
3. Stimare la probabilità di contrarre il tumore per una persona che consumi 35 sigarette al giorno.

Soluzioni

Esercizio 1. Definiamo gli eventi

$$\begin{aligned} M &:= \{ \text{ essere malati } \}, \\ P &:= \{ \text{ risultare positivi al test } \}, \end{aligned}$$

allora abbiamo che

$$\mathbb{P}(M) = 0.004, \qquad \mathbb{P}(P|M) = 0.95, \qquad \mathbb{P}(P^c|M^c) = 0.95$$

1. Poichè Ω è unione disgiunta degli eventi M ed M^c e questi sono tutti non trascurabili, possiamo applicare la formula della probabilità totale:

$$\mathbb{P}(P) = \mathbb{P}(P|M)\mathbb{P}(M) + \mathbb{P}(P|M^c)\mathbb{P}(M^c) = 0.95 \cdot 0.004 + 0.05 \cdot 0.996 = \frac{268}{5000} = 0.0536$$

2. Dato che sia M che P sono non trascurabili, possiamo usare la formula di Bayes:

$$\mathbb{P}(M|P) = \frac{\mathbb{P}(P|M)\mathbb{P}(M)}{\mathbb{P}(P)} = \frac{0.95 \cdot 0.004}{0.536} = \frac{19}{268} = 0.0709$$

3. Con il nuovo test si ha che $\mathbb{P}(P^c|M^c) = 0.99$, mentre le altre probabilità rimangono immutate. Si ha quindi che

$$\mathbb{P}(P) = \mathbb{P}(P|M)\mathbb{P}(M) + \mathbb{P}(P|M^c)\mathbb{P}(M^c) = 0.95 \cdot 0.004 + 0.01 \cdot 0.996 = \frac{344}{25000} = 0.01376$$

e

$$\mathbb{P}(M|P) = \frac{\mathbb{P}(P|M)\mathbb{P}(M)}{\mathbb{P}(P)} = \frac{0.95 \cdot 0.004}{0.01376} = \frac{95}{344} = 0.27616$$

Esercizio 2.

1. Se vogliamo un livello $\alpha = 0.05$ e una potenza di $1 - \beta = 0.95$ e $\sigma = 0.03$, $\delta = 0.05$, allora bisognerà effettuare

$$n \geq \bar{n} = 2\left(\frac{\sigma}{\delta}\right)^2 (q_{1-\alpha/2} + q_{1-\beta})^2 = 2\left(\frac{0.03}{0.05}\right)^2 (1.96 + 1.64)^2 = 9.33$$

e quindi saranno necessarie almeno 10 misurazioni.

2. Bisogna fare un test sulla differenza di medie di ipotesi $H_0 : \mu_X = \mu_Y$ e alternativa $H_1 : \mu_X \neq \mu_Y$. Calcoliamo

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{6.267 - 6.285}{0.0139} = -1.29$$

Se vogliamo fare un test di livello $\alpha = 0.05$ come nel punto 1., il valore critico è dato da $t_{1-\alpha/2}(\nu) = t_{0.975}(18) = 2.100$. Siccome $|t| < t_{1-\alpha/2}(\nu)$, accettiamo H_0: non ci sono quindi abbastanza elementi per affermare che le due sorgenti sono diverse. Siccome poi $t_{0.95}(18) = 1.734 > |t|$, possiamo riportare $P > 0.1$, che in effetti è un valore abbastanza elevato da indurci ad accettare l'ipotesi H_0 anche se non ci fosse stato fornito un livello α.

3. Per entrambi gli intervalli di confidenza ci serve il quantile $t_{1-\alpha/2}(\nu) = t_{0.975}(9) = 2.262$. Allora gli estremi dell'intervallo di confidenza per μ_X sono dati da

$$\bar{X} \pm s_{\bar{X}} t_{1-\alpha/2}(\nu) = 6.267 \pm \frac{0.0295}{\sqrt{10}} \cdot 2.262$$

e risulta uguale a $[6.246; 6.288]$. Gli estremi dell'intervallo di confidenza per μ_Y sono dati da

$$\bar{Y} \pm s_{\bar{Y}} t_{1-\alpha/2}(\nu) = 6.285 \pm \frac{0.0327}{\sqrt{10}} \cdot 2.262$$

e risulta uguale a $[6.262; 6.308]$.

Esercizio 3.

1. Gli intervalli di confidenza cercati sono $\hat{p}_i \pm q_{0.995} s_{\hat{p}_i}$, con $i = 1, 2$. Abbiamo che $\hat{p}_1 = \frac{618}{5215} = 0.1185$, $\hat{p}_2 = \frac{422}{67515} = 0.0063$, e

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 0.0045, \qquad s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.0003$$

   Allora l'intervallo di confidenza per p_1 ha estremi $0.1185 \pm 2.57 \cdot 0.0045$, ed è quindi uguale a $[0.1070; 0.1300]$. L'intervallo di confidenza per p_2 ha estremi $0.0063 \pm 2.57 \cdot 0.0003$, ed è quindi uguale a $[0.0055; 0.0071]$.

2. L'intervallo di confidenza cercato è $\hat{p}_1 - \hat{p}_2 \pm q_{0.995} s_{\hat{p}}$. Abbiamo che

$$\hat{p} = \frac{618 + 422}{5215 + 67515} = 0.0143, \qquad s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.0017$$

   Allora l'intervallo di confidenza ha estremi $0.1122 \pm 2.57 \cdot 0.0017$, ed è quindi uguale a $[0.1079; 0.1165]$.

3. Vogliamo fare un test di ipotesi $H_0 : p_1 = p_2$ contro l'alternativa $H_1 : p_1 \neq p_2$. Per effettuare il test possiamo utilizzare il test Z oppure il test χ^2, che producono gli stessi risultati. Siccome $n_2 \cdot \hat{p} > n_1 \cdot \hat{p} = 67515 \cdot 0.0143 = 965 > 5$, si puó usare il metodo del test Z:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = 65.78$$

   L'ultimo quantile nelle tavole è $2.99 = q_{0.99861}$, e si ha $|Z| > q_{0.99861}$, quindi $P < 0.00278$ (in realtà P è diversi ordini di grandezza più basso), quindi rifiutiamo H_0 in favore di H_1: sembra che il peso alla nascita induca tassi di mortalità diversi.

4. Nel punto 2. avevamo visto che 0 non era compreso nell'intervallo di confidenza al 99%, e questo significa che sicuramente nel test del punto 3. doveva risultare $P < 0.01$, come effettivamente è stato.

Esercizio 4. Partiamo calcolando le quantità:

$$\sum_{i=1}^{8} x_i = 295, \qquad \sum_{i=1}^{8} y_i = 2.377, \qquad \sum_{i=1}^{8} x_i^2 = 15525, \qquad \sum_{i=1}^{8} y_i^2 = 0.9123, \qquad \sum_{i=1}^{8} x_i y_i = 118.395$$

e quindi

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = 36.875, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 0.2971, \\ s_X^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = 663.84, \\ s_Y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right) = 0.0294, \\ s_{XY} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right) = 4.3919 \end{aligned}$$

1. Calcoliamo i coefficienti della retta di regressione:

$$\begin{aligned} b_1 &= \frac{s_{XY}}{s_X^2} = 0.00661 \\ b_0 &= \bar{Y} - b_1\bar{X} = 0.0531 \end{aligned}$$

   La retta di regressione è quindi $y = 0.00661x + 0.0531$.

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

$$\begin{aligned} s_{Y|X} &= \sqrt{\frac{n-1}{n-2}(s_Y^2 - b_1^2 s_X^2)} = 0.0209 \\ s_{b_1} &= \frac{s_{Y|X}}{\sqrt{(n-1)\cdot s_X^2}} = 0.000306 \end{aligned}$$

   Abbiamo allora

$$t = \frac{b_1 - 0}{s_{b_1}} = 21.62$$

   Bisogna confrontare $|t|$ con una legge di Student a $\nu = 8 - 2 = 6$ gradi di libertà. Abbiamo che $t_{0.999}(6) = 5.208 < |t|$, quindi $P < 0.002$. Con una P così bassa siamo portati a rifiutare H_0 e ad accettare H_1, il che significa che sembra esserci una relazione lineare significativa tra la probabilità di sviluppare un tumore e il consumo giornaliero di sigarette.

3. Se la regressione lineare è un modello valido , la probabilità di contrarre un tumore per una persona che consumi 35 sigarette al giorno è $y = b_1 x + b_0 = 0.2847$.

Esame di Statistica del 9 gennaio 2008 (Corso di Laurea in Biotecnologie, Universitá degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

Alessio Enrico	$23.5 + 3^-$
Begolo Daniela	31.5
Bellini Stefano	$18 + 2^-$
Bergamo Giorgia	24
Bortolotto Alberto	19.5
Cesarato Fabio	$30.5 + 3^+$
Cusinato Giulia	25
Dal Bello Simonetta	$19.5 + 3^-$
Dal Lago Marco	$17 + 3^-$
Duso Enrico	19.5
Ferro Giulia	$20.5 + 2^+$
Finotti Giulia	$18 + 3^+$
Francescato Federica	$20.5 + 3^+$
Genovesi Elisa	18
Gris Barbara	23.5
Leone Kevin	$19.5 + 3^-$
Magaraggia Mattia	17
Mosole Simone	26.5
Passoni Gabriella	$21 + 3^-$
Perin Giorgio	$29.5 + 3^+$
Polato Marco	$28.5 + 3^-$
Poles Lara	17.5
Ramponi Martina	$29.5 + 1^+$
Sciortino Marianna	18
Sudiro Cristina	21.5
Venturato Andrea	26
Vigolo Michele	$20 + 3^+$

Visione compiti corretti, registrazione voto e/o orali: lunedì 14 gennaio ore 17.00 aula 2BC/60 torre Archimede., oppure giovedí 17 gennaio ore 10 aula 2AB/45 torre Archimede.
Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.