Esame di Statistica del 7 luglio 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. La statura delle donne adulte negli Stati Uniti ha una distribuzione normale di media 64.5 pollici e deviazione standard 2.4 pollici.

1. Calcola la probabilità che una donna scelta a caso sia alta meno di 63 pollici;
2. meno di 70 pollici;
3. tra 63 e 70 pollici.
4. Alice è alta 72 pollici. Che percentuale della popolazione femminile adulta è più bassa di lei?
5. Calcola la probabilità che la media aritmetica della statura di due donne scelte a caso sia superiore a 66 pollici.

Esercizio 2. Un allevamento commerciale di salmoni dichiara che il peso medio dei suoi pesci sia superiore ai 3.8 kg.

1. Supponendo che il peso dei salmoni sia distribuito in modo normale, che test si deve usare per verificare se l'allevamento non stia facendo "pubblicità ingannevole"?
2. Supponiamo che in un campione casuale di 16 salmoni il peso medio sia risultato di 3.6 kg, con una deviazione standard di 0.6 kg. Che conclusione se ne trae al 5% di significatività? E all'1%?
3. Riportare limitazioni al valore P.
4. Calcolare l'intervallo di confidenza sul peso medio dei salmoni.

Esercizio 3. Vogliamo determinare se le cause intentate contro medici siano più frequenti per certi tipi di interventi che per altri. Studiando campioni casuali per tre tipi di intervento, si ottengono i seguenti dati:

	totale	n. cause
Chirurgia cardiaca	400	16
Chirurgia cerebrale	300	19
Appendicectomia	300	7

1. Calcolare l'intervallo di confidenza al 95% della proporzione totale delle operazioni che portano ad una causa giudiziaria.
2. Fare un test di livello $\alpha = 5\%$ per vedere se la percentuale di operazioni che porta ad una causa giudiziaria sia la stessa per i tre tipi di interventi.
3. Se necessario, individuare quali sono gli interventi che hanno tale percentuale uguale tra di loro e quali no.

Esercizio 4. I dati seguenti mettono in relazione la pressione sistolica di un gruppo di individui con il loro peso (a parità di stili di vita e di corporature):

peso in libbre (x)	162	184	166	150	192	155	184	202	174
pressione (y)	112	144	138	145	152	10	18	160	108

1. Trovare la retta di regressione della pressione rispetto al peso.
2. Eseguire un test per vedere se in effetti c'è una relazione lineare. Riportare limitazioni al valore P.
3. Determinare un intervallo che, con il 95% di confidenza, contenga la media delle pressioni di tutti gli individui (con corporatura e stile di vita simili ai precedenti) che pesano 180 libbre.
4. Determinare un intervallo che, con il 95% di confidenza, contenga la pressione di un singolo individuo che pesa 180 libbre.

Soluzioni

Esercizio 1.

1. L'altezza di una donna scelta a caso è una variabile aleatoria $X \sim N(64.5; 2.4^2)$. Allora

$$\begin{aligned}\mathbb{P}\{X < 63\} &= \mathbb{P}\left\{\frac{X-64.5}{2.4} < \frac{63-64.5}{2.4}\right\} = \mathbb{P}\{Z < -0.63\} = \mathbb{P}\{Z \le -0.63\} = F_Z(-0.63) = \\ &= 1 - F_Z(0.63) = 1 - 0.73565 = 0.26435\end{aligned}$$

dove $Z := (X - 64.5)/2.4 \sim N(0, 1)$.

2. Allo stesso modo,

$$\mathbb{P}\{X < 70\} = \mathbb{P}\left\{\frac{X-64.5}{2.4} < \frac{70-64.5}{2.4}\right\} = \mathbb{P}\{Z < 2.29\} = \mathbb{P}\{Z \le 2.29\} = F_Z(2.29) = 0.98899$$

3. Dobbiamo calcolare

$$\begin{aligned}\mathbb{P}\{63 < X < 70\} &= \mathbb{P}\{63 < X \le 70\} = \mathbb{P}\left\{\frac{63-64.5}{2.4} < \frac{X-64.5}{2.4} \le \frac{70-64.5}{2.4}\right\} = \\ &= \mathbb{P}\{-0.63 < Z \le 2.29\} = F_Z(2.29) - F_Z(-0.63) = 0.98899 - 0.26435 = 0.72464\end{aligned}$$

4. Siccome la popolazione femminile ha sempre legge $N(64.5; 2.4^2)$, la proporzione di popolazione femminile più bassa di 72 pollici si ottiene calcolando

$$\begin{aligned}\mathbb{P}\{X < 72\} &= \mathbb{P}\left\{\frac{X-64.5}{2.4} < \frac{72-64.5}{2.4}\right\} = \mathbb{P}\{Z < 3.13\} = \mathbb{P}\{Z \le 3.13\} = F_Z(3.13) > \\ &> F_Z(2.99) = 0.99861\end{aligned}$$

dove abbiamo dovuto approssimare il valore $F_Z(3.13)$ con l'ultimo valore presente in tavola. La percentuale di donne più basse di 72 pollici è dunque maggiore del 99.861%.

5. Se scegliamo a caso due donne, allora la media aritmentica delle altezze $Y = \frac{X_1+X_2}{2}$ è una combinazione lineare di gaussiane indipendenti, e ha quindi legge gaussiana con parametri

$$\begin{aligned}\mathbb{E}[Y] &= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2]}{2} = 64.5 \\ \text{Var}\,[Y] &= \frac{\text{Var}\,[X_1 + X_2]}{4} = \frac{\text{Var}\,[X_1] + \text{Var}\,[X_2]}{4} = \frac{2 \cdot 2.4^2}{4} = \frac{2.4^2}{2}\end{aligned}$$

quindi si ha

$$\begin{aligned}\mathbb{P}\{Y > 66\} &= \mathbb{P}\left\{\frac{Y-64.5}{\frac{2.4}{\sqrt{2}}} > \frac{66-64.5}{\frac{2.4}{\sqrt{2}}}\right\} = \mathbb{P}\{Z > 0.89\} = 1 - \mathbb{P}\{Z \le 0.89\} = \\ &= 1 - F_Z(0.89) = 1 - 0.81327 = 0.18673\end{aligned}$$

Esercizio 2.

1. Il test più adatto è un test t unilatero sulla media con ipotesi $H_0 : \mu = 3.8$ e alternativa $H_1 : \mu < 3.8$. Difatti se è vero che il peso è (maggiore o) uguale a 3.8 kg, ci potrebbero essere salmoni che pesano di meno, ma se i salmoni pesassero di più non si avrebbe "pubblicità ingannevole". Se invece un campione pesasse molto di meno del peso medio dichiarato, saremmo portati a credere che in realtà la media sia minore.

2. Abbiamo che $\bar{X} = 3.6$, $s_X = 0.6$, quindi $s_{\bar{X}} = 0.6/\sqrt{16} = 0.15$ e

$$t = \frac{\bar{X} - 3.8}{s_{\bar{X}}} = \frac{3.6 - 3.8}{0.15} = -1.33$$

I gradi di libertà sono $\nu = 16 - 1 = 15$, e il valore critico è della forma $t_\alpha(\nu)$. Per $\alpha = 0.05$ si ottiene allora $t_{0.05}(15) = -t_{0.95}(15) = -1.753$, e per $\alpha = 0.01$ si ottiene $t_{0.01}(15) = -t_{0.99}(15) = -2.602$. In entrambi i casi, $t > t_\alpha(\nu)$, e quindi accettiamo H_0: sebbene il peso dei salmoni sia più basso di quello dichiarato, non abbiamo abbastanza elementi per affermare che ci sia stata pubblicità ingannevole.

3. Il quantile nella tabella più vicino a t è $t_{0.05}(15) = -1.753$, quindi riportiamo $P > 0.05$.
4. Per gli intervalli di confidenza al 95% e al 99% ci servono i quantili di livello $1-\alpha/2$, con $\alpha = 0.05, 0.01$. Abbiamo che $t_{0.975}(15) = 2.131$, $t_{0.995}(15) = 2.946$; gli estremi dell'intervallo sono $\bar{X} \pm t_{1-\alpha/2}(\nu) s_{\bar{X}}$, e facendo i conti l'intervallo di confidenza al 95% risulta $[3.281; 3.919]$ e quello al 99% risulta $[3.159; 4.041]$.

Esercizio 3.

1. La stima di questa proporzione è
$$\hat{p} = \frac{16+19+7}{400+300+300} = 0.042$$
con errore standard della stima uguale a
$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.00634$$
Abbiamo infine $q_{0.975} = 1.96$, quindi gli estremi dell'intervallo sono $0.042 \pm 1.96 \cdot 0.00634$, e l'intervallo è $[0.0296; 0.0544]$.
2. Vogliamo fare un test di ipotesi $H_0 : p_1 = p_2 = p_3$ contro l'alternativa $H_1 : \exists i, j$ tali che $p_i \neq p_j$. Raccogliamo i dati osservati ed attesi nella seguente tabella:

	cause		no		totale
1	16	(16.8)	384	(383.2)	400
2	19	(12.6)	281	(287.4)	300
3	7	(12.6)	293	(287.4)	300
	42		958		1000

Per costruire la tabella attesa, abbiamo usato $\hat{p} = 0.042$, come calcolato nel punto precedente. Siccome tutti i numeri della tabella attesa sono maggiori di 5, si puó usare il metodo del χ^2:
$$\begin{aligned}\chi^2 &= \frac{(16-16.8)^2}{16.8} + \frac{(384-383.2)^2}{383.2} + \frac{(19-12.6)^2}{12.6} + \frac{(281-287.4)^2}{287.4} + \\ &+ \frac{(7-12.6)^2}{12.6} + \frac{(293-287.4)^2}{287.4} = 6.03\end{aligned}$$
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 (4 punti).
3. Facciamo ora dei confronti multipli per vedere se ci sono gruppi omogenei. Partiamo dai due gruppi che appaiono più vicini:

	causa		no		totale
1	16	(13)	384	(387)	400
3	7	(10)	293	(290)	300
	23		677		700

Stavolta testiamo $H_0 : p_1 = p_3$ contro $H_1 : p_1 \neq p_3$, e $\hat{p} = \frac{16+7}{400+300} = 0.033$. Costruiamo χ^2:
$$\begin{aligned}\chi^2 &= \frac{(16-13)^2}{13} + \frac{(384-387)^2}{387} + \\ &+ \frac{(7-10)^2}{10} + \frac{(293-290)^2}{290} \simeq 1.50\end{aligned}$$
Abbiamo che $\chi^2_{1-\alpha'}(1) \geq \chi^2_{0,975}(1) = 5.024 > \chi^2$, poichè $\alpha' = \alpha/k \leq 0.025$, dove $k \geq 2$ è il numero totale di test multipli. Accettiamo quindi H_0 e possiamo accorpare i due gruppi:

	sì		no		totale
1-3	23	(29.4)	677	(670.6)	700
2	19	(12.6)	281	(287.4)	8000
	42		958		1000

Si ha ora $\chi^2 = 4.85$. Siccome $\chi^2_{0,975}(1) = 5.024$, dobbiamo concludere che accettiamo H_0. Sembrerebbe quindi che tutti e tre i gruppi sono da considerare con lo stesso tasso, contro il risultato del punto 2! Questo può succedere, in quanto la disuguaglianza di Bonferroni ci fa fissare un α' abbastanza basso da garantirci di non commettere errori di prima specie con probabilità più alta di α; in realtà, la disuguaglianza di Bonferroni fornisce un α' volutamente "troppo" basso, il che porta a quantili più alti e quindi a volte capita di accettare ipotesi false con più facilità che con altre disuguaglianze.

Esercizio 4. Facendo i calcoli si ottiene

$$\sum_{i=1}^{6} X_i = 1569, \quad \sum_{i=1}^{6} Y_i = 1187, \quad \sum_{i=1}^{6} X_i Y_i = 208356$$

$$\sum_{i=1}^{6} X_i^2 = 275981, \quad \sum_{i=1}^{6} Y_i^2 = 159741$$

Partiamo calcolando le quantità:

$$\begin{aligned}
\bar{X} &= 174.3 \quad (0,5 \text{ punti}), \qquad \bar{Y} = 131.9 \quad (0,5 \text{ punti}),\\
s_X^2 &= 306.5 \quad (0,5 \text{ punti}),\\
s_Y^2 &= 398.6 \quad (0,5 \text{ punti}),\\
s_{XY} &= 177.7 \quad (0,5 \text{ punti})
\end{aligned}$$

1. Calcoliamo i coefficienti della retta di regressione:

$$\begin{aligned}
b_1 &= \frac{s_{XY}}{s_X^2} = 0.58\\
b_0 &= \bar{Y} - b_1 \bar{X} = 30.763
\end{aligned}$$

La retta di regressione è quindi $y = 0.58x + 30.763$.

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

$$\begin{aligned}
s_{Y|X} &= \sqrt{\frac{n-1}{n-2}(s_T^2 - b_1^2 s_X^2)} = 18.376\\
s_{b_1} &= \frac{s_{T|X}}{\sqrt{(n-1) \cdot s_X^2}} = 0.371
\end{aligned}$$

Abbiamo allora

$$t = \frac{b_1 - 0}{s_{b_1}} = 1.56$$

Bisogna confrontare $|t|$ con una legge di Student a $\nu = 9 - 2 = 7$ gradi di libertà. Il più basso valore in tavola è $t_{0.95}(7) = 1.894 > |t|$, quindi possiamo riportare $P > 0.12$. Con un valore P così alto, possiamo accettare H_0, concludendo che in effetti la relazione lineare non sembra essere significativa.

3. Supponendo che la relazione lineare sia valida, la pressione attesa per le persone che pesano 180 libbre è $y = b_1 x + b_0 = 135.18$, con un errore standard della stima di

$$s := s_{Y|X} \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_x^2}} = 6.48$$

e $t_{0.975}(7) = 2.364$, quindi l'intervallo di confidenza cercato ha estremi $y \pm t_{1-\alpha/2}(\nu)s = 135.18 \pm 2.364 \cdot 6.48$, ed è quindi $[119.9; 150.5]$

4. La pressione attesa per una persona che pesa 180 libbre è $y = b_1 x + b_0 = 135.18$, con un errore standard della stima di

$$s := s_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_x^2}} = 19.48$$

quindi l'intervallo di confidenza cercato ha estremi $y \pm t_{1-\alpha/2}(\nu)s = 135.18 \pm 2.364 \cdot 19.48$, ed è quindi $[89.1; 181.2]$

Esame di Statistica del 7 luglio 2006 (Corso di Laurea in Biotecnologie, Universitá degli Studi di Padova) (docente: Tiziano Vargiolu)

Hanno superato la prova:

Bazzo Irene	17
Cortese Matteo	17 + 3
Lanciai Federico	20
Mazzocco Giovanni	17
Procopio Maria	26 + 3
Stevan Martina	19
Verna Michela	22.5

Visione compiti corretti, registrazione voto e/o orali: martedì 11 luglio ore 15.30 aula LuM250 via Luzzatti (complesso Paolotti).
Verrà data precedenza alla registrazione voti a chi accetta il voto dello scritto e ha il bonus di + 3.