TUTORIAL EXERCISE PACKAGE - 2013

MATH1015 BIOSTATISTICS

Semester 1	**Tutorial Week 13**	2013

Given observations $(x_1, y_1), \ldots, (x_n, y_n)$, the correlation coefficient between x and y is $r = \dfrac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$ where $L_{xy} = \sum_i x_i y_i - \dfrac{1}{n}\left(\sum_i x_i\right)\left(\sum_i y_i\right)$, $L_{xx} = \sum_i x_i^2 - \dfrac{1}{n}\left(\sum_i x_i\right)^2$, and $L_{yy} = \sum_i y_i^2 - \dfrac{1}{n}\left(\sum_i y_i\right)^2$.

r^2 gives the percentage of variation in Y explained by the linear relationship with x.

The regression line is $\hat{y} = a + bx$ where $b = \dfrac{L_{xy}}{L_{xx}}$ and $a = \bar{y} - b\bar{x}$.

To test the significance of the regression line $H_0 : \beta = 0$, the test statistic is $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}} \sim t_{n-2}$ where $s^2 = \dfrac{L_{yy} - bL_{xy}}{n-2}$.

Tutorial discussion: Q3 to Q8 marked with *

Tutorial and Revision Questions

1. Multiple choice An estimate of the correlation coefficient $r = -0.7761$ tells us that:

(a) The relationship between x and y is strong and positive.

(b) The relationship between x and y is strong and negative.

(c) The relationship between x and y is weak and positive.

(d) The relationship between x and y is weak and negative.

(e) There is no relationship between x and y.

2. Multiple choice From Q1, the proportion of variation in y explained by the linear relationship between x and y is:

(a) 6.917 % (b) 96% (c) -77.6% (d) 60.2% (e) -20%

3. * J.B. Haldane is responsible for showing how carbon dioxide levels in blood influence breathing rates by affecting the acidity of the blood. In one experiment he administered varying doses of sodium bicarbonate with the following results:

Dose (in grams):	x	30	40	50	60	70	80	90	100
Breathing rate (in breath/min) :	y	16	14	13	13	11	12	9	9

From last week, we obtain:

$$L_{xx} = 4200,\ L_{yy} = 40.88,\ L_{xy} = -395,\ r = -0.953,\ b = -0.0940476,\ a = 18.23810$$

(a) What is the proportion of variation in Y that can be explained by the linear regression line of Y on X?

(b) Calculate the residuals for the first two pairs of observations.

(c) Test if the regression line is significant.

(d) Interpret the slope estimate b.

4. * **Multiple choice** (Practice Quiz 2 Q8) Summary statistics for two random samples from two independent normal populations are reported below:

$$\begin{array}{llll} \text{Sample 1:} & n_1 = 9 & \overline{x}_1 = 10 & s_1 = 5 \\ \text{Sample 2:} & n_2 = 11 & \overline{x}_2 = 12 & s_2 = 3 \end{array}$$

Estimates of mean and standard error for the sample mean difference $\overline{X}_1 - \overline{X}_2$ are closest to:

(a) -2 and 0.8864 (b) 2 and 0.9354 (c) -2 and 1.9039 (d) -2 and 0.9354 (e) -2 and 1.8041

5. * **Multiple choice** (Practice Quiz 2 Q11) A sample of size 200 will be taken at random from a population with binary outcomes. Given that the population proportion is 0.60, the approximate probability that the sample proportion will be *greater* than 0.58 is closest to:

(a) 0.2810 (b) 0.7190 (c) 0.5163 (d) 0.5900 (e) 0.7167

6. * **Multiple choice** (Practice Quiz 2 Q12) To test if the acceptance rate of production line 1 is higher than the acceptance rate of line 2, two random samples of 100 products each are selected from the two lines and the numbers of acceptable products are 76 and 68 respectively. The test statistic is closest to:

(a) 0.3984 (b) 0.8909 (c) 2.5198 (d) 0.7869 (e) 1.2599

7. * **Multiple choice** (Quiz 2A Q1) Suppose that $X_1, X_2, \cdots, X_9$ is an independent random sample of heights of certain plants following the normal distribution, $N(5, 3^2)$. Then $P(\sum_{i=1}^{9} X_i > 54)$ is closest to:

(a) 0.8413 (b) 0.9987 (c) 0.0013 (d) 0.9772 (e) 0.1587

8. * A random sample of 10 observations from a normal population of basal body temperatures with unknown mean μ gives the following summary: $\bar{x} = 36^\circ$ C and $s = 0.034^\circ$ C. Construct an interval estimate of the true mean basal body temperature. Use a confidence level of 80%. If the sample is converted to degrees Fahrenheit with the formula $F = \left(\frac{9}{5}\right) C + 32$, then adjust the sample mean and standard deviation appropriately and recalculate your CI.

Semester 1 | Problem Sheet Week 13 | 2013

---

1. (from 2002 exam) The unordered stem-and-leaf display given below reports the final examination marks of 40 students in a statistics course in 2001.

2	8 4
3	5 1 5
4	0 2 0 3
5	0 3 6 3 8
6	0 1 2 7 4 4 9
7	2 3 2 5 5 8 6 9 9
8	4 0 0 5 7 7
9	0 0 5 2

(i) Jonah estimated the median statistics mark using the average of 20th and 21st observations from the above display as $(64 + 69)/2 = 66.5$. Explain why she will receive no marks for this median estimate.

(ii) Calculate the correct median statistics mark and prepare the boxplot for this data set. Describe its shape.

(iii) What proportion of students receive high distinctions (HD), if the minimum mark for an HD is 85?

2. Recorded here is the frequency distribution of the blood types of 96 persons who have volunteered to donate blood at a plasma center.

Blood type	O	A	B	AB	Total
Observed frequency, O_i	38	43	10	5	96

What frequencies are expected under the hypothesis that

(a) the four blood types O, A, B, AB are equally likely.

(b) the four blood types O, A, B, AB are in the ratios 4:4:1:1.

Test each null hypothesis in (a) and (b).

3. (a) Four passing grades (HD, D, C, P) are allocated to students who gain 50 or more marks in an examination. The probabilities that a student will get a HD, a D, a C or a P in a course are (respectively) 0.06, 0.14, 0.23 and 0.42. If a student from this course is to be selected at random, what is the probability that he or she will not qualify for any of the above four passing grades? If two students are to be selected at random from the course, what is the probability that one with a HD and the other without any passing grade?

(b) Records show that the probability is 0.60 that a stolen car in a certain suburb of NSW will be recovered within one week. Let X be the number of cars recovered from n stolen cars in a particular week. Assuming the independence of events explain why X has a binomial distribution.

(a) Use binomial tables to find the probability that at least 7 of the 10 stolen cars will be recovered.

(b) Write down an expression for the exact probability that 12 of the 25 stolen cars will be recovered and use it to find this probability to 4 decimal places.

(c) Let $X \sim N(48, 6^2)$. Find $P(X > 63.5)$.

From a survey conducted in 2002, it was found that 25% of the undergraduates lived in rented housing. A welfare group believes that this percentage has increased in 2004. To test this claim a random sample of 192 students was taken and found that 64 of them lived in rented housing. Justify the claim of welfare group using a suitable argument based on hypothesis testing.

4. Ten cars of the same model are equipped to record both speed (x), in km/h, and petrol consumption (y), in km/4litres. The cars are driven along a course at different speeds. The resulting data are as follows:

Car	A	B	C	D	E	F	G	H	I	J
Speed (x)	25	30	35	40	45	50	55	60	65	70
Petrol (y)	34.5	32.4	35.0	30.6	31.7	32.2	30.4	28.5	25.7	26.6

$(\sum_{i=1}^{10} x_i = 475, \sum_{i=1}^{10} x_i^2 = 24,625, \sum_{i=1}^{10} y_i = 307.6, \sum_{i=1}^{10} y_i^2 = 9547.56, \sum_{i=1}^{10} x_i y_i = 14,234.5)$

(a) Find the proportion of variability in y explained by a linear regression of y on x. Briefly comment on the degree of linear association between the two variables x and y.

(b) Calculate the fitted linear least squares regression equation of y on x.

(c) Is it reasonable to estimate the petrol consumption using this model when the speed is 120km/h? Explain.

5. Try: Q11.21 to Q11.23 (P.227); Q11.36 (P.228); Q11.38 to Q11.41 (P.229).

GOOD LUCK