8112

THE UNIVERSITY OF SYDNEY

FACULTIES OF ARTS AND SCIENCE

STAT 2012: Statistical Tests

Semester 2, 2011 Time allowed: Two hours

Answer ALL Five questions.
For each hypothesis test, unless otherwise stated, you should state the two hypotheses, the test statistic, the test assumptions, the p-value and the conclusion. If an exact p-value cannot be found from the table, state the range of values from the table.

1. Carbon monoxide in cigarettes is thought to be hazardous to the fetus of a pregnant woman who smokes. If she cannot give up smoking completely, will the shift to smoking low-tar cigarettes help to reduce the carbon monoxide level in the blood? To investigate this hypothesis, blood samples were first drawn from 7 pregnant women after smoking ordinary cigarettes and the percentage of blood hemoglobin bound to carbon monoxide as carboxyhemoglobin (COHb) was taken from each woman. After a wide enough time interval, the same procedure was applied to the same group of women after they have shifted to smoke low-tar cigarettes. The percentages of COHb after smoking ordinary (x_i) and low-tar (y_i) cigarettes are reported in the following table.

Subject i	1	2	3	4	5	6	7	mean	variance
Ordinary x_i	4.0	5.0	6.3	5.8	5.0	4.2	5.2		
Low-tar y_i	4.2	3.5	4.9	5.1	5.6	3.7	5.2		
Reduction d_i	-0.2	1.5	1.4	0.7	-0.6	0.5	0.0	0.4714	0.6324

(a) (3 marks) Test at $\alpha = 0.05$ significance level if there is a reduction in blood COHb level after smoking *low-tar* cigarettes using the *sign test*.

(b) (2 marks) The test in (a) is repeated using the Z-test with the true variance for d_i being $\sigma_d^2 = 1$. If the rejection region is $\bar{d} \geq k$, find the value of k for $\alpha = 0.05$. Based on this rejection region, state whether the null hypothesis in (a) should be rejected.

(c) (3 marks) For the Z-test in (b), calculate the probability of committing a type II error $\beta(1)$ when the true mean difference $\mu_d = 1$.

(d) (2 marks) Under the assumption in (b), determine the sample size such that the *two-sided* 95% confidence interval for the true *mean* difference is 1 unit in width.

(e) (3 marks) Construct a *two-sided* 95% confidence interval for the *variance* of d_i. Based on this confidence interval, can we conclude that the variance differs from 1 at $\alpha = 0.05$?

2. (a) (8 marks) Suppose the low-tar cigarettes in Question 1 are called brand A cigarettes. The experiment in Question 1 is repeated with another brand (brand B) of low-tar cigarettes on a different group of pregnant women. The reduction in carboxyhemoglobin (COHb) for both brands is reported in the following table.

Brand	**Reduction in COHb**						
A	-0.2	1.5	1.4	0.7	-0.6	0.5	0.0
B	1.5	2.7	2.6	1.3	1.8	1.7	

Wilcoxon rank sum test is used to test if brand B is more effective in reducing the COHb level than brand A at $\alpha = 0.05$ significance level.

(i) Calculate *only* the test statistic W for the sum of ranks for brand B and the p-value. Hint: the zero data in Brand A should NOT be dropped.

(ii) Condition on the observed combine set of ranks, work out all cases for the *ranks of brand B* with the rank sum being at least the value W in (i). Hence calculate the exact p-value in (i) being the proportion of these cases.

(b) (5 marks) The experiment is further repeated with two more brands (C and D) of low-tar cigarettes on two more groups of pregnant women and the reduction in COHb for brands B to D are reported below:

Brand	**Reduction in COHb**					
B	1.5	2.7	2.6	1.3	1.8	1.7
C	1.6	2.4	0.7	1.2	2.0	
D	3.2	2.5	3.0	1.9	3.6	

Test at $\alpha = 0.05$ significance level if the reduction in COHb differs across the three brands (B to D) of low-tar cigarettes using the *Kruskal Wallis* test.

3. A researcher in a television rating company investigates if there are differences in television viewing habits among three cities in Australia. A sample of five adults from five different age groups is taken from each city and their numbers of hours spent watching TV in the previous week are reported in the following table.

Age	**Melbourne**	**Canberra**	**Sydney**	**Mean** $\bar{y}_{i\cdot}$
20-30	27	23	22	24.000
30-40	24	30	17	23.667
40-50	18	25	19	20.667
50-60	29	31	27	29.000
60-70	32	34	24	30.000
Mean $\bar{y}_{\cdot j}$	26.0	28.6	21.8	25.467

The following summary statistics are given:

$$\sum_{i=1}^{5}\sum_{j=1}^{3} y_{ij}^2 = 10104; \quad \sum_{i=1}^{5} \bar{y}_{i\cdot}^2 = 3304.252; \quad \sum_{j=1}^{3} \bar{y}_{\cdot j}^2 = 1969.2; \quad \bar{y} = \frac{1}{15}\sum_{i=1}^{5}\sum_{j=1}^{3} y_{ij} = 25.467;$$

(a) (6 marks) What type of experimental design was used in this study? Provide the ANOVA table.

(b) (4 marks) Test at $\alpha = 0.05$ significance level if the average number of hours of television watching differs across the three cities.

(c) (3 marks) Suppose that the information of age is dropped from the table. In other words, the five adults from each city are randomly selected regardless of their ages. Calculate *only* the revised p-value in (b). Explain briefly the new result. Hint: this is a *completely randomized* design and the new $SSR_{new} = SSR + SSB$ in (b).

4. Modern medical practice tells us not to encourage babies to become too fat. Medical research indicates that there is a positive correlation between the weight X (lb) of a female at age 1 and her weight Y (lb) at age 30. A random sample of medical files produced the following information for $n = 14$ females

x_i (lb)	21	25	23	24	20	15	25	21	17	24	26	22	18	19
y_i (lb)	110	127	115	125	130	102	145	127	118	131	140	121	98	105

The following summary statistics are given:

$$\sum_{i=1}^{n} x_i = 300, \quad \sum_{i=1}^{n} y_i = 1694, \quad \sum_{i=1}^{n} x_i^2 = 6572, \quad \sum_{i=1}^{n} y_i^2 = 207492, \quad \sum_{i=1}^{n} x_i y_i = 36763,$$

(a) (4 marks) Determine the least squares regression line with y as the dependent variable.

(b) (4 marks) Test at $\alpha = 0.05$ significance level if the regression model in (a) is significant. Hint: test if the slope of the regression line is zero.

(c) (2 marks) Determine the coefficient of determination R^2.

(d) (6 marks) Determine a 95% *prediction* interval for the weight of a woman at age 30 when her weight at age 1 is 12 lb. Should the regression model in (a) be used to predict the weight of a female at age 30 when her weight is only 12 lb at age 1? Explain briefly.

(e) (2 marks) Prove, in general, that the regression line always passes through the point of the two sample means $(\bar{x}, \bar{y})$.

********** **Please Turn Over** **********

5. The monthly rainfall in a city, observed over 10 years, yielded a mean of $\bar{y} = 100$ (mm) and a standard deviation of $s = 20$. The rainfall data were grouped into class intervals as shown below:

Interval (in mm)	**Frequency**
Less than 80	28
80 to 100	41
100 to 120	35
More than 120	16

The Chi-square test is conducted to test if the monthly rainfalls are normally distributed.

(a) (5 marks) Complete the following table *in your answer book* to calculate expected frequencies and hence the chi-square test statistic.

Interval	O_i	**Probability of event** π_i	$E_i = 120\pi_i$	$\frac{(O_i-E_i)^2}{E_i}$
Less than 80	28	$\Pr(Z < -1) =$ ______		
80 to 100	41	$\Pr(-1 < Z < 0) =$ ______		
100 to 120	35	$\Pr(0 < Z < 1) =$ ______		
More than 120	16	$\Pr(Z > 1) =$ ______		
Sum	120	1.0000	120	

Hint: use the *normal table* to find the probability $\Pr(Z > 1)$.

(b) (3 marks) Test at $\alpha = 0.05$ significance level if the monthly rainfalls are normally distributed.

********** **This is the Last Page** **********