GX2098

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II EXAMINATIONS 2002–2003

MODULE CODE: **MA 113, MA 228**
MODULE: STATISTICS AND PROBABILITY

External Examiner: Dr. D. Harrington
Internal Examiner: Dr. J.N. Sheahan

INSTRUCTIONS: **Answer the ten questions in PART A (30 marks) and two of the three questions B1, B2 and B3 in PART B (35 marks each).**

DURATION: **Two** hours

PART A

[Multiple choice. 30 marks] In each of questions **A1.** through **A10.** below, **write down one** choice of answer. For example, if in **A1.** below you think **A)** is the answer, you would write in your answer book **A1.A)**.

A1. In hypothesis testing, we commit a Type I error if we
A) reject the null hypothesis when it is true **B)** accept the null hypothesis when it is false.

A2. It is desired to take a sample of 100 students from a college that has 5,000 students. Suppose we divide the population into two strata, males and females, and that there are 3,000 male students and 2,000 female students in the college. **How many** of each sex should be included in the sample of 100 if we use *proportional allocation*?

A) 50 males and 50 females **B)** 60 males and 40 females **C)** 40 males and 60 females.

A3. Refer to question A2 above and suppose that the survey of 100 students is to be conducted to estimate the mean annual consumption, μ, of beer by students at the college. Suppose that the standard deviations of the number of units of beer consumed by males and females are $\sigma_1 = 100$ and $\sigma_2 = 50$, respectively. **How many** students from each stratum should be taken into the sample if *optimal allocation* is used?

A) 50 males and 50 females **B)** 60 males and 40 females **C)** 40 males and 60 females
D) 75 males and 25 females **E)** 70 males and 30 females **F)** 30 males and 70 females.

A4. Assume that the population of peoples' weights has mean $\mu = 60$ kg. and standard deviation $\sigma = 10$ kg.. Suppose that a footbridge will break down if the total weight on it exceeds 6,700 kg.. **What** is the approximate probability that it will break down if $n = 100$ people get on it all at once and each of them is carrying a case of beer that weights 5 kg.?
Note: If $Z \sim N(0,1)$, then $P(Z > 1) = 0.1587$, $P(Z > 2) = 0.0228$, $P(Z > 3) = 0.0013$.
A) 0 **B)** 0.1587 **C)** 0.0228 **D)** 0.0013 **E)** 0.05 **F)** 0.95.

A5. Suppose that the population of daily sales (in Euro) of a company has a normal distribution with mean $\mu = 1000$ and standard deviation $\sigma = 100$. Let $a = P$(sales on a random day will be less than 900) and let $b = P$(the mean sales on a random sample of 25 days will be less than 1060). **Then**

A) $a = 0.1587$ and $b = 0.9987$ **B)** $a = 0.0228$ and $b = 0.0228$ **C)** $a = 0.0228$ and $b = 0.1587$

D) $a = 0.1587$ and $b = 0.0228$ **E)** $a = 0.0013$ and $b = 0.0228$ **F)** $a = 0.0228$ and $b = 0.0013$.

Note: If $Z \sim N(0,1)$, then $P(Z > 1) = 0.1587$, $P(Z > 2) = 0.0228$, $P(Z > 3) = 0.0013$.

A6. A random sample of $n = 100$ students at NUI, Galway showed that 36 exercise regularly. **What** is an approximate 95.44% confidence interval for the proportion of all NUI, Galway students who exercise regularly? *Note:* If $Z \sim N(0,1)$, then $P(Z > 2) = 0.0228$.

A) 0.5 ± 0.096 **B)** 0.64 ± 0.096 **C)** 0.36 ± 0.096 **D)** 0.36 ± 0.048.

A7. **What** is the minimum number of Galwegians that should be sampled at random so that with probability (at least) 0.95 the sample proportion of smokers will not differ from the unknown population proportion of smokers by more than ± 0.04 ?

Note: If $Z \sim N(0,1)$, then $P(Z > 1.96) = 0.025$.

A) 100 **B)** 236 **C)** 484 **D)** 601 **E)** 1,068 **F)** 2,000.

A8. Suppose that based on a random sample of size 10 and the formula $\bar{x} \pm t_{n-1\,\alpha}\frac{s}{\sqrt{n}}$, a 95% confidence interval for the mean μ of a population was found to be $50 < \mu < 60$. We can then **say that**:

A) the probability is 0.95 that μ lies in the interval $(50, 60)$

B) of all possible samples of size 10 that could be taken from the population, 95% of the intervals that would be obtained (using the same formula as above) would contain μ

C) 95% of the means of all possible samples of size 10 that could be taken from the population would lie in the interval $(50, 60)$

D) if the appropriate t-test of the alternatives $H_0 : \mu = 53$ versus $H_1 : \mu \neq 53$ was conducted using a level of significance $\alpha = 0.05$, H_0 would be rejected

E) exactly two of statements A), B), C) and D) are true

E) exactly three of statements A), B), C) and D) are true

F) all four of statements A), B), C) and D) are true.

A9. Assume that the population of daily sales of a company has a normal distribution with unknown mean μ and standard deviation $\sigma = 40$ (ignore units). In order to test the alternatives $H_0 : \mu \leq 1000$ versus $H_1 : \mu > 1000$, a random sample of 16 daily sales will be taken and H_0 will be rejected if the sample mean sales $\bar{x}$ exceeds 1,020. **What** is the power of the test if in fact $\mu = 1,040$?

Note: If $Z \sim N(0,1)$, then $P(Z > 1) = 0.1587$, $P(Z > 2) = 0.0228$, $P(Z > 3) = 0.0013$.

A) 0.0013 **B)** 0.1587 **C)** 0.5 **D)** 0.8413 **E)** 0.9772 **F)** 0.9987.

A10. An investigation was conducted to test to see if a new drug is effective in reducing blood pressure. Fifteen subjects were randomly selected, and each had their blood pressure taken before being put on the drug, and again one month after being put on the drug. **Which** of the following methods of analysis is the most appropriate?

A) one-sided independent samples t-test **B)** one-sided paired samples t-test.

C) two-sided independent samples t-test **D)** two-sided paired samples t-test

E) chi-squared test for a single population proportion

F) chi-squared test for independence in a contingency table.

PART B

B1. Assume that the population of heights of people has (an approximately) normal distribution with unknown mean μ and standard deviation $\sigma = 3.0$. This question mainly concerns the z-test for testing the null hypothesis $H_0 : \mu = 67.0$ against the alternative hypothesis $H_1 : \mu \neq 67.0$ using $\alpha = 0.05$, and confidence intervals for μ. Some values from the standard normal tables that you should need somewhere below are:

$$z_{0.0274} = 1.92,\ z_{0.025} = 1.96,\ z_{0.0228} = 2.0 \text{ and } z_{0.0013} = 3.0.$$

Consider using the z-test along with $\alpha = 0.05$ to test the alternatives $H_0 : \mu = 67.0$, $H_1 : \mu \neq 67.0$. Note that this test rejects H_0 if $|z_{OBS}| > 1.96$ (where $z_{OBS} = \dfrac{\overline{x} - 67}{\sigma/\sqrt{n}}$),

or equivalently if $\overline{x} > 67 + 1.96\frac{\sigma}{\sqrt{n}}$ or $\overline{x} < 67 - 1.96\frac{\sigma}{\sqrt{n}}$.

Suppose that a random sample of $n = 9$ people was chosen, and their heights analysed by a statistical software package. The output is as follows:

TEST OF $H_0 : \mu = 67.0$ VERSUS $H_1 : \mu \neq 67.0$

n	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\lvert z_{OBS}\rvert$	p-value
9	70.0	1.0	3.0	0.0026

(a) *[2 marks]* Based on the printout above, **should** H_0 be rejected? **Give briefly** a reason for your answer.

(b) *[6 marks]* **Show** how the p-value 0.0026 was calculated by the software.

(c) *[6 marks]* **Calculate** the power of the test at $\mu = 67.04$.

(d) *[8 marks]* **Examine each of statements (i), (ii), (iii) and (iv) below separately**. Then **write** in your answer book whether it is True or False.

(i) If H_0 is true, then the value of $P(\overline{X} > 67 + 1.96\frac{\sigma}{\sqrt{n}}$ or $\overline{X} < 67 - 1.96\frac{\sigma}{\sqrt{n}})$ is 0.05.

(ii) The power of the test at $\mu = 90.0$ is larger than the power at $\mu = 66.0$.

(iii) The power of the test at $\mu = 68.0$ is larger than the power at $\mu = 66.0$.

(iv) In calculating the power of the test at $\mu = 68.0$, it is necessary to know the observed value, 70, of $\overline{X}$.

(e) *[4+4 =8 marks]* **Calculate** the 95% confidence interval $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ for μ, and **interpret** it carefully.

(f) *[5 marks]* If a $100(1-\alpha)\%$ confidence interval $\overline{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ for μ turned out to be $(67.0, 73.0)$, **what** was the confidence coefficient $1-\alpha$ used?

B2.

(a) Imagine that you are a shipping magnate who wishes to purchase one of two brands of paint to apply to all your ships. Naturally, you will choose the paint that leads to less rust on average. Let μ_1 and μ_2 respectively denote the population mean amount of rust per square metre of ships' surface with brand 1 and brand 2. Suppose the alternatives to be tested using $\alpha = 0.05$ are $H_0 : \mu_1 = \mu_2$, and $H_1 : \mu_1 > \mu_2$.

The design used on your behalf involves applying brand 1 paint to 3 randomly chosen ships in Galway Harbour and brand 2 paint to 3 ships in Dublin Harbour. Data are then collected after a year by taking rust measurements (from square metres of the ships' surfaces.) The data are as follows:

	Amount of rust		
Brand 1 paint	49	50	51
Brand 2 paint	46	47	48

(i) *[10 marks]* **Perform** in detail the independent samples t-test of the alternatives above, using $\alpha = 0.025$. To save you time, note that the two samples have the same variance!

Note: One of the following critical points is relevant: $t_{4,\,0.025} = 2.776$, $t_{6,\,0.025} = 2.447$.

(ii) *[4 marks]* A colleague claimed that you could get more information about the comparison of the two types of paint if, instead of the design used above, you applied paint 1 to the port side of three ships in Galway Harbour and paint 2 to the starboard side of the *same* three ships. Briefly but clearly **explain** why your colleague is correct. You must answer this question in a mature way, both in the context of describing the factor that the paired samples design suggested by your colleague would block out, and the statistical sense in which the latter design is 'better' for a given level of significance.

(b) *[9 marks]* We wish to test the null hypothesis H_0: the probabilities of 0, 1, 2, 3, and 4 or more, accidents at a certain crossroads on a random day are each equal to $\frac{1}{5}$. A random sample of 100 days showed the following frequencies of number of accidents.

Number of accidents	0	1	2	3	4 or more
Number of days	11	26	28	11	24

Suppose that a chi-squared goodness-of-fit test with $\alpha = 0.01$ is used to test to see if H_0 should be rejected. Accept that the observed value of the chi-squared goodness-of-fit statistic is 13.90. **Examine each of statements (i), (ii) and (iii) below separately** and then **state** in your answer book if it is True or False.

Note: $\chi^2_{5,\,0.01} = 15.086$, $\chi^2_{5,\,0.005} = 16.750$, $\chi^2_{4,\,0.01} = 13.277$, $\chi^2_{4,\,0.005} = 14.860$.

(i) If H_0 is true, the expected number of days on which no accident will occur is 20.

(ii) The p-value of the test is less than 0.005.

(iii) The critical value for the test is 13.277 and H_0 should be rejected.

(Question B2 is continued on next page)

(Question B2 continued from previous page)

(c) *[12 marks]* We wish to study the relationship between the weekly number of units of alcohol drunk by students at NUI, Galway and the Faculty of study. Each individual in a random sample of 300 students was asked their faculty of study and whether they drank less than 20 units, between 20 and 50 units, or more than 50 units of alcohol in the previous week. The results are given in the following frequency cross-tabulation.

		Faculty of Study							
		Arts	Cel. Studies	Comm.	Eng.	Law	Med.	Science	TOTAL
No. Units Consumed	<20	**10**	**6**	**12**	**9**	**10**	**10**	**23**	80
	20 to 40	**20**	**4**	**35**	**11**	**10**	**3**	**16**	99
	>40	**30**	**5**	**20**	**40**	**3**	**3**	**20**	121
	TOTAL	60	15	67	60	23	16	59	300

Suppose that a chi-squared test of independence with $\alpha = 0.05$ will be used to test the alternatives H_0 : Faculty of study and amount drunk by students are independent variables, H_1 : a relationship exists between faculty of study and alcohol consumption.

Accept that the observed value of the chi-squared contingency table test statistic is 56.129. **Examine each of statements (i), (ii), (iii) and (iv) below separately** and then **state** in your answer book if it is True or False. *Note:* $\chi^2_{12,\,0.05} = 21.026$, $\chi^2_{21,\,0.05} = 32.671$, $\chi^2_{28,\,0.05} = 41.337$.

(i) If H_0 is true, the estimated expected number of students who study Arts and drink less than 20 units in a random week is 16.

(ii) The estimated conditional probability that a student will drink more than 40 units *given that* he/she is an Engineering student is $\frac{2}{3}$.

(iii) The critical point of the test is 21.026.

(iv) The p-value of the test is < 0.05 and we should reject H_0.

B3.

(a) *[15 marks]* Concerning the linear regression model $\mu_{Y|x} = \alpha + \beta x$ relating a response random variable Y to a non-stochastic input variable x, suppose that we have available n data points (x_i, y_i), $i = 1, 2, \ldots, n$ in order to find the least squares regression line $\hat{y} = a + bx$. Recall that this least squares line (or least squares prediction equation or best-fitting line) is the line whose slope b and intercept a give the smallest value of

$$g(\alpha, \beta) := \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2.$$

Prove that the values b and a of β and α, respectively, that minimize $g(\alpha, \beta)$ are given by

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \text{ and } a = \bar{y} - b\bar{x}.$$

(b) We wish to study the relationship between x = performance (in terms of overall percentage mark) in final year at university and Y = the starting salary (in thousands of Euro) after graduation of NUI, Galway students. Assume that a model of the form $\mu_{Y|x} = \alpha + \beta x$, together with the usual assumptions, relates these two variables. Data from $n = 5$ graduates surveyed are as follows.

Performance (x_i)	60	75	70	65	80
Starting salary (y_i)	16	18	20	22	24

Note to save you time:

$$\sum x_i = 350,\ \sum y_i = 100,\ \sum x_i y_i = 7{,}060,\ \sum x_i^2 = 24{,}750,\ \sum y_i^2 = 2{,}040,$$
$$s_e = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2} = 2.921.$$

(i) *[6 marks]* **Show** that the least squares regression line is $\hat{y} = 3.2 + 0.24x$.

Hint: See part (a) above for formulae you'll require.

(ii) *[3 marks]* **State** whether the sample correlation coefficient r is positive or if it is negative. (You should not need to perform any calculations to answer this question.)

(iii) *[5 marks]* **State** (without providing details) which one of (A), (B), (C), (D), (E) and (F) below represents a 95% prediction interval for the starting salary of a graduate whose performance was 70. *Note:* A 95% prediction interval for the starting salary of a graduate whose performance is x_0 is given by

$$a + bx_0 \pm t_{n-2,\frac{\alpha}{2}} \times s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}}.$$

One of the following t-values is relevant:

$$t_{3,\,0.05} = 2.353,\ t_{3,\,0.025} = 3.182,\ t_{4,\,0.05} = 2.132,\ t_{4,\,0.025} = 2.776.$$

(A) $20 \pm 2.353 \times 2.921 \times \sqrt{\frac{6}{5}}$ **(B)** $20 \pm 3.182 \times 2.921 \times \sqrt{\frac{6}{5}}$

(C) $20 \pm 2.132 \times 2.921 \times \sqrt{\frac{6}{5}}$ **(D)** $3.2 \pm 2.776 \times 2.921 \times \sqrt{\frac{6}{5}}$

(E) $70 \pm 3.182 \times 2.921 \times \sqrt{\frac{1}{5}}$ **(F)** $20 \pm 3.182 \times 2.921 \times \sqrt{\frac{1}{5}}$.

(iv) *[3 marks]* **Write down** a point estimate of $\mu_{Y|75}$, the population mean starting salary of graduates who obtain 75% in their final year at university.

(v) *[3 marks]* If a 95% confidence interval for $\mu_{Y|75}$ was calculated, **would it** be wider or narrower than the 95% prediction interval discussed in (iii) above? (No explanation for your answer is needed.)

J.N.S.