GX2024

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II EXAMINATIONS 2003–2004

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

External Examiner: Dr. T.C. Bailey
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. A lecturer is performing a statistical test concerning the effectiveness of a new teaching technique. If he committed a Type I error by erroneously concluding that the technique is effective, **what** were the alternatives he was testing?
(a) H_0 : the technique is not effective, H_1 : the technique is effective,
(b) H_0 : the technique is effective, H_1 : the technique is not effective.

A2. From a college that has 7,000 male students and 3,000 female students, a stratified random sample of 110 students will be taken in order to estimate the mean monthly expenditure, μ (in €), on alcohol by students at the college. Suppose that the standard deviations of expenditure on alcohol 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) 100 males and 10 females **B)** 60 males and 50 females **C)** 10 males and 100 females.
D) 55 males and 55 females **E)** 50 males and 60 females **F)** 73 males and 37 females.

A3. The weights of apples in an orchard are normally distributed with mean $\mu = 120$ grams and standard deviation $\sigma = 9$ grams. Let X be the weight of a randomly selected apple and let $\overline{X}$ be the mean weight for a random sample of 81 apples. **Which** one of the following statements is true?
A) $P(|X - 120| < 9) = 0.5$ **B)** $P(|X - 120| < 9) = P(|\overline{X} - 120| < 1)$
C) $P(X > 120) = P(\overline{X} > 120)$ **D)** exactly two of statements A), B) and C) are true
E) all three of statements A), B) and C) are true.

A4. Student A will take a random sample of size 900 from an infinite population. Student B will take a random sample of size n from the same population. **How large** should n be if it is desired that the standard deviation of the mean $\overline{X}_A$ of A's sample be three times smaller than the standard deviation of the mean $\overline{X}_B$ of B's sample?

A) 16 **B)** 30 **C)** 50 **D)** 100 **E)** 150 **F)** 400 .

A5. Suppose that a population of marks of students has a normal distribution with mean $\mu = 60$ and standard deviation $\sigma = 10$. Let $a = P$(mark of a random student falls between 40 and 60) and let $b = P$(the mean mark for a random sample of 100 students will be between 58 and 62). **Then**

A) $a = 0.9544$ and $b = 0.9544$ **B)** $a = 0.6826$ and $b = 0.9544$

C) $a = 0.9544$ and $b = 0.6826$ **D)** $a = 0.6826$ and $b = 0.6826$

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. Assume that the standard deviation of the amount of copper precipitate from a chemical experiment is 4 grams. Approximately **how many** times should the experiment be conducted if one wants to be 99% confident that the sample mean amount of precipitate will be within ± 1.288 grams of the unknown population mean amount of precipitate?

A) 30 **B)** 40 **C)** 50 **D)** 64 **E)** 100 **F)** 1068 .

Note: If $Z \sim N(0,1)$, then $P(Z > 1.96) = 0.025$, $P(Z > 2.576) = 0.005$.

A7. A researcher conducted a large sample two-sided test of the null hypothesis that $\mu = 100$.

She reports a p-value of 0.034. **Which** one of the following is correct?

A) The null hypothesis is not rejected at $\alpha = 0.05$

B) The 95% confidence interval for μ would contain 100

C) The null hypothesis is not rejected at $\alpha = 0.01$

D) The 99% confidence interval for μ would contain 100 .

A8. Consider the following list of hypothesis tests: (i) "z-test for a single population mean", (ii) "z-test for a single population proportion", (iii) "t-test for the difference between two population means when the two samples taken are independent and random", (iv) "t-test for the difference between two population means when the random samples are paired", (v) "χ^2 test for a population variance", (vi) "χ^2 goodness-of-fit test". **Which** of these tests can be used to test if the proportion of smokers in Ireland differs from 0.3, when a random sample of 500 Irish people is taken?

A) Only test (iii) can be used **B)** only test (v) can be used only

C) only test (ii) can be used **D)** only test (vi) can be used

E) we have a choice between tests (iii) and (v)

F) we have a choice between tests (ii) and (vi).

A9. A random sample of 20 people was taken and each of them was weighed before and after being placed on a diet. **Which** of the tests mentioned in Q. A8. above can be used to test if the diet is effective in reducing the population mean weight of people? (Assume that the population of weight differences is normal).

A) Only test (ii) can be used
B) only test (iii) can be used only
C) only test (iv) can be used
D) only test (v) can be used
E) we have a choice between tests (iii) and (iv)
F) we have a choice between test (iii) and test (v).

A10. Suppose that based on a random sample of size 20 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 $4 < \mu < 7$. We can then **say that**:

A) of all possible samples of size 20 that could be taken from the population, 95% of the intervals that would be obtained (using the same formula as above) would contain μ
B) the probability is 0.95 that μ lies in the interval (4, 7)
C) 95% of the means of all possible samples of size 20 that could be taken from the population would lie in the interval (4, 7)
D) if the appropriate t-test of the alternatives $H_0 : \mu = 8$ versus $H_1 : \mu \neq 8$ 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
F) exactly three of statements A), B), C) and D) are true
G) all four of statements A), B), C) and D) are true.

PART B

B1. Assume that the population of weights of people (in kgs.) has an approximately normal distribution with unknown mean μ and standard deviation $\sigma = 10.0$. This question mainly concerns the z-test for testing the null hypothesis $H_0 : \mu = 60.0$ against the alternative hypothesis $H_1 : \mu < 60.0$ using $\alpha = 0.05$, and a confidence interval for μ. Note that in your work below, you will need most of the following values from the standard normal tables.

$z_{0.05} = 1.645$ [i.e. $P(Z > 1.645) = 0.05$], $z_{0.025} = 1.96$, $z_{0.0228} = 2.0$ and $z_{0.0013} = 3.0$.

Consider using a z-test along with $\alpha = 0.05$ to test the alternatives

$$H_0 : \mu = 60.0,\ H_1 : \mu < 60.0\,.$$

Suppose that a random sample of $n = 25$ people was chosen, and their weights analyzed by MINITAB. Part of the output is as follows.

Test of $H_0 : \mu = 60.0$ versus $H_1 : \mu < 60.0$.

n	Mean $\bar{x}$	z-value	p-value
25	58.0	−1.0	0.1587

Helpful note:

The test rejects H_0 if $z_{OBS} < -z_\alpha$ (where $z_{OBS} = \frac{\bar{x} - 60}{\sigma/\sqrt{n}}$) or equivalently if $\bar{x} < 60 - z_\alpha\frac{\sigma}{\sqrt{n}}$.

That is, the critical region of the test is $\left\{\bar{x} : \bar{x} < 60 - z_\alpha\frac{\sigma}{\sqrt{n}}\right\}$.

(Question B1 is continued on next page)

(Question B1 continued from previous page)

(a) *[3 marks]* Based on the printout above, **should** H_0 be rejected? Give briefly a reason for your answer (based on either $\bar{x}$, the z-value (i.e. z_{OBS}), or the p-value).

(b) *[2 marks]* **Show** how the z-value was calculated from some of the other values that appear above.

(c) *[10 marks]* **Calculate** the power of the above z-test when $\mu = 54.71$.

(d) *[10 marks]* Carefully **derive** the formula $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ for a 95% confidence interval for the mean μ of any normally distributed population that has known variance, and **calculate** this confidence interval for the output above.

(e) *[10 marks]* Now ignore the computer output above. **Find** the appropriate critical region so that the test would have a significance level of $\alpha = 0.0013$ based on a random sample of $n = 100$.

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	61	60	59
Brand 2 paint	48	47	46

(i) *[10 marks]* **Perform** in detail the independent samples t-test of the alternatives above, using $\alpha = 0.005$.

Note: The two samples have identical variance and one of the following critical points is relevant: $t_{4,\,0.005} = 4.604$, $t_{6,\,0.005} = 3.707$.

(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 200 days showed the following frequencies of number of accidents.

(Question B2 is continued on next page)

(Question B2 continued from previous page)

Number of accidents	0	1	2	3	4 or more
Number of days	40	50	40	32	38

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 4.2. **Examine each of statements (i), (ii) and (iii) below separately** and then **state** in your answer book if it is True or False.

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

(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 14.860 and H_0 should be rejected.

(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	**12**	**4**	**13**	**8**	**8**	**9**	**26**	80
	20 to 40	**30**	**4**	**35**	**11**	**10**	**3**	**16**	109
	> 40	**20**	**5**	**20**	**40**	**3**	**3**	**20**	111
	TOTAL	62	13	68	59	21	15	62	300

Suppose that a chi-squared test of independence with $\alpha = 0.01$ 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 56.01.

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: $\chi^2_{12,\,0.01} = 26.22$, $\chi^2_{21,\,0.01} = 38.93$, $\chi^2_{28,\,0.01} = 48.28$.

(i) If H_0 is true, the estimated expected number of students in the sample who study Medicine (Med.) and drink less than 20 units in a random week is 4.

(ii) The estimated conditional probability that a student will drink more than 40 units *given that* he/she is a Medical student is $\frac{1}{5}$.

(iii) The critical point of the test is 26.22.

(iv) The p-value of the test is < 0.01 and we should reject H_0.

B3.

(a) *[12 marks]* Consider the linear regression model $\mu_{Y|x} = \alpha + \beta x$ relating a response random variable Y to a non-stochastic input variable x, and suppose that we have available n data points (x_i, y_i), $i = 1, 2, \ldots, n$ measured on these two variables. Recall that the least squares regression line $\hat{y} = a + bx$ (or least squares prediction equation or best-fitting line) is the line whose slope b and intercept a correspond, respectively, to values of β and α which 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

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

(b) As financial analyst for a major firm, you have been asked to study the relationship between x = monthly average inflation rate and Y = monthly return on a particular stock in the New York Stock Exchange (ignore units). Assume that a model of the form $\mu_{Y|x} = \alpha + \beta x$, together with the usual assumptions, relates these two variables. Data from $n = 6$ years are as follows.

Inflation (x_i)	4.2	4.8	4.9	5.0	5.3	5.8
Return (y_i)	6	8	10	9	12	15

Note: $\sum x_i = 30$, $\sum y_i = 60$, $\sum x_i y_i = 308.2$, and $\sum x_i^2 = 151.42$.

(i) *[3 marks]* **Plot** a scatter diagram (scatter plot) to represent the data in the table above. (You need not use graph paper.)

(ii) *[5 marks]* **Show** that the least squares regression line $\hat{y} = a + bx$ is

$$\hat{y} = -18.85 + 5.77x.$$

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

(iii) *[5 marks]* If r denotes the sample correlation coefficient between the two variables, then by describing briefly (but clearly) in words what each of r and b represents, **explain** why it is true that $r = 0$ if and only if $b = 0$.

(iv) *[5 marks]* Based on your answer in (ii) above, **write down** a point estimate of $\mu_{Y|5}$, the population mean return when inflation is 5% and **briefly** explain why your answer does not equal the observed return 9 in the table above.

(v) *[5 marks]* Briefly **explain** why there is no need to carry out a test of the alternatives $H_0 : \beta = 5.77$, $H_1 : \beta \neq 5.77$ at level of significance 0.05, i.e. why it is obvious that we would not reject H_0. (A term like "statistical significant" should appear somewhere in your answer.)