GX2214

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SUMMER EXAMINATIONS 2000

M. SC. IN BIOTECHNOLOGY AND H. DIP. IN APPLIED SCIENCE AND MICROBIOLOGY.

STATISTICS

Professor C.M. Browne
Professor Emer Colleran
Cillian O'Murchú, B.Sc., M.Sc.

Time allowed: *Three Hours*
Answer any **five** questions
All questions (but not necessarily parts therein) carry equal weight.

Q1. (a) **Define** the branch of science STATISTICS.

(b) **Give** an example of a data set whose mean is 1000, the mode is 0 and whose median is 500.

(c) 150 dead fish are recovered from the stream following a pollution incident, and are measured to the nearest whole millimetre. Measurements obtained are expressed in the form of a frequency table:

Length of Fish (mm)	Frequency
100 – 109	7
110 – 119	16
120 – 129	19
130 – 139	31
140 – 149	41
150 – 159	23
160 – 169	10
170 – 179	3

Calculate the mean and standard deviation for the above grouped data.

(d) **Briefly explain** the difference between MEAN and MEDIAN. Give an example to show the difference between them.

Q2. (a) **Define** (with examples) each of the following concepts:

i. Sample Space
ii. Discrete Random Variable.

(b) In each draw of The National Lottery, six numbers are chosen at random from the forty-two integers 1, 2, ..., 42. Suppose you have your ticket bought for the next draw. (Your ticket contains six numbers from the above 42 integers – ignore the fact that one must actually buy a minimum of 'two lines of six numbers'.)
What is the probability that you will match three and the bonus number? That is, **what** is the chance that your ticket will contain three of the six numbers and the bonus number drawn by The National Lottery?

(c) One in five people die of cancer. If a sample of 12 recently deceased people was taken.
(i) **Find** the probability that exactly 2 of the deaths were due to cancer. [i.e. **find** $P(X=2)$ where $X=$ the number of deaths due to cancer in the random sample of $n=12$ people].
(ii) **Find** the probability that at most 2 of the deaths were due to cancer.

(d) In a sample of n recently deceased people suppose that the probability that none of them will have died of cancer is (approximately) 0.2. Calculate *n*.

Q3. A shellfish farmer is cultivating oysters under different conditions in two different sites, X and Y. a sample of ten oysters is taken from each site and weighed. The results are as follows.

	X	Y
Mean	69.9	67.00
Standard Deviation	4.5	4.8

(a) (i) **Test** the hypothesis that the mean weights of two year old oysters from each site are equal. i.e. **Perform** in detail the independent samples t-test of the alternatives above, using $\alpha = 0.025$. *Note:* One of the following critical points is relevant: $t_{17,.025} = 2.1098$, $t_{18,.025} = 2.1009$.

(ii) What two assumptions are we making here for the above test to be valid?

(b) (i) In the context of testing the alternative hypotheses above, **explain when and why** one sometimes prefers to employ a non-parametric test [rather than a parametric one].
(ii) **Name** any non-parametric test you know for testing to see if the population median weight of oysters from site X is equal to the corresponding median for site Y , for the data given above.

Q4. Throughout this question, assume that the population of heights of people has (an approximately) normal distribution with unknown mean μ and standard deviation $\sigma = 3$. This question mainly concerns the z-test for testing the alternatives H_0: $\mu = 68.0$ versus $H_1: \mu \neq 68.0$ using $\alpha = 0.05$, and confidence intervals for μ. Three z-values that you should need somewhere below are:

$$z_{.025} = 1.96,\ z_{.0228} = 2.0 \text{ and } z_{.05} = 1.645.$$

Consider using the z-test along with $\alpha = 0.05$ to test the alternatives H_0: $\mu = 68.0$ versus $H_1: \mu \neq 68.0$. Suppose that a random sample of $n = 9$ people was chosen, and their heights analysed by MINITAB. The output is as follows:

TEST OF H_0: $\mu = 68.0$ versus $H_1: \mu \neq 68.0$

n	mean	stdev	se mean	z-value	p-value
9	70.0	3.30	1.0	2.0	0.0456

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

(ii) **Show** how the z-value was calculated (from some of the entries that precede it in the above table).

(iii) **Define** the p-value of any statistical test, and for the output above, **show** how the p-value can be calculated from the z-value.

(iv) **Define** the power of the z-test of the above alternatives at a value of μ different from 68.

(v) **Calculate** a 95% confidence interval for μ and interpret it carefully.

Q5. (a) The following table shows the frequency distribution of the numbers of nematodes in all 60 squares of a counting chamber

Number of nematodes	0	1	2	3	4 or more
Frequency	5	10	15	30	0

(i) Let μ = the expected number of nematodes on a random square. **Show** that an estimate of μ is 2.16 .

(ii) Let X = the number of nematodes on a random square. Use $\alpha = .025$ and a goodness-of-fit procedure to **test** to see if a Poisson distribution is appropriate for X.
Note: One of the following critical points is relevant:

$$\chi^2_{2,.05} = 7.378,\ \chi^2_{3,.025} = 9.348,\ \chi^2_{4,.025} = 11.143,\ \chi^2_{5,.025} = 12.832\ .$$

(iii) **Explain**, without performing details, what adjustments you would make in the solution to (ii) if you were asked to test to see if X has a Poisson distribution with $\mu = 3$. (Note that now you are fitting a *particular* --rather than *some* -- Poisson distribution.)

Q5 continued on the next page

Q5. (b) A ornithologist believes that the 4 types of birds Rooks, Ravens, Magpies and Jays occur in the population in the ratio 5:3:5:1.Selecting n =280 birds the ornithologist classifies the birds as follows:

Birds	Rooks	Ravens	Magpies	Jays
Frequencies	92	58	105	25

Carry out a statistical test ($\alpha = 0.025$) to determine whether or not the data are consistent with the ornithologist's belief in the ratios of Rooks, Ravens, magpies, Jays.
Note: One of the following critical points is relevant:

$$\chi^2_{2,.05} = 7.378,\ \chi^2_{3,.025} = 9.348,\ \chi^2_{4,.025} = 11.143,\ \chi^2_{5,.025} = 12.832\ .$$

Q6. A researcher is trying to study the relationship between the average blood lead level measured in young children (Y) and the amount of lead used in the production of petrol (X) over a 10 year period. Her data is summarised below:

Petrol(x)	18	29	49	50	65	66	67	71	73	77
Blood (y)	9	11	13	15	16	16	17	19	20	21

[NOTE: $\sum x_i y_i = 9709$, $\sum x_i^2 = 35435$, $\sum y_i^2 = 2879$, $\sum x_i = 565$, $\sum y_i = 167$]

Assume that a population regression model of the form $\mu_Y/_x = \alpha + \beta x$, together with the usual assumptions, relates these two variables.

(a) Plot a scatter diagram to represent these data.

(b) Compute the least squares regression line of y on x to fit these data.

(c) Based on your answer in (b), write down a point estimate of $\mu_Y/_{x=55}$, (the population mean blood lead level, when the petrol lead level is 55), and calculate a 95% confidence interval for $\mu_Y/_{x=55}$.

(d) Calculate the sample correlation coefficient of x and y, and interpret its meaning.

(e) If the blood lead level for the eighth point above were 12 instead of 19, would you expect the correlation coefficient for the revised data to be closer to 1.0 than that obtained in (d) above? A simple **yes** or **no** will suffice as an answer.