Gx 1996

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SPRING EXAMINATIONS, 2003 — 2004

UNIVERSITY SCIENCE EXAMINATIONS

STATISTICS [MA419, MA851, BC852]

Exam Codes: 10A1, 1AM1, 3EV1,3EV2, 1AS1, 1CB1, 1MT1

Prof. C.M Brown,
Prof. Emer Colleran,
Paul Wilson, M.A., H.D.E.

Time allowed: **Three** Hours.
Answer any *four* questions.
All questions, but not necessarily parts therein, carry equal marks.
Z and t–Distribution Tables Are Attached.
A χ^2–Distribution table is available on page 37 of the Mathematical Tables.
Mathematical Tables and graph paper are available on request.
A formula sheet is given on pages 7 and 8

Question 1 is on the next page

1. (a) i. What is meant by a *random sample*

ii. A researcher wishes to estimate the quantities of various species of beetle in a given wood. He randomly selects pit–fall traps at various locations around the wood, sets these traps, and later collects and categorises the various species of beetle collected.
What is wrong with this approach?

(b) The widths of the orbital cavities of $n = 90$ owls were grouped into the frequency distribution shown below.

Table 1:

Width of Cavity	*Frequency*
3 – 8	6
8 – 13	16
13 – 18	33
18 – 23	16
23 – 28	14
28 – 38	5
80 – 100	8

i. Estimate the mean width of the orbital cavity of the above owls.

ii. Draw a histogram to illustrate the frequency distribution shown in Table 1. (It is not required that you use graph paper, but you may do so if you wish.)

(c) Briefly explain what is meant by *nominal*, *ordinal* and *continuous data*, giving an example of each. Which average is appropriate for each type of data? Explain why it is often unwise to use the mean as a measure of central tendency for ordinal data and why it is ridiculous to use the mean as a measure of central tendency for nominal data.

Question 2 is on the next page

2. (a) i. What is meant by saying that two events A and B are *independent*?

ii. The occurrence in a given region of farmland of Cow Parsley (*anthriscus sylvestris*) and the closely related Hogweed (*heracleum spondylium*)is known to be independent. If the probability of occurrence in an area of $100m^2$ of Cow Parsley is 0.6 and the probability of occurrence in an area of $100m^2$ of hogweed is 0.4, what is the probability that if a given $100m^2$ area of land is selected at random that

A. both cow parsley and hogweed will be found,

B. neither cow parsley nor hogweed will be found?

(b) Historical records of rainstorms in a town indicate that on the average there are three rainstorms per year. Assuming that the occurrence of rainstorms follows a Poisson distribution what is the probability that:

i. there are no rainstorms next year,

ii. there are exactly three rainstorms during the next two years,

iii. there are at least two rainstorms next year,

iv. there are at least two rainstorms next year, but none the year after?

(c) The probability that a given species of farmed salmon is infected by a virus is 0.3. If ten of these salmon are selected at random, what is the probability that:

i. *exactly* three of them are infected with the virus,

ii. none of them are infected by the virus

iii. at least one of them is infected with the virus.

(d) i. Suppose that the annual production of milk, per cow, has a normal distribution with mean $\mu = 1,200$ gallons and standard deviation of $\sigma = 250$gallons. If a cow is selected at random, what is the probability that it produces:

A. between 900 and 1,500 gallons,

B. less than 1,000 gallons,

C. between 1400 and 1450 gallons per annum?

ii. If 75% of cows produce less than x gallons of milk per annum, calculate x.

Question 3 is on the next page

3. (a) Explain briefly what is meant by: "The null hypothesis was rejected at a significance level of $\alpha = 0.05$".

(b) A factory manager believes that the mean weight of "one kilogram gram bags of sugar" being produced in fact differs from $1,000$ grams. He takes a random sample of 100 such bags, and finds that the bags in the sample have mean weight of $\mu = 1,002$ grammes with a standard deviation of 4 grams.

i. State a suitable null (H_0) and alternative (H_1) hypothesis.

ii. May the null hypothesis be rejected at a significance level of $\alpha = 0.05$.

iii. If a *type one error* was made above, what, in terms of this specific case, would this mean.

(c) i. State two conditions that should hold for an independent samples t-test to be valid.

ii. A shellfish farmer is cultivating mussels under different conditions in two different sites, X and Y. A random sample of 10 mussels is taken from site X and weighed. A random sample of 8 mussels is taken from site Y and weighed. We may assume that these samples are independent. The results are as follows:

	X	Y
Mean	38.4	31.00
Standard Deviation	2.5	4.8

A. Test the null hypothesis that the population mean weight of mussels from site X is greater than those from site Y.

B. Construct a 95% confidence interval for the mean weight of mussels from site X.

Question 4 is on the next page

4. (a) A biologist observed 250 specimens of a culture and obtained the following distribution for the counts of the bacteria.

Number of Bacteria	0	1	2	3	4
Occurrence	48	90	68	31	13

i. May we reject H_0 : Distribution is in the ratio $20:35:25:15:5$ at $\alpha = 0.05$?

ii. What problem would occur if the biologist had only sampled 100 specimens? birds?

(b) A sample of 240 fruit flies obtain from a cross was classified according to their wing and eye type as follows:

	Eye Type			
Wings	*Normal*	*Red rim*	*Blue rim*	**Total**
Normal	108	30	30	**168**
Curly	63	39	41	**143**
Total	**171**	**69**	**71**	**311**

Is there evidence, at $\alpha = 0.05$, that wing type and eye type are related.

(c) In relation to χ^2 test, what is meant by the term "expected count"? (Note that you are to explain what the term means, not how it is calculated.)

Question 5 is on the next page

5. (a) What is meant by saying that two variables, X and Y are *strongly* correlated. If two variables X and Y are strongly correlated, does this necessarily mean that one is causing the other. Illustrate your answer with an example.

(b) Draw scatter diagrams that illustrate:

i. Strong, positive, linear correlation,

ii. weak positive, linear correlation.

(c) What is meant by the term *least squares regression line*?

(d) The following figures refer to the temperature (in degrees Celsius) and the amount (in grammes) of a chemical substance extracted from %500 grammes of a mineral soil.

Temperature (x_i)	210	250	270	290	310	340	360
Amount (y_i)	2.1	5.8	8.1	10.8	12.8	14.3	16.0

i. Plot a scatter diagram (scatter plot) to represent these data. (There is no need to use graph paper, but you may do so if you wish.)

ii. Compute the least squares regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$.

iii. Calculate an estimate, r, for ρ (the population linear correlation between x and y).

iv. Based on your answer in (b), write down a point estimate of $\mu_{Y|210}$, the population mean amount of chemical extracted when the temperature of the soil is 210°.

v. Why does $\mu_{Y|210}$ differ from 2.1, the value of Y corresponding to a temperature of 210° given in the table above?

Formulæ

1. **Mean**

   $\bar{x} = \frac{\sum x_i}{n}$

2. **(Sample) Variance**

   $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$

3. **Binomial Distribution**

   $$P(X = r) = \binom{n}{r} p^r q^{(n-r)}$$

   $\mu = np \qquad \sigma = \sqrt{npq}$

4. **Poisson Distribution**

   $P(X = r) = \frac{\mu^r e^{-\mu}}{r!}$

   $\mu = \mu \qquad \sigma = \mu$

5. **One Sample Z–test**

   $z_{\text{obs}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

6. **Two Sample Z–test**

   $z_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

7. **One Sample t–test**

   $t_{\text{obs}} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$

8. **Independent Samples t–test**

   $t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

   where:

   $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

9. **χ^2 Goodness of fit test, Independent Samples test**

Continued on the next page

10. **Regression and Correlation**

$$S_{xx} = \sum_{i=1}^{i=n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{i=n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{i=n} x_i y_i - n\bar{x}\bar{y}$$

(a) **(Sample) Correlation coefficient**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

(b) **Equation of least squares regression line**

$y = \widehat{\alpha} + \widehat{\beta}x$

Where: $\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$, $\quad \widehat{\beta} = \frac{S_{xy}}{S_{xx}}$