OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER TWO EXAMINATIONS, 1998 — 1999

SECOND UNIVERSITY EXAMINATION

STATISTICS [MA238]

Dr. M.G. Kenward,
Dr. J.N. Sheahan,
P.J. Wilson, M.A., H.D.E.

Time allowed: **Two** Hours.
Answer any *five* questions.
All questions, but not necessarily parts therein, carry equal marks.
All relevant tables, and a sheet of formulæ, are supplied

1. (a) A researcher found 240 smokers in a random sample of 500 students. Construct an approximate 95% confidence interval for the population proportion of students that smoke.

   (b) When calculating the required sample size, n, necessary to estimate a population proportion p, to within d with probability (approximately) 0.98, when we already have a good estimate $\hat{p}$ of p, we solve the equation

$$2.33\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = d$$

   Explain why, if we have no idea of the true population proportion, we solve the equation

$$2.33\sqrt{\frac{1}{4n}} = d$$

   to calculate the required sample size necessary to estimate the population proportion to within the same degree of accuracy, d, and with at least 98% confidence.

   (c) If the researcher in part (a) of this question had no knowledge of the actual proportion of the population of students that smoke, what sample size would have been necessary to ensure that the sample proportion will be within 2% of the actual population proportion with probability (at least) 0.98?

2. (a) A manafacturer of pins claims that their large boxes contain more than 2,000 pins on average. It is known that the number of pins in a random box is normally distributed, with unknown mean μ and standard deviation $\sigma = 20$. In order to test the null hypothesis $H_0 : \mu \leq 2,000$ against the alternative hypothesis $H_A : \mu > 2,000$, it is decided to take a random sample of $n = 100$ large boxes, and to reject H_0 if $\bar{x} > 2,005$, where $\bar{x}$ is the sample mean number of pins per box.

   i. For this decision rule calculate:

      A. $\alpha =$ Prob[Type One error], when $\mu = 2,000$;

      B. $\beta =$ Prob[Type Two error], if in fact $\mu = 2,007$.

   ii. If it is wished to have a significance level of $\alpha = 0.02$, but to keep the sample size the same, what should the decision rule be?

   (b) Assume that in part (a) above H_0 was rejected. Which of the following statements are *possibly* true? (Do *not* elaborate. Simply state whether each of the statements (i) - (vi) below is *possibly* true.)

   i. The null hypothesis is false.

   ii. The null hypothesis is true.

   iii. The mean amount of pins in a box is at most 2,000.

   iv. The mean amount of pins in a box is greater than 2,000.

   v. A Type One error has been committed.

   vi. A Type Two error has been committed.

3. (a) It is believed that the daily number of accidents in a certain town is Poisson distributed. For each of 120 days, the number of road accidents that occur is recorded. It is decided to use a χ^2 goodness of fit test, with $\alpha = 0.01$, to investigate if indeed the number of accidents is Poisson distributed. The table of observed and expected frequencies is as follows:

Number of Accidents	0	1	2	3	≥ 4
Observed	27	39	33	15	6
Expected	27.2	40.4	29.9	14.8	7.7

   Note: The mean daily number of accidents in the above sample is 1.483.

   i. Show how the figure of 40.4 for the expected number of days on which exactly one accident occurs is calculated.

   ii. May we reject, at $\alpha = 0.05$, H_0 : *daily number of accidents in this town is Poisson distributed*?

   (b) 200 adults were classified according to sex and their view on a proposed constitutional amendment. The results are summarized in the table below.

	For	*Against*	*Undecided*
Male	46	43	20
Female	53	28	10

   Is there evidence, at $\alpha = 0.05$, of a relationship between the sex of a person and his or her view on the proposed amendment?

4. (a) The records of a university history department show that two of its essay correctors, A and B, are returning the same average mark of 60% for essays. It is known that the exam marks are normally distributed. The department is asked to check if either corrector's marks are more variable than the other. A random sample of 11 of A's papers and a random sample of 25 of B's papers are selected. It is found that the standard deviation of A's marks is 16%, whereas that of B's is 14%. Use the F-test to determine if:

   i. We may reject $H_0 : \sigma_A \leq \sigma_B$ at $\alpha = 0.05$?

   ii. We may reject $H_0 : \sigma_A \geq \sigma_B$ at $\alpha = 0.05$?

   iii. We may reject $H_0 : \sigma_A = \sigma_B$ at $\alpha = 0.05$?

(b) In comparing two population means using the independent samples t–test taught in this course, an assumption made is that the variances of the populations from which the independent samples are drawn are equal. A standard method of testing the above hypothesis of equality of variances is to use the F-test to test the null hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$, and then, if this null hypothesis is not rejected, to proceed with the independent samples t–test.

   i. State without elaboration whether non–rejection of H_0: $\sigma_X^2 = \sigma_Y^2$, (say at $\alpha = 0.05$), implies that the variances of populations X and Y are equal?

   ii. If not, why is the above method acceptable?

(c) Consider the following list of hypothesis tests:

   - Z-test for a single population mean,
   - Z-test for the difference between two population means when the two samples taken are independent and random,
   - t-test for a single population mean,
   - t-test for the difference between two population means when the two samples taken are independent and random,
   - F-test for two population variances when the two samples taken are independent and random,
   - χ^2 goodness of fit test.

   Now answer (i) and (ii) below separately. There is no need to elaborate — simply name the appropriate tests.

   i. Write down *two* tests from the above list that could be used to determine whether or not the population of male students in N.U.I., Galway has a mean height of 70 inches, when a random sample of 50 students is taken.

   ii. A random sample of 110 eggs is taken from a chicken farm. The eggs are weighed. It is found that $\bar{x}$, the mean weight of the sample of eggs, is 2 ounces, and that s_x^2, the variance of the sample, is $\frac{1}{4}$ ounce. Write down *one* test from the above list that could be used to investigate whether the population of weights of eggs from which the above sample is taken is normally distributed.

D 1588

5. A shellfish farmer is cultivating oysters under different conditions in two different sites, A and B. A random sample of 12 oysters is taken from site A, and weighed. A random sample of 9 oysters is taken from site B, and weighed. We may assume that these samples are independent. The results are as follows.

	A	B
Mean	68.9	67.00
Standard Deviation	4.5	4.8

(a) Show that s_p^2, the pooled variance, is 21.437, (and hence s_p= 4.63).

(b) Test the null hypothesis that the population mean weights of oysters from each site are equal.

(c) What two additional assumptions (i.e. in addition to the randomness and independence of the samples) about the weights of oysters from sites A and B do we make for the above test to be valid?

6. Six female teachers are asked to record the number of days during the school year on which they felt that they were subject to stress and they were also asked to record the number of days on which they experienced symptoms of stress over the same period. The results are shown in the table below.

Stress (x_i)	03	15	05	10	24	34
Symptoms (y_i)	100	109	62	81	115	121

Note: $\sum x_i$= 91, $\sum y_i$= 588, $\sum x_i y_i$= 9,929, $\sum x_i^2$= 2,091 $\sum y_i^2$= 60,152

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 (scatter plot) to represent these data.

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

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

(d) Using your scatter diagram from (a) above, does the sign of r calculated in (c) above appear to be correct. *Very* briefly justify your answer.

(e) Based on your answer in (b) write down a point estimate of $\mu_{Y|10}$, the population mean number of reported days of symptoms of stress of female teachers who have reported ten days subject to stress.

(f) Explain why your answer to (e) above differs from 81, the number of days of stress related symptoms reported by the teacher who experienced 10 days of stress.

(g) In (e) above we referred to a *point estimate* of $\mu_{Y|10}$, it is also possible to calculate a *confidence interval* for $\mu_{Y|10}$. (Do **not** calculate a confidence interval here.) Why can we not calculate an exact value of $\mu_{Y|10}$?