GD 2006

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II EXAMINATIONS 2003-2004

MA 238 –STATISTICS
Dr. T.C. Bailey
Dr. J. Newell

Time allowed: ***Two*** **hours.**
Answer the 10 Questions in Section A (20 marks)
and any 4 Questions from Section B (20 marks each).
Relevant tables and formulae are supplied

Section A – Compulsory

Each part is worth 2 marks with a **loss** of 1 mark for a wrong answer.

1. Which of the following statements is true regarding a population:

   a) it must refer to people;
   b) it is a collection of individuals or objects;
   c) neither of the above.

2. A sampling frame is:

   a) the list of units from which the sample is chosen;
   b) a table of random numbers;
   c) a non-probabilistic sampling method.

3. Sampling that divides the population in subgroups and chooses a proportionate number from each subgroup at random is called:

   a) cluster sampling;
   b) quota sampling;
   c) stratified sampling.

4. The estimated standard error of the mean is calculated as:

   a) $\frac{s}{\sqrt{n}}$;
   b) $\frac{\sigma}{\sqrt{n}}$;
   c) $\frac{\pi(1-\pi)}{n}$.

question continued ...

5. If the standard error of the sample mean is 5 with a sample size of 100, then in order to reduce the standard error of the mean to 2.5:

   a) decrease the sample size to 20;
   b) increase the confidence level;
   c) increase the sample size to 400.

6. If the value of a test statistic is in the critical region at the α=0.01 significance level:

   a) the P-value for this test is < 0.01 ;
   b) the P-value for this test is > 0.01;
   c) the value of the test statistic was not significant at the α% level.

7. Given that a *two-sided* test of H_0: $\mu = 50$. against H_1: $\mu \neq 50$ is not *significant* at the 5% level, which of these intervals is likely to be a 95% confidence interval for the true mean μ:

   a) [150 , 250];
   b) [30, 70];
   c) [20, 40].

8. Which of the following would be the most appropriate for measuring the strength of relationship between height (cm) and weight (kg) based on a random sample of subjects:

   a) a chi square test of association;
   b) a correlation coefficient;
   c) a two sample t-test?

9. In a simple linear regression, the intercept represents the:

   a) predicted value of Y when X = 0;
   b) variation around the line of best fit;
   c) estimated change in average Y per unit change in X.

10. In a simple linear regression, a 99% confidence interval for the population slope was calculated as [-10, -8]. This suggests that in the population:

   a) the response variable is not a useful predictor;
   b) there is no evidence of a non-zero slope;
   c) there is evidence of a non-zero slope.

Section B – Answer 4 Questions

Question 2.

a) Define the following terms in Statistics: a *population*, a *sample*, a *parameter*, a *statistic* and a *variable*. What does the term *inference* mean when applied to Statistics?

[6 Marks]

b)

(i) Explain the fundamental difference between the objectives addressed by probability theory as opposed to those addressed by statistical theory making reference to the respective roles of populations and samples.

[4 Marks]

(ii) What are meant by the terms *Descriptive* statistics and *Inferential* statistics?

[4 Marks]

c)

(i) List three ways in which stratified random sampling and cluster sampling differ.

[3 Marks]

(ii) Give a practical example where cluster sampling might be a better choice of sampling method over stratified random sampling.

[3 Marks]

Question 3.

a) The amount of rent paid by students in a large class follows a Normal distribution with population mean $\mu = €70$ and population standard deviation $\sigma = €3.5$.

(i) What range of values contains approximately 95% of all rents paid by students in this class?

[2 Marks]

(ii) Write down the approximate sampling distribution of the means of all possible samples of size 100 drawn randomly from this class.

[2 Marks]

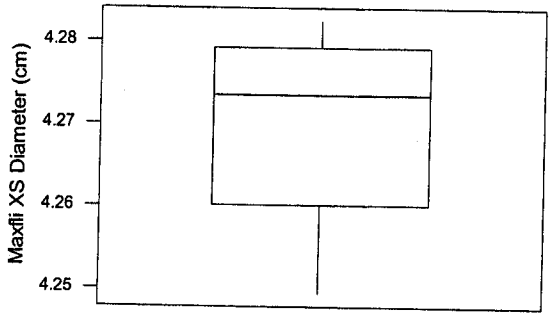
(iii) If 500 such samples were chosen at random and for each sample you calculated a 90% confidence interval for the true mean, how many such intervals would you expect to contain this population mean?

[2 Marks]

question continued ...

b) The United States Golf Association (USGA) requires that golf balls have a mean diameter of $\mu = 4.3$cm. A random sample of 13 new Maxfli XS golf balls was selected which provided the following sample statistics and boxplot:

	Maxfli XS
Sample size (n)	13
Sample mean ($\bar{x}$)	4.27
Sample standard deviation (s)	0.03



i) Construct and interpret a 95% confidence interval estimate of the true average diameter of Maxfli XS golf balls.

[8 Marks]

ii) Do you think that the Maxfli XS golf ball conforms to the requirement as stated by the USGA? Explain.

[2 Marks]

iii) Upon what distributional assumption does the calculation of the 95% confidence interval depend and does this assumption look valid based on the boxplot above?

[2 Marks]

iv) What information would a 95% prediction interval provide for you in this context?

[2 Marks]

Question 4.

An urban economist wanted to determine whether the mean price of a home differed in two large Irish cities. A random sample of 100 homes sold in each city resulted in the following summary statistics (in thousands of Euro):

	City A	City B
Sample size (n)	100	100
Sample mean ($\overline{x}$)	253.6	310.3
Sample standard deviation (s)	106.9	153.9

a) On the basis of the summary statistics above does it appear that there is a difference in the mean price for the samples of houses provided?

[2 Marks]

b) Is this an example of an observational or an experimental study? Explain.

[2 Marks]

c) Using a level of significance of 0.05, investigate whether there is evidence that the population mean price between the two cities is different. Your answer must include an appropriate null and alternative hypothesis, a suitable test statistic and critical region, a P-value, a decision and an interpretation.

[10 Marks]

d) In addition, calculate a 95% confidence interval for the true average difference in house price between the two cities and comment on your result.

[4 Marks]

e) Given that the sale prices of houses are typically skewed to the right, why do you think it is necessary to have large samples to test this hypothesis?

[2 Marks]

Question 5.

A survey was carried out on 1017 randomly selected people 18 years or older where they were asked whether they support the limited uses of marijuana when prescribed by physicians to relieve pain and suffering. The results of the survey, by age group, are as follows:

Opinion/ Age	**18-29**	**30-49**	**50 years or older**	Row Total
For	172	313	258	743
Against	52	103	119	274
Column Total	224	416	377	1017

a) Carry out a suitable hypothesis test (at $\alpha = 0.05$) to decide if there is any evidence of an association between age and opinion.

[10 Marks]

b) If you decide that there is an association between age and opinion, briefly describe the form of the association using column percentages.

[3 Marks]

c) What assumption must you make regarding the expected frequencies in order to carry out the test in part (i) and is this assumption valid in this analysis?

[2 Marks]

d) Based on these figures and using a 95% confidence interval what can you say regarding the true proportion of 18-29 year olds in favour of limited uses of marijuana when prescribed by physicians?

[5 Marks]

Question 6.

Data were collected on the number of calories per serving and the number of grams of sugar per serving for a random sample of 11 breakfast cereals. It was assumed that a simple linear regression model was an appropriate method to use in this analysis and a scatterplot of the data (with the least squares line of best fit superimposed) was prepared (Figure 1). The estimated regression coefficients, their respective standard errors, an estimate of the variability (*s*) of the estimated regression equation and some additional summary statistics are provided in Table 1.

question continued ...

Figure 1

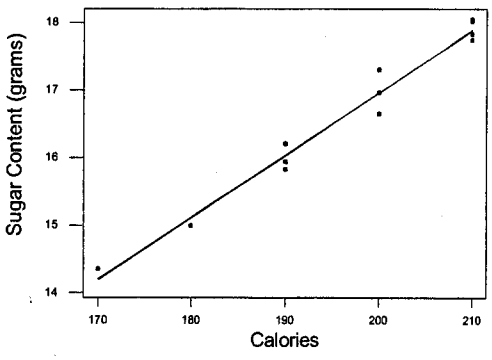


Table 1.
Regression Output for Breakfast Cereal Data

```
        The regression equation is

Sugar = - 1.50 + 0.09 Calories

Predictor        Coef     SE Coef
Constant       -1.50       0.921
Calories        0.09       0.004
```

$S = 0.20$, $r = 0.987$, $\overline{x} = 196.67$, $S_{xx} = 1866.67$

a) Identify the response and predictor variables in this analysis.

[2 Marks]

b) Does there appear to be a relationship between calorie and sugar content based on Figure 1 and the sample correlation coefficient?

[3 Marks]

c) Explain the line of best fit and in particular interpret the meaning of the slope (i.e. the regression coefficient) in this analysis.

[4 Marks]

d) Using a level of significance of 0.05, carry out a suitable hypothesis test to determine if there is evidence that calorie content is a good predictor of sugar content in breakfast cereals in general.

[6 Marks]

e) A new brand of breakfast cereal is on the market advertising a calorie content of 200. Provide a suitable 95% interval estimate of the likely sugar content of a future box of cereal of this brand.

[5 Marks]