OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II EXAMINATIONS 2002-2003

MA 238 –STATISTICS

Dr. D. Harrington.
Dr. J. Newell.

Time allowed: ***Two*** **hours.**
Answer Question 1 and any 4 others.
Each Question is worth 20 Marks while each part of Question 1 is worth 2 marks with a loss of 1 mark for a wrong answer.

Question 1.

1. A summary measure of the individual observations made by evaluation of a population is a(n)

   A) attribute;
   B) parameter;
   C) statistic.

2. The standard error of the mean

   A) is greater than the standard deviation of the population;
   B) increases as the sample size increases;
   C) measures the variability of the mean from sample to sample.

3. The main difference between stratified random sampling and cluster sampling is

   A) in cluster sampling all clusters are sampled while in stratified sampling only some strata are sampled;
   B) in stratified sampling all strata are sampled while in cluster sampling only some clusters are sampled;
   C) stratified samples are more representative.

4. The sampling distribution of the mean is a distribution of

   A) individual population values;
   B) individual sample values;
   C) statistics.

5. You and your friend decide to construct 95% confidence intervals for a given population mean where the population standard deviation is known. You wind up taking a sample of 49 random observations while your friend's sample is made up of 36 random observations. Which of the following is true?

A) Your friend's interval has a greater degree of confidence;
B) Your interval is narrower;
C) Your interval is wider.

6. A Type I error amounts to

A) rejecting the null hypothesis when it is false and should be rejected;
B) not rejecting the null hypothesis when it is false and should be rejected;
C) rejecting the null hypothesis when it is true and should not be rejected.

7. The _____ measures how close the computed sample statistic has come to the hypothesized population parameter.

A) level of significance;
B) confidence coefficient;
C) test statistic.

8. Given that a *two-sided* test of H_0: $\mu = 100$ against H_1: $\mu \neq 100$ is *significant* at the 5% level, which of these intervals is likely to be a 95% confidence interval for the true mean μ

A) [50 , 150];
B) [50, 70];
C) [99, 101].

9. If a p-value for an hypothesis test on a mean was given as 0.03, and the level of significance used was 5%, then the conclusion would be to

A) reject the null hypothesis;
B) not reject the null hypothesis;
C) accept the null hypothesis.

Question 1 continued.

10. In a simple linear regression, a 95% confidence interval for the population slope was calculated as [-1, 1]. This suggests that the

A) explanatory variable is a useful predictor of the response;
B) explanatory variable is not a useful predictor of the response;
C) response variable is a useful predictor.

Question 2.

(a)

What is meant by the term inference when applied to statistics?

[4 Marks]

(b)

What is the main difference between an observational and an experimental study in statistics? Give two reasons why randomisation is used in experimental studies.

[4 Marks]

(c)

A drug company believes that they have a new drug that will lower blood cholesterol levels amongst people at risk of coronary heart disease. They recruited the next 30 such people that registered at a local health practice and entered them in a 6 month trial of the new drug. At the end of the trial, 20 of the 30 people recruited had lower blood cholesterol levels than at the start of the trial.

(i) How impressed are you by these results and what reservations might you have regarding the study?

[4 Marks]

(ii) Suggest a more suitable sampling scheme and briefly outline why you consider it to be more suitable than the sampling scheme employed in this study?

[4 Marks]

(iii) How might you improve the design of this trial?

[4 Marks]

2113

Question 3.

(a)

(i) What does the central limit theorem say and why is it important when using a large sample to estimate a population mean?

[4 Marks]

(ii) If it is only possible to collect samples of size 20 what is the approximate sampling distribution of the mean in this case?

[2 Marks]

(iii) What additional information regarding the population distribution do you need in order for the approximate sampling distribution to be valid?

[1 Mark]

(iv) If 100 such samples of size 20 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 the true population mean?

[1 Marks]

(b) A company has a new process for manufacturing large artificial sapphires. The production of each gem is expensive, so the number available for examination is limited. In a trial run, 12 sapphires are produced. The mean weight for these 12 gems is 6.75 carats, and the sample standard deviation is 0.33 carats.

(i) Construct and interpret a 95% confidence interval estimate of the true average weight of sapphires made using the new process.

[8 Marks]

(ii) Do you think the manufacturer has the right to state that the average weight of sapphires is 6.8 carats? Explain.

[2 Marks]

(iii) Based on your interval estimate what can you say regarding the likely *actual* weight of the next sapphire produced?

[2 Marks]

Question 4.

(a) Explain what is meant by Independent and Dependent samples in the context of comparison of means problems in statistics.

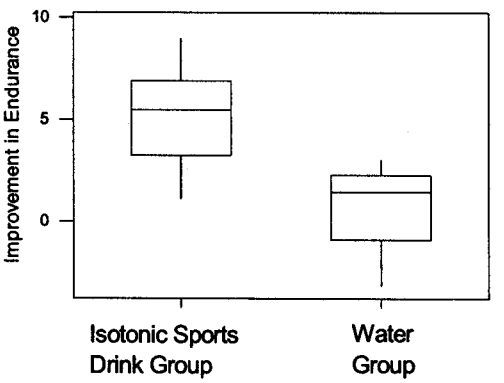
[2 Marks]

(b) A sports scientist at a large football club is interested in comparing the squad's endurance when using a particular isotonic sports drink compared to that if they used water. Due to time constraints he can only get access to a sample of 30 players. He chose the sample at random, made them carry out a test of endurance and measured their level of endurance. A week later he randomised the players into two groups and gave the first group water and the second group the isotonic sports drink in question.

Question 4 continued.

He then made the two groups carry out the identical test of endurance as before and recorded their endurance levels. He calculated the difference in endurance for each player where a positive difference indicated an increase in endurance.

A boxplot of the improvement in endurance levels for both groups is given overleaf.



	Improvement for those players using Water	Improvement for those players using the Isotonic Sports Drink
Sample size (n)	15	15
Sample mean ($\bar{x}$)	0.63	5.13
Sample standard deviation (s)	2.01	2.27

(i) On the basis of the boxplot above does it look like the average is a good summary statistic to compare these two groups?

[2 Marks]

(ii) Using a level of significance of 0.05, investigate whether there is evidence that the population mean endurance level between the two groups of players is different. Your answer must include an appropriate null and alternative hypothesis, a suitable test statistic and critical region, a decision and an interpretation.

[8 Marks]

(iii) In addition, calculate a 95% confidence interval for the true average difference in mean endurance level between the two groups of players and comment on your result.

[6 Marks]

(iv) How do you think the design of this study could be improved and briefly indicate the type of analysis you might carry out on your improved design?

[2 Marks]

GD 2113

Question 5.

(a) A sociologist is doing a study to see if there is a relationship between the age of a young adult (18 to 35 years old) and the type of movie preferred. A random sample of 93 adults revealed the following data.

Movie/ Age	**18 – 23 yr**	**24 – 19 yr**	**30 –35 yr**	Row Total
Drama				28
Science Fiction				41
Comedy				24
Column Total	29	33	31	93

(i) Carry out a suitable hypothesis test (at $\alpha = 0.01$) to decide if there is any evidence of an association between age and movie preference.

[10 Marks]

(ii) If you decide that there is an association between the age of a young adult and the type of movie preferred, briefly describe the form of the association.

[2 Marks]

(iii) 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]

(b) In a random sample of 400 males who were questioned regarding smoking habits, 46 said that they smoked while in a random sample of 380 females 76 were smokers. Based on these figures and using a 95% confidence interval what can you say regarding the difference in the proportion of male and female smokers?

[6 Marks]

Question 6.

A utility company needs to estimate the amount of natural gas that will be used by their customers. The consumption of natural gas required for home heating depends on the outdoor temperature. The daily gas consumption (MMcf) for each month in a particularly cold Midwestern city in addition to the midday temperature was recorded over nine months.

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 6 continued.

Figure 1

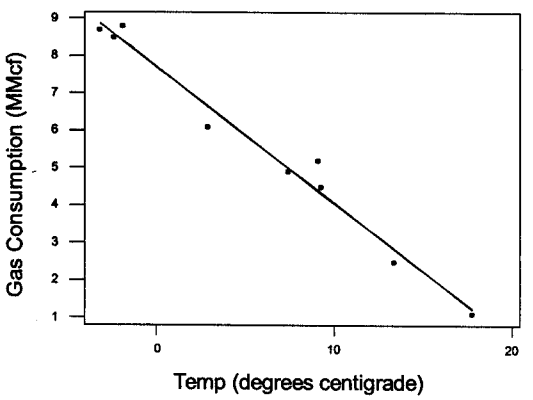


Table 1.
Regression Output for Gas Consumption Data

Gas Consumption = 7.70 - 0.36 Temperature

Predictor	Coef	SE Coef
Constant	7.70	0.19
Temperature	-0.36	0.02

S = 0.43, r = -0.99, $\bar{x}$ = 5.81, S_{xx} = 748.31

(i) Does there appear to be a relationship between temperature and gas consumption based on Figure 1 and the sample correlation coefficient?

[3 Marks]

(ii) 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]

(iii) What information is the intercept providing in this analysis?

[2 Marks]

(iv) Using a level of significance of 0.05, carry out a suitable hypothesis test to determine if there is evidence that temperature is indeed a good predictor of gas consumption.

[6 Marks]

(v) The meteorological office is predicting tomorrows midday temperature to be of 10°C in the city where the data was collected. Provide a suitable 95% interval estimate of tomorrow's likely gas consumption.

[5 Marks]

GD 2113