2580055

THE UNIVERSITY of York

Degree Examination 2007

ENVIRONMENT DEPARTMENT

MSc Environmental Science & Management
MSc Marine Environmental Management
MSc Environmental Economics & Environmental Management
MSc Environmental Economics

STATISTICS & QUANTITATIVE METHODS

Time allowed: **two hours**

There are three sections to this examination.

You must answer all parts of all questions in each section using the space in the paper.

SECTION 1 consists of 15 multiple choice questions. The total marks for this section is 15, one for each question.

SECTION 2 consists of four questions, each with several parts. The individual mark for each part of each question is indicated in the question. The total number of marks for this section is 60.

SECTION 3 consists of five questions, each with several parts. The individual mark for each part of each question is indicated in the question. The total number of marks for this section is 35.

Calculators will be provided

Pay adequate attention to spelling, punctuation and grammar, so that your answers can be readily understood

Section 1.

Tick the one box which best completes the statement in each question. Answer all 15 questions. There is 1 mark per question.

1. In a variable which exactly follows a normal distribution:

a. The median is less than the mean ☐

b. You cannot calculate the median ☐

c. The median is greater than the mean ☐

d. The median and the mean are the same ☐

2. The variable *Diet* describes the feeding habits of each species of mammal in a dataset as either carnivorous, insectivorous, omnivorous or herbivorous. *Diet* is an example of:

a. A continuous variable ☐

b. An ordinal variable ☐

c. A binary variable ☐

d. A categorical variable ☐

3. The variance of a variable is calculated using:

a. The sum of the squared differences of each observation from the mean ☐

b. The sum of the differences of each observation from the mean ☐

c. The sum of the squared differences of each observation from the median ☐

d. The maximum observed value minus the minimum observed value ☐

4. We often use samples to estimate parameters (e.g. the mean) for a wider population. Estimates of parameters will be:

a. More reliable if the sample size is large ☐

b. More reliable if the sample size is small ☐

c. Unaffected by sample size ☐

d. Most accurate for intermediate sample sizes ☐

5. The standard error of the mean measures:

a. How big a sample size we need to get an accurate estimate of the mean

b. How the estimate of the mean changes with sample size

c. The difference between two means

d. The standard deviation of an estimate of the mean

6. The probability of tossing a coin five times and getting five heads is:

a. 1/5

b. 1/25

c. 1/32

d. 1/64

7. Hypothesis testing under the scientific method usually involves:

a. Proving that a hypothesis is correct

b. Gathering data to support a hypothesis

c. Trying to falsify a null hypothesis

d. Trying to falsify an alternative hypothesis

8. *P* values measure:

a. The degree of support for the alternative hypothesis

b. The strength of evidence against the null hypothesis

c. The probability that the null hypothesis is correct

d. The probability that the alternative hypothesis is wrong

9. The standard normal distribution has:

a. Any positive mean and standard deviation

b. A mean of 1 and a standard deviation of 1

c. A mean of 1 and a standard deviation of 0

d. A mean of 0 and a standard deviation of 1

10. Degrees of freedom are a measure of:

a. The sample size

b. The number of observations which are free to vary

c. The number of parameters estimated

d. The number of predictor variables in a model

11. The main difference between parametric and non-parametric statistical tests is:

a. Parametric tests make assumptions, non-parametric tests don't

b. Non-parametric tests are always based on ranked data

c. Parametric tests assume your data is normally distributed

d. Parametric tests obtain *P* values from a specified Probability Density Function

12. The value of a correlation coefficient:

a. Can be anything between 0 and 1

b. Can be anything between -1 and 1

c. Will depend on the units in which you measured the variables

d. Will depend on the sample size

13. The value of a regression slope:

a. Can be anything between 0 and 1

b. Can be anything between -1 and 1

c. Will depend on the units in which you measured the variables

d. Will depend on the sample size

14. The *t* distribution is an example of a Probability Density Function:

a. With a shape that depends on the degrees of freedom

b. Which is not symmetrical

c. Which can only take positive values

d. Which is the same as the normal distribution

15. *t* tests are mainly used:

a. To test for differences in mean between any number of groups

b. To test hypotheses about single samples

c. To test for differences in mean between two groups

d. To test for relationships between pairs of variables

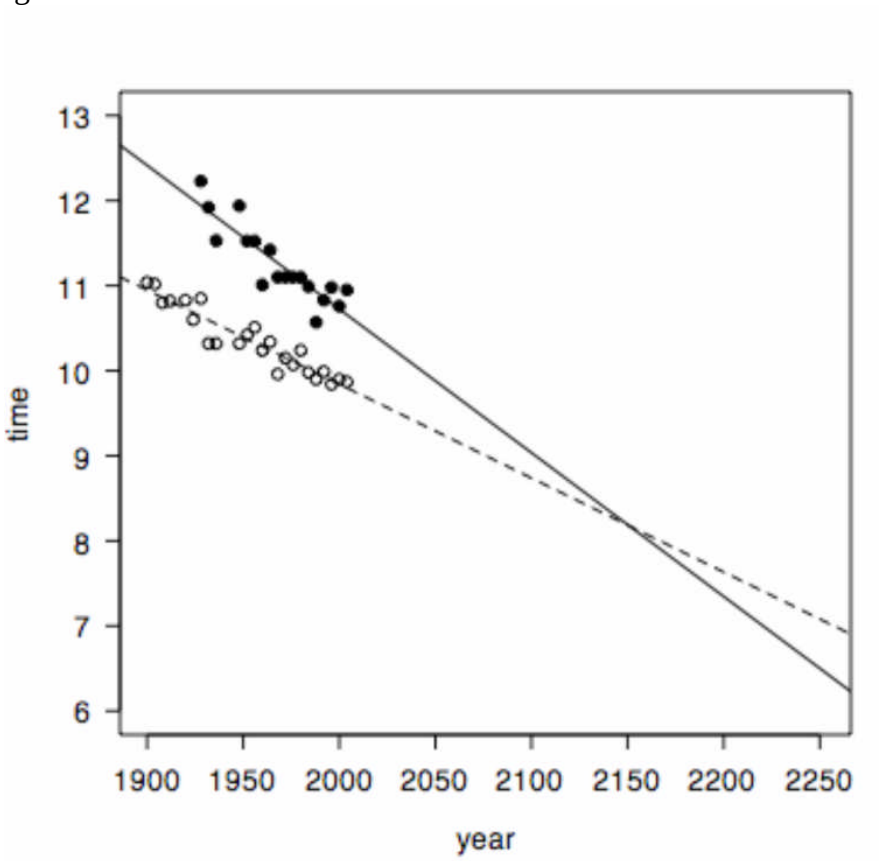
Section 2.

Answer all questions, using only the space provided. The marks available are shown in brackets after each part of each question. The total marks for this section are 60.

2.1 In an article in *Nature* from 2004, Tatem et al. showed how the winning times in both the men's and women's Olympic 100m race had become significantly faster over the course of the 20th Century (1900-2004 Olympics inclusive). They noted that women were getting faster quicker than men, such that the gap between men's and women's times had narrowed. They fitted regression models to the two sets of times, and used these to predict that in 2156 the fastest woman would run the 100m in a better time (8.08s) than the fastest man (8.10s). The regression equations were:

Men: $time = 31.94 - 0.011year$
Women: $time = 44.53 - 0.017year$

The results are presented in the figure below. Women's times are shown as solid symbols, and the regression is the solid line. Men's times are shown as open symbols, and the regression is the dashed line.



a. What is meant by the terms 'slope' and 'intercept' in regression? (2 marks)

b. Fill in the following table, using the regression equations given above. (2 marks)

	Slope	Intercept
Males		
Females		

c. Follow the logic used by the authors to predict the fastest male 100m time in the year 1066. (2 marks)

d. Use the same logic to predict when females will be able to cover 100m in zero seconds. (3 marks)

e. Your answers to questions c and d should have convinced you that the authors have used faulty logic. What precisely have they done which is generally not appropriate following a regression analysis? (4 marks)

2.2 In a study of the effect of grazing pressure on fruit production of a plant species, an ecologist assigned 120 plants at random to three grazing treatments, Heavy, Moderate and No grazing. At the end of the growing season, the mass of fruit produced was measured. The root diameter of each plant was also recorded as a measure of plant size.

a. Write down a suitable model formula (of the form *response(s) ~ predictor(s)*) to test the null hypothesis that the mass of fruit was not affected by grazing regime, and neither was it influenced by root diameter. (2 marks)

b. The ecologist analysed the data using Anova. The Anova table is given below, but most of the degrees of freedom are missing. Fill these in. (4 marks)

	d.f.	SS	MS	F	*P*
grazing		624.09	312.04	70.08	<0.0001
root diameter		2540.60	2540.60	570.61	<0.0001
grazing * root diameter		122.77	61.38	13.79	<0.0001
Residual		507.58	4.45		
Total	119	3371.53			

c. What is your main conclusion from this Anova table? (3 marks)

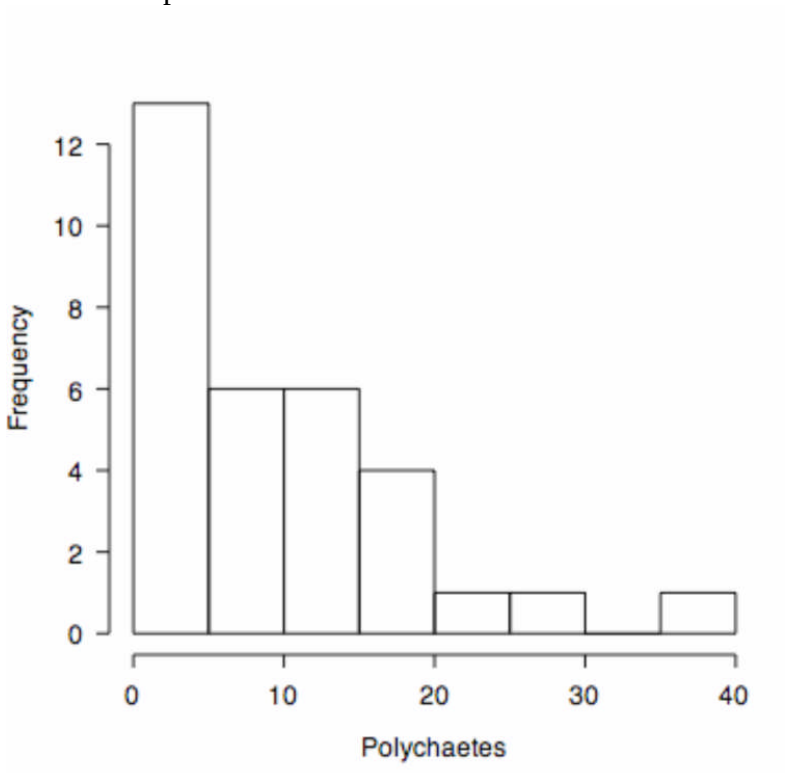
d. The table of coefficients is given below. Use this to write out separate regression equations for each grazing treatment. (6 marks)

Parameter	Estimate
Intercept	34.37
[grazing = 0]	-29.12
[grazing = Medium]	-14.07
[grazing = High]	NA
Root diameter	5.84
[grazing = 0]	4.28
[grazing = Medium]	2.30
[grazing = High]	NA

2.3 A marine biologist was interested in the rates of settlement of larval Polychaete worms onto each of four different types of artificial substrate in shallow waters. The substrates were experimentally controlled, and were randomly placed. There were eight replicates of each of the four types of substrate.

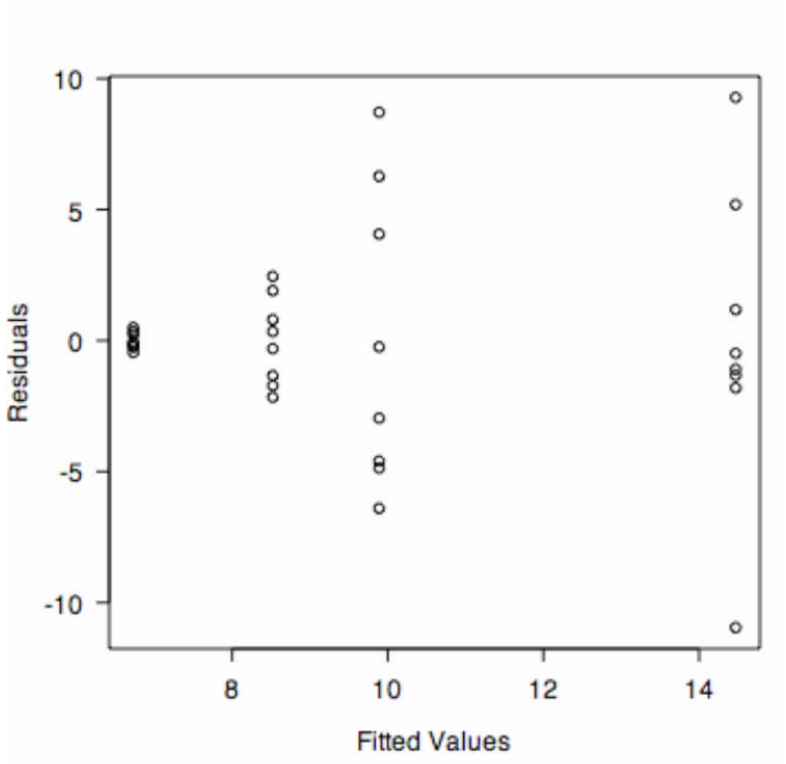
a. Write an appropriate model formula (of the form *response(s) ~ predictor(s)*) to test the null hypothesis that there was no difference in the number of Polychaete recruits found on each substrate type. (2 marks)

The number of polychaetes on each of the 32 artificial substrates was counted, and the frequency distribution is plotted below.



b. Describe the shape of this distribution. (2 mark)

c. The biologist analysed the raw data using a 1-way Anova, and before interpreting the output decided to check the assumptions of the test. Part of this process involved checking a plot of residuals versus fitted values, shown below.



How do you interpret this plot, as regards the assumptions of the test? (3 marks)

d. One way of proceeding would be to perform a transformation. Suggest an appropriate transformation, and identify the variable which you would transform. (2 marks)

e. Sometimes transformations are not possible. What is an alternative test which could test for differences between substrates, and which has less stringent assumptions than Anova? (1 mark)

f. The output from the Anova (using transformed data) is shown below. The output includes estimates of mean values for each substrate type. A Tukey test was conducted to test for pairwise differences between the mean values for each substrate; the results are summarised in the third table below which defines homogenous subsets.

	d.f.	SS	MS	F	*P*
Substrate	3	0.39	0.13	10.10	0.0001
Residuals	28	0.36	0.01		

$R^2 = 0.52$

Estimated means:

Substrate Type	Mean	S.E.
A	2.11	0.033
B	2.26	0.036
C	1.91	0.038
D	2.10	0.051

Homogeneous subsets

Substrate Type	subset 1	subset 2	subset 3
A		2.11	
B			2.26
C	1.91		
D		2.10	

Write a short paragraph summarising these results in the style you might use when writing up a scientific paper. (6 marks)

2.4 Generalized Linear Models are an extension of General Linear Models.

a. Which method of estimation is used by Generalized Linear Models, and how does this differ from General Linear Models? (2 marks)

b. Under what circumstances are Generalized Linear Models are particularly useful? (2 marks)

c. Manatees are marine mammals which are sensitive to human disturbance. In an attempt to quantify this, a marine conservationist surveyed a series of sites around the Caribbean, and recorded manatee presence or absence, as well as distance (in km) from the nearest major human settlement.

Write a suitable model formula (of the form *response(s) ~ predictor(s)*) to test the null hypothesis that proximity to human settlements has no effect on the likelihood of manatees being present at a site. (2 marks)

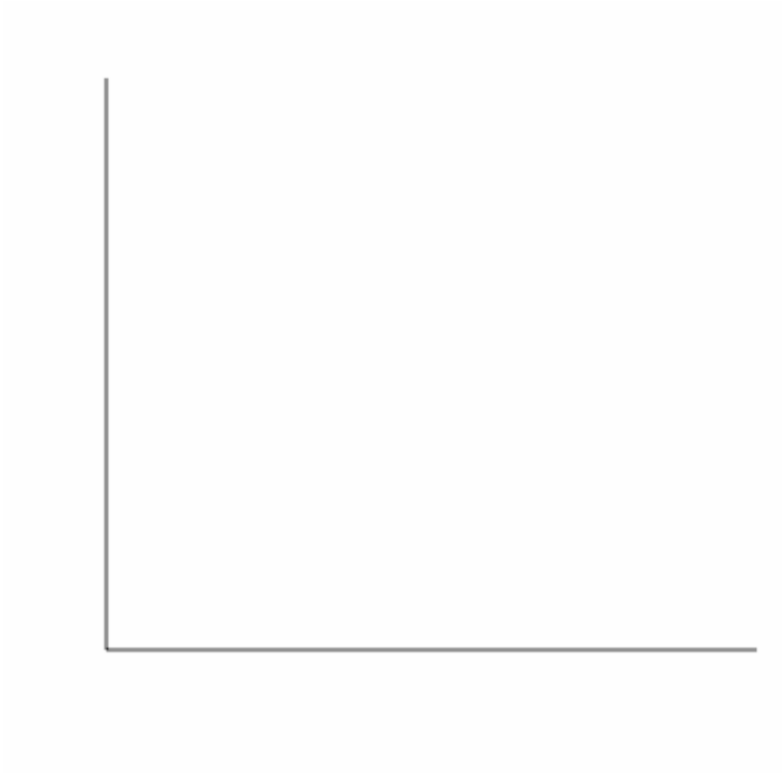
d. What would be a suitable statistical technique to test this hypothesis? (1 mark)

e. The conservationist carried out this analysis, and part of the output is reproduced below:

	b	s.e.	Wald statistic	*P*
Distance	0.561	0.238	2.358	0.0184

Use the information in this table to calculate the odds ratio for this test. (3 marks)

f. State your conclusions from this analysis and sketch the relationship suggested by the results on the axes provided below. Do not attempt to draw the relationship to scale, but show the broad shape of the relationship, and label axes and note any values that you think are important. (7 marks)



Section 3.

Answer all questions, using only the space provided. The marks available are shown in brackets after each part of each question. The total marks available for this section is 35.

3.1 A student was interested in the ectoparasites of badgers. She counted the number of fleas and ticks observed in a set amount of time on each of 20 individual badgers. Neither of the resulting variables (flea number and tick number) was normally distributed, and transformations did not appear to help.

a. Suggest, with justification, an appropriate statistical test to test the null hypothesis that there is no difference in the number of each kind of parasite across this population of badgers. (2 marks)

b. Suggest, with justification, an appropriate statistical test to test the null hypothesis that the numbers of fleas and ticks on a badger are unrelated. (2 marks)

3.2 A university administrator was concerned about possible gender bias in admissions to different departments. He collected data on the number of male and female undergraduates in each of 4 faculties, Arts, Engineering, Medicine and Natural Sciences.

a. Suggest an appropriate technique to test the null hypothesis that there is no difference in the proportion and female undergraduates between the different faculties. (2 marks).

b. Construct a table to illustrate how you would structure the data in order to perform this test. (3 marks)

3.3 A marine biologist was interested in classifying a series of sampling stations throughout the North Sea on the basis of the concentrations of several trace elements, together with a number of other chemical and physical properties (e.g. substrate type).

a. What is a technique which could be used to determine how similar the sampling stations were to each other on the basis of all measured variables? (2 marks)

b. She then noted whether each sampling station occurred in close proximity to an oil installation, or not (i.e. each sampling site was classified as either ‘oil platform’ or ‘not oil platform’). Suggest an appropriate statistical technique to test the general null hypothesis that the different kinds of sites cannot be distinguished on the basis of the chemical and physical variables measured. (2 marks)

3.4 An aquaculturist wanted to know whether sea bass responded in the same way as salmon to different stocking densities within cages. He set up an experiment in which each species was raised at each of 4 stocking densities. At the end of 4 weeks, the mean growth rate was recorded for each species at each density.

a. How many treatment combinations are there in this experiment? (1 mark)

b. Due to constraints of space, the experiment had to be conducted at three separate sites. 16 cages were available at each site. Give the name of a suitable design for this experiment, and describe what it would involve. (3 marks)

c. Write a model formula for an appropriate analysis of the experiment. (2 marks)

3.5 A fisheries biologist wanted to examine the potential fisheries benefits of marine reserves. She collected catch data (catch per unit effort, CPUE) from fishers who fished along the edge of existing marine reserves, and from fishers in the same region who fished in areas more distant from marine reserve. As a first analysis, the biologist decided to compare mean CPUE across the two groups of fishers.

a. What would be a suitable test to achieve this, and which variables would you include? (2 marks)

b. Assuming the design of the survey is statistically sound, what analysis-level assumptions does this test make? (2 marks)

c. List two further variables which could have been collected along with CPUE and which would result in a more informative future analysis. (2 marks)

END