FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

EXAMINATION PAPER 2016

MODULE CODE: MA4505

SEMESTER: Autumn 2016

MODULE TITLE: Applied Statistics for Administration 1

DURATION OF EXAM: 2.5 hours

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 100 marks (60% of Module Grade)

EXTERNAL EXAMINER: Prof. A. Marshall

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables are provided at the end of the exam paper.
Students must attempt any 4 questions from 5.

Question 1. (25 marks) Descriptive Statistics

Part A - Graphical Methods (15 Marks)

The exam results for a class of 60 students are tabulated below.

19	25	30	35	35	36	36	37	37	38
38	38	39	39	40	43	43	43	44	45
46	47	47	47	47	47	48	48	49	49
50	51	52	53	53	53	54	56	57	57
59	60	60	60	61	62	63	63	64	64
65	66	69	72	78	85	88	89	93	99

(i) (3 Marks) Summarize the data in the above table using a relative frequency table, a cumulative frequency table, and a cumulative relative frequency table. Use 9 class intervals, with 11 as the lower limit of the first interval.

(ii) (6 Marks) Draw a histogram for the above data. Comment on the shape of the histogram. Based on the shape of the histogram, what is the best measure of centrality and variability?

(iii) (6 Marks) Construct a box plot for the above data. Clearly demonstrate how all of the necessary values were computed.

Part B - Summary Statistics (5 Marks)

Data on the durations (measured in months) were collected for a random sample of product development projects. The durations for these development projects are as follows:

16	20	14	29	30	22	21	28

(i) (1 Mark) Calculate the mean project duration.

(ii) (1 Mark) Calculate the median project duration.

(iii) (2 Marks) Calculate the variance of the project durations.

(iv) (1 Mark) Calculate the standard deviation of the project durations.

Part C - Dixon Q Test (5 Marks)

Use the Dixon Q Test to determine if the lowest value is an outlier. You may assume a significance level of 5%.

$$131, 136, 101, 126, 123, 120, 132, 137$$

(i) (1 Mark) State the null and alternative hypotheses for this test.

(ii) (2 Marks) Compute the test statistic

(iii) (1 Mark) State the appropriate critical value.

(iv) (1 Mark) What is your conclusion to this procedure.

Question 2. (25 marks) Probability Distributions

Part A - Normal Distribution (10 Marks)

Assume that the diameter of a critical component is normally distributed with a mean of 500mm and a standard deviation of 12.5mm. You are required to estimate the approximate probability of the following measurements occurring on an individual component.

(i) (2 Mark) Greater than 517mm.

(ii) (3 Marks) Less than 495mm.

(iii) (2 Marks) Between 495mm and 513mm.

(iv) (3 Marks) The production manager reports that more than 99% of the output is between 475mm and 525mm. Do you agree with this statement? Justify your answer with the appropriate calculations.

Use statistical tables to determine the probabilities for the above exercises. You are required to show all of your workings.

Part B - Probability (10 Marks)

A new test has been developed to diagnose a particular disease. If a person has the disease, the test has a 95% chance of identifying them as having the disease. If a person does not have the disease, the test has a 1% chance of identifying them as having the disease. Suppose that 5% of the population have this disease. Suppose we select a person at random from the population.

(i) (4 Marks) What is the probability that the test will identify them as having the disease?

(ii) (3 Marks) What is the probability that the person has the disease given that the test identifies them as having the disease?

(iii) (3 Marks) What is the probability that the person has the disease given that the test identifies them as **not** having the disease?

Part C - Testing Distributional Assumptions (8 Marks)

Consider the results of a statistical analysis carried out on sample data for the random variable Y. These results are presented as output from a statistical computer program.

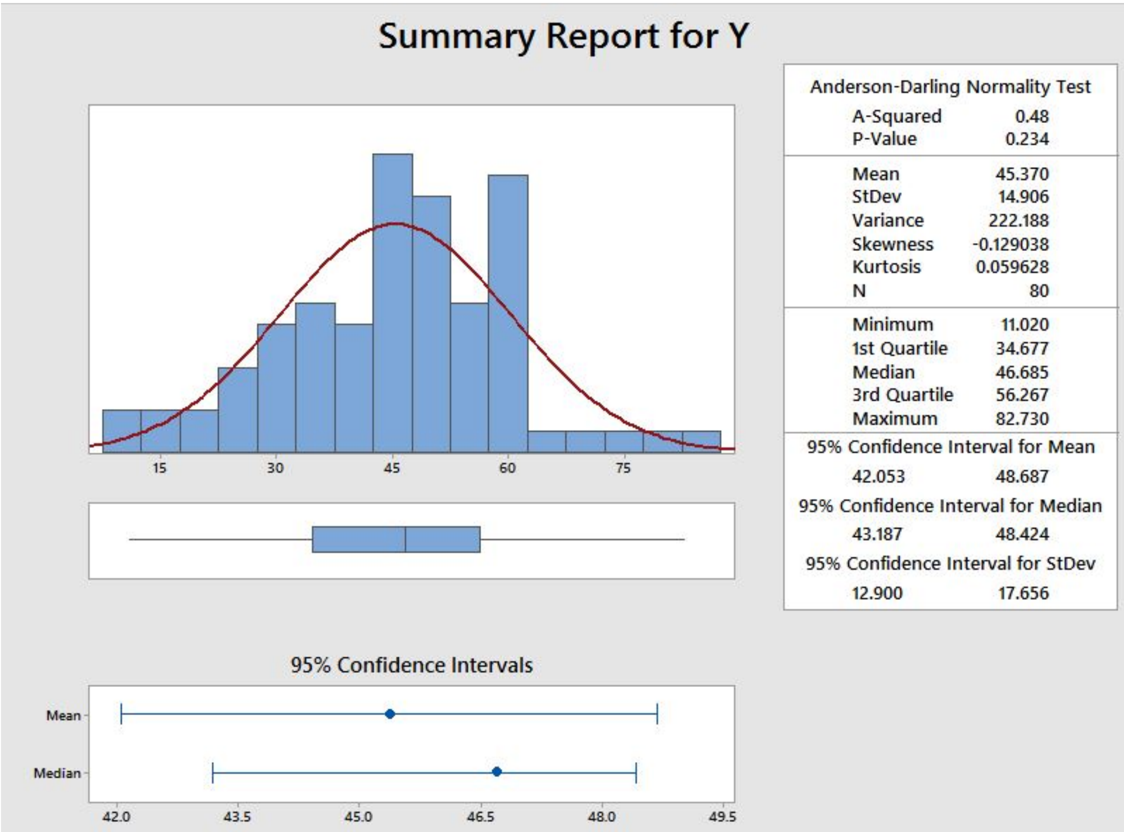
(i) (1 Mark) What sort of statistical analysis are we carrying out?

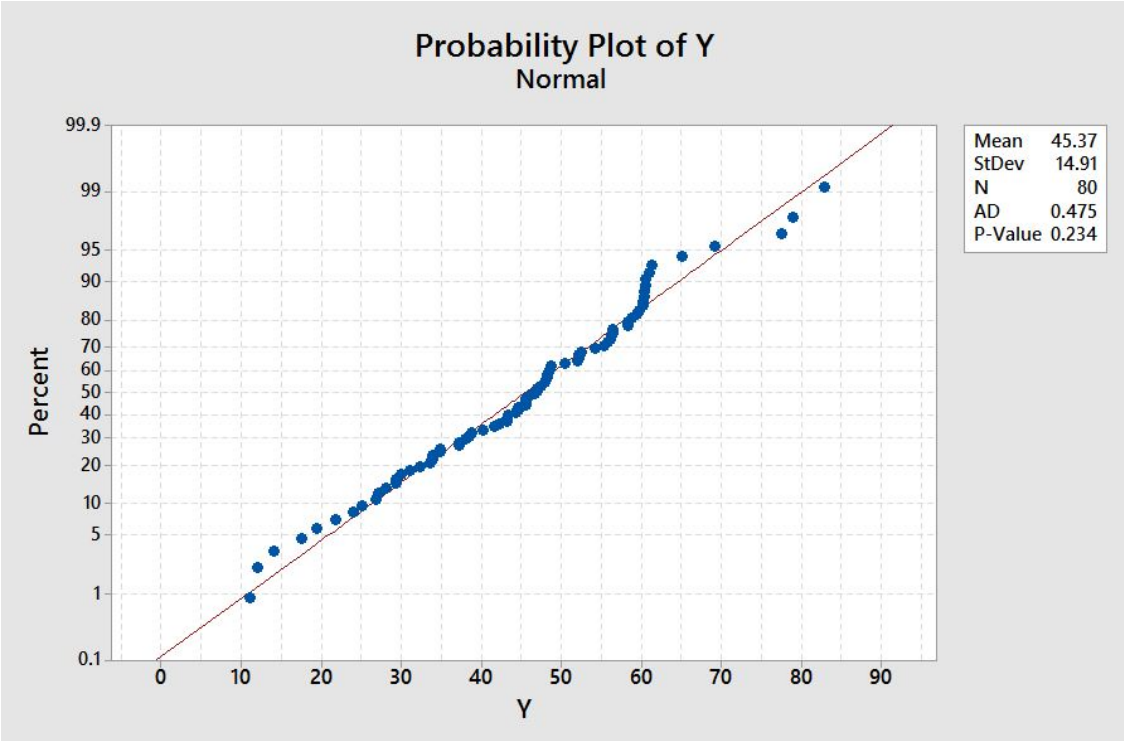
(ii) (1 Mark) What is the relevance of this analysis as part of an overall statistical study.

(ii) (2 Mark) Interpret the output of the Anderson-Darling Test. (*Remark: Use a 5% significance level. This test is a one-tailed procedure.*)

(iv) (1 Mark) What is the conclusion of this analysis for the variable X? Justify your answer with reference to 3 separate indications.

(*Computer results continue on the next page.*)





Question 3. (25 marks) Single Sample Inference Procedures

Part A - Single Sample Inference Procedures (15 Marks)

Mean blood iron concentration for children with adequate nutrition is taken to be 110mg/dl.

25 randomly selected children from a disadvantaged urban area were given blood tests. The mean concentration of iron from this sample was 98 mg/dl with a standard deviation of 25.5 mg/dl.

(i) (4 Marks) Calculate a 95% confidence interval for the mean concentration of iron for children in this area.

(ii) (2 Marks) Interpret this confidence interval. Do these results provide evidence that children in this area suffer from iron deficiency?

Test this hypothesis using a 5% level of significance.

(iii) (3 Marks) Formally state your null and alternative hypotheses.

(iv) (4 Marks) Compute the test statistic.

(v) (2 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Part B - Single Sample Proportion (10 Marks)

An environmental group states that fewer than 60% of industrial plants comply with air pollution standards. An independent researcher takes a sample of 400 plants and finds that 270 are complying with air pollution standards.

(i) (5 Marks) Carry out a hypothesis test to investigate the claim made by the environmental group. Clearly state your null and alternative hypotheses and your conclusion.

(i) (3 Marks) Compute the 95% confidence interval.

(ii) (2 Marks) By interpreting this confidence interval, state your conclusion about the environmental group's claim? Explain how you made this decision.

Question 4. (25 marks) Two Sample Inference Procedures

Part A - Two Sample Test for Means (10 Marks)

An exercise physiologist wants to determine if several short bouts of exercise provide the same benefit for cardiovascular fitness as one long bout of exercise.

50 volunteers are randomly assigned to group 1 and do standardised aerobic exercise on a stationary bicycle for 30 minutes once a day, 5 days a week. 40 volunteers are randomly assigned to group 2 and do the same exercise for 10 minutes, 3 times a day, 5 days a week. Cardiovascular fitness was measured by VO2 max (maximum oxygen consumption while exercising).

Group 1 The mean change in VO2 after 12 weeks of exercise was 2.1 for group 1 with a standard deviation of 1.7.

Group 2 The mean change in VO2 after 12 weeks of exercise was 0.7 for group 2 with a standard deviation of 1.

Test the hypothesis that there is no significant difference between two groups are the same.

(i) (3 Marks) Formally state your null and alternative hypotheses.

(ii) (4 Marks) Compute the test statistic.

(iii) (3 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Part B - Paired T Test (15 Marks)

A microbiologist measures the total growth in 24 hours of two strains of a germ culture in the same petri dish. Nine identical specimens are prepared. The growth rate for the eight samples for each strain are tabulated below:

Specimen	Strain 1	Strain 2
1	212	224
2	234	231
3	214	209
4	236	243
5	221	231
6	212	216
7	202	213
8	210	216
9	248	242

At a significance level of 5%, is there sufficient evidence to state that there is any difference in growth rates between the two strains.

(i) (3 Marks) Formally state the null and alternative hypotheses.

(ii) (3 Marks) Compute the mean and standard deviation of the case-wise differences.

(iii) (3 Marks) Compute the test statistic.

(iv) (3 Marks) State the appropriate critical value for this hypothesis test.

(v) (3 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Question 5. (25 marks) Linear Regression Models and Chi Square Tests

Part A - Chi Square Tests (10 Marks)

A market research survey was carried out to assess preferences for three brands of chocolate bar, A, B, and C. The study group was categorised by maturity level to determine any difference in preferences. The purpose of this study is to assess differing preferences for brands across age groups.

	Brand A	Brand B	Brand C	Total
Children	65	55	30	150
Teenages	35	75	40	150
Adults	50	20	30	100
	150	150	100	400

(i) (1 Mark) Formally state the null and alternative hypotheses.

(ii) (2 Marks) Compute each of the cell values expected under the null hypothesis.

(iii) (4 Marks) Compute the test statistic.

(iv) (1 Mark) State the appropriate critical value for this hypothesis test.

(v) (2 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Part B - Linear Regression (15 Marks)

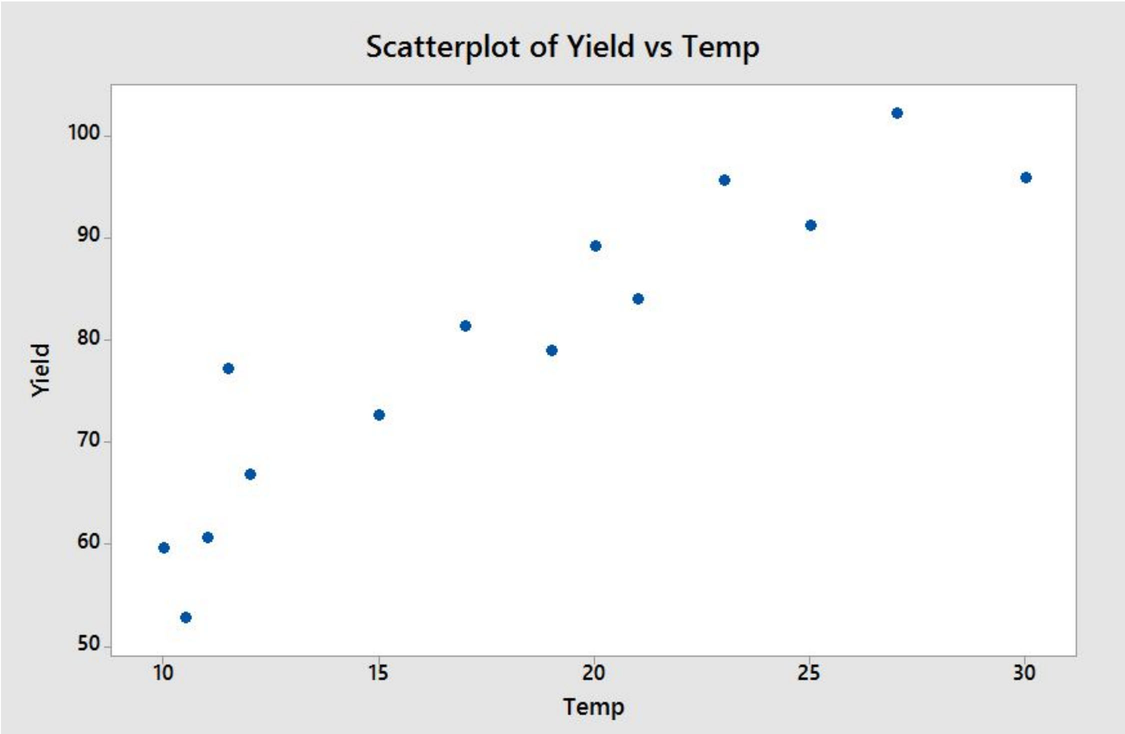
An experiment was conducted to study the relationship between baking temperature x (in units of 10 degrees Farenheit) and yield y (as percentage) of popular cake mix. Fourteen observations were made giving the following results.

Specimen	1	2	3	4	5	6	7
Temp	10.00	10.50	11.00	11.50	12.00	15.00	17.00
Yield	29.830	26.370	30.325	36.100	33.410	36.335	40.655

Specimen	8	9	10	11	12	13	14
Temp	19.00	20.00	21.00	23.00	25.00	27.00	30.00
Yield	39.475	44.555	42.015	47.795	45.560	51.045	47.900

- $S_{XX} = 570.5$
- $S_{XY} = 613.27$
- $\bar{Y} = 39.383$
- $S_{YY} = 751.1525$
- $\bar{X} = 18$

(This question is continued on the next page.)



(i) (2 Marks) Using the scatter plot, describe the relationship between the yield (Y) and the baking temperature (X).

(ii) (4 Marks) Calculate the correlation coefficient. Interpret your answer.

(iii) (6 Marks) Calculate the equation of the least squares regression line and interpret the value of the slope.

(iv) (3 Marks) Using this regression model, estimate the yield when the baking temperature is 16 degrees.

Formulae

Descriptive Statistics

- Sample Variance

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

Probability

- Conditional probability:

$$P(B|A) = \frac{P\,(A \text{ and } B)}{P\,(A)}.$$

- Bayes' Theorem:

$$P(B|A) = \frac{P\,(A|B) \times P(B)}{P\,(A)}.$$

Confidence Intervals

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(\hat{P}) &= \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}. \end{aligned}$$

Two samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}. \end{aligned}$$

Hypothesis tests

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(p) &= \sqrt{\frac{p \times (1 - p)}{n}} \end{aligned}$$

Two large independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{(\bar{p} \times (1 - \bar{p}))\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \end{aligned}$$

Two small independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \\ s_p^2 &= \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}. \end{aligned}$$

Paired sample

$$S.E.(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

Standard Deviation of case-wise differences (computational formula)

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}.$$

Chi Square Tests of Independence

$$\chi^2_{TS} = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Regression Estimates

$$\begin{aligned} S_{XY} &= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \\ S_{XX} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ S_{YY} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} \end{aligned}$$

Pearson's correlation coefficient

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}$$

Slope Estimate

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

Intercept Estimate

$$b_0 = \bar{y} - b_1\bar{x}$$

Critical Values for Dixon Q Test

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.410	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463

Critical Values for Chi Square Test

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
1	2.705	3.841	6.634	10.827
2	4.605	5.991	7.378	9.21
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.07	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.09
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209