Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2017	MATH1541-WE01

Title: Statistics

Time Allowed:	3 hours	
Additional Material provided:	Graph paper Tables: Normal distribution, t-distribution, Chi-squared distribution, F-distribution, Wilcoxon test, Mann-Whitney test.	
Materials Permitted:	You may keep one folder of notes at your desk.	
Calculators Permitted:	Yes	Models Permitted: Casio fx-83 GTPLUS or Casio fx-85 GTPLUS.
Visiting Students may use dictionaries: No		

Instructions to Candidates:	Credit will be given for the best **SIX** answers. All questions carry the same marks. This is an open-book examination: you may keep one folder of notes at your desk.

Revision:	

1. We consider an experiment designed to investigate the effects of the recipient on the growth of two species of eucalyptus seedlings (Simões, 1970). The recipients (R) and species involved (E) were:

R_1	large plastic bag
R_2	small plastic bag
R_3	plain
E_1	Eucalyptus citriodora
E_2	Eucalyptus grandis

The observed response is the average seedling height per recipient after a period of 80 days. For each treatment four replicates were produced. The data are reported below in form of a two-way-layout:

Recipient	Species							
	E_1				E_2			
R_1	26.2	26.0	25.0	25.4	24.8	24.6	26.7	25.2
R_2	25.7	26.3	25.1	26.4	19.6	21.1	19.0	18.6
R_3	22.8	19.4	18.8	19.2	19.8	21.4	22.8	21.3

(a) Carry out mean polish to estimate the overall mean, the effects of species and recipient, and the interaction effects.

(b) The overall standard deviation of the response is 2.94. Complete the following ANOVA table and give an interpretation.

Recipients	??
Species	19.08
Interaction	63.76
Residuals	23.09
Total	??

2. The following data set represents three-month certificate of deposit (CD) rates for 69 Long Island banks, as given in the August 23, 1989, issue of *Newsday*. The first 29 observations (displayed in the first three rows below) correspond to commercial banks, while the other ones correspond to thrift (Savings and Loan) institutions.

```
7.56 7.57 7.71 7.82 7.82 7.90 8.00 8.00 8.00 8.00
8.00 8.00 8.00 8.05 8.05 8.06 8.11 8.17 8.30 8.33
8.33 8.40 8.50 8.51 8.55 8.57 8.65 8.65 8.71

7.51 7.75 7.90 8.00 8.00 8.00 8.15 8.20 8.25 8.25
8.30 8.30 8.33 8.33 8.34 8.35 8.36 8.40 8.40 8.40
8.40 8.40 8.40 8.45 8.49 8.49 8.49 8.50 8.50 8.50
8.50 8.50 8.50 8.50 8.50 8.50 8.52 8.70 8.75 8.78
```

(a) Before carrying out any analysis: are there any conspicuous features of the data?

(b) Draw a stem and leaf plot displaying the distributions for both types of banks, using the same stem for both distributions. Comment on the shape of the distributions (in particular: relative position, modality).

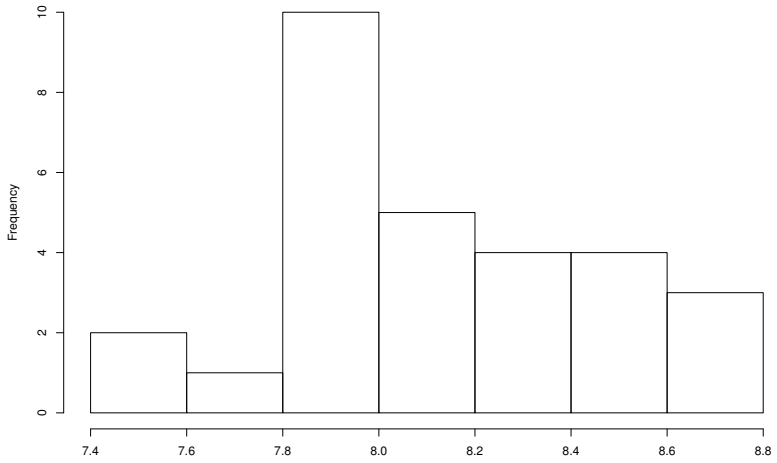
(c) Provide the five-number summary of the CD rates of the thrifts and use it to draw a boxplot, featuring the usual modification to show possible outliers.

(d) Below you find a histogram which corresponds to the data from one of the two subgroups.

i. Decide to which subgroup it belongs.

ii. What is the typical size of variation of the height of the largest bar when choosing *another* sample (of the same size from the same subpopulation)?

iii. How many bars would the histogram (at least) require if the bar width was set to 0.15?



3. The following data are the straight-line distance, x, and the shortest distance by road, y, between 20 different pairs of places in Sheffield. Of interest are the relationship between the two and the usefulness of straight-line distance as a predictor for road distance. Measurements are in miles.

x	9.5	5.0	23.0	15.2	11.4	11.8	12.1	22.0	28.2	12.1
y	10.7	6.5	29.4	17.2	18.4	19.7	16.6	29.0	40.5	14.2
x	9.8	19.0	14.6	8.3	21.6	26.5	4.8	21.7	18.0	28.0
y	11.7	25.6	16.3	9.5	28.8	31.2	6.5	25.7	26.5	33.1

The data may be summarised as follows: $\overline{x} = 16.13$, $\overline{y} = 20.86$, $s_x = 7.34$, $s_y = 9.61$, $r = 0.97$.

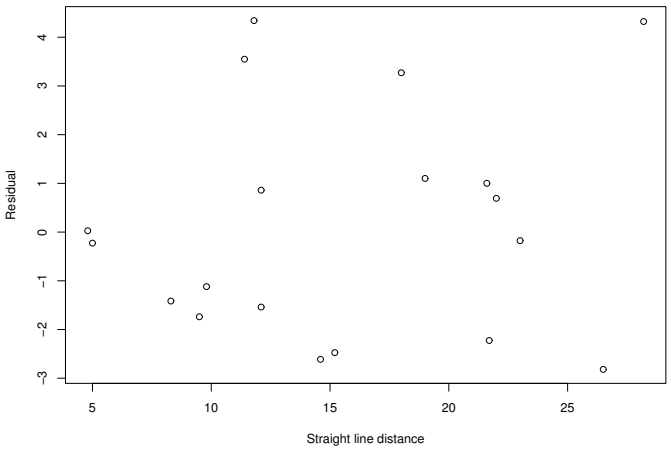
(a) Draw a scatter-plot of the data and add the regression line for predicting distance by road from straight-line distance. Interpret the result.

(b) Stating any necessary assumptions, calculate a prediction for the road distance between two places in Sheffield which are 25 miles apart by straight line distance.

(c) Stating any additional assumptions, find a range of values which would have approximately a 95% chance of containing the road distance between the same two places as in b).

(d) Below is a residual plot for the regression, which is missing the point corresponding to the last observation. Showing your working, calculate the coordinates of the missing point. Interpret the plot.

(e) What further plot would you produce to help validate the assumptions made previously? (You do not need to draw the plot.) Are there any assumptions which cannot be validated through diagnostic tools?



4. (a) The plot **shown on the next page** shows three histograms and three Gaussian quantile plots. Each histogram corresponds to one quantile plot but the order has been permutated. Match the histograms with the quantile plots and explain your working.

   (b) The point corresponding to one particular observation has been omitted from quantile plot 3. That particular observation had the value 90, and only six observations had a higher value. Showing your working, calculate the coordinates of that observation for this Gaussian quantile plot. The sample size was $n = 272$.

   (c) For the data considered in quantile plot 2, the mean and the standard deviation were $\bar{x} = 66.49$ and $s = 12.15$. Estimate the proportion of data exceeding the value 80, stating any assumptions made.

   (d) An expert has suggested that the data in quantile plot 1 actually follow an exponential distribution. The exponential distribution has height e^{-x} for $x \geqslant 0$. Show that the correct values for $q_{(i)}$ for drawing an exponential quantile plot for a sample size n are given by

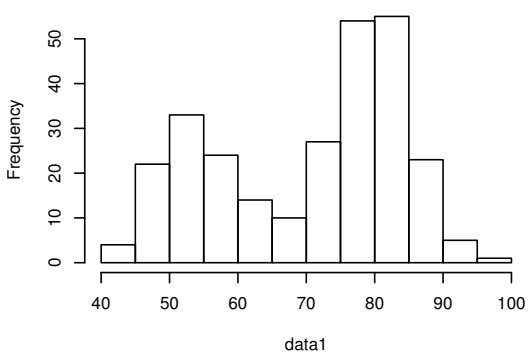
$$q_{(i)} = -\ln\left(1 - \frac{i}{n+1}\right)$$

   where ln is the natural logarithm, i.e. $\log_e$.

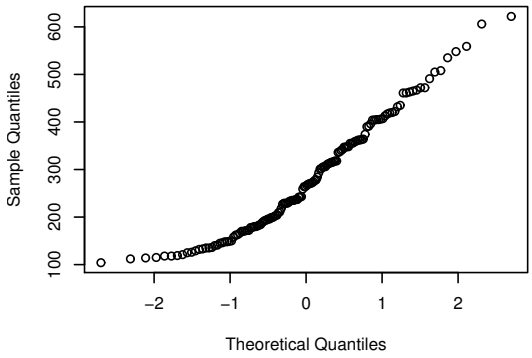
   (e) Compute four points of the quantile plot, namely for $x_{(1)} = 104, x_{(10)} = 126, x_{(50)} = 203$ and $x_{(144)} = 622$, and assess the correctness of the expert's opinion based on this preliminary analysis.

[Question 4 continues on the next page]

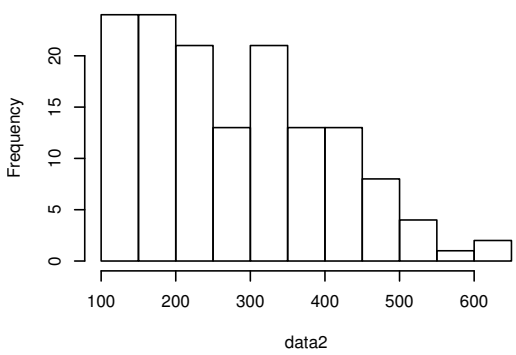
Histogram 1



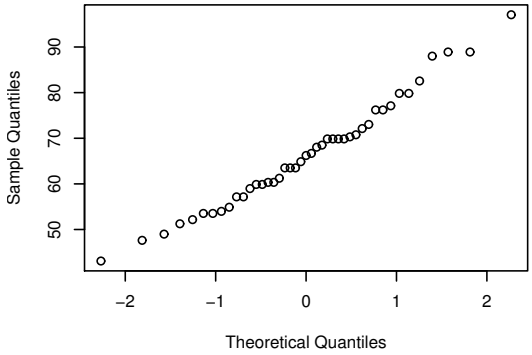
Quantile Plot 1



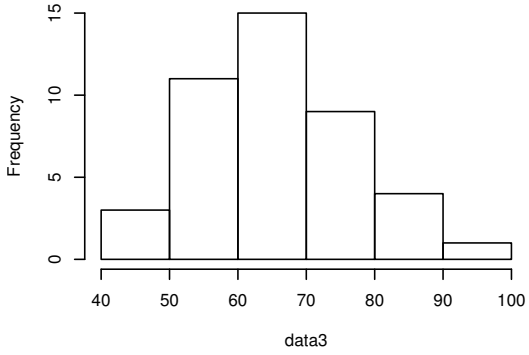
Histogram 2



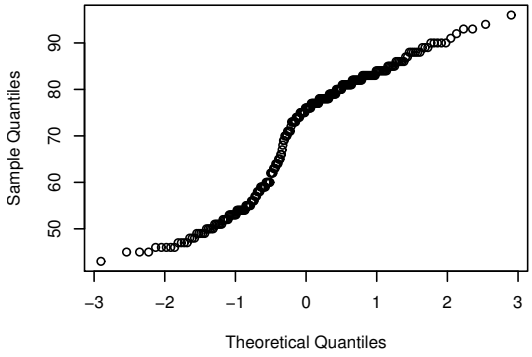
Quantile Plot 2



Histogram 3



Quantile Plot 3



5. Consider a function

$$f(x;\, c) = \begin{cases} cx^2, & x \in (0, 1) \\ 0, & x \in (-\infty, 0] \cup [1, +\infty) \end{cases}.$$

   (a) Find value $c_0 \in \mathbb{R}$ so that $f(x;\, c_0)$ can be an allowable density curve (i.e. valid probability density function)

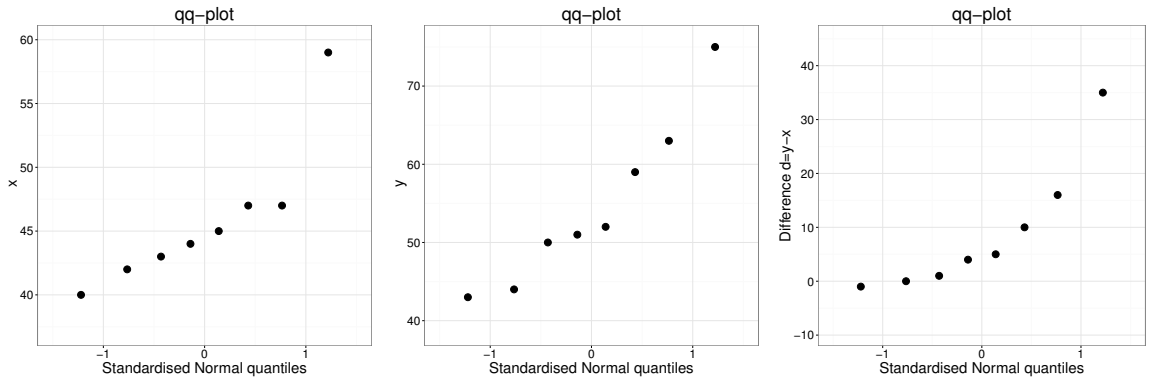
   (b) Consider the random variable X which has density equal to $f(x;\, c_0)$. Compute the following quantities: $\mathrm{E}(4X + 2)$, $\mathrm{Var}(4X + 2)$, $\Pr(0.25 < X < 0.5)$.

   (c) Consider the random variable X which has density equal to $f(x;\, c_0)$. Assume that Y follows the standard Normal distribution. Consider a random variable $Z = 2 + 4X + 3Y^2$. Compute the expected value $\mathrm{E}(Z)$.

6. A medical institute performed the following experiment: Eight smokers with chickenpox had their levels of carbon monoxide transfer measured on entry to hospital and then again after 1 week. The main question is whether one week of hospitalization has changed the carbon monoxide transfer factor. The data are presented in Table 1. We provide several qq-plots in Figure 1.

	CO transfer factor	
Patient (i)	Entry (x_i)	One week (y_i)
1	47	51
2	47	63
3	59	59
4	43	44
5	44	43
6	42	52
7	40	75
8	45	50

Table 1: CO transfer factor data



(a) Entry x_i (b) One week y_i (c) Difference between y and x ($d_i = y_i - x_i$)

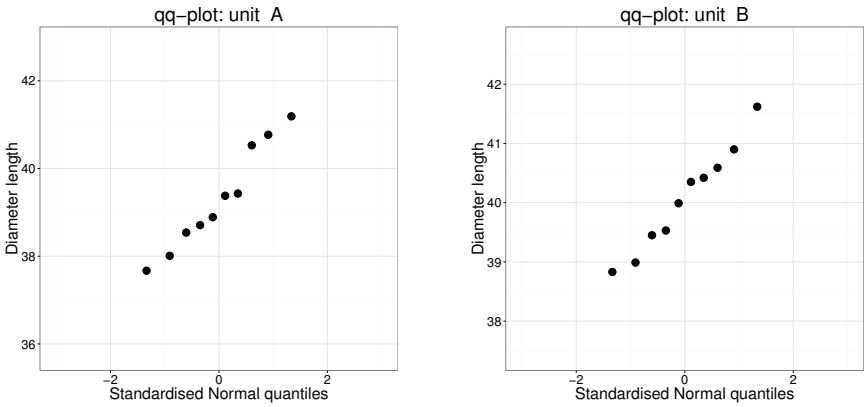
Figure 1: QQ-plots

Address the scientific question by analyzing the data, and report your results in an appropriate statistical manner. Use significance level 0.05. Justify your choice of test.

7. Two production units in a large company are set to produce baseball bats. At the end of a day's production, ten bats were randomly chosen from the production unit A and their lengths were measured. In a second experiment, ten bats were chosen from the production of unit B, and their lengths were measured. There is interest to find whether the two units produce bats of the same length. In Table 2, we present the lengths in inches. In Figure 2, we present the QQ-plots of the lengths for each unit.

Unit	Lengths				
A	38.54	38.01	39.43	41.19	37.67
	40.77	40.53	39.38	38.71	38.89
B	40.35	38.83	41.62	39.99	39.45
	40.59	39.53	38.99	40.42	40.90

Table 2



(a) Unit A (b) Unit B

Figure 2: QQ-plots

(a) Are the two samples drawn from populations with equal variances? Use significance level 0.05. Justify any assumptions you make.

(b) By constructing a confidence interval, infer whether the two units produce bats of different length on average. Use significance level 0.05. Justify any assumptions you make.

(c) Briefly, explain what a confidence interval is.

8. A random number generator (RNG) is a piece of software that generates random numbers distributed according to a given distribution. A company has designed a new RNG aiming to generate random numbers for the Binomial distribution $\text{Bin}(n = 6, p = 0.5)$. We have performed an experiment: we have generated a sample of 500 random numbers by using this new RNG. The frequency that each number appeared in the sample is presented in Table 3.

Value:	0	1	2	3	4	5	6
Observed frequency in the sample:	7	63	121	168	106	27	8

Table 3: Observed frequency of the values generated by the RNG

Test whether the new RNG generates numbers whose values are truly distributed according to $\text{Bin}(n = 6, p = 0.5)$. Use significance level 0.05. Justify any assumptions you make.