Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2018	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:	

ED01/2018

1. The Institute of Medicine (IoM) provides dietary guidelines, including Recommended Daily Allowances (RDA) for nutrients. The RDA is computed to be the intake level which is adequate to deliver correct nutrition for 97.5% of healthy adults.

   The IoM has 32 observations for intake of vitamin D, in International Units (IU), and the blood serum 25-hydroxyvitamin D (25(OH)D) levels (the chemical the body converts vitamin D into) achieved. Letting intake in IUs be x and 25(OH)D levels be y, the following summarise the data:

   $$\sum_{i=1}^{32} x_i = 18761.0 \qquad \sum_{i=1}^{32} y_i = 1828.9$$
   $$\sum_{i=1}^{32} x_i^2 = 17299775.0 \qquad \sum_{i=1}^{32} y_i^2 = 112354.4 \qquad \sum_{i=1}^{32} x_i y_i = 1195856.0$$

   (a) Calculate the mean, sample standard deviation and correlation summary statistics for both the dietary intake and serum 25(OH)D levels.

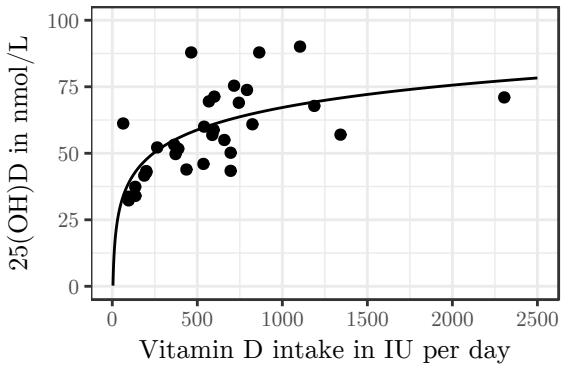
   (b) Hence calculate the least squares estimates $\hat{a}$ and $\hat{b}$ for the regression line $y = a + bx$. Also calculate the standard deviation of the residuals.

   (c) A typical vitamin D tablet contains 10 µg of vitamin D. Given that 1 IU = 0.025 µg, how much would you expect your serum 25(OH)D level to increase for each tablet you take? What, if any, assumptions are you making?

   (d) The scatterplot of the raw data (given below) led the IoM to hypothesise that the dose-response relationship may actually be logarithmic, $y = a + b \log x$. Using the fact that for this scenario:

   $$\hat{a} = -16.6,$$
   $$\hat{b} = 12.1,$$
   $$\sum \log x_i = 194.5,$$
   $$\text{and } \sum (\log x_i)^2 = 1204.6,$$

   we can calculate the RDA required to achieve 50 nmol/L of 25(OH)D as follows (which you should perform):

   i. Find the residual standard deviation for the new regression.

   ii. Find an expression (in terms of x) that gives the 25(OH)D value, y, which would have 97.5% of observations above it for an IU intake x.

   iii. Find the RDA value, in IUs, required to achieve 50 nmol/L of 25(OH)D.

   (e) A 2014 paper by Veugelers & Ekwaru notes that each observation the IoM used when calculating the RDA is in fact the average of a different randomised study. That is, each observation was the mean of a whole group of individuals whose dietary intake of vitamin D was fixed at the given x value. Give a one sentence explanation of whether the IoM correctly calculated the RDA according to their definition at the start of the question. If you say they did not, would you expect the RDA to be higher or lower and why?

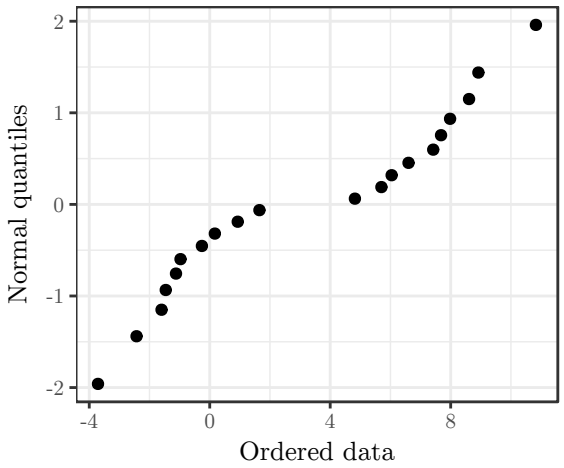
2. (a) Consider the following data which are in ascending order:

```
-3.71 -2.43 -1.60 -1.46 -1.12 -0.97 -0.26  0.17  0.93  1.65
 4.82  5.70  6.04  6.60  7.42  7.68  7.98  8.61  8.92 10.83
```

i. Draw a box plot of this data, carefully showing all calculations.

ii. List any features of the box plot that are or are not consistent with the data being Normally distributed.

iii. This is a Normal quantile plot of the above data:



Do you consider the data to be Normally distributed? Justify your answer.

iv. A friend suggests that the data may in fact be a mixture of two groups of Normal data. Construct a Normal quantile plot of just the 10 largest values, carefully showing all calculations. Would you agree with your friend that this group of 10 values by themselves are Normally distributed?

(b) Suppose we have two data sets, one with n observations and the other with m observations,

$$x_1, x_2, \ldots, x_n$$
$$y_1, y_2, \ldots, y_m$$

The sample mean of the first data set is $\bar{x}$ and the sample mean of the second data set is $\bar{y}$.

i. Is it always true that the average $(\bar{x} + \bar{y})/2$ of $\bar{x}$ and $\bar{y}$ is equal to the sample mean of the combined data set with $n + m$ observations? If no, then provide a simple counterexample. If yes, then show why.

ii. In the special case where $n = m$, is $(\bar{x} + \bar{y})/2$ equal to the sample mean of the combined data set with $n + m$ observations? If no, then provide a simple counterexample. If yes, then show why.

3. In a fictional election between two political parties in a country with 337 local voting regions (counties) and a population of tens of millions, there was a close overall result in favour of party A. County level vote data can be combined with census information to investigate any relationship between party vote and demographics.

   The data comprise the following variables, with names as used in an analysis in R:

Variable	Description
`Area`	Fictional county
`PctA`	Percentage of electorate voting for party A
`Elect`	The number of people eligible to vote in that county
`PctVote`	The percentage of eligible voters who cast a vote
`Earn`	Mean earnings in the county
`PctDeg`	Percentage of adults with a Bachelors degree or higher qualification
`Age`	Median age of the local population

Do not start this question until you have examined the plots and R output on the following page.

The fitted model shown uses all the variables, but was fitted without the data for observation 8, County H.

(a) Assess the relative value of the variables in the regression.

(b) Compute s_e.

(c) Find a 95% prediction interval for the percentage vote for party A in County H.

(d) Are there any concerning features of the diagnostic plots? What actions would you take to address these concerns and how would you justify your actions?

(e) The positive coefficient of 3.2×10^{-4} for `Earn`, implys a tendency to higher support for party A in higher income areas.

Your friend is puzzled by this conclusion, because when they fit a simple model $y = a + bx$ (where y is `PctA` and x is `Earn`) it results in $\hat{b} = -8.0 \times 10^{-4}$. In other words, a larger effect in the opposite direction to your model.

i. Why might this occur?

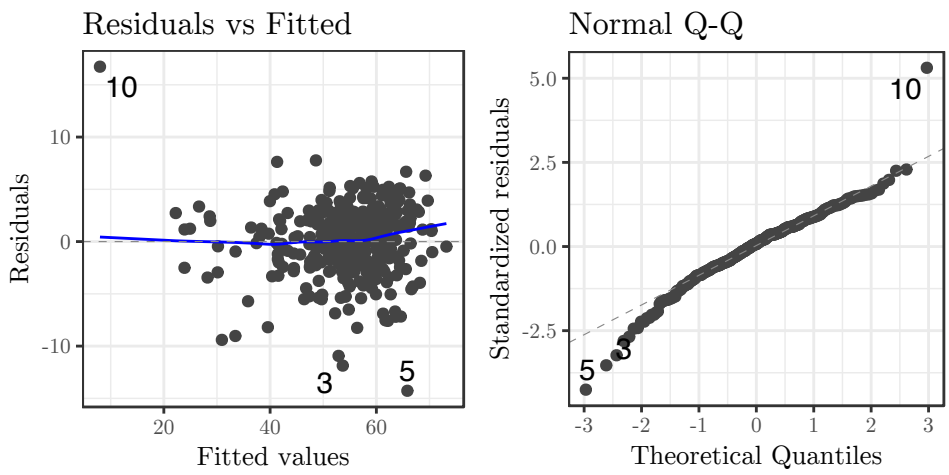
ii. What diagnostic plot or statistics should have been computed?

(f) If we take the fitted value prediction $\hat{y}_i$ for each county, then the overall predicted percentage vote for party A from this model is computed as:

$$V = \frac{\sum_{i=1}^{337} \left(\frac{\hat{y}_i}{100} \times \texttt{Elect}_i \right)}{\sum_{i=1}^{337} \texttt{Elect}_i} \times 100$$

This results in $V = 53.31173\%$ for this fictional data. If only the median age value is reduced by the *same* amount in every county, by how many years must it be reduced to cause the model to predict a hung result ($V = 50\%$)?

```
> head(election, 10)
             Area     PctA  Elect PctVote  Earn PctDeg Age
1        County A 60.28312  96760   76.34 34676   23.2  43
2        County B 70.83294  58063   72.33 27094   15.8  42
3        County C 41.80758 317924   64.08 29753   22.4  35
4        County D 50.30549 108342   75.60 37047   28.1  39
5        County E 51.56058 111647   63.54 28383   15.5  40
6        County F 51.69492 102395   78.54 35246   27.0  45
7        County G 52.92471  84164   78.16 32806   28.1  43
8        County H 57.54605 389507   68.69 28770   21.5  42
9        County I 50.54803  72487   75.69 32174   29.4  43
10       County J 24.71016   5987   73.58 63872   68.4  39

> sd(election[-8,-1])
        PctA        Elect      PctVote         Earn       PctDeg          Age
    9.816460 74658.590625     4.838368  7257.660321     7.764286     4.339923

> fitlm <- lm(PctA ~ Elect+PctVote+Earn+PctDeg+Age, election[-8,])
> summary(fitlm)

Call:
lm(formula = PctA ~ Elect + PctVote + Earn + PctDeg + Age, data = election[-8,
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-14.2732  -1.9222   0.2703   2.0665  16.7323

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.383e+01  3.295e+00  16.338  < 2e-16 ***
Elect       -9.240e-06  2.788e-06  -3.314  0.00102 **
PctVote      2.173e-01  6.563e-02   3.310  0.00104 **
Earn         3.182e-04  4.426e-05   7.188 4.41e-12 ***
PctDeg      -1.361e+00  4.076e-02 -33.404  < 2e-16 ***
Age          2.827e-01  6.967e-02   4.058 6.19e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.412 on 330 degrees of freedom
Multiple R-squared:  0.881, Adjusted R-squared:  0.8792
F-statistic: 488.5 on 5 and 330 DF,  p-value: < 2.2e-16
```



4. It is still not entirely understood what factors affect the lifetime of computer hard drives. As a result, drive reliability has been the subject of extensive studies, including an analysis in 2007 which Google performed of 100, 000 of the drives they used to store the search index. This suggested that, contrary to expectations, temperature is not such an important factor.

   The following is a toy sized (manufacturer anonymised) set of drive lifetimes in days at different operating temperatures.

Manufacturer	Temperature							
	Standard				High			
M_1	129	490	494	1131	248	972	1191	1833
M_2	18	109	197	692	31	104	129	196
M_3	103	1097	1289	1331	411	513	703	2069

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

   (b) The group standard deviations are as follows (to 1 decimal place):

Manufacturer	Temperature	
	Standard	High
M_1	416.8	653.6
M_2	301.0	68.2
M_3	577.1	772.9

   Assess whether an assumption of homogeneity is reasonable for this data set. If so, justify why and if not then determine what transformation may remedy the problem.

5. (i) (a) If X has Binomial distribution with parameters n and p, then let Y be a proportion such that $Y = X/n$. Define the margin of error for an estimate, and state this margin of error for proportions.

   (b) Show how in the case of proportions an upper bound can be found for the margin of error which does not depend on p, and comment on why this is extremely useful.

   (c) How large should n be to make sure the margin of error is less than 1%?

   (ii) A continuous random variable X has probability density function $f(x)$ given as

$$f(x) = \begin{cases} a(1-(x-c)^2) & c-1 \le x \le c+1 \\ 0 & \text{elsewhere} \end{cases}$$

   where a and c are real constants.

   (a) Show that a must be equal to $3/4$.

   (b) Give the definition of $\mathrm{E}[h(X)]$ where $h(X)$ is some function of the continuous random variable X.

   (c) Use the definition from part (b) to calculate $\mathrm{E}[(X-c)]$ directly, hence find $\mathrm{E}[X]$ and comment on its value.

   (d) Use the definition from part (b) to calculate $\mathrm{Var}[X]$ directly, and comment on whether it depends on c.

6. The following data classify 230 patients according to presence or absence of a rare but serious disease as diagnosed by a "gold standard" and by the results of a cheaper but less accurate diagnostic blood test.

		Has disease? Yes D^+	No D^-	Total
Test result positive	T^+	220	24	244
Test result negative	T^-	30	218	248
Total		250	242	492

   (a) Define and calculate the Sensitivity and Specificity of the test.

   (b) Define and calculate the False Positive and False Negative rates. Which one is arguably more of a problem in this example?

   (c) A patient from the general UK population who is selected at random is tested and receives a positive test result. In the UK it is known that about 1 in 1000 people have the disease. Calculate the probability the patient has the disease given the positive test result, $P(D^+|T^+)$. Comment on your answer.

   (d) The patient actually had the test done twice and received two positive results, represented by the event T^{++}. Assuming the test results are *conditionally independent given disease status*, implying that $P(T^{++}|D^+) = P(T^+|D^+)^2$ and $P(T^{++}|D^-) = P(T^+|D^-)^2$, calculate the probability that they have the disease after receiving two positive tests, $P(D^+|T^{++})$.

   (e) Derive a constraint on the number n of such positive test results that they would have to receive in a row, in order to ensure the probability that they did indeed have the disease is greater than p_0.

   (f) Using the results of part (e) or otherwise, find the lowest possible n for $p_0 = 0.4$.

7. (a) Two optical techniques were used to measure the wavelength of extreme ultraviolet light emitted from a particular star. Astronomers wished to find out whether the techniques differed. The following table gives the measured wavelengths of emitted light in nanometers. Analyse the data and address the scientific question (you may assume both data sets are normally distributed).

Technique A:	89.95	89.96	89.96	89.96	89.97	89.97	89.97	89.98
	89.98	89.98	90.00	90.02	90.03			
Technique B:	89.97	89.98	90.02	90.03	90.03	90.03	90.05	90.06

   (b) A sample of measurements $x_1, x_2, \ldots, x_n$ is taken and has mean $\bar{x}$ and standard deviation s_x. Suppose that each measurement is recoded as $z_i = b(x_i - a)$, where a and b are real constants with $b > 0$. Show that the mean and standard deviation for the recoded measurements are $\bar{z} = b(\bar{x} - a)$ and $s_z = bs_x$.

   (c) Explain how the results of part (b) could be used to simplify some of the calculations in part (a), and suggest appropriate values for a and b.

8. ***There is no question 8 on this paper.***