FAMILY NAME: ....
OTHER NAME(S): ....
STUDENT NUMBER: ....
SIGNATURE: ....

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

Semester 2, 2015

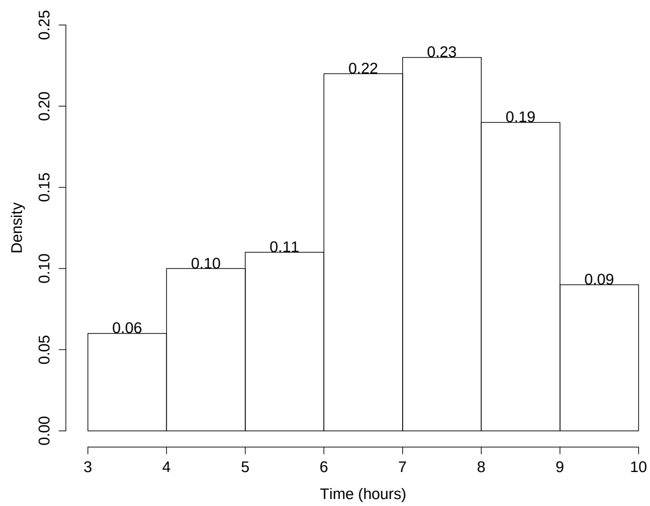
MATH2859/MATH2099

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

Probability, Statistics and Information

1. Answer in a separate book marked Question 1

The life of batteries for hand calculators produced by a particular manufacturer was tested. Each battery was put in use in a calculator that is programmed to do a continuous loop of typical calculations, and its lifetime was recorded. The lifetime (in hours) for a random sample of $n = 100$ batteries is shown by the following histogram.



a) Which class interval contains the first quartile of the observed sample of battery lifetimes? Which class interval contains the third quartile of the observed sample of battery lifetimes?

b) Are there any outliers by the $1.5 \times$ iqr-rule? Justify your answer.

c) i) For these 100 batteries, the sample mean lifetime is $\bar{x} = 6.92$ hours with a sample standard deviation $s = 1.64$ hours. Determine a 95%-confidence interval for the true mean lifetime of the batteries produced by this manufacturer.

ii) The above histogram does not look 'bell-shaped'. Explain why this is or is not a problem for building the previous confidence interval.

d) i) What is the observed proportion of batteries which last for more than 8 hours?

ii) The manufacturer claims that more than 33% of the batteries he produces last for more than 8 hours. Does the sample provide enough evidence to contradict this claim? Carry out a suitable hypothesis test at 5% level of significance. *(Write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language.)*

iii) Is the observed sample large enough for this test to be reliable? Justify.

e) In that factory, produced batteries are inspected for flaws by two quality inspectors. If a flaw is present, it will be detected by the first inspector with probability 0.9, and by the second inspector with probability 0.7. The inspectors function independently.

i) If a battery has a flaw, what is the probability that it will be found by at least one of the inspectors?

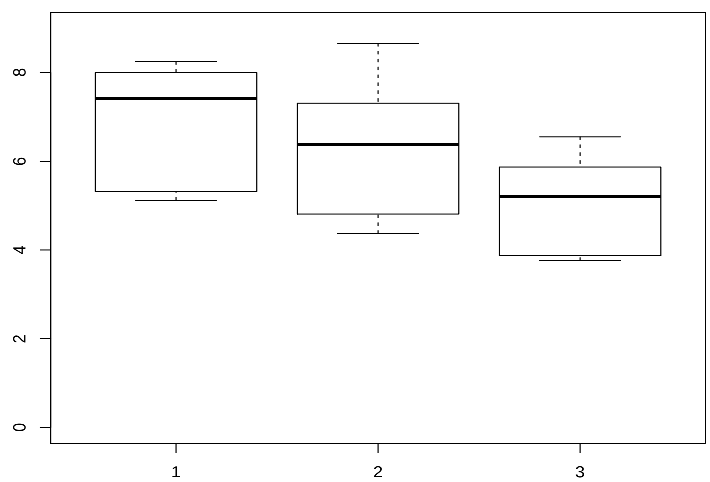
ii) Assume that both inspectors inspect every battery and that if a battery has no flaw, then neither inspector will detect a flaw. Assume also that the probability that a battery has a flaw is 0.10. If a battery is passed by both inspectors, what is the probability that it actually has a flaw?

2. Answer in a separate book marked Question 2

The table below gives the energy use of three gas ranges for six randomly selected days of a month (the units are in kilowatt-hours).

Range 1	Range 2	Range 3
8.25	8.66	6.55
5.12	4.81	3.87
5.32	4.37	3.76
8.00	6.50	5.38
6.97	6.26	5.03
7.86	7.31	5.87
$\bar{x}_1 = 6.920$	$\bar{x}_2 = 6.318$	$\bar{x}_3 = 5.077$
$s_1 = 1.387$	$s_2 = 1.586$	$s_3 = 1.103$

Comparative boxplots are given in the figure below.



a) State two features about the distribution of energy use that can be discerned from the boxplots.

b) Several assumptions need to be valid for an Analysis of Variance (ANOVA) to be an appropriate analysis. State one assumption whose validity can be assessed (at least approximately) based on the given information.

Assume from now on that these assumptions are valid.

c) An ANOVA table was partially constructed to summarise the data:

Source	df	SS	MS	F
Treatment	**(1)**	**(4)**	**(6)**	2.811
Error	**(2)**	**(5)**	1.885	
Total	**(3)**	38.88		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(6), stating how you computed the missing entries.

d) Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether there is a difference in energy use among the three different ranges. (*You can use the numerical values found in the above table; however, you are required to write the detail of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistic and p-value, conclusion in plain language.*)

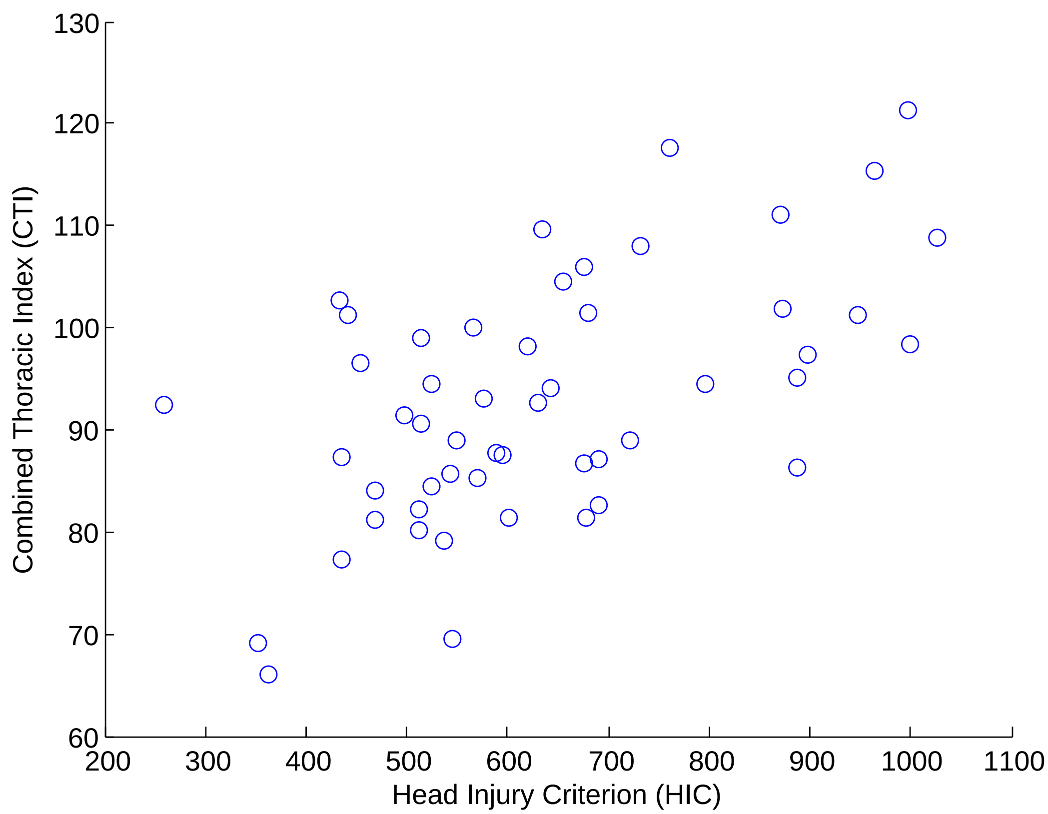
e) Using a significance level of $\alpha = 0.05$, carry out a test to see if the energy use on average for range 1 and range 2 are different. (*Write the detail of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistic and p-value, conclusion in plain language.*)

f) Two two-sample t-tests are carried out for comparing range 2 and range 3, and range 1 and range 3, with p-values 0.138 and 0.035, respectively. Does simultaneously analysing the three pairwise comparisons allow you to come to the same conclusion as the ANOVA F-test in (d), at overall significance level $\alpha = 0.05$?

3. Answer in a separate book marked Question 3

In motor vehicle crash testing, several measurements are taken on hybrid III crash test dummies such as velocity, acceleration and deflection of body parts. These measures are combined into indices such as the head injury criterion (`HIC`) and combined thoracic index (`CTI`). Due to a common cause, i.e., severity of crash, it is believed `HIC` and `CTI` are related linearly.

Data was collected for the 1998 New Car Assessment Program (NCAP) for 52 motor vehicles. A scatter plot of `HIC` and `CTI` data is shown below



The following summary statistics were obtained for `HIC`

$$\sum_{i=1}^{52} x_i = 33023 \quad \text{and} \quad s_{xx} = \sum_{i=1}^{52}(x_i - \bar{x})^2 = 1698115$$

The regression model is given by

$$CTI = \beta_0 + \beta_1 HIC + \epsilon.$$

Use the following regression output to answer the questions below.

The regression equation is

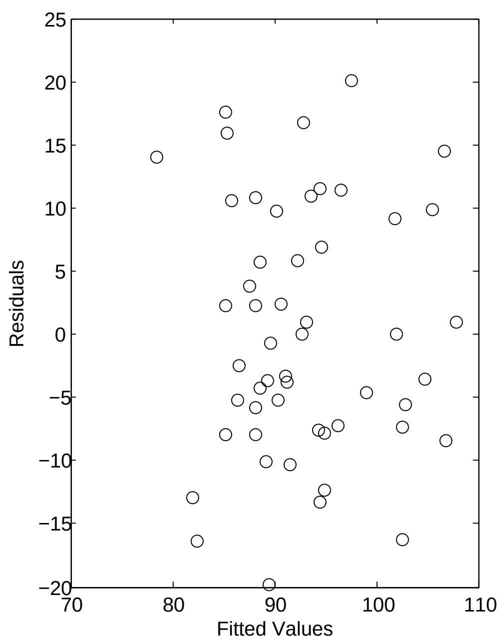
CTI = 68.5 + 0.0383*HIC

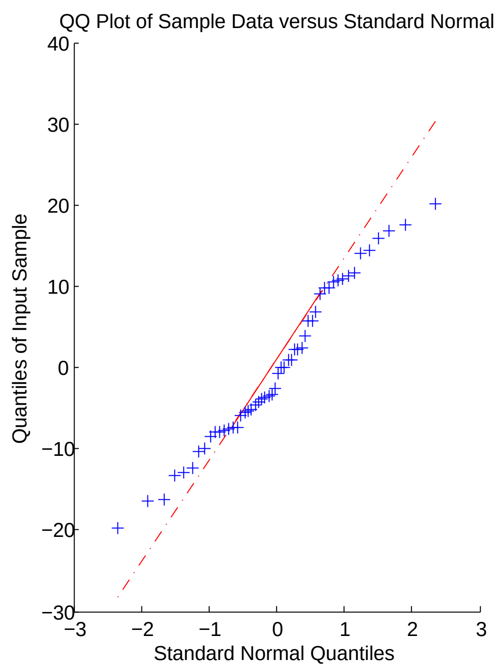
Predictor	Coef	SE Coef	T	P
Constant	68.518	5.0441	13.584	2.0897e-18
HIC	0.038321	0.0076395	5.0161	7.0309e-06

S = 9.96 R-Sq = 33.5% R-Sq(adj) = 32.1%

a) i) For the regression analysis to be valid, what is the assumed distribution of the error ϵ?

ii) Given the plots below and other given information, explain why the statistical results are at least approximately valid for this analysis?





Assume from now on that these assumptions are valid.

b) What is the expected change in CTI for a 1 unit change in HIC?

c) What proportion of variation in the response is explained by the variation in the predictor?

d) Determine the observed sample correlation coefficient between CTI and HIC.

e) Give the estimated value of σ, the standard deviation of the error term ϵ.

f) Perform a hypothesis test to determine whether the variable X is significant in this model, at the 5% level of significance. (*You can use the numerical values found in the above output, however you are required to write the details of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistics and p-value (you may present upper or lower bounds), conclusion in plain language.*)

g) Compute a 95% confidence interval for β_1.

h) Find a 99% confidence interval for the mean `CTI` when `HIC` is 1000. Explain how this interval is different to a prediction interval for `CTI` when `HIC` is 1000.