The University of Melbourne

Semester 2 Assessment — November, 1999

Department of Mathematics and Statistics

620-270 Applied Statistics

Exam duration: three hours
Reading time: fifteen minutes
This paper has ten (10) pages

Authorised materials:
Hand-held electronic calculators may be used.

Instructions to invigilators:
Statistical tables will be supplied.

Instructions to students:
There are twelve (12) questions. All questions may be attempted.
The number of marks for each question is indicated.
The total of the marks for all questions is 120; but the paper is regarded as marked out of 100 marks.
1. For the following data set

\[
\begin{array}{cccccccccccccc}
27 & 29 & 33 & 41 & 29 & 30 & 30 & 42 & 34 & 47 & 48 \\
34 & 38 & 26 & 39 & 35 & 41 & 40 & 27 & 31 & 38 \\
\end{array}
\]

(a) Find the sample median and the sample interquartile range.

(b) Find the sample mean and the sample standard deviation.  \(7\) marks

2. The following observations are salinity values for water specimens obtained from a particular region:

\[
11.3, 10.6, 5.5, 9.6, 12.2, 16.6, 9.2, 12.5, 7.9, 13.2, 11.3, 8.6, 12.7, 12.1, 9.0, 10.4, 13.7
\]

MINITAB gives the following descriptive statistics for these data:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SEMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>salinity</td>
<td>17</td>
<td>10.965</td>
<td>11.300</td>
<td>10.953</td>
<td>2.583</td>
<td>0.627</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>salinity</td>
<td>5.500</td>
<td>16.600</td>
<td>9.100</td>
<td>12.600</td>
</tr>
</tbody>
</table>

(a) Give a point estimate and find a 95% confidence interval for the mean salinity in the region.

(b) Give a point estimate and find a 95% confidence interval for the standard deviation of the salinity in the region.

(c) Test the hypothesis that the standard deviation of the salinity in the region is equal to 4 against a two-sided alternative.  \(9\) marks

The following formulae may be useful:

\[
\frac{\bar{X} - \mu}{S/\sqrt{n}} = t_{n-1} \quad \frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}
\]

3. In assessing the mineral content of a particular type of rock sample, two assay methods are available: G and S. To check the consistency of these two methods, each of ten rock samples was divided into two parts and tested by the respective methods with the following results:

<table>
<thead>
<tr>
<th>sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>method G</td>
<td>56</td>
<td>62</td>
<td>49</td>
<td>46</td>
<td>75</td>
<td>90</td>
<td>38</td>
<td>97</td>
<td>83</td>
<td>73</td>
</tr>
<tr>
<td>method S</td>
<td>58</td>
<td>66</td>
<td>51</td>
<td>46</td>
<td>74</td>
<td>93</td>
<td>40</td>
<td>99</td>
<td>88</td>
<td>80</td>
</tr>
</tbody>
</table>

It is desired to investigate the difference in means with the two methods. To do this two \(t\) statistics are proposed:

\[
t_a = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{and} \quad t_b = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n_d}}}
\]
(a) Use the MINITAB output below to compute both these $t$ statistics. Why are they so different?

(b) Which one is appropriate to use in this case? Why?

(c) Use the appropriate $t$-statistic to test for a difference in means in this case. State your conclusions clearly.

MTB > print c1-c3
Row   G   S   d
 1   56  58  -2
 2   62  66  -4
 3   49  51  -2
 4   46  46   0
 5   75  74   1
 6   90  93  -3
 7   38  40  -2
 8   97  99  -2
 9   83  88  -5
10   73  80  -7

MTB > desc c1-c3
Variable    N  Mean  Median  TrMean  StDev  SEMean
G           10  66.90  67.50  66.75  19.82   6.27
S           10  69.50  70.00  69.50  20.56   6.50
d           10  -2.600 -2.000 -2.500  2.319   0.733

Variable    Min  Max  Q1  Q3
G         38.00  97.00 48.25 84.75
S         40.00  99.00 49.75 89.25
d       -7.000  1.000 -4.250 -1.500

(10 marks)

4. A random sample of fifty observations from a bivariate normal population produced a sample correlation coefficient, $r = -0.24$.

(a) Give a rough indication of the appearance of the scatter plot.

(b) Determine an approximate 95% confidence interval for the population correlation coefficient $\rho$.

(c) Does this result provide significant evidence that the variables are not independent? Explain. (8 marks)

5. Independent random samples are obtained on $X_1$, $X_2$ and $X_3$ (representing strength measure on similar specimens — untreated, treated with additive $A$ and treated with additive $B$ respectively). It is assumed that $X_i \sim N(\mu_i, \sigma^2)$. The following table summarises the results:

<table>
<thead>
<tr>
<th></th>
<th>number of specimens tested</th>
<th>average strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>8</td>
<td>22.0</td>
</tr>
<tr>
<td>$X_2$</td>
<td>4</td>
<td>24.9</td>
</tr>
<tr>
<td>$X_3$</td>
<td>4</td>
<td>27.0</td>
</tr>
</tbody>
</table>

(a) Complete the analysis of variance table below (in which selected entries have been replaced by asterisks) and determine whether $H_0: \mu_1 = \mu_2 = \mu_3$ should be rejected using a test of size 0.05.
Analysis of Variance for strength

<table>
<thead>
<tr>
<th>source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive</td>
<td>**</td>
<td>*****</td>
<td>*****</td>
<td>****</td>
<td>****</td>
</tr>
<tr>
<td>error</td>
<td>**</td>
<td>*****</td>
<td>*****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>**</td>
<td>177.16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>21.985</td>
<td>2.626</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>24.862</td>
<td>3.508</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>27.002</td>
<td>2.616</td>
</tr>
</tbody>
</table>

Pooled StDev = 2.851

(b) Find an estimate of $\mu_3$ and a 95% confidence interval for $\mu_3$.

(c) The subcommands `tukey` and `fisher` with the `oneway` command produce the following output:

Tukey’s pairwise comparisons
Critical value = 3.73;
Family error rate = 0.050; Individual error rate = 0.021;

Intervals for (column level mean) - (row level mean)

1
1.729

2
-7.482

3
-7.458

*****

Fisher’s pairwise comparisons
Critical value = 2.16;
Family error rate = 0.116; Individual error rate = 0.050;

Intervals for (column level mean) - (row level mean)

1
0.895

2
-6.648

3
-6.495

*****

The Tukey and Fisher 95% confidence intervals for $\mu_i - \mu_j$ take the form

$$\bar{y}_i - \bar{y}_j \pm cs\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

For the Fisher intervals, $c = c_F = c_{0.975}(t_{\nu})$, and for the Tukey intervals, $c = c_T = c_{0.95}(Q_{k,\nu})/\sqrt{2}$.

i. Which is bigger: $c_F$ or $c_T$? Explain — relating your answer to family error rate and individual error rate.

ii. Why is the value $cs\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ called the least significant difference?

iii. Use the tables to find $c_F$ and $c_T$ in this case and hence find the confidence intervals for $\mu_1 - \mu_3$. (14 marks)
6. MINITAB was used to fit (i) a linear regression and (ii) a quadratic regression to the following data:

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>24</td>
<td>25</td>
<td>28</td>
<td>31</td>
<td>36</td>
<td>40</td>
<td>47</td>
<td>56</td>
</tr>
</tbody>
</table>

MTB > regr 'y' 1 'x';
SUBC> pred 10.

Predictor Coef StDev T P
Constant 15.786 2.334 6.76 0.001
x 4.4643 0.4621 9.66 0.000

S = 2.995 R-Sq = 94.0% R-Sq(adj) = 93.0%

Analysis of Variance
Source DF SS MS F P
Regression 1 837.05 837.05 93.31 0.000
Residual Error 6 53.82 8.97
Total 7 890.88

Fit StDev Fit 95.0% CI 95.0% PI
60.43 2.75 ( 53.69, 67.17) ( 50.47, 70.39) X

MTB > let c3=c1**2
MTB > regr 'y' 2 'x' 'x^2';
SUBC> pred 10 10;
SUBC> pred 10 100.

Predictor Coef StDev T P
Constant 24.0893 0.9543 25.24 0.000
x -0.5179 0.4865 -1.06 0.336
x^2 0.5535 0.0527 10.49 0.000

S = 0.6840 R-Sq = 99.7% R-Sq(adj) = 99.6%

Analysis of Variance
Source DF SS MS F P
Regression 2 888.54 444.27 949.58 0.000
Residual Error 5 2.34 0.47
Total 7 890.88

Fit StDev Fit 95.0% CI 95.0% PI
24.446 3.487 ( 15.481, 33.412) ( 15.310, 33.582)
74.268 1.462 ( 70.511, 78.025) ( 70.120, 78.416)

(a) Specify the fitted model in each case.
(b) Which model would you choose? Why? How would you determine the adequacy of your chosen model? Is it likely that a cubic or quartic polynomial would provide a better fit?
(c) You are asked to make a prediction for \( x = 10 \). Use the above MINITAB output to predict a future value of \( y \) when \( x = 10 \). In providing this prediction, what justification would you give for the model you have used, and what qualifications would you make about your prediction? (10 marks)
7. The data below were obtained in a study to calibrate four thermometers: T1, T2, T3 and T4. The study consisted of using each of the thermometers to measure the melting temperature of each of four chemical cells. The nature of the cells, and the study, were such that only one thermometer could be used for each of the (16) trials. The temperature was measured in degrees Celsius, but only the third and fourth decimal places are given, as the readings agreed up to the last two places.

![Table](image)

From these data the following, incomplete, analysis of variance table was obtained.

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>between cells</td>
<td>7</td>
<td>781.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>between thermometers</td>
<td></td>
<td>178.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>2593.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Complete the analysis of variance table.
(b) Find an estimate of and a 95% confidence interval for the error variance.
(c) Test for differences between the thermometers.
(d) Find a 95% confidence interval for the mean difference between T1 and T2.
(e) What conclusions do you reach from your analysis?

Assume an additive model with independent normally distributed errors having equal variances.

8. (a) In a randomised drug trial, 34/54 patients showed improvement with drug A, while 28/58 showed improvement with drug B. Find a 95% confidence interval for \( p_A - p_B \), the difference in improvement probabilities with the two drugs.
(b) Relate the confidence interval in (a) to the test for independence in a 2 \( \times \) 2 contingency table.

9. Choose three of the following five pairs, and for each of the three chosen pairs, discuss and explain their similarities and/or differences. For each pair, make at least three points.

(a) \( P \)-value and power;
(b) \( \bar{x} \) and \( \mu \);
(c) correlation and regression;
(d) confidence interval and prediction interval;
(e) utility and adequacy.
10. (a) Write a sentence or two on the importance of randomisation and blocking in the design of experiments.

(b) Sixteen plots are available in four blocks of four. It is required to run an experiment to compare four treatments. Explain how the treatments should be assigned to the plots.

(c) The experiment described in (b) is carried out, with results indicated in the MINITAB output below, together with an analysis of variance.

```
MTB > print 'y' 'block' 'trt' 'A' 'B'
        Row   y block trt A B
1      18.5  1   1  0  0
2      30.1  2   1  0  0
3      31.9  3   1  0  0
4      27.5  4   1  0  0
5      22.3  1   2  0  1
6      31.0  2   2  0  1
7      34.1  3   2  0  1
8      31.9  4   2  0  1
9      24.8  1   3  1  0
10     35.6  2   3  1  0
11     35.0  3   3  1  0
12     33.3  4   3  1  0
13     31.2  1   4  1  1
14     37.3  2   4  1  1
15     44.7  3   4  1  1
16     31.5  4   4  1  1
```

```
MTB > anova 'y'='block'+'trt'

Source   DF    SS      MS      F      P
block    3  326.312 108.771 17.38  0.000
trt      3  180.787  60.262  9.63  0.004
Error    9  56.311   6.257
Total    15 563.409
```

What are your conclusions?

(d) The treatments were actually two factors A and B at each of two levels — as indicated in the above output. An interactive model was fitted with the following results:

```
MTB > anova 'y' = 'block'+'A'|'B'

Source   DF    SS      MS      F      P
block    3  326.312 108.771 17.38  0.000
A        1  132.826 132.826 21.23  0.001
B        1   46.581  46.581  7.44  0.023
A*B      1   1.381   1.381  0.22  0.650
Error    9  56.311   6.257
Total    15 563.409
```

i. What are your conclusions?

ii. Give 95% confidence interval for the effect of factor A. [14 marks]
11. An experiment to investigate the effect of two methods of hanging beef carcasses (denoted by AT and TS) and the number of days for which the meat is aged (5, 10, 15, 20, 25, 30). There were six samples tested at each of the hang \times aged combinations: 72 samples in all. The dependent variable is a measure of meat quality. These data yielded the MINITAB analysis given below.

Note: the variable \((\text{aged-5})\) was used because the minimum acceptable ageing period is five days: the variable \((\text{aged-5})\) takes the values 0, 5, 10, 15, 20, 25.

(a) What models have been fitted?
(b) What does the non-significant interaction indicate?
(c) What are your conclusions?
(d) Write down the fitted model and hence predict the meat quality for a TS hung carcass aged for 14 days.

\[
\text{MTB} > \text{glm}$'mq'='(mhang)'+'(aged-5)';
\text{SUBC}> \text{covariate}(\text{aged-5});
\text{SUBC}> \text{means}'(mhang)'.
\]

\[
\begin{array}{|l|c|c|c|c|c|c|c|}
\hline
\text{Term} & \text{Coef} & \text{StDev} & \text{T} & \text{P} \\
\hline
\text{Constant} & 53.151 & 1.760 & 30.19 & 0.000 \\
\text{\textbf{(aged-5)}} & 0.3430 & 0.1163 & 2.95 & 0.004 \\
\text{\textbf{(aged-5)\text{*}\text{(mhang)}}} & -0.0618 & 0.1163 & -0.53 & 0.597 \\
\hline
\end{array}
\]

\[
\text{MTB} > \text{glm}$'mq'='(mhang)'+'(aged-5)';
\text{SUBC}> \text{covariate}(\text{aged-5});
\text{SUBC}> \text{means}'(mhang)'.
\]

\[
\begin{array}{|l|c|c|c|c|c|c|}
\hline
\text{Term} & \text{Coef} & \text{StDev} & \text{T} & \text{P} \\
\hline
\text{\textbf{Constant}} & 53.151 & 1.760 & 30.19 & 0.000 \\
\text{\textbf{(aged-5)}} & 0.3430 & 0.1163 & 2.95 & 0.004 \\
\text{\textbf{(aged-5)\text{*}\text{(mhang)}}} & -0.0618 & 0.1163 & -0.53 & 0.597 \\
\hline
\end{array}
\]

\[
\text{MTB} > \text{glm}$'mq'='(mhang)'+'(aged-5)';
\text{SUBC}> \text{covariate}(\text{aged-5});
\text{SUBC}> \text{means}'(mhang)'.
\]
12. An analysis was carried out to investigate the use of predictor variables $x_1, x_2, x_3, x_4$ and $x_5$ to predict $y$. Some of the resulting MINITAB output is given below.

**Correlations (Pearson)**

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>0.174</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>-0.373</td>
<td>0.758</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>-0.212</td>
<td>-0.753</td>
<td>-0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>-0.940</td>
<td>-0.432</td>
<td>0.195</td>
<td>0.506</td>
<td></td>
</tr>
<tr>
<td>x5</td>
<td>-0.753</td>
<td>-0.661</td>
<td>-0.038</td>
<td>0.760</td>
<td>0.916</td>
</tr>
</tbody>
</table>

The regression equation is

$$y = 158 - 4.11 x_1 + 1.52 x_2 + 0.847 x_3 - 2.82 x_4 - 0.375 x_5$$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5</td>
<td>10493.1</td>
<td>2098.6</td>
<td>390.31</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>34</td>
<td>182.8</td>
<td>5.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>10676.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(a) What criteria need to be considered in choosing the “best” model?

(b) Use this output to choose the “best” linear model for prediction of $y$ based on $x_1, x_2, x_3, x_4$ and $x_5$. Justify your choice — and comment on the utility and adequacy of your chosen model.

(c) Describe the likely outcome of the stepwise procedure to choose the “best” model.  

(8 marks)