There are 19 pages including this page.

The last two pages are a list of formulae that may be useful.

Tables of the Normal distribution can be found on page 14, the $t$ distribution can be found on page 15, the chi-square distribution can be found on page 16, and the studentized range distribution can be found on page 17.

Total marks: 80
1. Suppose $x_1, x_2, \ldots, x_n$ are $n$ observations from an exponential distribution with parameter $\lambda$. The exponential density function is

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

(a) (5 points) Find, $\hat{\lambda}$, the maximum likelihood estimate of $\lambda$.

(b) (2 points) The exponential density function is right-skewed. Suppose a normal probability plot was constructed for $n$ observations from an exponential distribution. Sketch what it would look like.
(c) (5 points) Describe how you would use the parametric bootstrap to find the sampling distribution of \( \hat{\lambda} \) from part (a).

(d) (3 points) Why is the bootstrap a useful procedure for part (c)?
2. A $100(1 - \alpha)\%$ confidence interval for a population mean is given by

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, $s$ is the sample standard deviation, and $t_{\frac{\alpha}{2}, n-1}$ is a value from the $t$ distribution with $n - 1$ degrees of freedom.

(a) (4 points) State the assumptions that lie behind the construction of this interval and describe how the assumptions can be checked.

(b) (3 points) Suppose that, based on some data collected for a study, a statistician reports that a 90% confidence interval for the mean is (4, 7). Give an explanation of what this means that a person who has not studied statistics should be able to understand.
(c) (3 points) Suppose that it is of interest whether or not the mean for the population from which the study data was collected is 3. What evidence is given by the confidence interval in part (b)?

(d) (2 points) Suppose that the statistician wanted a 90% confidence interval with a smaller margin of error than the one quoted above. How could the data collection be modified in order to obtain a confidence interval of approximately half the width?
3. A researcher wants to determine whether or not the addition of a nutrient to the diet of rats significantly affects their growth. She has 20 rats to test. She selects 10 rats to receive the nutrient by reaching into their cage and grabbing one at a time. The remaining 10 receive their feed without the nutrient additive. Otherwise, the living conditions of the two groups of rats is identical. After one month the researcher notices no significant difference in the weight gains of the two groups of rats. She concludes that the nutrient does not contribute to growth rate of rats.

(a) (3 points) Is this an observational study, sampling study, or experiment? Explain.

(b) (3 points) Comment on the validity of the researcher’s conclusion: that the nutrient does not contribute to growth rate of rats.
4. An experiment was conducted to determine whether a new drug is an improvement over a standard drug in reducing the duration of migraine headaches. 40 people who suffer from migraines were recruited, and each migraine sufferer was randomly allocated either the standard or new drug. The following summary statistics were recorded.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Sample size</th>
<th>Duration of migraine in minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>20</td>
<td>62.1 23.2</td>
</tr>
<tr>
<td>Standard</td>
<td>20</td>
<td>85.7 28.1</td>
</tr>
</tbody>
</table>

(a) (6 points) Carry out an hypothesis test to determine whether the new drug results in a lower mean duration of migraine headaches than the standard drug. Assume that the standard deviations in the two groups are the same.

(b) (3 points) Another way that the experiment could have been carried out would have been to observe each migraine sufferer for 2 headaches, and randomly allocate each drug to one of their headaches. Which design do you think is better and why?
5. The following are the results of a survey on student smoking habits obtained from a simple random sample of students in a city’s high schools.

<table>
<thead>
<tr>
<th></th>
<th>Student Smokes</th>
<th>Student Does Not Smoke</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both parents smoke</td>
<td>400</td>
<td>1380</td>
<td>1780</td>
</tr>
<tr>
<td>One parent smokes</td>
<td>416</td>
<td>1823</td>
<td>2239</td>
</tr>
<tr>
<td>Neither parent smokes</td>
<td>188</td>
<td>1168</td>
<td>1356</td>
</tr>
<tr>
<td>Total</td>
<td>1004</td>
<td>4371</td>
<td>5375</td>
</tr>
</tbody>
</table>

(a) (1 point) Let $p$ be the probability that a high school student from this city smokes. Estimate $p$.

(b) (3 points) Find a 99% confidence interval for $p$, the probability that a high school student from this city smokes, estimated in (a).

(c) (5 points) Do parents’ smoking habits influence their children’s smoking habits? Conduct a test to see if there is a relationship.
6. Three different interactive systems are to be compared with respect to their response times to an editing request. Ten sets of samples were taken for each response time, and the mean response time was recorded. Boxplots of the data and summary statistics for the three groups are given below.

<table>
<thead>
<tr>
<th>System</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.914</td>
<td>0.803</td>
<td>0.654</td>
</tr>
<tr>
<td>std dev</td>
<td>0.110</td>
<td>0.111</td>
<td>0.070</td>
</tr>
</tbody>
</table>

The following ANOVA table was obtained for testing the null hypothesis that the mean response time is the same for each system. Three values have been replaced with letters.

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>(A) 0.34041</td>
<td>0.17020</td>
<td>(B) 1.380e-05</td>
<td>***</td>
</tr>
<tr>
<td>Residuals</td>
<td>27</td>
<td>(C) 0.00977</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here is a normal probability plot of the residuals.
(a) (3 points) What information is available from the boxplots about the distributions of the response times?

(b) (3 points) Find the values of (A), (B), and (C).

(c) (7 points) Is there a significant difference between systems? Explain. If appropriate, use Tukey’s method to determine which systems differ. Use an overall significance level of 0.05.
(d) (6 points) State clearly the assumptions of the analysis of variance model and assess the model assumptions using the relevant features of the data and the R output.

(e) (3 points) System load also has an effect on response time. Here we have categorized system load as High or Low. In this comparison of systems, each system was tested at both High and Low system loads. (Five measurements were taken for each system and each load.) The following plot of the means was obtained. Describe the nature of the effects as indicated by the plot.
7. The real time needed to transfer a data set under a security protocol increases with the size of the data set. The following data were collected and a linear regression of time for the transfer (in seconds) on data set size (in bytes) was carried out.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Data set size (bytes)</th>
<th>Real time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>128</td>
<td>0.8145</td>
</tr>
<tr>
<td>2</td>
<td>256</td>
<td>0.7957</td>
</tr>
<tr>
<td>3</td>
<td>512</td>
<td>0.8002</td>
</tr>
<tr>
<td>4</td>
<td>1024</td>
<td>0.8016</td>
</tr>
<tr>
<td>5</td>
<td>2048</td>
<td>0.7698</td>
</tr>
<tr>
<td>6</td>
<td>4096</td>
<td>0.9112</td>
</tr>
<tr>
<td>7</td>
<td>8192</td>
<td>0.8306</td>
</tr>
</tbody>
</table>

Here is a plot of the data with the regression line and some of the regression output from R.

```
Call:
  lm(formula = time ~ size)

Residuals:
     1     2     3     4     5     6     7
0.012106 -0.007585 -0.004865 -0.007027 -0.045949  0.081205 -0.027885

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8015     0.0219     36.49 2.91e-07 ***
size       0.0007      0.0006     1.13  0.309

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04418 on 5 degrees of freedom
Multiple R-Squared: 0.2041,    Adjusted R-squared: 0.04488
F-statistic: 1.282 on 1 and 5 DF,  p-value: 0.3089
```
(a) (3 points) Predict the real time for data transfer for a data set of size 32768 bytes. Comment on how accurate you feel this prediction is.

(b) (1 point) What is the value of $r$, the sample correlation coefficient between $\text{time}$ and $\text{size}$?

(c) (3 points) Calculation of the regression coefficients for this problem is a minimization problem. Describe what is being minimized. (Feel free to refer back to the scatterplot.)
Some Assorted Formulae

If $X \sim \text{Bin}(n, p)$, $E(X) = np$ and $\text{Var}(X) = np(1 - p)$.

Some confidence intervals:

$$
\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
$$

$$
\bar{x} \pm t_{n-1; \alpha/2} \frac{s}{\sqrt{n}}
$$

$$
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
$$

$$
\left( \frac{(n - 1)s^2}{\chi^2_{n-1; \alpha/2}}, \frac{(n - 1)s^2}{\chi^2_{n-1; 1 - \alpha/2}} \right)
$$

$$
(\bar{y}_1 - \bar{y}_2) \pm t_{(df; \alpha/2)} \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}
$$

$$
(\bar{y}_1 - \bar{y}_2) \pm t_{(n_1 + n_2 - 2; \alpha/2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
$$

Some test statistics:

$$
z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}
$$

$$
t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
$$

$$
z_{\text{obs}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}
$$

$$
t_{\text{obs}} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}
$$

$$
t_{\text{obs}} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
$$

$$
\chi^2_{\text{obs}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
$$
Formulae continued ...

Some analysis of variance formulae:

\[ SS_{\text{Tot}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 \]

\[ SS_B = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y}_{..})^2 \]

\[ SS_W = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \]

\[ (\bar{y}_i - \bar{y}_{..}) \pm q(I, df, \alpha) \frac{s_p}{\sqrt{n}} \]

Some simple regression formulae:

\[ r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \]

\[ \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \]

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]