SB1: 3rd Assessed Practical

Week 4, HT 2020

- This practical sheet contains two sections. **Only the exercise in Section 2 is assessed** – it contributes 8.5% to your raw SB1 total mark.

- Write your answer to the exercise in Section 2 as a report.

- The deadline for submission, which is officially published in the Course Handbook (Part B synopses), is:

  3rd practical: 12 noon Monday week 8, Hilary Term 2020, at the Statistics Department reception, 24–29 St Giles’.

Your report should be clearly written. There are no marks awarded for presentation but there are marks awarded for clarity. Briefly justify any choices or assertions you make. When you make a statistical test do not just report the p-value but also report your conclusion using plain language. You should use captions for your tables and figures and include your commented R-code, preferably in an appendix.

**HINTS**

- You must hand-in a paper copy of your report by the deadline. Note that hand-in is at the Department of Statistics.

- Your report must be accompanied by a completed declaration of authorship form (on paper) saying that the report you are submitting is your own work. The declaration form will be available on the course material page: [http://www.stats.ox.ac.uk/current_students/bammath/course_material](http://www.stats.ox.ac.uk/current_students/bammath/course_material)

- Your report should probably be between 5 and 10 pages including any figures/tables. The best way to include your R code is as an appendix (not as part of the report). If you need, an extra page or two for R code is ok, but you should aim for the main report to be no more than 10 pages – part of writing clearly is writing concisely.

- Your report doesn’t need to be in LaTeX, but you can of course use LaTeX if you wish.

- There is plenty of time until the deadline. The report should only take a fraction of this time. But don’t leave writing your report to the last minute!
Section 1 - Example 1 - for practice, NOT ASSESSED

Some sample solutions are available in the solution script for this practical

http://www.stats.ox.ac.uk/~nicholls/CompStats/SB1-CompStats-Third-Prac.R

The data in the file http://www.stats.ox.ac.uk/~nicholls/CompStats/baseball.txt are a subset of data collected on professional American baseball players. Focus on the following variables:

1. salary - The yearly salary of each player (there are 59 NA's in salary - remove these rows)
2. homruns_career - The total number of homeruns a player made over their career
3. division - The division a player is in (either W or E)

(a) EDA and Smoothing

a.1 Examine the relationship between the log of outcome variable salary and explanatory variables homruns_career and division using exploratory plots.

a.2 Can we predict salaries with the predictor variable homruns_career? Perform a regression of response variable log_salary on the variable homruns_career. Plot the regression lines onto the data and comment on the fit. Give an interpretation of the model, keeping in mind that the salaries are reported on a log-scale. Comment on the effect of extreme observations in homruns_career.

a.3 Try a smoothing procedure, for example a Kernel smoother ksmooth() or local polynomial regression lqpoly(). Plot the fit for various smoothing parameters and discuss what you find. Investigate also lowess() and dplin().

a.4 Could anything be done to improve the fit?

(b) Rank sum test

b.1 Are the salaries similar in the two divisions? Examine whether the distributions of player salaries are the same in both leagues. Use the Wilcoxon test with the normal approximation and calculate by hand the value of the test statistic, its expected value and variance under the null hypothesis of no difference and give the test result.

b.2 The data contain ties. Explain how to simulate the distribution of the test statistic under the null. Implement this procedure and check that the significance levels you computed in the previous question are robust to the effect of ties.

b.3 Test your result by implementing the Wilcoxon test in R and give a point estimate (using the Lehmann-Hodges estimator) and a confidence interval for the difference in median salary between the two leagues. State any assumptions.

(c) Rank sign test

c.1 Consider the possibility that walks and runs have the same median. Why should we treat these data as paired?

C.2 Find a confidence interval for the difference in medians and use it to test for equal medians at 95%. What do you conclude? State any assumptions.
An 18×8 foot area of field planted with maize was split into 3×1 foot rectangles (so 48 rectangles in all). The contingency table below records the \( n = 48 \) live-weight measurements of Japanese beetle larvae found in the top foot of soil of each rectangle. The plough ran down the columns, and the seeds were planted in sequence down the columns.

<table>
<thead>
<tr>
<th>Northing</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>1</td>
<td>10.3</td>
<td>14.6</td>
<td>10.4</td>
<td>6.0</td>
<td>2.6</td>
<td>4.9</td>
<td>2.2</td>
<td>11.6</td>
</tr>
<tr>
<td>2</td>
<td>8.5</td>
<td>10.6</td>
<td>10.1</td>
<td>2.9</td>
<td>1.6</td>
<td>7.2</td>
<td>5.5</td>
<td>7.1</td>
</tr>
<tr>
<td>3</td>
<td>3.6</td>
<td>6.3</td>
<td>9.0</td>
<td>8.7</td>
<td>6.4</td>
<td>4.6</td>
<td>8.4</td>
<td>9.7</td>
</tr>
<tr>
<td>4</td>
<td>9.3</td>
<td>10.5</td>
<td>6.3</td>
<td>5.6</td>
<td>5.7</td>
<td>6.8</td>
<td>8.6</td>
<td>18.3</td>
</tr>
<tr>
<td>5</td>
<td>9.8</td>
<td>11.6</td>
<td>3.5</td>
<td>1.3</td>
<td>5.7</td>
<td>3.7</td>
<td>2.8</td>
<td>12.6</td>
</tr>
<tr>
<td>6</td>
<td>3.8</td>
<td>5.4</td>
<td>7.0</td>
<td>5.3</td>
<td>3.2</td>
<td>7.7</td>
<td>16.2</td>
<td>19.3</td>
</tr>
</tbody>
</table>

The data are available for download as a dataframe in the file

www.stats.ox.ac.uk/~nicholls/CompStats/blw.txt

How do the Northing and Easting values of the rectangular plot influence the distribution of beetle larvae weight measurements in that plot? Answer this question as clearly as you can by (selective) application of the permutation tests, Monte-Carlo tests, non-parametric rank-tests and smoothing methods we studied in the course and in Exercise 1 of this practical.
Section 2 ASSESSED EXERCISE

This problem asks you to carry out a computer experiment. You gather the data using simulation on the computer. Some R-scripts are available to do the simulation. Write a short report explaining the problem, clarifying and formalising the question and outlining and justifying your choice of statistical methods and presenting your results and conclusions. Draw attention to any assumptions you make in the course of selecting a test and any strengths or weakness of your methods. The total length of your answers should probably be between 5 and 10 pages including any figures/tables. There is no need to reproduce in your report the R-code you are given, but do paste your own code into an appendix.

The R-code in

http://www.stats.ox.ac.uk/~nicholls/CompStats/SB1-CompStats-Third-Prac-20.R

gives three main R-functions, Finv(u) is the inverse CDF of the density

\[ f(x) = \frac{1}{\pi} \frac{\sqrt{2/x}}{1 + x^2}, \quad x > 0 \]

This can be used to simulate \( Z \sim f \) using \( u = \text{runif}(1); Z = \text{Finv}(u) \) (method of inversion).

The functions \( m1(T, x0) \) and \( m2(T, x0) \) (with \( T \) a positive integer and \( x0 \) a positive real) simulate the same distribution, but are less reliable especially if \( T \) is small. They use different MCMC methods, but both simulate a Markov chain with \( T \) steps starting at \( x0 \) and return the \( T \)th state of the chain. This converges in distribution to \( f \), so when \( T \) is large, this return value should be distributed according to \( f \). As far as you are concerned \( m1() \) and \( m2() \) can otherwise be treated as black boxes that output unreliable samples - you do not need to understand how these functions work!

Some code that generates two data vectors \( X = (X_1, ..., X_K) \) (using \( m1() \)) and \( Y = (Y_1, ..., Y_K) \) (using \( m2() \)) is given at the end of the R-script. For each \( k = 1, ..., K \) a value \( x0[k] \sim \text{rt}(1, df = 1) \) is simulated from a \( t \)-distribution with 1 degree of freedom. The same value of \( x0 \) is input to the two samplers (they start in the same state) and we set \( X[k] = m1(T, x0[k]); Y[k] = m2(T, x0[k]) \). It follows that \( X[k] \) and \( Y[k] \) are correlated but conditionally independent given \( x0[k] \). If \( T \) is small this correlation will be strong.

1. In checking for bugs and other bias we often write different samplers and check they give the same distribution. Use non-parametric tests to check for \( X \sim Y \) at \( T = 10 \). You can try varying \( K \) but large values will give long run-times.

2. Suppose instead of initialising with \( x0[k] \sim \text{rt}(1, df = 1) \) we initialise with \( x0[k] = 10 \) for \( k = 1, ..., K \). How does this change your reasoning and conclusions?

3. How big does \( T \) have to be so we dont reject \( X \sim Y \) at \( T = 10000 \)? Can you say anything about the power of this test? It will depend on \( T \) and \( K \).

4. In this particular example we can simulate \( Z \sim f \) exactly using inversion. Do the Z-samples offer any further opportunity for testing?

In this practical we are interested in non-parametric tests and smoothing. Some of the questions of interest could be answered by other means. Marks will be awarded for effective use of the methods taught in the course.

NOTE: you may occasionally encounter numerical errors and warning messages from \( m1, m2 \) in very long runs. These do to some extent limit the length of the runs (ie \( K \) and \( T \)) we can consider.