BCB744 (BioStats): Summative Task 2

AJ Smit *University of the Western Cape*

Instructions

1. Produce full R scripts, including the meta-information at the top (name, date, purpose, etc.). Include comments explaining the purpose of the various tests/sections as necessary.
2. At the onset, *assume* that all assumptions are met (they might not be!). For each question, i) state the name of the parametric statistical test selected and write a sentence or two to justify your choice, and ii) write down the appropriate H_0.
3. Test the assumptions. For each statistical test (as per 1.), state the appropriate assumptions to be tested, mention how you will test the assumptions (incl. writing out the H_0 if necessary, or explaining the principle behind a graphical method), and explain the findings. Given the outcome of the test of assumptions, is your choice of test selected in (1.) still correct? If some/all the assumption tests are violated, how will you proceed?
4. Proceed with the analysis, and explain the findings in the light of the hypothesis tests stated at the beginning. ANOVAs, correlations, and regressions will require graphical support; this is not necessary for t-tests.
5. Pay attention to formatting. 10% will be allocated to the appearance of the script, including considerations of aspects of the tidiness of the file, the use of appropriate headings, and adherence to code conventions (e.g. spacing etc.).

1 The wheat yield data

Please see the file '**fertiliser_crop_data.csv**' for this dataset. The data represent an experiment designed to test whether or not fertiliser type and the density of planting have an effect on the yield of wheat. The dataset contains the following variables:

- Final yield (kg per acre)
- Type of fertiliser (fertiliser type A, B, or C)
- Planting density (1 = low density, 2 = high density)
- Block in the field (north, east, south, west)

Question 1 Do fertiliser type and planting density affect the yield of wheat? If so, which is the best density to plant wheat at, and which fertiliser produces the best yield?

Question 2 Does it matter if the wheat is planted in portions of the experimental fields that face north, east, south, or west?

2 The shells data

See the '**shells.csv**' file. This dataset contains measurements of shell widths and lengths of the left and right valves of two species of mussels, *Aulacomya* sp. and *Choromytilus* sp. Length and width measurements are presented in mm.

Question 3 Which species of mussel is the i) widest and ii) longest?

Question 4 Within each species of mussel, are the four different measurements correlated with each other?

Question 5 Considering *Aulacomya* sp. only, use a linear regression to predict the length of the left valve when the width of the left valve is 15 and 17 mm.

3 The health data

These data are in '**health.csv**'. Inside the file are several columns, but the ones that are relevant to this question are:

- 'Sex', which is the gender of the individuals assessed
- 'Substance', indicating the kind of drug abused by the individuals in question
- 'Mental_score', which is the outcome of a test designed to test the cognitive ability of individuals

Question 6 Do males and females suffer the same cognitive impairments if they abuse cocaine, alcohol, or heroin?

Question 7 Which drug is worst in terms of affecting the user's mental health?

4 The air quality data

Package **datasets**, dataset `airquality`. These are daily air quality measurements in New York, May to September 1973. See the help file for details.

Question 8 Which two of the four response variables are best correlated with each other?

Question 9 Provide a detailed correlation analysis for those two variables.

5 The crickets data

The file '**crickets.csv**' contains data for some crickets whose chirp rate was measured at several temperatures. The temperature was measured in °F, but please make sure you do all the calculations using °C instead.

Question 10 Does the chirp rate of the crickets depend on the temperature?

Question 11 Provide an equation that quantifies this relationship.

6 The SST data

The file '**SST.csv**' contains sea surface temperatures for Port Nolloth and Muizenberg in °C. The data are from 1 January 2010 to 31 December 2011.

Question 12 Do the temperatures differ between the two places?

Question 13 For each of the two sites, which month has the i) lowest and ii) highest temperature?

Question 14 For each of the two sites, is the winter temperature colder than the summer temperature?

Question 15 Same as Question 14, but use 95% confidence intervals to approach this problem (and provide the supporting graphs).

Hint: The **lubridate** package (and others) offers convenient ways to work with time series (i.e. in this case coding a variable for month).

That's all, Folks!