BIO9029 Professional and Research Skills in the Life and Environmental Sciences

Data analysis test

Your name: Your student ID:

Instructions

Answer all the questions. The points awarded for each question are indicated within the questions. A maximum of 60 points can be awarded and the points will be then converted into the final mark in %.

Please write the answers just below the question. Also, paste any figure that the question may ask to generate. Some answers require you to include the code, please write the code in `Courier New` font. Do not paste screenshots as these won't yield any marks. Do not paste code where this is not explicitly requested. Perform the analysis using R. Any code using a language other than R won't yield marks.

The answers that require calculated values are awarded a maximum of 2 points. Values that are correct but poorly presented are awarded only 1 point. The answers that require figures are penalised if the figure is poorly presented.

The test is divided into two parts, each dealing with a different dataset. The datasets can be downloaded from Blackboard, BIO9029 module site, Assessments tab, Assessment 3a – datasets. Make sure to download the two datasets with your name: YOUR NAME (YOUR ID) weights.csv and YOUR NAME (YOUR ID) supplement.csv.

Complete both parts within a maximum of 1 hour and 30 minutes. Submit the test via the Turnitin link provided and also email your work to slssubmissions@lincoln.ac.uk at the same time.

General Instructions to Candidates

1. In sitting this test you agree to **comply** with the University of Lincoln Code of Conduct in Examinations.
2. You **must** submit your answers as a PDF or MS Word Document to Turnitin on Blackboard **before** the submission time: failure to do so will be classified as misconduct in examinations.
   We strongly recommend you submit 15 minutes prior to the deadline.
3. You **must** also send a copy of your work to the slssubmissions@lincoln.ac.uk at the same time. You must place the Module Code and your Student Id in the Subject Field of the Mail.
4. This assessment is an **open resource format**: you may use online resources, lecture and seminar notes, text books and journals.
5. **No collaboration or interaction** with other candidates or individuals using any means of communication or device is permitted during online tests
6. All work will be **subject to plagiarism and academic integrity checks**. In submitting your assessment you are claiming that it is your own original work; if standard checks suggest otherwise, Academic Misconduct Regulations will be applied.

7. **The duration of the online test will vary for those students with Personal Academic Study Support (PASS)**. Extensions do not apply, but Extenuating Circumstances can be applied for in the normal way.

Part 1 - descriptive statistics and statistical tests

The dataset in the file "YOUR NAME (YOUR ID) weights.csv" contains about 1000 records of the weight in grams of rodents in a laboratory that studies rodents' diets. In the population under study, a mutant phenotype is present which seems to be slightly heavier than the wild type. For each individual recorded, the dataset reports:

- the weight in grams = "Weight"
- the phenotype = "Phenotype":
  - "Wild Type" or
  - "Mutant"

The primary aim of the analysis is to find out if the weight of the mutant is significantly higher than the wild type.

As the laboratory is interested in knowing if the possible weight gain is linked to carotenes in the diet, a secondary aim of the research is to find out if a diet supplemented with carotenes can also influence weight. To test this, the rodents were fed on 4 diets with 4 different levels of carotenes supplied: "None", "Low", "Normal" and "Rich". The level of carotenes is recorded under the variable "Carotenes". You will have to test if any of the levels of the carotene is associated with a significantly different weight.

Import the "YOUR NAME (YOUR ID) weights.csv" data, assign the name `dataset` to the resulting data frame and answer the questions below.

1. Define an ordered subset of `dataset` named `sub` using the following:

   ```
   sub<-head(dataset[order(dataset$Weight,decreasing=TRUE),],20)
   ```

   The subset `sub` includes the 20 heaviest rodents from `dataset`.

   a) Write a one-line function `tapply` applied to the subset `sub` to determine the number of records for each level of carotenes in the diet. (1 point)
   b) How many rodents in the subset `sub` have low levels of carotenes? (2 points)
   c) How many rodents have normal levels? (2 points)

2. What is the mean weight for each phenotype for the entire dataset?
   a) Write a one-line function `tapply` applied to the entire `dataset` to determine the mean weight of each phenotype. (1 point)
   b) What is the mean weight of the wild-type phenotype? Round to 2 decimal places. (2 points)

   c) What is the mean weight of the mutant phenotype? Round to 2 decimal places. (2 points)

3. Make a histogram of weight for the entire dataset using the function `hist`. Do not include any main title and give a sensible title to the x-axis, inclusive of the units. Report the one-line function and paste the resulting figure. (5 points)

4. Make a boxplot of the weight by phenotype. Include the outliers and give a sensible title to the y-axis, inclusive of the units. Report the one-line function and paste the resulting figure. (5 points)

5. Visually check if the weight of the dataset is normally distributed.
   a) Write a one-line code to make a Q-Q plot of the weight for the entire dataset. (1 point)
   b) Write a one-line code to add a red line to highlight the trend. (1 point)
   c) Paste the resulting figure. (2 points)
   d) Write a short sentence to comment on whether or not the plot suggests that the data are normally distributed and explain why. (1 point)

6. Test if the difference in the mean of the wild-type and mutant groups is significant by performing an unpaired two-samples t-test, including checking any prerequisites.
   a) A prerequisite of the t-test is the normality of the distribution of each group. Write the name of the statistical test used to verify this. (1 point)
   b) Write a one-line command using `tapply` to test the normality of both groups. (1 point)
   c) What is the p-value of the normality test for the wild-type group? Round to 4 decimal places. (2 points)
   d) What is the p-value of the normality test for the mutant group? Round to 4 decimal places. (2 points)
   e) Write a short sentence to explain how you deduce that the weight is normally distributed in both groups. (1 point)
   f) Apply the F-test to test if the variances of the two groups are equal (which is another prerequisite for the t-test) using the function `var.test`. Write a one-line command to perform the F-test. (1 point)
   g) Write the p-value resulting from the F-test. Round to 4 decimal places. (2 points)
   h) Having checked the prerequisite, write a one-line command to perform the unpaired two-samples t-test. (1 point)
   i) Is the mean weight of mutants significantly different from the wild-type and why? Write a short sentence to explain. (1 point)

7. Test if any carotene level is associated with a significantly different mean weight using suitable statistical tests. You can assume without testing that the weight is normally distributed with similar variances for all subsets grouped by carotenes level.
   a) Name the suitable test for comparing the mean of carotene level groups. (1 point)
   b) Write a one-line code to perform the test above. (1 point)
   c) What is the resulting p-value of the test? Round to 3 decimal places. (2 points)
   d) Write a short sentence to interpret the results of the test and conclude if there are any carotene level groups that have a significantly different mean weight. (1 point)
   e) Name a suitable post-hoc test to perform multiple pairwise comparisons of the carotene level groups. (1 point)
   f) Represent the multiple pairwise comparisons visually and paste the plot. (2 points)

Part 2 – curve fitting

The same laboratory above is testing the effect of a food supplement on the weight in grams of rodents. The hypothesis is that the mutant rodents respond better to low doses of the supplement, which ultimately leads to weight gain.

Different doses of the supplement have been given to small groups of the wild type and mutant rodents and the average weight of each dose group has been recorded in the file "YOUR NAME (YOUR ID) supplement.csv". The variables are:

"Dose" = the dose of the supplement in ng/g

"WT" = the average weight in grams for each dose group of the wild-type rodents

"Mutant" = the average weight in grams for each dose group of the mutant rodents

The data hints at higher doses of the supplement yielding higher weights and the trend is sigmoidal.

Import the file, assign the name `supplement` to the resulting data frame and answer the question below.

8. The weight data in the two series (wild-type and mutant) broadly follow the trend of the 4-parameters sigmoid equation below:

```
y<-function(x,a,b,c,d){d+(a-d)/(1+(x/c)^b)}
```

a) Write a one-line command to fit the wild-type data to the `y` function. (1 point)
b) Write a one-line command to fit the mutant data to the `y` function. (1 point)
c) Write the best estimates for the parameters a, b, c and d of the wild-type data. Round to 2 decimal places. (2 points)
d) Write the best estimates for the parameters a, b, c and d of the mutant data. Round to 2 decimal places. (2 points)
e) Which of the 4 parameters can be used to assess if the mutants respond better to low doses of the supplement? (1 point)
f) Compute and report the confidence intervals of the parameter above for both the wild-type and the mutant. Round to 2 decimal places. (2 points)
g) Based on the calculation above, can you confirm if the mutants respond better on average to low doses of the supplement? Explain your answer in a short sentence. (1 point)
h) Plot weight versus supplement dose for both the mutant and the wild-type rodents. Include the curves of best fit. Use different colours, lines and point shapes to distinguish the two series well. No legend is required but please make use of accurate axis labels, inclusive of units. (5 points)