BB839 Exam 2019

27 November, 2019

Introduction

This exam includes four questions that test different aspects of your learning during this course: data wrangling, data visualisation and statistics. Each question is broken down into a number sub questions.

You must answer questions 1-3; you must answer EITHER question 4 OR 5

Your work should be handed in as a single PDF. Each question should be clearly indicated

For each question you must provide the R code you used to answer the question. The code should include comments to explain what you are doing. The code should be provided as text using a fixed-width font such as `Courier`. The rest of your answers should be in another font (e.g. Times New Roman, Cambria, Ariel). You can use the Microsoft Word template provided alongside these questions as guidance.

- Plots and tables should have captions.
- Plots should be produced using `ggplot`.
- Axis labels are important.
- Reporting of any statistics should be appropriate to the type of analysis you have done.
- Reporting of methods and results should be written in the style of a scientific paper.

If you don't understand what is required for any of these questions you are encouraged to ask for help.

Hand in deadline is Friday 10th January 2020 at 12:00 CET

You MUST submit your work via Blackboard (not email!)

1) Data wrangling life history data (10 points)

The Anage database (`anage_data.csv`) is a large collection of data on the life history of animals. It includes information on life span, body size, generation time etc. for species from a wide range of taxonomic groups. In this question you will be using this data to produce a graph and some summary information. You may need to `filter`, `select`, `mutate` or otherwise manipulate the data before you use it. You may also like to rename some of the data columns for ease of use.

The length of gestation in humans is 9 months, but it varies across mammals. Gestation time (`Gestation.Incubation..days.`) is positively associated with birth weight (`Birth.weight..g.`), which makes sense since larger bodies take longer to build. It is interesting to consider if there are differences in this relationship between rodents (Order Rodentia) and their flying cousins - bats (Order Chiroptera).

a) Produce a graph showing the relationship between between the log transformed birth weight (x-axis) and log-transformed gestation length in days (y-axis) for bats and rodents.

b) Produce a table showing the minimum, maximum, mean and median gestation times (in days) for rodents and bats.

c) Which are the species (latin names) with the minimum and maximum gestation lengths in Rodentia and Chiroptera?

d) Explain why an ordinary least squares regression would not be the best approach to analyse this relationship.

2) Industrial melanism (10 points)

Industrial melanism is a famous example of an evolutionary effect where dark pigmentation (melanism) evolves via natural selection where the environment is polluted with soot deposits. Darker individuals have a higher fitness in areas where their camouflage matches the polluted background surfaces better.

Some biologists have claimed that the knot grass moth (*Acronicta rumicis*) shows industrial melanism. You have collected data on the percentage of moths collected in light traps that are of the dark morph (`melanism.csv`). Your expectation is that you will find a higher percentage of dark morphs in the city where there is more pollution.

a) plot the data (e.g. with a box plot)

b) carry out a randomisation test to determine if there is a significant difference in the frequency of the dark morph between city and countryside. Write (i) a brief method description and (ii) a summary of the results.

3) Power in a planned experiment (10 points)

You are planning an experiment testing how the behaviour of small fish is affected by exposure to danger from predators. There are two treatments: (i) "exposed to predator cues" (from video footage of large predatory fish) and (ii) a "control" treatment, which is a safe predator-free environment. You would like to be able to detect a 20% reduction in swimming distance.

You record behaviour as the distance in meters travelled by fish as they swim around their tank within a 5 minute observation window. You hypothesise that when exposed to predator cues the fish will cover less distance and will tend to hide among the rocks and plants in the tank more. You have 5 fish tanks available for your study.

You have some preliminary data from a pilot study (`fishFear.csv`) which shows the distances moved by several individuals in a single pilot study.

a) Summarise the pilot study data to obtain mean and standard deviation.

b) Conduct a power analysis based on the pilot study data to estimate the number of samples required to carry out your experiment with 80% power. Describe the results of this power analysis.

c) Describe how you might carry out your study, given your findings and the facilities available to you.

4) The Titanic (20 points)

The `titanic.csv` data set includes a range of data on a subset of the passengers on the Titanic that sank in 1920 after hitting an iceberg. There were approximately 2200 passengers and crew on board and >1400 of these people died.

Your task is to analyse the data to find out how passenger class (`Pclass`) and gender (`Sex`) influenced survival probability. Survival is indicated with the numeric values 0 (died) and 1 (survived). Passenger class has values of 1, 2 or 3 and Sex is recorded as `male` or `female`.

a) Produce an appropriate graph of the raw data that illustrates survival differences among passenger classes and genders.

b) Fit a suitable statistical model to estimate survival probability among passenger class and genders. Describe the method and then summarise the results produced by the model.

c) Make a plot showing the model's estimates and their 95% confidence intervals.

5) Elephant poaching (20 points)

Illegal poaching activity is a major problem for African elephant conservation. You are provided with some data from some nature reserves across southern Africa (`elephantPoaching.csv`).

The data includes the number of elephants killed in 2016 (`nkilled`) and the number of scouts (the guards that protect the elephants) (`scoutPerKm`). A philanthropist billionnaire has been working in the area to provide additional tools to some reserves such as drones, high-tech communications and additional vehicles. This funding is recorded in the dataset column `funding` as `funded` (the ordinary parks with no extra funding are recorded as `normal`).

Use an appropriate statistical model to explore the relationship between elephants killed and the amount of cover provided by guards. Does this relationship differ depending on how well equipped the guards are?

a) Plot the data to show the relationship between the the number of scouts per unit area of park and the number of elephants killed. Colour code the points by whether the park management received extra funding or not.

b) Fit a suitable statistical model to estimate the statistical relationship between guards per km, funding, and elephant deaths. Describe the method and then summarise the results produced by the model.

c) Produce a plot that shows (in addition to the raw data points) the fitted values produced by your model and the uncertainty in those estimates.