Exercises

Exercise 5: Basic statistics in R

Although this short course is primarily focussed on introducing you to R, it wouldn't be complete if we didn't have a quick peek at some of R's statistical roots. Having said that, this will be a very brief overview with very little in the way of theory so don't worry if you get a little lost - this is just a taster and something you will build on in subsequent courses. During this exercise you will practice using the linear modelling framework in R to analyse datasets with various structures. You'll learn how to fit linear models, validate these models and finally interpret models. For more information about linear models (and other statistical approaches) see Chapter 6.

1. Download the datafile *'prawnGR.CSV'* from the **Data** link and save it to the `data` directory. Import these data into R and assign to a variable with an appropriate name. These data were collected from an experiment to investigate the difference in growth rate of the giant tiger prawn (*Penaeus monodon*) fed either an artificial or natural diet. Have a quick look at the structure of this dataset and plot the growth rate versus the diet using an appropriate plot. How many observations are there in each diet treatment?

2. You want to compare the difference in growth rate between the two diets using a two sample t-test. Before you conduct the test, you need to assess the normality and equal variance assumptions. Use the function `shapiro.test()` to assess normality of growth rate for each diet separately (Hint: use your indexing skills to extract the growth rate for each diet `GRate[diet=='Natural']` first). Use the function `var.test()` to test for equal variance (see `?var.test` for more information or Section 6.1 of the book for more details). Are your data normally distributed and have equal variances?

3. Conduct a two sample t-test using the `t.test()` function (Section 6.1of the book). Use the argument `var.equal = TRUE` to perform the t-test assuming equal variances. What is the null hypothesis you want to test? Do you reject or fail to reject the null hypothesis? What is the value of the t statistic, degrees of freedom and p value? How would you summarise these summary statistics in a report?

4. An alternative (but equivalent) way to compare the mean growth rate between diets is to use a linear model. Use the `lm()` function to fit a linear model with `GRate` as the response variable and `diet` as an explanatory variable (see Section 6.3 for a very brief introduction to linear modelling). Assign (`<-`) the results of the linear model to a variable with an appropriate name (i.e. `growth.lm`).

5. Produce an ANOVA table for the fitted model using the `anova()` function i.e. `anova(growth.lm)`. Compare the ANOVA p value to the p value obtained using a t-test. What do you notice? What is the value of the F statistics and degrees of freedom? How would you summarise these results in a report?

6. Assess the normality and equal variance assumptions by plotting the residuals of the fitted model (see Section 6.3 for more details). Split the plotting device into 2 rows and 2 columns using `par(mfrow=c(2,2))` so you can fit four plots on a single device. Use the `plot()` function on your fitted model (`plot(growth.lm)`) to plot the graphs. Discuss with an instructor how to interpret these plots. What are your conclusions?

7. Download the datafile *'Gigartina.CSV'* from the **Data** link and save it to the `data` directory. Import the dataset into R and assign the dataframe an appropriate name. These data were collected from a study to examine the change in `diameter` of red algae *Mastocarpus stellatus* spores grown in three different diatom cultures and a control group grown in artificial seawater (`diatom.treat` variable). Use the function `str()` to examine the dataframe. How many replicates are there per diatom treatment? Use an appropriate plot to examine whether there are any obvious differences in diameter between the treatments.

8. You wish to compare the mean diameter of *Metacarpus* grown in the four treatment groups (`ASGM`, `Sdecl`, `Sexpo`, `Sstat`) using a one-way analysis of variance (ANOVA). What is your null hypothesis?

9. We will conduct the ANOVA using the linear model function `lm()` once again. Make sure you know which of the variables is your response variable and which is your explanatory variable (ask an instructor if in doubt). Fit the linear model and assign the model output to a variable with an appropriate name (i.e. `gigartina.lm`).

10. Produce an ANOVA table using the `anova()` function. What is the value of the p value? Do you reject or fail to reject the null hypothesis? What is the value of the F statistic and degrees of freedom? How would you report these summary statistics in a report?

11. Assess the assumptions of normality and equal variance of the residuals by producing the residual plots as before. Don't forget to split the plotting device into 2 rows and 2 columns before plotting. Discuss with an instructor whether the residuals meet these assumptions. Do you accept this model as acceptable?

12. Let's compare the treatment group means to determine which treatment group is different from other treatment groups. In general, you should be careful with these types of post-hoc comparisons, especially if you have a large number of groups (There are much better ways to do this, but that's for another course!). In this case we only have 4 groups, and therefore we will use Tukey's Honest significant difference to perform the comparisons and control for type 1 error rate (rejecting a true null hypothesis).

13. We will use the function `TukeyHSD.lm()` from the `mosaic` package to perform these comparisons (you will need to install this package first and then use `library(mosaic)` to make the function available). Which groups are different from each other if we use the p-value cutoff (alpha) of $p < 0.05$?

14. We can also produce a plot of the comparisons to help us interpret the table of comparisons. Use the `plot()` function with the `TukeyHSD.lm(gigartina.lm)`. Ask if you get stuck (or Google it!).

15. Download the *'TemoraBR.csv'* file from the **Data** link and save it to the `data` directory. Import the dataset into R and as usual assign it to a variable. These data are from an experiment that was conducted to investigate the relationship between temperature (`temp`) and the beat rate (Hz) `beat_rate` of the copepod *Temora longicornis* which had been acclimatised at three different temperature regimes (`acclimitisation_temp`). Examine the structure of the dataset. How many variables are there? What type of variables are they? Which is the response (dependent) variable, and which are the explanatory (independent) variables?

16. What type of variable is `acclimitisation_temp`? Is it a factor? Convert `acclimitisation_temp` to a factor and store the result in a new column in your dataframe called `Facclimitisation_temp`. Hint: use the function `factor()`. Use an appropriate plot to visualise these data (perhaps a coplot or similar?).

17. We will analyse these data using an Analysis of Covariance (ANCOVA) to compare the slopes and the intercepts of the relationship between `beat_rate` and `temp` for each level of `Facclimatisation_temp`. Take a look at the plot you produced in Q16, do you think the relationships are different?

18. As usual we will fit the model using the `lm()` function. You will need to fit the main effects of `temp` and `Facclimatisation_temp` and the interaction between `temp` and `Facclimatisation_temp`. You can do this using either of the equivalent specifications:

    `temp + Facclimatisation_temp + temp:Facclimatisation_temp` or

    `temp * Facclimatisation_temp`

19. Produce the summary ANOVA table as usual. Is the interaction between `temp` and `Facclimatisation_temp` significant? What is the interpretation of the interaction term? Should we interpret the main effects of `temp` and `Facclimatisation_temp` as well?

20. Assess the assumptions of normality and equal variance of the residuals in the usual way. Do the residuals meet these assumptions? Discuss with a instructor.

21. Write a short summary in you R script (don't forget to comment this out with `#`) describing the interpretation of this model. Report the appropriate summary statistics such as F values, degrees of freedom and p values.

22. (Optional) refit the model using the square root transformed `beat_rate` as the response variable. Does the interpretation of the model change? Do the validation plots of the residuals look better?

End of Exercise 5