Babraham Bioinformatics

Introduction to Statistics with R: Exercises

Version 2024-02

Contents

Introduction to Statistics with R: Exercises **1**
Version 2024-02 ..... *1*
Exercise 1: Power **3**
Activated T cells 3
Mice weight 3
Arachnophobes 3
Exercise 2: t-test **4**
Working memory 4
Coffee 4
Exercise 3: ANOVA **5**
Crop yield 5
Neutrophils 5
Exercise 4: Correlation **6**
Roe deer 6
Exam anxiety 6
Exercise 5: Non-parametric **7**
Oxbridge rivalry: T-shirts [Mann-Whitney] 7
Botulinum [Wilcoxson paired] 7
Creatine [Kruskal-Wallis] 7
Violin [Friedman] 8
Dominance [Spearman Rank] 8
Exercise 6: Qualitative data **9**
Cats & dogs 9
Cane toads 9
Exercise 7: Mixed **10**
Iris flowers 10
Recycling 10
Income data 11
Coffee: part 2 11

Exercise 1: Power

Activated T cells

- A researcher activates T cells with an antibody. The cells are then marked by a fluorescent antibody which recognises a T-cell receptor. They count the cells and categorise them as polarised or not polarised. They hypothesise that the polarisation varies between WT and KO cells, the scientist starts by running a pilot study.

	Polarised	Not Polarised
WT	10	31
KO	14	21

- Providing the observed difference between WT and KO cells is of scientific interest, what sample size is needed to achieve 80% power?
  - **Hint:** use `power.prop.test()`

Mice weight

- A researcher wants to compare the weights of WT and KO mice. They do not have any data yet on the KO but they have values for the WT. A difference of interest would be at least 10%.

Weight	27.2	25.5	26	29.1

- Given that you will be using a t-test, what sample size is needed to be able to spot a 10% difference with 80% power?
  - **Hint:** in R use `mean()` and `sd()`

Arachnophobes

- You want to know whether real tarantulas and pictures of a tarantula results in the same level of anxiety in arachnophobes.
- You run a pilot study, where 10 arachnophobes were asked to perform 2 tasks:
  - Task 1: Group1 (n=5): to play with a real, live tarantula
  - Task 2: Group 2 (n=5): to look at pictures of the same tarantula
- Anxiety scores were measured for each group (0 to 100).
- Use the data (spider.data.csv) to calculate the values for a power calculation and run a power calculation. You will use an independent t-test for your analysis.
  - **Hint:** use `group_by()` and `summarise()`, with `mean()` and `sd()`
- If, after carefully running your power analysis, you realise that the pilot study actually involved 5 arachnophobes, who each looked at both a picture and a real spider (i.e. creating a paired design, so using a paired t-test), how many arachnophobes would you need to achieve a power of 80%?
  - **Hint:** calculate differences using `pivot_wider()`, `mutate()`, and `summarise()`

Exercise 2: t-test

Working memory

- A group of rhesus monkeys (n=15) performs a task involving memory after having received a placebo. Their performance is graded on a scale from 0 to 100. They are then asked to perform the same task after having received a dopamine depleting agent.
- Is there an effect of treatment on the monkeys' performance?
    - Load working.memory.csv and look at the structure of the data.
    - Work out the difference: DA.depletion – placebo and assign the difference to a column: working.memory$difference
    - Plot the difference as a stripchart with a mean
    - Add confidence intervals as error bars
        - Clue: `stat_summary(…, fun.data=mean_cl_normal)` (Hmisc package)
    - Check the assumptions for a parametric test.
    - Run the paired *t*-test: `t_test(var ~ 1, mu=0)`
    - **Extra task 1:** Run the paired *t*-test using `t_test(paired=TRUE)`
        - **Hint:** you will need to convert the data to long format using `pivot_longer()` and sort by the ID column using `arrange(Subject)` so that the correct pairing is used in the test.
    - **Extra task 2:** Plot the original data showing the pairing between the two groups using `ggplot() + geom_line()` or either `ggline()` or `ggpaired()` from the ggpubr package. Add p-values to the graphs.

Coffee

- You are tired and want a decent coffee but are not sure which type to have – you have the option of Robusta or Arabica beans. You find some data where they have been tasted and rated so decide to use this to base your decision on.
- Which beans should you choose for your coffee?
    - Explore the data (coffee.bean.species.csv)
    - Check the assumptions
    - Run a t-test
    - **Extra task:** produce a final plot of the results including p-values on the graph
- *Data from the CORGIS Dataset Project https://corgis-edu.github.io/corgis/csv/coffee/*

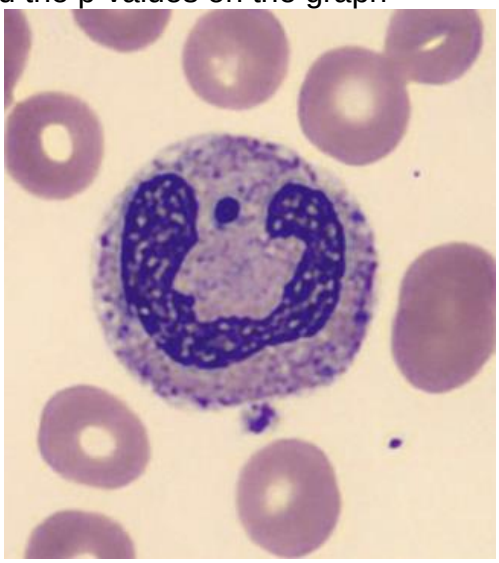
Exercise 3: ANOVA

Crop yield

- A crop researcher wants to test the effect of three different fertiliser mixtures on crop yield. They carry out an experiment using three fertiliser types (1, 2, and 3) and measure crop yield (crop.yield.csv).
- Do the fertilisers have an impact on yield and, if so, which fertiliser mixture gives the highest yield?
- **Hint:** you will need to convert the fertiliser type to a factor:
  - `mutate(fertilizer = as.factor(fertilizer))`
- *From Bevans, R. (2023, June 22). One-way ANOVA | When and How to Use It (With Examples). Scribbr. https://www.scribbr.com/statistics/one-way-anova/*

Neutrophils

- A researcher is looking at the difference between 4 cell groups. They have run the experiment 5 times. Within each experiment, they have neutrophils from a WT (control), a KO, a KO + Treatment 1 and a KO + Treatment2.
- Is there a difference between KO with/without treatment and WT?
  - Plot the data so that you have an idea of the consistency of the results between the experiments (use data neutrophils.long.csv).
  - Check the assumptions
  - Run the repeated measures ANOVA and post-hoc tests
    - `anova_test(dv =, wid =, within =) -> res.aov`
    - `get_anova_table(res.aov)`
    - `pairwise_t_test(p.adjust.method =)`
  - Choose a graphical presentation consistent with the experimental design
  - **Extra task**: add the p-values on the graph



Exercise 4: Correlation

Roe deer

- You want to know if there is a relationship between parasite burden (PL) and body mass (BM) in roe deer. Explore the data (Roe.deer.csv) and test this relationship.
  - Build a fit for males and females separately
    - `data %>% filter() lm(y~x, data=)`
  - Plot the 2 lines of best fit on the same graph
    - `coefficients() geom_abline()`
  - Check the assumptions visually from the data and with the output for models
    - `par(mfrow=c(2,2)) plot(fit.male) plot(fit.female)`
  - Run the correlation test
    - `deer %>% group_by() %>% cor_test()`

Exam anxiety

- Is there a relationship between time spent revising and exam anxiety (exam.anxiety.csv)? And, if yes, are males and females different? How good is the model?
  - Build a fit for the boys and a fit for the girls
    - `data %>% filter() lm(y~x, data=)`
  - Plot the 2 lines of best fit on the same graph
    - `coefficients() geom_abline()`
  - Check the assumptions visually from the data and with the output for models
    - `par(mfrow=c(2,2)) plot(fit.male) plot(fit.female)`
  - Filter out misbehaving values based on the standardised residuals
    - `rstandard() cooks.distance() add_column()`
  - Plot the final (improved) model
    - `bind_rows()`

Exercise 5: Non-parametric

Oxbridge rivalry: T-shirts [Mann-Whitney]

- A group of Cambridge students hypothesise that group body odour is less disgusting when associated with an in-group member (Cambridge) versus an out-group member (Oxford). To test this, they designed a study where two groups of Cambridge University students are presented with one of two worn T-shirts with university logos. They are asked to score these on a scale of 1 to 7, with 7 being the most disgusting

- Can Cambridge students tell the difference between worn T-shirts from Oxford or Cambridge? Explore the data (smelly.teeshirt.csv) and answer the question with a non-parametric approach `wilcox_test()`
- What do you think about the design?
- **Extra task:** add the p-value on the graph

Botulinum [Wilcoxson paired]

- A group of 9 children with muscle spasticity (or extreme muscle tightness limiting movement) in their right upper limb underwent a course of injections with botulinum toxin to reduce spasticity levels. A neurologist (blinded) assessed levels of spasticity pre- and post-treatment for all 9 children using a 10-point ordinal scale. Higher ratings indicated higher levels of spasticity.
- Question: do botulinum toxin injections reduce muscle spasticity levels?
    - Explore the data (botulinum.long.csv)
    - Work out and plot the difference (after – before)
    - Answer the question with a non-parametric approach
        - `wilcox_test(paired = TRUE)`

Creatine [Kruskal-Wallis]

- Creatine is a supplement popular among body builders, to test whether it has an impact on weight gain, three groups were tested: No creatine; Once a day; and Twice a day.
- Does the average weight gain depend on the creatine group to which people were assigned? Explore the data (creatine.csv) and answer with a non-parametric approach
    - `kruskal_test(y~x)` produces omnibus part of the analysis
    - `dunn_test(y~x)` produces pairwise comparisons (dunn.test package)

Violin [Friedman]

- An auction house is putting three violins, A, B, and C, up for bidding. Ten violinists are blindfolded are asked to rate the instruments and each player plays the violins in a randomly determined sequence (BCA, ACB, etc.).
- After each violin is played, the violinist (Rater) rates the instrument on a 10-point scale of overall excellence (Likert: 1=lowest, 10=highest).
- Question: which violin is the best according to the 10 violinists?
    - Explore the data (violin.csv) and answer the question with a non-parametric approach
    - `friedman_test(y~x|id)`
    - `wilcox_test(y~x, paired = TRUE, p.adjust.method = "bonferroni")`

Dominance [Spearman Rank]

- Six male colobus monkeys were ranked for social dominance and the eggs of *Trichirus* nematode per gram of monkey faeces were measured (dominance.csv).
- Is social dominance associated with parasitism?
    - `cor_test(method = "spearman")`

Exercise 6: Qualitative data

Cats & dogs

- We previously tested whether cats are more likely to line dance given different rewards: food or affection. Repeat this analysis for dogs & compare to the results from cats, graphically represent the results (dogs.csv).
- **Hint:** need to restructure the data to long format and calculate proportions to plot
  - `dogs %>% mutate(Total = No+Yes) %>% pivot_longer(cols= 2:4, names_to =, values_to =)`
  - `dogs.long %>% group_by(Training) %>% mutate(fraction = count/count[Dance =="Total"]) %>% filter(Dance != "Total") %>% ungroup()`
  - `dogs.long %>% ggplot(aes(x=Training, y=fraction, fill=Dance))+ geom_col(colour="black")+ scale_fill_brewer(palette = 1)`

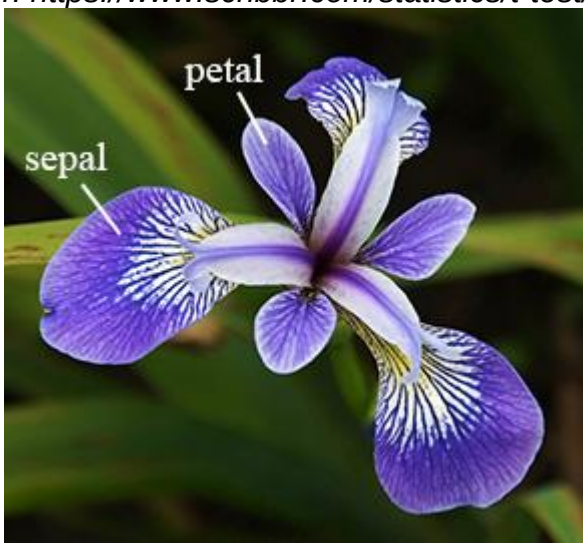
Cane toads

- A researcher decided to check the hypothesis that the proportion of cane toads with intestinal parasites was the same in 3 different areas of Queensland (cane.toad.csv).
- Questions:
  - Is the proportion of cane toads infected by intestinal parasites the same in 3 different areas of Queensland? Produce graphical and statistical answers.
  - Is the proportion of infected cane toads different in Bowen than in the other 2 areas?
- **Hint:** to do pairwise comparisons will need:
  - `column_to_rownames() pairwise_fisher_test()`
- *From Statistics Explained by Steve McKillup*

Exercise 7: Mixed

Iris flowers

- You want to know whether the mean petal length of iris flowers differs according to their species. You find two different species of irises growing in a garden and measure 25 petals of each species. Explore the data (iris.species.csv) and test if there is a difference.
- Next, you want to know whether the mean petal length of iris flowers correlates with the mean sepal length in the virginica species. Explore the data (iris.sepal.csv) and test for a correlation.
- *From Bevans, R. (2023, June 22). An Introduction to t Tests | Definitions, Formula and Examples. Scribbr. https://www.scribbr.com/statistics/t-test/*



Recycling

- A city wants to encourage more of its residents to recycle their household waste. They decide to test two interventions: an educational flyer (pamphlet) or a phone call. They randomly select 300 households and randomly assign them to the flyer, phone call, or control group (no intervention). They'll use the results of their experiment to decide which intervention to use for the whole city. Six months after the intervention, the city looks at the outcomes for the 300 households (recycling.csv).
- Which intervention should the city use to maximise recycling?
- *From Turney, S. (2023, June 22). Chi-Square Test of Independence | Formula, Guide & Examples. Scribbr. https://www.scribbr.com/statistics/chi-square-test-of-independence/*

Income data

- A social researcher is interested in the relationship between income and happiness. They survey 500 people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to 10 (income.data.csv).
- Generate a linear model describing the relationship between income and happiness.
- How happy would you expect someone earing 50k to be, based on your model?
- *From Bevans, R. (2023, June 22). Simple Linear Regression | An Easy Introduction & Examples. Scribbr. https://www.scribbr.com/statistics/simple-linear-regression/*

Coffee: part 2

- You enjoyed your previous coffee so much that you decide to do some more research, this time looking at which country the beans came from (coffee.country.csv). Your local coffee shop sells beans from Ethiopia, Costa Rica, China, and Indonesia – which beans do you buy?