ADS2 Practical 2.14

Rob Young

Semester 2, 2023/24

This practical will not use any new code or functions that you have yet seen in R, but instead you will need to combine many of techniques you have already learnt. You may well find that writing code that runs as you expect is more challenging than designing it, but please do not be nervous about asking questions.

Please also remember that the answers here are only suggestions. There may be many ways of answering the questions here and you don't need to approach everything in the same manner as I have! If you have any questions, please write on this week's discussion board.

Learning objectives

After completing this practical you will be able to:

- Perform bootstrapping analysis to perform hypothesis testing and generate confidence intervals.
- Interpret the output of these analyses and how they differ from standard statistical testing in R.

Does enhancer activity differ with histone modification or transcriptional activity?

Enhancers are long-range regulatory elements which act in the genome to positively drive gene expression. These can be discovered and predicted using a number of techniques, including whether they contain active epigenetic marks or whether they are transcriptionally active. In one of my previous papers (Young et al. 2017, PMID: 29284524) we measured enhancer activity and investigated whether the epigenetic mark or transcriptional marks were associated with higher or lower enhancer activity. Although it does not matter for the purposes of this practical, please note that this paper was highly controversial (see `https://doi.org/10.1101/048629` and the associated linked reviews).

We are going to use bootstrapping to investigate whether there are differences in the median enhancer activity across these groups. You can access these data from the file `Reporter_assay_4-1-15.txt` on Blackboard Learn. The relevant columns that we will need are 'ave', which contains a quantitative measure of enhancer activity, 'Epigenetic_status' which is a binary variable describing whether each enhancer has the Active or Repressed enhancer mark and 'Transcription_status' which is a similar binary variable describing whether each enhancer shows a transcription mark or not.

1. Plot the data and examine it.

Can you see the four groups and understand the difference between them?

2. Generate the first bootstrap sample

Lets look first at the epigenetic status of these enhancers. Can you calculate the median difference between those which are marked as active and those which are marked as repressed?

Can you generate one bootstrap sample of a median difference? Remember, you will need to write this as code which can be automatically run many times later.

3. Generate a large number of bootstraps

Now try to run many replicates of the code you wrote above to generate a distribution of median differences. Can you plot this distribution?

4. Make a statistical inference

Where does your observed difference fall on this distribution? Do you think there is a significant difference between enhancer activity across groups with different epigenetic marks?

5. Explore the number of replicates

The choice of replicates in a bootstrap sample can be difficult to choose - it may depend on your underlying biological assumptions or even the power of your computer! In this case, we don't have too much data so we can play around with different numbers of bootstrap samples. Does it make a difference to your final result?

6. Now, see whether you can repeat this procedure by looking at the 'Transcription_status' variable.

What conclusion do you reach here? Are there any differences or similarities in the data you sampled for when generating these bootstraps?

Which movie genres are most popular with Chinese university students?

A recent survey (https://www.statista.com/statistics/1284497/china-popular-movie- genres-among-university-students/) has reported the favourite movie genre of university students in China (Full confession: these numbers were reported as percentages but to make this practical more straightforward we are going to assume that they are number of students who reported preferring each genre).

We are going to investigate whether there are significant differences in the popularity of genres. You can access these data from the file `movie_data.txt` on Blackboard Learn.

1. Again, plot the data and examine it first.

Can you see any differences across genres? Which do you think might be statistically significant?

2. Generate a first bootstrap sample.

Lets look at comedy, for example. 73 students preferred this genre out of a total of 267 response. From these two numbers, can you generate a vector to sample from in order to generate a bootstrap sample?

3. Generate a confidence interval by bootstrapping many times.

4. Repeat this bootstrapping for the entire dataset

Do you need to sample from different datasets for each genre?

5. Explore the differences across genres, are there any differences that you can detect?

Please do this in the way you think is appropriate - you could try plotting the confidence intervals as in the lecture or making judgements on the overlap between confidence intervals. Please think about how you would prepare your null and alternative hypotheses in this setting and how you could estimate statistical significance.