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?

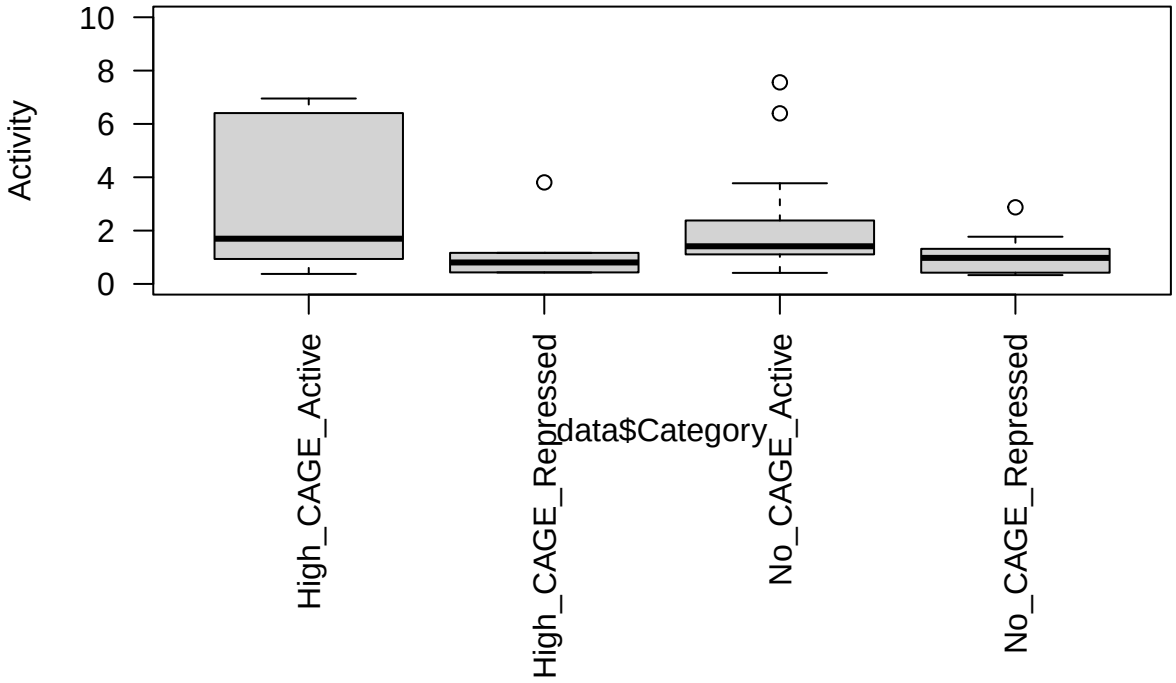
Enhancers are long-range regulatory elements which act in the genome to positively drive gene expression. These can be discovered and predicted using a number of techniques, including whether they contain active epigenetic marks or whether they are transcriptionally active. In one of my previous papers (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?

```
data<-read.table("Reporter_assay_4-1-15.txt", header = T)
par(mar=c(11,4,4,2))
boxplot(data$ave~data$Category, ylim = c(0,10), las = 2, ylab = "Activity")
```



It looks to me like the two 'Active' categories might have higher activity than the two 'repressed' categories. I'm not sure if there will be any difference between the two 'Active' categories (High_CAGE and No_CAGE). It also looks to me as if there might be no difference between the two 'Repressed' categories.

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?

```
active_activity<-median(subset(data$ave, data$Epigenetic_status == "Active"))
repressed_activity<-median(subset(data$ave, data$Epigenetic_status == "Repressed"))
median_diff<-active_activity - repressed_activity
median_diff
```

```
## [1] 0.6368301
```

The median difference in activity between the active and repressed categories is 0.64.

Now its time to make a bootstrap sample for this comparison. If we are going to use the `sample` command, which subsets or pools of the data should we use to sample from?

The null hypothesis is that there is no difference between these two groups of enhancers. We should therefore pool the activity measurements from these two groups. Each sample should match the same number of observations as in the initial dataset - 32 active and 15 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.

```
length_active<-nrow(subset(data, data$Epigenetic_status == "Active"))
length_repressed<-nrow(subset(data, data$Epigenetic_status == "Repressed"))
bootstrap_active<-median(sample(data$ave, length_active, replace = TRUE))
bootstrap_repressed<-median(sample(data$ave, length_repressed, replace = TRUE))
bootstrap_median<-bootstrap_active - bootstrap_repressed
bootstrap_median
```

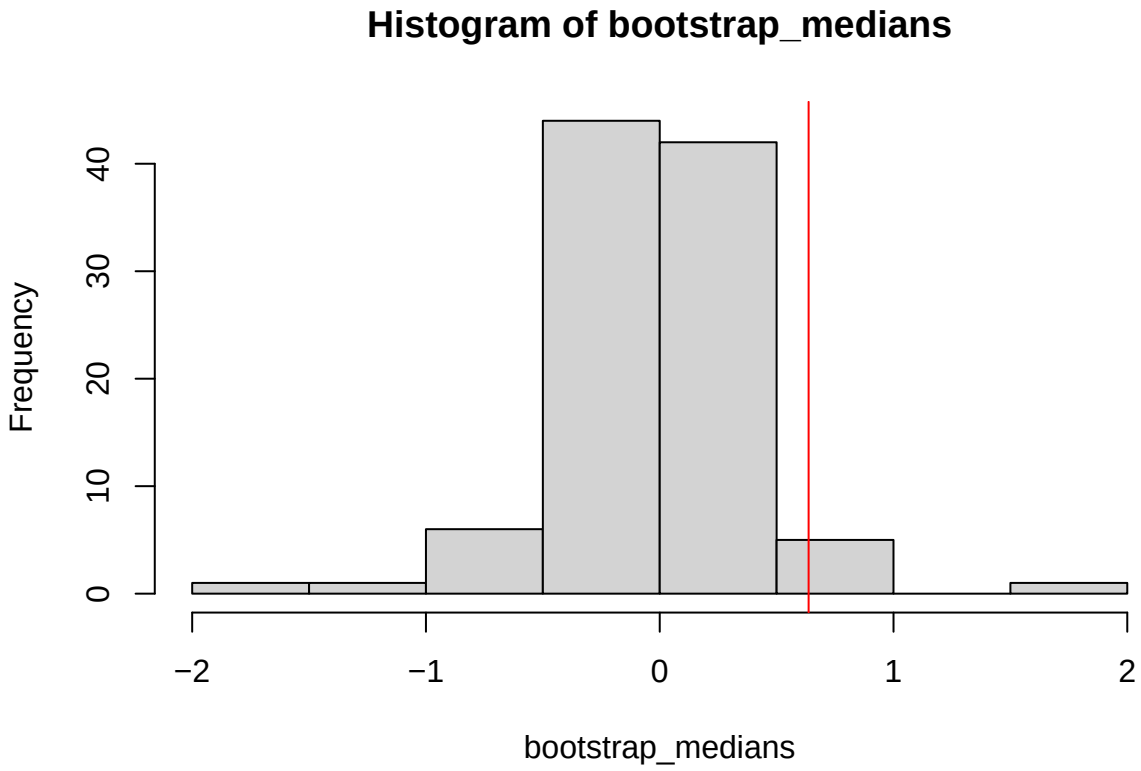
```
## [1] 0.3067601
```

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?

I generated 100 bootstraps here and drew this histogram:

```
set.seed(45)
bootstrap_medians<-vector()
for (a in 1:100){
  bootstrap_active<-median(sample(data$ave, length_active, replace = TRUE))
  bootstrap_repressed<-median(sample(data$ave, length_repressed, replace = TRUE))
  bootstrap_median<-bootstrap_active - bootstrap_repressed
  bootstrap_medians<-c(bootstrap_medians, bootstrap_median)
}
hist(bootstrap_medians)
abline(v = median_diff, col = 'red')
```



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?

```
sig_bootstraps = length(subset(bootstrap_medians, bootstrap_medians >= median_diff))
sig_bootstraps/100
```

```
## [1] 0.03
```

This result indicates that 3 out of my 100 bootstraps had a larger median difference than the observed value and the associated p-value is therefore 0.03. As this is less than the standard threshold of $p = 0.05$ I reject the null hypothesis and therefore conclude that there **is** a significant difference in activity between the groups - those marked as active show a higher enhancer activity than those marked as repressed.

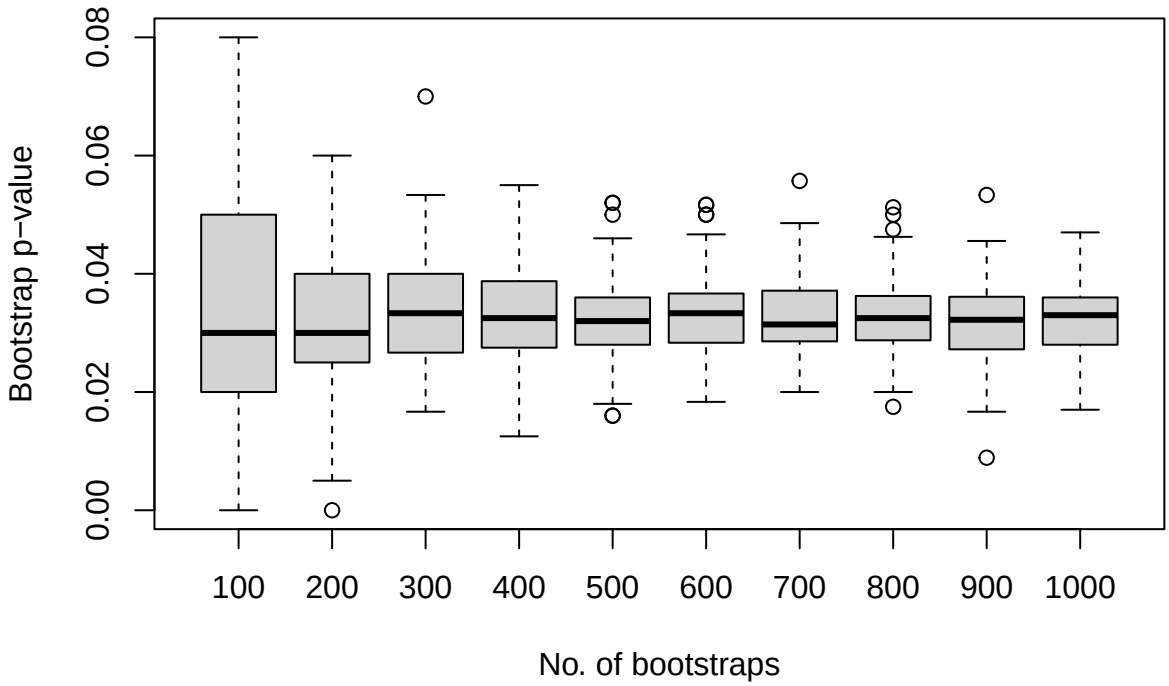
If you look carefully at my code you will see that I have included the line `set.seed` so that I generate the same random sampling each time. I noted when preparing this that I did not always reject the null hypothesis at $p < 0.05$, depending on the random sampling that I did so don't worry if you get a different answer to me here - we will explore this in the next section.

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?

I decided to increase my number of bootstraps from 100 to 1,000 in steps of 100. For each number of replicates, I performed the bootstrap 100 times (I bootstrapped the bootstraps!) to see how many times I received a significant result at each threshold. Please note that this following code took several minutes to run on my computer - it will complete, but you will need to be patient!

```
all_sig_results <- vector()
all_bootstrap_replicates <- vector()
for (bootstrap_replicate in seq(100, 1000, 100)) {
    for (a in 1:100) {
        sig_bootstraps <- 0
        for (b in 1:bootstrap_replicate) {
            bootstrap_active <- median(sample(data$ave, length_active,
                replace = TRUE))
            bootstrap_repressed <- median(sample(data$ave, length_repressed,
                replace = TRUE))
            bootstrap_median <- bootstrap_active - bootstrap_repressed
            if (bootstrap_median >= median_diff) {
                sig_bootstraps <- sig_bootstraps + 1
            }
        }
        sig_result <- sig_bootstraps/bootstrap_replicate
        all_sig_results <- c(all_sig_results, sig_result)
        all_bootstrap_replicates <- c(all_bootstrap_replicates,
            bootstrap_replicate)
    }
}
boxplot(all_sig_results ~ factor(all_bootstrap_replicates), ylab = "Bootstrap p-value",
    xlab = "No. of bootstraps")
```



In this plot you can see that the bootstrap p-value hovers around 0.03 but there is no systematic change as I increase the number of bootstrap values. The only pattern you can see is the reduced variance in p-value with higher numbers of bootstraps but no systematic change to the likely result.

Does this result surprise you?

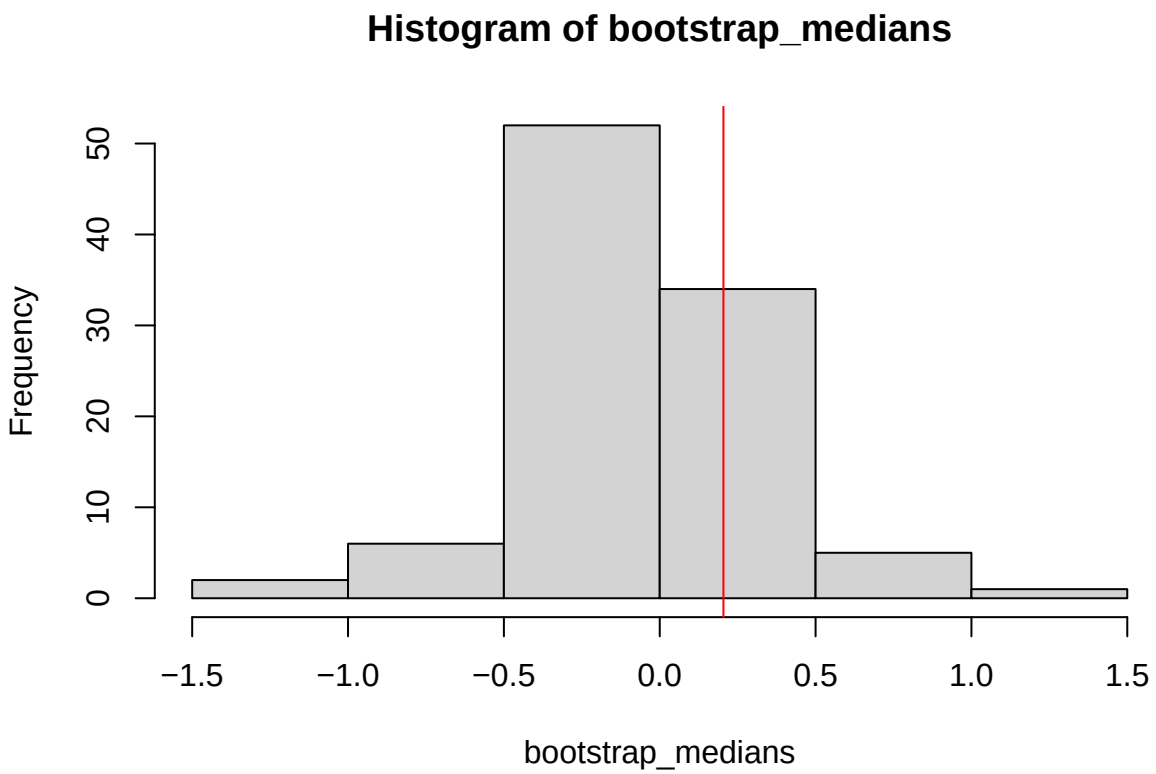
6. Now, see whether you can repeat this procedure by looking at the 'Transcription_status' variable.

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

```
active_activity<-median(subset(data$ave, data$Transcription_status == "Active"))
repressed_activity<-median(subset(data$ave, data$Transcription_status == "None"))
median_diff<-active_activity - repressed_activity
median_diff
```

```
## [1] 0.2043462
```

```
length_active<-nrow(subset(data, data$Transcription_status == "Active"))
length_repressed<-nrow(subset(data, data$Transcription_status == "None"))
set.seed(45)
bootstrap_medians<-vector()
for (a in 1:100){
  bootstrap_active<-median(sample(data$ave, length_active, replace = TRUE))
  bootstrap_repressed<-median(sample(data$ave, length_repressed, replace = TRUE))
  bootstrap_median<-bootstrap_active - bootstrap_repressed
  bootstrap_medians<-c(bootstrap_medians, bootstrap_median)
}
hist(bootstrap_medians)
abline(v = median_diff, col = 'red')
```




```
sig_bootstraps = length(subset(bootstrap_medians, bootstrap_medians >= median_diff))
sig_bootstraps/100
```

```
## [1] 0.22
```

I generated samples very similarly to as above (I actually copy and pasted the code before modifying it!) but just replaced the `Epigenetic_status` column with the `Transcription_status` column. I also modified the length of each bootstrap sample to correspond to the numbers in this column.

Here I have reported a p-value = 0.22, which prevents me from rejecting the null hypothesis at the standard p-value = 0.05 threshold. I can therefore conclude that there is no difference in enhancer activity between regions which are transcriptionally active and those which are not.

Which movie genres are most popular with Chinese university students?

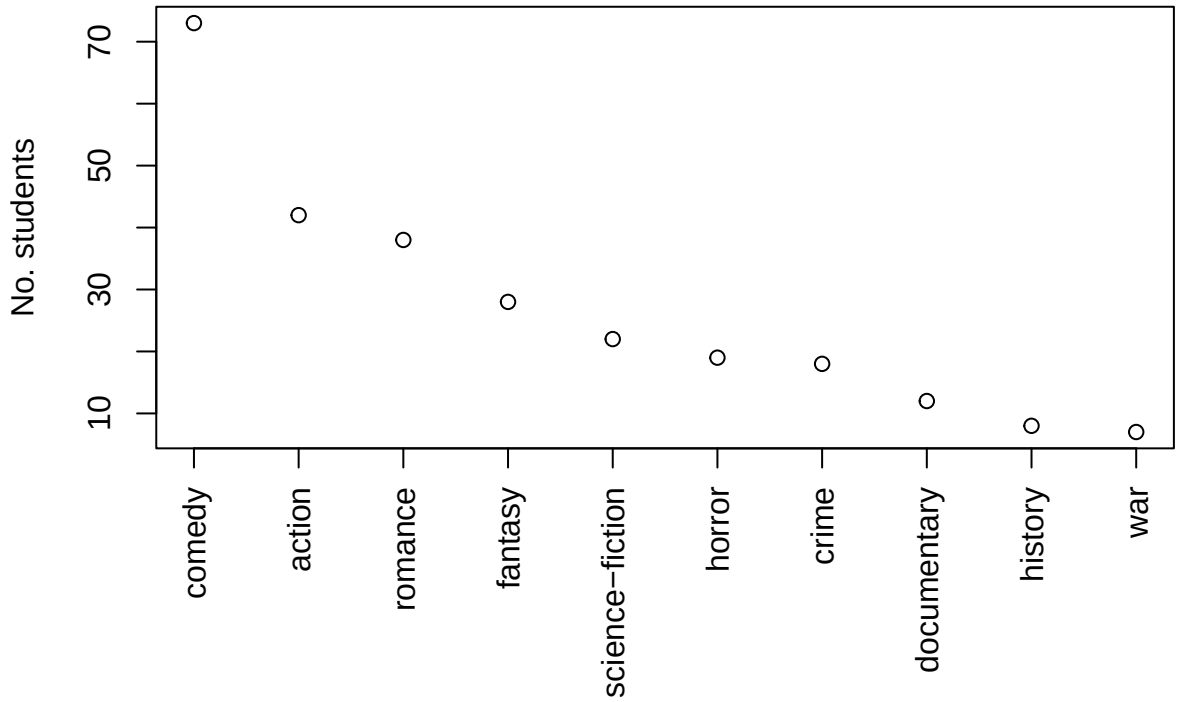
A recent survey (https://www.statista.com/statistics/1284497/china-popular-movie- genres-among-university-students/) has reported the favourite movie genre of university students in China (Full confession: these numbers were reported as percentages but to make this practical more straightforward we are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

```
movie_data<-read.table("movie_data.txt", header = T)
par(mar=c(7,4,4,2))
plot(movie_data$students, ylab = "No. students", xaxt = "n", xlab = "")
axis(side = 1, at = seq(1,nrow(movie_data),1), labels = movie_data$genre, las = 2)
```



Comedy and action appear to be the most popular while history and war movies are least popular. I can't tell just from this plot which ones are significantly different from each other.

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?

```
comedy_fans<-73
total_fans<-267
not_comedy_fans<-total_fans - comedy_fans
obs_sample<-c(rep("comedy", comedy_fans), rep("not_comedy", not_comedy_fans))
summary(factor(obs_sample))
```

```
##     comedy not_comedy
##         73        194
```

```
bootstrap_sample<-sample(obs_sample, length(obs_sample), replace = T)
summary(factor(bootstrap_sample))
```

```
##     comedy not_comedy
##         67        200
```

The observed sample should represent the total number of movie fans, which I have categorised here as to `comedy` and `not_comedy`. As before, my first bootstrap sample looks similar but is not identical to the observed sample. You might have done this differently.

3. Generate a confidence interval by bootstrapping many times.

This might take a bit of coding. You can check whether you did this correctly if your confidence interval spans the observed datapoint and whether the average of your bootstrap values is close to your observation.

```
bootstrap_new <- vector()
for (b in 1:100) {
    bootstrap_sample <- sample(obs_sample, length(obs_sample),
        replace = T)
    bootstrap_new <- c(bootstrap_new, length(subset(bootstrap_sample,
        bootstrap_sample == "comedy")))
}
lower_ci <- as.numeric(quantile(bootstrap_new, 0.025))
upper_ci <- as.numeric(quantile(bootstrap_new, 0.975))

comedy_fans
```

```
## [1] 73
```

```
lower_ci
```

```
## [1] 63
```

```
upper_ci
```

```
## [1] 86.525
```

4. Repeat this bootstrapping for the entire dataset

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

```
obs_values <- vector()
lower_cis <- vector()
upper_cis <- vector()

for (a in 1:nrow(movie_data)) {
    genre <- movie_data[a, 1]
    observed_fans <- movie_data[a, 2]
    not_observed_fans <- total_fans - observed_fans

    obs_sample <- c(rep(genre, observed_fans), rep("not_genre",
        not_observed_fans))
    bootstrap_new <- vector()
    for (b in 1:100) {
        bootstrap_sample <- sample(obs_sample, length(obs_sample),
            replace = T)
        bootstrap_new <- c(bootstrap_new, length(subset(bootstrap_sample,
            bootstrap_sample == genre)))
    }
    lower_ci <- quantile(bootstrap_new, 0.025)
    upper_ci <- quantile(bootstrap_new, 0.975)

    obs_values <- c(obs_values, observed_fans)
    lower_cis <- c(lower_cis, lower_ci)
    upper_cis <- c(upper_cis, upper_ci)
}
```

As you can see, for each genre I have made a separate `obs_sample` variable from which I generate my

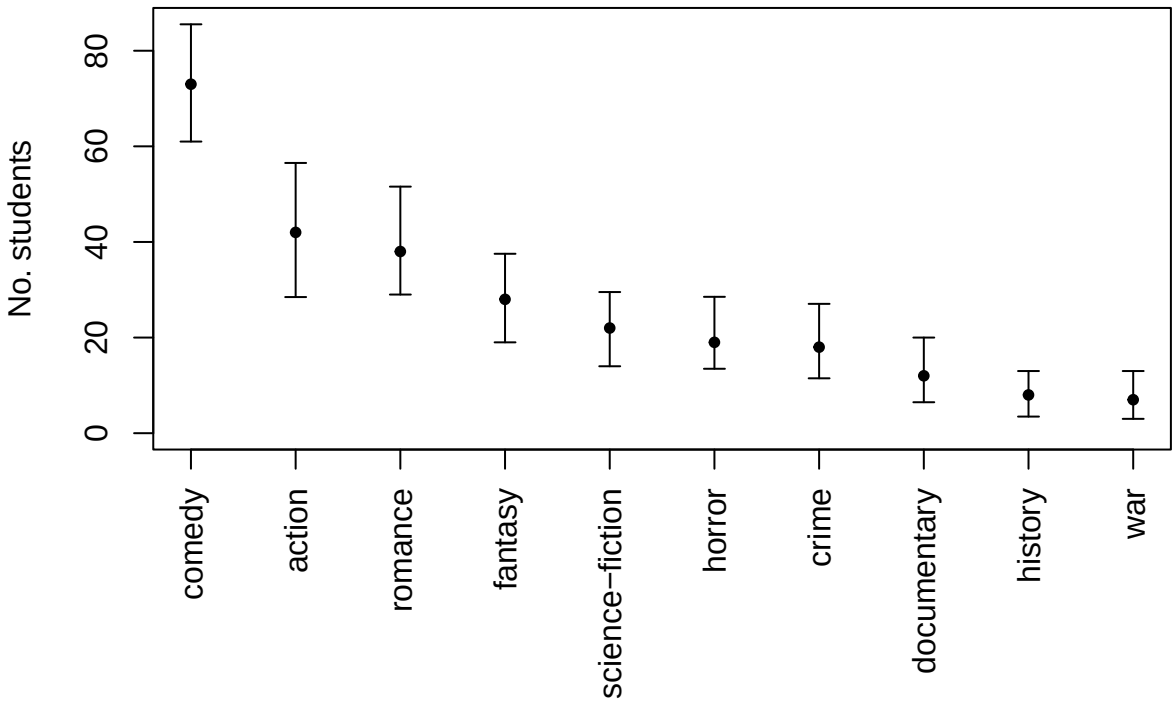
bootstrap values. This variable had the same number of fans and not_fans as in the original sample. I could then generate a bootstrap sample which was the same length as the total number of students. I did this because I am interested in the number of fans of each movie genre, rather than directly exploring any specific hypotheses at this stage.

This took me quite a bit of code as you can see! Did you find a more efficient way to explore these data?

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.

```
ymax<-ceiling(max(upper_cis)*100)/100
par(mar=c(7,4,4,2))
plot(obs_values, xaxt = "n", ylim = c(0,ymax), xlab = "", pch = ".", ylab = "No. students")
axis(side = 1, at = seq(1,nrow(movie_data),1), labels = movie_data$genre, las = 2)
for (a in 1:length(lower_cis)){
    lines(x = c(a,a), y = c(lower_cis[a], upper_cis[a]))
    lines(x = c(a-0.1,a+0.1), y = c(lower_cis[a], lower_cis[a]))
    lines(x = c(a-0.1,a+0.1), y = c(upper_cis[a], upper_cis[a]))
}
points(x = seq(1,nrow(movie_data),1), y = obs_values, pch = 20)
```



Here I just took the confidence intervals I generated above and plotted them in a similar manner to the example in the lecture. My null hypothesis is that there is no difference between genres, while my alternative hypothesis is that there is some difference between genres - I hope you can see here that this is deliberately vague here, I am not making any direct claims or testing differences between pairs of genres as this point.

As the confidence intervals defined here represent a 95% certainty that the true proportion of new cases is within this interval, I will conclude that pairs of genres with non-overlapping confidence intervals are

significantly different at the standard p-value threshold of 0.05. I could alter this by changing the value of the confidence intervals above, e.g. if I plotted the 90% confidence interval I would define non-overlapping samples as being significantly different with a p-value of 0.1.

There are a large number of differences which can be reported here (partly why I recommended against specific comparisons) but to me the biggest patterns are (1) comedy is by far the most popular genre, (2) there is a gradual pattern of reduced popularity across all other genres, but no particular outlier, e.g. the 95% confidence interval for horror, crime, documentary and history are all within the war confience interval, suggesting no significant differences across these groups.

What are your thoughts on these data? Do you agree with my conclusions or did you spot something different and potentially more interesting?