ADS2 Problem Set 2.14

Rob Young

2023/24

Part 1: Quidditch League

Dataset

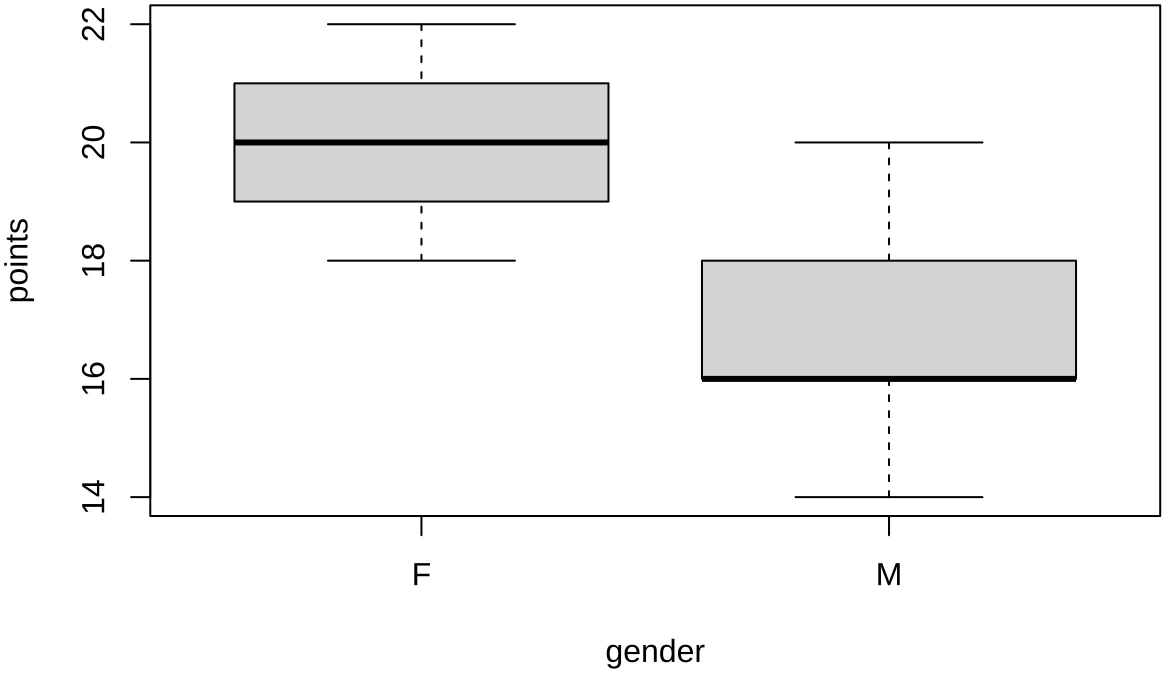
For a study on the influence of gender on physical exercise on university campuses, you analyse a dataset of player scores in a university Quidditch league (expressed at points in an open-ended scale).

The dataset is provided in file quidditch_league.csv

Import and plot the data

First of all, let's look at the data. One thing you are curious about is whether women are better Quidditch players than men. Plot the scores for both groups. You will need the "gender" and "points" columns.

```
data <- read.csv("quidditch_league.csv")
data$gender <- factor(data$gender)
boxplot(data$points ~ data$gender, xlab = "gender", ylab = "points")
```



Are women better than men at quidditch?

Is there a difference between women's and men's ability to play Quidditch? Conduct a test that would answer the question, but think first:

- What is your Null and Alternative Hypothesis?
- What parameter are you comparing between groups?
- What kind of test would you use?
- Do you have to run a bootstrap test, or is there another test available?

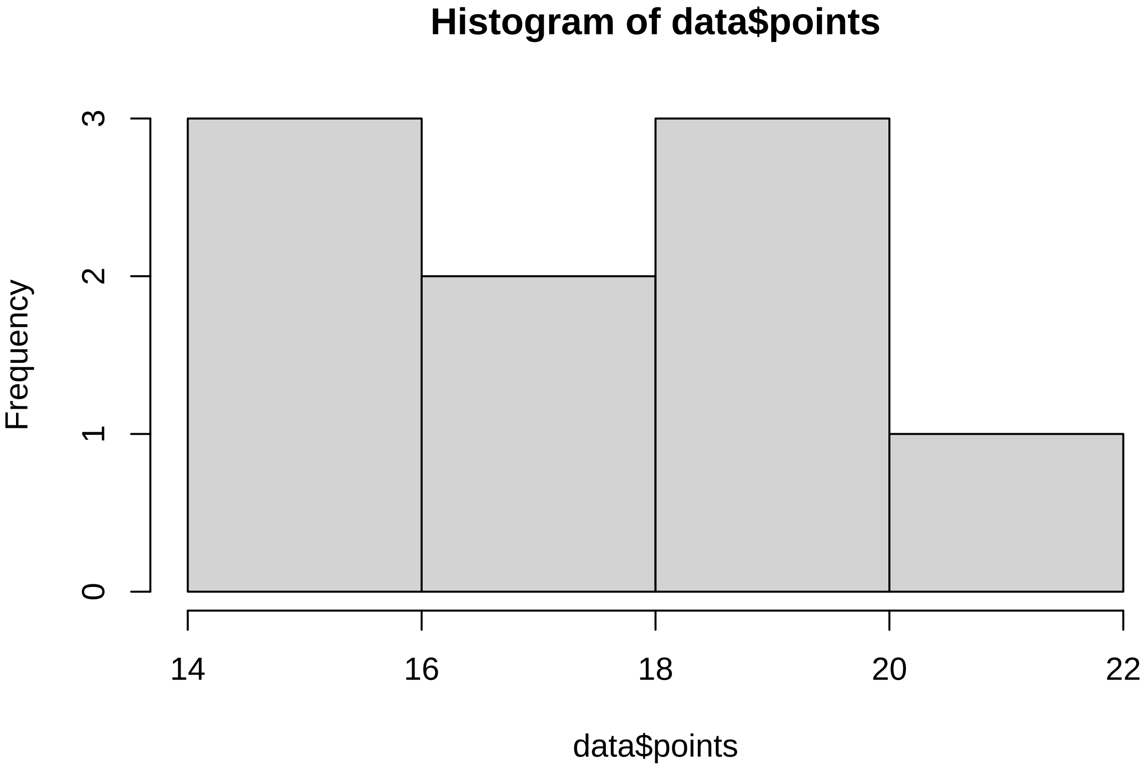
Whatever method you decide to use, use it to come to a conclusion about gender and Quidditch ability. From the above plot it looks like women are better than men at quiditch. However, lets check whether this is a statistically valid decision.

I first assess whether the data is normally distributed which would allow me to perform parametric tests such as a t-test. I performed a Shapiro-Wilks test to test for normality which was non-significant. However, the histogram below suggests that these data are non-normal.

```r
shapiro.test(data$points)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$points
## W = 0.94753, p-value = 0.663
```

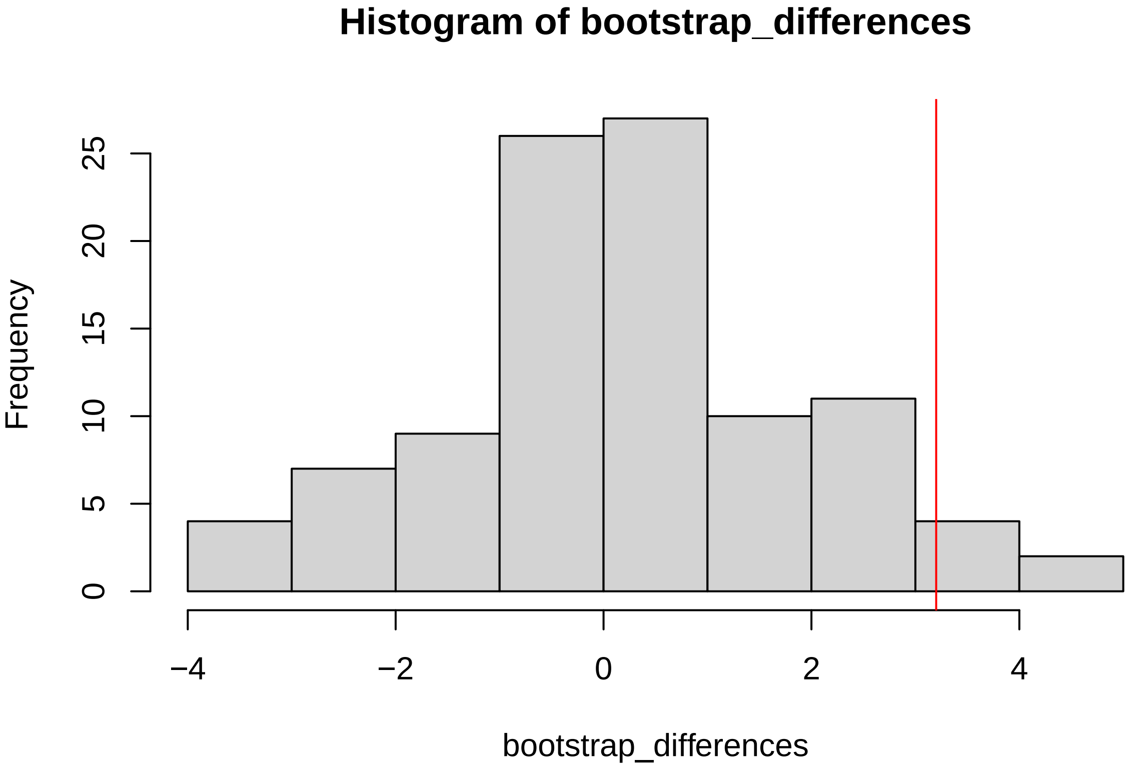
```r
hist(data$points)
```



I think the lack of significance in the Shapiro-Wilks test is due to a lack of statistical power rather than the data being truly normally distributed. To be conservative in my analyis, I therefore use a Bootstrap test to investigate this in detail.

My null hypothesis is that there is no difference in the mean points between men and women while the alternative hypothesis is that the difference observed in the above boxplot is statistically significant. I compare the difference between the means of these groups to a distribution of differences calculated from permutations of the data. For each permutation, I will randomly sample with replacement from the total number of points in the dataset. I will assign the first five random samples to be male and the remainder as female (how or in what order you do this exactly does not matter if you are randomly sampling, this just seemed like a more straightforward approach to me). I next calculate the difference between the means for the sampled ‘male’ and female’ groups and repeat this procedure 100 times. To get an idea of this permutation distribution, I plot the histogram of these values and then add a line describing the location of the observed value.

```r
set.seed(84)
male <- mean(subset(data$points, data$gender == "M"))
female <- mean(subset(data$points, data$gender == "F"))
male_count <- nrow(subset(data, data$gender == "M"))
difference <- female - male
bootstrap_differences <- vector()
for (a in 1:100) {
    temp_data <- data[sample(nrow(data)), 4]
    temp_male <- temp_data[0:male_count]
    temp_female <- temp_data[(male_count + 1):length(temp_data)]
    temp_diff <- mean(temp_female) - mean(temp_male)
    bootstrap_differences <- c(bootstrap_differences, temp_diff)
}
hist(bootstrap_differences)
abline(v = difference, col = "red")
```



```r
print(length(subset(bootstrap_differences, bootstrap_differences >
    difference)))
```

```
## [1] 2
```

I then obtain a p-value for this test by determining how many of my bootstrap values have a more extreme value than my observed value. In this case, this length is 2 which corresponds to p=0.02.

I can therefore conclude that the p-value is significant (at the conventional p = 0.05 threshold), the null hypothesis is rejected and women **are** better at quidditch than men.

Does this depend on who is a woman and a man though?

Now for the second, trickier question. Here is a bit of background on the dataset. It was sent to you by a friend at another university, but all your friend had was players' first names and their numbers of points.

In order to get from names to Gender, you had to guess a bit. This is tricky, because some names are popular across genders. Luckily, there are tools, based on large datasets, that take a name as input and return a probable Gender, and a number associated with that prbability. For instance, according to https://genderize.io/ the name Rob is probably male, with a 0.98 probability (suggesting 98% of people called Rob are men). We have included the probability of a person being male, based on their name, to the data frame as prob_male.

For this dataset, you used the genderizer database and assigned gender as follows. If the probability of being male for a name was 50 percent or above, you assigned "M" as a gender, if it was less than that, then you assigned "F".

But you may have been wrong about some of the people in the dataset. Would this being wrong affect the outcome of your study and the conclusion you drew earlier?

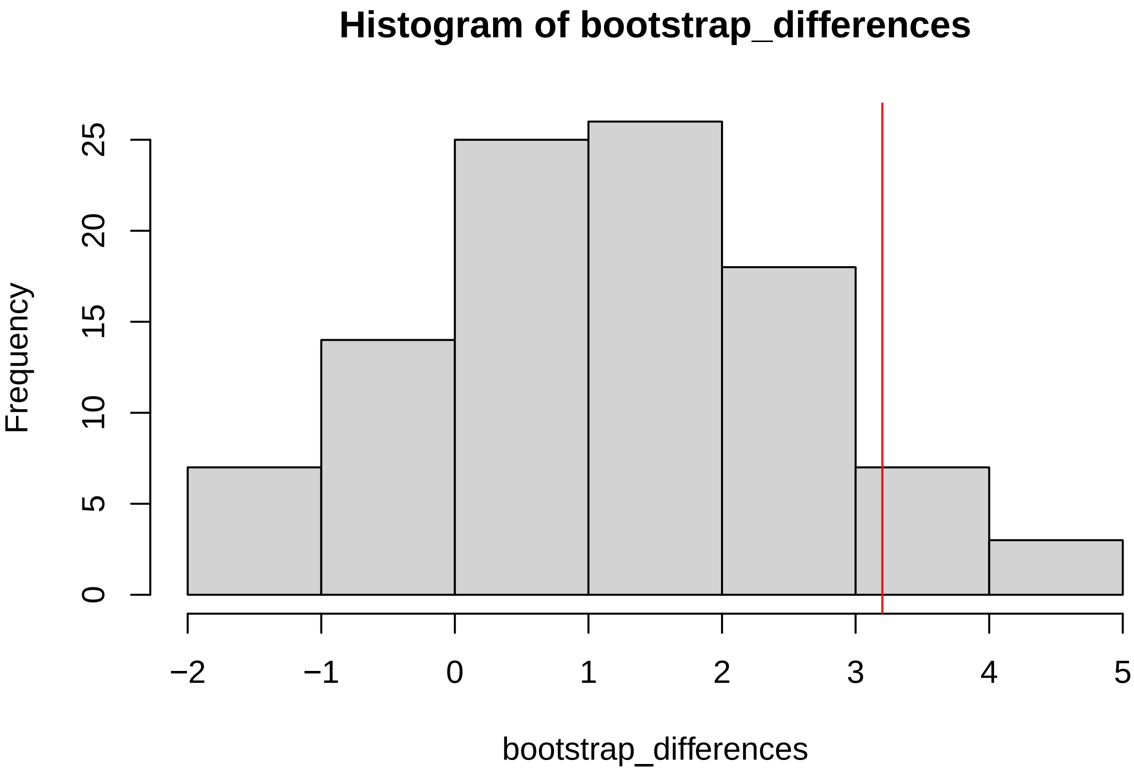
- This is a trickier question, and here bootstrapping is definitely needed!
- But this bootstrap is a bit different from what we have seen in previous examples. We do want to re-assign "F" and "M" labels, but you would probably not want to just redistribute them. Redistributing would mean losing the rich layer of information that is provided in the prob_male column. How do you proceed?
- And what is your conclusion?

Now, lets look at the second part which is more complicated. We have a quantitative probability of each individual being male or female. In my previous bootstrapping, I ignored these results and simply permuted the male and female labels randomly.

Now I want to randomly sample from these labels but take into account these probability measures. To do this, I first generated a background probability distribution for each individual where the number of 'male' labels represented the probability of that individual being a male. The number of 'female' labels came from 100 - prob(male). I assigned the score for that individual to all these 100 instances.

Next, I sampled from my male pool (where each individual was represented in proportion to the probability of their being male) without replacement to calculate a mean sampled 'male' score. The number of samples I extracted was equal to the original number of individuals assigned as 'male'. I repeated the same for my female pool and then calculated a sampled difference. This process was repeated 100 times to generate a bootstrapped distribution and the p-value extracted as above.

```
set.seed(19)
all_genders <- vector()
all_points <- vector()
for (a in 1:nrow(data)) {
    this_row_prob_male <- 100 * data[a, 3]
    this_row_points <- data[a, 4]
    genders <- c(rep("M", this_row_prob_male), rep("F", 100 -
        this_row_prob_male))
    points <- rep(this_row_points, 100)
    all_genders <- c(all_genders, genders)
    all_points <- c(all_points, points)
}
all_males <- subset(all_points, all_genders == "M")
all_females <- subset(all_points, all_genders == "F")

bootstrap_differences <- vector()
for (a in 1:100) {
    temp_male <- sample(all_males, male_count, replace = F)
    temp_female <- sample(all_females, nrow(data) - male_count,
        replace = F)
    temp_diff <- mean(temp_female) - mean(temp_male)
    bootstrap_differences <- c(bootstrap_differences, temp_diff)
}
hist(bootstrap_differences)
abline(v = difference, col = "red")
```



```
print(length(subset(bootstrap_differences, bootstrap_differences >
    difference)))
```

```
## [1] 7
```

I then obtain a p-value for this test by determining how many of my bootstrap values have a more extreme value than my observed value. In this case, this length is 7 which corresponds to p=0.07.

I can therefore conclude that the p-value is not significant (at the conventional p = 0.05 threshold), and the null hypothesis is not rejected. We cannot state that there is a significant difference between men and women when we take this extra complexity into account. Notice here that, by using bootstrapping, we are able to devise a more complex null background which better takes into account the complexity of real-world data. It also changes our conclusion from above such that this uncertainty in assigning gender results in us no longer rejected the null hypothesis.

Part 2 (Optional): Are hockey players born in January?

In his book “Outliers”, the science writer Malcolm Gladwell observes that a lot of professional male ice hockey players in Canada are born in January. (More than would be expected by chance). His explanation for this has to do with the way boys are selected to become hockey players: Once a year (in December), there is a selection process where the best players of any age group get selected into better teams (which includes better coaching, more opportunities to play etc.)

For young children, one major predictor of hockey skill is size, with taller children having an advantage. Imagine a selection happening in December among 7 year-old boys. By the time of the selection, the boys born in January will be almost eight years old, whereas the boys born in November or December will just have turned 7, so will be smaller on average.

Could this explain the dominance of January births among professional players?

In this exercise, we will use a bootstrapping approach twice: First, to create a “virtual selection” of young ice-hockey players. Then, to test whether children born early in the year are indeed over-represented in this selection.

This exercise is quite open-ended (there are several possible ways of solving it), and therefore difficult. Don’t be discouraged if you find it hard. Do what you can. Identify where you are stuck.

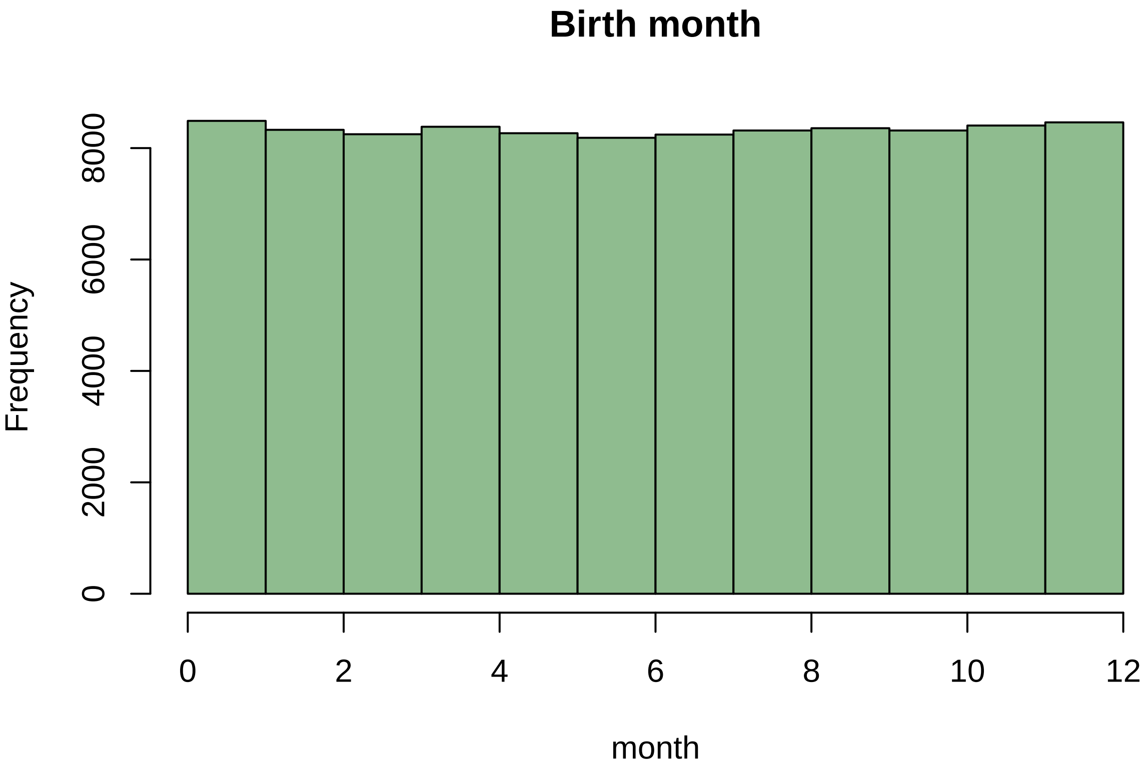
Create an in-silico hockey team

Assume that from the age of 6, every year, half of the players are selected into a better team and continue to play hockey seriously. Of course, size is not the only thing that matters, there is also talent. To model this, we assume that for the children who are median height or taller, the probability of being selected is 0.6; and for shorter children it’s 0.4.

Assume that 6 rounds of selection take place this way (until the children are 12 - maybe after that age, skill becomes more important than height).

If we start with 100000 children who love hockey at age six (and have not undergone prior selection), we would expect the distribution of birth months to look something like the this:

```r
teamsize = 1e+05
all_birthmonths = sample(1:12, teamsize, replace = TRUE)
hist(all_birthmonths, 0:12, main = "Birth month", xlab = "month",
    col = "darkseagreen")
```



After 6 rounds of selection, as described above, there should be around 1563 players left. What would a histogram of their birth months look like?

Here are some numbers that may be helpful:6-year-old children are on average 116cm tall (standard deviation: 4cm). Between the ages of 6 and 12, children grow by 6cm a year (sd: 0.5 cm)

Is the distribution of birth months unusual?

This is much trickier and I built it up over several stages.

First, I generated the starting height distribution from the instructions (average 116cm tall, standard deviation: 4cm).

```
all_heights = rnorm(teamsize, mean = 116, sd = 4)
```

I modified this command to simulate the growth over a year. I then calculated the increased height of all children over a year by adding this growth to their starting height. Note here that I also divided by this annual growth by `12 - birthmonth` to control for the number of months each child had to grow. I realised later while preparing this that I only needed to perform this comparison in the first year. For all subsequent years, each child had the full 12 months from when they were selected in December to grow.

```
growth = rnorm(all_birthmonths, mean = 6, sd = 0.5)/12
all_heights = all_heights + (growth * (12 - all_birthmonths))
median_height <- median(all_heights)
```

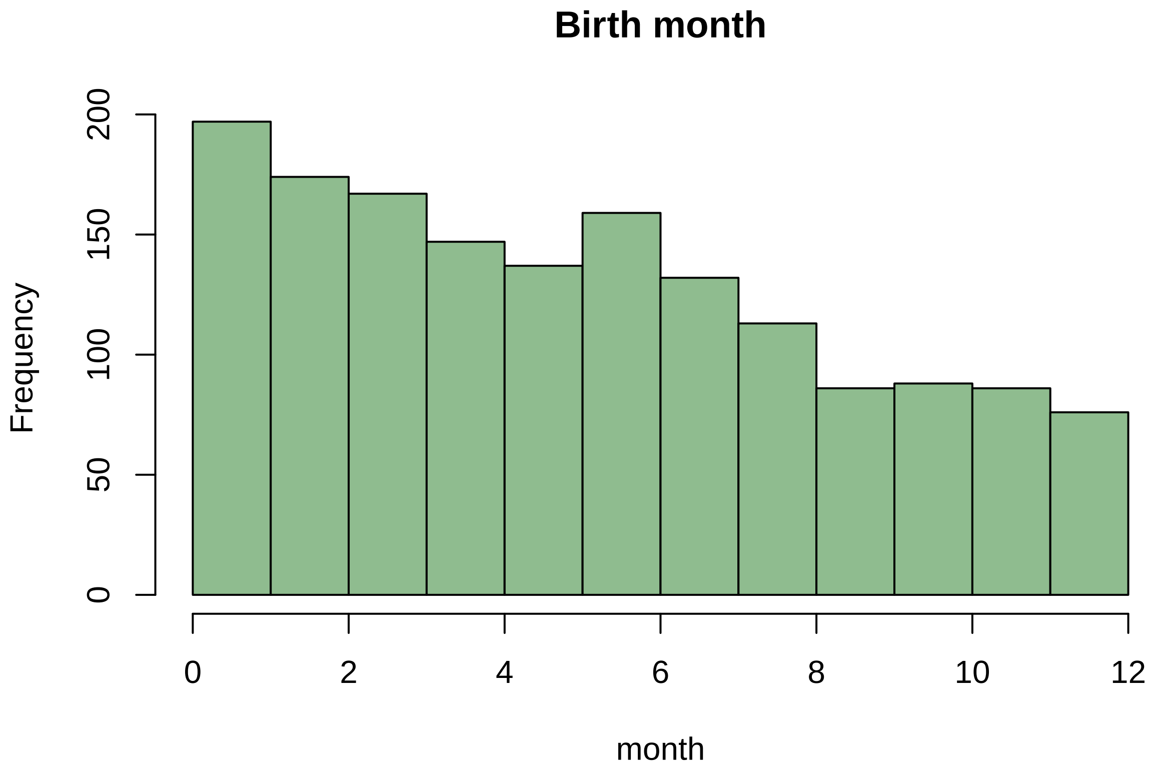
I combined all these new data (birthmonth, height) into a single vector called this_year_data which I am going to sample from. I used the this_year_probs vector to calculate a probability of sampling each individual based on whether their height was greater or less than the median.

The final four lines in this section were used to reformat the sampled data. I needed to extract the birthmonth and heights into separate vectors which were encoded as numbers. Note that these vectors were called the same as earlier up in the code so that they are being overwritten. I needed this because I wanted each year to inherit the data from last year so by modifying these vectors they would keep track of the children being selected each year.

```
this_year_birthmonths <- all_birthmonths
this_year_heights <- all_heights
this_year_data <- paste(this_year_birthmonths, this_year_heights,
    sep = ":")
this_year_probs <- rep(0.6, length(this_year_birthmonths))
this_year_probs[this_year_heights < median_height] <- 0.4

selected_data <- sample(this_year_data, size = length(this_year_birthmonths)/2,
    prob = this_year_probs)

selected_birthmonths <- unlist(strsplit(selected_data, ":"))[seq(1,
    length(selected_data) * 2, 2)]
selected_heights <- unlist(strsplit(selected_data, ":"))[seq(2,
    length(selected_data) * 2, 2)]

all_birthmonths <- as.numeric(selected_birthmonths)
all_heights <- as.numeric(selected_heights)
```

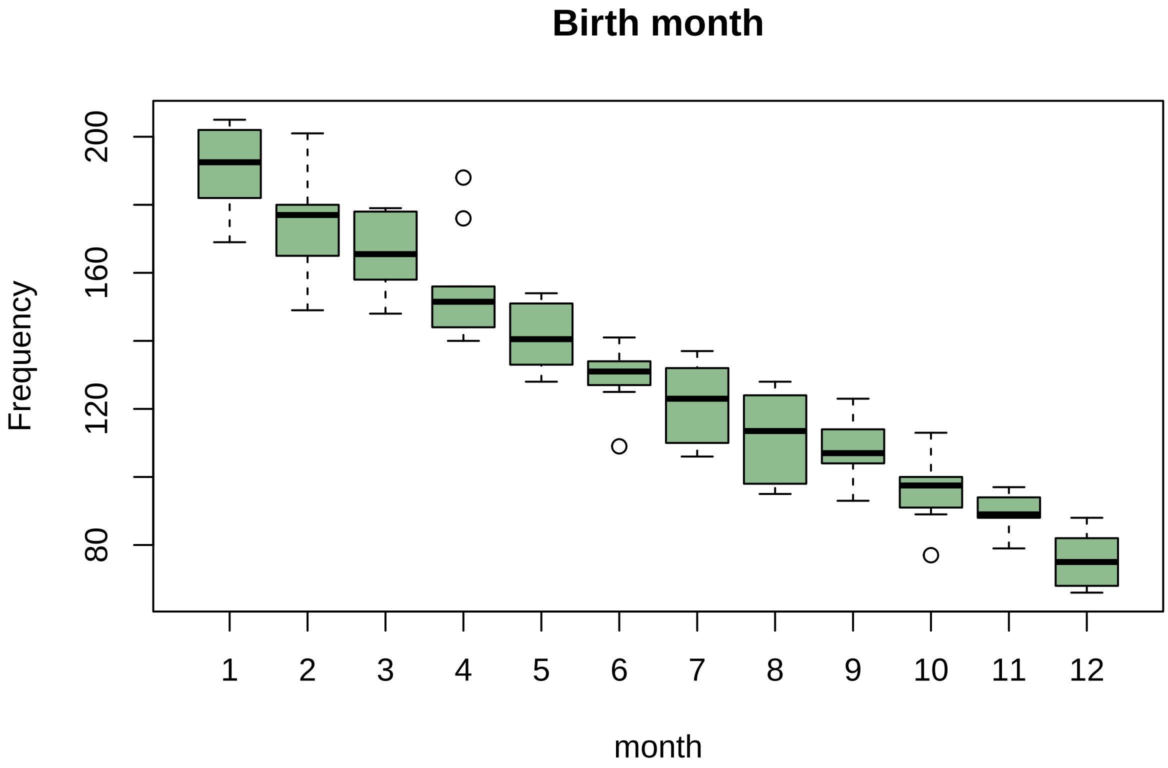
Finally, I added a for loop to perform this selection for 6 years.

```
birthmonth_data <- vector()
frequency_data <- vector()
teamsize = 1e+05
birthmonth = sample(1:12, teamsize, replace = TRUE)
all_birthmonths = birthmonth  # starting point
all_heights = rnorm(teamsize, mean = 116, sd = 4)
for (a in 1:6) {

    growth = rnorm(all_birthmonths, mean = 6, sd = 0.5)/12
    if (a == 1) {
        all_heights = all_heights + (growth * (12 - all_birthmonths))
    } else {
        all_heights = all_heights + (growth * 12)

    }
    median_height <- median(all_heights)

    this_year_birthmonths <- all_birthmonths
    this_year_heights <- all_heights
    this_year_data <- paste(this_year_birthmonths, this_year_heights,
        sep = ":")
    this_year_probs <- rep(0.6, length(this_year_birthmonths))
    this_year_probs[this_year_heights < median_height] <- 0.4

    selected_data <- sample(this_year_data, size = length(this_year_birthmonths)/2,
        prob = this_year_probs)
    selected_birthmonths <- unlist(strsplit(selected_data, ":"))[seq(1,
        length(selected_data) * 2, 2)]
    selected_heights <- unlist(strsplit(selected_data, ":"))[seq(2,
        length(selected_data) * 2, 2)]

    all_birthmonths <- as.numeric(selected_birthmonths)
    all_heights <- as.numeric(selected_heights)
}
```



I can see that there is a clear pattern where the earlier months are over-represented. But, like the first example, I'm still not sure whether there are significant differences between these groups. To test this, I bootstrapped these values by performing the analysis 10 times (I could have done 100, but my computer would have taken too long to complete this). Each time I recorded the number of children selected from each birthmonth at the end of the six years selected. I plotted the results as a boxplot:

```
birthmonth_data <- vector()
frequency_data <- vector()
for (bootstrap in 1:10) {
    teamsize = 1e+05
    birthmonth = sample(1:12, teamsize, replace = TRUE)
    all_birthmonths = birthmonth  # starting point
    all_heights = rnorm(teamsize, mean = 116, sd = 4)
    for (a in 1:6) {

        growth = rnorm(all_birthmonths, mean = 6, sd = 0.5)/12
        if (a == 1) {
            all_heights = all_heights + (growth * (12 - all_birthmonths))
        } else {
            all_heights = all_heights + (growth * 12)

        }
        median_height <- median(all_heights)

        this_year_birthmonths <- all_birthmonths
        this_year_heights <- all_heights
        this_year_data <- paste(this_year_birthmonths, this_year_heights,
            sep = ":")
        this_year_probs <- rep(0.6, length(this_year_birthmonths))
        this_year_probs[this_year_heights < median_height] <- 0.4

        selected_data <- sample(this_year_data, size = length(this_year_birthmonths)/2,
            prob = this_year_probs)
        selected_birthmonths <- unlist(strsplit(selected_data,
            ":"))[seq(1, length(selected_data) * 2, 2)]
        selected_heights <- unlist(strsplit(selected_data, ":"))[seq(2,
            length(selected_data) * 2, 2)]

        all_birthmonths <- as.numeric(selected_birthmonths)
        all_heights <- as.numeric(selected_heights)
    }
    for (b in 1:12) {
        this_month_freq <- length(subset(all_birthmonths, all_birthmonths ==
            b))
        birthmonth_data <- c(birthmonth_data, b)
        frequency_data <- c(frequency_data, this_month_freq)
    }
}
boxplot(frequency_data ~ birthmonth_data, main = "Birth month",
    xlab = "month", col = "darkseagreen", ylab = "Frequency")
```



I could now perform a number of statistical tests on these data to investigate different null/alternative hypotheses. Personally, however, I also like this visual description of the data where you can see both the magnitude of the difference between birth months and make an estimate of the statistical significance by judging whether certain months overlap with each other or now (with a standard p-value threshold of 0.05, you would expect the 95% confidence intervals drawn here to be non-overlapping for groups which are significantly different from each other).

I thought it was interesting that this pattern looked fairly linear where there was a similar decrease in the number of children selected across each month during the year. Pairwise comparisons, e.g. between months 1 and 2, generally did not look significant but there may be more significant differences between children born at least 2 or 3 months apart. You may of course have explored these data in a different way or made different comparisons. This is just as valid as what I have done here and shows the extraordinary flexibility of bootstrapping as a technique.

If we have talked so much about the power of bootstrapping, can you think of some drawbacks or why it is not used for all statistical tests?