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.

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.

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?

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.

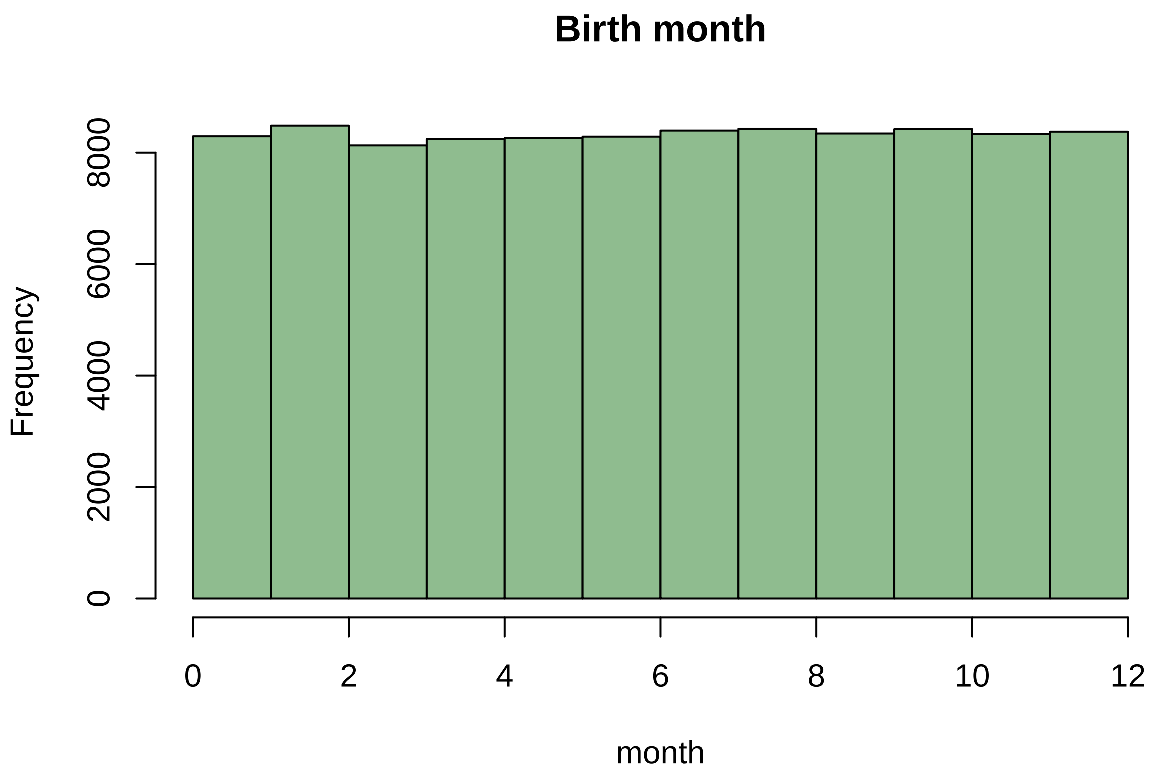
Create an in-silico hockey team

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")
```

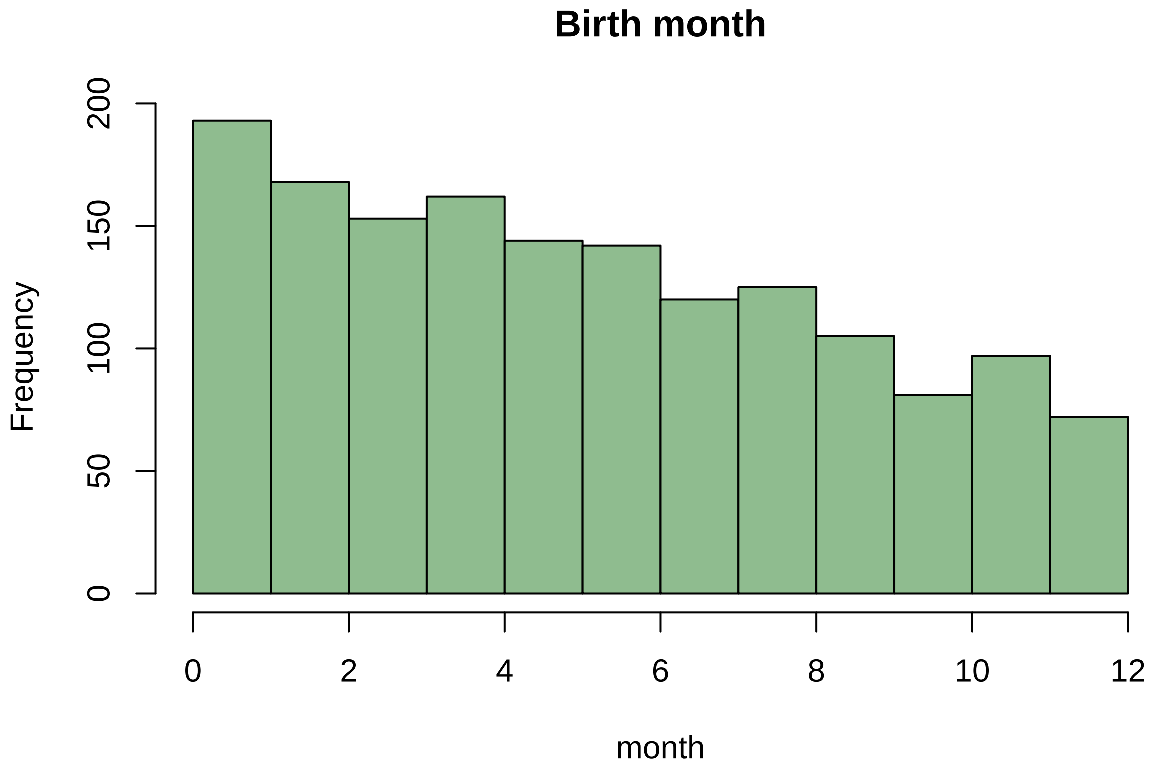


After 6 rounds of selection, as described above, there should be around 1563 players left. What would a histogram of their birth months look like?

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?

Here is what the distribution of birth months looked like when I did it earlier.



But it will look a bit different for you - after all, we used a random process to generate our “team”.

Do you see some birth months being over-represented? Do you think this is effect is statistically significant?

Test for this. Before you do this, sit down and think very carefully:

- What is your Null Hypothesis? What is your Alternative Hypothesis?
- What would your data look like if H0 was true?
- Can you design a statistical test (bootstrap-based or otherwise) and generate a p-value?