Statistics with Sparrows - 05

Julia Schroeder

2022

HO 05

Comparing means

Learning aims

- Understand that effect size is most important when comparing means estimates
- Get insight into why the p-value is only secondary
- Learn how to execute a t-test
- Learn how to judge the biological significance of the results
- Apply the effect of sample size on mean estimate precision

Hypothesis testing

What do you remember about hypothesis testing? I am sure there are some concepts in your brains about p-values and 0.05?

p-value: — the probability of observing a particular result (and more extreme results) when the null hypothesis is true (i.e. there is no effect).

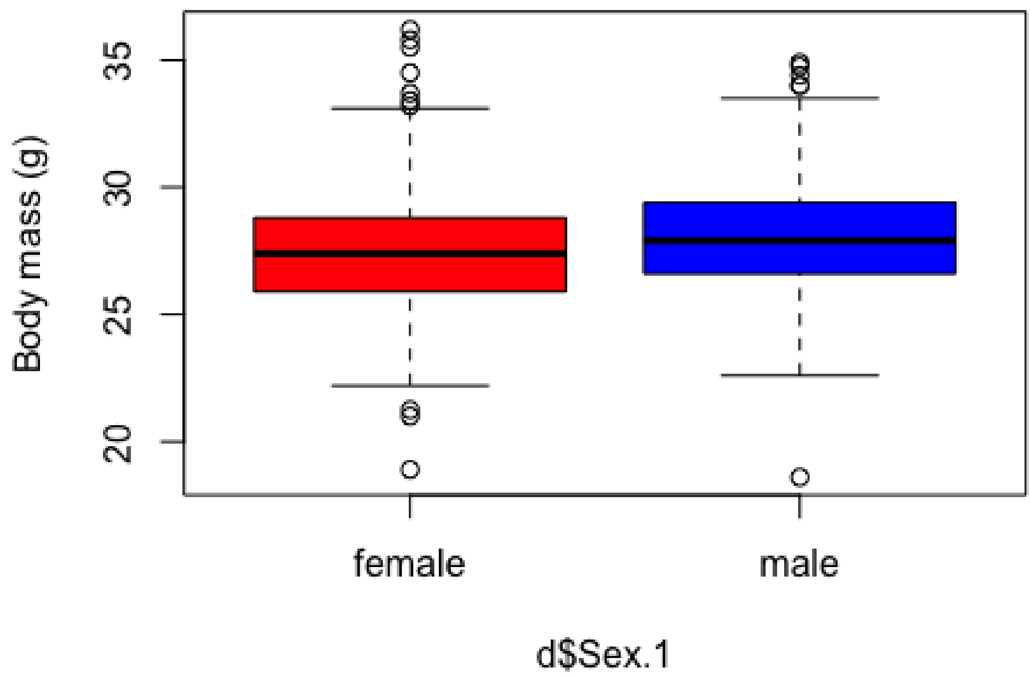
Ok, let's try this out. But first, - you know, housekeeping:

```
rm(list=ls())

d<-read.table("SparrowSize.txt", header=TRUE)
```

We will use it to test for a difference in female and male body mass in house sparrows. Let's see what we can gleam from a boxplot:

```
boxplot(d$Mass~d$Sex.1, col = c("red", "blue"), ylab="Body mass (g)")
```



It looks like males are slightly heavier than females, but how can we tell if that difference means something? We first have to find out our hypothesis. In this case, it is that the DIFFERENCE between males and females in body mass is different from zero. Our null hypothesis is that the difference is zero. If that is not supported, then there are two options: If this difference is positive, it means males have larger body mass, if it is negative, it means males have lower body mass than females.

```
t.test1 <- t.test(d$Mass~d$Sex.1)
t.test1

##
##  Welch Two Sample t-test
##
## data:  d$Mass by d$Sex.1
## t = -5.5654, df = 1682.9, p-value = 3.039e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7669117 -0.3672162
## sample estimates:
## mean in group female   mean in group male
##             27.46852             28.03558
```

R is quite helpful here in that it tells us the alternative hypothesis: that the difference between the two groups is not equal to 0. That is what we found, because the t-value is large, given the degrees of freedom (1683). This means that with a very high probability, males are heavier than females. Should we accept the alternative hypothesis then?

Yes, but wait, does it actually make biological sense? The p-value is super small, most people would be very excited. But I want to teach you to not focus too much on the p-values, instead look at 95%CIs and the parameter estimates. The male mean is 28.0g and the female mean is 27.5g. The difference is about half a gram. However, that's the mean difference. But can we say something about the precision of this effect? Look at the output, it gives us a 95%CI of the difference between males and females: 95% of the differences between males and females fall between -0.77g and -0.37g (males heavier). This gives us a good indication about how important this difference is in biology.

5% of the times, however, the difference will be outside of this interval. That's a type 1 error. There is a 5% chance that this data is actually not representing the real world, and that the difference between the sexes is actually 0.

Large datasets are more likely to pick up on small effect sizes (remember the square root law). Let's see if we would reduce our dataset to the 50 first rows, could we still detect a difference between male and female body mass?

```
d1<-as.data.frame(head(d, 50))
length(d1$Mass)

## [1] 50

t.test2 <- t.test(d1$Mass~d1$Sex)
t.test2

##
##  Welch Two Sample t-test
##
## data:  d1$Mass by d1$Sex
## t = 0.33484, df = 26.84, p-value = 0.7403
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9361255  1.3011254
## sample estimates:
## mean in group 0 mean in group 1
##         27.5200         27.3375
```

Oh. All of the sudden, we find no difference! Just because we took a smaller dataset. The lesson here is that with large datasets, you are more likely to encounter a statistically significant effect, but whether or not this effect is actually meaningful, is not something you can understand by looking at the p-value. In the first t.test, the p-value was very significant, however the real difference (effect size) was very small (0.3g more or less is not a lot, it is actually

Exercises:

1) Test if wing length in 2001 differs from the grand-total mean - Test if male and female wing length differ in 2001 - Test if male and female wing length differ in the full dataset - Test if male and female tarsus differs in the full dataset - Report all your results in a table and share with your group. Discuss what and how to present. Find at least two different ways of presenting your resutls in a table. Explore different table options.

2) Now, run a batch of tests - check for each year whether any of the measures differs from the grand total mean.

Write a word document for (2). In this word document, present the results from (2) using what you've learned in (1), in a delightful and easy to understand way, as you would in a scientific publication. Discuss output with your GTA.

3) Then, see if you understand how to employ the t.test properly in R. There are eleven cohorts (years) in this dataset. Test whether the first five years differ in terms of Tarsus, Mass, Wing, and Bill from the latter six years. To do this, you will first need to create a two-level variable that denotes to wich group a certain observation belongs. Then you can test against this. Do it in two ways - once with , and once with ~.

To do that you will have to re-arrange the data, which might be tricky and take quite some of your time, a lot of brain power, and a lot of de-bugging! All this is something you will need to do quite often over the course of this year. So do try to wrap your head around it. Use all your mighty plyr functions for this. There will be various, vastly different solutions for this one - as long as the end result is the same, that's perfectly fine.

Call on your group for help. Discuss your possible solutions in your group, and with staff.