Biostatistics Portfolio II

Contents

Preface

This is the second part of the biostatistics portfolio. The first part mainly served to ease you into the concepts and provide an introduction to programming in R. If you haven't followed the first part, at least make sure you install RStudio.

This part aims to:

- Bridge the gap between year 1 and year 3;
- Apply the methods that you have learned so far;
- Show some actual analyses of real biological research.

You've now had one course in statistics, but probably little opportunity to ever put it to use. That is what this portfolio is hopefully going to address. Some chapters require you to use what you learned in year 1, others will show you some more advanced methods applied to data of courses you are currently following.

If you really want to understand every detail of analysis, consider pursuing a degree in statistics. For everyone else, just try and understand the gist of it, that is more than enough!

⚒ The portfolio, as well as the online version of the course, are a work in progress. Please check back later for updates. Official information will always be communicated through Brightspace.
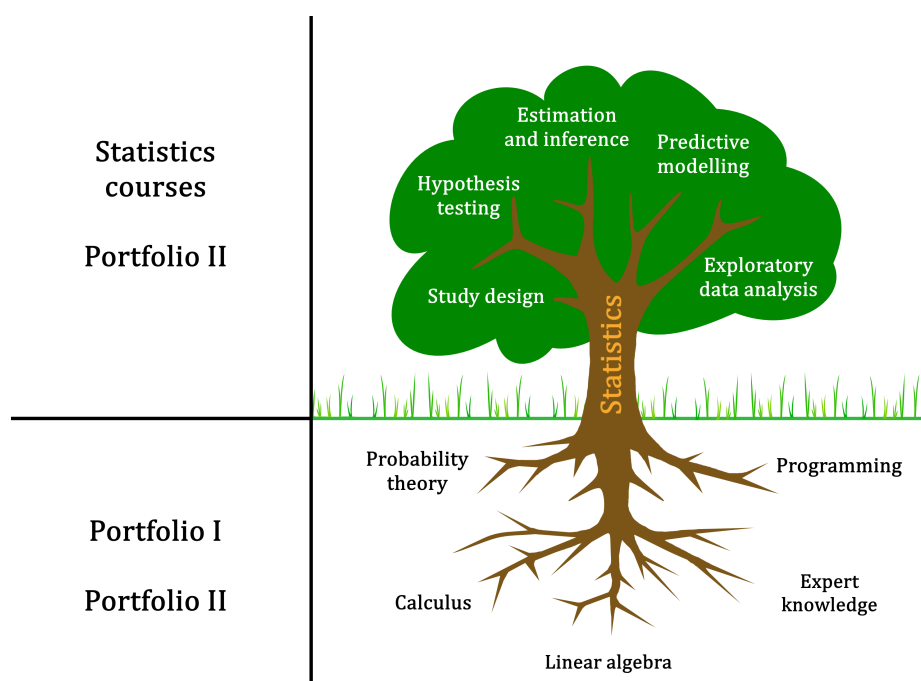
Figure 1: In this part, some of the applications of statistics will be shown in the context of courses you are following at the IBL.

Chapter 1

Bird Song

In this chapter, an advanced analysis is shown of the volume of different areas of the bird brain related to song. Your assignment is to read the explanation, have a look at the analysis used, and then perform **data cleaning** on your own data. Once your data is ready, copy the analysis and write a conclusion.

1.1 Introduction

Statistical analysis is about explaining the different sources of variance. A source of variance can be of interest, like the different areas of the brain, but it can also simply be a source of potential errors, like which student is performing the measurements, or which particular bird you are measuring. By correctly accounting for the different sources of variance, we end up with a powerful model that gives trustworthy standard errors and p-values.

In this analysis I want to demonstrate three things:

1. **A complicated experimental design** like this one requires a complicated model called a mixed model;
2. **Comparing the groups with a mixed model** can fortunately still be fairly easy to do in R, and its results match up with the expectations.
3. **Data cleaning** is part of every analysis, and has a great impact on the results. This is the part you will be doing by yourself. The analysis you can then copy from my example.

This chapter will demonstrate how (1) study design is a crucial part of scientific research, which dictates the type of model you should use; (2) even if you do not

fully understand the (statistical) methods used in papers, you can still judge the validity of one's approach, and read the results; (3) even simple data cleaning can be a time-consuming task that requires careful consideration of how the data came to be.

Throughout this chapter I will use example data from a different year. You data should be (more or less) the same. Any differences you can iron out in the data cleaning assignment.

1.2 A Complex Experimental Design

The experiment consisted of measuring the total volume of different parts of the brain related to song. This experimental design has two difficulties that I want to address, before showing which model I used. Hopefully that will make it clear *why* I used the model shown in the next paragraph, even if you are not familiar with it.
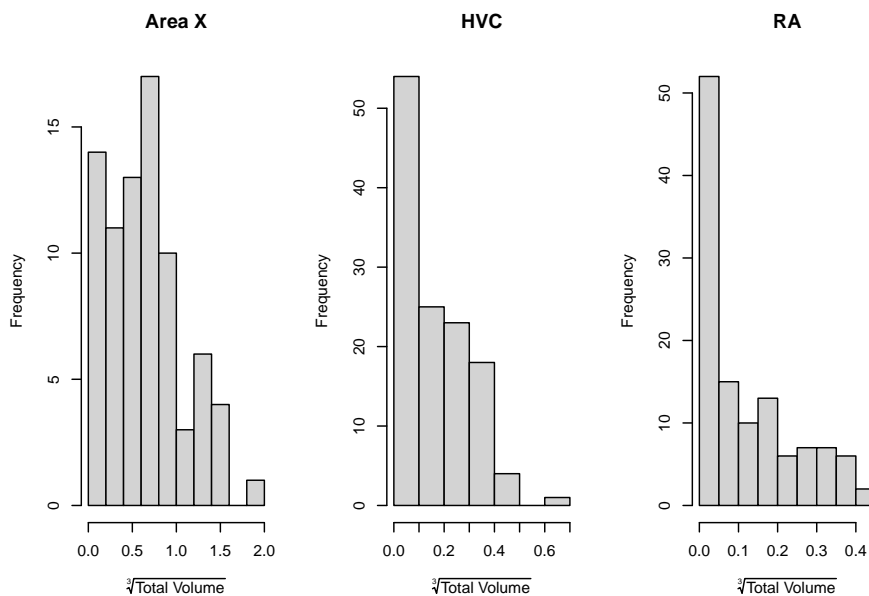
1.2.1 The Outcome Is Skewed

If a distribution has more deviations in one direction than the other, we call it skewed:

Histogram of Total Volume
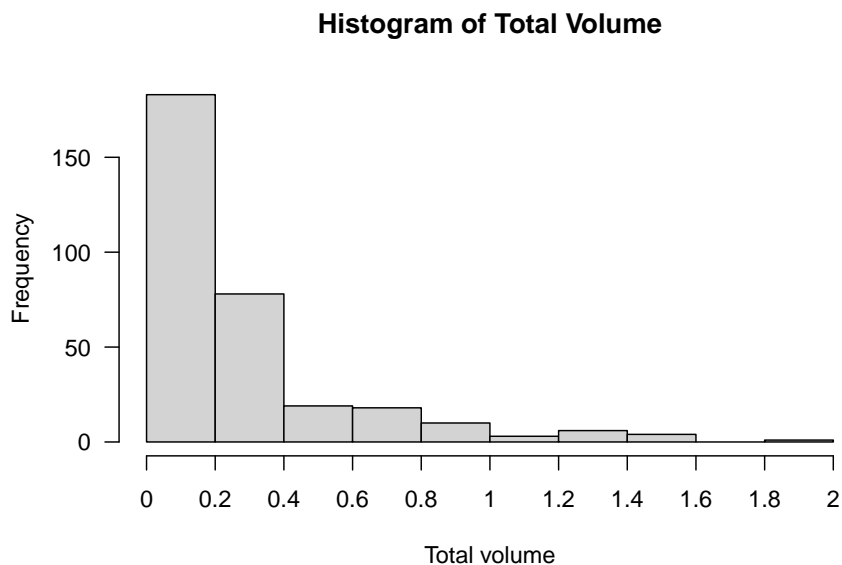
This histogram of the volume displays strong right-skew: Volumes of 0–0.2 are most common, 0.2–0.4 less common, and any larger volumes are exceedingly

rare.

The model we're going to use in the next paragraph assumes (like many statistical tests and models) that if we account for the effects of the explanatory variables, what remains should be more or less normally distributed noise. With a response variable this strongly skewed, it is unlikely that assumption will hold. Even if we split by area of the brain, the distribution of volume is still right-skewed:

Why is volume so strongly right-skewed? Because the volume grows cubically (x^3): Volume is a three-dimensional measurement, determined by the length, width and depth. Therefore, we could consider *transforming* the outcome as follows:

Histogram of $\sqrt[3]{\text{Volume}}$
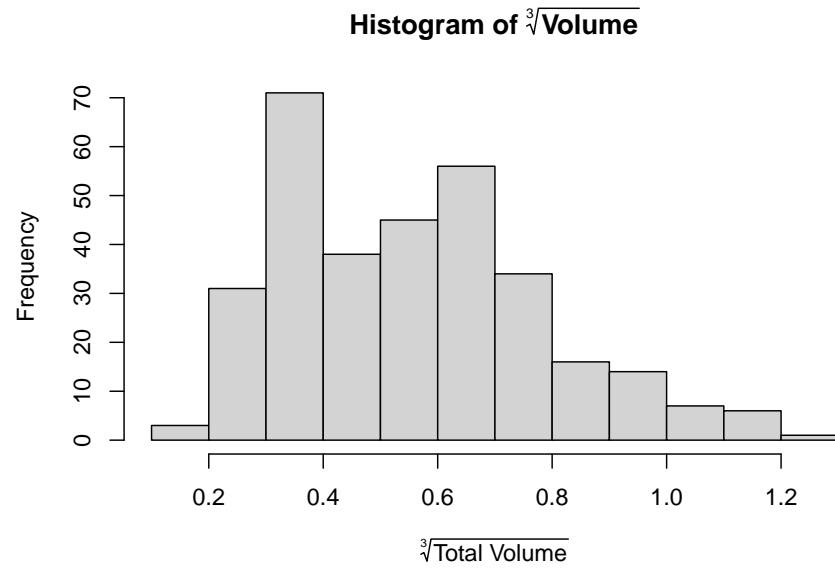
We can also split by area again and see that there is indeed no more right-skew:
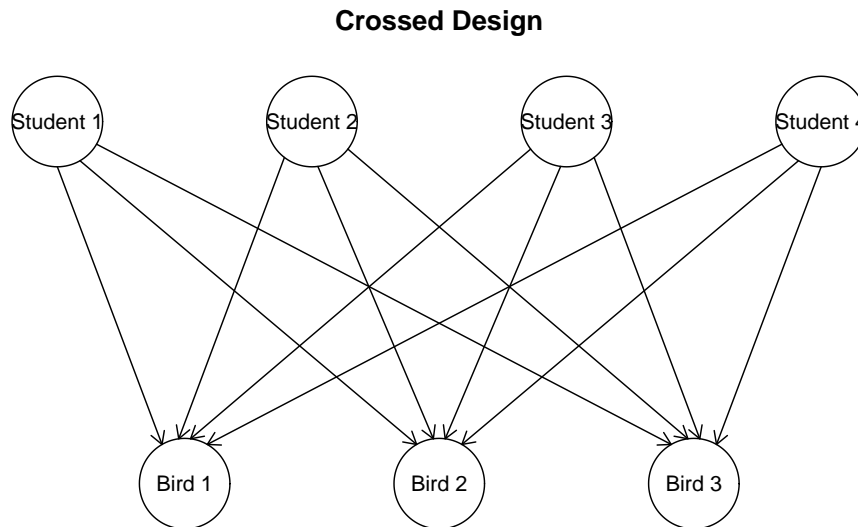
Crossed Design
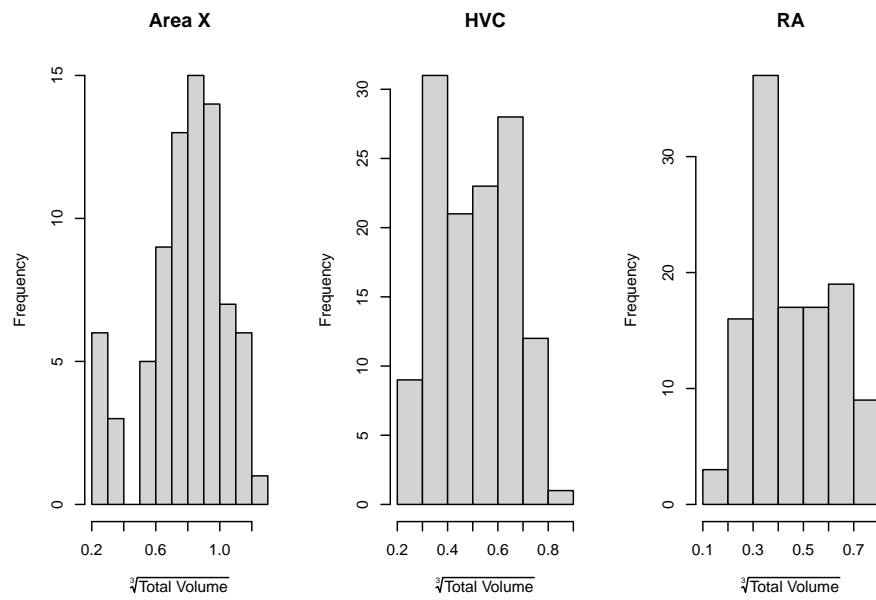
Figure 1.1: A simplified version of the experiment. Each bird has had different parts of its brain measured by different students.

1.2.2 The Measurements are Dependent

Here is a very brief description of **the problem** and **the solution**:

- **Most tests and models assume independent measurements**;
- Each bird in the data set has been measured more than once (different brain areas). These measurements come from the same individual and are therefore, **measurements of the same bird are not independent**;
- In the first year's course about statistics you learned to **compare dependent data** through a paired *t*-test. A paired *t*-test can only compare two sets of paired measurements;
- A more general way to deal with dependent data is **through a mixed model**.
- This works **by estimating a random effect for birds**. This is like assuming that these birds come from some larger population of birds we could have used in the experiment. Instead of estimating a separate effect for each bird, we estimate the variance between birds and subtract the effect of the individual birds from the effects of interest.

Another potential source of dependence is… you! No matter how hard we try to perform measurements according to protocol, there will always be minor, but

systematic differences in the way each student performs the measurement. If each student performed one measurement, this would not pose any problem. However, as shown in figure 1.1, each student has performed several measurements. If the differences in measurement between students are sufficiently large, this poses a problem. Therefore, **measurements of the same student might be correlated**.

Fortunately, both problems can be tackled with the same solution: A **mixed model**.

1.3   Analysis With a Mixed Model

To fit a mixed model, we are going to use a package called `lmerTest`. The entire procedure then boils down to:

1. Install package `lmerTest` if you don't have it;
2. Load package `lmerTest`;
3. Fit a mixed model with `lmer()`;

   - Use the transformation described above;
   - Specify which variables cause the dependence.

1.3.1   Install Package `lmerTest`

```r
install.packages("lmerTest")
```

1.3.2   Load Package `lmerTest`

```r
library("lmerTest")
```

1.3.3   Fit a Mixed Model With `lmer()`

Let's call this model `model1`:

```r
model1 <- lmer(Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID) + (1 | StudentID),
               data = Birdsong)
```

- `lmer` is short for linear mixed effects regression—a mixed model;
- `Volume` is transformed as $\sqrt[3]{x} = x^{\frac{1}{3}}$ as explained earlier;

- `Area` and `Hemisphere` are the effects of interest;
- Using the `(1 | ...)` notation, I am telling the function I want to estimate the variance between birds and the variance between students;
- Using the `data` argument, I am telling the function where to find all these variables.

That's the whole model! You might not have been able to think of that yourself, but if you understand why this particular model was used, that is more than enough.

1.3.4 Interpreting the Output

Let's start by producing a summary of the model:

```
summary(model1)
```

There's a lot going on in the summary, and most of it is not relevant for this assignment. Therefore, the only thing I've show here is the `Random effects` and `Fixed effects` tabs.

```
Random effects:
 Groups    Name        Variance  Std.Dev.
 StudentID (Intercept) 0.0003434 0.01853
 BirdID    (Intercept) 0.0262298 0.16196
 Residual              0.0066445 0.08151
Number of obs: 322, groups:  StudentID, 75; BirdID, 68

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  0.723953   0.023065 99.726496  31.387   <2e-16 ***
AreaHVC     -0.204927   0.013914 63.788016 -14.728   <2e-16 ***
AreaRA      -0.257940   0.014450 60.295630 -17.850   <2e-16 ***
HemisphereR  0.005896   0.010697 46.569237   0.551    0.584
```

Variance between students is extremely small (compare its value of `0.0003434` to that of birds, or the residual (remaining) variance). This is actually good news, it means that the measurements do not depend much at all on who is measuring. You could therefore opt to leave out the random effect for students:

```
model2 <- lmer(Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID), data = Birdsong)
```

1.3.5   Decide on the right model

Lastly, we use a test to decide which of these models fits the data better:

```
anova(model1, model2)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: Birdsong
Models:
model2: Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID)
model1: Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID) + (1 | StudentID)
       npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
model2    6 -484.18 -461.53 248.09  -496.18
model1    7 -482.54 -456.12 248.27  -496.54 0.357  1      0.5502
```

First we are notified that the models were refitted using maximum likelihood. This is because measures like AIC, BIC and deviance cannot be calculated from *restricted* maximum likelihood (the usual fitting procedure for mixed models). In this table we see several things:

- `model2` uses 6 parameters, while `model1` uses 7;
- `model2` has a lower AIC (note the minus sign);
- `model2` has a lower BIC;
- `model2` has a marginally lower log-likelihood;
- `model2` has a marginally higher remaining deviance (again, note the minus sign);
- A chi-squared test on these two values of the deviance produces $\chi^2 = 0.3124$;
- This test has one degree of freedom (the difference in the number of parameters used by the models);
- The corresponding p-value for this χ^2-test on one degree of freedom is $p = 0.5762$;
- This is non-significant for any reasonable level of significance;

That's a lot to consider at once, and most of these terms you have probably never heard of before. Fortunately, the important part is the easiest to interpret: Neither model fits the data significantly better than the other ($p = 0.5762$).

The usual thing to do in a situation like this is to go with the model that uses fewer parameters. After all, if you can fit the data just as well with a simpler model, what is the benefit of making it more complex?

1.3.6 Interpret the final output

```
summary(model2)
```

```
Random effects:
 Groups    Name         Variance Std.Dev.
 BirdID   (Intercept) 0.026279 0.16211
 Residual              0.006966 0.08346
Number of obs: 322, groups:  BirdID, 68

Fixed effects:
             Estimate Std. Error          df t value Pr(>|t|)
(Intercept)  0.722294   0.022743 101.208286   31.759   <2e-16 ***
AreaHVC     -0.202879   0.013069 258.000162  -15.523   <2e-16 ***
AreaRA      -0.256395   0.013540 258.830858  -18.936   <2e-16 ***
HemisphereR  0.005295   0.009895 255.119269    0.535    0.593
```

Here we see that the individual difference between birds is actually quite large: Its variance of `0.026279` is more than $3\times$ the same size of the residual variance of `0.006966`.

We also see that there is a significant difference in volume between the reference group (area X) and the other groups:

- `(Intercept)` is the value of $\sqrt[3]{\text{volume}}$ for the reference group (area X);
- `AreaHVC` is on average -0.20 lower with a significant p-value of $p < 2 \cdot 10^{-16}$;
- `AreaHVC` is on average -0.26 lower with a significant p-value of $p < 2 \cdot 10^{-16}$;
- `HemisphereR` does not differ significantly, with an average difference of 0.0053 between left and right and a p-value of $p = 0.593$.

This matches up nicely with what we saw in the boxplots, but now we have estimates and p-values that have been corrected for the individual differences between birds!

1.4 Assignment: Data Cleaning

So far, I have only covered the experimental design and the model used. But I left out an important part of analysis: Data cleaning.

In the example data above, there were simple errors in the data entry, a problematic outlier, and low quality data mislabeled as 0. Here I am going to show how you can clean your data set to make it ready for analysis.

Your assignment is to clean your own data set following the example here. When you're done, copy the analysis shown above and report on the conclusion.

1.4.1   Read the data and look at the structure

```
## 'data.frame':    376 obs. of  5 variables:
## $ Student_ID  : int  42 42 42 42 42 73 73 73 73 73 ...
## $ Zanggebied  : chr  "RA" "RA" "RA" "RA" ...
## $ Hemisfeer   : chr  "R" "R" "R" "R" ...
## $ Vogel_ID    : chr  "OG164" "PK179" "PK190" "PK198" ...
## $ Totaal_volume: num  0.1104 0.0276 0.2304 0.0204 0.3012 ...
```

In order for this to work:

1. If you haven't already, install RStudio;
2. Download the data. If it is an Excel (.xlsx) file, open it in Excel and save as comma-separated values (.csv) first;
3. Open RStudio and create a new R markdown file. **Save this file to the same folder as the data!**
4. Go to **Session** > **Set Working Directory** > **To Source File Location**. R now knows where to look for your data;
5. Create a new code chunk and run the following code (**replace the name of the data** with the name of yours):

```
Birdsong <- read.csv("Birdsong2020.csv")
str(Birdsong)
```

Here we see the following:

- The data is stored as a `data.frame`, and consists of 376 observations of 5 variables;
- Those variables are:
    - `Student_ID`: An integer value (whole number);
    - `Zanggebied`: A character value;
    - `Hemisfeer`: A character value;
    - `Vogel_ID`: A character value;
    - `Totaal_volume`: A numeric value (any number).

1.4.2   Shorten variable names and change variable types

The first step I made is change the variable names to be slightly shorter (and in English) and convert the categorical variables to factors:

```
colnames(Birdsong)  <- c("StudentID", "Area", "Hemisphere", "BirdID", "Volume")
Birdsong$StudentID  <- factor(Birdsong$StudentID)
Birdsong$Area       <- factor(Birdsong$Area)
Birdsong$Hemisphere <- factor(Birdsong$Hemisphere)
Birdsong$BirdID     <- factor(Birdsong$BirdID)
```

Why does that matter? Well, have a look what happens when you check the structure again:

```
str(Birdsong)
```

```
## 'data.frame':    376 obs. of  5 variables:
##  $ StudentID : Factor w/ 75 levels "1","2","3","4",..: 38 38 38 38 38 65 65 65 65 65 ...
##  $ Area      : Factor w/ 5 levels "Area X","HCV",..: 5 5 5 5 5 1 1 1 1 1 ...
##  $ Hemisphere: Factor w/ 2 levels "L","R": 2 2 2 2 2 1 1 1 1 1 ...
##  $ BirdID    : Factor w/ 69 levels "Bk 145","BK 145",..: 38 59 60 61 68 64 19 49 51 50 ...
##  $ Volume    : num  0.1104 0.0276 0.2304 0.0204 0.3012 ...
```

Now we see that there were apparently 75 students, and… wait, *5* areas?

```
summary(Birdsong$Area)
```

```
## Area X    HCV    HVC     Ra     RA
##    125      5    128      1    117
```

Ah, some people entered `Ra` instead of `RA`, and some entered `HCV` instead of `HVC`. Let's fix their mistakes:

### 1.4.3   Fix errors in data entry

```
Birdsong$Area[Birdsong$Area == "Ra"]  <- "RA"
Birdsong$Area[Birdsong$Area == "HCV"] <- "HVC"
Birdsong$Area <- droplevels(Birdsong$Area)
```

The first two lines correct the wrong entries, and the third line removes the now unused levels `Ra` and `HCV`.

1.4.4 Plot your data

And now we're good to go! The first thing you should do in any analysis is to try and create one or several relevant plots of your data. No amount of looking at numbers is going to give you the same, easy insight as a figure:

```r
boxplot(Volume ~ Hemisphere * Area, data = Birdsong)
```
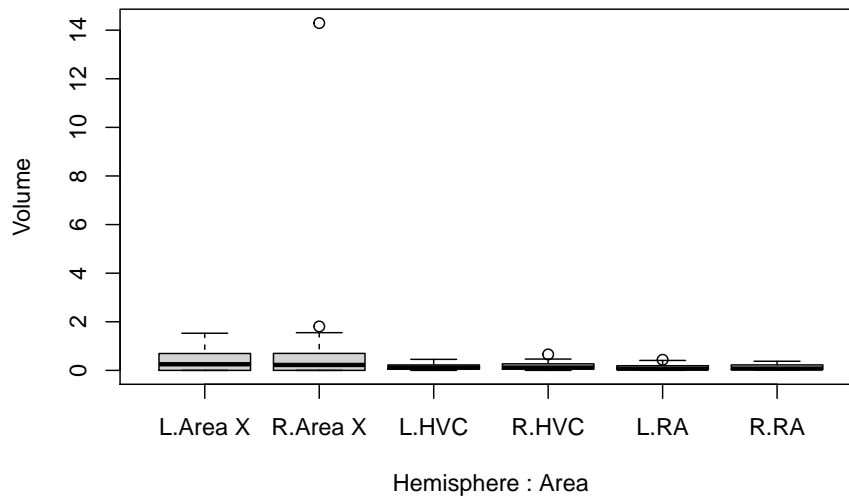

We can already make several useful observations:

- Area X seems to have a larger mean[1] and variance[2] than the rest;
- The left and right hemispheres do not seem to differ much;
- There is a suspiciously large value for area X of the right hemisphere. Is this an unrealistic value?

1.4.5 Inspect suspicious value(s)

Let's inspect the suspicious value further. Since it is the larger value, I can use `which.max` to find it:

[1]If the mean is larger, the location of the boxplot is higher.
[2]If the variance is larger, the size of the boxplot is larger.

```
suspicious <- which.max(Birdsong$Volume)
Birdsong$Volume[suspicious] / mean(Birdsong$Volume)
```

```
## [1] 53.86181
```

This value is 53.9× larger than the mean total volume. Keep in mind that volume is a three-dimensional measurement. It grows at a rate of x^3 with the length, width and depth of the brain. Therefore, at least some extreme observations are expected... So let's use a transformation that accounts for this:

```
boxplot(Volume^(1/3) ~ Hemisphere * Area, data = Birdsong)
```
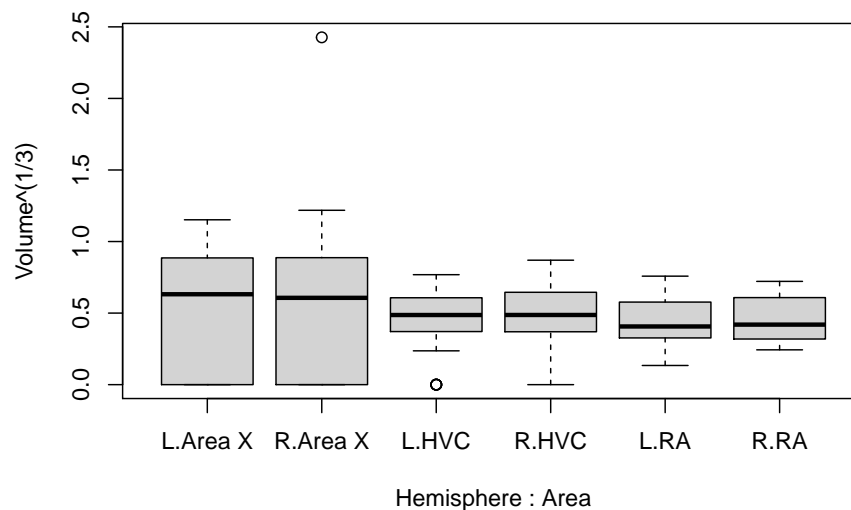

Even with a transformation of volume$^{\frac{1}{3}}$, the suspiciously large value still sticks out quite a lot.

### 1.4.6 Remove unrealistic value(s)

It is hard to say what this value is supposed to be. It could have been a misplaced decimal separator, but should it then be 1.4292, or 0.14292? Since we cannot say for sure what it should have been, let's omit this observation from further analysis:

```
Birdsong <- Birdsong[-suspicious, ]
```

If this analysis were to be used in a publication, we would have to go further:
How much does this observation affect the results? Can we use a form of multiple
imputation to fill in the missing value we have now created? But for now we'll
leave it at this.

Let's create the boxplots again to see what the data look like without the sus-
picious value:

```
boxplot(Volume^(1/3) ~ Hemisphere * Area, data = Birdsong)
```
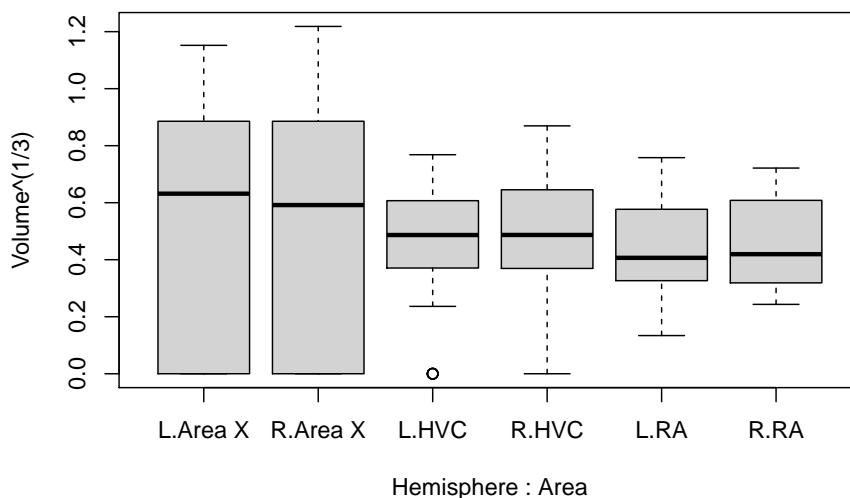

This looks far more reasonable. Now let's address the next elephant in the room:
There appears to be far greater variance in the volume of area X, than in RA
or HVC. But notice how measurements of zero almost exclusively occur in area
X. We can confirm this as follows:

```
by(Birdsong, Birdsong$Area, function(x){sum(x == 0)})
```

```
## Birdsong$Area: Area X
## [1] 45
## ------------------------------------------------------------
## Birdsong$Area: HVC
## [1] 8
```

```
## ------------------------------------------------------------
## Birdsong$Area: RA
## [1] 0
```

How do we deal with this? There are several ways to do so, and the best way depends on the problem being studied, as well as the reason for measurements being 0. We could, for example, consider these zeroes to be 'censored', meaning that they are not actually zero, but just too small to be picked up by the measurement. These kinds of considerations are what make statistical modeling difficult.

### 1.4.7 Remove values incorrectly marked "0"

Fortunately, the cause of the zeroes was in most cases simply a low quality sample, unrelated to the actual volume. It is still suspicious that this happened considerably more often in area X than elsewhere, but for the purpose of this exercise let's consider these measurements to be missing completely at random. Whatever they were, they weren't 0, and this is skewing the results strongly otherwise.

Therefore, in the next step we set these values to 'not available' (`NA`):

```
Birdsong$Volume[Birdsong$Volume == 0] <- NA
```

One final time we inspect the boxplots:

```
boxplot(Volume^(1/3) ~ Hemisphere * Area, data = Birdsong)
```
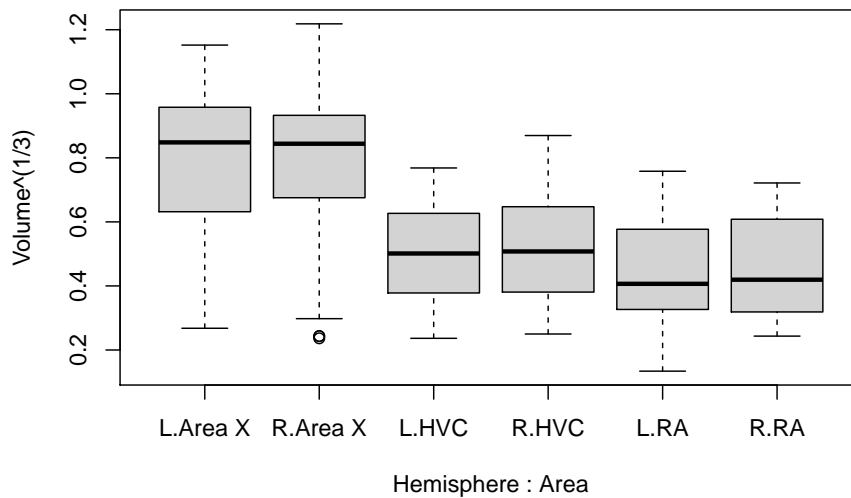
And now we have a fairly good idea of what we should expect in the results: HVC is lower than area X, and RA is even lower. The difference between hemispheres is extremely small and unlikely to be picked up by the model.

1.4.8   Other changes

Are there any other strange values, incorrect labels, or other problems in your version of the data set? Adjust the data in R as you see fit before moving on to the analysis.

1.4.9   Run the analysis

Once you're over here, you can copy the models I've shown above:

```
model1 <- lmer(Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID) + (1 | StudentID),
               data = Birdsong)
model2 <- lmer(Volume^(1/3) ~ Area + Hemisphere + (1 | BirdID),
               data = Birdsong)
```

Do both models run without errors? Can you compare the models like shown before? What do you conclude from the output?