Stats with Sparrows - 14

Julia Schroeder

14

Housekeeping!

```
rm(list=ls())
setwd("H:/StatsWithSparrows")

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

We will examine the whole ANOVA bits again, from a bit of a differnt perspective. We will look at *repeatability*. We run the ANOVA on wing length again with BirdID as explantory factor:

```
d1<-subset(d, d$Wing!="NA")
model3<-lm(Wing~as.factor(BirdID), data=d1)
anova(model3)

## Analysis of Variance Table
##
## Response: Wing
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(BirdID)  617 8147.3 13.2047  8.1734 < 2.2e-16 ***
## Residuals         1077 1740.0  1.6156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Another way of saying what we found here would be that individual bird have consistent wing length. They differ less between multiple measures than they differ from each other. Another way of saying that is saying bird's wing length is *repeatable*. The statistical term is called *repeatablity*, or r, or "intraclass correlation coefficient", and it is important in several ways. It can be used to describe biological things, as we did here. It can however, also, be use to assess the quality of a method - to test for individual observer repeatability, as we did in the lecture.

The repeatability can be calculated in different ways. The simplest way is using the SS and MS of an ANOVA. A very good biologist and statistician, Kate "C" Lessells, published a paper on this, it is called "Unrepeatable repeatabilities: a common mistake". This paper was rejected in many journals, and ultimately was published in a bird journal, "The Auk". It was published in 1987, and since has been cited 2381 times. It has been very successful in teaching generations of biologists how to do the correct statistics. I strongly urge you to read it, maybe best today, as now your knowledge on ANOVA is fresh, and you will understand it much easier!

However, for now, I will also explain how to calculate the repeatability. It is given as

$$r = \frac{s_A^2}{(s_W^2 + s_A^2)}$$

Where s_A is the among-group variance and s_W is the within-group variance. So what we really do here is we calculate the fraction of variance of the total variance $(s_W^2 + s_A^2)$ that is explained by among-group differences. A over total. That's something really important for you to remember. We look at the % of variance explained by among-group differences. How can we do that? Remember how the ANOVA really only describes the variance partitions. We only used the sums of squares because the denominator canceled out. However, remember how calculating SS of among-groups was really complicated, because we had to weigh for the sample size of each group? Yes. That's what this paper is about - this weighing is not as easy as one migh think. It's easy for balanced group sizes. But if they are not balanced, you're in trouble. You can read this up in the Lessells and Boag paper I mentioned above, but for the record:

$$s_W^2 = MS_W$$

and

$$s_A^2 = \frac{MS_A\text{-}MS_W}{n_0}$$

The paper from Kate Lessells pointed out that many people used to just assume that we could just ignore n_0. That is wrong, because MS are not the variance unless we account properly for sample sizes of the groups. So we won't make this mistake. We know where the MS in the function come from, but what exactly is n_0 then? It is most often, actually not the sample size. It would be the sample size n if all group sizes were equal, say, we'd have 10 birds observed once, 10 twice, 10 thrice ect. That would be called a balanced dataset. However, much to our distress, in ecology and evolution, most datasets are far from balanced. Our datasets are often "dirty" and heterogeneous. As we saw earlier, we don't have balanced sample sizes for our BirdID ANOVA here. So we need to calculate n_0. It is calculated by a complicated function:

$$n_0 = [\frac{1}{a\text{-}1}[\sum_{i=1}^{a} n_i \text{-}(\frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i})]$$

This looks horribly complicated. But, as most things in life, it is not. Check this out: a is the number of groups. n_i is the sample size in each group. In our BirdID example, it's 618. 618 individual birds, and n_1 would be 1, n_2 is 10, and so forth. We know this from this code:

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
d1 %>%
  group_by(BirdID) %>%
  summarise (count=length(BirdID))
```

```
## # A tibble: 618 × 2
##    BirdID count
##     <int> <int>
## 1       4     1
## 2      11    10
## 3      13     1
## 4      16     1
## 5      22     9
## 6      23     4
## 7      24     2
## 8      28     1
## 9      29     1
## 10     32     3
## # ... with 608 more rows
```

We could also find it out a in a more elegant way, using this code:

```
d1 %>%
  group_by(BirdID) %>%
  summarise (count=length(BirdID))  %>%
    summarise (length(BirdID))
```

```
## # A tibble: 1 × 1
##   `length(BirdID)`
##              <int>
## 1              618
```

Cool. So we know now what a and n_i are. Next, we are told to do lots of sums. First, let's look at the inner most bracket, the fracture. We sum the square values of each group's n, and divide this sum by the unsquared sum of the ns of each group. That's easy enough, don't you think? Dplyr to the rescue:

```
d1 %>%
  group_by(BirdID) %>%
  summarise (count=length(BirdID)) %>%
  summarise (sum(count))
```

```
## # A tibble: 1 × 1
##   `sum(count)`
##          <int>
## 1         1695
```

Uhh. That's the denominator = 1695. The numerator is similar, but we square the count (n) of each group first:

```
d1 %>%
  group_by(BirdID) %>%
  summarise (count=length(BirdID)) %>%
  summarise (sum(count^2))
```

```
## # A tibble: 1 × 1
##   `sum(count^2)`
##            <dbl>
## 1           7307
```

Cool. That's 7307. Ok, R, this is easy, the fraction is:

```
7307/1695
```

```
## [1] 4.310914
```

Good. What's next? We're supposed to substract this fraction from the denominator. Clever, huh, we don't have to calculate that monster again. We certainly can do that:

```
1695-7307/1695
```

```
## [1] 1690.689
```

Even better. The last part is easy. a is 618, so it's really 1 over 617. We can do that:

```
(1/617)*(1695-7307/1695)
```

```
## [1] 2.740177
```

And that's our n_0, excellent. It is actually pretty close to 3, which you can see as a sort of a centrality measure of how many observations we have for each group. Because the differences are so extreme, we can't use a mean or so, we've got to use this monster of n_0.

But since you've become so handy with sums of squares, you probably have some understanding where this comes from, and what it corrects for.

Let's finally calculate the repeatability! I already forgot the equation - I have really bad memory, I can only remember we needed to divide the part of variation that's explained by among group differences by the total variation, corrected for n_0. Here is the equation again:

$$r = \frac{s_A^2}{(s_W^2 + s_A^2)}$$

with

$$s_W^2 = MS_W$$

and

$$s_A^2 = \frac{MS_A\text{-}MS_W}{n_0}$$

and (monster)

$$n_0 = [\frac{1}{a\text{-}1}[\sum_{i=1}^{a} n_i \text{-}(\frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i})]$$

which is 2.74.

Here is out ANOVA result:

```
model3<-lm(Wing~as.factor(BirdID), data=d1)
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: Wing
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(BirdID)  617 8147.3 13.2047  8.1734 < 2.2e-16 ***
## Residuals         1077 1740.0  1.6156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, really, the repeatability is:

```
((13.20-1.62)/2.74)/(1.62+(13.20-1.62)/2.74)
```

```
## [1] 0.7229006
```

W00oot!!11!1!11. That was hard work. The result of r is a fraction. We can multiply it with 10 and then get a percentage. Now we know that 72% of the variation in wing length is

determined by between-individual differences. That equally means that individuals are relatively consistent in their wing length - most of the variation comes from differences between individuals. Very cool. The concept of repeatability is an important one, and you should remember this for the rest of your life, as everything, really in this course. Over the time you spend here in Silwood, you will come across other methods of calculating the repeatability, but the principles stay the same: we look at the ratio of variance explained among vs between groups.

Exercise:

Calculate the repeatability of body mass within individual birds. The N0 will be different, so you will have to re-do the dplyr stuff for a dataset subsetted to no NAs in body mass!