Lesson 3 AY23

Nicholas Clark

Questions from previous lesson?

Notation:

$$
\begin{aligned}
i &= \text{group} \\
j &= \text{observation} \\
y_{i,j} &= \text{response from the } j\text{th observation from the } i\text{th group} \\
\bar{y} &= \text{Overall sample mean} \\
\bar{y}_i &= \text{Sample mean of the ith group} \\
n &= \text{total sample size} \\
n_i &= \text{group size of the ith group}
\end{aligned}
$$

From last class we had

```
library(tidyverse)
scent.dat<-read.table("http://www.isi-stats.com/isi2/data/OdorRatings.txt",header=TRUE)
```

which we fit to the model:

$$
\begin{aligned}
y_{i,j} &= \mu + \epsilon_{i,j} \\
\epsilon_{i,j} &\sim \text{iid } F(0, \sigma)
\end{aligned}
$$

Note here we don't have any structure to our population mean, μ. Recall that μ is a *parameter* in our model. In order to fit this we could simple set:

$$\hat{\mu} = \bar{y}$$

One way to answer the question of how appropriate this model is for our data is to look at the standard error of our residuals, which is an estimate of the standard deviation of $\epsilon_{i,j}$. To find this we could do:

```
ybar=mean(scent.dat$rating)
sq.resid=(ybar-scent.dat$rating)^2
SSE=sum(sq.resid)
```

Recalling from our previous classes, we then want to divide by $n-1$ in order to have an unbiased estimate of σ^2, or we find the SE of the residuals by:

```
sqrt(SSE/(nrow(scent.dat)-1))
```

```
## [1] 1.271447
```

As our book mentions, the $n-1$ is also the number of degrees of freedom in our model which can be though of as the number of independent values in our model (which I always found a bit confusing). But if we know our mean is 4.48, we only have $n-1$ unique data points. The final data point is determined through knowing $\hat{\mu} = 4.48$

An alternative model is:

$$y_{i,j} = \mu_i + \epsilon_{i,j}$$
$$\epsilon_{i,j} \sim \text{iid } F(0, \sigma)$$

Note here we've put additional structure on μ. It would make sense to estimate $\hat{\mu_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1,j}$ and $\hat{\mu_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2,j}$. Our Sum of square errors now is:

```
group1 = scent.dat %>% filter(condition=="scent")%>%summarise(scent.mean=mean(rating))
group2 = scent.dat %>% filter(condition=="noscent")%>%summarise(non.scent.mean=mean(rating))
SSE1 = scent.dat %>% filter(condition =="scent")%>%
  mutate(SqErr = (rating-group1$scent.mean)^2)%>%summarize(SSE.group1=sum(SqErr))
SSE2 = scent.dat %>% filter(condition =="noscent")%>%
  mutate(SqErr = (rating-group2$non.scent.mean)^2)%>%
  summarize(SSE.group2=sum(SqErr))
SSE1+SSE2
```

```
##   SSE.group1
## 1   55.95833
```

Now note here we are estimating both μ_1 and μ_2 so we lose two degrees of freedom, so our estimate of our SE is:

```
sqrt((SSE1+SSE2)/(nrow(scent.dat)-2))
```

```
##   SSE.group1
## 1   1.102944
```

So we see that our error in our model has went from 1.27 to 1.10. So we have explained more of our variation by using two means instead of one and, perhaps more importantly in this case, we've addressed our scientific question.

In terms of our model parameters, our **effect size** is $\mu_1 - \mu_2$ which we estimate by:

While the seperate means model is useful, our book also considers a model that looks like:

$$y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}$$
$$\epsilon_{i,j} \sim \text{iid } F(0, \sigma)$$

How is this different? Can we fit this model?

In R we can do:

```
scent.dat <- scent.dat %>% mutate(condition = as.factor(condition))
contrasts(scent.dat$condition)=contr.sum
lm.mod=lm(rating~condition,data=scent.dat)
summary(lm.mod)
```

```
##
## Call:
## lm(formula = rating ~ condition, data = scent.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8333 -0.8333 -0.1250  0.8750  2.1667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4792     0.1592  28.136  < 2e-16 ***
## condition1   -0.6458     0.1592  -4.057 0.000191 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 46 degrees of freedom
## Multiple R-squared:  0.2635, Adjusted R-squared:  0.2475
## F-statistic: 16.46 on 1 and 46 DF,  p-value: 0.0001907
```

To see what this is fitting, consider the model:

$$
\begin{aligned}
&y_{i,j} = \mu + x_{i,j}\beta + \epsilon_{i,j} \\
&x_{i,j} = 1 \text{ if observation i,j is in scent group 0 otherwise} \\
&\epsilon_{i,j} \sim \text{iid } F(0, \sigma)
\end{aligned}
$$

This is the standard model that R would fit. In this case what is μ?

```
contrasts(scent.dat$condition)=contr.treatment
lm.mod=lm(rating~condition,data=scent.dat)
summary(lm.mod)
```

```
##
## Call:
## lm(formula = rating ~ condition, data = scent.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8333 -0.8333 -0.1250  0.8750  2.1667
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8333     0.2251  17.027  < 2e-16 ***
## condition2    1.2917     0.3184   4.057 0.000191 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 46 degrees of freedom
## Multiple R-squared:  0.2635, Adjusted R-squared:  0.2475
## F-statistic: 16.46 on 1 and 46 DF,  p-value: 0.0001907
```

Using the books parameterization, if we want to determine whether scents matter, what parameters are we looking at?

If we want to use Rs default parameterization what are we looking at?

In order to determine if the scents matter, we want to examine how much our group means vary from our overall means. In other words, we want a measure of:

$$\sum_{i,j}(\bar{y}_i - \bar{y})^2 = n_j \sum_{i}(\bar{y}_i - \bar{y})^2$$

This statistic is called the Sums of Squares due to Treatment, or Sums of Squares due to model for the two means model. If this number is really big, what can we say about our effect size?.

Note that the calculations in some cases can be extremely simplified as $n_1 = n_2$ we must have $\bar{y}_1$ and $\bar{y}_2$ to be equadistant from $\bar{y}$, so we end up with $\sum_{i,j}(\bar{y}_i - \bar{y})^2 = n_1\hat{\alpha}_1^2 + n_2\hat{\alpha}_2^2$. Now this doesn't always hold true, so remember what we are doing is $\sum_{i,j}(\bar{y}_i - \bar{y})^2$ and we will be fine. Note here we must have $\alpha_1 + \alpha_2 = 0$ so our degrees of freedom are 1.

The calculations we have done can be summarized as:

Let's draw a picture

One measure of how well our model explains the variation is the oft-mis quoted R^2. From our previous courses, what is R^2?

Another that we can compare is the effect size, or the differences in group means, compared to the standard error of ther residuals. Again a picture:

If our model does a good job of explaining our data most of our variability will be due to the treatment we are applying or examinig. So if we look at the amount of variation that's remaining (SSE) and compare it to the SSM the ratio should be small. Note that SST=SSM+SSE.

Let's look at Exploration 1.2 (Page 50 in our text). The data can be obtained from:

```
dung_data <- read.table("http://www.isi-stats.com/isi2/data/DungBeetles.txt", header=T)
```

1. Explain how this is an expertiment rather than an observational study. Identify the response variable and the explanatory variable. What are the treatments?

2. Record the overall mean and standard deviation for the times. Use these values to write out a "single-mean" statistical model for predicting the time to reach the edge. Then write out the fitted model.

3. Using R find the standard error of the null model. How does this relate to the SD of the times on page 52? Hint: if you use lm(time~1,data=dung_data) what model are you fitting?

4. Using your lm object from above, run anova() on that object. Examine the Sums of Squares, confirm that it is equal to the equation given in problem 6 in our text on page 52

5. Now let's consider the separate means model. Using R come up with the means of the treatment groups

6. Write out the statistical model for separate means, then write out the fitted model

7. Fit the model in R, What is the SE of the residuals of the separate means model?

8. Use this and the equation on the bottom of page 52 to find the SSError

9. Using your lm object, run predict(yourlm). This gives us $\hat{y}$. Use these predicted values to confirm that SSError can also be found through the equation given on the top of page 53.