Lesson 4

Nicholas Clark

Questions from previous exploration?

Notation:

$$j = \text{group}$$
$$i = \text{observation}$$
$$y_{i,j} = \text{response from the } i\text{th observation from the } j\text{th group}$$
$$\bar{y} = \text{Overall sample mean}$$
$$\bar{y}_j = \text{Sample mean of the jth group}$$
$$n = \text{total sample size}$$
$$n_j = \text{group size of the jth group}$$

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:

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

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))
```

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:

1

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

Note here we've put additional structure on μ. It would make sense to estimate $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_i} y_{i,1}$ and $\hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_i} y_{i,2}$. 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
```

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))
```

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:

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

How is this different? Can we fit this model?

In R we can do:

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

To see what this is fitting, consider the model:

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

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)
```

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:

$$\sum_{i,j}(\bar{y}_j - \bar{y})^2$$

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 $\sum_{i,j}(\bar{y}_j - \bar{y})^2 = n_1\hat{\alpha}_1^2 + n_2\hat{\alpha}_2^2$. Now this doesn't always hold true, so remember what we are doing is $\sum_{i,j}(\bar{y}_j - \bar{y})^2$ and we will be fine. Note here we must have $\alpha_1 + \alpha_2 = 0$ so our degrees of freedom are 1.

This calculation is called Sums of square due to model or SSM.

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: