Lesson 5

Nicholas Clark

What is a Statistic?

Review

Last class we fit the model:

$$
\begin{aligned}
i &= \text{Scent status} \\
j &= \text{Student} \\
y_{i,j} &= \text{Store rating from the } j\text{th observation exposed to the } i\text{th scent} \\
y_{i,j} &= \mu + \alpha_i + \epsilon_{i,j} \\
\epsilon_{i,j} &\sim F(0, \sigma)
\end{aligned}
$$

Which we obtained estimates of:

```
library(tidyverse)
scent.dat<-read.table("http://www.isi-stats.com/isi2/data/OdorRatings.txt",header=TRUE)
contrasts(scent.dat$condition)=contr.sum
scent.mod<-lm(rating~condition,data=scent.dat)
coef(scent.mod)
```

$$
\begin{aligned}
\hat{\mu} &= 4.48 \\
\hat{\alpha}_{scent} &= 0.645 \\
\hat{\alpha}_{noscent} &= -0.645
\end{aligned}
$$

From here we also obtained SSTotal, $\sum_{i,j} (y_{i,j} - \bar{y})^2$ which is also SSE from the null model

```
sum((scent.dat$rating-mean(scent.dat$rating))^2)
```

Which can be decomposed into SSTreatment (or SSModel) and SSError. Recall SSE=$\sum_{i,j} (y_{i,j} - \hat{y_{i,j}})^2$

```
sum((scent.dat$rating-scent.mod$fitted.values)^2)
```

And SSM = $\sum_{i,j} (\bar{y} - \hat{y_{i,j}})^2$

```
sum((scent.mod$fitted.values-mean(scent.dat$rating))^2)
```

So the amount of variation that the model explains is $\frac{SSM}{SST} \approx 0.26$

So what?

But is it possible that scent actually doesn't explain the variation in the model? What if we just got unlucky and randomly assigned people who like stores to the scent group and folks that despise shopping to the no scent group?

To answer this, we want to conduct step 4 of our statistical investigation process **Draw inferences beyond the data**. Up to this point we have merely been exploring the data.

Recall that a null hypothesis and an alternative hypothesis are:

So here H_0 can be written as:

And H_a can be written as:

Our parameter that addresses our hypothesis is:

And our statistic that will be used to test our hypothesis is:

If we just got unlucky and assigned our folks who like stores to the scent data but there is no scent effect, which hypothesis is true?

So, we want to get a feel for how likely is it that we just got unlucky. If the null hypothesis was true than how we labled our students (scent or noscent wouldn't matter)

```
head(scent.dat)
```

So, again, if there was no effect for scent, it wouldn't matter that these six were called 'scent'. We could call them noscent or call the first scent, the second no scent, the third scent, etc. If H_0 is true, condition label doesn't matter.

So let's rearrange them:

```
scent.dat.rearr = scent.dat %>% mutate(new_cond=sample(scent.dat$condition))%>%
  select(-condition)
```

Now under this labeling sceme we would have observed a difference of:

```
diffs=scent.dat.rearr %>% group_by(new_cond) %>%
  summarize(avg=mean(rating))%>%select(avg)
diffs[2,]-diffs[1,]
```

Now if I shuffled these labels again obviously I'd get something different. We can do this 1000 times to build up an empirical estimate of the difference of the sample means assuming that the there is no difference in the

two groups:

```
M <- 1000
emp.est<-data.frame(shuffle=seq(1,M),stat=NA)#Blank data set to fill in
for(i in 1:M){
  scent.dat.rearr = scent.dat %>% mutate(new_cond=sample(scent.dat$condition))%>%
  select(-condition)
  diffs=scent.dat.rearr %>% group_by(new_cond) %>%
  summarize(avg=mean(rating))%>%select(avg)
  emp.est[i,]$stat=as.numeric(diffs[2,]-diffs[1,])
}
```

So, assuming there is no difference in the means, we could see values that fall as:

```
ggplot(data=emp.est,aes(x=stat))+geom_histogram()
```

Recall that our actual difference, $\hat{\alpha}_{scent} - \hat{\alpha}_{noscent} = 1.3$. Assuming H_0 is true, where would this value fall on our histogram?

We can quantify now rare this is by:

```
emp.est %>%filter(abs(stat) > 1.29)%>%summarize(num=n()/1000)
```

So what does this mean?

Statistic, simulate, strength of evidence

There are a ton of different choices for statistics to use. It was convenient for us to use $\hat{\alpha}_{scent} - \hat{\alpha}_{noscent}$, but there are other, standardized statistics that have natural distributions we will take advantage of once we start making more assumptions about $\epsilon_{i,j}$. One is the **pooled t- statistic**, which is:

```
M <- 1000
emp.est<-data.frame(shuffle=seq(1,M),stat=NA)#Blank data set to fill in
for(i in 1:M){
  scent.dat.rearr = scent.dat %>% mutate(new_cond=sample(scent.dat$condition))%>%
  select(-condition)
  sim.lm<-lm(rating~new_cond,data=scent.dat.rearr)
  diff.means=2*coef(sim.lm)[2]
  est.se = summary(sim.lm)$sigma
  #summary(lm(speed~dist, cars))$r.squared for R^2
  emp.est[i,]$stat=diff.means/(est.se*sqrt(2/24)) #Form t-statistic
}
```

Which again we can visualize:

```
ggplot(data=emp.est,aes(x=stat))+geom_histogram()
```

Our data had a realized t-statistic value of $1.29/(1.10\sqrt{2/24}) = 4.03$, which we can see is quite rare:

```
emp.est %>% filter(abs(stat)>4.03)%>% summarise(tot=n()/M)
```

The t-stat is nice in that under H_0 it is centered at zero and has sd=1, but perhaps more importantly if we

make an additional assumption we have a known distribution for the t stat. Our assumption is $\epsilon_{i,j} \sim N(0, \sigma)$. This can actually be relaxed a bit since we'll never know the actual distribution of ϵ. Our book gives these as 3 conditions:

- Samples are independent of one another
- Sample SDs for each group are roughly equal
- Distributions of the data are roughly symmetric with no outliers

If this is true, then $t_{stat} \sim t_{n_1+n_2-2}$. With this in hand we can analytically calculate the probability of having observed something as extreme or more under H_0.

```
t.stat=4.03
2*(1-pt(t.stat,48-2))
```

A picture:

```
x=data.frame(val=rt(10000,46))
ggplot(data=emp.est,aes(x=stat))+geom_histogram(aes(y=..density..))+
  geom_density(data=x,aes(x=val),lwd=2,col="red",adjust=4)
```

Once we have made our additional assumptions we can now easily find things like confidence intervals. Most CIs have the form:

Here our multiplier is $t_{n_1+n_2-2}(1-\alpha/2)$ and the standard error of the statistic is $\hat{\sigma}\sqrt{1/n_1+1/n_2}$

```
alpha=.05
n1=24
n2=24
SSE = sum((scent.dat$rating-scent.mod$fitted.values)^2)
multiplier=qt(1-alpha/2,n1+n2-2)
stat.se = SSE/(n1+n2-2)*sqrt(1/n1+1/n2)
1.29+stat.se*multiplier
1.29-stat.se*multiplier
```

Now that we have our results we can conduct step 5, formulate conclusions, and step 6, look back and ahead. Look very carefully at pages 72 and 73 to see how our book makes a conclusion. What I really like is the last sentence, "The scent model explained 26 % of the variability in the ratings, probably meaninful enough for a store manager to care as this corresponds to about a 1 point increase on a seven point scale"

Note the CI on page 72 of our text is wrong. It should be multiplied by 2.013 not 2.103.

Exploration