Lesson 6

Clark

Common Errors on Exploration 1

Writing out model as lm~0

Confounding definition

Confounding can still occur in an experiment by chance

Dealing with Factors in R

Multiple Groups

Fish consuption and Omega-3. The research question is, "Is there an association between the amount of fish consuption and omega-3 fatty acids level in the blood"

30 generally healthy adults volunteered to participate in a study and were asked a single question, How often do you eat tuna or other non-friend fish: 1 or fewer times a month, 2-3 times per month, 1 time per week, or morethan 2 times a week? The individuals then had their omega-3 fatty acids measured directly via a blood test.

Why is this an observational study?

Our sources of variation diagram is:

Our associated statistical model is:

We can fit this model using:

```
library(tidyverse)
omega.dat<-read.table('http://www.isi-stats.com/isi2/data/FishOmega3.txt',fill=TRUE)


omega.df<-data.frame(omega3=as.numeric(as.character(omega.dat$V1[-1])),omegaper=as.numeric(as.character

omega.df <- omega.df %>% filter(omegaper < 9)

group.means<-omega.df%>%group_by(fishcat)%>%summarize(mean=mean(omegaper),tot=n())
group.means
```

```
## # A tibble: 5 x 3
##   fishcat               mean   tot
##   <fct>                <dbl> <int>
## 1 ">2 times/week  "     5.28     5
## 2 1 or fewer times/mo.  3.77     6
## 3 "1 time/week  "       5.10     5
## 4 "2-3 times/month  "   4.08     6
## 5 "2x/week   "          5.65     5
```

Note that this can also be fit by:

```
contrasts(omega.df$fishcat)<-contr.sum
omega.lm<-lm(omegaper~fishcat,data=omega.df)
summary(omega.lm)
```

```
##
## Call:
## lm(formula = omegaper ~ fishcat, data = omega.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05940 -0.42013 -0.08127  0.50630  1.25220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7742     0.1553  30.736  < 2e-16 ***
## fishcat1      0.5089     0.3189   1.596  0.12478
## fishcat2     -1.0087     0.2979  -3.386  0.00266 **
## fishcat3      0.3249     0.3189   1.019  0.31936
## fishcat4     -0.6976     0.2979  -2.341  0.02867 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8039 on 22 degrees of freedom
## Multiple R-squared:  0.5057, Adjusted R-squared:  0.4158
## F-statistic: 5.627 on 4 and 22 DF,  p-value: 0.002825
```

What is being fit here?

Note here that μ is NOT the overall mean

```
mean(omega.df$omegaper)
```

```
## [1] 4.71103
```

What this value actually is the mean of the means

```
mean(group.means$mean)
```

```
## [1] 4.774224
```

Why might we want to use the mean of the means?

To see how well our multiple group model fits the data we can see how much of our variation our model accounts for.

To do this we first find SST, recall SST is:

```
SST=sum((omega.df$omegaper-mean(omega.df$omegaper))^2)
```

Instead of directly finding SSM we will find SSE first:

SSE is:

```
SSE=sum((omega.df$omegaper-omega.lm$fitted.values)^2)
```

So, SSM can be found from:

```
SSM=SST-SSE
SSM
```

```
## [1] 14.54518
```

Therefore R^2 is

```
1-SSE/SST
```

```
## [1] 0.5056906
```

What can we conclude at this point?

But we probably want to say something about the population, not about our study, so we want to test some hypothesis:

One statistic we can use to test this hypothesis is R^2. So we have to get a feel for the distribution of R^2 under H_0.

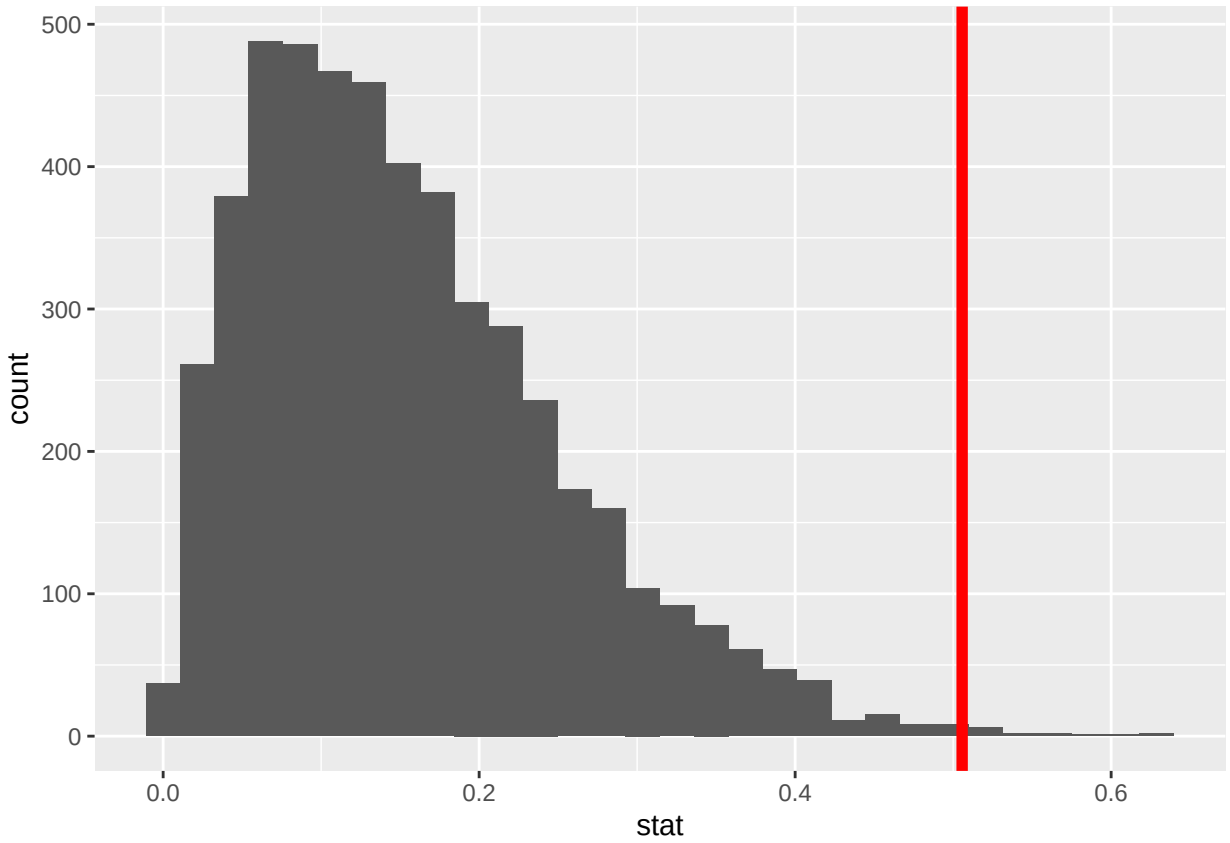
```
M<-5000
stats.df<-data.frame(trial=seq(1,M),stat=NA)

for(j in 1:M){
  omega.df$shuffled.cat<-sample(omega.df$fishcat)
  shuff.lm<-lm(omegaper~shuffled.cat,data=omega.df)
  stats.df[j,]$stat<-summary(shuff.lm)$r.squared
}

#So if labels don't matter we get R2 values such as:

stats.df %>% ggplot(aes(x=stat))+geom_histogram()+
  geom_vline(xintercept=SSM/SST,color="red",lwd=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
stats.df %>% filter(stat>SSM/SST)%>%summarise(perc=n()/M)
```

```
##     perc
## 1 0.0032
```

So our strength of evidence against H_0 is pretty strong

While R^2 is ok to use as a statistic, it turns out that there is a better choice. If we had two groups what statistic would we use?

With multiple groups we can generalize this to an F statistic

```
dfmod=4 #We have five means, so our model is estimating 4 alpha values
dfred=nrow(omega.df)-1-4 #Start with n observations, estimate mu, estimate 4 alpha values
Fstat=(SSM/dfmod)/(SSE/dfred)
Fstat
```

```
## [1] 5.626635
```

```
#When H0 is true, how rare would it be to get an Fstat of 5.6?
M<-5000
stats.df<-data.frame(trial=seq(1,M),stat=NA)
for(j in 1:M){
  omega.df$shuffled.cat<-sample(omega.df$fishcat)
  shuff.lm<-lm(omegaper~shuffled.cat,data=omega.df)
```

```
  SSE.shuf<-sum((omega.df$omegaper-shuff.lm$fitted.values)^2)
  SSM.shuf<-SST-SSE.shuf
  stats.df[j,]$stat<-(SSM.shuf/dfmod)/(SSE.shuf/dfred)
}
stats.df %>% filter(stat>Fstat)%>%summarise(perc=n()/M)
```

```
##     perc
## 1 0.0026
```

Assuming validity conditions are met, the f stat also has a convenient distribution, an F distribution. The F distribution has two parameters, one that is degrees of freedom for model, the other is degrees of freedom for error

We can then calculate:

```
1-pf(Fstat,dfmod,dfred) #pretty darn close to our simulation
```

```
## [1] 0.00282529
```

Let's look at page 91 of our text, are the validity conditions met?

We could also do:

```
#Note we can also do:
anova(omega.lm)
```

```
## Analysis of Variance Table
##
## Response: omegaper
##           Df Sum Sq Mean Sq F value   Pr(>F)
## fishcat    4 14.545  3.6363  5.6266 0.002825 **
## Residuals 22 14.218  0.6463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's break down this table