Lesson 4: Odor and Consumer Behavior

CDT I.M. Smrt

1) What is a Parameter? a Statistic?

Review

The last few classes we fit the single-means and multiple-means models to the Afghan schools data. The book also fit a similar model to determine if using different scents would affect consumer behavior. The model for the consumer behavior is as follows:

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

Using the `OdorRatings.txt` file, the book calculated the following estimates:

```
library(tidyverse)
scent.dat<-read.table("http://www.isi-stats.com/isi2/data/OdorRatings.txt",header=TRUE)
# scent.dat<-read.table("https://raw.githubusercontent.com/jkstarling/MA376/main/OdorRatings.txt",heade
# setwd("/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons_AY23-1/Block I/LSN4/")
# scent.dat <- read.table("OdorRatings.txt", header = TRUE)
scent.dat <- scent.dat %>% mutate(condition = as.factor(condition))
contrasts(scent.dat$condition)=contr.sum
scent.mod<-lm(rating~condition,data=scent.dat)
summary(scent.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
```

$$\hat{\mu} = 4.48$$
$$\hat{\alpha}_{scent} = 0.645$$
$$\hat{\alpha}_{noscent} = -0.645$$

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

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

```
## [1] 75.97917
```

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)
```

```
## [1] 55.95833
```

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

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

```
## [1] 20.02083
```

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.

2) Recall that a null hypothesis and an alternative hypothesis are:

The null hypothesis H_0 is the "by chance alone" explanation for the observed results The alternative hypothesis H_a typically corresponds to the research conjecture (e.g. the imposed treatments explain variation in the response variable).

3) So here H_0 and H_a can be written as:

H_0:

H_a:

4) Our *parameter* and *statistic* that addresses our hypothesis are:

5) In Section 1.1, p. 38, the authors found that the difference between the two group means is 1.3. Verify this with the data.

```
# df <- ...
# mydiff <- ...
# mydiff
```

6) 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 then how we labeled our students (`scent` or `noscent` wouldn't matter)

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

```
##   condition rating
## 1     scent      5
## 2     scent      7
## 3     scent      4
## 4     scent      4
## 5     scent      4
## 6     scent      5
```

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 `noscent`, the third `scent`, etc. If H_0 is true, condition label doesn't matter. So let's rearrange them:

```
# set.seed(376)
# scent.dat2 <- ...
```

7) Now under this labeling scheme we would have observed a difference of:

```
# diffs=scent.dat2 %>% ...
# diff(diffs$mn)
```

8) Now if we shuffle these labels again (with a different seed #) we would get something different. We can simulate this 1000 times to build an empirical estimate of the difference of the sample means assuming that the there is no difference in the two groups. After simulating, assuming there is no difference in the means, plot the results in a histogram.

```
# set.seed(376)
# M <- 1000
# stat <- rep(NA, M)
# for(i in 1:M){
#   scent.dat3 = ...
#   diffs=scent.dat3 %>%
#     ...
#   stat[i] <- ...
# }
# p <- ggplot(...
# show(p)
```

9) 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? Plot it. How can we quantify how rare this event is? So what does this mean?

```
# p + geom_vline(...) + ...
```

```
# ...
```

Statistic, simulate, strength of evidence (3S Strategy) 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 **two-sample pooled t- statistic**, which is the *t-statistic*:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{(SE\ of\ residuals) \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

10) Redo the simulation using the t-statistic and plot your results.

```
# M <- 1000
# stat <- rep(NA, M)
# for(i in 1:M){
#   scent.dat3 <- scent.dat %>%
#     mutate(new_cond=sample(scent.dat$condition))%>%
#     select(-condition)
#   sim.lm <- lm(...)
#   diff.means=...
#   est.se =...
#   stat[i] <- ...
# }
```

```
# ggplot(data=as.data.frame(stat), aes(x=stat)) + geom_histogram()
```

11) Calculate our observed t-statistic. How rare is observing the t-statistic?

```
# my.tstat <- ...
# my.tstat
# ...
```

Assumptions for the t-statistic.

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 (the larger SD is not more than twice the size of the smaller).
- Distributions of the data are roughly symmetric or both sizes are at least 20 without strong skewness or outliers in the distributions.

12) 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.

```
# ...
```

13) Draw a picture with the results from our last random sample and the t-distribution.

```
# x=data.frame(val=rt(10000,46))

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

14) Once we have made our additional assumptions we can now easily find things like confidence intervals. The two-sample t-confidence interval is:

$$(y_1 - y_2) \pm t^* \times (std\, error\, of\, residuals)\sqrt{1/n_1 + 1/n_2}$$

Here our multiplier is $t^* = t_{n_1+n_2-2}(1-\alpha/2)$ and the standard error of the statistic is $\hat{\sigma}\sqrt{1/n_1 + 1/n_2}$. Calculate the 95% CI for the scent data.

```
# alpha <- ...
# n1 <- ...
# n2 <- ...
# SSE <- sum((scent.dat$...
# tstar <- qt(...)
# stat.se <- SSE/(n1+n2-2)*sqrt(1/n1+1/n2)
# upperCI <- mydiff + ...
# lowerCI <- mydiff - ...
# c(lowerCI, upperCI)
```

15) Steps 5 and 6 Now that we have our results we can conduct step 5, formulate conclusions, and step 6, look back and ahead. Look very carefully at page 65 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 meaningful enough for a store manager to care as this corresponds to about a 1 point increase on a seven point scale"