Lesson 2

Nicholas Clark

Admin

The begining of our text focuses on experiments vs. observational studies. Why is this important?

At West Point, as well as at most universities, prior to conducting an experiment, your **study protocol** must be reviewed by an Institutional Review Board or IRB. The point of the IRB is to protect the rights of the subjects of a study as well as to ensure that inferences made from the study are statistically valid.

A **double blind** study is:

Why is this important?

Our book talks about a study on store ratings and wants to determine whether a rating is influenced by exposure to a scent. Are there ethical issues with this study?

The first model they consider is

$$\begin{aligned} i &= \text{ Student} \\ y_i &= \text{rating of student } i \\ y_i &= \mu + \epsilon_i \end{aligned}$$

What is the sources of variation diagram associated with this model?

What does ϵ_i represent in this model?

Are there any assumptions we are making on ϵ_i?

The book says that the fitted model is:

$$y_i = 4.48 + \epsilon_i$$
$$\epsilon_i \sim F(0, 1.27)$$

Note here I use the generic F to stand for some distribution, I'm not making any distributional assumptions on ϵ_i. How did the book find $\hat{\mu} = 4.48$ and the standard error of the residiuals as 1.27?

What assumption are we making when we use this model? What would our causal diagram look like?

What is the treatment variable? Let's sketch out the sources of variation diagram

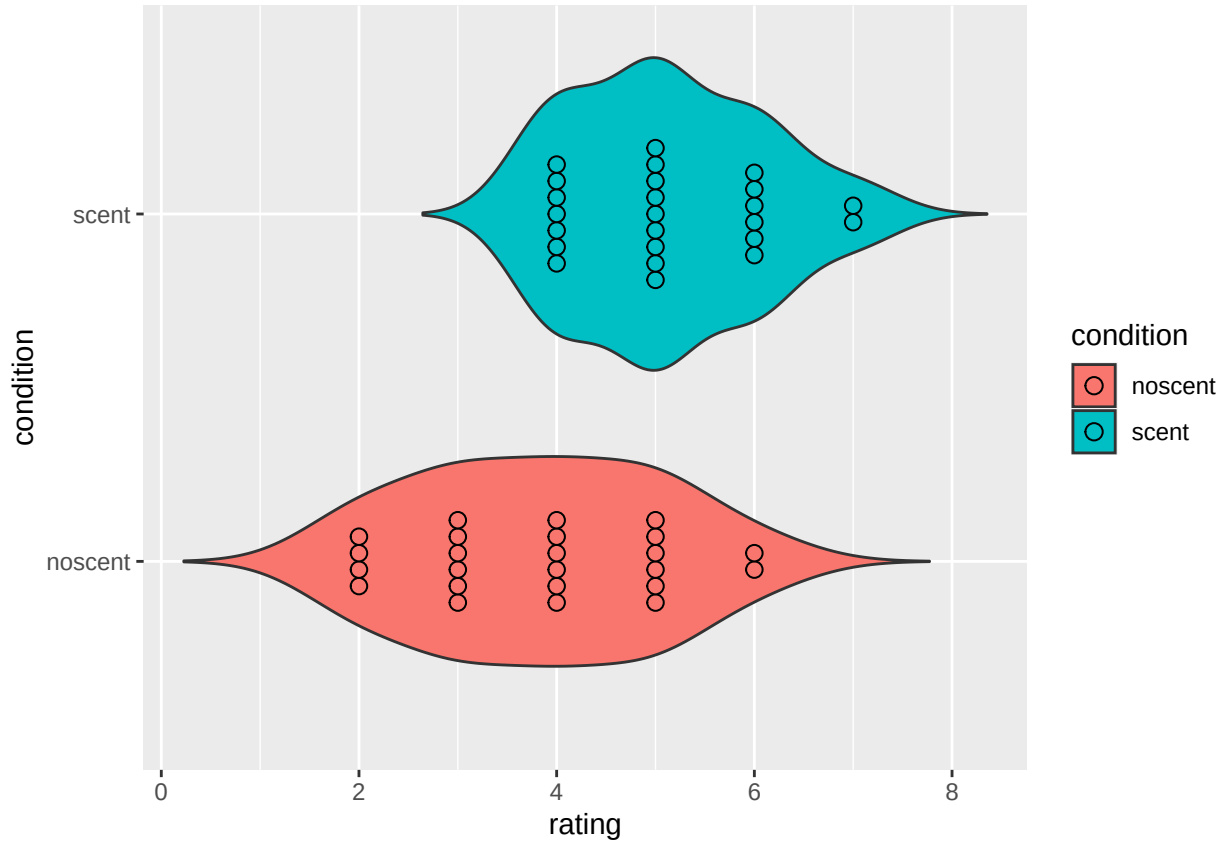
Our proposed diagram is:

We can visualize:

```
library(tidyverse)

dat=read.table("http://www.isi-stats.com/isi2/data/OdorRatings.txt",header=T)

dat %>% ggplot(aes(x=condition, y=rating,fill=condition)) +
  geom_violin(trim = FALSE)+
  geom_dotplot(binaxis='y', stackdir='center')+
  coord_flip()
```



A statistical model that could be used to address the scientific question is:

How could we fit this model? Well, getting the estimates for μ_1 and μ_2 shouldn't be hard.

```
dat %>% group_by(condition)%>%summarize(samp.mus=mean(rating),sds=sd(rating))
```

```
## # A tibble: 2 x 3
```

```
##   condition samp.mus   sds
##   <chr>        <dbl> <dbl>
## 1 noscent       3.83 1.24
## 2 scent         5.12 0.947
```

```
scent.model=lm(rating~0+condition,data=dat)
summary(scent.model)
```

```
##
## Call:
## lm(formula = rating ~ 0 + condition, data = 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|)
## conditionnoscent   3.8333     0.2251   17.03   <2e-16 ***
## conditionscent     5.1250     0.2251   22.76   <2e-16 ***
## ---
## 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:  0.9461, Adjusted R-squared:  0.9438
## F-statistic:   404 on 2 and 46 DF,  p-value: < 2.2e-16
```

Note that the standard error from this output does not match the standard error given on the top of page 39. Why do you think that is? How could we match the standard error given on page 39?

Looking at the output, (ignoring p values for now), what appears to be happening? How certain are we? How could we be sure?

What could be a confounding variable for this study?

The most important part of thinking of confounding is given in figure 1.1.5.

This is in our text, but it bears repeating: The goal of random assignment is to reduce the chances of

there being any confounding variables in the study. By creating groups that are expected to be similar with respect to all variables (other than the treatment variable of interest) that may impact the response, random assignment attempts to eliminate confounding. A key consequence of not having variables confounded with the treatment variable in a randomized experiment is the potential to draw cause-and-effect conclusions between the treatment variable and the response variable.

https://www.vox.com/science-and-health/2018/6/20/17464906/mediterranean-diet-science-health-predimed

Think - If our investigators wanted to know if there was a difference between `scent` and `noscent` what would we be testing in terms of our parameters?

In the United States, the 1963 Equal Pay Act requires that men and women be given equal pay for equal work and Title VII of the Civil Rights Act of 1964 prohibits discrimination on the basis of race, color, relgion, sex, and national origin. How successful have these acts been?

WageRace contains observations from 1987 for a sample of 25,632 males between the age of 18 and 70 who worked full-time along with their years of educaiton, years of experience, race, whether they worked in a standard metropolitan area, and the region of US where they worked.

Primary research question is whether wages for blacks differ significantly from wages for non-blacks?

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

Identify the observational units in the study. How many are there?

```
nrow(wage.dat)
#head(wage.dat) This gives the first couple of entries
```

Is the wages variable a quantitative or categorical variable?

```
ggplot(wage.dat,aes(x=wage))+geom_histogram(bins=100)
ggplot(wage.dat,aes(y=wage))+geom_boxplot()+coord_flip()
```

Why are we looking at histograms and boxplots rater than a bar graph?

Does anything stand out to you about the boxplot that is less obvious in the histogram?

Which visual, the histogram or boxplot, do you like better? Why?

Which is larger, the mean or the median? How do you know?

Do the wages appear to follow a normal distribution? How do you know?

In this study, the researchers were most interested in whether race explained differences in wages.

Which variable is the explanatory variable? Which is the response variable?

Do you think the explanatory variable explains some variation in the response variable? Do you think it explains all of the variation in the response variable? Why or why not?

```
ggplot(wage.dat,aes(y=wage,x=race))+geom_boxplot()+coord_flip()

wage.dat%>%group_by(race)%>%
  summarise(n=n(),mean=mean(wage),StDev=sd(wage),Minimum=min(wage),Median=median(wage),Maximum=max(wage
```

Consider whether there appears to be an association between wage and race: Does the wage distribution differ substantially between blacks and non-blacks? What is the difference in the mean weekly wages? Can we conclude wage discrimination?

```
ggplot(wage.dat,aes(y=wage,x=educ))+geom_boxplot()+coord_flip()
wage.dat%>%group_by(educ)%>%
  summarise(n=n(),mean=mean(wage),StDev=sd(wage),Minimum=min(wage),Median=median(wage),Maximum=max(wage
```

Suggest an easy way to improve this graphical display to better focus on a trend of increasing salaries with increasing education.

Describe the association between education and wage. Is it as you would have predicted? Explain.

What would need to be true for education level to provide an alternative explanation for why non-blacks in this sample tended to earn more than blacks?

```
ggplot(wage.dat,aes(y=wage,x=educ,fill=race))+geom_boxplot()+coord_flip()
wage.dat%>%group_by(educ,race)%>%
  summarise(mean=mean(wage),StDev=sd(wage))
```

Is there a difference in the average wage between blacks and non-blacks in the "beyond college" group? Is this difference larger or smaller than when we did not take the education level into account?

Do the lower average wages for blacks compared to non-blacks appear to be consistent across each of the education levels?

If you were to compare the average weekly wage for blacks to the average weekly wage for non-blacks in the same education group, roughly how large would you say that difference is?

How do you respond to the argument that the wage disparity between blacks and non-blacks is really an issue of education level?

Sources of variation diagram:

```
birthwt.dat<-read.table("http://www.isi-stats.com/isi2/data/births.txt",header=T,fill=T)
```

Explore:

```
ggplot(aes(x=weight),data=birthwt.dat)+geom_histogram(bins=100)
```

Filter out the unknowns

```
birthwt.clean<-birthwt.dat %>% filter(weight < 8166)
ggplot(aes(x=weight),data=birthwt.clean)+geom_histogram(bins=100)
```

summary statistics

```
birthwt.clean%>%summarise(N=n(),Mean=mean(weight),StDev=sd(weight),Min=min(weight),Max=max(weight))
```

If we used the mean to predict future newborn weight how well would we do?

The statistical model would be:

A residual is the value $y_i - \hat{y}_i$ for $i = 1, \cdots, n$. We can find the residuals two different ways:

```
birthwt.resid<-birthwt.clean%>% mutate(resid=weight-mean(weight))%>%select(resid)
ggplot(aes(x=resid),data=birthwt.resid)+geom_histogram(bins=100)

birthwt.resid%>%summarise(Mean=mean(resid),StdDev=sd(resid))
```

What is going on? Have we explained any variation?

```
ggplot(aes(x=weight,color=full),data=birthwt.clean)+geom_histogram(fill="white", alpha=0.5, position="i

birthwt.clean%>%group_by(full)%>%
  summarise(N=n(),Mean=mean(weight),StDev=sd(weight),Min=min(weight),Max=max(weight))
```

Our predicted model becomes:

Standard error:

```
model<-lm(weight~0+full,data=birthwt.clean)
summary(model)
```

Does Mom's BMI impact weight?

```
birthwt.bmi<-birthwt.clean%>%filter(mom.BMI < 90)
model<-lm(weight~0+full*mom.BMI,data=birthwt.bmi)
summary(model) #I think there's an error in book
```

Sources of variation diagram: