Lesson 3

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 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

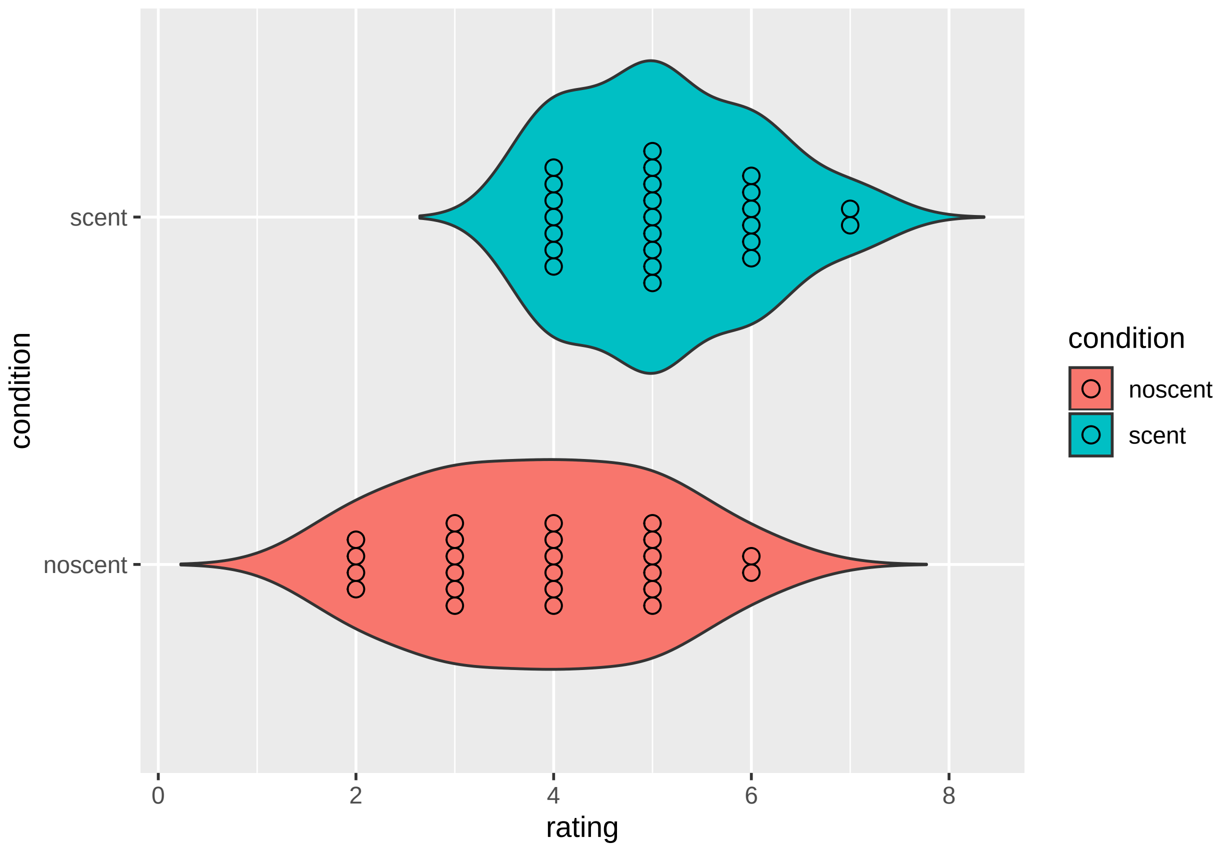
Our proposed diagram is:

We can visualize:

```r
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
##   <fct>        <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?