Lesson 1

Nicholas Clark

Contact Information

COL Nick Clark

nicholas.clark@westpoint.edu

Thayer Hall Room 226B (Back room, just keep going back...)

845-938-0267

Test

Course Overview

Applied Statistics

-Learn Statistics by doing

-Hands on/interactive

-R/Rstudio

-Course Overview

-Graded Events

Introductions

Getting started with R studio/knitr

Berkely Admissions Data

In the early 1970s, the University of California at Berkeley was concerned with possible discrimination against women in its graduate admissions process. Data about the applicants for the 1972–73 school year were recorded from several programs, including their sex and whether or not they were accepted (Bickel & O'Connell, Science, 1975).

```
#install.packages("tidyverse")
library(tidyverse)
berkley.data<-read.table("http://www.isi-stats.com/isi2/data/Berkeley.txt",header=T)
```

Observational Unit:

Response Variable:

Values Response Variable can take on:

Visualize

```
ggplot(data=berkley.data,aes(Accepted))+geom_histogram(stat="count")
```

Explore

Visualize the table:

```
#install.packages
library(ggmosaic)
ggplot(data=no.program)+
  geom_mosaic(aes(x=product(Accepted,Sex),fill=Accepted))
```

What's your thoughts?

What would the mosaic plot look like if sex was not associated with acceptance?

What other variables might help us predict whether or not someone would be accepted?

```
ggplot(data=berkley.data)+
  geom_mosaic(aes(x=product(Accepted,Sex),fill=Accepted))+
  facet_grid(Program~.)
```

What are your thoughts looking at this plot?

How can it be that females have higher acceptance rates than males in Program A AND in Program F but when we combine the two, the overall acceptance rate is noticeably smaller for females?

In examining gender and acceptance what is **confounding?**

What else could contribute to an admissions decision other than sex and program?

Sources of Variation Diagram:

Six Steps of a statistical investigation

- Ask a Research Question
- Design a study and collect data
- Explore the data
- Draw inferences beyond the data
- Formulate conclusions
- Look back and ahead

Did we do this with the Berkley Data? What would you tell the college administrators?

Let's brainstorm how to write a statistical model for the Berkley data