EDA Intro

Probability

Learning Objectives

- Understand and explain the role of randomness in designing studies and drawing conclusions
- Recognize and simulate probabilities as long-run frequencies
- Construct two-way tables and tree diagrams to evaluate marginal, joint, and conditional probabilities

```
library(readr)
library(car)
injury <- read.csv("head_injury.csv")
injury$Helmet <- factor(injury$Helmet)
injury$Injury <- factor(injury$Injury)
```

Background

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., Journal of the American Medical Association, Vol. 295, No. 8, we can see the results from a random sample 3562 skiers and snowboarders involved in accidents.

1. What are the observational units in this study?

2. What is the explanatory variable? Is it categorical or quantitative?

3. What is the response variable? Is it categorical or quantitative?

4. Is this an experiment or an observational study?

5. In this study, they saw that 576 of these skiers and snowboarders had a head injury. Of those with a head injury 96 wore a helmet and of those without a head injury 656 did not wear a helmet. Use this information to complete the following two-way table.

|Head Injury|No Head Injury|Total|
Wore Helmet| | | |
|Did Not Wear Helmet| | | | Total| 576 | 2986 | 3562|

7. What is the probability a selected skier/snowboarder has a head injury?

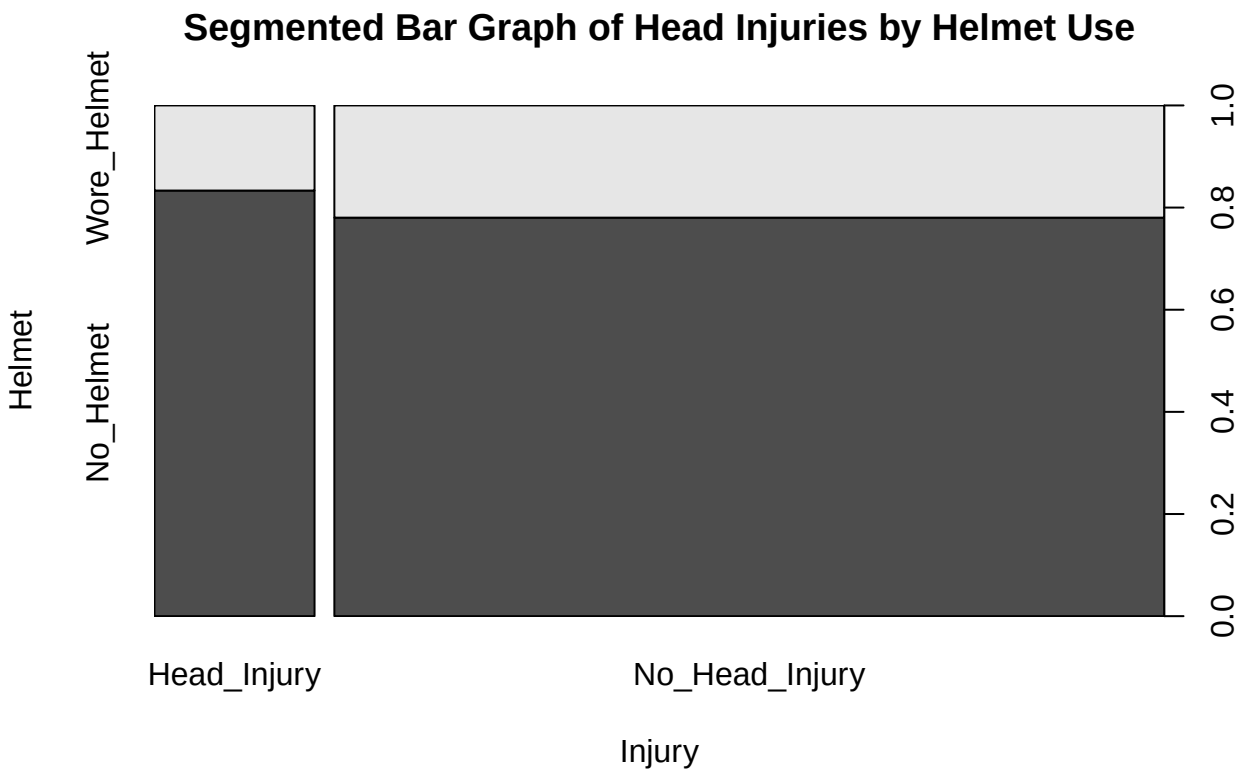
8. What is the probability a selected skier/snowboarder with a head injury was wearing a helmet?

9. What is the probability a selected skier/snowboarder that did not wear a helmet had no head injury?

10. Use your answers to questions 6 - 9 to create a tree diagram.

Let's look at a segmented bar graph to assess the data.

```
plot(Helmet~Injury, data=injury, main="Segmented Bar Graph of Head Injuries by Helmet Use")
```



11. Do the probabilities found above match what is seen in the bar graph?

The following dataset is from the Current Population Survey in 1985. The following table summarizes the data.

Variable	Description
educ	Number of years of education
south	Indicator variable for living in a southern region: S = lives in south, NS = does not live in south
sex	Gender: M = male, F = female
exper	Number of years of work experience (inferred from age and education)
union	Indicator variable for union membership: Union or Not
wage	Wage (dollars per hour)
age	Age (years)
race	Race: W = white, NW = not white
sector	Sector of the economy: clerical, const (construction), management, manufacturing, professional, sales, service, other
married	Marital status: Married or Single

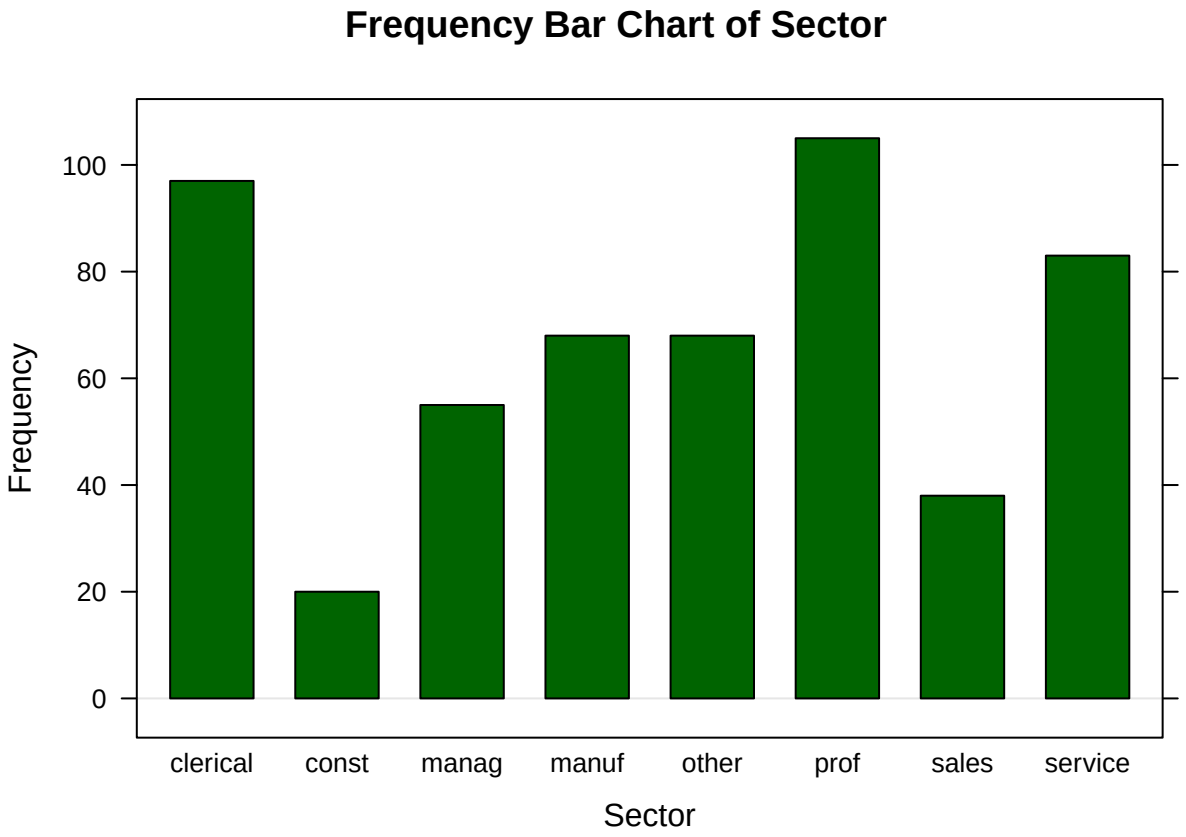
12. Which variables are categorical?

13. Which variables are quantitative?

A bar chart is used to plot a single categorical varialbe. We can plot the counts for each category in a frequency bar chart and the proportion in each categorcy in a relative frequecy bar chart. Here we will create a bar chart of sector.

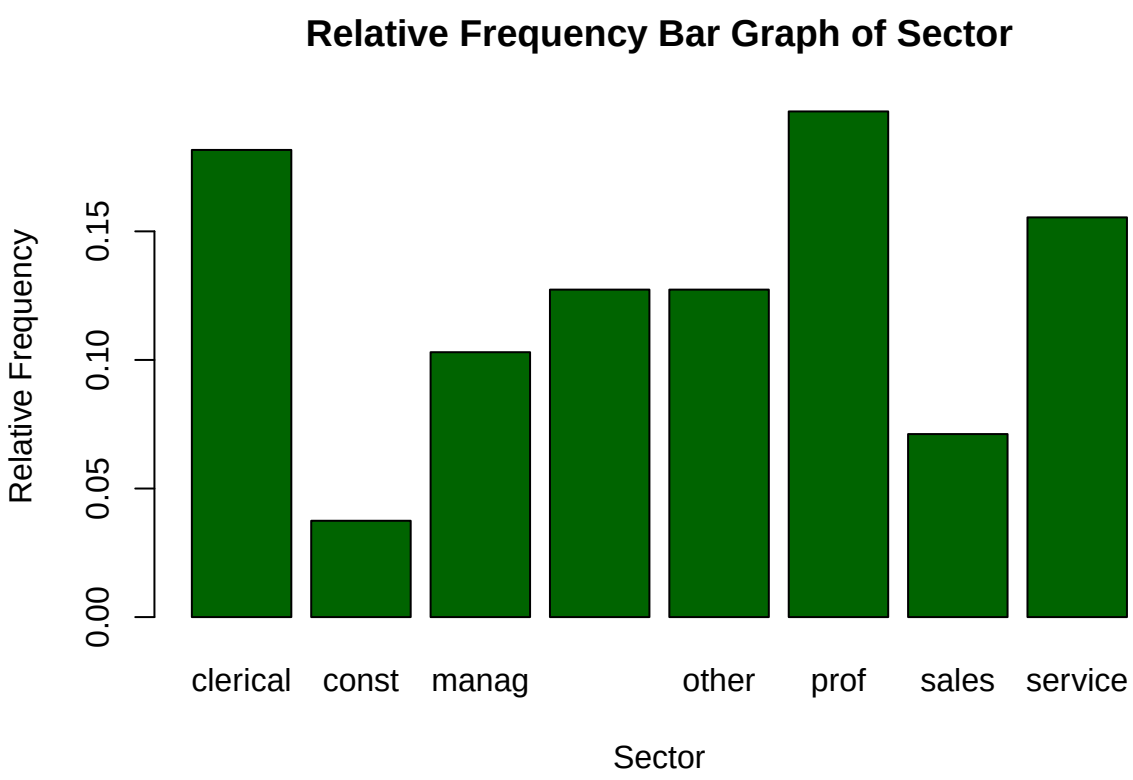
```
cps <- read.csv("cps.csv")
cps$sector <- factor(cps$sector)
cps$sex <- factor(cps$sex)
```

```
barchart(cps$sector, #This specifies the dataset and the variable
         horizontal = FALSE, #Turn the bars so they are vertical
         main = "Frequency Bar Chart of Sector", #Give your chart a title
         xlab = "Sector", #Label the x axis
         ylab = "Frequency", #Label the y axis
         col = "darkgreen") #change the color of the bars
```



15. Which Sector has the largest number of people in it?

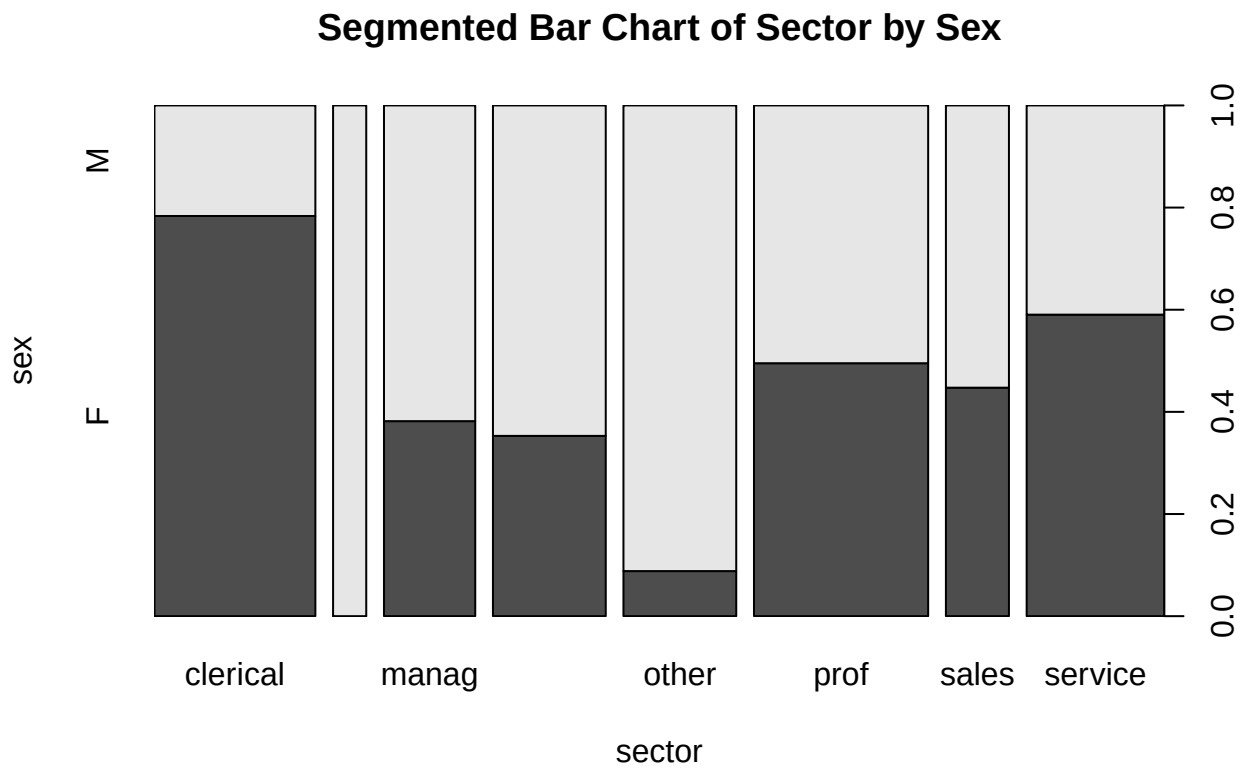
```
barplot(table(cps$sector)/nrow(cps), #divide the frequency counts by the total
        main = "Relative Frequency Bar Graph of Sector", #Give your chart a title
        xlab = "Sector", #Label the x axis
        ylab = "Relative Frequency", #Label the y axis
        col = "darkgreen")
```



16. How does this plot differ from the plot above?

To visually display two categorical variables we will use a segmented bar chart.

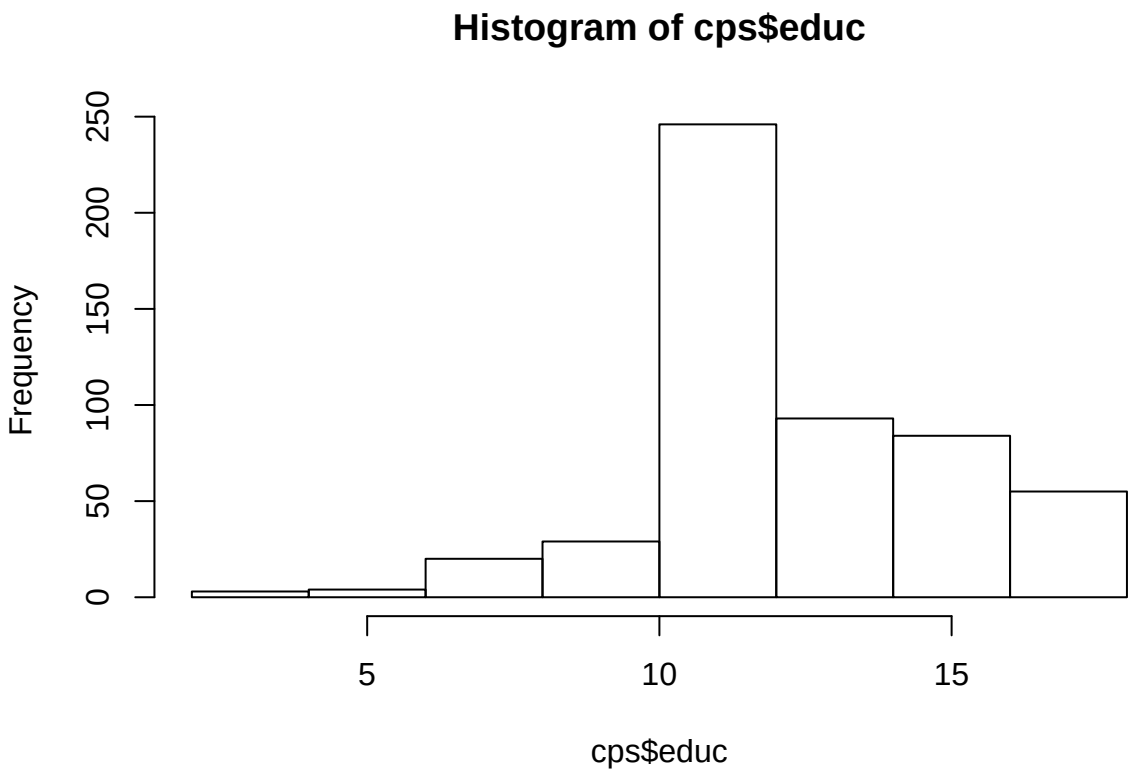
```
plot(sex~sector, data = cps, main="Segmented Bar Chart of Sector by Sex")
```



17. Using the segmented bar chart, which sector has about the same proportion of males and females?

To plot quantitative variables, we can use a histogram or boxplot. This is a histogram of the variable education.

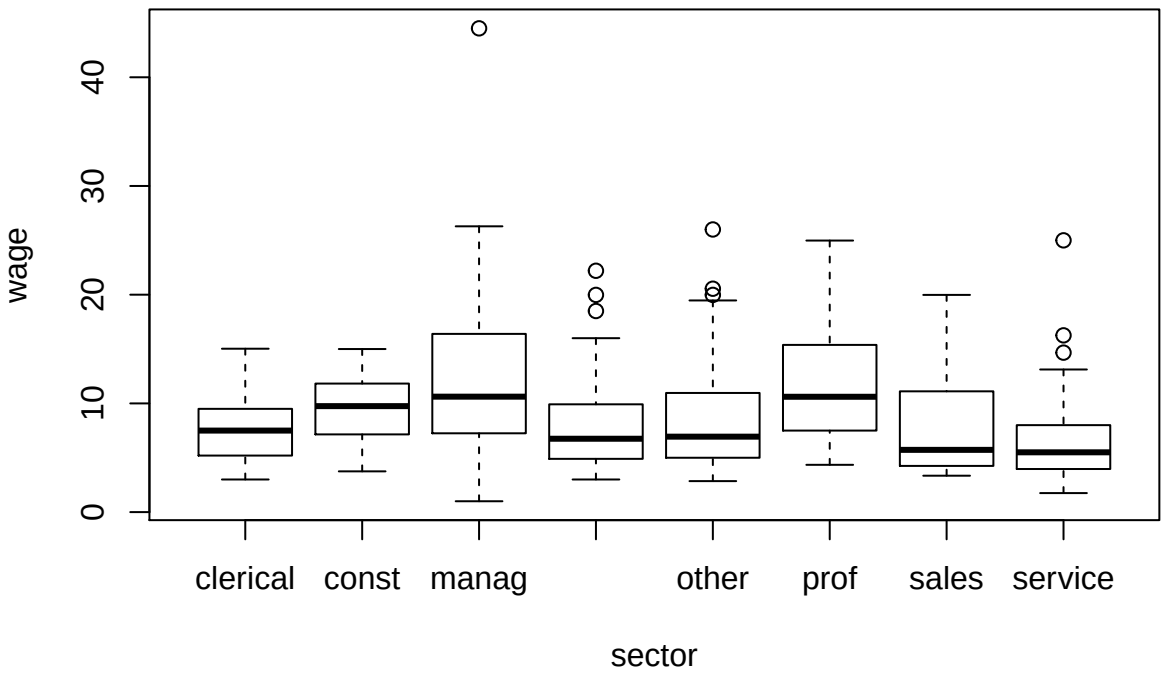
```
hist(cps$educ)
```



18. What is the most common level of Education?

The boxplot

```
boxplot(wage~sector, data=cps)
```



19. Compare Service and Sales using the four characteristic to comparing distributions.