Exploratory Data Analysis - Categorical Variables

Learning Outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question
- Plots for a single categorical variable: bar plot
- Plots for association between two categorical variables: segmented bar plot, mosaic plot
- Recognize and simulate probabilities as long-run frequencies
- Construct two-way tables to evaluate conditional probabilities

Terminology Review

In today's activity we will review summary measures and plots for categorical variables. Some terms covered in this activity are...

- Proportions
- Bar plots
- Segmented bar plots
- Probability
- Conditional Probability
- Two-way tables

To review these concepts see Section 2.1 in the textbook.

The data set we will use for this activity is from the Current Population Survey in 1985. The CPS is a survey sponsored by the Census Bureau and the Bureau of Labor Statistics to track labor force statistics for the United States population. 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

Vocabulary Review

1. What are the observational units?

2. Which variables are categorical?

3. What types of plots can be used to display categorical data?

An important part of understanding data is to create visual pictures of what the data represents. In this activity we will create graphical representations of categorical data.

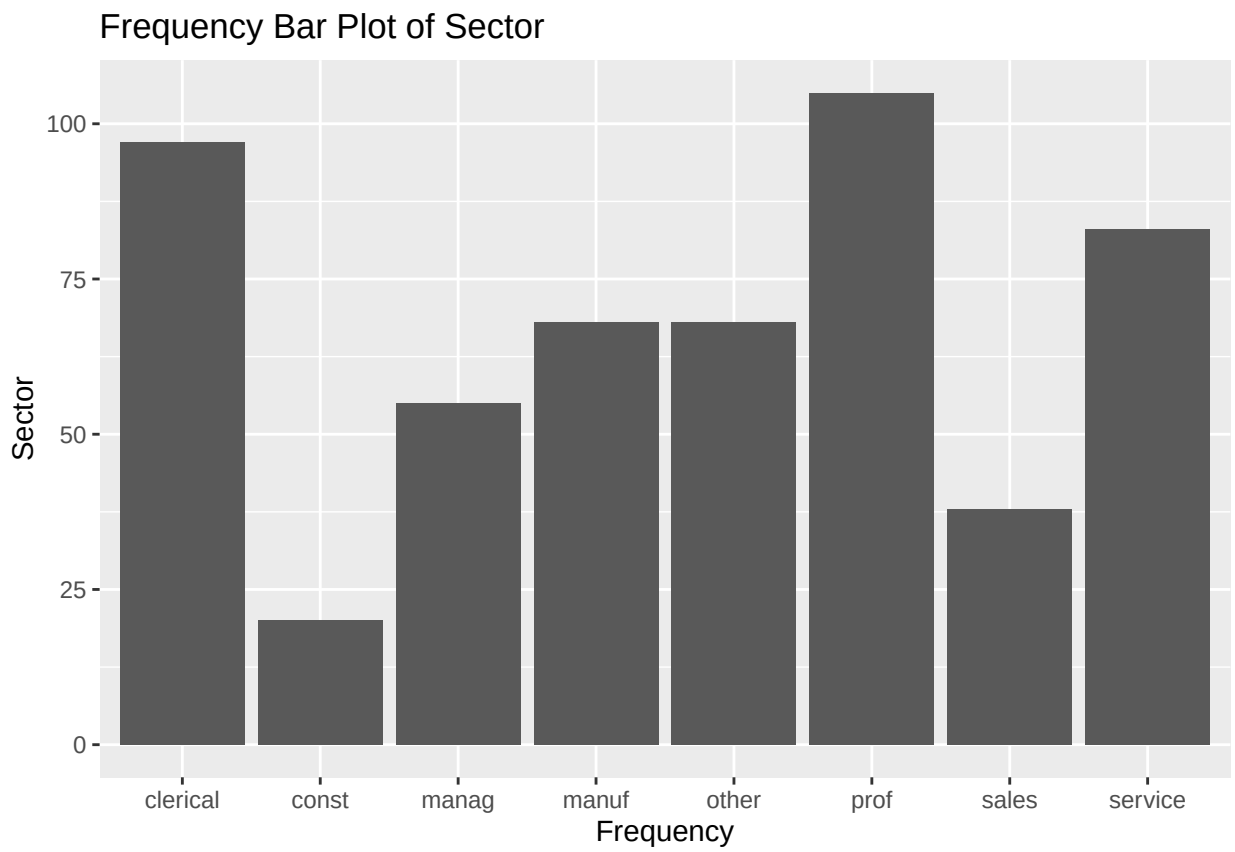
Displaying a single categorical variable.

If we wanted to know how many people in our data set were in each sector, we would create a bar plot of the variable sector.

```
cps <- read.csv("../data/cps.csv") #This will read in the dataset
cps$sector <- factor(cps$sector) #When a variable is categorical, need to make it a factor
cps$sex <- factor(cps$sex)
```

```
ggplot(data = cps,   #This specifies the dataset
       aes(y = sector)) +   #This specifies the variable
  geom_bar(stat = "count") +  #Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Sector",  #Give your plot a title
       x = "Sector",   #Label the x axis
```

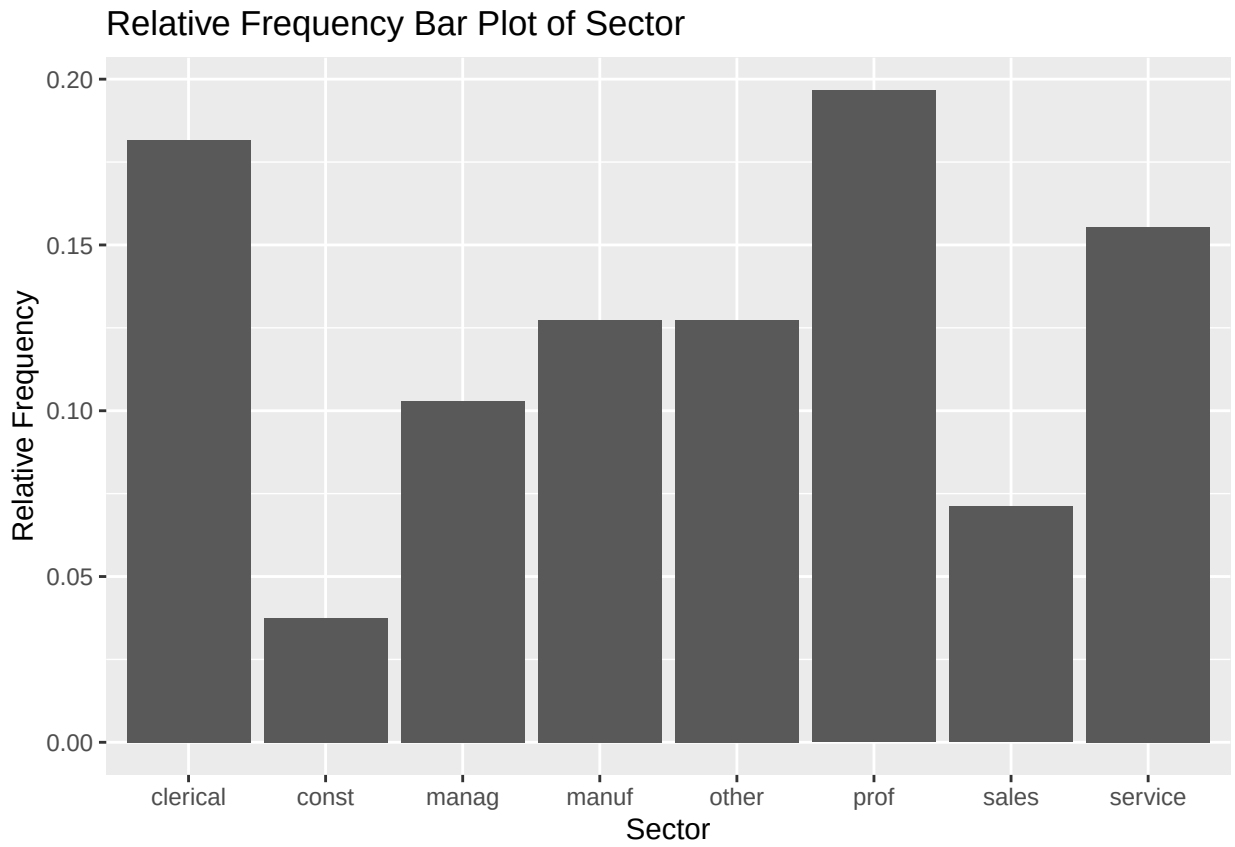
```
      y = "Frequency")  + #Label the y axis
  coord_flip()  #Turn the bars so they are vertical
```



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

We could also choose to display the data as a proportion in a relative frequency bar plot. To find the relative frequency divide the count in each sector by the sample size. This is the sample proportion.

```
ggplot(data = cps,   #This specifies the dataset
       aes(x = sector)) +   #This specifies the variable
  geom_bar(aes(y = ..prop.., group = 1)) +  #Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Sector",  #Give your plot a title
       x = "Sector",   #Label the x axis
       y = "Relative Frequency")  #Label the y axis
```

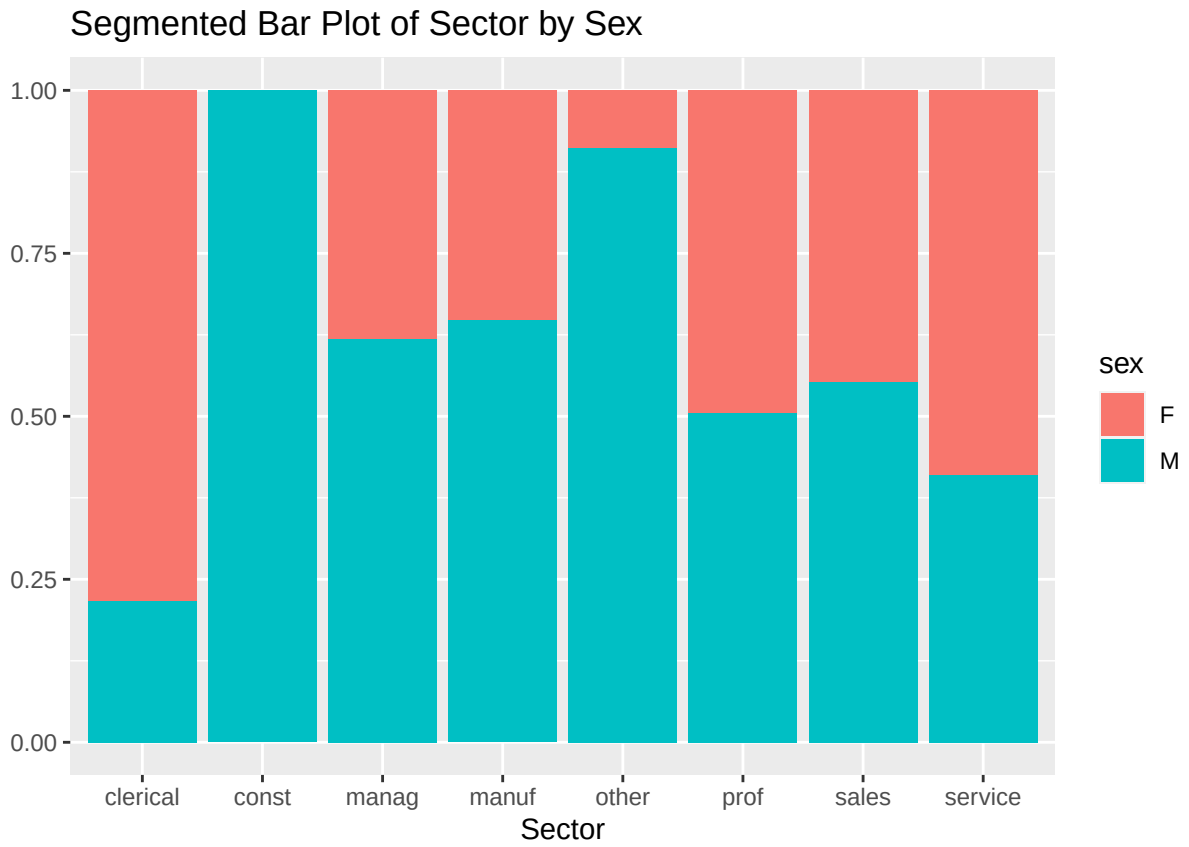


5. How does is this plot the same as the frequency bar plot? How does it differ?

Displaying two categorical variables

To see the differences in proportion of each sector between males and females we would create a segmented bar plot of sector segmented by sex.

```
ggplot(data = cps,   #This specifies the dataset
       aes(x = sector, fill = sex)) +   #This specifies the variables
  geom_bar(stat = "count", position = "fill") +  #Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Sector by Sex",  #Make sure to title your plot
       x = "Sector",   #Label the x axis
       y = "")  #Remove y axis label
```



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

7. Which sector has the highest proportion of females?

8. Which variable is the explanatory variable? Which is the response variable?

Probability

9. A study was reported in which ninth grade Minnesota teens were asked whether they had gambled at least once a week in the past year. The sample consisted of 49.1% boys. The proportion of boys who had gambled at least once per week during the past year was 0.229, while among non-boys this proportion was only 0.045.
   Let B = the event the person is a boy, and C = the event the person is a weekly gambler.

   a. Draw a segmented bar plot of gambling segmented by sex.

b. Identify what each numerical value represents in probability notation.

c. Create a two-way hypothetical table to represent the situation.

d. Find P(B and C).

e. What does this probability represent in the context of the problem?

f. Find the probability that a selected non-gambler is a non-boy.

g. What is the notation used for the probability calculated in part f?

10. In a computer store, 30% of the computers in stock are laptops and 70% are desktops. Five percent of the laptops are on sale, while 10% of the desktops are on sale. Let L = the event the computer is a laptop, and S = the event the computer is on sale.

    a. Identify what each numerical value represents in probability notation.

b. Create a two-way table to represent the situation.

c. Calculate the probability that a randomly selected computer will be a desktop, given that the computer is on sale.

d. What is the notation used for the probability calculated in part c.

e. Find $P(S|L^C)$.