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 and tree diagrams to evaluate marginal, joint, and conditional probabilities

The following data set 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

Vocabulary Review

1. What are the observational units?

2. Which variables are categorical?

3. Which variables are quantitative?

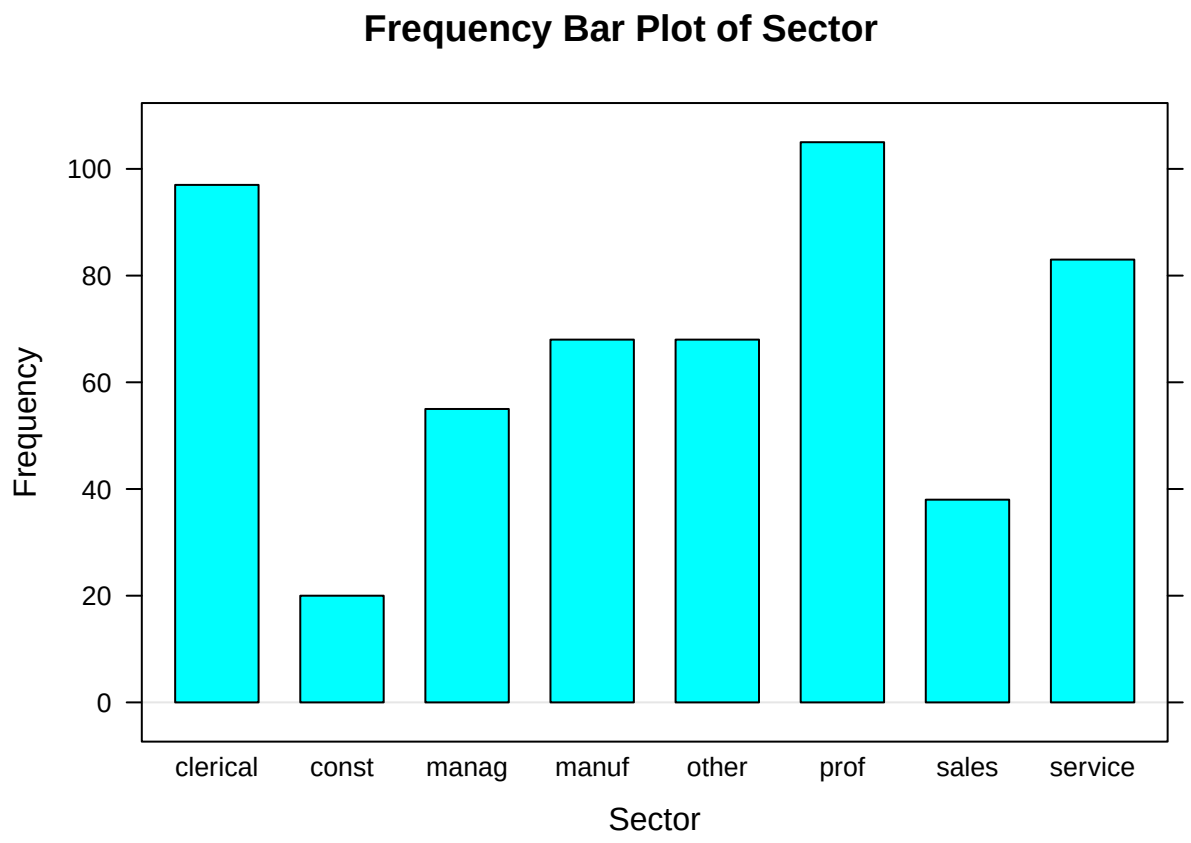
An important part of understanding data is to create visual pictures of what the data represents. In this activity we will create graphical representations of categorical data.

Displaying a single categorical variable.

A bar plot is used to plot a single categorical variable. We can plot the counts for each category in a frequency bar plot and the proportion in each category in a relative frequency bar plot. 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 you need to set up as a factor
cps$sex <- factor(cps$sex)
```

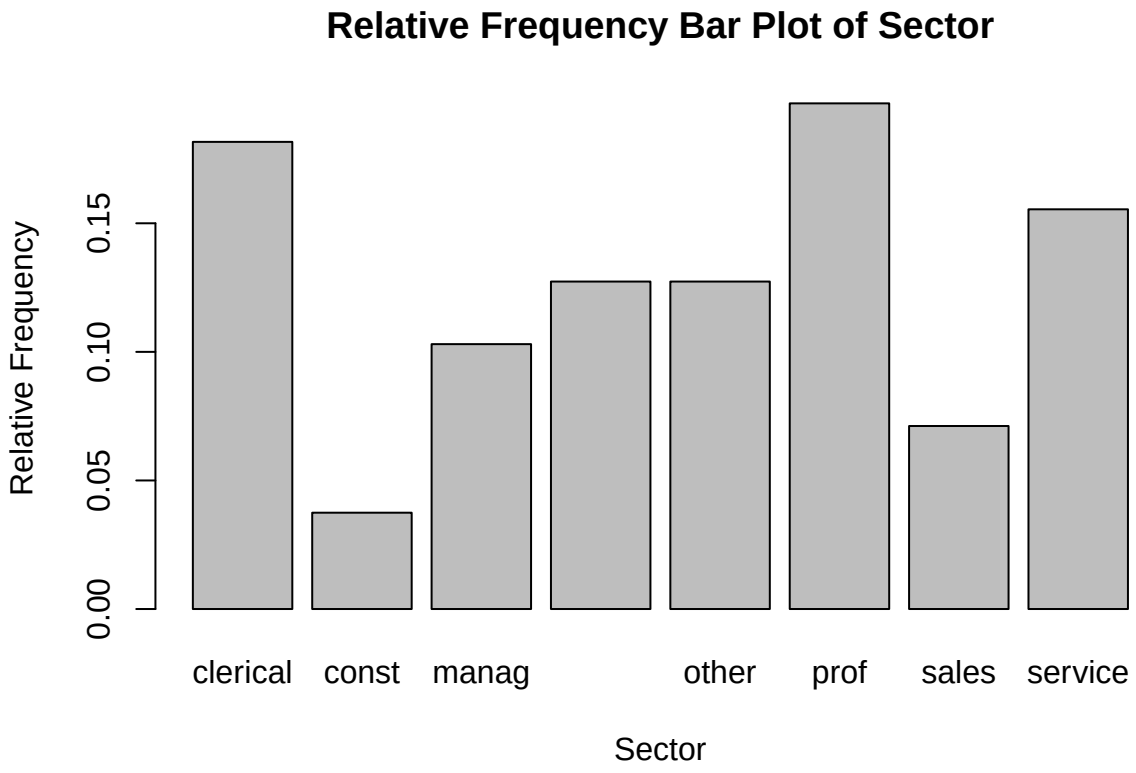
```
barchart(cps$sector, #This specifies the dataset and the variable
         horizontal = FALSE, #Turn the bars so they are vertical
         main = "Frequency Bar Plot of Sector", #Give your plot a title
         xlab = "Sector", #Label the x axis
         ylab = "Frequency", #Label the y axis
  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



3. 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.

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



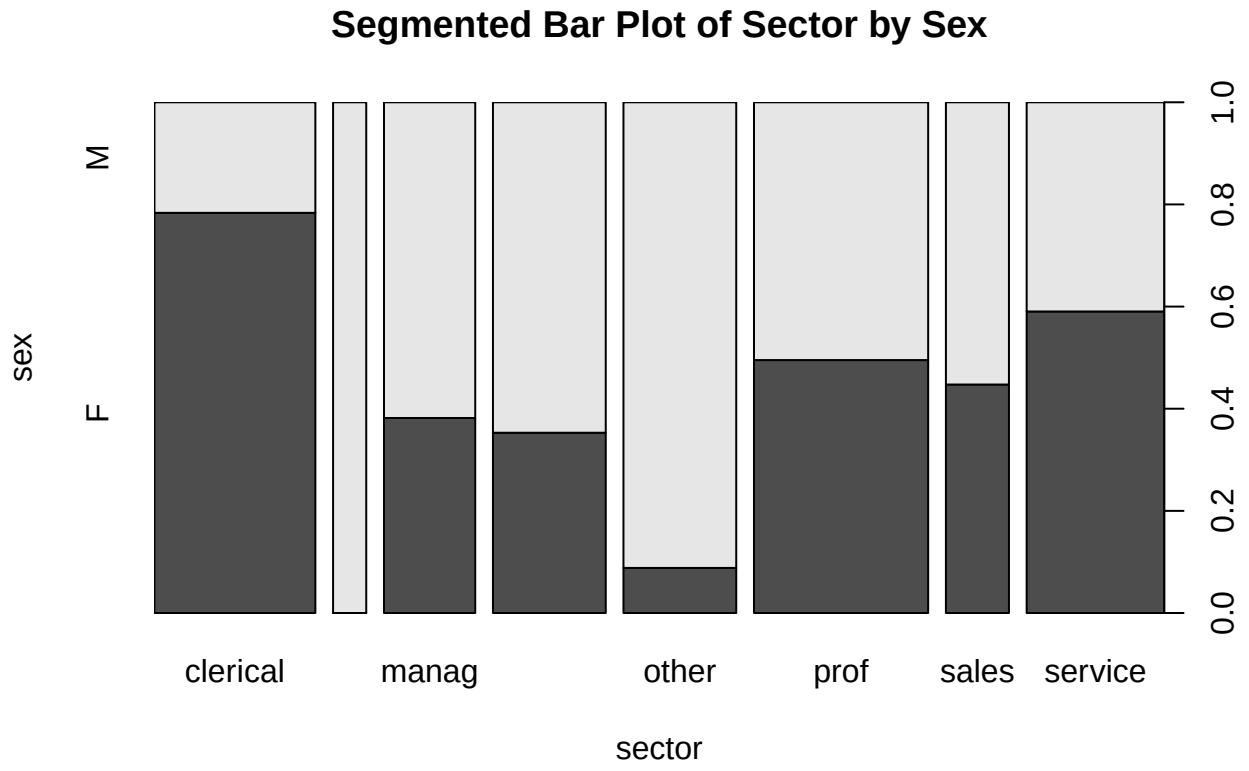
4. How does this plot differ from the plot above?

Displaying two categorical variables

To visually display two categorical variables we will use a segmented bar plot or a mosaic plot. In a segmented bar plot each bar sums to 100% or a proportion of 1. Typically the explanatory variable will be plotted on the y-axis and the response variable on the x-axis.

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. In this plot we are comparing the variable sex (explanatory variable) for the different sectors (response variable).

```
plot(sex~sector #response~explanatory allows us to plot two variables
     , data = cps, main="Segmented Bar Plot of Sector by Sex" #Make sure to title your graph
     )
```



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

6. Which sector has the highest proportion of females?

Probability

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

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

	Boys	Non-Boys	Total
Gambled			
Did Not Gamble			
Total			100,000

c. Find $P(BandC)$.

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

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

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

8. 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 tree diagram 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 $P(S|L^C)$.