ACTIVITY 3

Current Population Survey

3.1 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

3.2 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 Sections 2.1 and 2.2 in the textbook.

3.3 "Current" Population Survey: 1985

The data set we will use for this activity is from the Current Population Survey (CPS) 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 describes the variables in the data set:

Variable	Description
`educ`	Number of years of education
`south`	Whether lives in southern region of the US: `S` = lives in south, `NS` = does not live in south
`sex`	Sex: `M` = male, `F` = female
`exper`	Number of years of work experience (inferred from age and education)
`union`	Whether union member: `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 represent. In this activity, we will create graphical representations of categorical data.

`R` code

`R` is a free statistical analysis software program we will use in Stat 216. Please see D2L for instructions on how to download or access a version of `R` on your laptop, or plan to use the school computers for some parts of assigned out of class work for this course.

Throughout these activities, we will often include the `R` code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into `R` using the `read.csv()` function.

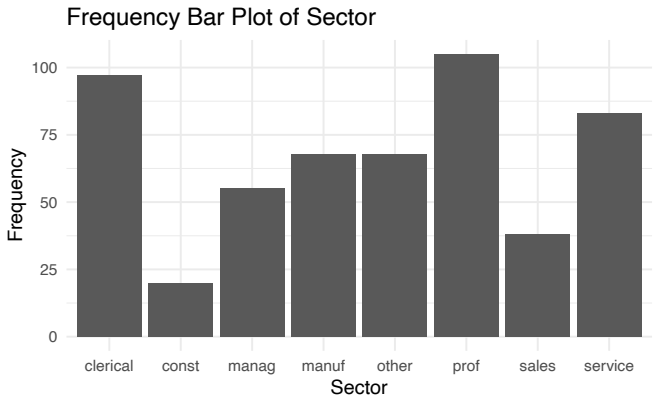
The # sign is not part of the `R` code. It is used by these authors to add comments to the `R` code and explain what each call is telling the program to do. `R` will ignore everything after a # sign when executing the code.

```
cps <- read.csv("data/cps.csv") #This will read in the data set
```

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 frequency bar plot of the variable `sector`.

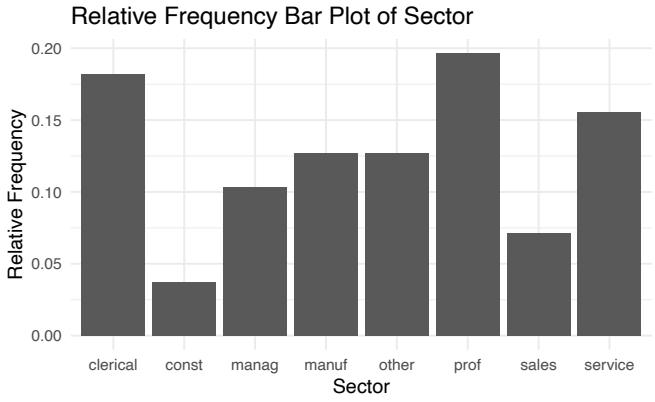
```
cps %>% #Data set piped into...
ggplot(aes(y = sector)) +   #This specifies the variable
  geom_bar(stat = "count") +  #Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Sector",  #Give your plot a title
       x = "Frequency",   #Label the x axis
       y = "Sector")  + #Label the y axis
  coord_flip()  #Turn the bars so they are vertical
```



4. Which sector of the economy has the largest number of people in it? Approximately how many people are in this sector?

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. These are sample proportions.

```
cps %>% #Data set piped into...
ggplot(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
```

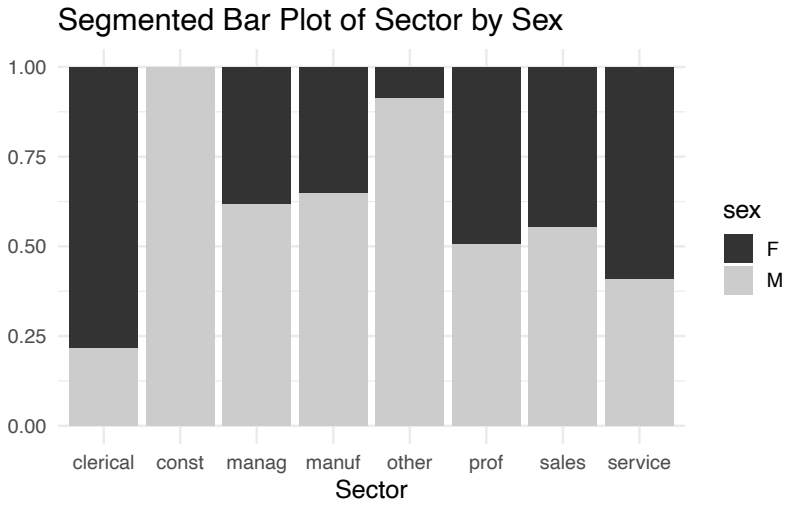


5. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

Displaying two categorical variables

To examine the differences proportion of males and females across sectors, we would create a segmented bar plot of `sector` segmented by `sex`.

```
cps %>% #Data set piped into...
ggplot(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
    scale_fill_grey()  #Make figure black and white
```



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 bar plot treating as the explanatory variable? Which is the response variable?

3.4 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 sex segmented by gambling. Make sure to clearly label your axes and legend.

b. Identify what each numerical value given in the problem represents in probability notation.

   0.491 =

   0.229 =

   0.045 =

c. Create a hypothetical two-way table to represent the situation. Recall that in a two-way table, the explanatory variable should be your column headers (similar to the x-axis in a segmented bar graph!) while the response variable becomes the row headers.

	Total
Total	100,000

d. Find P(B and C). What does this probability represent in the context of the problem?

e. Find the probability that a selected non-gambler is a non-boy. What is the notation used for this probability?

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 given in the problem represents in probability notation.

    0.30 =

    0.70 =

    0.05 =

    0.10 =

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

	Total
Total	100,000

c. Calculate the probability that a randomly selected computer will be a desktop, given that the computer is on sale. What is the notation used for this probability?

d. Find $P(S|L^C)$. What does this probability represent in context of the problem?

3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity.