STAT 216 Activity Coursepack

Fall 2020

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University. Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written complete workbook will provide an optimal study guide for exams!

Fall 2020 Calendar of In-Class Activities

Placeholder

ACTIVITY 1

Martian Alphabet

Placeholder

1.1 Learning outcomes

1.2 Terminology review

1.3 Can you read "Martian"?

1.3.1 Steps of the statistical investigation process

1.4 Take home messages

1.5 Additional notes

ACTIVITY 2

Study Design

Placeholder

2.1 Learning outcomes

2.2 Terminology review

2.3 Types of sampling bias

2.4 Study design

2.5 Additional notes

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 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

3.3.1 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.

3.3.2 `R` code

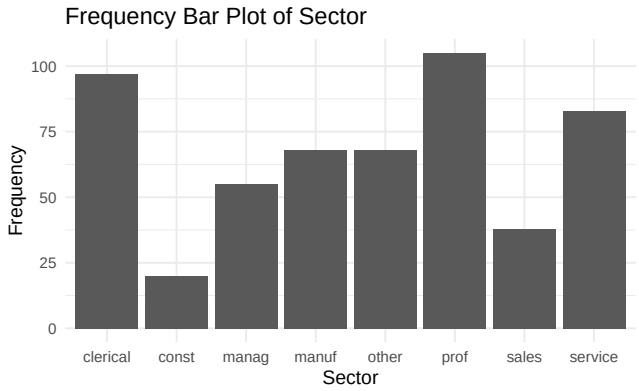
Throughout these activities, we will often include the `R` code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into `R` using the `read.csv()` function, and tell `R` to treat the `sector` and `sex`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

```r
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)
```

3.3.3 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.

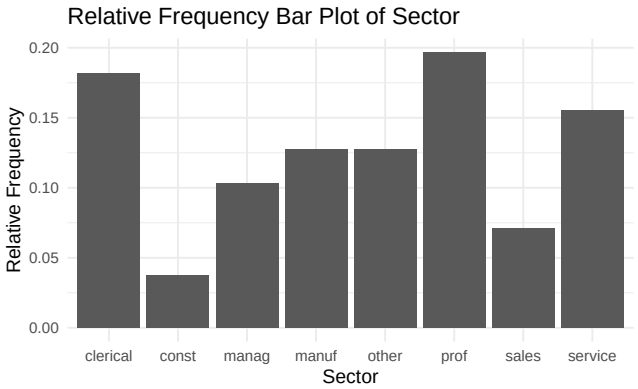
```
ggplot(data = cps,   #This specifies the dataset
       aes(y = sector)) +   #This specifies the variable
  geom_bar(stat = "count") +  #Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Sector",  #Give your plot a title
       x = "Frequency",   #Label the x axis
       y = "Sector")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

```
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
```

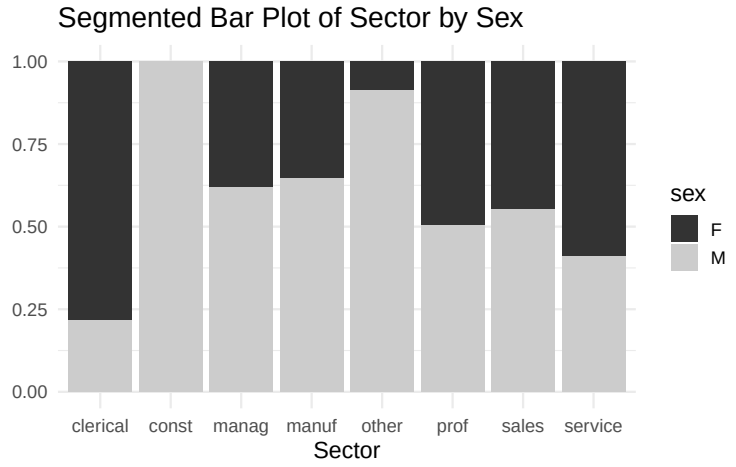


5. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

3.3.4 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
    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 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.

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

   0.491 =

   0.229 =

   0.045 =

c. Create a two-way hypothetical 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 represents in probability notation.

$0.30 =$

$0.70 =$

$0.05 =$

$0.10 =$

b. Create a 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.

ACTIVITY 4

IMDb Movie Reviews

Placeholder

4.1 Learning objectives

4.2 Terminology review

4.3 Movies released in 2016

4.4 Vocabulary review

4.5 Summarizing a single quantitative variable

4.6 Displaying a single quantitative variable

4.7 Displaying a single categorical and single quantitative variable

4.8 Additional notes

ACTIVITY 5

Movie Profits

Placeholder

5.1 Learning objectives

5.2 Terminology review

5.3 Movies released in 2016

5.3.1 Vocabulary review

5.3.2 Correlation

5.3.3 Slope

5.3.4 Residuals

5.3.5 Coefficient of determination (R-squared)

5.3.6 Multivariate plots

5.4 Additional notes

ACTIVITY 6

Handedness of Male Boxers

Placeholder

6.1 Learning objectives

6.2 Terminology review

6.3 Steps of the statistical investigation process

6.4 Handedness of male boxers

6.4.1 Summary statistics review

6.4.2 Ask a research question

6.4.3 Design a study and collect data

6.4.4 Summarize and visualize the data

6.4.5 Use statistical analysis methods to draw inferences from the data

6.4.6 Communicate the results and answer the research question

6.4.7 Revisit and look forward

6.5 Additional notes

ACTIVITY 7

Winter Sports Helmet Use and Head Injuries

Placeholder

7.1 Learning objectives

7.2 Terminology review

7.3 Helmet Use and Head Injuries

7.3.1 Vocabulary review

7.3.2 Ask a research question

7.3.3 Summarize and visualize the data

7.3.4 Use statistical analysis methods to draw inferences from the data

7.3.5 Types of errors

7.4 Additional notes

ACTIVITY 8

COVID-19 and Air Pollution

Placeholder

8.1 Learning outcomes

8.2 Terminology review

8.3 COVID-19 and air pollution

8.3.1 Vocabulary review

8.3.2 Ask a research question

8.3.3 Summarize and visualize the data

8.3.4 Use statistical inferential methods to draw inferences from the data

8.3.5 Communicate the results and answer the research question.

8.3.6 Revisit and look forward

8.4 Additional notes

ACTIVITY 9

Weather Patterns and Record Snowfall

Placeholder

9.1 Learning objectives

9.2 Terminology review

9.3 Weather Patterns and Record snowfall

9.3.1 Quantitative variables review

9.3.2 Ask a research question.

9.3.3 Summarize and visualize the data

9.3.4 Use statistical inferential methods to draw inferences from the data

9.3.5 Communicate the results and answer the research question

9.3.6 Revisit and look rorward

9.4 Additional notes

ACTIVITY 10

Hand Dexterity

Placeholder

10.1 Learning outcomes

10.2 Terminology review

10.3 Hand dexterity

10.3.1 Vocabulary review

10.3.2 Conditions for the least squares line

10.3.3 Ask a research question

10.3.4 Summarize and visualize the data

10.3.5 Use statistical inferential methods to draw inferences from the data

10.3.6 Communicate the results and answer the research question

10.3.7 Revisit and look forward

10.4 Additional notes