STA 218: Statistics for Management

Al Nosedal.
University of Toronto.

Fall 2018

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Definitions

A **variable** is some characteristic of a population or sample. We usually represent the name of a variable using uppercase letters such as X, Y, and Z.

The **values** of the variable are the possible observations of the variable.

Data are the observed values of a variable.

Types of Data

There are three types of data: interval, nominal, and ordinal.

- **Interval** data are real numbers, such as heights, weights, incomes, and distances. We also refer to this type of data as **quantitative** or **numerical**.
- The values of **nominal** data are categories. For example, responses to questions about marital status nominal data. Nominal data are also called **qualitative** or **categorical**.
- **Ordinal** data appear to be nominal, but the difference is that the order of their values has meaning. For example, at the completion of most university courses, students are asked to evaluate the course.

Calculations for Types of Data

- **Interval Data**. All calculations are permitted on interval data. We often describe a set of interval data by calculating the average.
- **Nominal Data**. Because the codes of nominal data are completely arbitrary, we **cannot** perform any calculations on these codes.
- **Ordinal Data**. The most important aspect of ordinal data is the order of the values. The only permissible calculations are those involving a ranking process.

Exercise

For each of the following examples of data, determine the type.

a. The number of miles joggers run per week.

b. The starting salaries of graduates of MBA programs.

c. The months in which a firm's employees choose to take their vacations.

d. The final letter grades received by students in a Statistics course.

Solution

a. Interval.
b. Interval.
c. Nominal.
d. Ordinal.

Exercise

A sample of shoppers at a mall was asked the following questions. Identify the type of data each question would produce.

a. What is your age?

b. How much did you spend?

c. What is your marital status?

d. Rate the availability of parking: excellent, good, fair, or poor.

e. How many stores did you enter?

Solution

a. Interval.
b. Interval.
c. Nominal.
d. Ordinal.
e. Interval.

Distribution of a variable

The distribution of a variable tells us what values it takes and how often it takes these values.
The values of a categorical variable are labels for the categories.
The distribution of a categorical variable lists the categories and gives either the count or the percent of individuals that fall in each category.

Summarizing Qualitative Data

Frequency distribution. A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes.

Relative frequency of a class $= \frac{\text{Frequency of the class}}{n}$

where n represents the total number of observations.

More about Qualitative Data

A **relative frequency distribution** gives a tabular summary of data showing the relative frequency for each class.

A **percent frequency distribution** summarizes the percent frequency of the data for each class.

Bar charts and pie charts

A **bar chart**, is a graphical device for depicting qualitative data summarized in a frequency, relative frequency, or percent frequency distribution. On one axis of the graph, we specify the labels that are used for the classes (categories). A frequency, relative frequency, or percent frequency scale can be used for the other axis of the graph.

The pie chart provides another graphical device for presenting relative frequency and percent frequency distributions for qualitative data.

Toy Example

The response to a question has three alternatives: A, B, and C. A sample of 120 responses provides 60 A, 24 B, and 36 C.
a)Show the frequency, relative frequency and percent frequency distributions.
b) Construct a pie chart.
c) Construct a bar graph.

Solution

Class	Frequency	Relative Freq.	Percent Freq.
A	60	60/120	0.50
B	24	24/120	0.20
C	36	36/120	0.30

Homework

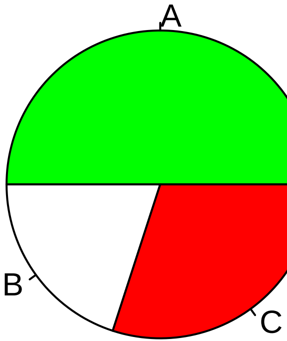
This course uses R. R is an open-source computing package which has seen a huge growth in popularity in the last few years. R can be downloaded from https://cran.r-project.org

Please, download R and bring your laptop next time.

Solution (pie chart)

```
## Step 1. Entering Data;
counts=c(60,24,36);
classes=c("A","B","C");

## Step 2. Making pie chart;
pie(counts,classes,col=c("Green","White","Red"))
```

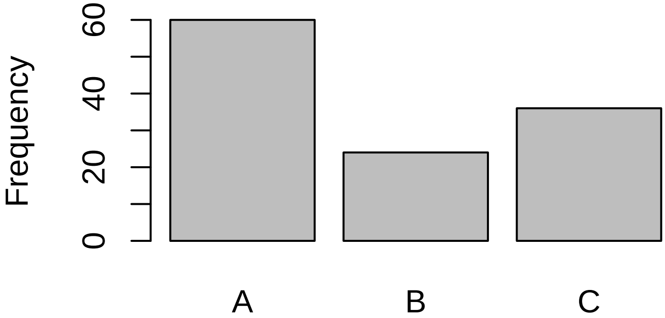
Solution (pie chart)



Solution (bar chart)

```
## Step 1. Entering Data.

counts=c(60,24,36);
classes=c("A","B","C");

## Step 2. Making bargraph.
barplot(counts,names.arg=classes,ylab="Frequency")
```

Solution (bar chart)



Example. Never on Sunday?

Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2008:

Day	Births
Sunday	7,534
Monday	12,371
Tuesday	13,415
Wednesday	13,171
Thursday	13,147
Friday	12,919
Saturday	8,617

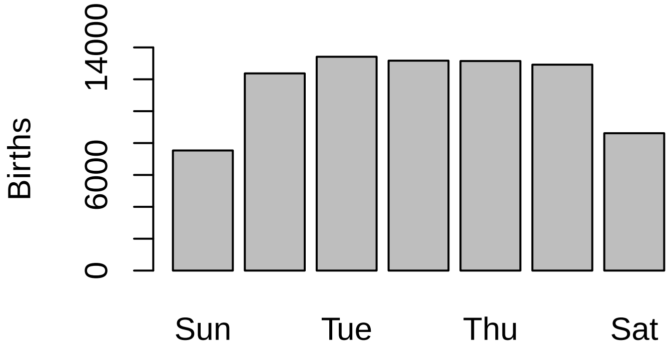
Example. Never on Sunday? (cont.)

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

Solution (bar chart)

```r
## Step 1. Entering Data;
births=c(7534,12371,13415,13171,13147,12919,8617);
names=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat");

## Step 2. Making bargraph.
barplot(births,names.arg=names,
ylim=c(0,14000),ylab="Births");
```

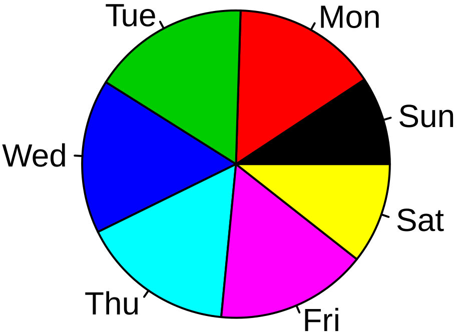
Solution (bar chart)



Solution (pie chart)

```
## Step 1. Entering Data;
births=c(7534,12371,13415,13171,13147,12919,8617)
names=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat")


## Step 2. Making pie chart.

pie(births,names,col=c(1:7))
```

Solution (pie chart)



Example. Never on Sunday?

Solution.
It would be correct to make a pie chart but a pie chart would make it more difficult to distinguish between the weekend days and the weekdays. Some births are scheduled (e.g., induced labor), and probably most are scheduled for weekdays.

Example. What color is your car?

The most popular colors for cars and light trucks vary by region and over time. In North America white remains the top color choice, with black the top choice in Europe and silver the top choice in South America. Here is the distribution of the top colors for vehicles sold globally in 2010.

Color	Popularity (%)
Silver	26
Black	24
White	16
Gray	16
Red	6
Blue	5
Beige, brown	3
Other colors	

What color is your car? (cont.)

a) Fill in the percent of vehicles that are in other colors.

b) Make a graph to display the distribution of color popularity.

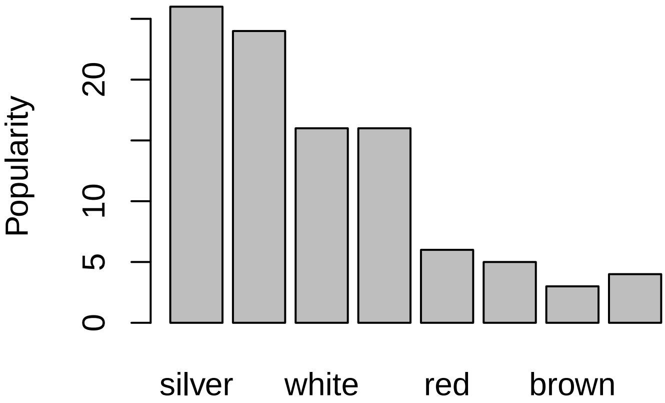
Solution

a) Other $= 100 - (26 + 24 + 16 + 16 + 6 + 5 + 3) = 4.$

Solution (pie chart)

```
## Step 1. Entering Data;
popularity=c(26,24,16,16,6,5,3,4);
color=c("silver","black","white",
"gray","red","blue","brown","other");


## Step 2. Making bar graph.

barplot(popularity,names.arg=color,ylab="Popularity")
```

Solution (bar chart)



Exercise

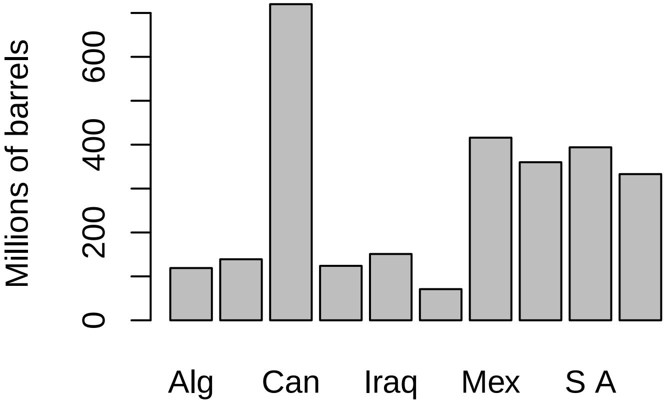
The following table lists the top 10 countries and amounts of oil (millions of barrels annually) they exported to the United States in 2010.

Country	Oil Imports (millions of barrels annually)
Algeria	119
Angola	139
Canada	720
Colombia	124
Iraq	151
Kuwait	71
Mexico	416
Nigeria	360
Saudi Arabia	394
Venezuela	333

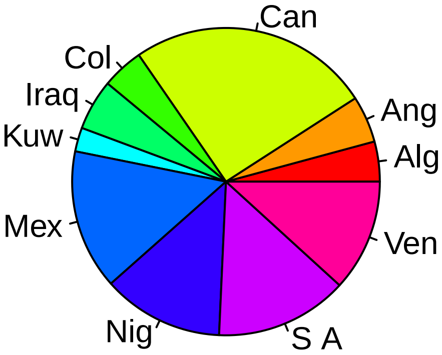
Exercise (cont.)

a. Draw a bar chart.
b. Draw a pie chart.

Solution (bar chart)



Solution (pie chart)



Newspaper Readership Survey

A major North American city has four competing newspapers: the *Globe and Mail (G & M)*, *Post*, *Star*, and *Sun*. To help design advertising campaigns, the advertising managers of the newspapers need to know which segments of the newspaper market are reading their papers. A survey was conducted to analyze the relationship between newspapers read and occupation. A sample of newspaper readers was asked to report which newspaper they read - *Globe and Mail (1), Post (2), Star (3), Sun (4)* - and indicate whether they were blue-collar workers (1), white-collar workers (2), or professionals (3).

Newspaper Readership Survey

Some of the data are listed here.

Reader	Occupation	Newspaper
1	2	2
2	1	4
3	2	1
⋮	⋮	⋮
352	3	2
353	1	3
354	2	3

Determine whether the two nominal variables are related.

Solution

By counting the number of times each of the 12 combinations occurs, we produced the following Table.

	Newspaper				
Occupation	**G & M**	**Post**	**Star**	**Sun**	**Total**
Blue Collar	27	18	38	37	120
White Collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	89	112	81	72	354

Solution

Table of Row Relative Frequencies for our example.

	Newspaper				
Occupation	**G & M**	**Post**	**Star**	**Sun**	**Total**
Blue Collar	0.23	0.15	0.32	0.31	1
White Collar	0.27	0.40	0.19	0.14	1
Professional	0.26	0.40	0.17	0.16	1
Total	0.25	0.32	0.23	0.20	1

Solution

```
# Step 1. Entering data;

news.tab=matrix(c(0.23,0.27,0.26,0.15,0.40,0.40,
0.32,0.19,0.17,0.31,0.14,0.16),nrow=3,ncol=4);

news.tab;
```

Solution

```
##      [,1] [,2] [,3] [,4]
## [1,] 0.23 0.15 0.32 0.31
## [2,] 0.27 0.40 0.19 0.14
## [3,] 0.26 0.40 0.17 0.16
```

Solution (R code)

```
# Giving names to columns and rows;

colnames(news.tab)=c("GandM","Post","Star","Sun");

rownames(news.tab)=c("Blue Collar",
"White Collar", "Professional");

news.tab;
```

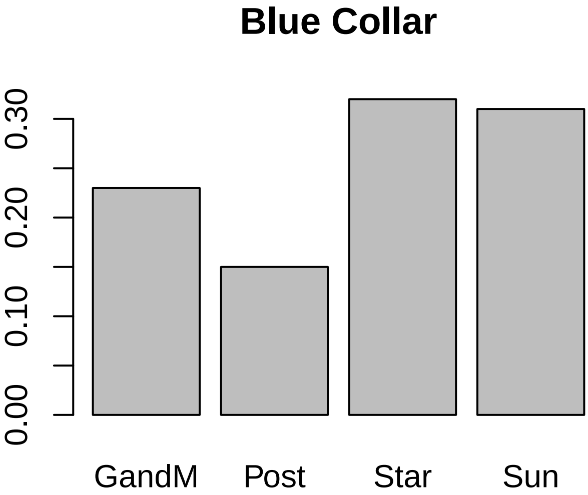
Solution (R code)

```
##                 GandM Post Star  Sun
## Blue Collar   0.23 0.15 0.32 0.31
## White Collar  0.27 0.40 0.19 0.14
## Professional  0.26 0.40 0.17 0.16
```

Solution (R code)

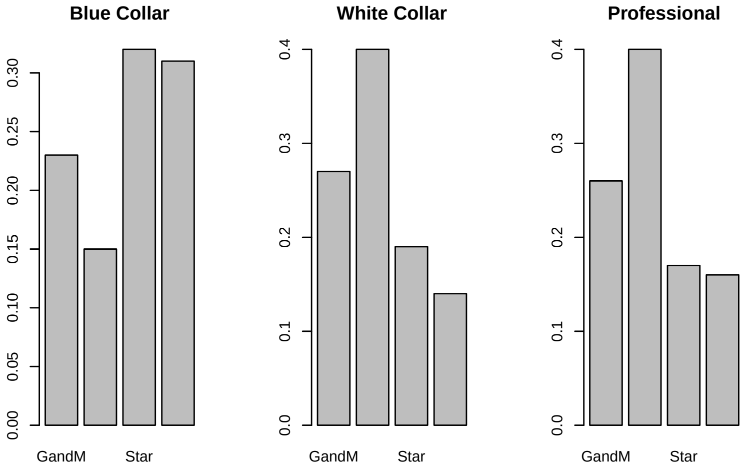
```
# Step 2. Bar chart for Blue Collar;

barplot(news.tab[1, ]);

title("Blue Collar");
```

Solution (R code)



Solution (R code)

```
par(mfrow=c(1,3));

barplot(news.tab[1, ]);

title("Blue Collar");

barplot(news.tab[2, ]);

title("White Collar");

barplot(news.tab[3, ]);

title("Professional");
```


Blue Collar
0.00
0.05
0.10
0.15
0.20
0.25
0.30
GandM
Star
White Collar
0.0
0.1
0.2
0.3
0.4
GandM
Star
Professional
0.0
0.1
0.2
0.3
0.4
GandM
Star

Reading data from txt files

Now, we will learn how we can create tables from our data and calculate relative frequencies.

```
# Step 1. Entering data;
# url of news data;
news_url = "http://www.math.unm.edu/~alvaro/news.txt"
# import data in R;
news_data= read.table(news_url, header = TRUE);

news_data[1:5, ];
```

Reading data from txt files

```
##   Reader Occupation Newspaper
## 1      1          2         2
## 2      2          1         4
## 3      3          2         1
## 4      4          3         2
## 5      5          1         3
```

Creating table of frequencies (R Code)

```
# Step 2. Making table of frequencies;

xtabs(~ Occupation + Newspaper, data = news_data);
```

Creating table of frequencies (R Code)

```
##            Newspaper
## Occupation  1  2  3  4
##          1 27 18 38 37
##          2 29 43 21 15
##          3 33 51 22 20
```

Creating table of relative frequencies

```
# Step 3. Making table of relative frequencies;

freq.tab=xtabs( ~ Occupation + Newspaper,
data = news_data);

rel.freq.tab=prop.table(freq.tab,margin=1);

rel.freq.tab;

# margin=1 is telling R to compute relative frequencies
# with respect to ROW totals;
# margin =2 would do the same with respect to
# COLUMN totals;
```

Creating table of relative frequencies

```
##            Newspaper
## Occupation         1         2         3         4
##          1 0.2250000 0.1500000 0.3166667 0.3083333
##          2 0.2685185 0.3981481 0.1944444 0.1388889
##          3 0.2619048 0.4047619 0.1746032 0.1587302
```

Solution (R code)

```r
# Giving names to columns and rows;

colnames(rel.freq.tab)=c("GandM","Post","Star","Sun");

rownames(rel.freq.tab)=c("Blue Collar",
"White Collar", "Professional");

rel.freq.tab;
```

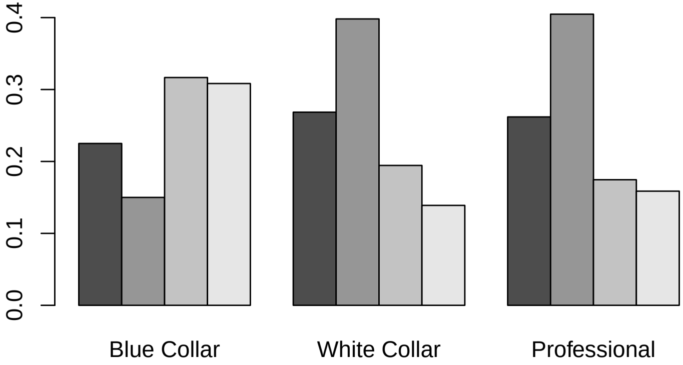
Solution (R code)

```
##                Newspaper
## Occupation         GandM      Post      Star       Sun
##   Blue Collar  0.2250000 0.1500000 0.3166667 0.3083333
##   White Collar 0.2685185 0.3981481 0.1944444 0.1388889
##   Professional 0.2619048 0.4047619 0.1746032 0.1587302
```

Making bar charts

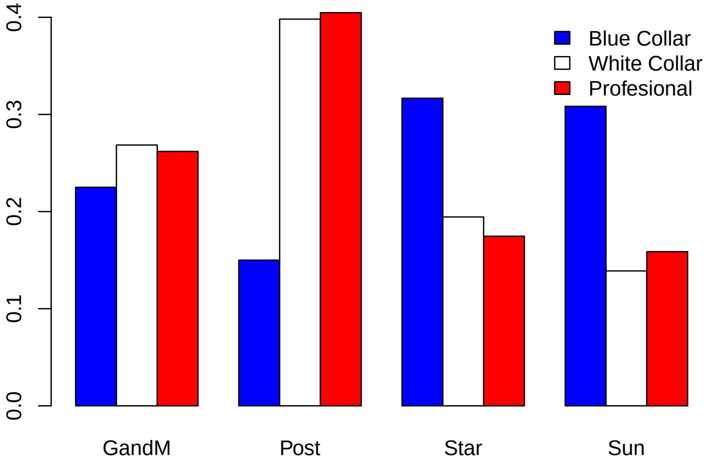
```
# Step 4. Graphing table of row relative
# frequencies;

barplot(t(rel.freq.tab),beside=T);
```

Making bar charts



Making bar charts (another version)

```r
# Making bar charts;

barplot(rel.freq.tab,beside=T,
col=c("blue","white","red"));

legend("topright",
c("Blue Collar","White Collar","Profesional"),
bty="n",fill=c("blue","white","red") );

# legend wil add a legend to our bar chart;
# bty="n" means no box for our legend;
```


0.0
0.1
0.2
0.3
0.4
Blue Collar
White Collar
Profesional
GandM
Post
Star
Sun