Kyle Thomas

Math 341

1. For each of the following, what are the individuals (or observational units) and what is the variable? Is the variable categorical or quantitative?

   a) People in a city are asked if they support a new recycling law.

   The people in the city are the observational units. The variable is if they support the new recycling law, which is yes or no, therefore categorical.

   b) Record the percentage change in the price of a stock for 100 stocks publicly traded on Wall Street.

   Observational units are the 100 publicly traded stocks on wall street. The variable is the percentage change, which is quantitative.

   c) Record whether or not the literacy rate is over 75% for each country in the world.

   The observational units are the literacy rate of each country in the world. The variable is yes or no, if it is above or below 75%, therefore categorical.

2. In each of the following, a relationship between two variables is described. In each case, we can think of one variable as helping to explain the other. Identify the explanatory variable and the response variable.

   a) Lung capacity and number of years smoking cigarettes.

   The explanatory would be the number of years smoking cigarettes and the response variable would be the lung capacity.

   b) Blood alcohol content (BAC) and number of alcoholic drinks consumed

   The explanatory variable would be the number of alcoholic drinks consumed, while the blood alcohol content would be the response variable.

3. The American Academy of Pediatrics recommends that parents begin reading to their children soon after birth and that parents set limits on screen time. A new study reinforces these recommendations. In the study, 27 four-year-olds were presented with stories in three different formats: audio (sound only), illustrated (sound and pictures), and animated (sound and animation). During the presentations, a magnetic resonance imaging (MRI) machine measured each child's brain connectivity. The researches found a "Goldilocks effect," in which audio was too cold (with low brain connectivity as the children strained to understand) and animation was too hot (with low brain connectivity as the animation did all the work for the children). The highest connectivity (just right!) was found with the illustrated format, which simulates reading a book to a child.

a) What is the explanatory variable? Is it categorical or quantitative?

The explanatory variable would be the type of format the story was told to the child, which is categorical.

b) What is the response variable? Is it categorical or quantitative?

The response variable would be to observe the effect it has on the Childs brain connectivity, which is quantitative.

c) How many observational units (or individuals) are there?

There are 27 four-year-olds.

4. Suppose we want to investigate the question, "Does voter turnout differ by political party?" How might we collect data to answer this question? What would the observational units be? What would the variable(s) be?

One way to collect data for this question would be to

5. As of 2019, there were 217 countries listed by the World Bank. Data from these countries is in the AllCountries.csv file in the Data Files page on the class Canvas site. Select a random sample of 5 of these countries. Write the names of the countries you got and explain how you got them.

My program randomly selected Zimbabwe, Luxembourg, Kazakhstan, Vietnam, Albania. First my program opened the AllCountries.csv file. Then it selects 5 random countries, and prints them.

```
countries <- read.csv(file = "~/Desktop/AllCountries.csv", head = TRUE)
x <- sample(seq(1,nrow(countries)), 5, replace =FALSE)
print(countries[x,1])
```

6. A recent study shows that just one session of cognitive behavioral therapy can help people with insomnia. In the study, forty people who had been diagnosed with insomnia were randomly divided into two groups of 20 each. People in one group received a one hour cognitive behavioral therapy session while those in the other group received no treatment. Three months later, 14 of those in the therapy group reported sleep improvements while only 3 people in the other group reported improvements.

a) What are the observational units in this study?

The observational units in this study are the forty people diagnosed with insomnia.

b) What are the relevant variables? Identify each as categorical or quantitative.

The variables are the treatment received by one group (categorical) and the sleep improvements reported (categorical).

c) Indicate explanatory and response variables.

The explanatory variables are the therapy sessions while the response variables are to observe the sleep improvements.

d) If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

There would be 40 rows for each participant and there would be two columns to indicate if they were selected for the therapy sessions (yes or no) and another to indicate if they felt improvements in their sleeping (yes or no).

7. For each of the following, state whether the data are best described as a population or a sample.

a) To estimate the size of trout in a lake, an angler records the weight of 15 trout she catches over a weekend.

The data is a sample because it is only a portion of the fish in the lake. There is almost no way to get the population of the lake without catching every fish.

b) A subscription-based music app tracks its total number of active users.

The data is the population because it includes all of their active users.

8. For each of the following, describe the sample and describe a reasonable population the sample could represent.

a) A sociologist conducting a survey at a mall interviews 120 people about their cell phone use.

The sample is 120 people at the mall. The reasonable population could represent people from ages 15 to 60.

b) Five hundred Canadian adults are asked if they are proficient on a musical instrument.

Sample is the 500 Canadian adults. The reasonable population could be Canadian adults.

c) A cell phone carrier sends a satisfaction survey to 200 randomly selected customers.

The sample is the 200 randomly selected cell phone customers. The reasonable population could be everyone with that cell phone provider.

d) The Nielsen Corporation attaches databases to televisions in 1000 households throughout the US to monitor which shows are being watched and produce the Nielsen Ratings for television.

The sample is the 1000 households in the US with the databases. The reasonable population could be every household throughout the US with a TV.

9. For each of the following, describe the sample, the population of interest, a reasonable population we can generalize to given the sample, and any the bias in the sampling.

a) To estimate the proportion of Americans who support changing the drinking age from 21 to 18, a random sample of 100 college students are asked the question, "Would you support a measure to lower the drinking age from 21 to 18?"

The sample would be the 100 college students that were asked. The population of interest would be college students. A reasonable population would be 18 to 22 year olds. There would be a lot of bias in the sampling because the question directly affects the population of interest.

b) To estimate the average number of tweets from all twitter accounts in 2019, a certain actor randomly selected 10 of her followers and counted their tweets.

The sample would be the 10 random followers the actor selected. The population of interest would be her followers. A reasonable population could be twitter users in general. The only potential bias is that the actor selected another celebrity who follows her who tweets more frequently than a less public figure.

c) To investigate interest across all residents of the US in a new type of ice skate, a random sample of 1500 people in Minnesota are asked about their interest in the product.

The sample would be the 1500 Minnesota people asked. The population of interest would be people who live in Minnesota. A reasonable population would be in snowy and cold places. The bias could be that people from Minnesota prefer a different type of skate than someone from New York or California, which have much larger populations.

10. In each of the following, state whether or not the sampling method described produces a random sample from the given distribution.

a) The population is the approximately 25,000 protein-coding genes in human DNA. Each gene is assigned a number (from 1 to 25,000) and computer software is used to randomly select 100 of these numbers yielding a sample of 100 genes.

Yes, this produces a random sample because the computer program should not have any bias on which number it selects.

b) The population is incoming students at a particular university. The name of each incoming student is thrown into a hat, the names are mixed, and 20 names are drawn without replacement from the hat.

Yes, this produces a random sample because the names are randomly being selected out of a hat.

c) The population is all employees at a company. All employees are emailed a link to a survey.

No, this would not be random sampling because everyone is included in the sample.

d) The population is adults between the ages of 18 and 22. A sample of 100 students is collected from a local university, and each student at the university had an equal chance of being selected for the sample.

Yes, this would be a random sample, although it did not explain how each student was selected, it did say each student had an equal chance of being selected.

11. In each of the following, decide whether we should trust the results of the study. Is the method of data collection biased? If it is, explain why.

a) In order to collect data to estimate the average number of hours a week that all college students study, ask a random sample of students at the library on a Friday night, "How many hours a week do you study?"

I think we should trust the results of the study because the students are at the library on the weekend studying. The data collection would also be biased because the students at the library studying on a Friday are more likely to study more than those who are not at the library.

b) Take a random sample of one type of printer and test each printer in the sample to see how many pages of text each will print before the ink runs out. Use the average from the sample to estimate how many pages, on average, all printers of this type will print before the ink runs out.

I think we should trust this sample because they are doing the test on the same type of printer. The only biased could be the text on the pages being different, which was not stated if they are the same or not.

12. This problem is about the used Honda Civic data linked to our class Canvas site under pages>Data Files>UsedHondaCivics2015.csv. The 22 cars were advertised for sale online (Kelly Blue Book kbb.com) within 50 miles of the bookstore at Cal Poly San Luis Obispo, California on August 17, 2015.

a) What are the five variables given in the data file? Identify each as categorical or quantitative. (Choose the best answer, even though some can go either way. Or justify why you chose one over the other.)

year - categorical

Mileage - quantitative

Price - quantitative

Type - categorical

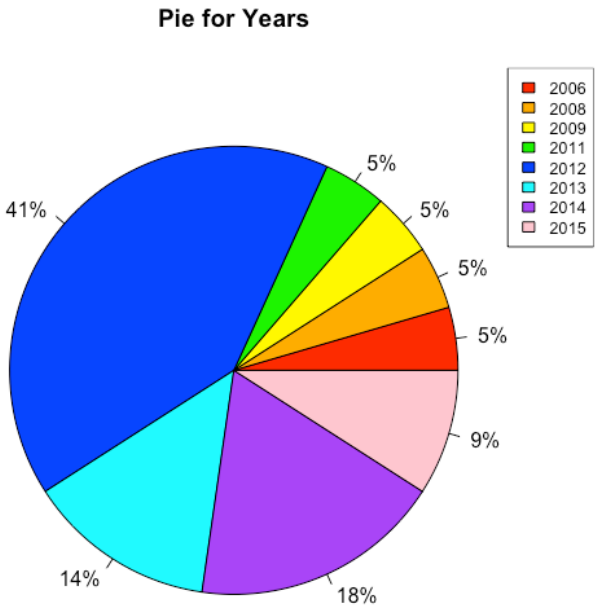
Age - quantitative

b) For each categorical variable, create a frequency table and a relative frequency table.

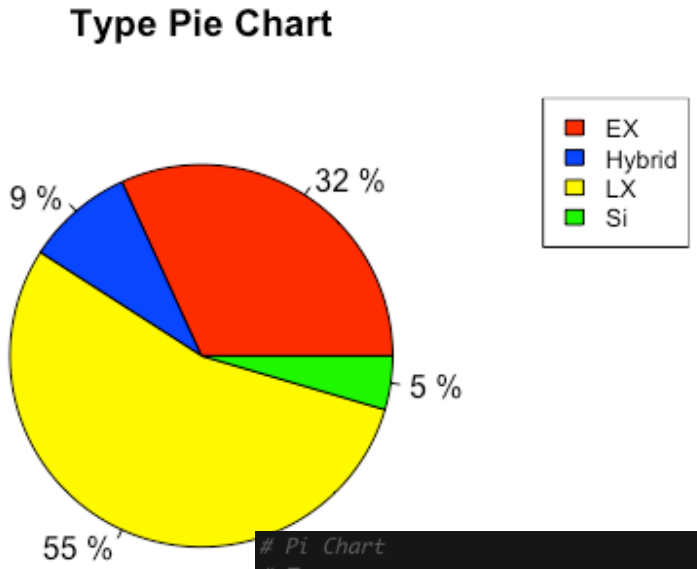
```
hondaData <- read.csv(file = "~/Desktop/UsedHondaCivics2015.csv", head = TRUE)
summary(hondaData)

f <- table(hondaData$year,hondaData$type)

rf = f / 22

f

rf
```

```
> f

      EX Hybrid LX Si
  2006  0      0  0  1
  2008  1      0  0  0
  2009  0      0  1  0
  2011  0      0  1  0
  2012  3      2  4  0
  2013  1      0  2  0
  2014  1      0  3  0
  2015  1      0  1  0
>
> rf

             EX     Hybrid         LX         Si
  2006 0.00000000 0.00000000 0.00000000 0.04545455
  2008 0.04545455 0.00000000 0.00000000 0.00000000
  2009 0.00000000 0.00000000 0.04545455 0.00000000
  2011 0.00000000 0.00000000 0.04545455 0.00000000
  2012 0.13636364 0.09090909 0.18181818 0.00000000
  2013 0.04545455 0.00000000 0.09090909 0.00000000
  2014 0.04545455 0.00000000 0.13636364 0.00000000
  2015 0.04545455 0.00000000 0.04545455 0.00000000
```

c) For each categorical variable, create a pie chart, showing the categories and percentages in each sector of the circle.

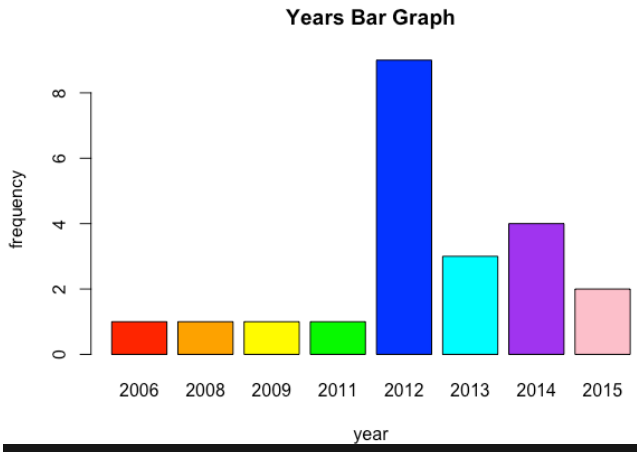



```
# Pi Chart
# Years
tableYear = table(hondaData$year)
slices<-(tableYear)
lab <- c(2006, 2008, 2009, 2011, 2012, 2013, 2014, 2015)
colors <- c("red", "orange", "yellow", "green", "blue", "cyan", "purple", "pink")
percent <- round(slices/sum(slices)*100)
lab <- paste(percent)
lab <- paste(lab, "%", sep="")
pie(slices, labels = lab, col = colors,main="Pie for Years")
legend("topright",c("2006", "2008", "2009", "2011", "2012", "2013", "2014", "2015"), cex=0.8, fill=colors)
```

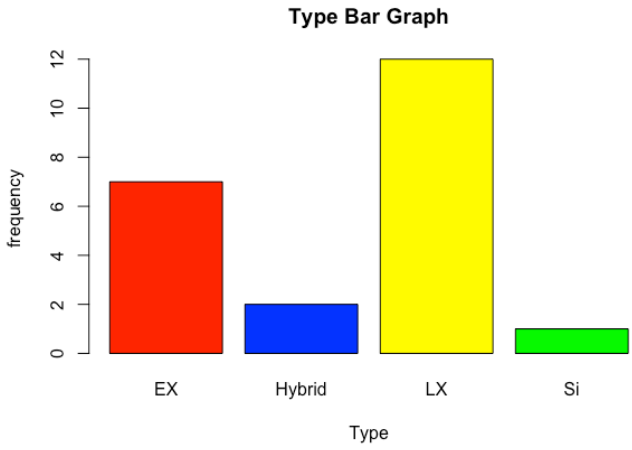



```
# Pi Chart
# Type
tableType <- table(hondaData$type)
block <- c(tableType)
colors <- c("red", "blue", "yellow", "green")
percent <- round(block/sum(block)*100)
typeLab <- paste(percent)
typeLab <- paste(typeLab, "%", sep=" ")
pie(block, main="Type Pie Chart", col=colors, labels=typeLab)
legend("topright", c("EX", "Hybrid", "LX", "Si"),cex=0.8, fill = colors)
```

d) For each categorical variable, create a bar graph, showing the categories along the horizontal axis and frequencies along the vertical axis.




```
# Bar Graph
# Year
# tableYear = table(hondaData$year)s
lab <- c(2006, 2008, 2009, 2011, 2012, 2013, 2014, 2015)
block <- (tableYear)
colors <- c("red", "orange", "yellow", "green", "blue", "cyan", "purple", "pink")
barplot(block,main="Years Bar Graph",ylab="frequency",xlab="year",beside=TRUE,col=colors)
```




```
# Bar Graph
# Type
tableType <- table(hondaData$type)
block <- c(tableType)
class <- c(1,12,7,2)
colors <- c("red", "blue", "yellow", "green")
barplot(block,main="Type Bar Graph", ylab="frequency",xlab="Type",beside=TRUE,col=colors)
```