Math 183 HW2, Name: Chandler Blaid Burgess , **PID:** A98029477

1.8 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[58]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent? A UK Resident

(b) How many participants were included in the survey? 1691

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Sex - categorical, not ordinal

age - numerical, discrete

marital - categorical, not ordinal

grossIncome - Categorical, ordinal

Smoke - Categorical, not ordinal

amtWeekends - numerical, discrete

amtweekdays - numerical, discrete

1.14 Cats on YouTube. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.

(a) Percentage of all videos on YouTube that are cat videos. Population parameters

(b) 2%. Sample statistic

(c) A video in your sample. Observation

(d) Whether or not a video is a cat video. Variable

1.10 Cheaters, scope of inference. Exercise 1.9 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

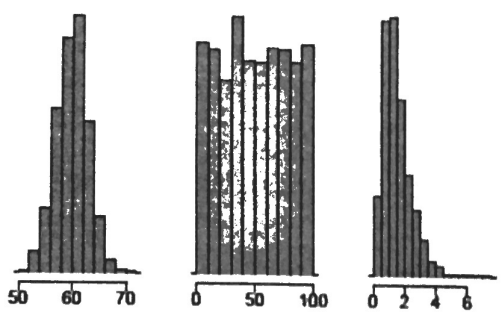
(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

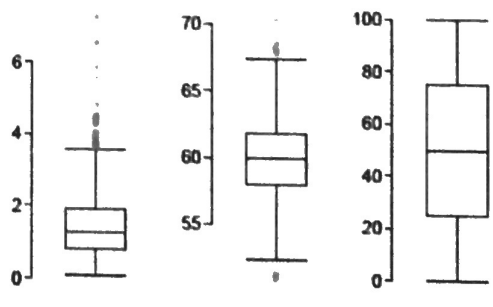
a) The population of interest is children between the ages of 5 and 15. The sample in this study is 160 children between the ages of 5 and 15.

b) It is not stated that the 160 children were chosen at random, thus the sample is not necessarily representative of the population and cannot be generalized. Since the sample set was not randomized, biases could have been introduced into the data, preventing a causal relationship to be infered.

1.50 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.



(a) (b) (c)



(1) (2) (3)

[2] unimodal, Symmetric

[3] multimodal, symmetric

[1] unimodal, Right skew

1.28 Reading the paper. Below are excerpts from two articles published in the *NY Times*:

(a) An article titled *Risks: Smokers Found More Prone to Dementia* states the following:[61]

> "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking, 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

(b) Another article titled *The School Bully Is Sleepy* states the following:[62]

> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

---

[61] R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". In: *New York Times* (2010).
[62] T. Parker-Pope. "The School Bully Is Sleepy". In: *New York Times* (2011).

a) This is an observational study, thus we cannot causation of dementia, from smoking. We can, however, say that there is an association.

b) The statement of causation of bullying from sleep disorders is not justified, because this is based on an observational study. Thus, we can at best assert that sleeping disorders and bullying are corelated.

1.32 Vitamin supplements. In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.[64]

(a) Was this an experiment or an observational study? Why?

(b) What are the explanatory and response variables in this study?

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

a) No, it was an experiment because the data was collected in a way that was guided by the hand of the researchers.

b) Explanatory => amount of vitamin C
Response => length of cold

c) yes.

d) yes.

e) No, the number of patients that fail to take the pills will be equal in each group, because the patients were blinded.

1.46 **Medians and IQRs.** For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(a) (1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

(b) (1) 3, 5, 6, 7, 9
(2) 3, 5, 7, 8, 9

(c) (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

(d) (1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

a) Medians & IQR's are equal

b) (2) has a higher median & IQR

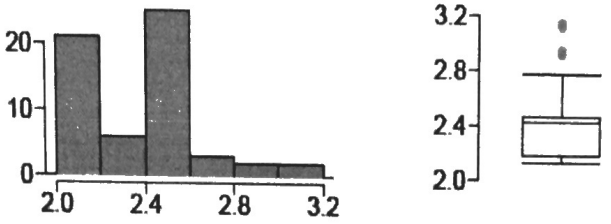
c) (2) has a higher median, however, the IQR's are equal

d) (2) has a higher median & IQR

1.60 **A new statistic.** The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a) $\frac{\bar{x}}{median} = 1$ Symmetric
(b) $\frac{\bar{x}}{median} < 1$ Skewed Right
(c) $\frac{\bar{x}}{median} > 1$ Skewed Left

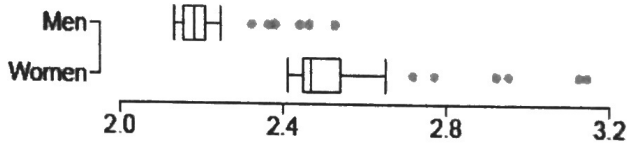
1.54 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown below.



a) That the shape of the graph is bimodal & the width of the spread of values. The box-plot shows the outliers in the data.

b) There may be central tendancies specific to males and to females.

c) men's times were generally lower than womens, however, they both have outliers, skewing them to the right.

1.56 Distributions and appropriate statistics, Part II . For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

a) The data is right-skewed, because of the number of houses over $6 million. The median & IQR would be best to resist these outliers.

b) The data is symmetrical, because every $300k encompasses a quarter of the housing market. This data has very few outlies, so the mean, median and the IQR & standard deviation can each be used to describe the graph equally.

c) The data is skewed right because of the excessive drinkers. The median & IQR would be best, to deal with those outliers.

d) The data will be skewed right, because the executives make a significant amount more than the average. To resist these outliers, median & IQR would be best.

1.66 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[70]

		Political ideology			
		Conservative	Moderate	Liberal	Total
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910

(a) What percent of these Tampa, FL voters identify themselves as conservatives?

(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

(c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

a) 40.88% ($^{372}/_{910}$)

b) 30.55% ($^{278}/_{910}$)

c) 6.26% ($^{57}/_{910}$)

d) conservatives = 15.32% ($^{57}/_{372}$)

Moderates = 33.06% ($^{120}/_{363}$)

Liberals = 57.71% ($^{101}/_{175}$)

e) No, because based on your political ideology you are more likely to hold a certain opinion, as shown above.

Questions on R

In the "HW Data Sets" subfolder of the "Homework" folder (on Triton Ed), you will find the data set email.txt. You can find a complete description of this data set, and all others used in this course, in the Descriptions PDF in the same directory. All code that you are asked to write should be in R. In brackets after each question, you will see what things should be included in your answer. For example, on R2, you should include the code you used to find your answer, as well as your answer (which is simply a number).

R1. Write a line of code that reads this text file in and stores it in the data frame called emails. [code]

[> emails ← read.csv("email.txt", header = TRUE, sep = "", row.names = NULL)]

R2. How many spam emails are in the data set? [code, answer]

[> sum(emails$spam), 620]

R3. Create a contingency table for spam emails vs. number of attachments. [code, table]. In general, how are the dimensions of the table related to the two variables you are using?

[> table(emails$spam, emails$attach)

	0	1	2	3	4	5	6	7	8	9	10	20	21
0	3147	77	56	10	2	2	2	2	1	0	1	0	1
1	491	81	34	9	1	2	0	0	0	1	0	1	0

]

R4. In words, describe the skew of the histogram of the number of exclamation points in the email messages themselves (not in the subjects). [code, skew] (You need not include the histogram. Try typing ?hist at the command prompt to learn how to better use the histogram function. Its default may not give you a helpful picture.)

[>hist(emails$exclaim_mess), right-skew]

R5. How many emails have more than 50,000 characters in them? [code, answer]

[>sum(emails$num_char > 50), 74]