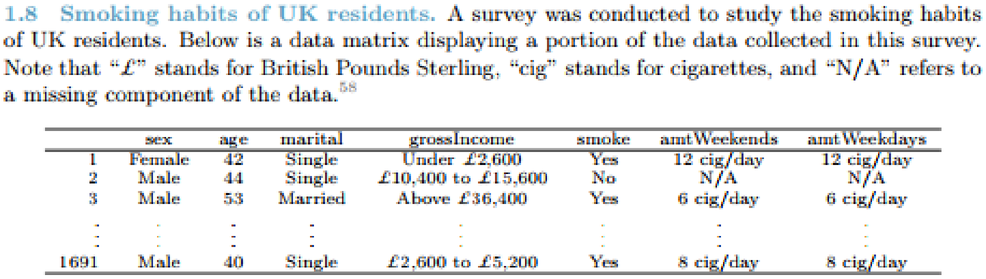
Data606 - Homework 1

Name: Joshua Sturm
Date: 08/25/2017

Chapter 1 problems - Page 57

1.8—



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

(a) Each row represents a resident or case.
(b) There were 1,691 participants.

(c)

Smoking Habits of UK Residents

	Numerical (continuous)	Numerical (discrete)	Categorical (nominal)	Categorical (ordinal)
sex			✓	
age		✓		
martial			✓	
grossIncome				✓
smoke			✓	
amtWeekends		✓		
amtWeekdays		✓		

1.10— Cheaters, scope of inference. Exercise 1.5 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 all children, and the sample used was 160 children aged 5 to 15.

(b) I don't believe these results can be generalized to the general population. Different cultures have differing views on (what constitutes) honesty. Also, a person's honesty may change with age (maturity, better understanding of the consequences, etc), so to use a sample with such a large range of ages may not be optimal. Furthermore, a sample size of 160 is not that large.

Since this is an experiment and not an observational study, it is possible to establish a causal relationship.

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: "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: "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?

(a) Since this is a prospective observational study and not an experiment, we can't form a causal relationship.

(b) The statement is not justified, because it is not clear which symptom caused the other. Furthermore, there could be confounding variables, which would effect the reliability of the study.

A better conclusion would be that there *may* be some association between sleep disorders and behavioral problems, but no definite evidence that one causes the other.

1.36— A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure

representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.
(a) What type of study is this?
(b) What are the treatment and control groups in this study?
(c) Does this study make use of blocking? If so, what is the blocking variable?
(d) Does this study make use of blinding?
(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

(a) This study is a randomized experiment.

(b) The treatment group is the one assigned to exercise, and the control group is the one told to not exercise.

(c) This study does make use of blocking, and the blocking variable is the age of the people in the study.

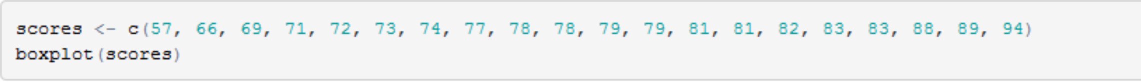
(d) This study does not use blinding. The researcher knows which group each person belongs to, and since the control group was told to *not* to exercise without being given a placebo, they, too, know which group they belong to.

(e) This is a randomized experiment, so it may be possible to establish a causal relationship. However, as mentioned in (d), this study is not blind, so there may be bias in the study which would effect its ability to be generalized to the rest of the population.

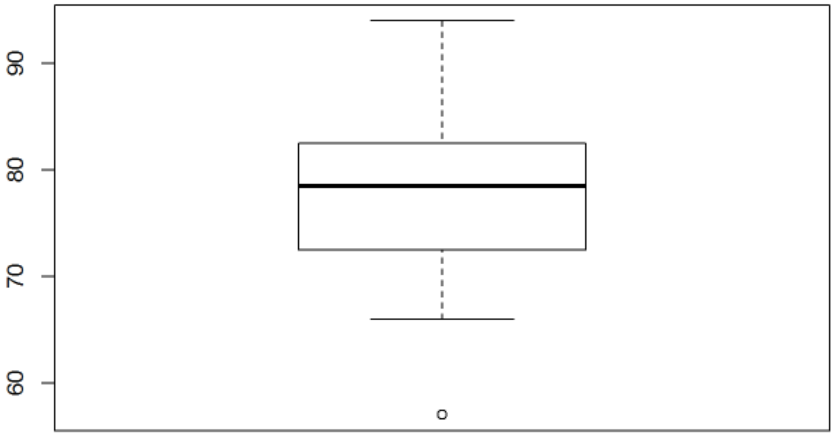
(f) In its current form, I wouldn't fund this study, since it's difficult to control the bias. If the researcher were to make the experiment double-blind, it would be more appealing.

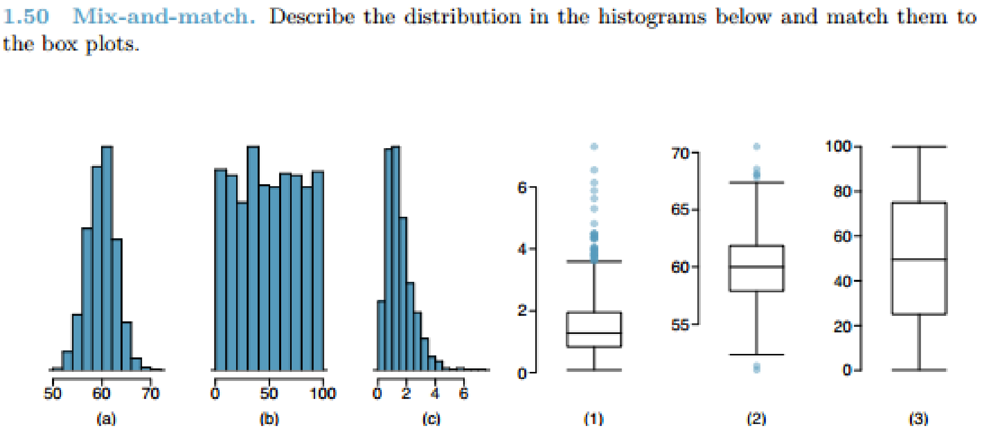
1.48— Below are the final exam scores of twenty introductory statistics students.
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94
Create a box plot of the distribution of these scores.

Begin by creating a single column vector named *scores* with each score as a case, and then plot it using the *boxplot* function in r.



```
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
boxplot(scores)
```





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

1.50

(a) is a symmetric unimodal distribution, and matches (2).

(b) is a multimodal distribution (has multiple humps), and matches (3).

(c) is a right skewed unimodal distribution, and matches (1).

1.56— 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.

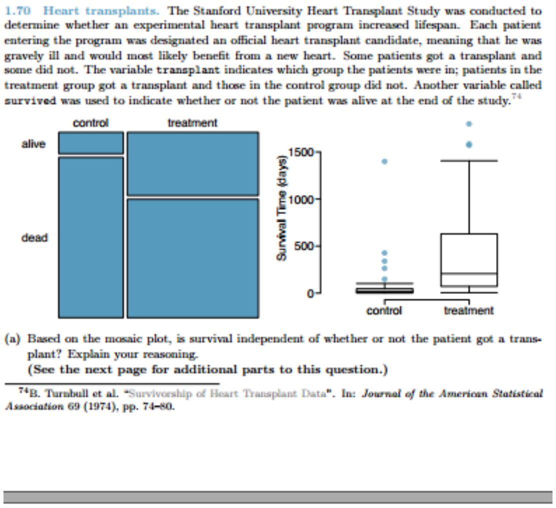
(a) Housing prices in a country where 25% of the houses cost below \$350,000, 50% of the houses cost below \$450,000, 75% of the houses cost below \$1,000,000 and there are a meaningful number of houses that cost more than \$6,000,000.
(b) Housing prices in a country where 25% of the houses cost below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

(a) Since the data is more concentrated on the left, the distribution is likely right skewed. Furthermore, the data would be best represented by the median and IQR due to the extreme (outliers) homes that cost >\$6,000,000, which would bias the data.

(b) The data seems to be pretty evenly distributed. Assuming that the "very few houses" at the extreme end are not enough to skew the data, it would be a symmetric distribution, best represented by the mean and standard deviation.

(c) Since the majority don't drink, and only a few do excessively, the histogram would taper out to the right, so the distribution would be right skewed. The data would be best represented by the median and IQR to reduce the bias of both extreme outliers.

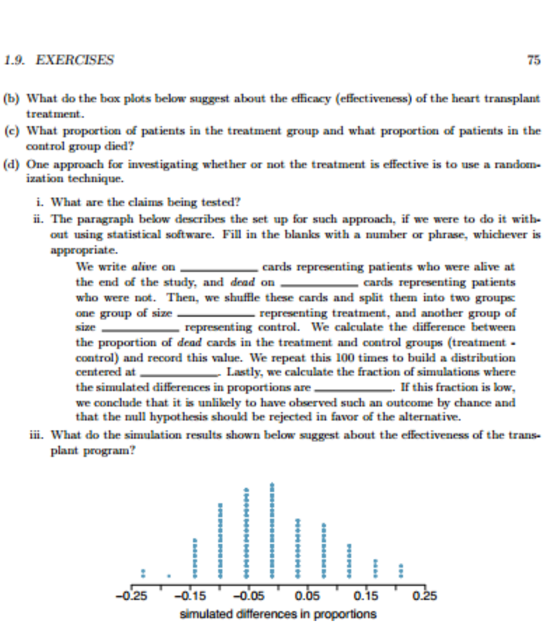
(d) The question is phrased in a vague manner - it could be symmetrical or right skewed. Either way, it is better to use the median and IQR to reduce the effects of the outliers.



1.70 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study.[74]

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
(See the next page for additional parts to this question.)

[74] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.



1.9. EXERCISES 75

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write *alive* on ______ cards representing patients who were alive at the end of the study, and *dead* on ______ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size ______ representing treatment, and another group of size ______ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at ______. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ______. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

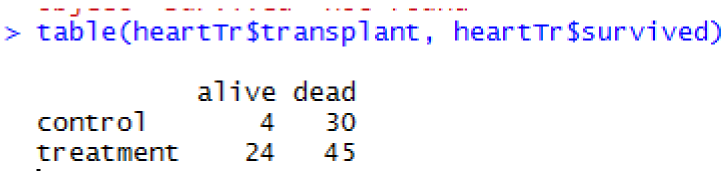
iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

1.70—

(a) Since a larger portion of the treatment group survived compared to those in the control group, the two variables are likely dependent.

(b) Aside from a few outliers in the control group, most people in the treatment survived for significantly longer than those in the control group. Additionally, the spread of days survived in the control group is not that large; that is, most survived for around the same amount of time.

(c)



```
> table(heartTr$transplant, heartTr$survived)

           alive dead
  control      4   30
  treatment   24   45
```

Proportion that died in the control group: $\frac{30}{34} \approx 88\%$.
Proportion that died in the treatment group: $\frac{45}{69} \approx 65\%$.

(d)(i) H_0 : Survivability is independent of whether the patient had a stent or not.

H_A : Survivability is dependent on whether the patient had a stent implanted.

(ii) We write alive on **28** cards representing patients who were alive at the end of the study, and dead on **75** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size **69** representing treatment, and another group of size **34** representing control. We calculate the difference between the proportion of dead cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at ≈ -0.23 . Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ≤ -0.23. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative. Note that the responses for this were taken from the table in (c).

(iii) Since there were only two instances where the fraction was less than or equal to -0.23, we can reject the null hypothesis in favour of the alternative, and conclude that it is not likely due to chance, and having a stent implanted does indeed affect survivability.