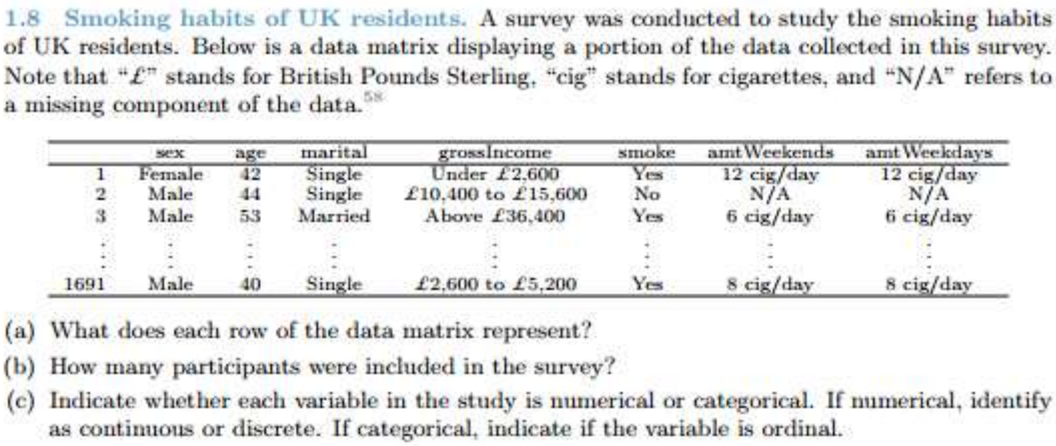
Leland Randles, DATA606

Homework, Chapter 1



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?

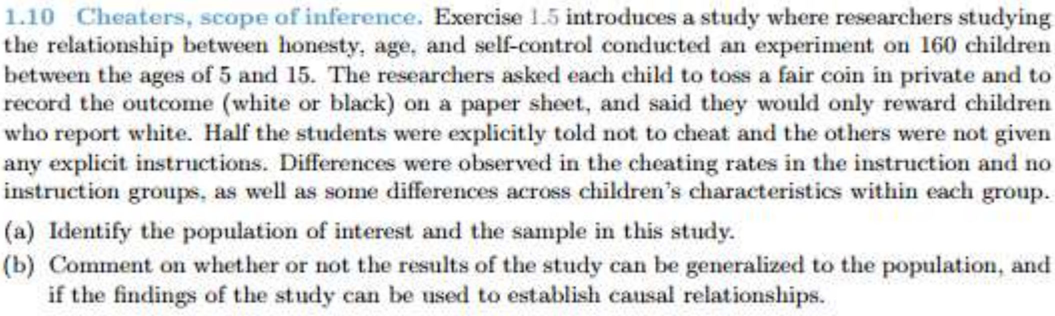
(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

1.8 Answers:

(a) Each row of the data matrix represents a UK resident

(b) 1,691

(c) sex: categorical, not ordinal
age: numerical, discrete (at least as long as it is represented as an integer…conceptually, it is continuous)
marital: categorical, not ordinal
grossIncome: categorical, ordinal (because it has been binned, conceptually income is numerical and discrete, assuming no fractional pound sterlings)
smoke: categorical, not ordinal
amtWeekends: numerical, discrete (data set assumes no partial cigarettes; all whole numbers)
amtWeekdays: numerical, discrete (data set assumes no partial cigarettes; all whole numbers)



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.

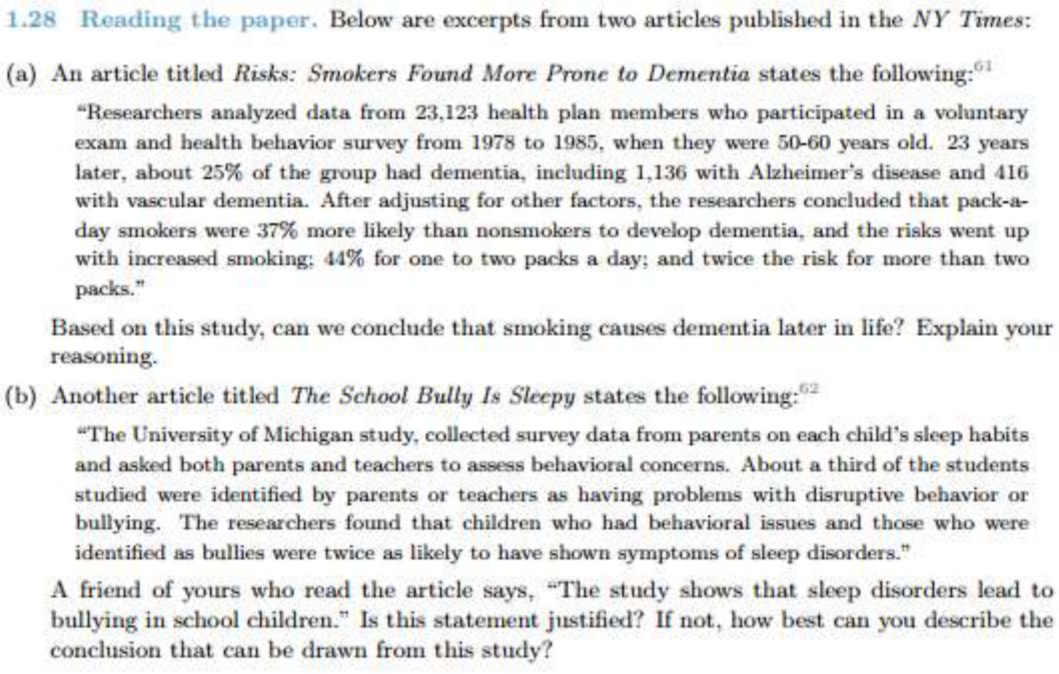
(a) Identify the population of interest and the sample in this study.

(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.

1.10 Answers:

(a) Children between the ages 5 and 15 is the population of interest; the sample is a specific group of 160 children between the ages of 5 and 15.

(b) The degree to which the results of the study can be generalized to the population depends on multiple factors such as 1) the effectiveness of the sampling, 2) the degree to which experimental design factors were adhered to (Controlling, Randomization, Replication and Blocking), and 3) a statistical analysis of whether the results could be a product of chance or due to a strong relationship between the variables. In this case, I think the only causal relationship which could potentially be established is the effect of the explicit instructions not to cheat on the propensity for cheating.



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]

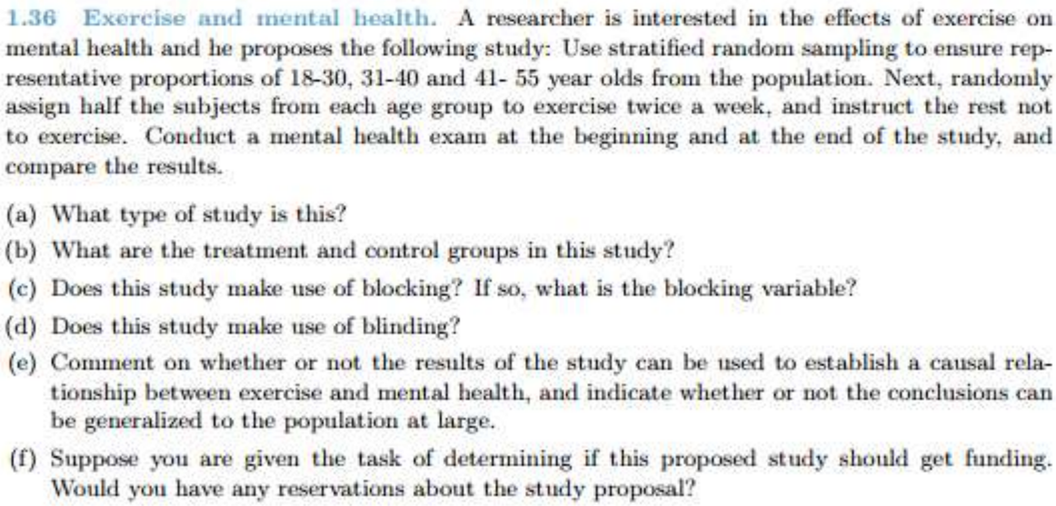
> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. 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?

1.28 Answers:

(a) I don't believe we can conclude that smoking "causes" dementia, at least not without much more information about the specifics of the research. For one thing, I would want to know the confounding variables. If obesity is heavily related to dementia, and people who smoke are way more likely to be obese, then it could be that obesity is the driver, not smoking. But the article does say the researchers "adjusted for other factors", so without seeing the specifics, it is tough to say. Also, this was not a randomized experiment, so there's the issue of the fact that this was a voluntary exam and health behavior survey, which means there could be some bias.

(b) The statement is not justified. We might be able to say there is a correlation between sleep disorders and bullying, but not causation.



1.36 Exercise and mental health. 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?

1.36 Answers:

(a) A randomized experiment

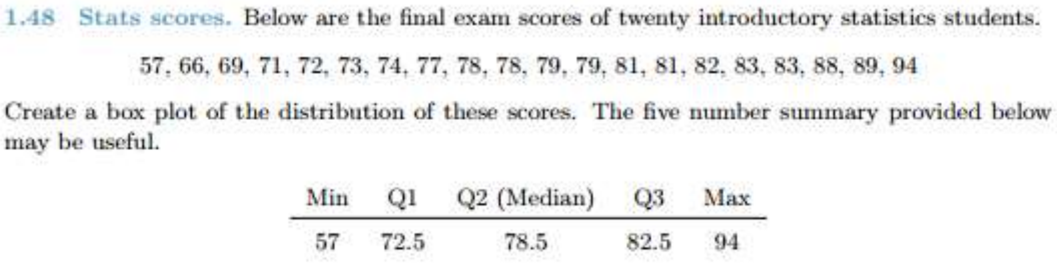
(b) The treatment group are the subjects who exercise twice a week

(c) Yes, the study is using blocking and the blocking variable is age range

(d) It does not, the subjects know which group they are in (though it is not clear whether they know that there is another group or not)

(e) Yes, if there is a statistically significant difference between the control group and the treatment group we could conclude that it was a causal factor for individuals included in the study. If the difference indicated meets the chosen significance level to not be the result of chance, then it could be generalized to the population at large.

(f) It isn't completely clear from the paragraph how strictly the principles of experimental design (Controlling, Randomization, Replication, and Blocking) are being used outside of the fact that they are blocking by age range. I'd want to know more specifics about this. I would also want to know the length of the study, and would propose more frequent mental health exams besides the beginning and end of the study.



1.48 Stats scores. Below are the final exam scores of twenty introductory statistics students.

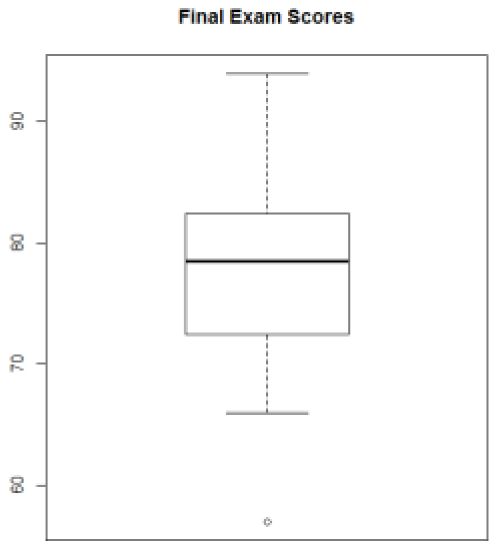
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

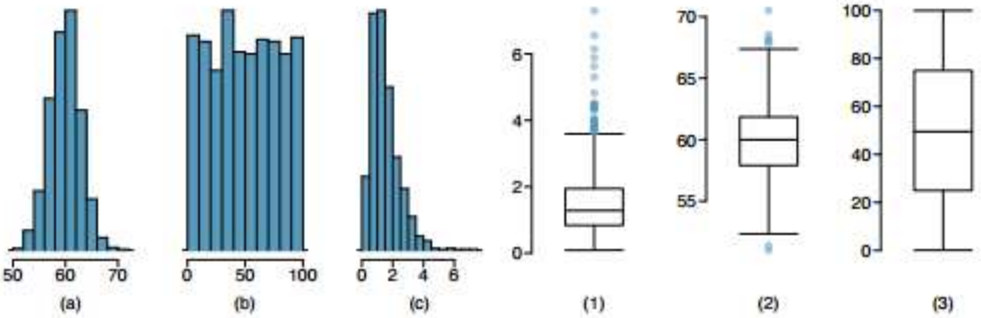
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

1.48 Answers:

```
x <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83,
83, 88, 89, 94)
boxplot(x, main = "Final Exam Scores")
```



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



1.50 Answers:

Histogram (a) is unimodal and minimally skewed (looks like a normal distribution).

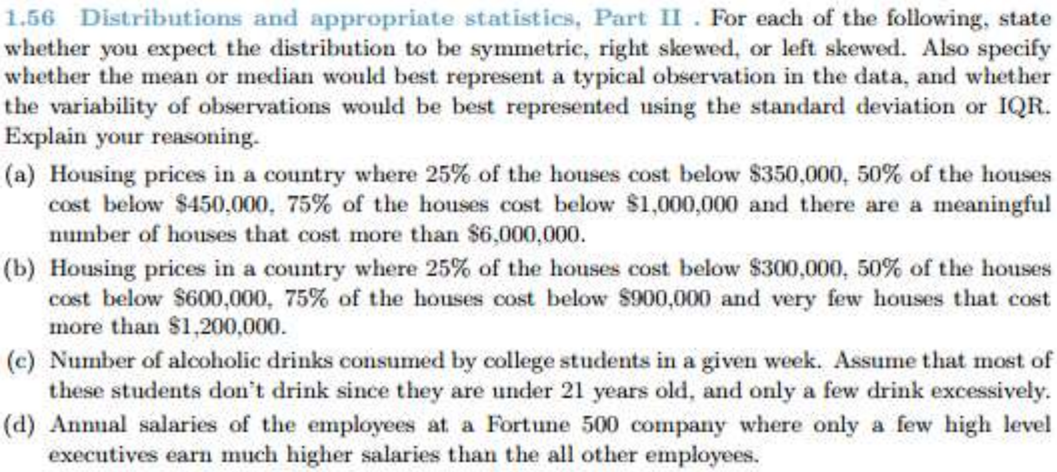
Histogram (b) is multimodal and not skewed

Histogram (c) is unimodal and skewed to the right

Histogram (a) goes with Boxplot (2)

Histogram (b) goes with Boxplot (3)

Histogram (c) goes with Boxplot (1)



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.

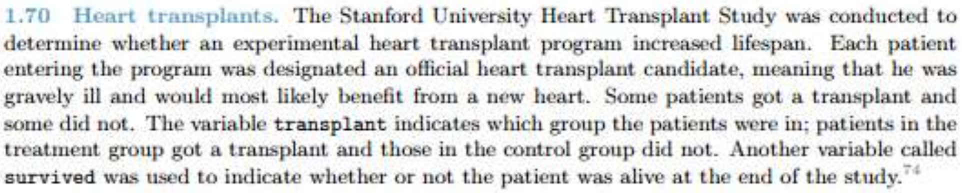
1.56 Answers:

(a) I would expect the distribution to be right-skewed. I think the median would best represent a typical observation because the average would be heavily affected by the houses over $6 mil.

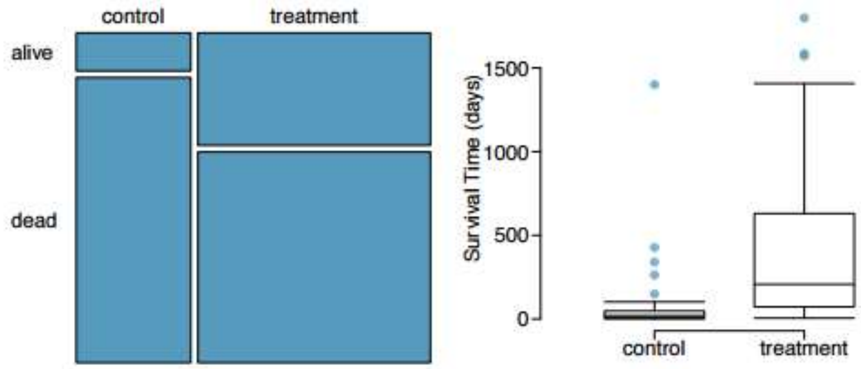
(b) I would expect the distribution to be almost symmetric, and I think the mean would best represent a typical observation.

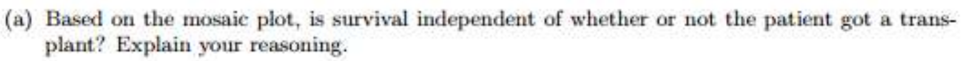
(c) I would expect the distribution to be left-skewed. I think the median would best represent a typical observation because the average would be heavily affected by the excessive drinkers.

(d) I would expect the distribution to be right-skewed. I think the median would best represent a typical observation because the average would be heavily affected by the high-level executive salaries.

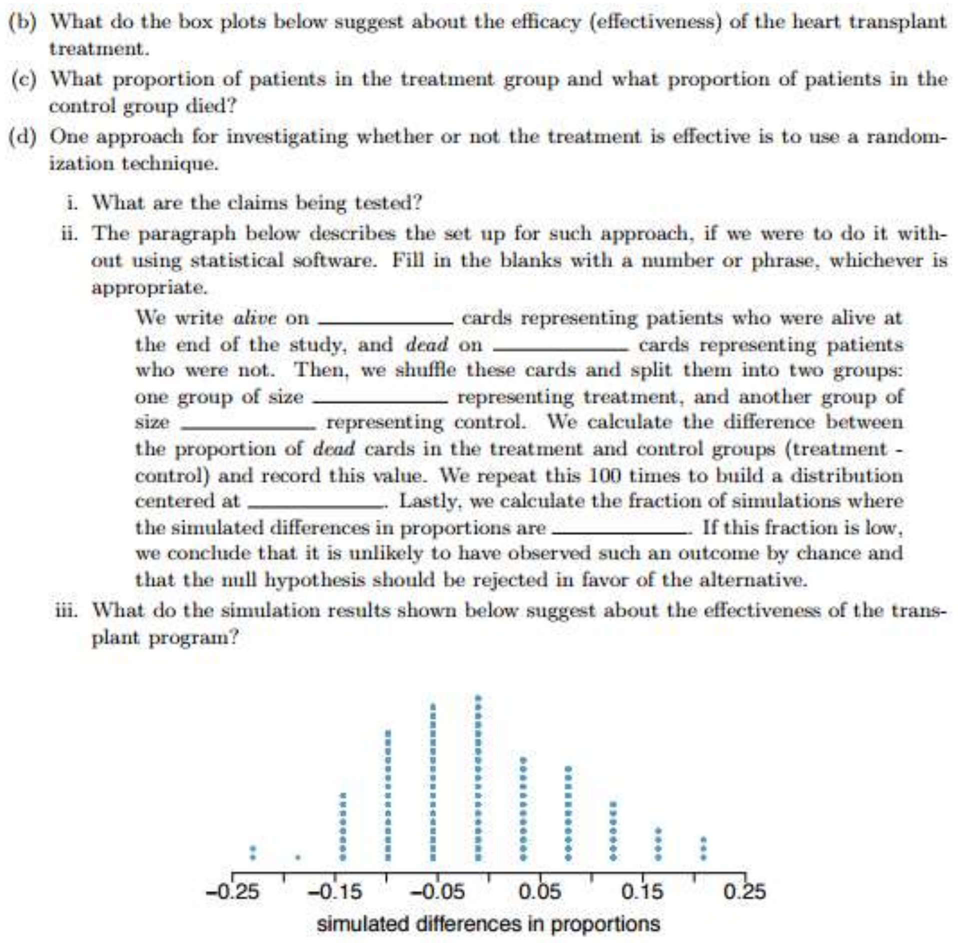


1.70 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most 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.



(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 Answers:

(a) It is hard to judge significance based on a mosaic plot, but based on the plot, I would say that survival is NOT independent of whether or not the patient got a transplant.

(b) It suggests that treatment does have a significant impact on survival/lifespan.

(c) 45 out of 69 (65.2%) of the treatment group died and 30 out of 34 (88.2%) of the control group died.

(d) i) That treatment and survival time are independent (independence model), or that treatment and survival time are not independent (alternative model).

ii) 6 blanks, in order: 28, 75, 69, 34, 0, significantly different.

iii) It suggests that it is very unlikely that the difference is the result of chance alone.