HW1-Intro_To_Data

Michael Y.

February 10, 2019

Homework - Chapter 1 - Intro to Data -

Exercises 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70 (pp.55-75)

Datasets for this chapter:

1.1 - migraine

1.2 - sinusitis

1.7 - iris

1.8 - smoking

1.13 - gss2010

1.15 - gpa_study_hours

1.16 - countyComplete

1.19 - cia_factbook

1.38 - mammals

1.49 - cia_factbook

1.51 - pm25_2011_durham

1.54 - marathon

1.61 - [simulated data]

1.63 - countyComplete

1.64 - countyComplete

1.65 - antibiotics_in_children

1.66 - immigration

1.67 - dream

1.68 - ppp_201503

1.69 - avandia

1.70 - heart_transplant

Exercise 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.

```
#install.packages("openintro")  ## already installed
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

```
#data(package='openintro') ## this just lists the names of the datasets -- the relevant one is "smoking
data(smoking)
#str(smoking)         ## the smoking dataset includes more columns than are shown in the textbook
smoking_subsetdf = subset(smoking,
                          select=c(gender,age,maritalStatus,grossIncome,smoke,amtWeekends,amtWeekdays))
names(smoking_subsetdf)[1]="sex"
names(smoking_subsetdf)[3]="marital"
str(smoking_subsetdf)
```

```
## 'data.frame':    1691 obs. of  7 variables:
##  $ sex        : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 1 2 2 2 1 ...
##  $ age        : int  38 42 40 40 39 37 53 44 40 41 ...
##  $ marital    : Factor w/ 5 levels "Divorced","Married",..: 1 4 2 2 2 2 2 4 4 2 ...
##  $ grossIncome: Factor w/ 10 levels "10,400 to 15,600",..: 3 9 5 1 3 2 7 1 3 6 ...
##  $ smoke      : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2 2 ...
##  $ amtWeekends: int  NA 12 NA NA NA NA 6 NA 8 15 ...
##  $ amtWeekdays: int  NA 12 NA NA NA NA 6 NA 8 12 ...
```

```
head(smoking_subsetdf)
```

```
##      sex age  marital      grossIncome smoke amtWeekends amtWeekdays
## 1   Male  38 Divorced   2,600 to 5,200    No          NA          NA
## 2 Female  42   Single      Under 2,600   Yes          12          12
```

```
## 3   Male  40  Married 28,600 to 36,400    No          NA          NA
## 4 Female  40  Married 10,400 to 15,600    No          NA          NA
## 5 Female  39  Married   2,600 to 5,200    No          NA          NA
## 6 Female  37  Married 15,600 to 20,800    No          NA          NA
```

```
tail(smoking_subsetdf)
```

```
##         sex age  marital      grossIncome smoke amtWeekends amtWeekdays
## 1686 Female  31   Single  5,200 to 10,400    No          NA          NA
## 1687   Male  22   Single   2,600 to 5,200    No          NA          NA
## 1688 Female  49 Divorced   2,600 to 5,200   Yes          20          20
## 1689   Male  45  Married  5,200 to 10,400    No          NA          NA
## 1690 Female  51  Married   2,600 to 5,200   Yes          20          20
## 1691   Male  31  Married 10,400 to 15,600    No          NA          NA
```

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

Each row represents the responses from an individual participant in the survey with respect to the attribute identified at the top of the respective column.

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

The study included **1691** participants.

(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* [vs. *nominal*].

sex: *categorical nominal* (not ordinal).
levels = Female, Male

age: *numerical* - although age **should** be considered *continuous* (e.g., allowing for fractional years), in the present dataframe it has *actually* been **stored** *discretely*, as integer (i.e., full-years only):

```
class(smoking_subsetdf$age)
```

```
## [1] "integer"
```

marital: *categorical nominal* (not ordinal), as it doesn't make sense to impose an ordering on the following:
levels = Divorced,Married,Separated,Single,Widowed

grossincome: *categorical* : (see the following discussion regarding *nominal* vs. *ordinal*):

```
summary(smoking_subsetdf$grossIncome)
```

```
## 10,400 to 15,600 15,600 to 20,800   2,600 to 5,200 20,800 to 28,600
##              268              188              257              155
## 28,600 to 36,400  5,200 to 10,400     Above 36,400          Refused
##               79              396               89              108
##      Under 2,600          Unknown
##              133               18
```

In theory, this variable *should* be **ordinal**, because the income bands *can* be ranked.

However, in this dataset, this column has *actually* been stored as a ***non-ordered*** factor:

Is smoking_subsetdf$grossIncome stored as a **factor**?

```
is.factor(smoking_subsetdf$grossIncome)
```

```
## [1] TRUE
```

```
str(smoking_subsetdf$grossIncome)
```

```
##  Factor w/ 10 levels "10,400 to 15,600",..: 3 9 5 1 3 2 7 1 3 6 ...
```

Is smoking_subsetdf$grossIncome stored as an ***ORDERED*** **factor**?

```
is.ordered(smoking_subsetdf$grossIncome)
```

```
## [1] FALSE
```

```
summary(smoking_subsetdf$grossIncome)
```

```
## 10,400 to 15,600 15,600 to 20,800   2,600 to 5,200 20,800 to 28,600
##              268              188              257              155
## 28,600 to 36,400  5,200 to 10,400     Above 36,400          Refused
##               79              396               89              108
##      Under 2,600          Unknown
##              133               18
```

The way in which the levels have *actually* been assigned fails to make use of the ordering, as the levels are stored **alphabetically**, disregarding the numerical semantics associated with each band:

```
for (i in 1:10) { cat(paste("grossIncome level", i,":",
                            levels(smoking_subsetdf$grossIncome)[i]), sep="\n")}
```

```
## grossIncome level 1 : 10,400 to 15,600
## grossIncome level 2 : 15,600 to 20,800
## grossIncome level 3 : 2,600 to 5,200
## grossIncome level 4 : 20,800 to 28,600
## grossIncome level 5 : 28,600 to 36,400
## grossIncome level 6 : 5,200 to 10,400
## grossIncome level 7 : Above 36,400
## grossIncome level 8 : Refused
## grossIncome level 9 : Under 2,600
## grossIncome level 10 : Unknown
```

This *can* be fixed by (manually) designating an *ordering* on each of the levels, e.g. :

```
ordering=c(9,3,6,1,2,4,5,7,8,10)
orderedlevels = levels(smoking_subsetdf$grossIncome)[ordering]
for (i in 1:10) { cat(paste("grossIncome ordered level", i,":",
                            orderedlevels[i]), sep="\n")}
```

```
## grossIncome ordered level 1 : Under 2,600
## grossIncome ordered level 2 : 2,600 to 5,200
## grossIncome ordered level 3 : 5,200 to 10,400
## grossIncome ordered level 4 : 10,400 to 15,600
## grossIncome ordered level 5 : 15,600 to 20,800
## grossIncome ordered level 6 : 20,800 to 28,600
## grossIncome ordered level 7 : 28,600 to 36,400
## grossIncome ordered level 8 : Above 36,400
```

```
## grossIncome ordered level 9 : Refused
## grossIncome ordered level 10 : Unknown
```

An ***ordered* factor** *can* thus be created (where I have arbitrarily placed "Refused" and "Unknown" at the end of the list):

```
orderedincome = ordered(smoking_subsetdf$grossIncome,levels=orderedlevels)
# Is this an ORDERED factor?
is.ordered(orderedincome)
```

```
## [1] TRUE
```

```
summary(orderedincome)
```

```
##      Under 2,600   2,600 to 5,200  5,200 to 10,400 10,400 to 15,600
##              133              257              396              268
## 15,600 to 20,800 20,800 to 28,600 28,600 to 36,400     Above 36,400
##              188              155               79               89
##          Refused          Unknown
##              108               18
```

So, **grossIncome** can thus be converted from a *nominal* factor to an *ordered* factor:

```
smoking_subsetdf$grossIncome = orderedincome
# Is grossIncome NOW stored as an ORDERED factor?
is.ordered(smoking_subsetdf$grossIncome)
```

```
## [1] TRUE
```

```
str(smoking_subsetdf$grossIncome)
```

```
##  Ord.factor w/ 10 levels "Under 2,600"<..: 2 1 7 4 2 5 8 4 2 3 ...
```

```
summary(smoking_subsetdf$grossIncome)
```

```
##      Under 2,600   2,600 to 5,200  5,200 to 10,400 10,400 to 15,600
##              133              257              396              268
## 15,600 to 20,800 20,800 to 28,600 28,600 to 36,400     Above 36,400
##              188              155               79               89
##          Refused          Unknown
##              108               18
```

Thus, although the grouping of income into bands fundamentally creates a **categorical ordinal** variable, the way in which it was stored in the dataframe did not make use of the ordinality, limiting its use to **nominal** unless changed to an **ordered factor** as shown.

smoke: *categorical nominal* (not ordinal) - the values are [Yes|No]

amtWeekends: *numerical discrete* (for those who smoke, the number is given as full cigarettes per day, stored as an integer)

amtWeekdays: *numerical discrete* (for those who smoke, the number is given as full cigarettes per day, stored as an integer)

Presumably one doesn't smoke a fractional portion of a cigarette (though, as a non-smoker, I can't be certain of this.) Thus the figures for amount smoked are not stored as a *continuous* variable.

Exercise 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.

The **population** of interest is **all children** between the ages of 5 and 15.
The **sample** in this study is a subset of the children who attended the *CUS ("Centro Universitario Sportivo") Summer Camp of Padua, Italy* in July 2008 and who were present on the two specific days on which the experiment was conducted.

(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.

The findings should **not** be generalized to the population because the specific nature of the children who participated in the experiment is unlikely to be representative of all children.
The day camp is comprised of children living in a specific area (Padua, Italy) whose families are presumably able to afford a weekly fee of approximately 100 euros to cover their attendance. (No indication is given as to whether fees are waived in the case of children from poorer families for whom the above fee would be a hardship.)
This could result in a biased sample, e.g., by excluding children from poorer families.
Among the children attending the camp, some did not participate in the experiment because their parents did not grant permission. This could further bias the results, as no indication is given as to the characteristics of those children who were excluded vs. those who participated.
While the study might be generalizable to children in *Northern* Italy of similar socioeconomic background to those who participated, further generalization would not be supportable absent repetition of the experiment across diverse locations. (Indeed, the study may not be generalizable across the whole of Italy, as the large difference in economic performance between wealthy North and the impoverished South is well-known.)

Because this is an experiment (rather than an observational study), it **is** appropriate to use the findings of the study to establish causal relationships.
Here, the research question is *"Does explicitly telling children not to cheat affect their likelihood to cheat?"*
The results of the experiment suggest that doing so does indeed reduce cheating, moreso in girls than in boys, with noticeable differences in other subgroups as well.

Exercise 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-aday 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.

As this was an observational study, rather than than an experiment, it is ***not*** appropriate to reach a causal conclusion.
Indeed, in the original journal article,
https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/226695/ioi05150_333_339.pdf the authors do not claim "causation', but rather they refer to an increased"association" of dementia with midlife smoking.

(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?

Again, this is an *observational* ***study***, not a *controlled, randomized* ***experiment***.
It displays an ***association*** between sleep disorders and behavioral problems/bullying, but this is not adequate to prove ***causation***.
Indeed, in the original journal article, https://www.sciencedirect.com/science/article/pii/S1389945711001572 the authors write: "Our study was cross-sectional and cannot prove causality"

They add, "Evidence is growing to support the idea that SDB ['sleep-deprived breathing'] may cause or contribute to disruptive behavior disorders."

The authors conclude:
"Our findings do not prove a cause-and-effect relationship but raise the possibility that addressing the underpinnings of childhood sleepiness may offer a largely untapped opportunity to reduce the common problem of aggressive behavior in schoolchildren."

Thus, the friend's statement is **not** justified.

Exercise 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?

This is a randomized, controlled experiment to examine the effects of exercise on mental health.

(b) What are the treatment and control groups in this study?

The **treatment** group consists of the half of the participants (within each age group) who are *advised to exercise* twice each week during the course of the study.
The **control** group consists of the other half of the participants, who are *advised to refrain from exercise* during the study.

(c) Does this study make use of blocking? If so, what is the blocking variable?

Yes, the experiment makes use of **age**-based blocking, as the participants are grouped by age before being placed into treatment vs. control groups. This ensures that an equal number of participants *within each age group* will be assigned to exercise vs. non-exercise groups.

(d) Does this study make use of blinding?

No, because the experiment requires the participants to exercise or not exercise, it cannot be conducted without the participants knowing to which group they have been assigned.

(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.

Because this is a randomized, controlled experiment, the results ***can*** be used to establish a causal relationship. As the experimental design indicates that the selection of the participants and assignment into groups should use "stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population," it **would** be appropriate to generalize the conclusions to the population at large (where such population consisted of individuals between the ages of 18 and 55.)

(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?

My first reservation would be, how can you ensure that those subjects who are randomly placed into the treatment group and are directed to exercise *actually do so*, and those placed into the control group, and thus directed to *not* exercise, do *not* do so.

Another reservation would be whether confounding effects bearing no relationship to exercise or not could impact the mental health of a substantial number of participants.
Such confounding effects may include:

- medications taken by a participant (and changes to such medications during the course of the study)
- major life event(s) affecting individual participants during the course of the study, e.g., illness of a family member, marriage/divorce, job change/loss, etc., which could impact the mental health assessment
- other factors that the study proposer hasn't considered

Additionally, there is no indication given as to the expected length of the study. The proposer doesn't state whether he/she wants to run this study for a month, a year, etc. (Presumably a longer study would require more funding.)

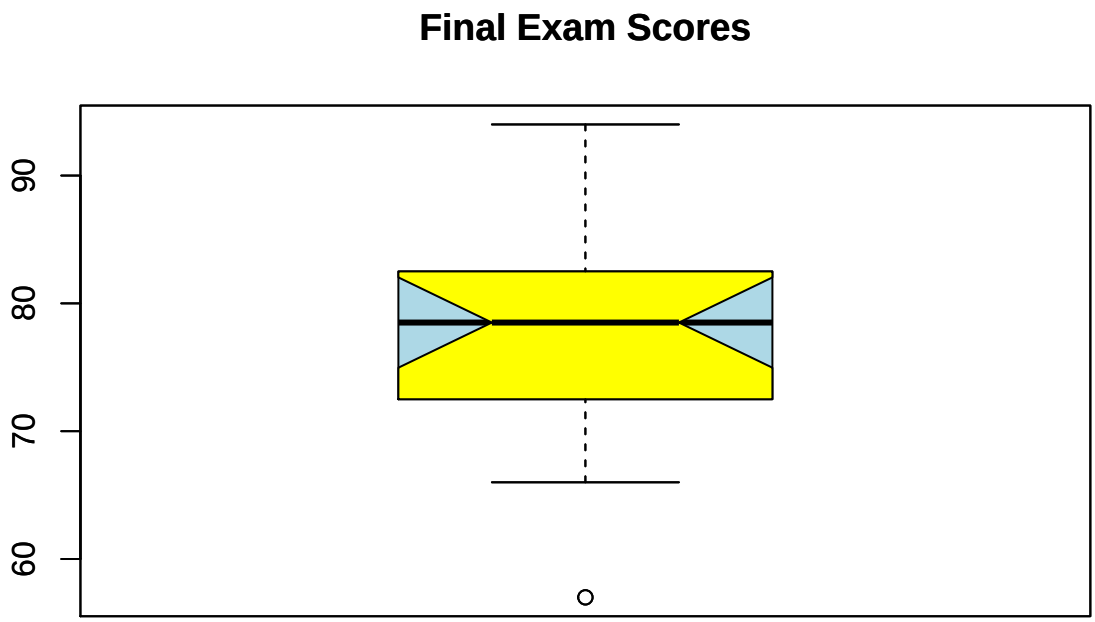
Exercise 1.48 - Stats scores.

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

```
scores <- c(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.

```
boxplot(scores, main=c("Final Exam Scores"), col="lightblue")
boxplot(scores, main=c("Final Exam Scores"), col="yellow" , notch = TRUE, add = TRUE)
```



The summary provided below may be useful.

```
summary(scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   72.75   78.50   77.70   82.25   94.00
```

Exercise 1.50 - Mix-and-match.

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

(a) This is a **normal** distribution, centered around x=60. The corresponding boxplot is **(2)**.

(b) This is a **uniform** distribution, extending from x=[0,100]. The corresponding boxplot is **(3)**.

(c) This is a right-skewed distribution, resembling a **lognormal** distribution, with median just above x=1. It exhibits right skew, so the mean is above the median. The corresponding boxplot is **(1)**.

Exercise 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.

This distribution exhibits a strong right skew, in that there are significant number of very expensive homes which "pull" the right tail of the distribution. As such, the mean value of a house would be greater (perhaps substantially) than the median value.

In this case, the best measure to represent a "typical" observation in the data would be the **median**, and the best measure of variability would be the **IQR**, as these two measures are robust to outliers.

(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.

This distribution appears to be closest to a "Normal" (Gaussian) distribution, where the mean is close to the median = 600,000. The IQR of (300k,900k) is symmetric around the median, and the absence of many expensive houses means that the values in the lower quartile (0,300k) are closely reflected in the upper quartile (900k,+) .

For such a symmetric distribution, the **mean** would be best represent a typical observation in the data (i.e., its measure of central tendency) and the **standard deviation** would best represent the variability of the observations.

(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.

This distribution would be extremely **right-skewed**, because many college students may not drink at all, resulting in a sizable result at zero, while (as is stated above, only a few drink "excessively."
In such case, the **median** would best represent a typical observation, and the **IQR** would be represent the variability.
[Note: there may be extreme variability between campuses, with heavier drinking expected at institutions known to be "party schools," and presumably less drinking at more studious (or, religiously-oriented) schools.]

(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

Again, this distribution would be extremely **right-skewed** with a few high-earners pulling the "right tail" of the distribution, while the rank-and-file workers would be earning considerably less.
In such case, the **median** would best represent a typical observation, and the **IQR** would be represent the variability.

Exercise 1.70 - Heart transplants.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:openintro':
##
##     diamonds
```

```
library(openintro)
data(heartTr)
table(heartTr$transplant, heartTr$survived)
```

```
##
##             alive dead
##   control       4   30
##   treatment    24   45
```

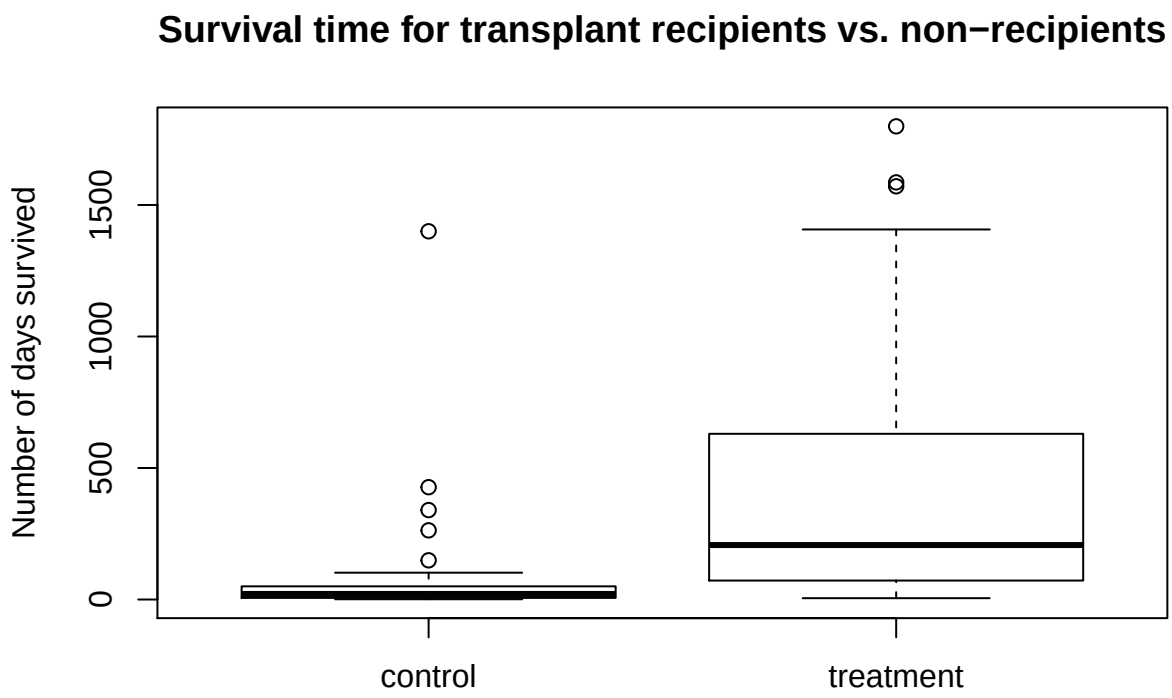
The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan.
Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a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

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

No, survival is **not** independent of whether or not a patient received a transplant.
Very few (4) members of the **control** group (who did **not** receive a transplant) were still alive at the conclusion of the study. A significantly larger number of members (24) of the **treatment** group (each of whom **did** receive a heart transplant) did remain alive as of the end of the study.

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

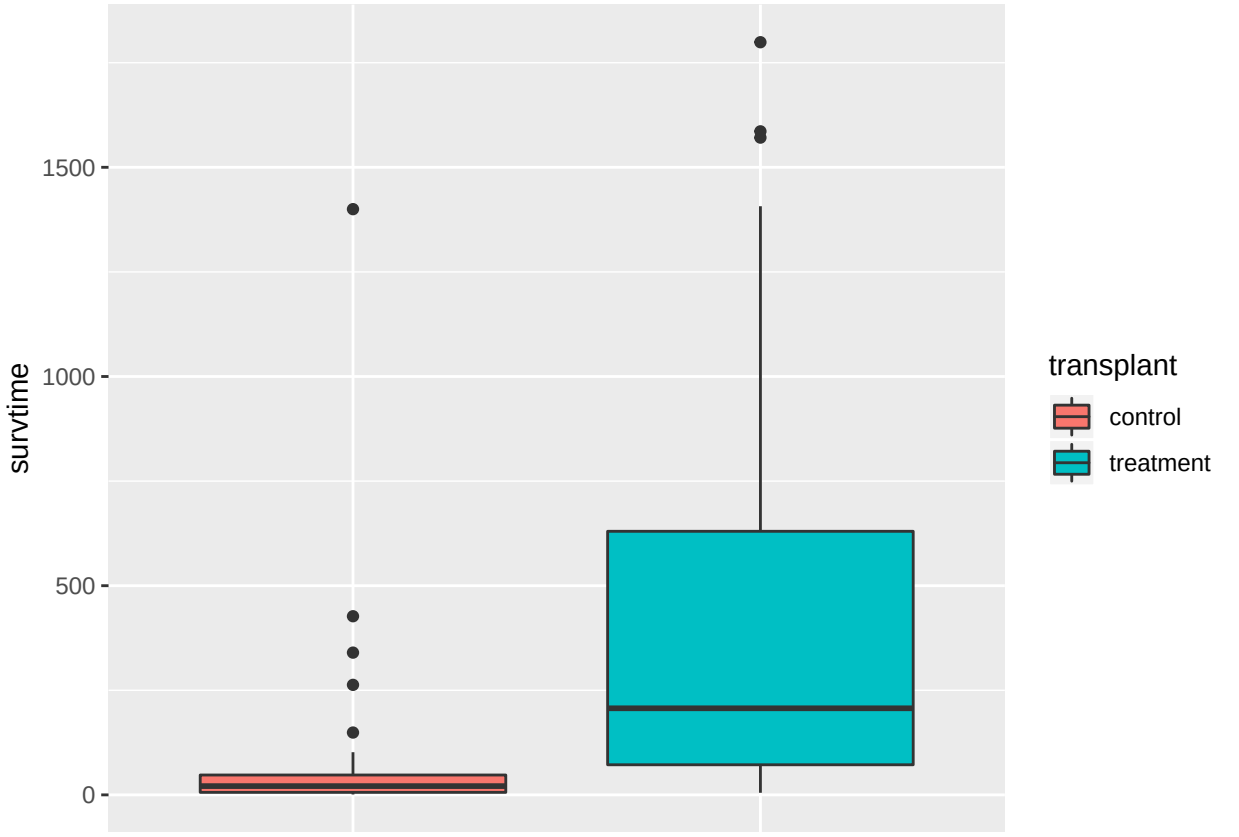
```
boxplot(survtime~transplant, data=heartTr, ylab="Number of days survived",
        main="Survival time for transplant recipients vs. non-recipients")
```



The boxplots show that members of the **treatment** group (i.e., those who **did** receive a transplant) lived much longer than members of the **control** group, who did **not**. This suggests that the heart transplant treatment is effective in extending survival.

Trying ggplot2, to compare the quality of the boxplot (same as above plot, but nicer-looking)

```
ggplot(heartTr, aes(x=transplant, y=survtime, fill=transplant)) +
  geom_boxplot() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



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

In the **treatment** group, 45/69 = **65.2** percent of the patients died, while in the **control** group, 30/34 = **88.2** percent died.
Thus, the treatment group exhibited an improvement of **23 percent**.

(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?

Under the **null hypothesis** H_0, whether a participant remained alive as of the end of the study is **independent** of whether he/she received a heart transplant (***treatment*** group) or not (***control*** group); the observed difference would only happen rarely.

Under the **alternative hypothesis** H_A, receipt of a heart transplant **does** have an effect on longevity, and the observed difference **was** actually due to the fact that ***a heart transplant significantly increases lifespan*** (for those who are ill enough to be in need of a transplant), which explains the large difference of -23% in the rate of survival to the end of the study.

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 ___***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 ___***zero***___.

Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ___***less than the (-23 percent difference) observed in the actual study***___ . 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?

```
#Recreate the histogram from the text
# confusion matrix - col and row sums
TotAlive = 28
TotDead =  75
TotTreatment = 69
TotControl = 34
# simulation results - by counting the number of dots in each column in the textbook historgram
SimTreatAlive = c(rep(24,2) ,
                  rep(23,1) ,
                  rep(22,8) ,
                  rep(21,15),
                  rep(20,18),
                  rep(19,19),
                  rep(18,12),
                  rep(17,11),
                  rep(16,7) ,
                  rep(15,4) ,
                  rep(14,3))
#SimTreatAlive
SimTreatDead = TotTreatment - SimTreatAlive
#SimTreatDead
PercentTreatAlive = SimTreatAlive / TotTreatment
#PercentTreatAlive
PercentTreatDead  = SimTreatDead  / TotTreatment
#PercentTreatDead

SimControlAlive = TotAlive - SimTreatAlive
#SimControlAlive
SimControlDead = TotControl - SimControlAlive
#SimControlDead

PercentControlAlive = SimControlAlive / TotControl
#PercentControlAlive
PercentControlDead  = SimControlDead  / TotControl
#PercentControlDead
```

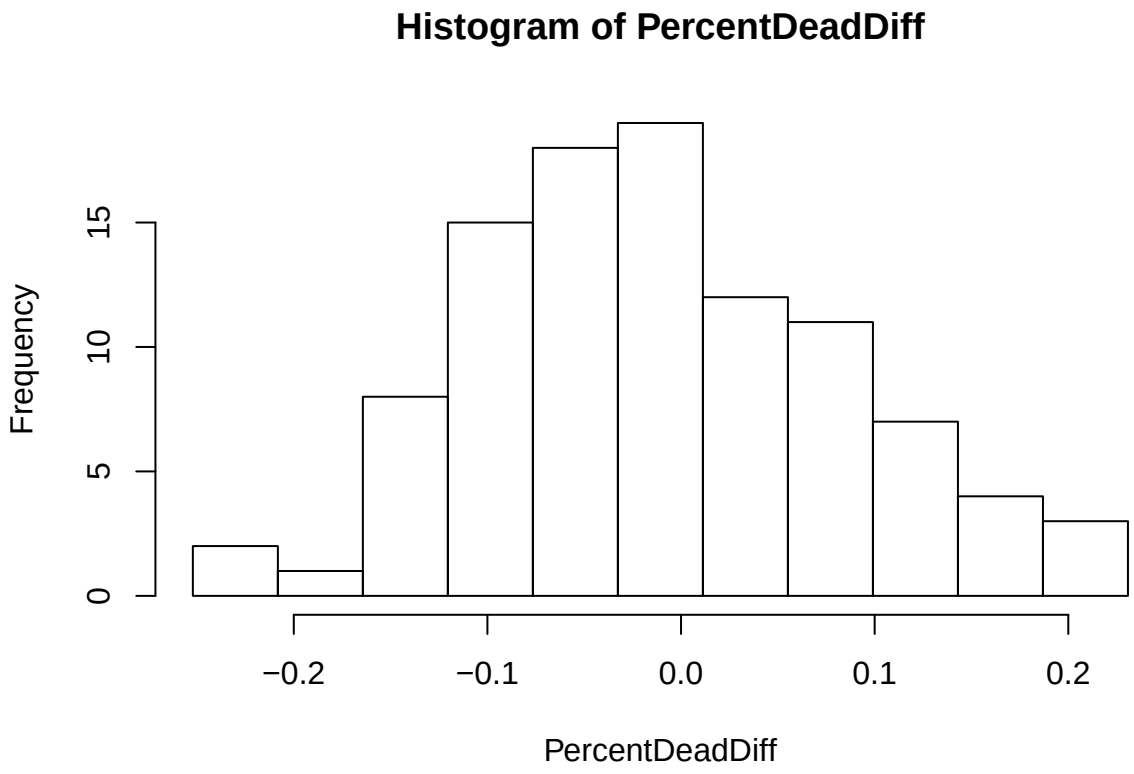
```r
# compute the difference in percent dead, between Treatment and Control, from simulation results
PercentDeadDiff = PercentTreatDead - PercentControlDead
# get the unique values for the percentages
uniq = unique(PercentDeadDiff)
#uniq

# find out how far apart the above values are, for positioning of breaks
dif = diff(uniq)

# put the breaks halfway between each value
breaks = uniq - dif[1]/2
#breaks

# append one more break on the far right
breaks[length(breaks)+1]=breaks[length(breaks)]+dif[length(dif)]
#breaks

# draw the histogram
hist(x = PercentDeadDiff, breaks = breaks)
```



First, the stacking of results in vertical bars above specific points reflects the fact that because of the specific number of participants involved (n=103), with the above counts of alive vs. dead and treatment vs. control, it is not possible to obtain exactly zero as the difference in the proportions.
The peak value on the chart corresponds to a simulated result with 19 members of the treatment group alive; the x-value below this bar is

$$\frac{50}{69} - \frac{25}{34} = .7246 - .7352 = -.0107$$

There are 19 dots in this column, which indicates that 19 of 100 simulations returned a result with 19 members of the **treatment** group alive and 50 dead, while 9 members of the **control** group survived vs. 25 dead.

Each column to the **left** represents the simulations where the number of **survivors** in the **treatment** group **increased** by 1, while the number of **survivors** in the **control** group was **reduced** by one.

Thus, the second-tallest column, containing 18 dots, represents the **18** simulations in which **20** members of the **treatment** group lived, as did **8** members of the **control** group. The difference in proportions (on the x-axis) can be computed as

$$\frac{49}{69} - \frac{26}{34} = .7101 - .7647 = -.0546$$

Continuing to the left, each bar plot represents the number of simulations where the number of survivors in the treatment group is 21, then 22, then 23, and finally at the leftmost point on the graph, **24**.

This leftmost column reflects the number of simulations in which the count of survivors and deaths in each of the treatment and control groups ***happens to match the results from the actual study***, where **24** members of the **treatment** group survive, but only 4 members of the control group do.

This corresponds to the difference in proportions computed as

$$\frac{45}{69} - \frac{30}{34} = .6522 - .8824 = -.2302$$

So, out of 100 simulations where we distrubute **28** *alive* cards and **75** *dead* cards across **69** *treatment* slots and **34** *control* slots, we have only ***2*** simulations which correspond to the values from the actual study (and none which exceed it.)

Accordingly, the "p-value" of this simulation is $0.02 < 0.05$, which provides evidence at 95% confidence to **reject** the null hypothesis H_0 in favor of the alternative H_A, which states that

the receipt of a heart transplant **does** have a significant effect on longevity, and the observed difference of ***-23 percent*** **was** actually due to the fact that ***a heart transplant significantly increases lifespan*** (for those who are ill enough to be in need of a transplant).