DATA 606 - Assignment 01

Kavya Beheraj

February 11, 2018

OpenIntro Statistics: Chapter 1

Introduction to Data

Questions: 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70

```
install.packages(c('openintro','OIdata','devtools','ggplot2','psych','reshape2','knitr','markdown','shi
```

```
## Installing packages into 'C:/Users/Kavya/Documents/R/win-library/3.4'
## (as 'lib' is unspecified)
```

```
devtools::install_github("jbryer/DATA606")
```

```
## Skipping install of 'DATA606' from a github remote, the SHA1 (ab793d5d) has not changed since last i
##   Use `force = TRUE` to force installation
```

```
library(openintro)
```

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

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

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

1.8 | Smoking habits of UK residents

(a) Each row of the data matrix represents one observation – one participant of the study.

(b) 1,691 participants were included in the study.

(c)

- "sex" is categorical
- "age" is numerical (discrete)
- "marital" is categorical
- "grossIncome" is categorical (ordinal)
- "smoke" is categorical
- "amtWeekends" is numerical (discrete)
- "amtWeekdays" is numerical (discrete)

1.10 | Cheaters, scope of inference

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

(b) It would be hard to generalize results to the population because the description does not describe the sampling method that the researchers used and how successful they were in implementing the method. It would also be hard to establish causal relationships because we don't know how much the outcome could be due to random chance, and how much is influenced by the experimental setup.

1.28 | Reading the paper

(a) Based on the data, we can only conclude that health plan members between the ages of 50-60 who voluntarily admitted to smoking had a greater incidence of dementia later in life. The data is constrained by being self-reported (so people may under-report their smoking) and only focusing on people who were already at an older age. To conclude that smoking causes dementia, we would need to have data on these participants at different stages of life and compare long-term outcomes.

(b) The statement is not justified, since the data appears to be observational (not experimental), so we can't draw causal conclusions. We can conclude that students who were observed to have behavioral issues or observed to be bullies were also observed to have symptoms of sleep disorders.

1.36 | Exercise and mental health

(a) This study is a randomized experiment.

(b) The treatment is exercise twice a week, and the control is no exercise.

(c) Yes, this study makes use of blocking. The blocking variable is age.

(d) No, it does not appear that the study makes use of blinding. Both the participants and the researcher know whether the participant will get the treatment.

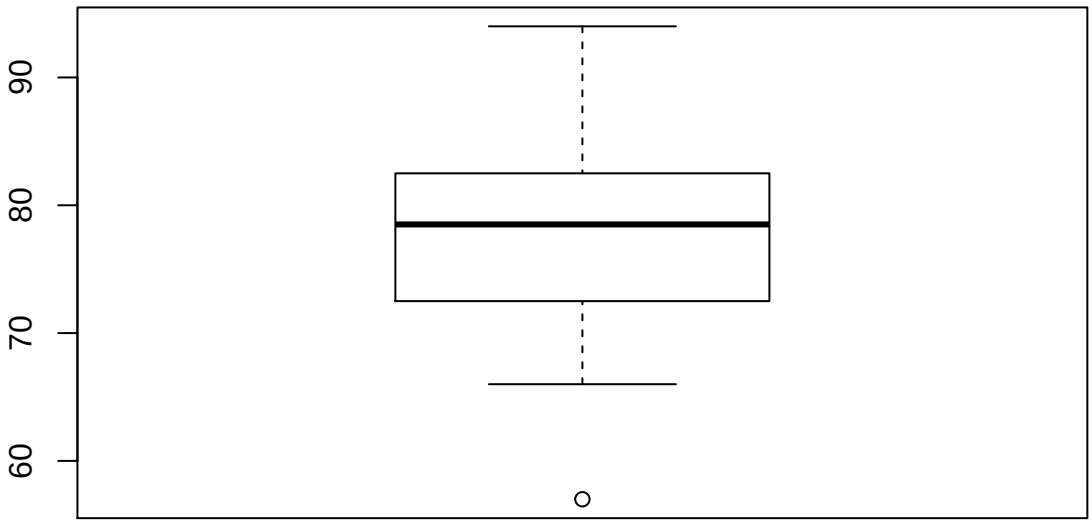
(e) Since the study is a randomized experiment, we can make a causal statement and generalize to the population. However, the lack of a blind or double-blind design might affect the outcome. We also don't know the size of the population, which could affect the study's replication.

(f) I would have reservations about the experiment not including a blind or double-blind design, and whether the researchers are controlling for other variables like the existence of a pre-existing mental health condition.

1.48 | Stats scores

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

boxplot(statsscores)
```



1.50 | Mix-and-match

(a) The distribution is normal and unimodal with a mean of about 60. This matches with boxplot **(2)**.

(b) The distribution is symmetric and multimodal with a mean of about 50. This matches with boxplot **(3)**.

(c) The distribution is right-skewed and unimodal with a mean of about 1. This matches with boxplot **(1)**.

1.56 | Distributions and appropriate statistics, Part II

(a) I expect the distribution to be right-skewed. The median would best represent the data because the majority of observations fall to left of the mean. The variability would be best represented by the IQR because it would account for outliers, like the houses that cost more than $6m, while representing the bulk of the data.

(b) I expect the distribution to be symmetric because it seems like the same number of houses are in each quartile. Both the mean and median would work to represent the data, since they will likely be the same in a normal distribution. Similarly, the standard deviation or IQR could both work to represent variability.

(c) I expect the distribution to be right-skewed – fewer students have a higher number of drinks in a week, and the majority have about 0. Since the distribution is skewed, the median and IQR better represent the data.

(d) I expect the distribution to be close to a normal distribution. Since the number of executives earning a very high salary is small, those executives are outliers. However, since we know there are outliers, we should

use the median and IQR to represent the data.

1.70 | Heart transplants

(a) No, it does not appear that survival is independent of transplants, since significantly more people with transplants survived than the control group, who got no transplant.

(b) The box plots suggest that the treatment was efficacious, since it appears to have increased the average survival time of patients compared to the control.

(c) About 88% of patients in the control group and 65% of patients in the treatment group died.

```
30/34
```

```
## [1] 0.8823529
```

```
45/69
```

```
## [1] 0.6521739
```

(d)

- **(i)** Our null hypothesis is that a heart transplant has no effect on a gravely ill patient's survival rate. Our alternate hypothesis is that a heart transplant increases a gravely ill patient's survival rate.
- **(ii)** 24; 75; 34; 69; 0; different from the actual outcome
- **(iii)** In our actual outcome, the difference in proportions between the treatment and control (0.65 - 0.88) was -0.23. However, in our simulation, the majority of outcomes showed little to no difference in proportions, and very few were near -0.23. Based on this, we need to accept our null hypothesis that a heart transplant has no effect of survival.
