# IS606 Chapter 1 Homework

*Daniel Dittenhafer*

*September 6, 2015*

**1.8 Smoking habits of UK residents (p57)**

**(a) What does each row of the data matrix represent?** Each row of the data matrix represents a case in the sample.

**(b) How many participants were included in the survey?** Assuming row 1691 is the last row of the data, then 1691 participants were included in the survey.

**(c) Indicate whether each variable in the study is numerical or categorical.**

- sex: Categorical
- age: Numerical, Discrete
- marital: Categorical
- grossIncome: Categorical, Ordinal
- smoke: Categorical
- amtWeekends: Numerical, Discrete
- amtWeekdays: Numerical, Discrete

**1.10 Cheaters, scope of inference (p58)**

**(a) Identify the population of interest and the sample in this study.** The population is children between ages 5 and 15. The sample was 160 children within the age range.

**(b) Comment on whether or not the results of the study can be generalized to the population, and whether the findings of the study can be used to establish a causal relationship.** The sample size is small compared to the assumed population (all children 5-15). Also, it isn't clear how the sample was selected. As such, I am skeptical that the results could or should be generalted to the population.

The study is described as an experiment, with the explanatory variable being the instruction to cheat (or lack of instruction), and response variable being the cheating outcome. Based on the information in the 1.5 exercise write-up, a causal relationship based on the treatment alone is not supported. Gender, age or possibly maturity could be confounding variables in this study.

**1.28 Reading the paper (p62)**

**(a) Risks: Smokers Found More Prone to Dementia** Based on the study, I take away the conclusion that smoking (implied to be tobacco) is correlated to some degree with dementia. This may be related to a chemical response from nicotine, but may not be. It could be driven instead by another chemical in the smoke, or some other factor common to smokers. The description of the study suggests an observational study structure which cannot show a causal connection.

**(b) The School Bully Is Sleepy** The statement "The study shows that sleep disorders lead to bullying in school children" is not justified. The conclusion I draw is that there is a correlation between sleep disorders and behavioral issues. There may be another factor which is driving both the sleep disorder and the behavioral issue.

### 1.36 Exercise and mental health (p64)

**(a) What type of study is this?** This is a randomized experiment.

**(b) What are the treatment and control groups in this study?** The treatment group is the the group instructed to exercise twice a week. The control group is the group instructed not to exercise.

**(c) Does this study make use of blocking? If so, what is the blocking variable?** Yes, this study makes use of blocking in respect to the age of the participants. (p24)

**(d) Does this study make use of blinding?** No, this study does not make use of blinding. The participants know if they are exercising or not.

**(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health?** Given the exeriment design, a causal relationship conclusion is possible (though not as strong as if blinding was used some how). The study proposal does not indicate how big the sample size will be, therefore I am unable to conclude that the study could 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 proposed sample size should be included in the study proposal. I would expect a funding request to have factored in the cost of obtaining the sample data including the size of the sample.
- The length of and intensity of the exercise should be better defined, otherwise the treatment group may have widely different exercise treatments.
- Twice a week exercise might not be enough to elicit a response. This might be wasted money... a more frequent exercise regime might produce a stronger difference.
- The lack of blinding makes the causal relationship if any much weaker. Those subjects in the control group might think differently based on their knowledge of the group membership.

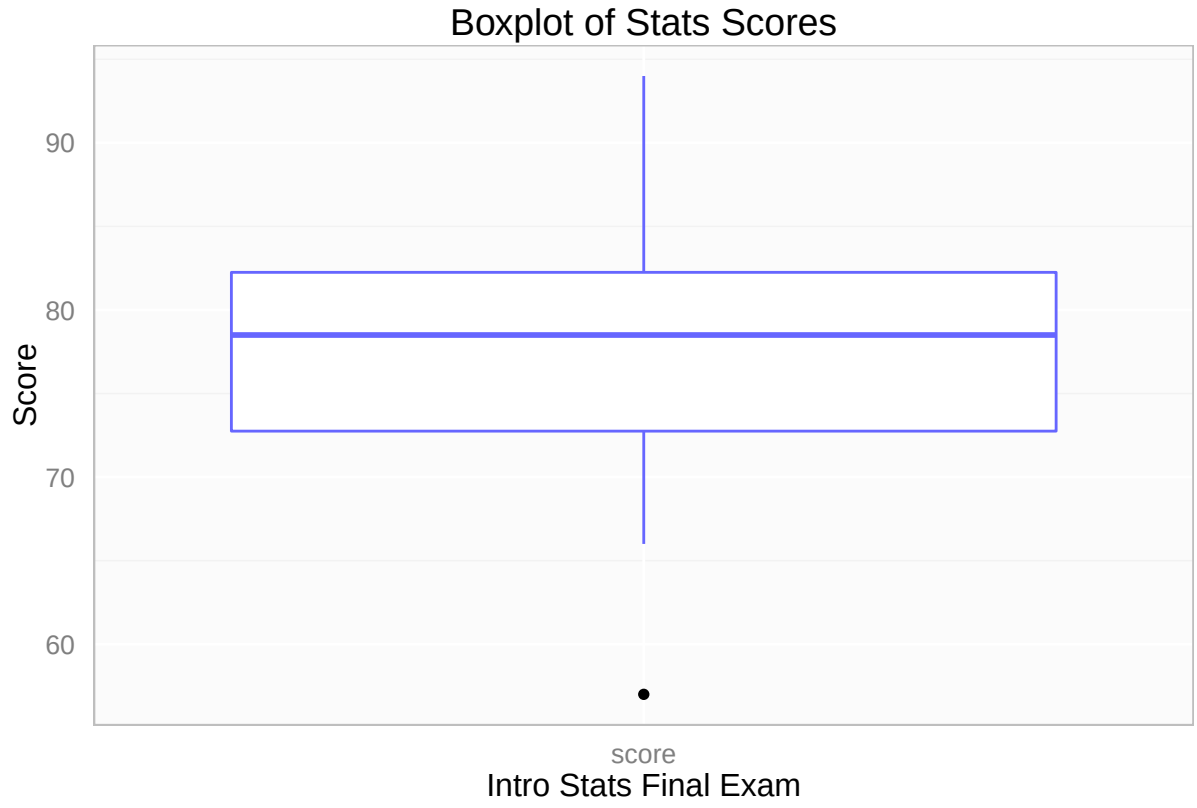
The points listed above, if addressed by the researcher in a better way, might reduce reservations about the study, but the lack of blinding is a significant aspect for which I don't see a resolution.

### 1.48 Stats scores (p66)

```
# Include ggplot2
library(ggplot2)
myTheme <- theme(axis.ticks=element_blank(),
                 panel.border = element_rect(color="gray", fill=NA),
                 panel.background=element_rect(fill="#FBFBFB"),
                 panel.grid.major.y=element_line(color="white", size=0.5),
                 panel.grid.major.x=element_line(color="white", size=0.5))
```

```r
# define a vector for the scores
data <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
scores <- data.frame(scores=data,
                     type=rep("score", length(data)))
# Define the boxplot
bp <- ggplot(data=scores, aes(x=type, y=scores)) +
  myTheme +
  geom_boxplot(colour="#6666FF") +
  labs(title="Boxplot of Stats Scores", x="Intro Stats Final Exam", y="Score")

bp
```




### 1.50 Mix-and-match (p67)

**(a)** The distribution is mostly symmetric, unimodal and matches the (2) boxplot.

**(b)** The distribution is fairly evenly distributed, and (depending on scale) could be multimodal. It matches the (3) boxplot.

**(c)** The distribution is right skewed, unimodel and matches the (1) boxplot.

### 1.56 Distributions and appropriate statistics (p69)

**(a)** Q1 - 25% - Below $350,000; Q2 - 50% - Below $450,000 (Median); Q3 - 75% - Below $1,000,000; Meaningful # of houses that cost more than $6,000,000.

- Distribution: Right skewed due to the "meaningful" number of houses that cost significantly more than $1,000,000.
- Median would best represent the typical observation due to the fact that most houses are in the $350K - $1M range. I would expect the mean to be higher than then median, and potentially higher than the third quartile depending on the cost of the houses in the upper quartile.
- As a result of potentially shifted mean, the variance and standard deviation could be inflated as well. As such, the IQR would be a better representation of the variability of the observations.

**(b)** Q1 - 25% - Below $300,000; Q2 - 50% - Below $600,000 (Median); Q3 - 75% - Below $900,000; Very few houses that cost more than $1,200,000.

- Distribution: Fairly symmetric due to the reasonably similar portion of houses in each of the quartiles, though slightly right skewed due to the few houses costing more than $1.2M.
- Median seems to still be the best representation of the typical observation, though in this data mean would be close due to the distribution of the data.
- Again, IQR seems the better represenation of variability, though standard deviation would make a close estimate of variability due to the distribution of the data.

**(c) Number of alcoholic drinks** Assuming # of drinks/week as follows:

Bottom 75% don't drink or drink minimally (0-2 drinks/week) because 18 - 20 year old is under legal age in USA. Upper 25% drink (3-6 drinks/week) Only a few excessively ($> 6$ drinks/week).

- Distribution: Left skewed due to the low # of drinks in much of data (long left tail).
- Median would be best representation of typical # of drinks for typical student since most students are below drinking age and would be in the lower 2 quariles.
- Again, IQR seems the better represenation of variability because it wouldn't be influenced by the few excessive drinkers.

**(d) Annual salaries**

- Distribution: Left skewed due to the long left tail of lower salaries and only a few very high salaries on the right.
- Median would best represent the typical salary. Mean would be influenced by the very high salaries and would show a non-typical higher average salary.
- Again, IQR is best for variability. Standard deviation would be influenced by the few high salaries.

**1.70 Heart transplants (p74)**

**(a) Based on mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.** The mosaic plot shows more survival in the treatment group relative to the control group. Also because of the wider size of the treatment group, it appears there were more subjects in this group. The mosaic alone is not enough to claim survival is independent or dependent, but it appears that survival rate improves with, and is not purely independent of, the transplant treatment.

**(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment?** The boxplots show the control group having a consistently short survival time (much less than 250 days) with only some outliers with extended survival times, while the treatment group has a much larger quartile range with a median close to 250 days and up above 500 days for the third quartile.

Considering this interpretation of the boxplots, the transplant treatment appear to be effective in extending the survival time of the subjects significantly.

**(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?** The following R code shows loading the Heart Transplant data and computing the proportions for each group:

```
library(dplyr, quietly=TRUE, warn.conflicts=FALSE)
# Load the data from our copy downloaded from course site.
heart <- read.table("../data/heartTr.csv", sep=",", stringsAsFactors=FALSE, header=TRUE)
# Using dplyr to group and count.
aggHeart <- tally(group_by(heart, transplant, survived))

# Compute totals and died for each group
totalControl <- sum(aggHeart[aggHeart$transplant == "control",]$n)
diedControl <- sum(aggHeart[aggHeart$transplant == "control" &
                            aggHeart$survived == "dead",]$n)

totalTreatment <- sum(aggHeart[aggHeart$transplant == "treatment",]$n)
diedTreatment <- sum(aggHeart[aggHeart$transplant == "treatment" &
                              aggHeart$survived == "dead",]$n)

# Control porportion died
ratioControlDied <- diedControl / totalControl

# Treatment porportion died
ratioTreatmentDied <- diedTreatment / totalTreatment
```

The proportion of the treatment group which died is 0.6521739, and the proportion of the control group which died is 0.8823529.

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

(i) The claim being tested is that a heart transplant increased lifespan.

(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 betwen 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**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **-0.230179 or less**. 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) Given that there are only 2 of the 100 simulated differences at approximately -0.23, this suggests it would be unlikely to have observed the treatment survival outsomes by chance and therefore the transplant program is effective.