DATA 607 Week 1 graded problems

James Topor

January 30, 2016

Chapter 1, problems 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70

1.8

a) Each row represents a case

b) There were 1691 participants

c)

Sex: Categorical, not ordinal

Age: Numerical, discrete

Marital: Categorical, not ordinal

Gross Income: Categorical, ordinal

Smoke: Categorical, not ordinal

amtWeekends: Numerical, discrete

amtWeekdays: Numerical, discrete

1.10

a) The population is children between the ages of 5 and 15. The sample size is 160.

b) The results of the study cannot be generalized to the population since the children do not appear to have been randomly selected or randomly assigned to the “No Cheating” / “No Instruction” groups. Similarly, no causal relationship can be established since the children were neither randomly selected nor randomly assigned to the “No Cheating” / “No Instruction” groups. The only result we can derive from this study is a correlation statement for the sample.

1.28

a) We cannot conclude that smoking causes dementia later in life based on this study since it was an observational study. Observational studies cannot derive causal relationships; they can only be used to established an association between explanatory and response variables.

b) The statement “Sleep disorders lead to bullying in schoolchildren” is not justified by this study. The study relied on anecdotal data collected via surveys. As such, the data collected by the researchers may not be representative of the broader population and therefore cannot be used to derive causal relationships. At best, any conclusion that might be drawn from this study would be limited to a correlation statement for the sample.

1.36

a) What type of study is this?

This is an experiment that uses blocking to separate potential participants by age group. Then, participants are chosen randomly from each of the age group blocks. Finally, half of the selected participants are randomly assigned to the treatment and control groups, respectively.

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

The treatment group is instructed to exercise twice per week. The control group is instructed to do no exercise each week.

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

Yes the study makes use of blocking. The blocking variable is a person’s age.

d) Does this study make use of blinding?

No, the study does not make use of blinding since the participants need to be instructed as to whether they should exercise twice per week 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, and indicate whether or not the conclusions can be generalized to the population at large.

The study can be used to establish a causal relationship between exercise and mental health since it is an experiment that uses both random sampling and random assignment. Furthermore, since both random sampling of participants from the broader population AND random assignment of participants to the treatment and control groups was utilized, 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?

I would have reservations about funding the study because I personally believe that many other confounding variables might influence an individual’s mental health. Without knowing how the researchers plan to address the multitude of other potential variables that might influence their results (e.g, diet, genetic factors, education, etc..) I would be unlikely to fund the study.

1.48

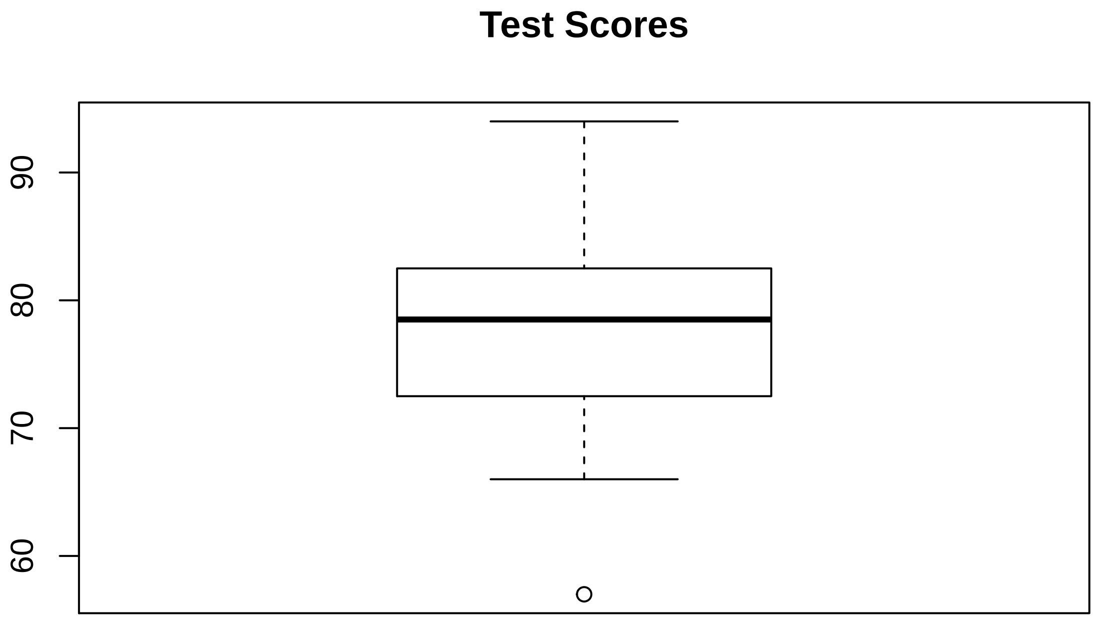
Create a box plot:

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

summary(vec)
```

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

```r
boxplot(vec, main = 'Test Scores')
```



1.50

a) Symmetric; corresponds to boxplot # 2
b) Multimodal; corresponds to boxplot # 3
c) Right skewed; corresponds to boxplot # 1

1.56

a) The distribution is right skewed since the median is $450K and there are many homes that cost more than $6 million while the IQR is $1,000,000 - $350K = $650K. As such, the median is a better representation of a typical observation than is the mean. Similarly, due to the skew the IQR will be a better representation of the variability of the data than will the standard deviation.

b) This distribution is symmetric since the IQR is spread proportionally on either side of the mean and there are few houses that cost more than $1 million. Since the distribution is symmetric, the mean and standard deviation are better representations of the data than are the median and IQR.

c) This distribution is right skewed since it is zero-bounded and a large number of samples will have a value of zero since the legal drinking age is greater than the age of most college students. As such, the median and IQR are better representations of the data than are the mean and standard deviation.

d) Distributions of salaries tend to be right-skewed since many lower level employees will earn similar salaries while a smaller number of senior employees and executives will earn higher salaries. As such, the median and IQR would be better representatives of the data than would the mean and standard deviation.

1.70

a) Based on the mosaic alone I do not believe it is possible to determine whether or not survival is independent of whether or not the patient received a transplant. The mosaic alone does not provide any insight into the size of the treatment and control groups: we'd need to look elsewhere for that information. Without any sense of the size of these groups we can't make any judgment as to whether the proportionality described by the mosaic for each group is indicative of a correlation or causal relationship.

b) The boxplots indicate that there is likely a meaningful relationship between receiving a transplant and survival time. As such, there appears to be some degree of efficacy in performing a transplant since the patient's survival time is measurably greater than that of patients who did not receive a transplant.

c) Proportion of Control group deaths

```
30/34
```

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

Proportion of Treatment group deaths

```
45/69
```

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

d)

(i) **H0** = an experimental heart transplant program has NO EFFECT on increased lifespan. In other words, lifespan is independent of whether the patient received a transplant.
HA = the experimental heart transplant program increased the lifespan of patients. In other words, lifespan **IS NOT** independent of whether the patient received a transplant.

(ii)

- Write ALIVE on **28** cards
- Write DEAD on **75** cards

- 69 cards represent treatment
- 34 represent control
- Distribution centered at ZERO
- Calculate the fraction of simulations where the simulated differences in proportions are **at least 23%** (23% being the observed difference in proportions from the actual study).

(iii) The simulated results show that it is unlikely that a difference in proportions of 23% would have occurred if the variables were independent. As such, we should reject the null hypothesis **H0** in favor of **HA**.
Therefore, the data provide convincing evidence for the alternative hypothesis of the experimental heart transplant program increasing the lifespan of patients. The observed difference between the two proportions was due to a real effect of the experimental transplant treatment.