Andrew Goldberg

9/6/15

IS606 Assignment 1

* Graded: 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70

**1.8**

a. Each row of the data matrix is a ‘case’, and in this case, they are UK residents who responded to a survey.

b. n=1691

c.

sex = categorical, nominal

age = numerical, discrete

married = categorical, nominal

grossIncome = categorical, ordinal

smoke = categorical, nominal

amtWeekends = categorical, ordinal

amtWeekdays = categorical, ordinal

**1.10**

a. children between the ages of 5 and 15, n = 160

b. There’s quite a bit of mental development between the ages of 5 and 15, so I would expect some differences in behavior, but the results should be able to be generalized to the population of 5 to 15 year olds. Not the full population, including adults, however. Since the study is experimental, the findings can be used to establish causal relationships.

**1.28**

a.

Since the survey is observational, you cannot conclude a causal relationship. I suppose, theoretically, if scientists could adjust for ALL factors, it could be causal, but they are pretty limited in the factors they can actually adjust for. That said, due the size of the sample and the amount of factors can adjust for, I would feel comfortable thinking about it as causal until proven otherwise.

b.

Since it’s observational, you cannot assume a causal link between sleep habits and behavioral problems and bullying. I would describe the conclusion of the study as symptoms of sleep disorders are a potential factor leading to child behavioral problems and bullying.

Further, I’d be very skeptical of the research overall, as parents and teachers have a motivation to blame behavioral problems on sleep disorders, instead of questioning ability as caregivers.

**1.36**

a. (stratified) randomized experiment

b. treatment and control groups are the same; half of each of the 18-30, 31-40, and 41-55 year old age groups

c. Blocking is used by splitting up the groups into the three age groups (or blocks)

d. no blinding

e. Since the study is experimental, it can be used to establish causal relationships. The conclusions can be generalized to similar populations at large. IE 18-30 year sample to the population of 18-30 year olds.

f. It doesn’t mentioning if the researchers are blinded or the sample size, and there is no spelled out way to ensure that the participants actually exercised or rested

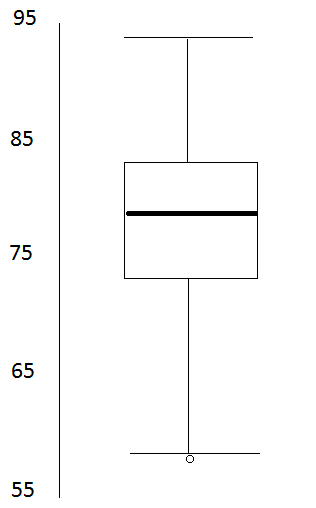
**1.48**

Min = 57, q1 = 72.5, Median = 78.5, q3 = 82.5, Max = 94

IQR = 82.5 – 72.5 = 10, so max whiskers = 15

72.5 – 15 = 57.5, so min whisker doesn’t reach min data point (57)

82.5 + 15 = 97.5



**1.50**

a. symmetric and unimodal, 2

b. I would argue its symmetric, and maybe multimodal, 3

c. right skewed and unimodal, 1

**1.56**

a. There is a rightward skew as the housing gets very very expensive. Therefore the center would be best described by the median and iqr.

b. The distribution should be closer to symmetric, since there’s roughly equal (25%) of housing between each price range. Therefore, the center would best be described by the mean and standard deviation.

c. The distribution should have a rightward skew, since most don’t drink and fewer and fewer students drink at higher levels. Therefore, the center would best be described by the median and iqr.

d. This distribution should have a rightward skew, as it would be less and less frequent to have very highly paid employees. Therefore, the center would best be described by the median and iqr.

**1.70**

a. the vertical lines separating alive from dead between the treatment and control are not very close to being even, which suggests the likelihood of surviving is linked to treatment, and the variables are dependent.

b. In general, the treatment is more effective than the control, but roughly half of the treatment group enjoys a much longer survival time, while the other half survive roughly less than a year.

c. It looks like approximately 60% in those in the treatment group died, and 85% of the control group died.

d.

i. Claims being tested in the simulation are if the null hypothesis is true, or if the survived and treatment variables are independent.

ii.

all,

all,

proportion of patients who got treatment,

proportion of patients in control group,

0,

At least 25% (% survived in treatment - % survived in control)

iii. It appears that none of the results of the simulation reached above 25% (and only a few around 20% if I misread the mosaic plot), suggesting that we can reject the independent model in favor of the alternative model, that there is strong evidence that survival rates are dependent on treatment