Homework #5, Mei Maddox.

Mei Maddox

Submit the solution on Canvas into the corresponding assignment (e.g. "Homework #1") in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide only code and output. NO NEED TO COPY THE PROBLEM FORMULATION (!)

Problem 1

1) a) The explanatory variable is marijuana use. The response variable is number of traffic fatalities. This study is an observational study. A potential lurking variable is proportion of sample which uses marijuana.

   b) An experimental study on "marijuana -> accidents" link cannot be ethically conducted.

   - define target population: all US drivers who get in accidents
   - sampling frame: List of registered drivers from the DMV. This raises issues regarding confidentiality. If only a few DMV registries are used, then there are issues with undercoverage. Even if we somehow obtained a representative sample, there would be ethical issues of forcing the marijuana treatment on some (especially if we aren't accounting for any health issues or legallity in their respective state) and then knowingly putting the subjects at risk by having them drive.

2) The jury is randomly selected, but then there is a self-selection process where the randomly selected jurly can choose to participate or not. This can lead to nonresponse bias as individuals who might have more conservative views would be more likely to participate.

3)

- a) The beta testing participants where from a convenience sample, only friends of the engineer. The convenience sampling leads to undercoverage as the sample population was not representative of the target population. There is also non-response bias as the focus group are volunteers. Furthermore, there may be response bias as the beta sample participants may be less inclined to provide negative feedback as they are his friends (i.e. the only non-techy person merely said it felt "engineered").

- b) Having the CEO come in to describe the "new" product they are going to be testing out violates the blind experiment as the participants now know which product they are being given. The moderator is less enthused and makes the statement about corrupt data because by generating excitement (and understanding) of the new product, the CEO has influenced the results of the study because they are no longer unbiased. Therefore, the data must be tossed.

Problem 2 (make sure to include code with outputs)

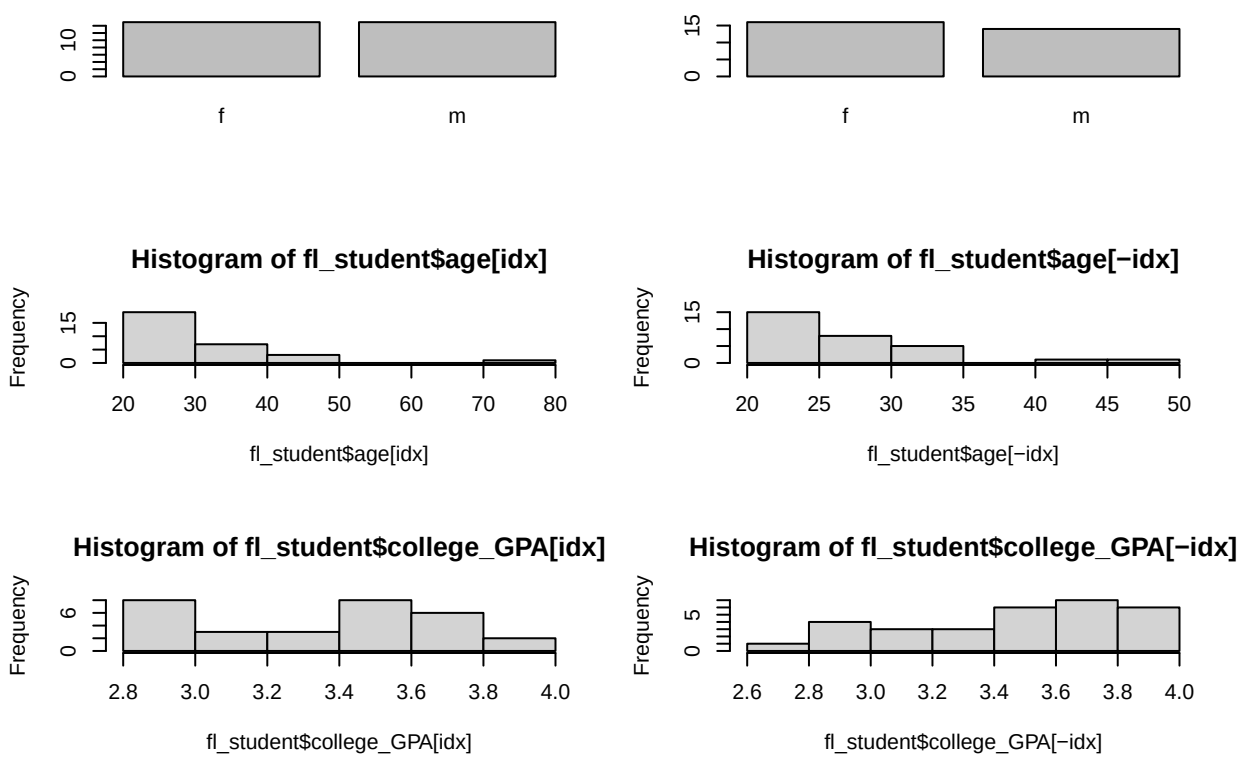
1) She should randomly assign the students to the treatment groups to balance out each group on any potential lurking variables. Note, I assumed that one of the two teaching methods was the "current" or "most prevalent" method and would act as a baseline/control. If this is not the case, then she might want to have a third group who undergoes this treatment to act as a control.

```
fl_student <- read.csv("https://img1.wsimg.com/blobby/go/bbca5dba-4947-4587-b40a-db346c01b1b3/downloads/

print(nrow(fl_student))
```

```
## [1] 60
```

```
idx <- sample(1:60, 30)
par(mfrow=c(3,2))
barplot(table(fl_student$gender[idx]))
barplot(table(fl_student$gender[-idx]))
hist(fl_student$age[idx])
hist(fl_student$age[-idx])
hist(fl_student$college_GPA[idx])
hist(fl_student$college_GPA[-idx])
```




```
par(mfrow=c(1,1))
```

2) If the volunteers could commit to an extended study, I'd have her teach the same material but in different methods for the different groups over the course of three-five sessions. At the end of each session would be a mini-quiz to test comprehension during the class. At the end of the three sessions, there would be a cumulative quiz. Furthermore, at the start of each session she could provide a small questionnaire regarding confidence/understanding in the particular subject being taught and an identical one at the end of of the session. Ideally this would contain some open-ended responses such as "favorite"/"worst" aspect of the teaching method and suggestions for improvements. She could

measure the effectiveness of the teaching methods by comparing quiz scores and comparing the before and after responses of the questionnaires.

3) Because participation is voluntary and promotion is done via announcement post, there is potential for undercoverage and undersampling. Only individuals who keep up-to-date on online announcements from the given source she posted on would be reached, and further self-selection would likely have occurred, with students who are unhappy with the current teaching method might be more likely to participate. If participation is open to the entire student body, then those who are familiar with the professor and/or have interest in the material they teach might be more willing to participate. The results of the study can be generalized to some student populations but not others depending on who the target population is. If it is all the students in her class, then yes. If it is the entire student body, then no.

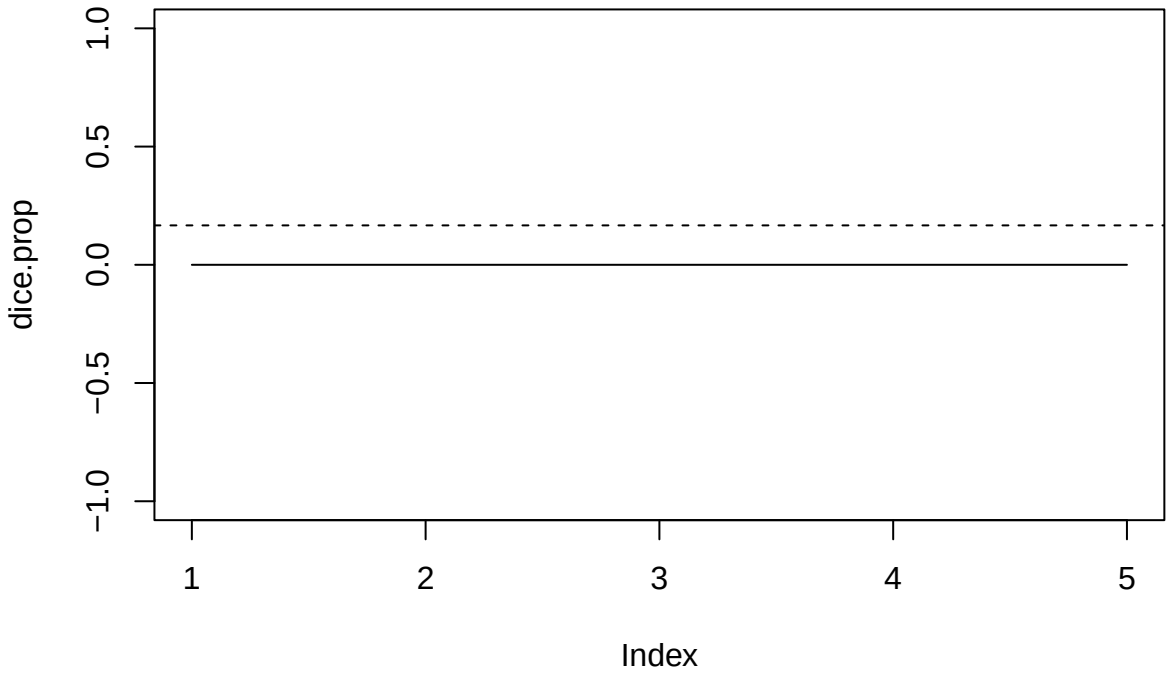
Problem 3

1. For the later runs, I ran them first without seeds to get a better idea of trends and then decided upon using a seed for creating the pdf document.

```
roll.dice <- function(nrolls, seed=NULL){
  set.seed(seed)
  dice.out <- sample(1:6, nrolls, replace=T)
  dice.prop <- numeric(nrolls)
  for( i in 1:nrolls ){
    dice.prop[i] <- mean(dice.out[1:i] == 6)
  }
  plot(dice.prop, type="l")
  abline(h=1/6, lty=2)
}
```

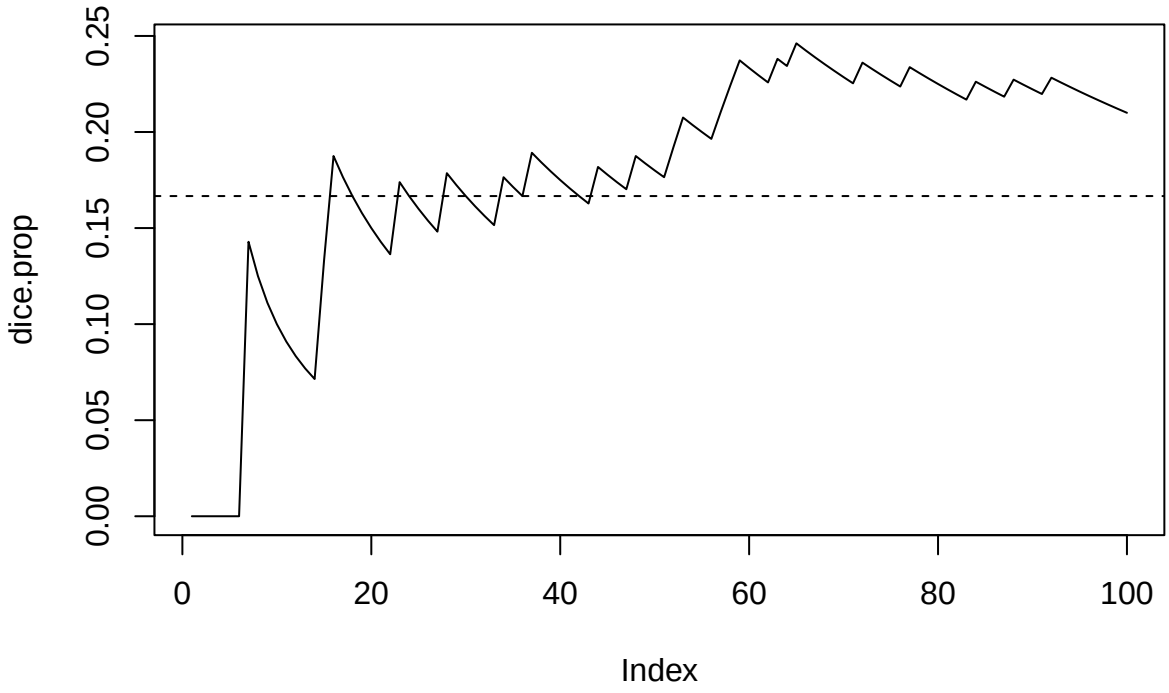
a) The proportion of rolling a 6 starts over the course of 5 trials creates an unpredictable line graph. Sometimes it remains a straight line at 0, sometimes it spikes and drops everywhere seemingly missing the true probability, sometimes it seems to result in a reasonable proportion. Either way, it's a toss up and is not consistently representative of the true probability.

```
roll.dice(5, 1)
```



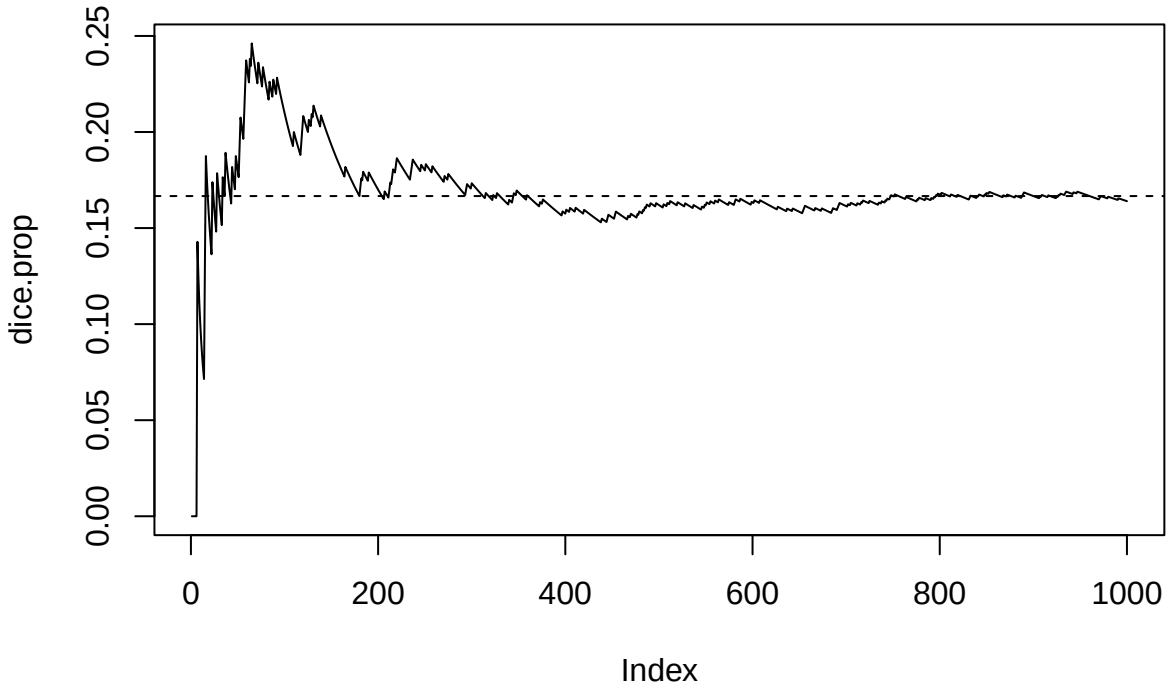
b) The proportion of rolling a 6 starts far away from the true probability, but over more trials the spiking is gradually pulled toward the true probability.

```
roll.dice(100, 1)
```



c) The proportion of rolling a 6 starts far away from the true probability, but over more trials the spiking is gradually pulled toward the true probability. Eventually, the spiking appears to level out and hover right around the true probability.

```
roll.dice(1000, 1)
```



2.

```
n.rep <- 10000
no.6.row <- numeric(n.rep)
for(j in 1:n.rep){
  dice.out <- sample(1:6, 100, replace=T)
  seq.6 <- logical(length(dice.out))
  for(i in 3:length(dice.out)){
    seq.6[i] <- all(dice.out[(i-2):i] == c(6, 6, 6))
  }
  no.6.row[j] <- sum(seq.6)
  #print(paste(j, no.6.row[j]))
}
x <- mean(no.6.row >= 1)
```

If we roll a fair die 100 times it is 0.3224 likely to get three 6's in a row at least once.

Problem 4

4.9 a) Observational study because assigning to treatments (smoke or not smoke) is unethical and would be hard to enforce over several years.

b) Observational study because assigning subjects an SAT score doesn't measure the subjects intellectual potential and therefore cannot be compared against any college GPA score they may recieve. Subjects' SAT scores must be observed and then compared to their college GPA score.

c) Experimental study because coupon placement would be easy for a mail-order company to manipulate and directly measure results from. Treatments can be easily assigned as the mail-order company can randomly choose which version of the coupon book a subject receives in the mail. This would allow them to condclude causal relationships.

4.22 a) All US employers.

b) In order to calculate the nonresponse rate, the number of total polls sent out is needed.

c)

- Undercoverage: Employers who do not use one of the 300 career service centers on college campuses are not being included in the sample
- Non-response bias: Completing a poll is voluntary. Employers who do not track or keep a record of employee education are unlikely to respond.

4.26 a) Teenagers without easy access to the Internet are less likely to respond to the survey

b) Perhaps those that have bought alcohol online would be less likely to respond due to concerns of repercussions.

c) If parents are nearby or would have access to their child's response, teens are more likely to lie.

4.35 a) This study is an experimental study because subjects are being assigned treatments (feed manipulation or no feed manipulation).

b) Single Facebook user

c)
- explanatory variable: positive feed manipulation vs. no feed manipulation (control)
- response variables: percentage of positive words written and percentage of negative words written by a subject

d) Facebook perhaps didn't notify users of participation so that their responses wouldn't be biased. Just knowing Facebook is potentially manipulating their feed could make subjects in a bad mood which might influence their responses to the content they see. Or they could decide to ignore the content and that would also influence the results. The subjects, however, may have argued that not notifying them is unethical because although it may have been part of terms of agreement, they were not knowingly aware that this is something they were agreeing to. They might argue manipulation is not cool or that it infringies upon their privacy.

4.40 a) - i) The treatments are regular large dosage of vitamin C and regular no dosage of vitamin C given in the form of identical looking pills where one contains vitamin C and the other is a placebo. - ii) I would randomly assign the subjects to either treatment A or treatment B. - iii) Having identical pills, one vitamin C and the other placebo makes the study blind (subjects don't know which group they are in). To make the study double-blind, all operators who have contact with the subjects also do not know what treatment the subjects were given.

b) An observational study which concludes that individuals who take vitamin C regularly have less colds, on average, may be misleading as it does not account for any lurking variables which may exist between the group who takes it regularly and those who don't.

4.46 a) Retrospective studies are observational studies which subjects are recruited based on having and not having the outcome of interest. The researchers then look back and compare their exposure status, given their outcome.

b)

- cases: number of subjects who have eye cancer
- control: number of subjects who do not have eye cancer

c)
- i) $16/118 = 0.1355932$
- ii) $46/475 = 0.0968421$

4.48 This study was a prospective study because participants were recruited to first collect baseline exposure data (smoker vs. non-smoker) before any subjects had developed the outcome of interest (death). Later in the future, a follow-up with the subjects revealed outcome of interest.

4.52 a) His claim is based on a small amount of trials which is not conducive for predicting the true probability.

b) To ensure that the cumulative proportion of heads falls very close to 1/2, several more trials would need to be conducted.