Homework 6

Miles Tweed

October 17, 2020

Problem 1

Part 1

The most obvious example of sampling bias is that only user of the platform will be included in the sample. It could be possible that the users of that platform find certain content acceptable on average that the general population would find unacceptable. This study could also suffer from non-response bias.

Part 2

a

The CEO introduced sampling bias when he initially tested his platform on a group that consisted of only engineers that he thought would give a certain type of criticism. This sample was not representative of the general population therefore the bias was due to under-coverage. Also, in his initial sample only engineers or people involved in the project took part which is self-selection bias. As a result, the product was not user friendly for the general public and felt "engineered" to people who were not engineers.

b

He is providing extra support to that one group by explaining the platform to them. This could influence their responses. The moment that the CEO walked into the room the subjects immediately knew that they were testing the actual product so they were no longer blind to the conditions of testing.

Part 2

Problem 2

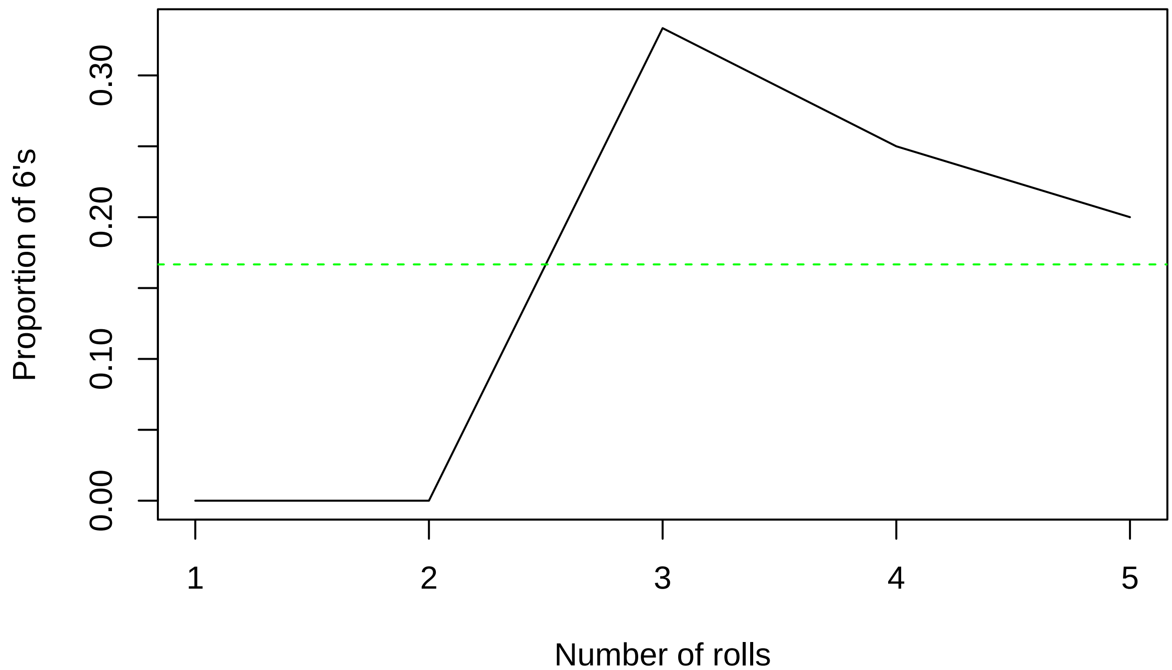
Part 1

```r
roll <- function(n = 1, seed = NA) {
  if (!is.na(seed)){
    set.seed(seed)
  }
  r <- sample(1:6, n ,replace = TRUE)
  props <- vector()
  for (i in 1:n){
    props <- append(props, sum(r[1:i] == 6)/i)
  }
  plot(x = 1:n, y = props, type = 'l',
       xlab = "Number of rolls", ylab = "Proportion of 6's")
  abline(h = (1/6), lty = 2, col = "green")
}
```

a

In 5 rolls the proportion spikes once to 0.33 when one 6 is found in three rolls. This proportions drops back down to around 0.2 in the remaining two rolls. However, it does not reach the theoretical value of $P = \frac{1}{6} \approx 0.1667$ (green line).

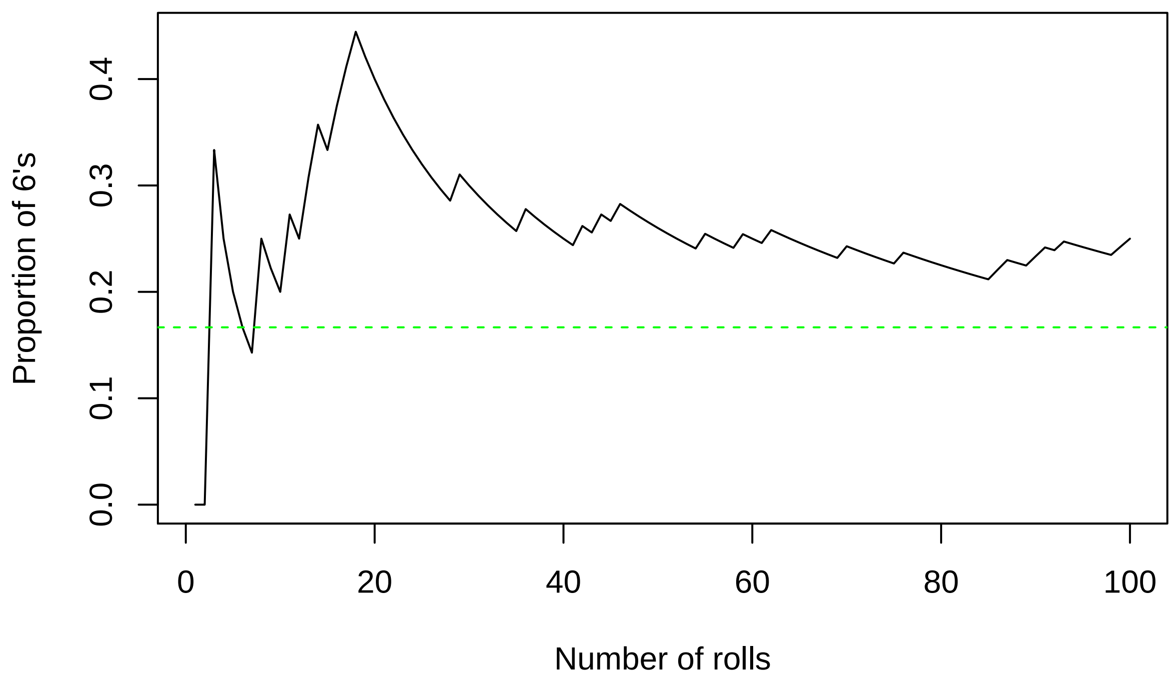
```
roll(n=5, seed = 1234)
```



###

b Again the proportion has a spike (or several) in the beginning as more sixes are rolled but the trend starts to settle after about 40 rolls. However, it still does not reach the theoretical value of $P = \frac{1}{6} \approx 0.1667$ (green line).

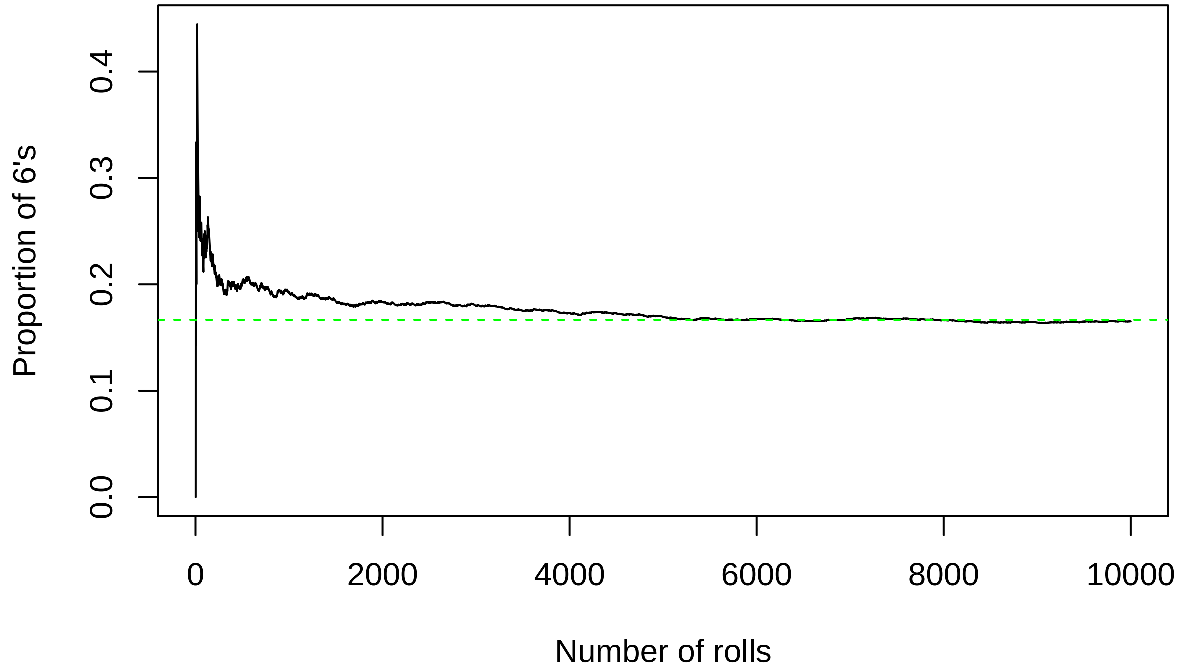
```
roll(n=100, seed = 1234)
```



###

c Finally, there is once again a spike early on and the value settles to the theoretical value of $P = \frac{1}{6} \approx 0.1667$ after about 5000 rolls.

```
with_seed <- roll(n=10000, seed = 1234)
```



Part 2

```
num.666.found <- vector()
for (n in 1:10000){
  run.tot <- 0
  r <- sample(1:6, 100, replace = T)
  for (i in 1:100){
    if (i+2 <= 100){
      if (r[i] == 6){
        if (r[i+1] == 6){
          if (r[i+2] == 6){
            if(r[i+3] != 6 | is.na(r[i+3])){
              run.tot <- run.tot + 1
            }
          }
        }
      }
    }
  }
  num.666.found <- append(num.666.found, run.tot)
}

print(paste("After 10000 trials the probability of",
            "at least 1 instance of 666 in 100 rolls is:"
            ,mean(num.666.found >= 1)))
```

```
## [1] "After 10000 trials the probability of at least 1 instance of 666 in 100 rolls is: 0.3214"
```

Problem 3

4.35

a) The researchers had control over the factors of the study and chose the individuals that made up each group and the treatments that the groups received.

b) The experimental units are the Facebook users selected for the study.

c) The explanatory variable is the amount of positive content on the users news feed. The response variables are (i) the percentage of positive words produced by the person and (ii) the percentage of negative words produced by the person following exposure.

d) Being actively exposed to more negative content could effect the users psyche. This could negatively affect other aspects of the user's life over the course of the experiment.

4.40

a) To test this you could give large regular doses of vitamin C to subjects over the course of a year or longer and compare the number of colds that the subjects report with those whose vitamin C intake is unchanged. (i) The treatments are regular large doses of Vitamin C or the lack thereof. (ii) I would randomly select half of the subjects to receive a regular large dose of vitamin C while the remaining subjects would have unchanged vitamin C intake. (iii) The study could be double blind if every subject receives an identical pill to take where some contain the large dose of vitamin C while the other is a placebo. The study will be double blind as long as neither the subjects nor the people distributing the pills know which contains the vitamin C.

b) The results of these studies can be misleading for many reason. First, the amount of vitamin C intake is a difficult thing to quantify since the subject's diet is a major factor that is difficult to track accurately. Also, the subject's exposure to factors that could lead them to have a cold are impossible to control which makes the link between vitamin C intake and developing a cold impossible to determine with any certainly.

4.46

a) This means that the researchers studied historical information about the subjects in order to observe any association between response and explanatory variables.

b) Cases were the subjects that had eye cancer (the outcome of interest) while controls were subjects that did not (the other possible outcome).

c) Approximately 13.6% of those with eye cancer and 9.7% of those without eye cancer used mobile phones.

4.48

I believe this would be a prospective study, because the researchers began the study with with the goal to follow up with the subjects twenty years later. They collected information from them at the beginning of the study in order to explain a response variable that would be assessed in twenty years.

5.2

a) Each successive toss of the coin is independent of the previous and the actual number of times that a coin is heads versus tails can vary therefore he did not take into count the action over the long term. If he had flipped the coin 100 times instead of 10 the proportion of heads tails would likely be closer to 0.5 if it was a fair coin.

b) You would have to flip the coin a large number of times.

5.16

a) There are two possible outcomes for the ten questions so the total number of possible outcomes is $2^{10} = 1024$.

b) There is only one outcome in which someone would not get at least one question wrong (i.e. getting every question right). Therefore, the probability of getting at least one question wrong is $P(A) = \frac{1023}{1024}$ and the complement to this probability is $P(A^c) = 1 - P(A) = 1 - \frac{1023}{1024} = \frac{1}{1024}$.

c) Since each outcome for every question is equally likely the probability of getting at least one question wrong with random guessing is $\frac{1023}{1024} = 0.99902 \approx 0.999$.

5.23

a) There are four possible outcomes (YS, NS, YD, ND).

b) Using the total number of events with the outcome of interest and dividing that by the total number of outcomes: (i) $P(D) = \frac{2111}{577006} \approx 0.00366$ (ii) $P(N) = \frac{164128}{577006} \approx 0.284$

c) $P(ND) = \frac{1601}{577006} \approx 0.00277$

d) If $P(N)$ and $P(D)$ were independent the probability of them both occurring would be $P(ND) = P(N) \cdot P(D) = \frac{2111}{577006} \cdot \frac{164128}{577006} \approx 0.00104$. This shows that $P(N)$ and $P(D)$ are not independent events and wearing a seat belt does have and effect on the probability of surviving an accident.

5.27

a) There are two possible outcomes for each customer (making a sale or not) and because there are three potential customers the total number of possible outcomes is $2^3 = 8$. (YYY, YYN, YNY, NYY, YNN, NYN, NNY, NNN)

b) Since there is only one outcome in which no sale is made the probability of making at least one sale to the three is $P = \frac{7}{8} = 0.875$.

c) These calculations assume that the events of an individual customer buying something is independent of whether or not another customer buys something. This assumption may be unrealistic if the customers have some influence on each other as could be the case if they were friends or family.