Homework 04 - Producing Data

Problem 1

A health study is being conducted on a group of volunteers (487 smokers and 1513 non-smokers) to determine the effect of a new drug.

Part a)

Estimate how many smokers are in a simple random sample of size 200.

The proportion of smokers in the population is $487/2000 = 0.2435$. In a simple random sample of size 200, we expect approximately $200 \times 0.2435 = 48.7 \approx 49$ smokers.

Part b)

Using R, determine 10 simple random samples, each of size 200, and record the number of smokers in each of the samples. Let's agree to label the smokers with numbers 1 through 487 and the nonsmokers with numbers 488 through 2000.

- Run the following code.
- Explain what the indicated line of code does in the loop.

```r
population = 1:2000

smoker_count = rep(0, 10)

for (k in 1:10) {
  tmp_sample = sample(population, size = 200)

  # This line counts how many subjects in the sample are smokers (labeled 1-487)
  # by checking how many sampled values are less than 488, and stores the count.
  smoker_count[k] = sum( tmp_sample < 488 )
}

(smoker.df = data.frame(smokers = smoker_count, nsmokers = 200 - smoker_count))
```

```
   smokers nsmokers
1       45      155
2       58      142
3       43      157
```

```
4         51        149
5         47        153
6         44        156
7         44        156
8         36        164
9         48        152
10        42        158
```

The indicated line `smoker_count[k] = sum(tmp_sample < 488)` counts how many subjects in the *k*-th sample are smokers. Since smokers are labeled 1 through 487, any sampled label less than 488 belongs to a smoker. The `sum()` function counts how many such labels exist in the sample, and that count is stored as the *k*-th entry of `smoker_count`.

Part c)

Using R, determine the mean and the standard deviation of the number of smokers in the samples. How does the sample mean conform (i.e., is it far/close) to your calculation for the predicted population mean?

```r
(avg_smoker <- mean(smoker_count))
```

```
[1] 45.8
```

```r
(sd_smoker <- sd(smoker_count))
```

```
[1] 5.846176
```

The sample mean of 45.8 is close to our predicted value of 48.7 smokers per sample. This makes sense because we used simple random sampling.

Part d)

The head researcher would like to choose 20 subjects for comprehensive medical imaging. Using R, perform a single stratified random sample having 10 smokers and 10 nonsmokers, and display the labels for the selected subjects in ascending order.

```r
samp_smokers = sample(1:487, 10)
samp_nonsmokers = sample(488:2000, 10)

(smoker.df = data.frame(smokers = sort(samp_smokers), nonsmokers = sort(samp_nonsmokers)))
```

```
   smokers nonsmokers
1       19        606
2       71        610
3       74        687
4       86        717
5      138        740
6      148        816
7      191        824
8      197        893
9      323        916
10     374       1031
```