d-worksheet : Producing Data

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.

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

The proportion of smokers to the total sample is 487/2000. Therefore, there would be (487/2000) * 200 = 48.7 which is approximately 49 smokers out of a sample of 200.

b. Determine ten simple random samples of size 200 and record the number of smokers in each of the samples. Let's agree to label the smokers 1 through 487 and the nonsmokers 488 through 2000.

```
population <- seq(1,2000)
smoker_count <- rep(0,10)
for (k in 1:10)
{
# the call to sample()should include the vector to be sampled (population)
# along with the number to draw and whether or not to replace each
# subject after it's drawn (set replace = FALSE)

# the call to sum here should be an expression in terms of tmp_sample
# e.g., sum(tmp_sample < 3) counts the number of entries in tmp_sample
# strictly less than 3.

  tmp_sample <- sample(population, 200, replace = FALSE)
  smoker_count[k] <- sum(tmp_sample < 488)
}

# we'll make a data frame to store the counts nicely, wrapping
# it in parenthesis displays
(smoker.df = data.frame(trial= 1:10, nsmokers = smoker_count))
```

```
##    trial nsmokers
## 1      1       45
## 2      2       61
## 3      3       48
## 4      4       45
## 5      5       57
## 6      6       41
## 7      7       49
## 8      8       47
## 9      9       45
## 10    10       38
```

c. Find the mean and the standard deviation of the number of smokers in the samples. How does the sample mean and sample standard deviation conform to your calculation for the predicted population mean?

Since our predicted number of smokers in the trial falls between the sample mean plus/minus the sample standard deviation of the sample, we can say that our predicted number is accurate.

```
# You're on your own for this one *cough* *cough* mean(), sd() *cough*
mean(smoker_count)
```

```
## [1] 47.6
```

```
sd(smoker_count)
```

```
## [1] 6.883152
```

d. The head researcher would like to choose 20 subjects for comprehensive medical imaging. Perform a single stratified random sample having 10 smokers and 10 nonsmokers and display the labels for the selected subjects. *Hint: This is very similar to the process in D2.10 and not too different from what you did above.*

```
# should be two different calls to sample()
# sort the two samples if you're feeling fancy
(strat_smokers<-sample(population[1:487], 10))
```

```
##  [1] 447 435 450 126  29 277 157  23 242  40
```

```
(strat_nonsmokers<-sample(population[488:2000], 10))
```

```
##  [1] 1618  774  496 1726 1628 1447 1017 1298 1556  590
```