Lab 10

Lab Overview

This lab will look at sampling with unequal weights using natural gas deliveries. The dataset contains natural gas deliveries (in million cubic feet) for 56 utilities in Montana, Wyoming, Idaho, Utah, Colorado, North Dakota, and South Dakota.

```
NG <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/EIA_NG.csv')
NG
```

```
## # A tibble: 56 x 4
##    Company                              total2018 total2017 weight
##    <chr>                                    <dbl>     <dbl>  <dbl>
##  1 PUB SERVICE CO OF COLORADO             253050.   225147. 0.231
##  2 QUESTAR GAS COMPANY                    185972.   175728. 0.180
##  3 COLORADO INTERSTATE GAS COMPANY LLC    105885.    93805. 0.0962
##  4 NORTHWESTERN ENERGY                     80145.    76020. 0.0780
##  5 INTERMOUNTAIN GAS COMPANY               68606.    70353. 0.0722
##  6 BLACK HILLS ENERGY                      62054.    51519. 0.0528
##  7 MONTANA DAKOTA UTILITIES CO             49411.    44459. 0.0456
##  8 COLORADO SPRINGS UTILITIES              40102.    32914. 0.0338
##  9 WBI ENERGY TRANSMISSION INC             29335.    25123. 0.0258
## 10 ATMOS ENERGY CORPORATION                18016.    17137. 0.0176
## # ... with 46 more rows
```

```
N <- nrow(NG)
```

Rather than explicitly calculating the variance for each scenario, we will use the repeated sampling technique from earlier in the class.

1. (4 points)

Take a simple random sample of size 20 and estimate the total natural gas delivered in this region for the year 2018.

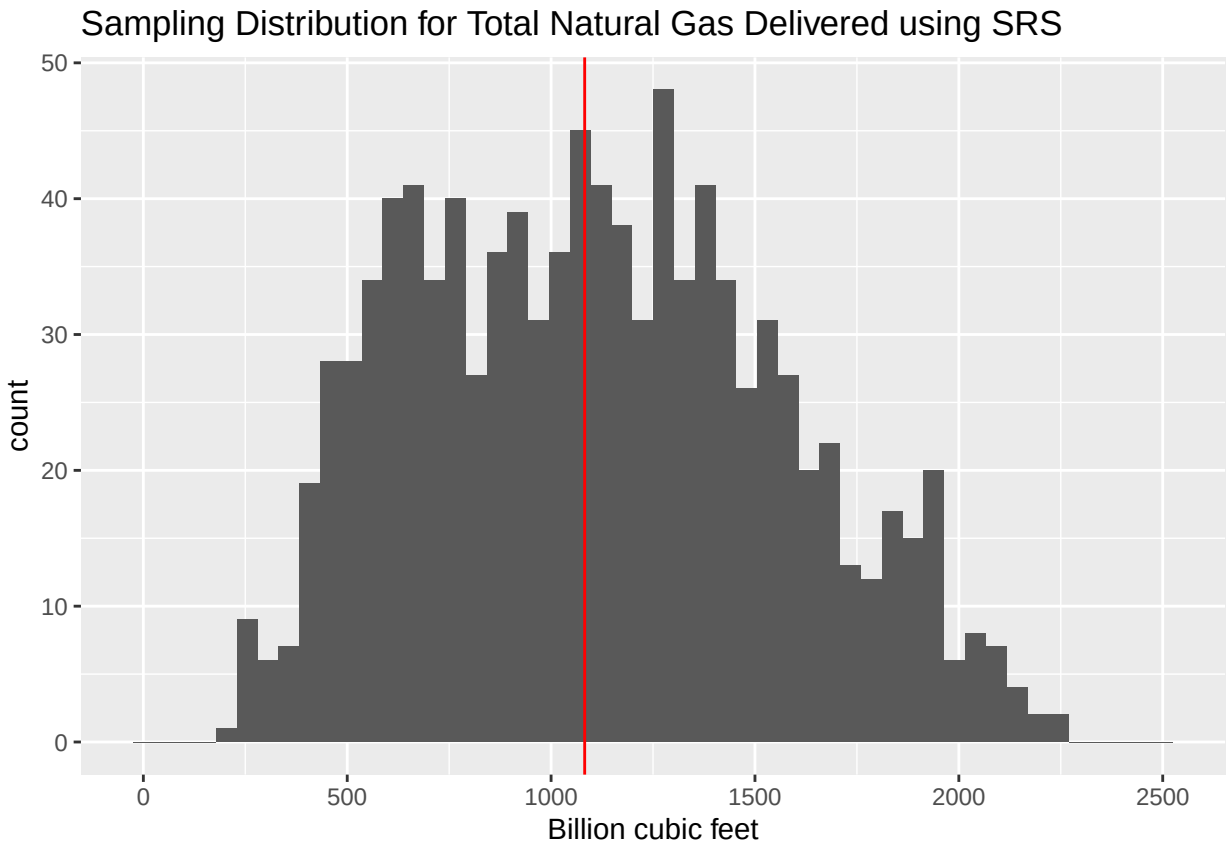
```
total <- mean(sample(NG$total2018, 20)) * N
```

Based on this sample, our estimate of the total natural gas delivered is 793 billion cubic feet.

2. (4 points)

Now repeat the process from part 1 and take 1000 different samples of size 20 and plot the resultant point estimator along with true value.

```
estimated_totals <- replicate(1000, mean(sample(NG$total2018, 20)) * nrow(NG))
tibble(x = estimated_totals / 1000) %>% ggplot(aes(x=x)) + geom_histogram(bins = 50) +
  geom_vline(xintercept = sum(NG$total2018)/1000, col = 'red') +
  ggtitle('Sampling Distribution for Total Natural Gas Delivered using SRS') +
  xlab('Billion cubic feet') + expand_limits(x=c(0,2500))
```



3. (4 points)

Take a sample of size 20 with unequal probability using the `weight` column in the dataset. Then estimate the total natural gas delivered in this region for the year 2018 using the Hansen-Hurwitz framework.

```
indices_hh <- sample(1:N, size = 20, replace = T, prob = NG$weight)
total_hh <- mean(NG$total2018[indices_hh] / NG$weight[indices_hh])
```

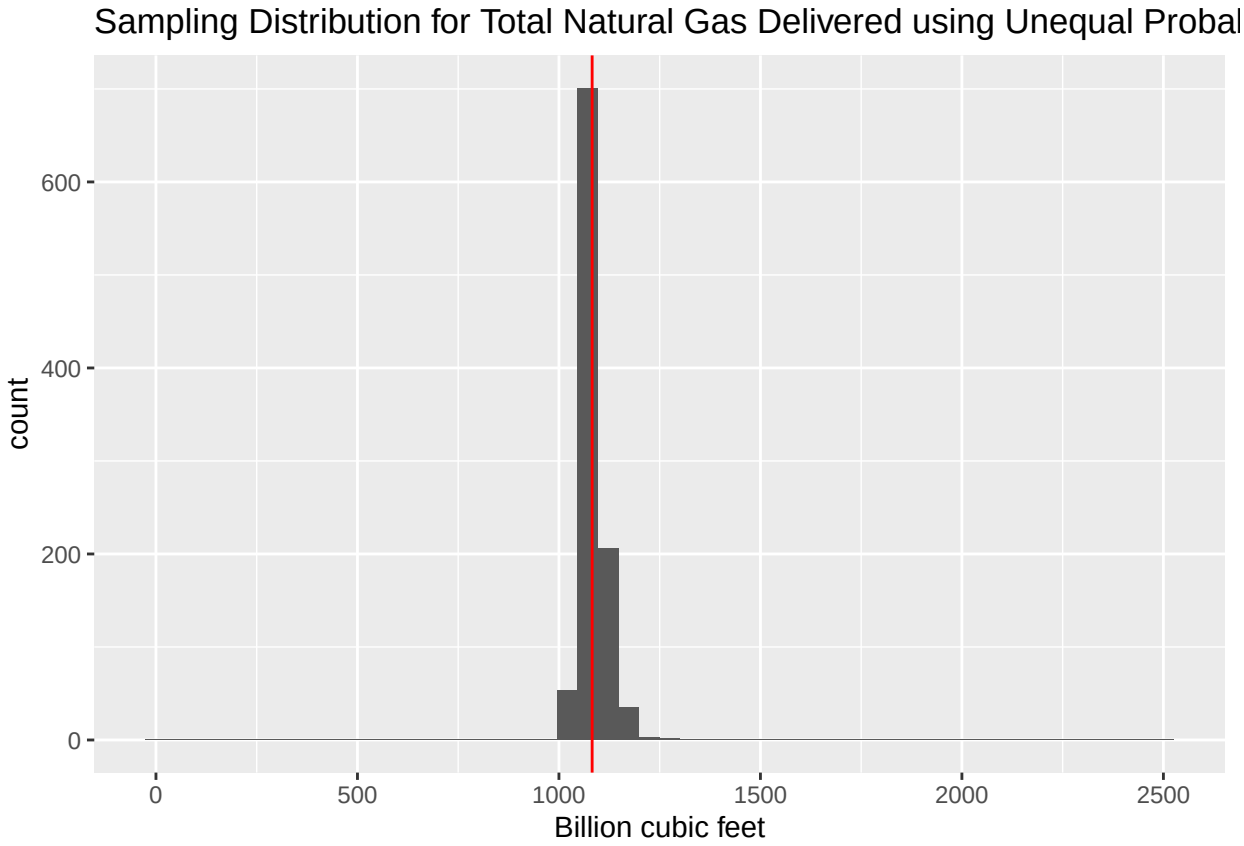
Based on this sample, our estimate of the total natural gas delivered is 1061 billion cubic feet.

4. (4 points)

Now repeat the process from part 3 and take 1000 different samples of size 20 and plot the resultant point estimator along with true value.

```
estimated_totals_hh <- rep(0, 1000)
for (i in 1:1000){
  indices_hh <- sample(1:N, size = 20, replace = T, prob = NG$weight)
  estimated_totals_hh[i] <- mean(NG$total2018[indices_hh] / NG$weight[indices_hh])
}


tibble(x = estimated_totals_hh / 1000) %>% ggplot(aes(x=x)) + geom_histogram(bins = 50) +
  geom_vline(xintercept = sum(NG$total2018)/1000, col = 'red') +
  ggtitle('Sampling Distribution for Total Natural Gas Delivered using Unequal Probability Sampling') +
  xlab('Billion cubic feet') + expand_limits(x=c(0,2500))
```



5. (4 points)

Comment on the strengths and weaknesses of the approaches in part 1 and part 3.

The SRS approach is unbiased, but has large variance. The unequal probability sample in part 3 requires careful selection of sampling weights; however, the variance is greatly reduced.