Homework 1

Xinyi Lin

9/14/2019

```
library(tidyverse)
library(pps)
```

Question 1

Problem a

Based on the table of random digits, the IDS of selected sample are 38, 44, 46, 52, 57, 62, 63, 65, 78, 104, 130, 145, 147, 154, 155, 176, 177, 178, 183, 196.

Problem b

```
sample_id = c(38, 44, 46, 52, 57, 62, 63, 65, 78, 104, 130, 145, 147, 154,
155, 176, 177, 178, 183, 196)
sampleQ1 = read.csv("./HW1dataQ1.csv") %>%
  dplyr::filter(ID %in% sample_id)
mean(sampleQ1$SBP)

## [1] 117.05
```

The mean Systolic Blood Pressure(mmHg) $\overline{y}$ is 117.05.

Problem c

```
var(sampleQ1$SBP)

## [1] 80.15526
```

The sample variance $s^2 = 80.155$, the finite population correction$f = \frac{n}{N} = 0.1$ and the estimation of the sampling variance of $\overline{y}$ is:

$$\widehat{Var(\overline{y})} = \frac{1-f}{n} s^2 = \frac{0.9}{20} \times 80.155 = 3.607$$

.

Problem d

```
117.05-qt(0.975,19)*sqrt(3.607)

## [1] 113.0749

117.05+qt(0.975,19)*sqrt(3.607)

## [1] 121.0251
```

As the 95% confidence interval for the estimated mean($\overline{y}$) is $\overline{y} \pm t_{0.975,n-1}\sqrt{\widehat{Var(\overline{y})}}$, the 95% confidence interval for the estimated mean is (113.0749, 121.0251).

Problem e

As the sampling variance of $\overline{y}$ is $Var(\overline{y}) = \frac{1-f}{n}S^2$. When sample size increase, S^2 does not change, while n and f increase, which decrease $Var(\overline{y})$, so $Var(\overline{y})$ gets smaller.

Question 2

```
dataQ2 = read.csv("./HW1dataQ2.csv")
set.seed(3)
ppss(dataQ2$area, 4)

## [1] 1 2 3 5
```

If we use `ppss` function directly, the IDs of selected sample is 10005, 10008, 10009, 10012 when the seed is 3.

Problem a

```
round(dataQ2$area*4/110361, 4)

##  [1] 0.2604 1.7686 0.7772 0.1244 0.3878 0.0420 0.0830 0.3551 0.0713 0.1301
```

For farm 10005, the probability of selection is: $\pi_1 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times7185}{110361} = 0.2604 < 1.$

For farm 10008, the probability of selection is: $\pi_2 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times48795}{110361} = 1.7686 > 1.$

For farm 10009, the probability of selection is: $\pi_3 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times21444}{110361} = 0.7772 < 1.$

For farm 10011, the probability of selection is: $\pi_4 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times3432}{110361} = 0.1244 < 1.$

For farm 10012, the probability of selection is: $\pi_5 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times10700}{110361} = 0.3878 < 1.$

For farm 10014, the probability of selection is: $\pi_6 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times1159}{110361} = 0.0420 < 1.$

For farm 10015, the probability of selection is: $\pi_7 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times2291}{110361} = 0.0830 < 1.$

For farm 10016, the probability of selection is: $\pi_8 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times9798}{110361} = 0.3551 < 1.$

For farm 10017, the probability of selection is: $\pi_9 = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times1961}{110361} = 0.0713 < 1.$

For farm 10019, the probability of selection is: $\pi_{10} = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{4\times3590}{110361} = 0.1301 < 1.$

As $\pi_2 > 1$, the farm 1008 should be sampled with certainty.

Problem b

```
new_area = dataQ2$area[-2]
cumsum(new_area)

## [1]  7185 28629 32061 42761 43920 46211 56009 57976 61566
```

The cumulative counts of area are: 7185, 28629, 32061, 42761, 43920, 46211, 56009, 57976, 61566. As the random start number is 10000, the second farm in the new sample is selected. $10000 + 20522(61566/3) = 30522$ and $30522 + 20522 = 51044$, so the third and seventh farm in the new sample are selected.

As the second farm is removed, the actual IDs of selected sample is 10008, 10009, 10011, 10016.