Final - Part B

Wesley Chang

Fall 2020, Professor Subir Ghosh

```
library(tidyverse)

# import the names and student numbers of every person in the STAT 130 class
# into data frame stat130
stat130 <- as_tibble(read.csv("stat130.csv"))
stat130
```

```
## # A tibble: 38 x 2
##    number student
##     <int> <chr>
##  1      1 Christian
##  2      2 Zachary
##  3      3 Wesley
##  4      4 Laura
##  5      5 Morgan
##  6      6 Quyen
##  7      7 Lin
##  8      8 Shalom
##  9      9 Daniel
## 10     10 Huijin
## # ... with 28 more rows
```

```
# set seed to output consistent results
# comment this if variable sample results are desired
set.seed(130)
```

Question 1

Draw a simple random sample without replacement. Present the names of 10 students.

Answer

```
# SRS of ten rows from the data frame stat130
sample10 <- stat130 %>% sample_n(10)

# final answer
sample10
```

```
## # A tibble: 10 x 2
```

```
##    number student
##     <int> <chr>
##  1      3 Wesley
##  2      5 Morgan
##  3     11 Yihao
##  4     20 Jade
##  5     24 Alexander
##  6     37 Alexandra
##  7     28 Baron
##  8     36 Sawanee
##  9      1 Christian
## 10      2 Zachary
```

The SRS WOR of size `n = 10` is found in `sample10`. I used the `sample_n()` function to draw the names of 10 students from the population of 38 students, without replacement. The names of the ten students are presented above.

Question 2

Draw a stratified random sample with proportional allocation, where the strata are:

```
  Stratum 1: Students 1-19
  Stratum 2: Students 20-38
```

Present the names of 10 students by strata

Answer

```
# add column to data frame stat130
# that describes which stratum each student belongs to
q2_stratified <- stat130 %>%
  mutate(stratum = case_when(
    stat130$number <= 19 ~ "Stratum 1",
    stat130$number <= 38 ~ "Stratum 2"
  ))

# determine population total from the number of rows in data frame q2_stratified
PopTotal <- nrow(q2_stratified)
# determine the proportion of each stratum over the population (n/N or 10/38)
noverN <- 10 / PopTotal

# create empty list to store the students sampled from each stratum
q2_sample_list <- vector(mode = "list")

# for loop to take a SRS of (n/N * population size of each stratum) students from each stratum
# store the results in list q2_sample_list
for (i in 1:2) {
  temp <- q2_stratified %>% filter(stratum == paste0("Stratum ",i))

  sample_num_per_stratum <- round(nrow(temp) * noverN)
  q2_sample_list[[i]] <- temp %>% sample_n(sample_num_per_stratum)
}
```

```r
# combine all data frames in q2_sample_list into data frame q2_final_sample
q2_final_sample <- q2_sample_list %>% bind_rows()

# final answer
q2_final_sample
```

```
## # A tibble: 10 x 3
##    number student     stratum
##     <int> <chr>       <chr>
##  1      5 Morgan      Stratum 1
##  2     19 Antoinette  Stratum 1
##  3     18 Tiffany Ngo Stratum 1
##  4     15 Thymi       Stratum 1
##  5      3 Wesley      Stratum 1
##  6     28 Baron       Stratum 2
##  7     32 Yanlin      Stratum 2
##  8     21 Alan        Stratum 2
##  9     31 Alice       Stratum 2
## 10     27 Tiffany Tom Stratum 2
```

The stratified random sample with proportional allocation of size `n = 10` is found in `q2_SRS_10`. Since the given population sizes of strata 1 and 2 both represent 50% of the population, for a desired sample size of `n = 10` we allocate 5 students to each stratum. We then use the `sample_n()` function to draw the names of 5 students from each strata, and the results are presented above.

Question 3

Draw a stratified random sample with proportional allocation, where the strata are:

```
Stratum 1: Students 1-16
Stratum 2: Students 17-28
Stratum 3: Students 29-34
Stratum 4: Students 35-38
```

Present the names of 10 students by strata

Answer

```r
# add column to data frame stat130
# that describes which stratum each student belongs to
q3_stratified <- stat130 %>%
  mutate(stratum = case_when(
    stat130$number <= 16 ~ "Stratum 1",
    stat130$number <= 28 ~ "Stratum 2",
    stat130$number <= 34 ~ "Stratum 3",
    stat130$number <= 38 ~ "Stratum 4"
  ))

# determine population total from the number of rows in data frame q3_stratified
PopTotal <- nrow(q3_stratified)
# determine the proportion of each stratum over the population (n/N or 10/38)
```

```r
noverN <- 10 / PopTotal

# create empty list to store the students sampled from each stratum
q3_sample_list <- vector(mode = "list")

# for loop to take a SRS of (n/N * population size of each stratum) students from each stratum
# store the results in list q3_sample_list
for (i in 1:4) {
  # filter out select rows from the data frame by stratum and pass into data frame temp
  temp <- q3_stratified %>% filter(stratum == paste0("Stratum ",i))

  # multiply the number of rows in temp (which is the population total of each stratum)
  # by the proportion n/N
  sample_num_per_stratum <- round(nrow(temp) * noverN)

  # take a SRS of a size determined by prop above from filtered stratum
  q3_sample_list[[i]] <- temp %>% sample_n(sample_num_per_stratum)
}

# combine all data frames in q3_sample_list into data frame q3_final_sample
q3_final_sample <- q3_sample_list %>% bind_rows()

# final answer
q3_final_sample
```

```
## # A tibble: 10 x 3
##    number student     stratum
##     <int> <chr>       <chr>
##  1      3 Wesley      Stratum 1
##  2     16 Camilla     Stratum 1
##  3     13 Julia       Stratum 1
##  4     15 Thymi       Stratum 1
##  5     21 Alan        Stratum 2
##  6     25 Kamille     Stratum 2
##  7     27 Tiffany Tom Stratum 2
##  8     34 Xunyang     Stratum 3
##  9     29 Brian       Stratum 3
## 10     35 Jiaying     Stratum 4
```

To draw a stratified random sample with proportional allocation, we need to first proportionally allocate sample sizes to each strata. We do this by dividing the population size of each strata by the entire population size. Then, we multiply the proportion of each strata by the desired total sample size `n=10`. Once rounded, this gives us sample sizes of 4, 3, 2, and 1 for strata 1-4, respectively, for a total of sample size of `n = 10` in the final sample. The names of students are then drawn from each strata using the `sample_n()` function, and are displayed above.

Question 4

Draw a probability sample of 3 clusters from the 13 clusters given below:

`Cluster 1: 1-3` `Cluster 2: 4-6` `Cluster 3: 7-9`

```
Cluster 4: 10-14      Cluster 5: 15-19      Cluster 6: 20-24
Cluster 7: 25-26      Cluster 8: 27-28      Cluster 9: 29-30
Cluster 10: 31-32     Cluster 11: 33-34     Cluster 12: 35-36
Cluster 13: 37-38
```

How would one draw the 3 clusters in the sample from the 13 in the population to have in the total of 10 students?

Present the names of 10 students in the total from 3 clusters.

Answer

In a one stage cluster sample, the population is divided into clusters, and a predetermined `n` amount of clusters is chosen to be sample. We then take a SRS of all clusters to select the `n` clusters to include in our sample. Once selected, every individual in the cluster is included in the sample. For example, if we want to sample three clusters from the cluster above and our SRS of the sample gives us clusters 1, 7, and 13, we would sample every individual within these clusters. Our sample would include students 1-3, 25-26, and 37-38, for a total of 7 students.

However, since we are given the constraint that the sample total must have 10 students, we must alter the one stage cluster sampling process. Clusters 1-3 each contain three students. Clusters 4-6 each contain five students. Clusters 7-13 each contain two students. Therefore, to have a sample with 3 clusters for a total of ten students, we would need to pick one cluster from each of the above groups. For example, if the sample contains students from cluster 1 with three students, cluster 5 with five students, and cluster 7 with two students, the total amount of students sampled would be ten. In essence, we stratify the clusters based on their cluster sizes (two, three, and five) and then take a cluster sample from each of the strata.

To get a total of 10 students from 3 clusters, I would first separate the clusters into three strata such that the sum of the students in each of three clusters selected (one from each strata) is 10:

Strata 1 (3 students)	Strata 2 (5 students)	Strata 3 (2 students)
Cluster 1	Cluster 4	Cluster 7
Cluster 2	Cluster 5	Cluster 8
Cluster 3	Cluster 6	Cluster 9
-	-	Cluster 10
-	-	Cluster 11
-	-	Cluster 12
-	-	Cluster 13

I would then take a SRS of one cluster from each stratum to, and the resulting three clusters should total ten students.

```r
# add column to data frame that lists the cluster associated with each student
q4_pop <- stat130 %>%
  mutate(cluster = case_when(
        stat130$number <= 3 ~ "Cluster 1",
        stat130$number <= 6 ~ "Cluster 2",
        stat130$number <= 9 ~ "Cluster 3",
        stat130$number <= 14 ~ "Cluster 4",
        stat130$number <= 19 ~ "Cluster 5",
        stat130$number <= 24 ~ "Cluster 6",
        stat130$number <= 26 ~ "Cluster 7",
        stat130$number <= 28 ~ "Cluster 8",
        stat130$number <= 30 ~ "Cluster 9",
        stat130$number <= 32 ~ "Cluster 10",
```

```
        stat130$number <= 34 ~ "Cluster 11",
        stat130$number <= 36 ~ "Cluster 12",
        stat130$number <= 38 ~ "Cluster 13",
  ))

# add column to data frame that lists the stratum associated with each student and cluster
q4_pop <- q4_pop %>%
  mutate(stratum = case_when(
        stat130$number <= 9 ~ "Stratum 1",
        stat130$number <= 24 ~ "Stratum 2",
        stat130$number <= 38 ~ "Stratum 3"
  ))

# create empty list to store the clusters sampled from each stratum
q4_sample_list <- vector(mode = "list")

# for loop to sample (by SRS) 1 cluster from each stratum, and store the data frame in list q4_sample_l
for (i in 1:3) {
  q4_sample_list[[i]] <- sample(q4_pop %>% filter(stratum == paste0("Stratum ",i)) %>% group_split(clus
}

# combine all data frames in q4_sample_list into one data frame
q4_final_sample <- q4_sample_list %>% bind_rows()

# final answer, sample of 10 students from 3 clusters
q4_final_sample
```

```
## # A tibble: 10 x 4
##    number student     cluster   stratum
##     <int> <chr>       <chr>     <chr>
##  1      7 Lin         Cluster 3 Stratum 1
##  2      8 Shalom      Cluster 3 Stratum 1
##  3      9 Daniel      Cluster 3 Stratum 1
##  4     15 Thymi       Cluster 5 Stratum 2
##  5     16 Camilla     Cluster 5 Stratum 2
##  6     17 Sarah       Cluster 5 Stratum 2
##  7     18 Tiffany Ngo Cluster 5 Stratum 2
##  8     19 Antoinette  Cluster 5 Stratum 2
##  9     27 Tiffany Tom Cluster 8 Stratum 3
## 10     28 Baron       Cluster 8 Stratum 3
```

The final answer for this question is displayed in `q4_final_sample`.