Sampling: Midterm I
Leslie Gains-Germain

1. (a) $\hat{t}_{SRS} = 5690$. See R code below.

```
branches.srs <- c(50, 67, 53, 63, 64, 64, 41, 38, 35, 41, 40, 39, 42, 75, 88, 71,
                  65, 40, 38, 41, 44, 39)
that <- 110*mean(branches.srs)
that

## [1] 5690
```

   (b) $SE(\hat{t}_{SRS}) = \sqrt{\hat{V}(\hat{t}_{SRS})} = 81.6726$. See R code below.

```
se.that <- sqrt(110 * (110-22) * sd(branches.srs) / 22)
se.that

## [1] 81.7
```

   (c) A 95% t-based confidence interval for t_{SRS} is $(5520, 5860)$. See R code below. We are 95% confident that the true total number of apples on all apple-bearing branches of these four trees is between 5520 and 5860 apples.

```
tstar <- qt(0.975, 21)
ci <- c(that-tstar*se.that, that+tstar*se.that)
ci

## [1] 5520 5860
```

   (d) At a 5% significance level, there is evidence that the true total number of apples on all apple-bearing branches of these four trees is different than 5500 apples. We make this conclusion because the 95% confidence interval for t_{SRS} does not contain 5500.

   (e) $\hat{t}_{STR} = 5931.167$. See R code below. After rounding, I would estimate the population total t_{STR} to be 5931.

```
branches.str1 <- c(53, 64, 61, 58, 66, 57)
branches.str2 <- c(46, 40, 38, 42, 40, 39)
branches.str3 <- c(66, 80, 74, 89)
branches.str4 <- c(40, 38, 39, 33, 44, 35)
that.str <- 30*mean(branches.str1)+
            25*mean(branches.str2)+
            26*mean(branches.str3)+
            29*mean(branches.str4)
that.str

## [1] 5931
```

(f) $SE(\hat{t}_{STR}) = \sqrt{\hat{V}(\hat{t}_{STR})} = 51.1827$. See R code below.

```
se.that.str <- sqrt(30 * (30-6) * sd(branches.str1) / 6 +
                      25 * (25-6) * sd(branches.str2) / 6 +
                      26 * (26-4) * sd(branches.str3) / 4 +
                      29 * (29-6) * sd(branches.str4) / 6)
se.that.str

## [1] 51.2
```

(g) A 95% t-based confidence interval for t_{STR} is $(5823, 6039)$. See R code below. We are 95% confident that the true total number of apples on all apple-bearing branches of these four trees is between 5823 and 6039 apples.

```
tstar.str <- qt(0.975, 22-4)
ci.str <- c(that.str-tstar.str*se.that.str, that.str+tstar.str*se.that.str)
ci.str

## [1] 5824 6039
```

(h) At a 5% significance level, there is evidence that the true total number of apples on all apple-bearing branches of these four trees is different than 5500 apples. We make this conclusion because the 95% confidence interval for t_{STR} does not contain 5500.

(i) Yes, stratification did seem to improve the estimation process because the variance of the total estimate found using the stratified SRS is less than the variance of the total estimate found using SRS. As a result, the confidence interval for t_{STR} is narrower than the confidence interval for t_{SRS}.

```
ci[2]-ci[1]

## [1] 340

ci.str[2]-ci.str[1]

## [1] 215
```

2. (a) A percentile-method 95% confidence interval for t_{SRS} is $(5045, 6380)$. A percentile-method 95% confidence interval for t_{STR} is $(5490, 5930)$. The output is shown below, and the R code for this section is shown in the R code appendix.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = branches.srs, statistic = samptotal, R = Brep)


Bootstrap Statistics :
    original  bias    std. error
t1*     5690  -0.392         345
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = boottotal, conf = 0.95)

Intervals :
Level      Normal              Basic
95%   (5014, 6367 )   (5000, 6335 )

Level     Percentile            BCa
95%   (5045, 6380 )   (5095, 6440 )
Calculations and Intervals on Original Scale
```

```
STRATIFIED BOOTSTRAP


Call:
boot(data = branches.str, statistic = samptotal, R = Brep, strata = stratum)


Bootstrap Statistics :
    original  bias    std. error
t1*     5710   -0.55         113
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = boottotal, conf = 0.95)

Intervals :
Level      Normal              Basic
95%   (5489, 5932 )   (5490, 5930 )

Level     Percentile            BCa
95%   (5490, 5930 )   (5490, 5930 )
Calculations and Intervals on Original Scale
```

(b) The bootstrap standard error of $\hat{t}_{SRS}$ is 345.1742. A t-based 95% confidence interval for t_{SRS} is then $5690 \pm 2.0796 * 345.1742$, or $(4972, 6408)$ after rounding. The bootstrap standard error of $\hat{t}_{STR}$ is 113.0012. A t-based 95% confidence interval for t_{STR} is then $5710 \pm 2.1009 * 113.0012$, or $(5472, 5948)$ after rounding. The code is shown below. Recall that `that` and `that.str` were defined in problem 1.

```
ci.tbootsrs <- c(5690-tstar*345.1742, 5690+tstar*345.1742)
ci.tbootstr <- c(5710-tstar.str*113.0012, 5710+tstar.str*113.0012)
```

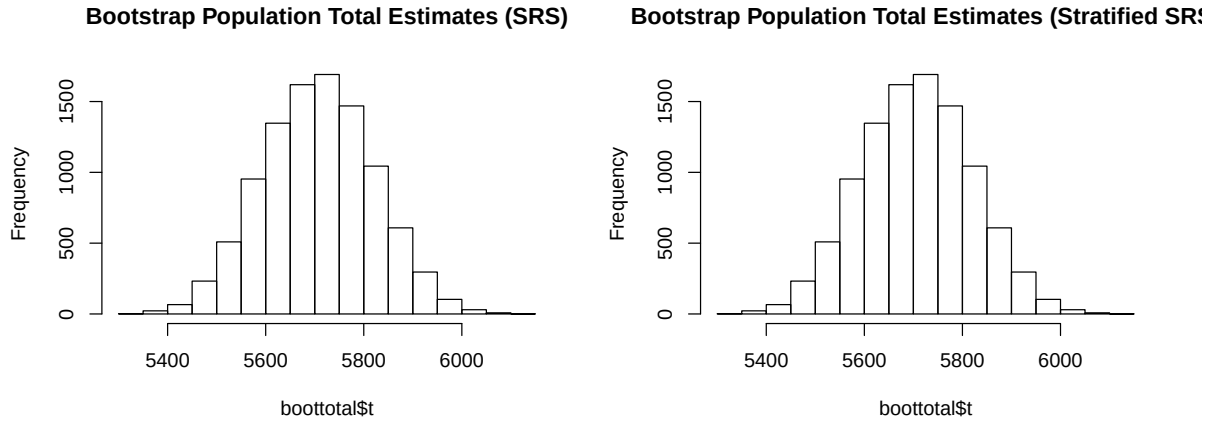
I'd like to point out that the estimate for t_{STR} found by the `boot` function is different than the estimate for t_{STR} we found in problem 1 part (e). The reason for this is because the `boot` function doesn't take into account the fact that we know the true stratum sizes in the population (it only knows the sample sizes that we took from each stratum). So, the program estimates the true stratum sizes based on the number of samples taken from each stratum. For example, 6 apples were chosen from stratum 1, so the program estimates the true stratum size to be $110 * 6/22 = 30$. The R code shown below explains

how the `boot` function arrived at 5710 as the estimate for t_{STR} (rather than 5931 found in part 1(e). I think it would be good to investigate how to incorporate the true stratum sizes, when they are known, into the `boot` function.

```
110*(6*mean(branches.str1)+6*mean(branches.str2)+4*mean(branches.str3)+
       6*mean(branches.str4))/22

## [1] 5710
```

(c) Yes, I think the t-based confidence interval is a reasonable alternative to the percentile-based confidence interval because the bootstrap distributions of totals are mostly symmetric for both the SRS and the stratified SRS sampling plans (see below). I also noticed that the intervals obtained via the percentile method are similar to the t-based intervals found above, which is another indication that the t-based interval is a reasonable alternative to the percentile-based confidence interval.



3. (a) $\widehat{\bar{y}_U} = 12.9297$. See my work below.

```
ybar1 <- 175/25
ybar2 <- 400/25
ybar3 <- 375/16
ybar4 <- 176/16
ybarhat <- (300*ybar1+300*ybar2+100*ybar3+100*ybar4)/800
ybarhat

## [1] 12.9
```

(b) The standard error of $\widehat{\bar{y}_U}$ is 1.2990. See my work below.

```
var.1 <- 300*(300-25)*100/25
var.2 <- 300*(300-25)*100/25
var.3 <- 100*(100-16)*400/16
var.4 <- 100*(100-16)*400/16
var.ybarhat <- 1/(800^2)*(var.1+var.2+var.3+var.4)
se.ybarhat <- sqrt(var.ybarhat)
se.ybarhat

## [1] 1.3
```

(c) A 95% confidence interval for $\bar{y}_U$ is $(10.3435, 15.516)$. My work is shown below.

```
tstar <- qt(0.975, 82-4)
ci.ybar <- c(ybarhat-tstar*se.ybarhat, ybarhat+tstar*se.ybarhat)
ci.ybar

## [1] 10.3 15.5
```

(d) If proportional allocation had been used, the sample sizes for strata 1 and 2 would have been 31 $(300/800 * 82)$. The sample sizes for strata 3 and 4 would have been 10 $(100/800 * 82)$.

(e) If optimum allocation had been used, the sample sizes for strata 1 and 2 would be 25, and the sample sizes for strata 3 and 4 would be 16. Clearly optimum allocation was used. See my work below.

```
den <- 300*10+300*10+100*20+100*20
n12 <- 82*300*10/den
n12

## [1] 24.6

n34 <- 82*100*20/den
n34

## [1] 16.4
```

4. We solve for n in the following equation:

$$n = \frac{1}{1/n_0 + 1/N}$$

where $n_0 = \frac{z^2 p(1-p)}{d^2}$. Since we have no prior estimate for p, I will use $p = 0.5$ to be conservative.

I will assume that the α level is 0.05. Solving the equation, I find that a sample size of 305 is needed so that $\hat{p}$ will be within 0.04 of p with probability at least 0.95. The R-code for the calculations is in the appendix.

5. (a) $\bar{y}_U = 3$ and $S^2 = 13.5$. My work is shown in the R code below.

```
y <- c(0, 3, 9, 3, 0)
ybarU <- mean(y)
ybarU

## [1] 3

Ssq <- var(y)
Ssq

## [1] 13.5
```

   (b) The expected value of $\bar{y}$ is the sum of the observed $\bar{y}$ for each sample times the probability of that sample. $E[\bar{y}] = 2.6$. My work is shown below.

```
y1 <- c(1,2,3)
y2 <- c(2,3,4)
y3 <- c(1,3,5)
e.ybar <- mean(y1)*0.4 + mean(y2)*0.4 + mean(y3)*0.2
e.ybar

## [1] 2.6
```

   (c) $V[\bar{y}] = \sum_S P(S)(\bar{y}_s - 2.6)^2$. $V[\bar{y}] = 0.24$, and my work is shown below.

```
var.ybar <- 0.4*(mean(y1)-e.ybar)^2+0.4*(mean(y2)-e.ybar)^2+0.2*(mean(y3)-e.ybar)^2
var.ybar

## [1] 0.24
```

   (d) $Bias[\bar{y}] = E[\bar{y}] - 3 = 2.6 - 3 = -0.4$.

6. (a) The probability that I am selected to be in the sample is $1000/100000000 = 0.00005$.

(b) The probability that I am not in any of the 2000 samples is $(1 - 0.00005)^{2000} = 0.9048$.

(c) The probability of being in at least one sample is 1 minus the probability that you are not in any of the samples. We solve for x in the following equation.

$$P(atleastone) = 1 - (1 - 0.00005)^x = 0.5$$

$$(1 - 0.00005)^x = 0.5$$

$$x = ln(0.5)/ln(1 - 0.00005)$$

$$x = 13862.6$$

So, $13,863$ samples must be selected for me to have a 0.5 probability of being in at least one sample.

R code appendix

```
require(boot)
set.seed(99)

#branches.srs (vector of responses) defined in previous problem

Brep = 10000
N=110

samptotal <- function(y, i) N*mean(y[i])

boottotal <- boot(data=branches.srs, statistic=samptotal, R=Brep)
boottotal
boot.ci(boottotal, conf=0.95)
```

```
set.seed(99)
branches.str <- c(branches.str1, branches.str2, branches.str3, branches.str4)
stratum <- c(rep(1, 6), rep(2, 6), rep(3, 4), rep(4, 6))

boottotal <- boot(data=branches.str, statistic=samptotal, strata=stratum, R=Brep)
boottotal
boot.ci(boottotal, conf=0.95)
```

```
n0 <- 1.96*0.5*0.5/(0.04^2)
n <- 1/((1/n0)+1/41660)
```