Satoshi Ido

3488706

26 February 2023

STAT 522 HW5

1.

(a) Draw a stratified sample of 200 objects from the population with proportional allocation, stratified by shape. Answer the questions below for the stratified sample. These questions are like (c) – (d) of Homework 2.

```
# Create sample design
STR_sample <- svydesign(id = ~1, strata = ~STR$shape,
                        weights = ~STR$sampwt,
                        fpc = ~STR$popsize,
                        data = STR)
```

(b) Using the sample, estimate the average area for objects in the bin. Give a 95% CI.

The mean is 27.78

The 95% CI of the estimate is [25.91808, 29.64192].

HW5: Stratified area mean

```
> print(STR_area_mean)
      mean     SE
area 27.78 0.9442
> confint(STR_area_mean,  df = 198)
        2.5 %   97.5 %
area 25.91808 29.64192
```

HW3: SRS area mean

```
> print(SRS_area_mean)
      mean     SE
area 28.99 1.1429
> # Confidence Intervals for Model Parameters
> confint(SRS_area_mean, df = 199)
        2.5 %   97.5 %
```

```
area 26.73618 31.24382
```

a). How does the 95% CI for this sample compare to the 95% CI from your SRS sample results in Homework 3?

The CI for this sample is underestimated than that of the SRS sample from homework 3.

(c) Using the sample, estimate the total number of gray objects in the population, along with the 95% CI.

The total number of gray objects is 7200, and the 95% CI is [5919.006, 8489.994].

HW5: Stratified average number of gray objects

```
> print(STR_gray_total)
           total     SE
colorblack 12800 654.15
colorgray   7200 654.15
> confint(STR_gray_total, df = 198)
               2.5 %    97.5 %
colorblack 11510.006 14089.994
colorgray   5910.006  8489.994
```

HW3: SRS number of gray objects

```
> print(SRS_gray_total)
           total     SE
colorblack 13000 672.84
colorgray   7000 672.84
> confint(SRS_gray_total, df = 199)
               2.5 %    97.5 %
colorblack 11673.189 14326.811
colorgray   5673.189  8326.811
```

a. How does the 95% CI for this sample compare to the 95% CI from your SRS sample results in Homework 3?

The CI for this sample([5910, 8489]) has a smaller range than that of the SRS sample from HW3 ([4673, 8326]). However, the stratification does not have much smaller variance, meaning the stratification method is not necessarily effective.

2.

(a) Take an SRS of 150 players from the file.

```
     SRS2_sample <- svydesign(ids = ~1,
                              weights = ~SRS2$sampwt,
                              fpc = rep(797, 150),
                              data = SRS2)
```

(b) Take a stratified random sample of 150 players from the file, using proportional allocation with the different teams as strata.

```
STR2_sample <- svydesign(id = ~1, strata = ~STR2$team,
                         weights = ~STR2$sampwt,
                         fpc=~STR2$popsize,
                         data = STR2)
```

(c) For each sample, estimate the proportion of players who are pitchers and give a 95% CI.

The SRS estimation of the proportion for a pitcher is 0.520 and the 95% CI is [0.44713114, 0.59286886] (=0.14573772).

```
> print(SRS2_pitchers_mean)
          mean     SE
posP  0.520000 0.0369
> confint(SRS2_pitchers_mean, df = 149)
           2.5 %     97.5 %
posP  0.44713114 0.59286886
```

The stratified estimation of the proportion for a pitcher is 0.448683, and the 95% CI is [0.373065636, 0.52429948] (=0.151233844).

```
> print(STR2_pichers_mean)
          mean     SE
```

```
posP  0.448683 0.0382
> confint(STR2_pichers_mean, df = 120) # df = 150(sample size) - 30(number of
strata)
            2.5 %     97.5 %
posP  0.373065636 0.52429948
```

a. How do the estimates compare across the two samples? How do the 95% CI's compare?

It is clear that the estimate with stratification adjusts the overestimation of SRS estimate. However, the stratification does not have much smaller variance (0.145 vs 0.151), meaning the stratification method is not necessarily effective.

(d) For each sample, estimate the mean of the `logsal` variable and give a 95% CI.

a. How do the estimates compare across the two samples? How do the 95% CI's compare?

```
> print(SRS2_logsal_mean)
confint(SRS2_logsal_mean, df = 149)
         mean     SE
logsal 13.788 0.0836
> confint(SRS2_logsal_mean, df = 149)
          2.5 %   97.5 %
logsal 13.62322 13.95355
```

```
> print(STR2_logsal_mean)
         mean     SE
logsal 13.837 0.0873
> confint(STR2_logsal_mean, df = 120)
          2.5 %   97.5 %
logsal 13.66432 14.00984
```

The estimates with SRS of `*logsal*` is 13.788 and 95% CI is [13.62322, 13.95355]. Meanwhile, the estimate with a stratification is 13.838 and 95% CI is [13.66432, 14.00984].

The mean of `*logsal*` with a stratification is 13.838 which is slightly higher than that of SRS (=13.788), and CI of a stratification is [13.66432, 14.00984] while that of SRS is [13.62322, 13.95355]. By just comparing the results from two approaches, it is statistically unclear if the stratification approach creates any adjustment.

(e) Examine the variances of `logsal` in each stratum using the population data. Do you think optimal allocation would be worthwhile for this problem? Why or why not?

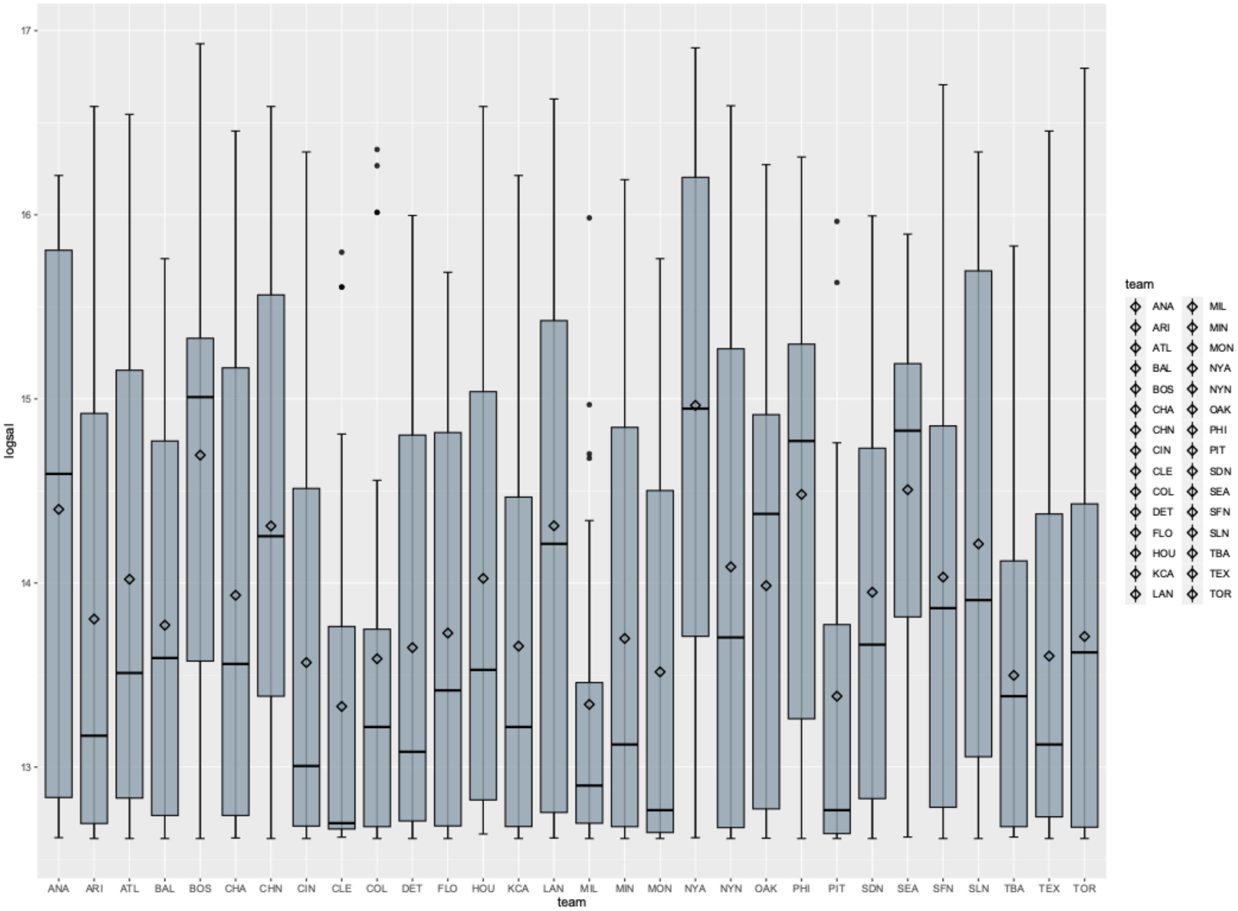
Yes. It is worth taking the stratification approach to apply an optimal allocation to the sampling since each strata, which is the baseball team, has various means (=`*logsal*`) and variances. Stratification improves precision by creating subpopulations (=strata) to incorporate variations pertaining to stratum within their stratum respectively (=baseball team).

Below is the variance of each stratum. In this case, strata represent baseball teams.

```
> print(var_team, n = 30)
# A tibble: 30 × 4
   team   mean   var    sd
   <chr> <dbl> <dbl> <dbl>
 1 ANA    14.4 1.99  1.41
 2 ARI    13.8 1.75  1.32
 3 ATL    14.0 1.92  1.38
 4 BAL    13.8 1.24  1.11
 5 BOS    14.7 1.61  1.27
 6 CHA    13.9 1.64  1.28
 7 CHN    14.3 1.46  1.21
 8 CIN    13.6 1.33  1.15
 9 CLE    13.3 1.00  1.00
10 COL    13.6 1.48  1.22
11 DET    13.6 1.39  1.18
12 FLO    13.7 1.19  1.09
13 HOU    14.0 1.87  1.37
14 KCA    13.7 1.25  1.12
15 LAN    14.3 1.72  1.31
16 MIL    13.3 0.843 0.918
17 MIN    13.7 1.46  1.21
18 MON    13.5 1.24  1.11
19 NYA    15.0 1.85  1.36
20 NYN    14.1 1.81  1.35
21 OAK    14.0 1.36  1.17
22 PHI    14.5 1.59  1.26
```

```
23 PIT    13.4 0.930 0.964
24 SDN    13.9 1.26  1.12
25 SEA    14.5 1.07  1.04
26 SFN    14.0 1.44  1.20
27 SLN    14.2 1.75  1.32
28 TBA    13.5 0.741 0.861
29 TEX    13.6 1.10  1.05
30 TOR    13.7 1.19  1.09
```

Below is the boxplot showing the distribution of data and any outliers


logsal
ANA
ARI
ATL
BAL
BOS
CHA
CHN
CIN
CLE
COL
DET
FLO
HOU
KCA
LAN
MIL
MIN
MON
NYA
NYN
OAK
PHI
PIT
SDN
SEA
SFN
SLN
TBA
TEX
TOR
team
team
ANA
ARI
ATL
BAL
BOS
CHA
CHN
CIN
CLE
COL
DET
FLO
HOU
KCA
LAN
MIL
MIN
MON
NYA
NYN
OAK
PHI
PIT
SDN
SEA
SFN
SLN
TBA
TEX
TOR