Stat 138: Introduction to Sampling Designs
Problem Set 1

Anne Christine Amores

March 12, 2025

Problems

1. An SRS of size 30 is taken from a population of size 100. The sample values are given below, and in the data file srs30.dat.

8 5 2 6 6 3 8 6 10 7 15 9 15 3 5 6 7 10 14 3 4 17 10 6 14 12 7 8 12 9

a. What is the sampling weight for each unit in the sample?

Under SRSWOR, the probability of inclusion is

$$\pi_i = \frac{n}{N} = \frac{30}{100} \qquad i = 1,\ 2,\ ...,\ 30$$

Thus, the sampling weight for each unit in the sample is

$$w_i = \frac{1}{\pi_i} = \frac{100}{30} \approx \boxed{3.3333}.$$

b. Use the sampling weights to estimate the population total, t.

$$\hat{t} = \sum_{i \in S} w_i y_i = \frac{100}{30} \sum_{i \in S} y_i = \frac{100}{30}(8 + 5 + 2 + 6 + 6 + 3 + ... + 12 + 9) \approx \boxed{823.3333}$$

c. Give a 95% CI for t. Does the fpc make a difference for this sample?

For the population total t, an approximate 95% CI is given by

$$\left[\ \hat{t} - t_{0.025,n-1} SE(\hat{t}),\ \ \hat{t} + t_{0.025,n-1} SE(\hat{t})\ \right]$$

$$= \left[\hat{t} - t_{0.025,29}\sqrt{N^2(1-\frac{n}{N})\frac{s_y^2}{n}},\ \hat{t} + t_{0.025,29}\sqrt{N^2(1-\frac{n}{N})\frac{s_y^2}{n}} \right]$$

$$= \left[823.3333 - 2.045\sqrt{100^2(1-\frac{30}{100})\frac{s_y^2}{30}},\ 823.3333 + 2.045\sqrt{100^2(1-\frac{30}{100})\frac{s_y^2}{30}} \right],$$

$$\text{where } s_y^2 = \frac{\frac{1}{30-1}\sum_{i\in S}(y_i - 8.2333)^2}{30}.$$

$$= \boxed{[\ 698.4670,\ 948.1996\ ]}$$

Ignoring the fpc, the resulting approximate 95% CI is given by

$$= \left[823.3333 - 2.045\sqrt{100^2\frac{s_y^2}{30}},\ 823.3333 + 2.045\sqrt{100^2\frac{s_y^2}{30}} \right],$$

$$\text{where } s_y^2 = \frac{\frac{1}{30-1}\sum_{i\in S}(y_i - 8.2333)^2}{30}.$$

$$= \boxed{[\ 674.0896,\ 972.5770\ ]}$$

Thus, the fpc does make a difference in this case, as it resulted in a narrower confidence interval.

---

2. The percentage of patients overdue for a vaccination is often of interest for a medical clinic. Some clinics examine every record to determine that percentage; in a large practice though, taking a census of the records can be time-consuming. Cullen (1994) took a sample of the 580 children served by an Auckland family practice to estimate the proportion of interest.

a. What sample size in an SRS (without replacement) would be necessary to estimate the proportion with 95% confidence and margin of error 0.10?

The desired precision of the estimate of the proportion is expressed as

$$P\left[|\hat{p} - p| \leq e\right] = 1 - \alpha$$

$$\Rightarrow P\left[-e \leq \hat{p} - p \leq e\right] = 1 - \alpha$$

$$\Rightarrow P\left[\frac{-e}{\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}} \leq \frac{\hat{p}-p}{\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}} \leq \frac{e}{\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}}\right] = 1-\alpha$$

$$\Rightarrow z_{\frac{\alpha}{2}} = \frac{e}{\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}}$$

Solving for n from the "mother equation",

$$\Rightarrow n = \frac{N}{\frac{e^2}{z^2_{\alpha/2}\frac{p(1-p)}{N-1}} + 1}$$

Since we do not know the value of p, let us use the value of p that will maximize the sample size, i.e., $p = 0.5$.

$$\Rightarrow n = \frac{580}{\frac{0.1^2}{1.96^2\frac{0.5(1-0.5)}{580-1}} + 1}$$

$$\Rightarrow n = 82.51836928$$

$$\Rightarrow n \approx \boxed{83}$$

b. Cullen actually took an SRS with replacement of size 120, of whom 27 were *not* overdue for vaccination. Give a 95% CI for the proportion of children not overdue for vaccination.

Since 27 out of the 120 children in the sample were not overdue for vaccination, $\hat{p} = \frac{27}{120} = 0.225$.

And since SRS was done with replacement, we do not need to use the fpc.

An approximate 95% CI for the proportion of children not overdue for vaccination is given by

$$[\ \hat{p} - z_{0.025}SE(\hat{p}),\ \hat{p} + z_{0.025}SE(\hat{p})\ ]$$

$$= \left[\ \hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ \right]$$

$$= \left[\ 0.225 - 1.96\sqrt{\frac{0.225(1-0.225)}{120}},\ 0.225 + 1.96\sqrt{\frac{0.225(1-0.225)}{120}}\ \right]$$

$$= \boxed{[\,0.1503,\ 0.2997]}$$

3. The Special Census of Maricopa County, Arizona, gave 1995 populations for the following cities:

City	Population
Buckeye	4,857
Gilbert	59,338
Gila Bend	1,724
Phoenix	1,149,417
Tempe	153,821

Suppose that you are interested in estimating the percentage of persons who have been immunized against polio in each city and can take an SRS of persons. What should your sample size be in each of the 5 cities if you want the estimate from each city to have margin of error of 4 percentage points? For which cities does the finite population correction make a difference?

Since we do not have an approximation of p, let us use $p = 0.5$. And since α was not specified, let us use $\alpha = 0.05$.

$$P\left[|\hat{p} - 0.5| \leq 0.04\right] = 1 - 0.05$$

$$\Rightarrow P\left[-0.04 \leq \hat{p} - p \leq 0.04\right] = 0.95$$

Ignoring the fpc,

$$\Rightarrow P\left[\frac{-0.04}{\sqrt{\frac{N}{N-1}\frac{p(1-p)}{n}}} \leq \frac{\hat{p} - 0.5}{\sqrt{\frac{N}{N-1}\frac{p(1-p)}{n}}} \leq \frac{0.04}{\sqrt{\frac{N}{N-1}\frac{p(1-p)}{n}}}\right] = 1 - \alpha$$

$$\Rightarrow z_{0.025} = \frac{0.04}{\sqrt{\frac{N}{N-1}\frac{p(1-p)}{n}}}$$

Solving for n from the "mother equation",

$$\Rightarrow n = \frac{z_{0.025}^2 \frac{N}{N-1} p(1-p)}{0.04^2}$$

Computing the sample size for each city **WITHOUT** the fpc,

Buckeye

$$n = \frac{1.96^2 \frac{4,857}{4,857-1} 0.5(1-0.5)}{0.04^2} = 600.37361 \approx \boxed{601}$$

Gilbert

$$n = \frac{1.96^2 \frac{59,338}{59,338-1} 0.5(1-0.5)}{0.04^2} = 600.2601159 \approx \boxed{601}$$

Gila Bend

$$n = \frac{1.96^2 \frac{1,724}{1,724-1} 0.5(1-0.5)}{0.04^2} = 600.5983749 \approx \boxed{601}$$

Phoenix

$$n = \frac{1.96^2 \frac{1,149,417}{1,149,417-1} 0.5(1-0.5)}{0.04^2} = 600.2505222 \approx \boxed{601}$$

Tempe

$$n = \frac{1.96^2 \frac{153,821}{153,821-1} 0.5(1-0.5)}{0.04^2} = 600.2539023 \approx \boxed{601}$$

Computing the sample size for each city **WITH** the fpc,

Buckeye

$$n = \frac{N}{\frac{e^2}{z_{\alpha/2}^2 \frac{p(1-p)}{N-1}} + 1} = \frac{4,857}{\frac{0.04^2}{1.96^2 \frac{0.5(1-0.5)}{4,857-1}} + 1} = 534.3256357 \approx \boxed{535}$$

Gilbert

$$n = \frac{59,338}{\frac{0.04^2}{1.96^2 \frac{0.5(1-0.5)}{59,338-1}} + 1} = 594.2487268 \approx \boxed{595}$$

<u>Gila Bend</u>

$$n = \frac{1,724}{\frac{0.04^2}{1.96^2 \frac{0.5(1-0.5)}{1,724-1}} + 1} = 445.4238674 \approx \boxed{446}$$

<u>Phoenix</u>

$$n = \frac{1,149,417}{\frac{0.04^2}{1.96^2 \frac{0.5(1-0.5)}{1,149,417-1}} + 1} = 599.937222 \approx \boxed{600}$$

<u>Tempe</u>

$$n = \frac{153,821}{\frac{0.04^2}{1.96^2 \frac{0.5(1-0.5)}{153,821-1}} + 1} = 597.9206435 \approx \boxed{598}$$

Summary of Sample Sizes

City	Sample Size without FPC	Sample Size with FPC
Buckeye	601	535
Gilbert	601	595
Gila Bend	601	446
Phoenix	601	600
Tempe	601	598

Looking at the table, the fpc only made a noticeable difference for **Buckeye** and **Gila Bend**, significantly decreasing the required sample size to achieve the same precision.

4. *Forest data.* The data in file forest.dat consist of a subset of the measurements from 581,012 30x30m cells from Region 2 of the U.S. Forest Service Resource information System. The original data were used in a data mining application, predicting forest cover type from covariates. Data-mining methods are often used to explore relationships in very large data sets; in many cases, the data sets are so large that statistical software packages cannot analyze them. Many data-mining problems, however, can be alternatively approached by analyzing probability samples from the population. In these exercises, we treat forest.dat as a population.

a. Select an SRS of size 2000 from the 581,012 records.

Importing the dataset and renaming the columns,

```
library(readxl)
forest <- read_excel("C:/Users/amore_6ou078y/Downloads/forest.xlsx",
    col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
```

```
View(forest)
colnames(forest) <- c("elevation", "Aspect", "Slope", "Horiz", "Vert",
        "HorizRoad", "Hillshade_9am", "Hillshade_Noon", "Hillshade_3pm",
                "HorizFire", "Wilderness1", "Wilderness2", "Wilderness3",
                        "Wilderness4", "Cover")
head(forest)
```

```
## # A tibble: 6 x 15
##   elevation Aspect Slope Horiz  Vert HorizRoad Hillshade_9am Hillshade_Noon
##       <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>         <dbl>          <dbl>
## 1      2596     51     3   258     0   510           221            232
## 2      2590     56     2   212    -6   390           220            235
```

```
## 3      2804    139     9   268    65  3180            234             238
## 4      2785    155    18   242   118  3090            238             238
## 5      2595     45     2   153    -1   391            220             234
## 6      2579    132     6   300   -15    67            230             237
## # i 7 more variables: Hillshade_3pm <dbl>, HorizFire <dbl>, Wilderness1 <dbl>,
## #   Wilderness2 <dbl>, Wilderness3 <dbl>, Wilderness4 <dbl>, Cover <dbl>
```

Obtaining an SRS of size 2000,

```r
set.seed(10)
srs_forest <- forest[sample(nrow(forest), size = 2000, replace = FALSE), ]
head(srs_forest)
```

```
## # A tibble: 6 x 15
##   elevation Aspect Slope Horiz  Vert Horiz Hillshade_9am Hillshade_Noon
##       <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>         <dbl>          <dbl>
## 1      2840     20     6    42     6   566           216            228
## 2      2690     95    11     0     0  1605           238            223
## 3      2759     22    17     0     0   752           207            200
## 4      3140     51    27   400   219  1981           222            172
## 5      3170     29     6    30     1  1288           218            226
## 6      2780    148    16    60    -3  3416           240            237
## # i 7 more variables: Hillshade_3pm <dbl>, HorizFire <dbl>, Wilderness1 <dbl>,
## #   Wilderness2 <dbl>, Wilderness3 <dbl>, Wilderness4 <dbl>, Cover <dbl>
```

b. Using your SRS, estimate the percentage of cells in each of the 7 forest cover types, along with 95% CIs.

Computing for $\hat{p}$ for all 7 forest cover types,

```r
n <- nrow(srs_forest) # Sample size
p_hat1 <- sum(srs_forest$Cover == 1)/n
p_hat2 <- sum(srs_forest$Cover == 2)/n
p_hat3 <- sum(srs_forest$Cover == 3)/n
p_hat4 <- sum(srs_forest$Cover == 4)/n
p_hat5 <- sum(srs_forest$Cover == 5)/n
p_hat6 <- sum(srs_forest$Cover == 6)/n
p_hat7 <- sum(srs_forest$Cover == 7)/n
p_hat_table <- data.frame(
  Cover_Type = c("Spruce/Fir", "Lodgepole Pine", "Ponderosa Pine"
            , "Cottonwood/Willow", "Aspen", "Douglas-fir", "Krummholz"),
  p_hat = c(p_hat1, p_hat2, p_hat3, p_hat4, p_hat5, p_hat6, p_hat7))
```

Getting an approximate 95% CI for cover type proportions,

```
z_alpha <- 1.96
p_hat_table$Lower_CI <- p_hat_table$p_hat - z_alpha *sqrt((p_hat_table$p_hat
        * (1 - p_hat_table$p_hat)) / n)
p_hat_table$Upper_CI <- p_hat_table$p_hat + z_alpha * sqrt((p_hat_table$p_hat
        * (1 - p_hat_table$p_hat)) / n)
print(p_hat_table)
```

```
##          Cover_Type  p_hat     Lower_CI    Upper_CI
## 1        Spruce/Fir 0.3655 3.443943e-01 0.386605740
## 2    Lodgepole Pine 0.4965 4.745871e-01 0.518412929
## 3    Ponderosa Pine 0.0645 5.373429e-02 0.075265714
## 4 Cottonwood/Willow 0.0020 4.196098e-05 0.003958039
## 5             Aspen 0.0125 7.630721e-03 0.017369279
## 6       Douglas-fir 0.0270 1.989639e-02 0.034103614
## 7         Krummholz 0.0320 2.428646e-02 0.039713540
```

c. Estimate the average elevation in the population, with 95% CI.

Getting an estimate of the average elevation in the population and computing for a 95% CI,

```
mean_elevation <- mean(srs_forest$elevation)
se_elevation <- sd(srs_forest$elevation) / sqrt(nrow(srs_forest))
z_alpha <- 1.96
lower_CI <- mean_elevation - z_alpha * se_elevation
upper_CI <- mean_elevation + z_alpha * se_elevation
cat(paste("A 95%% CI for the average elevation in the population is (",
            round(lower_CI, 4), ",", round(upper_CI, 4), ")", sep = ""))
```

```
## A 95%% CI for the average elevation in the population is (2950.3235,2974.8495)
```