# Homework 1, STAT 472/572

**Due:** Mar 02/05/24 Monday

**Problem 1.4** Potential jurors in some jurisdictions are chosen from a list of country residents who are registered voters or licensed drivers over age 18. In the fourth quarter of 1994, $100,300$ jury summons were mailed to Maricopa County, Arizona, residents. Approximately $23,000$ of those were returned from the post office as undeliverable. Approximately 7000 persons were unqualified for service because they were not citizens, were under 18, were convicted felons, or other reason that disqualified them from serving on a jury. An additional $22,000$ were excused from jury service because of illness, financial hardship, military service, or other acceptable reason. The final sample consists of persons who appear for jury duty; some unexcused jurors fail to appear.

Describe the target population, sampling frame, sampling unit, and observation unit. Discuss any possible sources of selection bias or inaccuracy of responses.

**Problem 1.7** A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.

Describe the target population, sampling frame, sampling unit, and observation unit. Discuss any possible sources of selection bias or inaccuracy of responses.

**Problem 2.1** Let $N = 6$ and $n = 3$. For purposes of studying sampling distributions, assume that all population values are known.

$$\begin{aligned} y_1 &= 98, \qquad y_2 = 102, \qquad y_3 = 154, \\ y_4 &= 133, \qquad y_5 = 190, \qquad y_6 = 175. \end{aligned}$$

We are interested in $\bar{y}_U$, the population mean. Two sampling plans are proposed.

- Plan 1. Eight possible samples may be chosen.

| Sample Number | Sample, $\mathcal{S}$ | $P(\mathcal{S})$ |
|---|---|---|
| 1 | $\{1,3,5\}$ | 1/8 |
| 2 | $\{1,3,6\}$ | 1/8 |
| 3 | $\{1,4,5\}$ | 1/8 |
| 4 | $\{1,4,6\}$ | 1/8 |
| 5 | $\{2,3,5\}$ | 1/8 |
| 6 | $\{2,3,6\}$ | 1/8 |
| 7 | $\{2,4,5\}$ | 1/8 |
| 8 | $\{2,4,6\}$ | 1/8 |

- Plan 2. Three possible samples may be chosen.

| Sample NUmber | Sample, $\mathcal{S}$ | $P(\mathbb{S})$ |
|---|---|---|
| 1 | $\{1,4,6\}$ | 1/4 |
| 2 | $\{2,3,6\}$ | 1/2 |
| 3 | $\{1,3,5\}$ | 1/4 |

a. What is the value of $\bar{y}_U$?

b. Let $\bar{y}$ be the mean of the sample values. For each sampling plan, find $E[\bar{y}]$, $Var(\bar{y})$, $Bias(\bar{y})$ and $MSE(\bar{y})$.

c. Which sampling plan do you think is better? Why?

**2.5.** 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?

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

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

**Problem 2.18** n 2005, the Statistical Society of Canada (SSC) had 864 members listed in the online directory. An SRS of 150 of the members was selected; the sex and employment category (industry, academic, government) was ascertained for each person in the SRS, with results in file ssc.dat.

a. What are the possible causes of selection bias in this sample?

b. Estimate the percentage of members who are female, and give a 95% CI for your estimate.

c. Assuming that all members are listed in the online directory, estimate the total number of SSC members who are female, along with a 95% CI.

**Problem 2.19** The data set agsrs.dat also contains information on other variables. For each of the following quantities, plot the data, and estimate the population mean for that variable along with a 95% CI.

a. Number of acres devoted to farms in 1987.

b. Number of farms, 1992.

c. Number of farms with 1000 acres or more, 1992. (use variable largef92)

d. Number of farms with 9 acres or fewer, 1992. (use variable smallf92)

**Problem 2.30 extra credits** In an SRSWR, a population unit can appear in the sample anywhere between 0 and $n$ times. Let

$$Q_i = \text{number of times unit } i \text{ appears in the sample.}$$

and

$$\hat{t} = \frac{N}{n} \sum_{i=1}^{N} Q_i y_i.$$

a. Argue that the joint distribution of $Q_1, Q_2, \ldots, Q_N$ is multinomial with $n$ trials and $p_1 = p_2 = \cdots = p_N = 1/N$.

b. Using (a) and properties of the multinomial distribution, show that $E[\hat{t}] = t$.

c. Using (a) and properties of the multinomial distribution, find $Var(\hat{t})$.

# Homework 2, STAT 472/572

**Due:** Mar 02/19/24 Monday

**Problem 1 (textbook chapter 3, problem 3).** Consider a population of 6 students. Suppose we know the test scores of the students to be

| Student | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Score | 66 | 59 | 70 | 83 | 82 | 71 |

(a) Find the mean $\bar{y}_U$ and variance $S^2$ of the population.

(b) How many SRS's of size 4 are possible?

(c) List the possible SRS's. For each, find the sample mean. Using Equation (2.9), find $Var(\bar{y})$.

(d) Now let stratum 1 consist of students 1–3, and stratum 2 consist of students 4–6. How many stratified random samples of size 4 are possible in which 2 students are selected from each stratum?

(e) List the possible stratified random samples. Which of the samples from (c) cannot occur with the stratified design?

(f) Find $\bar{y}_{str}$ for each possible stratified random sample. Find $Var(\bar{y}_{str})$, and compare it to $Var(\bar{y})$.

**Problem 2 (textbook chapter 3, problem 5).** The survey in Example 3.4 collected much other data on the subjects. Another of the survey's questions asked whether the respondent agreed with the following statement: "When I look at a new issue of my discipline's major journal, I rarely find an article that interests me." The results are as follows:

| Discipline | Agree (%) |
|---|---|
| Literature | 37 |
| Classics | 23 |
| Philosophy | 23 |
| History | 29 |
| Linguistics | 19 |
| Political Science | 43 |
| Sociology | 41 |

(a) What is the sampled population in this survey?

(b) Find an estimate of the percentage of persons in the sampled population that agree with the statement, and give the standard error of your estimate.

**Problem 3.** Proportional allocation was used in the stratified sample in Example 3.2. It was noted, however, that variability was much higher in the West than in the other regions. Using the estimated variances in Example 3.2, and assuming that the sampling costs are the same in each stratum, find an optimal allocation for a stratified sample of size 300.

**Problem 4.** Set seed as 9231, select a stratified random sample of size 300 from the data in the file agpop.dat, using your allocation in Problem 3. Estimate the total number of acres devoted to farming in the United States, and give the standard error of your estimate. How does this standard error compare with that found in Example 3.2?

# Homework 3, STAT 472/572

**Due:** Mar 03/04/24 Monday

*Chapter 4*

**Problem 1.** Foresters want to estimate the average age of trees in a stand. Determining age is cumbersome, because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter of all 1132 trees and find that the population mean equals 10.3. They then randomly select 20 trees for age measurement.

| Tree No. | Diameter, $x$ | Age, $y$ | Tree No. | Diameter, $x$ | Age, $y$ |
|---|---|---|---|---|---|
| 1 | 12.0 | 125 | 11 | 5.7 | 61 |
| 2 | 11.4 | 119 | 12 | 8.0 | 80 |
| 3 | 7.9 | 83 | 13 | 10.3 | 114 |
| 4 | 9.0 | 85 | 14 | 12.0 | 147 |
| 5 | 10.5 | 99 | 15 | 9.2 | 122 |
| 6 | 7.9 | 117 | 16 | 8.5 | 106 |
| 7 | 7.3 | 69 | 17 | 7.0 | 82 |
| 8 | 10.2 | 133 | 18 | 10.7 | 88 |
| 9 | 11.7 | 154 | 19 | 9.3 | 97 |
| 10 | 11.3 | 168 | 20 | 8.2 | 99 |

(a) Draw a scatterplot of $y$ vs. $x$.

(b) Estimate the population mean age of trees in the stand using ratio estimation and give an approximate standard error for your estimate.

(c) Repeat (b) using regression estimation.

(d) How do the estimators in (b) and (c) compare?

**Problem 2.** The data set *agsrs* also contains information on the number of farms in 1987 for the SRS of $n = 300$ counties from the population of the $N = 3078$ counties in the United States. In 1987, the United States had a total of $2,087,759$ farms.

(a) Plot the data.

(b) Use ratio estimation to estimate the total number of acres devoted to farming in 1992 (variable *acres92*), using the number of farms in 1987 (variable *farms87*) as the auxiliary variable.

(c) Repeat (b), using regression estimation.

(d) Which method gives the most precision: ratio estimation with auxiliary variable *acres87*, ratio estimation with auxiliary variable *farms87*, or regression estimation with auxiliary variable *farms87*? Why?

# Homework 4, STAT 472/572

**Due:** Mar 03/27/24

**Problem 1.** Use the data set *agsrs.dat* and consider *acres92* as the response variable. For problems (a) and (b), define two domains using variable *farms92*:

(a) Counties with fewer than 600 farms.

- Estimate the total number of acres devoted to farming (*acres92*).
- Provide standard errors and construct a 95% confidence interval for the estimator in this domain.

(b) Counties with 600 or more farms.

- Estimate the total number of acres devoted to farming (*acres92*).
- Provide standard errors and construct a 95% confidence interval for the estimator in this domain.

(c) Poststratify the sample *agsrs* into the four census regions NC, NE, S, and W given in Table 3.1. Estimate the population total, standard error and construct a 95% confidence interval for the estimator using the poststratification method.

**Problem 2.** Simulation practice (refer to simulationexample for reference), remove missing values of acres92 from data *agpop* before selecting a random sample
(a) Use data agpop to draw a stratified random sample of size 300 according to optimal allocation (NC: 69; NE: 7; S: 122; W: 102).
(b) Calculate standard error estimator $SE(\hat{t}_{str})$ for variable acres92.
(c) Repeat the above for 1000 times, draw a histogram of the 1000 standard error estimates you got. Record mean and median of the 1000 standard error estimates.

# Homework 5 STAT 472/572

**Due:** Mar 04/15/24

**Problem 1.** A language school owner takes an SRS of 10 of the 72 Introductory Spanish classes offered by the school. Each student in each of the sampled classes is given a vocabulary test and is also asked whether he or she is planning a trip to a Spanish-speaking country in the next year. The data are in file *spanish.dat*.

(a) estimate the total number of students planning a trip to a Spanish-speaking country in the next year, and give a 95% CI (use unbiased estimator).

(b) estimate the mean vocabulary test score for Introductory Spanish students in the language school, and give a 95% CI (use ratio estimator).

**Problem 2.** Gnap (1995) conducted a survey to estimate the teacher workload in Maricopa County, Arizona, public school districts. Her target population was all first through sixth grade full-time public school teachers with at least one year of experience. In 1994, Maricopa County had 46 school districts with 311 elementary schools and 15,086 teachers. Gnap stratified the schools by size of school district; the large stratum, consisting of schools in districts with more than 5000 students, is considered in this exercise. The stratum contained 245 schools; 23 participated in the survey. All teachers in the selected schools were asked to fill out the questionnaire. Due to nonresponse, however, some questionnaires were not returned. The data are in file *teachers*, with psu information in *teachmi*.

You may need to merge data *teachers* and *teachmi*, and remove missing values NA at this point.

```
ex.data<-merge(teachers,teachmi,by=c("dist","school"))
```

(a) Why would a cluster sample be a better design than an SRS for this study? Consider issues such as cost, ease of collecting data, and confidentiality of respondent. What are some disadvantages of using a cluster sample?

(b) Calculate the mean and standard deviation of *hrwork* for each school in the "large" stratum. Construct a graph of the means for each school and a separate graph of the standard deviations. Does there seem to be more variation within a school, or does more of the variability occur between different schools?

(c) Construct a scatterplot of the standard deviations versus the means for the schools, for the variable *hrwork*. Is there more variability in schools with higher workloads? Less? No apparent relation?

(d) Estimate the average of *hrwork* in the large stratum in Maricopa County, along with its standard error. Use *popteach* in *teachmi* for the $M_i$'s.

**Problem 3:** For each of the following situations, say what unit might be used as psu. Do you believe there would be a strong clustering effect? Would you sample psus with equal or unequal probabilities?

a. You want to estimate the percentage of patients of U.S. Air Force optometrists and ophthalmologists who wear contact lenses.

b. Human taeniasis is acquired by ingesting larvae of the pork tapeworm in inadequately cooked pork. You have been asked to design a survey to estimate the percentage of inhabitants of a village who have taeniasis. A medical examination is required to diagnose the condition.

c. You wish to estimate the total number of cows and heifers on all Ontario dairy farms; in addition, you would like to find estimates of the birth rate and stillbirth rate.

**Problem 4:** The file *statepop*, contains information on total number of farms, number of veterans, and other items.

The file statepop.csv contains data from an unequal-probability sample of 100 counties from the 1994 County and City Data Book (U.S. Census Bureau, 1994), selected with probability proportional to population. The selection probabilities are given in variable *psii*. Sampling was done with replacement, so large counties occur multiple times in the sample: Los Angeles County, with the largest population in the United States, occurs four times. Let $t_i$ represent the number of veterans in county $i$ (variable *veterans*).

a. Draw a scatterplot of $t_i$ vs. $\psi_i$ for the counties in the sample. Do you expect pps sampling to work well for estimating the total number of vetarans in the United States?

b. Calculate $u_i = t_i/\psi_i$ for each county in the sample. Calculate the mean $\bar{u}$ and standard deviation $s_u$ for the 100 counties in the sample (use direct calculation by formulas).

c. Using $\bar{u}$ and $s_u$ from (b), calculate a 95% CI for the estimated total number of veterans (use direct calculation by formulas).

d. Check your results by estimating the total number of veterans along with a 95% CI using R.

**Problem 5:** You are asked to design a survey to estimate the total number of cars without permits that park in handicapped parking places on your campus. What variables (if any) would you consider for stratification? For clustering? What information do you need to aid in the design of the survey? Describe a survey design that you think would work well for this situation.

# Homework 6 STAT 472/572

**Due:** Mar 04/29/24

Problem 1. (a) For the supermarket example in Section 6.1, suppose that the $\psi_i$s are the same as in the example, but that each store has $t_i = 75$. What is $E[\hat{t}_\psi]$ and $V(\hat{t}_\psi)$?
(b) comparing the results with those in the example, discuss why there are differences if any.

Problem 2. Use the data in file nhanes for this exercise. The sagittal abdominal diameter (variable bmdavsad), measures the distance from the small of the back to the upper abdomen.
(a) Draw a histogram of bmdavsad, using the weights. Do the data appear to be normally distributed?
(b) Estimate the mean value of bmdavsad for the domain of adults age 20 and over, along with a 95% CI.
(c) Find the minimum, 25th, 50th, and 75th percentiles, and maximum of bmdavsad. Calculate the same quantities separately for each gender (variable riagendr). Use these to construct side-by-side boxplots of the data as in Figure 7.6.
(d) Construct a weighted bubble plot with smoothed trend line for y = bmdavsad and x = bmxbmi. Does there appear to be a linear relationship? What other features do you see in the data?

Problem 3. Kosmin and Lachman (1993) had a question on religious affiliation included in 56 consecutive weekly household surveys; the subject of household surveys varied from week to week from cable TV use, to preference for consumer items, to political issues. After four callbacks, the unit nonresponse rate was 50%; an additional 2.3% refused to answer the religion question. The authors say:

```
Nationally, the sheer number of interviews and careful research
design resulted in a high level of precision ...  Standard error
estimates for our overall national sample show that we can be
95% confident that the figures we have obtained have an error
margin, plus or minus, of less than 0.2%.  This means, for example,
that we are more than 95% certain that the figure for Catholics
is in the range of 25.0% to 26.4% for the U.S. population ...
```

(a) Critique the preceding statement.

(b) If you anticipated item nonresponse, do you think it would be better to insert the question of interest in different surveys each week, as was done here, or to use the same set of additional questions in each survey? Explain your answer. How would you design an experiment to test your conjecture?

Problem 4. Gnap (1995) conducted a survey on teacher workload which was used in Exercise 15 of Chapter 5.

(a) The original survey was intended as a one-stage cluster sample. What was the overall response rate?

(b) Would you expect nonresponse bias in this study? If so, in which direction would you expect the bias to be? Which teachers do you think would be less likely to respond to the survey?

(c) Gnap also collected data on a random subsample of the nonrespondents in the "large" stratum, in file *teachnr.dat*. How do the respondents and nonrespondents differ?

(d) Is there evidence of nonresponse bias, when you compare the subsample of nonrespondents to the respondents in the original survey?

(e) Use an appropriate imputation method for the missing values in *teachnr.dat*, then calculate the means and variances for the four variables again, compare to the results from (c).

(f) Use an appropriate imputation methods for the missing values in *teachers.dat*, then calculate the means and variances for the four variables again, compare to the results from (c).