

Assignment 1

Deadline to hand in: Oct. 6 in class

- Q. 1 For the following survey, describe the target distribution, sampling frame, sampling unit, and observation unit. Discuss any possible source of selection bias or inaccuracy of responses.

The article, "What Readers Say about Marijuana" (Parade, July 31, 1994, p.16) reported "More than 75 % of the readers who took part in an informal PARADE telephone poll say marijuana should be as legal as alcoholic beverages". The telephone poll was announced on page 5 of the June 12 issue; readers were instructed to "Call 1-900-773-1200, at 75 cents a call, if you would like to answer the following questions. Use touch-tone phones only. To participate, call between 8 a.m. EDT [Eastern Daylight Time] on Saturday June 11, and midnight EDT on Wednesday, June 15."

- Q. 2 Let the index set for the population is

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

The values of y_i are

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

Consider the following sampling scheme:

S	P(S)
$\{1,3,5,6\}$	1/8
$\{2,3,7,8\}$	1/4
$\{1,4,6,8\}$	1/8
$\{2,4,6,8\}$	3/8
$\{4,5,7,8\}$	1/8

- (a) Let \bar{y} be the mean of the sample values. Find
 (i) $E[\bar{y}]$ (ii) $V[\bar{y}]$ (iii) $Bias(\bar{y})$ (iv) $MSE[\bar{y}]$
- (b) Find the probability of selection π for each unit i .
- (c) What is the sampling distribution of $\hat{\tau} = 8\bar{y}$?
- Q. 3 A botanical researcher wishes to design a survey to estimate the total number of birch trees in a study area. The study area has been divided into 1000 units or plots. From previous experience, the variance in the numbers of birth trees per plot is known to be approximately 45. Using simple random sampling, what sample size (number of plots) should be used to estimate the total number of trees in the study area to within 500 trees of the true value with 95% confidence? To within 1000 trees? To within 2000 trees?

Q. 4 For this question, use "agpop.csv" file from the Blackboard.

The following R commands will be useful:

```
read.csv(agpop.csv, header=T), set.seed(number), sample(x, n, replace=F)
```

Using R to select a SRSWOR sample and obtain parameter estimates. Suppose we are interested in the change in the number of acres devoted to farms through the years.

- (a) First, remove the observations with missing data in them. Select a SRS of size 100 from the sampling frame (agpop.csv) using 754 as the random seed. Provide R program for this selection process.
 - i. What is the sampling fraction for this problem?
 - ii. What is the probability of selecting this sample?
 - iii. What is the probability of selecting a particular county in the population?
- (b) Estimate the mean number of acres per county devoted to farms and standard errors of the mean for each of the three years using the sample created in (a).
- (c) Is the FPC needed in calculating the standard errors of the mean?
- (d) Using your answers from (b), estimate the total number of acres devoted to farms in the US in 1992 and obtain a 99% confidence interval for this estimate.
- (e) Now suppose that we are interested in what proportion of the counties have tendency to have more large farms than small farms. Use the same sample that we obtained in (a) to do the following (and provide the R outputs):
 - i. Obtain a table with the distribution of your sample data for counties for which the majority of farms are large and the majority of farms are small. Obtain an estimate of the proportion we are interested in and the standard error of the estimated proportion.
 - ii. Calculate an approximate 90% confidence interval for the population proportion of counties with more large farms than small farms.
 - iii. Provide an interpretation of the estimate you obtained in (ii) in words.