Table 1:

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

Table 2:

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

Stat 579: **Assignment 1 Due 2/1 Tuesday in class**

For problem 1 and 2, describe the target population, sampling frame, sampling unit, and observation unit. Discuss any possible sources of selection bias or inaccuracy of responses.

1. A student wants to estimate the percentage of mutual funds whose shares went up in price last week. She selects every tenth fund listing in the mutual fund pages of the newspaper and calculates the percentage of those in which the share price increased.

2. To estimate how many books in the library need rebinding, a librarian uses a random number table to randomly select 100 locations on library shelves. He then walks to each location, looks at the book that resides at that spot, and records whether the book needs rebinding or not.

3. Let's look at an artificial situation in which we know the value of y_i and x_i for each of the $N = 8$ units in the whole population. The index set for the population is

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

The values of y_i and x_i are listed in Table 1. Consider the following sampling scheme (Table 2):

a) Find the probability of selection π_i for each unit i. Is this a simple random sample? Why or why not?

b) What is the sampling distribution of $\hat{t} = 8\bar{y}$?

c) Find $E[\hat{t}]$, $V[\hat{t}]$ and $MSE[\hat{t}]$. Is $E[\hat{t}]$ an unbiased estimator of t?

4. Use randomization theory to show that $\hat{V}(\bar{y}_S) = (1-\frac{n}{N})\frac{s^2}{n}$ is an unbiased estimator of $V(\bar{y}_S) = (1 - \frac{n}{N})\frac{S^2}{n}$.

Stat 579: **Assignment 2 Due 2/17 Thursday in class**

1. The Academic Performance Index *(api)* is computed for all California schools based on standardized testing of students. The data sets contain information for all schools with at least 100 students and for various probability samples of the data. *apipop* contains the entire population. *apistrat* contains a sample stratified by stype. *apiclus1* contains a cluster sample of school districts. *apiclus2* contains a two-stage cluster sample of schools within districts. Also you need to create a simple random sample *apisrs* data from the population apipop.

(a) Compare mean and total estimates of variable "enroll" along with a 95% confidence interval from *apistrat, apiclus1, apiclus2* and *apisrs* to those from *apipop*, do you think your stratification or cluster sampling was worthwhile for sampling from this population? Briefly discuss the reason using theory.

(b) Investigate how R deal with the missing data in your analysis.

2. The data in the file coots.data come from Arnold's (1991) work on egg size and volume of American Coot eggs in Minnedosa, Manitoba. In this data set, we look at volume of a subsample of eggs in clutches (nests of eggs) with at least two eggs available for measurement. Install package "SDaA" and use command library(SDaA), coots.dat is available then. Variable *clutch* is the name of of the cluster (nests of eggs). Variable *csize* is the size of cluster (i.e. M_i).

(a) Calculate the mean egg volume

(b) Use the jackknife to calculate the variance of the mean egg volume

Stat 579: **Assignment 3 Due 3/10 Thursday in class**

1. For data anthuneq (use library(SDaA) to find data), consider height as the dependent variable and finger as the independent variable, find the regression line using svyglm. Make sure you include the weights into your design. Find the jackknife variance estimator for the slope $\hat{B}_1$.

2. Use the stratified sample from the API to examine whether the proportion of teachers with only emergency qualifications (emer) affects academic performance (measured by 2000API). What confounding variables measuring socioeconomic status of students should be included in the model?

3. Read "diagnostics 1" and "diagnostics 2" on my website and write a brief report. You report should include: (1) what should be considered in multiple regression diagnostics; and (2) what is the difference of the diagnostics for survey data and classical diagnostics in multiple regression; (3) what is the key idea of diagnostics for survey data.

Stat 579: **Assignment 4 Due 3/31 Thursday in class**

1. Consider a two-way random effects model. For simplicity, consider the case of one observation per cell. In this case, the observations $y_{ij}, i = 1, \cdots, m, j = 1, \cdots, k$ satisfy

$$y_{ij} = u + \xi_i + \eta_j + \epsilon_{ij} \tag{1}$$

for all i, j. u is the overall mean; $\xi_i, i = 1, \cdots, m, \eta_j, j = 1, \cdots, k$ are independent random effects such that $\xi_i \sim N(0, \sigma_1^2), \eta_j \sim N(0, \sigma_2^2)$; and ϵ_{ij}'s are independent errors distributed as $N(0, \tau^2)$. The random effects and errors are independent. Show that the two-way random effects model can be written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$.

2. Consider one-way balanced random effect model

$$y_{ij} = u + \alpha_i + \epsilon_{ij} \tag{2}$$

where $i = 1, \cdots, m, j = 1, \cdots, k, mk = n$; u is the overall true mean; $\alpha_i \sim N(0, \sigma^2)$, $\epsilon_{ij} \sim N(0, \tau^2)$; the random effects are independent with the errors. The log-likelihood function is given on page 30 of the lecture notes for mixed models. Use the log-likelihood function to find the estimates $\hat{u}, \hat{\sigma}^2, \hat{\tau}^2$

3. Derive an expression for $-2\log R$, where R is the likelihood ratio $R = \frac{L(\boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}^{(2)})}{L(\hat{\boldsymbol{\theta}})}$, under the one-way random effects model (2), (where $k_i = 5$ for all i's) for testing $H_0 : \sigma^2 = 0$ and $H_0 : \sigma^2 = 2$ respectively. What is the asymptotic distribution of the likelihood-ratio test, that is, the asymptotic distribution of $-2\log R$? Study empirically the (asymptotic) size of the likelihood-ratio test and compare it with the normal levels. For the empirical study, let the true parameters be $u = .5$ and $\tau^2 = 1.0$; and consider sample sizes $m = 2, 10, 50$ and $k_i = 5$ for all i in all cases. Run 1000 times.

Stat 579: **Assignment 5 Due April 14 Thursday in class**

Write a brief analysis report about the tumor size study (oncology2.dat). Use "onc.mixed.pdf" as a reference. Use either SAS or R to do the coding.

Stat 579: **Assignment 6**

To measure the performance of the estimators, we use squared error loss in estimating f by

$$\mathrm{L}(\hat{f}) = \frac{\sum_{i=1}^{n}(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i))^2}{n}.$$

Another commonly used measurement is mean squared test set error defined as

$$\mathrm{PE}(\hat{f}) = \frac{\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2}{n}.$$

The loss measures the precision of the estimators for training data and the test error is a measure of prediction precision. The loss measure is widely used in nonparametric smoothing and the test error is often used for machine learning problems.

The following regression function, Friedman#1 (Friedman(1991) and Breiman (1999, 2001)) is for use in your simulation study

$$f(x) = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon, \tag{3}$$

where ε is standard normal and x_i are uniformly distributed between 0 and 1 for $i = 1, 2, \ldots, 10$.

Generate 100 replicate random samples of size 200 for training and 100 replicate random samples of size 200 for testing. For each sample, apply random forests estimator to calculate the losses and test set errors.