Survey Sampling, Fall, 2006, Columbia University
Homework assignments (2 Sept 2006)

Assignment 1, due lecture 3 at the beginning of class

1. Lohr 1.1

2. Lohr 1.2

3. Lohr 1.3

4. Download data from the CBS News/New York Times Teenage Problems Poll, May, 1994:

   - Go to `www.columbia.edu/acis/eds/dgate.html` and search for "Teenage Problems".
   - Click to download the data. In doing this, you will need to follow the instructions and set up an account at the Inter-University Consortium for Political and Social Research (ICPSR).
   - Download the codebook (in PDF form) and the data (in ASCII form). Save the data as `teenage.txt` in the directory of your computer where you will be doing your computations.

   Read into R the responses to the questions, "Do you smoke?", "How many kids in your school smoke?", and the age and sex of the respondent:

   - Search the codebook pdf file for "smoke" to find the two relevant questions on smoking.
   - Open R.
   - Adapt `www.stat.columbia.edu/~gelman/teaching/sampling.course/teenage0.R` to read the relevant columns of the data matrix into R.

   (a) Tabulate the responses to the "Do you smoke?" question and the "How many kids in your school smoke?" question and comment on any discrepancies. (Use the `table` function in R. For help on this function, type `?table` from the R console.)

   (b) Give an estimate and standard error for the proportion of teenagers in the U.S. who smoke (based on the "Do you smoke?" question). The survey has weights (see the note on page 7 of the codebook pdf file), but you can ignore them for now. Just compute the simple average and then use the basic method from introductory statistics to estimate the standard error of a proportion.

   (c) Run a regression of smoking on age and sex using the `lm` function in R. Summarize the regression results (using the `display` function) and discuss whether they make sense.

   (d) Use the regression to predict the smoking status of a randomly-selected sixteen-year-old girl.

Assignment 2, due lecture 5 at the beginning of class

1. Lohr 2.1

2. Lohr 2.4. For each part, you don't have to write a lot; just do the following: (i) give the probability of selection for the units in the population, and thus determine whether all units have equal probability of selection; (ii) if probabilities of selection are equal, state whether this satisfies the rules for simple random sampling.

3. Lohr 2.8 (b, c, d)

4. Suppose you perform systematic sampling of every 10th name from a list of length 352 (picking a random number between 1 and 10 as your starting point). Answer True or False (with explanation) to each of the following statements:

   (a) All items in the list have equal probability of being in the sample.

   (b) Your sample mean $\bar{y}$ is an unbiased estimate of the mean $\bar{Y}$ from the population of 352 units.

5. A simple random sample is conducted of 1000 families in a city that contains 50,000 families. Of the families sampled, 400 have no children, 300 have one child each, 200 have two children each, and 100 have three children each.

   Give an estimate and a standard error of the total number of children in the city.

Assignment 3, due lecture 7 at the beginning of class

1. Lohr 3.1

2. Download the survey on Work, Family, and Well-Being in the United States.

   (a) Load into R the data on height, weight, and sex of the respondents. Use `table` and `hist` to show the distribution of responses for each of these items, numerically and graphically.

   (b) Check the respondents and clean the data as necessary. Report on what you did here.

   (c) Estimate the average weight (in pounds) of the population. Give an estimate and standard error, using each of the following methods.

       i. The simple mean, $\bar{y}$

       ii. The ratio estimate, using the information that the average height in the population is 66.4 inches.

       iii. The regression estimate, using the information that the average height in the population is 66.4 inches.

3. Complete the following sentence:
   Ratio estimation is a special case of regression estimation in which ...

4. A forest resource manager is interested in estimating the number of dead fir trees in a 200-acre area. Using an aerial photo, he divides the area into 200 one-acre plots and estimates the number of dead fir trees in each plot. The total number of dead fir trees in the 200 plots estimated from the photo count is 4200. He then selects a simple random sample of 5 plots and counts the exact number of dead trees in each of these. The data for these five plots are given below.

```
photo count:    12 30 24 30 24
ground count:   18 42 24 36 30
```

   (a) Compute the ratio estimate of the total number of dead fir trees in the 200-acre area. Also get a standard error for the estimate.

(b) From past experience, the manager expects to pick up about two-thirds of the affected trees from an aerial photo. Assume $b = 1.5$ and compute the appropriate regression estimate of the total number of dead fir trees in the 200-acre area. Also get a standard error for the estimate.

(c) Which of the above estimates (a), (b) are unbiased?

Assignment 4, due lecture 9 at the beginning of class

1. Continuing with the last exercise in the previous assignment, estimate the design effects of the estimates (a) and (b).

2. Lohr 3.4

3. Continuing with on Work, Family, and Well-Being in the United States:

   (a) Make a scatterplot of weight vs. height from the survey data (using the `plot` function in R). On the scatterplot, display the ratio line and the regression line. (Use the `lines` function in R to do this.) Compute the mean squared errors of the weights compared to the ratio line and the mean squared errors of the weights compared to the regression line.

   (b) Estimate the average weight (in pounds) of the women *without* children. Give an estimate and standard error. Although your estimate is just the sample mean, this is a ratio estimate (as discussed in Section 3.3 of Lohr) and your standard error should reflect this.

   (c) Estimate the average weight (in pounds) of the women *with* children. Give an estimate and standard error. Although your estimate is just the sample mean, this is a ratio estimate (as discussed in Section 3.3 of Lohr) and your standard error should reflect this.

4. The following data show the stratification of all the farms in a county and the average acres of corn per farm in each stratum.

Stratum, h	Number of farms, N_h	Average corn acres, $\overline{Y}_h$	Standard deviation, S_h
1	800	5.0	8.0
2	200	15.0	20.0

Population mean $\overline{Y} = 7.0$, population standard deviation $S_y = 12.1$

   (a) For a sample of 100 farms, compute the sample sizes in each stratum under proportional allocation. Compute the mean and variance of $\overline{y}_W$ under this design.

   (b) For a sample of 100 farms, compute the sample sizes in each stratum under optimal allocation. Compute the mean and variance of $\overline{y}_W$ under this design. Verify that your variance is lower than the variance from part (a).

Assignment 5, due lecture 11 at the beginning of class

1. Lohr 4.5

2. Lohr 4.6

3. Lohr 4.11

4. Lohr 4.17

5. A survey is done of adults in a (hypothetical) city, with the following results for a certain binary outcome y. Broken down by demographics:

Category	Population size	Sample size	Sample proportion
1	2 million	100	0.1
2	2 million	300	0.5
3	4 million	200	0.6

(a) Estimate the population mean and give a standard error.

(b) Compute the design effect (compared to a simple random sample of size 600).

Assignment 6, due lecture 13 at the beginning of class

1. A group of 100 rabbits is being used in a nutrition study. A prestudy weight is recorded for each rabbit. The average of these weights is 3.1 points. After two months, the experimenter takes a simple random sample of 3 rabbits and weighs them:

rabbit	prestudy weight	current weight
1	3.1	4.2
2	3.0	4.0
3	2.9	3.8

(For convenience in calculation, I have set these these three points to fall exactly on a straight line.)

(a) Give the ratio estimate (and its standard error) for the average of the current weights of the 100 rabbits.

(b) Give the regression estimate (and its standard error) for the average of the current weights of the 100 rabbits.

2. Respondents to a telephone poll are asked whether they have regular telephone service, and how many telephone lines they have.

(a) Respondents are given sampling weights as follows. State whether (i) or (ii) below makes more sense.

i. Respondents with 2 or more phone lines get a weight of 2; respondents with 1 phone line get a weight of 1; respondents with intermittent service get a weight of 1/2

ii. Respondents with 2 or more phone lines get a weight of 1/2; respondents with 1 phone line get a weight of 1; respondents with intermittent service get a weight of 2

The survey researcher is interested in the proportion of respondents who say "yes" to a certain question. Suppose the survey results are as follows:

telephone service	number of respondents	proportion of respondents who say "yes"
2+ telephone lines	400	0.7
1 telephone line	1000	0.5
intermittent service	100	0.3

(b) Give a weighted estimate of the proportion of "yes" responders in the population.

(c) Using the weights, estimate the proportion of the population in each of the three strata above.

(d) Give a standard error for your estimate in (b) above. (Hint: to get the standard error, treat the weighted estimate as a poststratified estimate.)

(e) Estimate the design effect for your weighted estimate (compared to a simple random sample of size 1500 from the population).

3. Do your part for an experiment evaluating survey questions. (We will discuss details in class.)

Assignment 7, due lecture 15 at the beginning of class

1. Perform an interview of a friend. (We will discuss details in class.)

2. Groves (not Lohr) 8.2

3. Groves (not Lohr) 9.2

4. Lohr 5.1

5. Lohr 5.3

6. Complete the following sentences. Be precise.

   (a) Stratified sampling is a special case of cluster sampling in which . . .

   (b) Simple random sampling is a special case of cluster sampling in which . . .

Assignment 8, due lecture 17 at the beginning of class

1. Lohr 5.13

2. Lohr 5.15

3. The following values were obtained for time to wakeup (in minutes) in a systematic sample of day surgical patients (every twentieth patient) at a hospital:
   32, 28, 34, 26, 23, 24, 22, 21, 18, 21.

   (a) Estimate the mean time to wakeup over the study period.

   (b) Use an appropriate estimator for variance under systematic sampling to estimate the variance of this estimated mean. Make it clear what you did.

   (c) Give an approximate 95% confidence interval for the mean.

   (d) Estimate the variance you would have calculated if you had treated the data as coming from a simple random sample. Calculate a design effect.

   (e) Do the simple random sampling and systematic sampling estimates of variance appear to be different? Plot the data and refer to your plot to explain any differences that you may see.

4. A new bottling machine is being tested by a company. During a test run, the machine fills 24 cases, each containing 12 bottles. The company wishes to estimate the average number of ounces of fill per bottle. A two-stage cluster sample is employed using 4 cases, with 5 bottles randomly selected from each. The results are given in the table below.

Case	Average ounces of fill for sample	Sample sd within the case
1	7.5	0.4
2	8.8	0.5
3	7.8	0.3
4	7.2	0.3

(a) Estimate the average number of ounces per bottle and give a 95% confidence interval for this estimate.

(b) Estimate the design effect of this survey (compared to a simple random sample of 20 bottles).

5. A health inspector wants to estimate the number of factory employees in a certain industry who have been exposed to a particular toxin. Suppose the industry has 200 factories, with 100 employees per factory. The plan is to do cluster sampling: first sample a factories at random, then, for each factory sampled, get a list of the employees and perform blood tests on a random sample of b of them, for a total sample size of $n = ab$. It is believed that approximately 10% of the employees have been exposed to the toxin.

Finally, suppose that the cost of the survey is $200 for each factory sampled, plus and $20 for each employee sampled (and assume the blood test is perfectly accurate).

(a) Suppose the intraclass correlation is zero. Approximately how large a sample size n is needed to estimate the proportion of employees exposed to the toxin, so that the 95% interval has the form $x \pm 1\%$?

(b) Suppose the intraclass correlation is 0.2. What is the most cost-effective design (that is, choice of a and b) so that the 95% interval for the proportion exposed has the form $x \pm 1\%$?

Assignment 9, due lecture 19 at the beginning of class

1. A sociologist wants to estimate the total number of retired people residing in a certain city. She decides to sample city blocks and then to sample households within blocks. She does the first stage of sampling by selecting a simple random sample of 4 blocks from the 300 blocks of the city. The number of households in the 4 sampled blocks are 18, 14, 9, and 12.

The sociologist performs a simple random sample of 3 households in each of these 4 blocks. She obtains the following result:

Block	Number of households in block	Number of households sampled	Number of retired residents in the sampled households
1	18	3	1, 0, 2
2	14	3	0, 3, 0
3	9	3	1, 1, 2
4	12	3	0, 1, 1

(a) Estimate the total number of retired residents in the city (and explain how you got your estimate).

(b) Give a standard error for your estimate. (Be careful to do this right, taking the clustering into account!)

2. Lohr 6.1 (a, b, c)

3. Lohr 6.4

4. Lohr 6.13

5. A survey is done sampling 20 out of the 150 units within a company (using probability proportional to a "measure of size," which is last year's revenue from that unit of the company), and then 50 employees are sampled at random within each sampled unit.

Suppose the goal is to get inference about the population of employees in the company (for example, you want to estimate the proportion of the employees who are feeling depressed). Then what should the weight be for each respondent? Define your notation carefully so it is clear what you are saying.

6. A survey is being conducted of public school students in a state. Suppose there are 5000 public schools in the state with 1000 students each. The plan is to pick a simple random sample of schools and then, within each of those schools, pick a simple random sample of students. Suppose that the cost of the survey is $100 per school sampled and $10 per student sampled, and there is a total budget of $10,000. Each student in the survey will be given a standardized test, and their test scores will be recorded.

The survey designers are considering three possible designs: sampling 2 schools, 10 schools, or 50 schools.

(a) Give the number of students per school under each design.

(b) Suppose that S_a^2, the variance of the school means, is 10^2, and S_b^2, the average of the within-school variances, is 20^2. Using this information, give the variance of the sample mean under each design. Which design is best?

(c) The survey is performed. It is desired to estimate the variance of the test scores of all 5 million students. How can you estimate this variance using the data you have collected? Use appropriate notation.

Assignment 10, due lecture 21 at the beginning of class

1. Lohr 7.5

2. Lohr 7.9

3. A large consumer-goods corporation regularly conducts marketing surveys; a typical question asked in a survey is, "How much did your household spend on hair-care products last month?" The corporation estimates that, for the goal of estimating $\overline{Y}$ (for example, the average amount of money spent on hair-care products per household in the target area), the estimate $\overline{y}$ has a coefficient of variation of about 20% and a relative bias of about -10% (that is, a standard error of about $0.2\overline{Y}$ and bias of about $-0.1\overline{Y}$). These values include both sampling and nonsampling errors.

(a) Sketch the sampling distribution of $\bar{y}$ for a quantity with true value $\overline{Y} = \$20$.

(b) What is the approximate probability that the estimate $\bar{y}$ lies within 10% of the true value? (Hint: shade in the appropriate area of the sampling distribution you just drew and use the normal approximation.)

(c) Explain why it makes sense for the corporation to estimate the bias and standard deviation as multiples of $\overline{Y}$ rather than as dollar values.

(d) If the corporation doubled the sample size of its surveys, approximately what would happen to the bias and what would happen to the standard error?

4. Answer the following questions using the survey on Work, Family, and Well-Being in the United States.

- Load into R the data on height, weight, sex, age, ethnicity, and employment status. Characterize ethnicity as white/black/hispanic/other.

- Check the data and clean if necessary. Report on the checks that you did and on any cleaning that was necessary.

- Fit a logistic regression (using the `glm` function in R) to predict whether someone is employed, given their sex, age, ethnicity, and height. Divide age into categories—don't just treat is as a single continuous predictor variable.

(a) Report if any data cleaning was necessary.

(b) Display the estimated logistic regression coefficients and their standard errors.

(c) Using your model, estimate the probability that a 6-foot-tall white male, age 42, is employed.

5. A linear regression is performed to predict a measurement y given the following three predictors: age, female (an indicator that is 0 for men and 1 for women), and age $\times$ female. The model's predictions are as shown on the graph below. [The graph needs to be added here.]

For each of the four coefficients (the constant term, age, sex, and age $\times$ sex), state whether it is positive or negative. Give one sentence for each to explain your answer.

Assignment 11, due lecture 23 at the beginning of class

1. Lohr 8.1

2. Lohr 8.2 (a, d, e, f, g)

3. Groves (not Lohr) 6.1

4. Groves (not Lohr) 6.2

5. The following is the outcome from fitting a logistic regression predicting presidential vote ($0 =$ Democrat, $1 =$ Republican) in a survey of voters in 1972:

```
glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1972))
            coef.est coef.se
(Intercept)  0.67     0.18
```

```
female      -0.25      0.12
black       -2.63      0.27
income       0.09      0.05
```

The variables `female` and `black` are indicators, and `income` is on a 1–5 scale.

(a) What is the probability that a white woman with income category 3 voted for the Republican candidate?

(b) Draw a curve of the estimated probability of supporting the Republican candidate as a function of income, with separate lines indicating white women, white men, black women, and black men.

Assignment 12, due lecture 25 at the beginning of class

1. Impute the missing data in a survey of interest to you. Discuss what you did and your choices.

2. Download data from the CBS News/New York Times Teenage Problems Poll, May, 1994. Where you have missing data, you can either exclude cases or recode to reasonable intermediate values as appropriate; just state clearly what you are doing.

   (a)   i. Estimate the proportion of teenagers who participate in organized team sports.
         ii. Get an standard error for your estimate. (This standard error should account for the weighting, which for this survey is pretty much from poststratification.)

   (b) Run a logistic regression of this outcome on age, sex, race, and parents' education level.

         i. Explain the regression results (the coefficient estimates and standard errors).
         ii. For a white hispanic boy whose parents are both college graduates, plot a curve showing the probability that he participates in organized team sports as a function of his age.
         iii. Compare weighted and unweighted regressions here and discuss the differences. Explain why the the two analyses differ (or why they do not differ).

   (c)   i. Estimate the proportion of teenagers who know someone who's been shot (and give a standard error for this estimate).
         ii. Compare teenagers who participate in organized team sports to teenagers who do not participate in organized team sports. Estimate the difference between the two groups in the proportion who know someone who's been shot. Is this difference statistically significant at the 5% level?

Assignment 13, due lecture 27 at the beginning of class

1. The country of Kalorama has 1 million adults, with 10,000 living in each of 100 administrative districts. The King of Kalorama wishes to know the proportion of adults who support legalized gambling, and so he performs a two-stage cluster sample: first he draws a SRS of 5 districts, then he draws a SRS of 100 adults within each district. The results are as follows: 50, 60, 55, 45, and 65 adults support legalized gambling in the 5 districts sampled.

   (a) Estimate $\overline{Y}$, the proportion of adults in Kalorama who support legalized gambling. Estimate the standard error of your estimate.

(b) The King now decides to obtain a more precise estimate of $\overline{Y}$. Estimate what the standard error of your estimate of $\overline{Y}$ would be if he interviews all 10,000 adults in each of the 5 districts in the sample.

(c) In going from (a) to (b), the King is increasing his sample size by a factor of 100, but the standard error decreases only slightly. This surprises the King, because he remembers from introductory statistics that the standard error should be proportional to $1/\sqrt{n}$. Explain to the King, in non-technical terms, what is happening here.

2. Get the details of a survey of interest to you. Discuss how it could be improved and the costs of these improvements.

3. Alcoholics Anonymous World Service does three kinds of surveys:

   - At their annual convention, about 5000 people attend, and they do a pencil-and-paper survey of the conference attenders.

   - Every four years, AA does a cluster sample, sending a mail survey to about 600 AA meeting groups at a certain date, and asking each person at the meeting at that date to fill out a copy of the survey and send it in.

   - They occasionally pay for questions on national random-digit-dialing telephone polls asking questions such as, "Do you go to AA meetings? If so, how often?" and "What do you think of AA?"

AA is concerned about their demographics: their membership is almost all white, and the average age of the members is increasing.

Discuss how results from the three different kinds of surveys can be used to understand these demographic changes. How would you expect the results from the three surveys to differ, and what is the best use of each?