Survey Sampling, Fall, 2006, Columbia University
Detailed plan of lectures (2 Sept 2006)

lecture 1: Introduction, chapter 1 of Lohr, Chapters 2 and 3 of Groves

1. Some examples of sample surveys
   (a) National public opinion polls
   (b) Home radon surveys
   (c) Post office surveys
   (d) New York City telephone surveys
   (e) Traffic exposure of Australian schoolchildren
   (f) Alcoholics Anonymous membership survey
2. Theory and practice of sample surveys: what is known and what is unknown
   (a) Sampling theory (how to sample from a list)
   (b) Estimation theory (statistical inference)
   (c) Implementation issues, for example:
      i. What fraction of telephone numbers are residential?
      ii. How many people have answering machines?
      iii. How many callbacks?
      iv. How long should a questionnaire be?
      v. How to discourage nonresponse and untruthful responses?
   (d) Combination of above (for example, how to design surveys so that one can reasonably handle the nonresponses)
   (e) Example of research topics: post-stratification of surveys by sex, ethnicity, age, education, region of country, etc.
3. Outline of course

(a) Topics

(b) Course organization

4. Basic terminology of sampling

(a) Links in a chain: superpopulation, target population, surveyed population, sampled population, respondents, responses

(b) Frame, sampling units, clusters, strata, observational units

i. Sampling unit: the individual being sampled (e.g., in a telephone survey, would be "telephone number" not "person")

ii. Observational unit: the person you ask questions to (more generally, the unit on which you take measurements)

iii. Frame: the list of all the sampling units

iv. Cluster: you first sample clusters, then units within sampled clusters (for example, for the radon survey, they first sample counties, then houses within the sampled counties)

v. Stratum: the population is divided into strata, and you sample from within each stratum

5. Key issues in sampling

(a) Census vs. sample

(b) Representative sample

(c) Systematic vs. random sampling

(d) Simple vs. multistage designs

(e) Sampling

i. Selection: how you actually get the sample

ii. Measurement: take data

iii. Estimation of population quantities from sample information

(f) Goals of sampling

  i. Accurate estimation (low bias, low variance)

  ii. Low cost

  iii. Practicality

  iv. Simplicity and reliability of analysis

(g) Kinds of sampling

  i. Haphazard

  ii. Judgment

  iii. Quota

  iv. Probability sampling: every unit in the population has a known, nonzero probability of being selected

(h) Kinds of probability selection methods:

  i. Equal vs. unequal probabilities of selection

  ii. Unstratified vs. stratified

  iii. Element vs. cluster sampling

  iv. Unstratified vs. stratified sampling

  v. Random vs. systematic selection

  vi. One-phase vs. multi-phase sampling

6. The sampling distribution (example: sample of size 2)

   (a) List of sample space and probabilities

   (b) Distribution of $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{2}(y_1 + y_2)$ and
   $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{1}[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2] = \frac{1}{2}(y_1 - y_2)^2$

7. Relation between sampling theory and practice

   (a) Sampling theory is mostly about variance, practice is mostly about bias

   (b) Theory is useful when using complicated designs

(c) Theory is also useful for more elaborate data analysis (going beyond estimating the population mean)

lecture 2: Review of probability and statistics, up to regression forecasting

1. Example: random-digit dialing
2. Question: how can you have sampling with equal probs (epsem) but not SRS?
3. Conditions of simple random sampling:
   (a) No unit can be selected more than once
   (b) Probability of selection is equal for all units (epsem)
   (c) Probability of selection is equal for all pairs, triplets, etc. of units
   (d) Sample size n is fixed
4. Review of probability
   (a) Probability of compound events (cluster sampling)
      i. Example: Suppose you have 1000 schools and 100 kids within each school. You sample 20 schools and then 5 kids within each sampled school. What is the probability that any given kid gets sampled?

   (b) Mean, var, sd, cov, coeff of variation
      i. A random variable u can have the values $u_1, \ldots, u_k$ with probabilities $p_1, \ldots, p_k$, with $p_1 + \ldots + p_k = 1$ (that is, $\sum_{i=1}^k p_i = 1$).
      ii. The mean of u is $\mathrm{E}(u) = \sum_{i=1}^k p_i u_i$.
      iii. The variance of u is $\mathrm{var}(u) = \sum_{i=1}^k p_i (u_i - \mathrm{E}(u))^2$.
      iv. The standard deviation of u is $\mathrm{sd}(u) = \sqrt{\mathrm{var}(u)}$.
      v. The coefficient of variation of u is $\mathrm{c.v.}(u) = \mathrm{sd}(u)/\mathrm{E}(u)$.
      vi. A pair of random variables (u, v) can have the values $(u_1, v_1), \ldots, (u_k . v_k)$ with probabilities $p_1, \ldots, p_k$, with $p_1 + \ldots + p_k = 1$.

    vii. The covariance of (u, v) is $\mathrm{cov}(u, v) = \sum_{i=1}^{k} p_i(u_i - \mathrm{E}(u))(v_i - \mathrm{E}(v))$.

    viii. The correlation of (u, v) is $\mathrm{corr}(u, v) = \frac{\mathrm{cov}(u,v)}{\mathrm{sd}(u)\mathrm{sd}(v)}$.

(c) Mean and var of $ax + by$

    i. We use notation a, b, c for constants and $u, v, \ldots$ for random variables.

    ii. $\mathrm{E}(au + bv + c) = a\mathrm{E}(u) + b\mathrm{E}(v) + c$

    iii. $\mathrm{var}(au + bv + c) = a^2\mathrm{var}(u) + b^2\mathrm{var}(v)2 + 2ab\mathrm{sd}(a)\mathrm{sd}(b)\mathrm{corr}(a, b)$

(d) Mean and var of $x_1 + \cdots + x_n$

(e) Mean and var of arbitrary functions $g(x)$ or $g(x, y)$

(f) Mean and var of uv

    i. Write $u = u_0(1 + \delta_u)$, $v = v_0(1 + \delta_v)$

    ii. u_0 and v_0 are the mean of u and v

    iii. δ_u has mean 0, sd $\sigma_u/u_0|$, where σ_u is the sd of u

    iv. Similarly for δ_v

    v. $uv = u_0v_0(1 + \delta_u)(1 + \delta_v) = u_0v_0(1 + \delta_u + \delta_v + \delta_u\delta_v)$

    vi. Suppose the coeff of var of u and v is small

        A. Then δ_u and δ_v are small

        B. And $\delta_u\delta_v$ is *really* small

        C. So $uv = u_0v_0(1 + \delta_u + \delta_v + \delta_u\delta_v) \approx u_0v_0(1 + \delta_u + \delta_v)$

        D. $\mathrm{E}(uv) \approx u_0v_0(1 + 0 + 0) = uv$

        E. $\mathrm{var}(uv) \approx u_0^2v_0^2\mathrm{var}(\delta_u + \delta_v) = u_0^2v_0^2(\sigma_u^2/u_0^2 + \sigma_v^2/v_0^2 + 2\sigma_{uv}/|u_0v_0|)$

(g) Mean and var of u/v

    i. $\frac{u}{v} = \frac{u_0}{v_0}\frac{1+\delta_u}{1+\delta_v} \approx \frac{u_0}{v_0}(1 + \delta_u)(1 - \delta_v + \cdots)$

    ii. We're using the expansion, $\frac{1}{1+x} = 1 - x + \cdots$

    iii. Continuing: $\frac{u}{v} \approx \frac{u_0}{v_0}(1 + \delta_u - \delta_v + \cdots)$

    iv. $\mathrm{E}(\frac{u}{v}) \approx \frac{u_0}{v_0}$

v. $\mathrm{var}(\frac{u}{v}) \approx (\frac{u_0}{v_0})^2(\sigma_u^2/u_0^2 + \sigma_v^2/v_0^2 - 2\sigma_{uv}/|u_0 v_0|)$

(h) Computing using simulation

i. Think of u and v as random variables with a joint distribution

ii. Random draws of (u, v)

iii. For each draw of (u, v), compute the quantity of interest (for example, u/v)

iv. Compute mean and sd from the simulated values of the quantity of interest

5. Review of statistics

(a) Parameters, data, estimates, bias, standard error

General terms	Example (random sampling with replacement)
parameter θ	$\overline{Y}$ = average of the N units in the population
data y	$y = (y_1, \dots, y_n)$
Probability distribution for y	random sampling with replacement
Estimate $\hat{\theta} = \hat{\theta}(y)$	$\bar{y}$
sampling distribution of $\hat{\theta}$	
mean $\mathrm{E}(\hat{\theta})$	$\mathrm{E}(\bar{y}) = \overline{Y}$
"unbiased" if $\mathrm{E}(\hat{\theta}) = \theta$	yes
variance $\mathrm{var}(\hat{\theta})$	$\mathrm{var}(\bar{y}) = \frac{1}{n}\mathrm{var}(Y) = \frac{1}{n}\sigma_Y^2$
standard error $\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{var}(\hat{\theta})}$	$\mathrm{sd}(\bar{y}) = \frac{1}{\sqrt{n}}\sigma_Y$
Estimated standard error: $\widehat{\mathrm{se}}(\hat{\theta})$	estimate σ_y by s_y
	$\widehat{\mathrm{se}}(\bar{y}) = s_y/\sqrt{n}$
confidence intervals $[\hat{\theta} \pm 2 * \mathrm{se}(\hat{\theta})]$	$[\bar{y} \pm 2 * s_y/\sqrt{n}]$

(b) Sample surveys: notation of y, Y, x, X, etc.

i. Population values $Y_1, \dots, Y_N$, population mean $\overline{Y}$, population total Y.

ii. Sample values $y_1, \dots, y_n$, sample mean $\bar{y}$, sample total y.

(c) Central limit theorem, confidence intervals based on normal and t distributions

i. If $\hat{\theta}$ is an unbiased estimate of θ, and $\widehat{\mathrm{var}}(\theta)$ is an unbiased estimate of $\mathrm{var}(\theta)$, then the usual normal or t-interval is approximately correct.

6. Review of regression forecasting

(a) Linear regression, estimates, std errors, residual sd

(b) Forecasting: deterministic and probabilistic

(c) Forecasting using random simulations

lecture 3: Simple random sampling, Sections 2.1–2.5 of Lohr, Section 4.3 of Groves, Chapter 20 of Gelman and Hill

1. Simple random sampling

   (a) $\mathrm{E}(\bar{y}) = \frac{1}{n}\mathrm{E}(y_1 + \ldots + y_n) = \frac{1}{n}[\mathrm{E}(y_1) + \ldots + \mathrm{E}(y_n)]$

   (b) For any i, all values of Y are equally likely, so $\mathrm{E}(y_i) = \frac{1}{N}Y_1 + \ldots + \frac{1}{N}Y_N = \overline{Y}$

   (c) Thus, $\mathrm{E}(\bar{y}) = \overline{Y}$

   (d)

$$\begin{aligned}
\mathrm{var}(\bar{y}) &= \mathrm{var}[\frac{1}{n}(y_1 + \ldots + y_n)] \\
&= \frac{1}{n^2}\mathrm{var}(y_1 + \ldots + y_n) \\
&= \frac{1}{n^2}[\mathrm{var}(y_1) + \ldots + \mathrm{var}(y_n)] + 2\mathrm{cov}(y_1, y_2) + 2\mathrm{cov}(y_1, y_3) + \ldots + 2\mathrm{cov}(y_{n-1}, y_n)] \\
&= \frac{1}{n^2}[\sum_{i=1}^{n}\mathrm{var}(y_i) + \sum_{i \neq j}\mathrm{cov}(y_i, y_j)]
\end{aligned}$$

   So we need to figure out $\mathrm{var}(y_i)$ and $\mathrm{cov}(y_i, y_j)$.

   (e) $\mathrm{var}(y_i) = \frac{N-1}{N}S_Y^2 = \sum_{i=1}^{N}\frac{1}{N}(Y_i - \overline{Y})^2$

   (f) $\mathrm{cov}(y_i, y_j)$ is the same for any pair of i, j (if $i \neq j$). For simplicity, do $\mathrm{cov}(y_1, y_2)$:

$$\begin{aligned}
\mathrm{cov}(y_1, y_2) &= \mathrm{E}[(y_1 - \mathrm{E}(y_1))(y_2 - \mathrm{E}(y_2))] \\
&= \mathrm{E}[(y_1 - \overline{Y})(y_2 - \overline{Y})] \\
&= \sum_{i \neq j}\frac{1}{N(N-1)}(Y_i - \overline{Y})(Y_j - \overline{Y}) \\
&= \sum_{i \neq j}\frac{1}{N(N-1)}[\sum_{i=1}^{N}\sum_{j=1}^{N}(Y_i - \overline{Y})(Y_j - \overline{Y}) - \sum_{i=1}^{N}(Y_i - \overline{Y})^2] \\
&= \sum_{i \neq j}\frac{1}{N(N-1)}[[\sum_{i=1}^{N}(Y_i - \overline{Y})][\sum_{j=1}^{N}(Y_j - \overline{Y})] - \sum_{i=1}^{N}(Y_i - \overline{Y})^2] \\
&= \sum_{i \neq j}\frac{1}{N(N-1)}[0 * 0 - (N-1)S_Y^2] \\
&= -\frac{1}{N-1}S_Y^2
\end{aligned}$$

   It makes sense that this covariance is negative.

(g) So var($\bar{y}$) $= \frac{1}{n^2}[n\frac{N-1}{N}S_Y^2 - n(n-1)\frac{1}{N}S_Y^2] = \frac{1}{n}\frac{N-n}{N}S_Y^2 = \frac{1}{n}(1-\frac{n}{N})S_Y^2$

(h) Finally, we need an estimate for S_Y^2. It turns out that s_y^2 is an unbiased estimate: $\mathrm{E}(s_y^2) = S_Y^2$

i. We will use the following identity: $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - A)^2 - n(\bar{y} - A)^2$, for any A.

Derivation:

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - A - (\bar{y} - A))^2 \\
&= \sum_{i=1}^n [(y_i - A)^2 + (\bar{y} - A)^2 - 2(y_i - A)(\bar{y} - A)] \\
&= \sum_{i=1}^n (y_i - A)^2 + n(\bar{y} - A)^2 - 2(\bar{y} - A)\sum_{i=1}^n (y_i - A) \\
&= \sum_{i=1}^n (y_i - A)^2 + n(\bar{y} - A)^2 - 2(\bar{y} - A)n(\bar{y} - A) \\
&= \sum_{i=1}^n (y_i - A)^2 - n(\bar{y} - A)^2
\end{aligned}$$

ii. Now, we can work out the expectation of $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$:

$$\begin{aligned}
\mathrm{E}(s_y^2) &= \mathrm{E}(\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2) \\
&= \mathrm{E}(\frac{1}{n-1}[\sum_{i=1}^n (y_i - A)^2 - n(\bar{y} - A)^2]) \\
&= \frac{1}{n-1}[\sum_{i=1}^n \mathrm{E}(y_i - A)^2 - n\mathrm{E}(\bar{y} - A)^2].
\end{aligned}$$

We now substitute in $A = \overline{Y}$ (we can use any value of A that we want; we use this particular value for analytical convenience) and make use of the fact that $\mathrm{E}(\bar{y}) = \overline{Y}$ and, for any i, $\mathrm{E}(y_i) = \overline{Y}$:

$$\begin{aligned}
\mathrm{E}(s_y^2) &= \frac{1}{n-1}[\sum_{i=1}^n \mathrm{E}(y_i - \overline{Y})^2 - n\mathrm{E}(\bar{y} - \overline{Y})^2] \\
&= \frac{1}{n-1}[\sum_{i=1}^n \mathrm{var}(y_i) - n\mathrm{var}(\overline{y})] \\
&= \frac{1}{n-1}[n\frac{N-1}{N}S_Y^2 - n\frac{1}{n}(1 - \frac{n}{N})S_Y^2] \\
&= \frac{1}{n-1}[n\frac{N-1}{N} - \frac{N-n}{N}S_Y^2] \\
&= S_Y^2.
\end{aligned}$$

(i) So, in simple random sampling, s_y^2 is an unbiased estimate of S_Y^2.

(j) Approximate 95% confidence interval for $\overline{Y}$ is $[\bar{y} \pm 2\sqrt{(1-f)}s_y/\sqrt{n}]$.

2. Example: sampling from the telephone book

3. Sampling with and without replacement

    (a) Sampling with replacement is less efficient than simple random sampling (because you might get the same unit more than once).

    (b) Simple random sampling (without replacement) is the standard, but if n/N is small, it does not matter much.

4. Finite population correction.

    (a) Consider 3 claims:

        i. All that matters is n/N (naive view of non-statisticians)

        ii. All that matters is n (what you learned in intro statistics class)

        iii. N is relevant because of f

    (b) "Effective sample size" n^* is defined so that $\text{var}(\bar{y}) = s_y^2/n^*$. Thus, $\frac{1}{n^*} = \frac{1}{n}(1 - \frac{n}{N}) = \frac{1}{n} + \frac{1}{N}$.

    (c) For simple random sampling, n^* is always larger than n

5. Variance estimates

6. Inference for proportions as a special case

    (a) Y_i's (and y_i's) are 0 or 1.

    (b) Define population proportion $P = \overline{Y}$ and sample proportion $p = \bar{y}$

    (c) Then

$$
\begin{aligned}
S_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{(Y)})^2 \\
&= \frac{1}{N-1}((NP)(1-P)^2 + (N-NP)P^2 \\
&= \frac{N}{N-1}P(1-P) \\
&\approx P(1-P)
\end{aligned}
$$

(d) Similarly, $s_y^2 = \frac{n}{n-1}p(1-p)$.

(e) As before, approximate 95% confidence interval for $\overline{Y}$ is $[\bar{y} \pm 2\sqrt{1-f}s_y/\sqrt{n}]$.

(f) Important special point: if p is not too close to 0 or 1, then $s_y \approx 1/2$. So, if f is small, the 95% confidence interval for $\overline{Y}$ is approximately $[\bar{y} \pm 1/\sqrt{n}]$.

(g) For example: an opinion poll of 1000 people, the confidence interval is approximately $[\bar{y} \pm 1/\sqrt{1000}] = [\bar{y} \pm 0.03]$.

7. 3.5 vs. 2.1 kids

   (a) Student demonstration

   (b) Discussion—return to this topic later as "sampling with unequal probabilities"

8. Question: why do sampling with replacement?

   (a) Sampling with replacement is sometimes easier in practice (for example: sampling fish in a lake).

   (b) Sampling with replacement might be easier, and it does not matter if n/N is small.

9. Estimating the population total

   (a) $Y = N\overline{Y}$

   (b) Mean, sd, and all confidence intervals for $\overline{Y}$ are multiplied by N

10. Question: how can you do simple random sampling without knowing N?

    (a) Someone else does the sampling and gives you the sample but does not tell you N

    (b) Sample with equal probability, then subsample n from your first sample

    (c) Sampling from a list with blanks

11. Basic sample size calculations

    (a) How large should n be so that $\text{sd}(\bar{y}) \leq 2000$?

       i. Solve for n: $\frac{1}{n}(1-f)S_Y^2 \leq 2000^2$:

       ii. In terms of the effective sample size n^*, $\frac{1}{n^*}S_Y^2 \leq 2000^2$

iii. So $n^* = (S_Y/2000)^2$ is the lowest possible value for n^*.

iv. Now solve for n: $\frac{1}{n} + \frac{1}{N} = 1/n^*$

v. For example: if $S_Y = 10,000$ and $N = 103$, then $n^* = 25$, so $\frac{1}{n} + \frac{1}{100} = \frac{1}{25}$, and so $n = 1/(\frac{1}{25} + \frac{1}{103}) = 20.1$

vi. n must be at least 20.1, so $n = 21$ is the lowest possible sample size.

(b) How large should n be so that the 95% confidence interval for y has a width of less than 2000?

i. Solve for n: $2(1.96\sqrt{\frac{1}{n}(1-f)S_Y^2}) \leq 2000$

ii. Can correct "1.96" for t distribution if desired

(c) In reality, S_Y is not known—it must be guessed or estimated.

lecture 4: Simple random sampling, Sections 2.6–2.9 of Lohr, Section 4.6 of Groves, Sections 20.1–20.3 of Gelman and Lohr

1. Frame problems: how to fix

(a) Need a list or an algorithm ("virtual list") for picking sample units

(b) Requirements:

i. Must pick one unit at a time

ii. Every unit is on the list exactly once

iii. Nothing else is in the list

(c) Potential problems (consider examples from the U.S. Census)

i. Units not on the list

ii. Blanks in the list

iii. Duplicates

iv. Clusters

(d) Potential solutions

i. Ignore the problem

ii. Redefine the population to fit the frame

iii. Correct the list

iv. Separate stratum for elements not included in the frame

v. Dealing with clusters of elements

vi. Rejecting blanks that are sampled

vii. Correcting duplicate listings (illustrate with class)

2. Example: sampling deliveries in postal routes

3. Systematic sampling

4. Looking forward

   (a) Multstage sampling designs (cluster sampling)

   (b) Sampling with unequal probabilities

5. Estimating things other than the population mean or total

6. Nonresponse

   (a) Example: NYC Social Indicators Survey

   (b) Unit and item nonresponse

   (c) Response rates in actual surveys

   (d) Survey design and nonresponse

   (e) Analyzing surveys with weighting adjustments for nonresponse

7. More sample size calculations

lecture 5: Review

1. Introduction to sampling

2. Review of probability and statistics

3. Data analysis using R

4. Simple random sampling

lecture 6: Ratio and regression estimation, Sections 3.1–3.2 of Lohr

1. Example: QMSS enrollment

   (a) # accepted and planning to attend, 16 June 1999: 7

   (b) # of entering students, Fall 1999: 11

   (c) # accepted and planning to attend, 20 June 2000: 14

   (d) # of entering students, Fall 2000: ?

2. Example: accounting survey

3. Ratio estimation: notation and setup

   (a) Estimating $\overline{Y}/\overline{X}$

      i. X_i and Y_i are two different survey responses

      ii. Estimate is $\bar{y}/\bar{x}$

      iii. Examples:

         A. Estimating the average income of males

         B. Estimating the average revenue per hour worked by part-time employees

      iv. Why do it? Because you want to know $\overline{Y}/\overline{X}$.

   (b) Estimating $\overline{Y}$ using $\overline{X}$

      i. Estimate of $\overline{Y}$ is $\frac{\bar{y}}{\bar{x}}\overline{X}$

      ii. You can only do this if you know $\overline{X}$

      iii. Examples:

         A. Estimating average sales this year using average sales last year

         B. Estimating average calories of fish using weight of fish

iv. Why do it?

A. Because this can be a more efficient estimate (lower variance) compared to simply using $\bar{y}$

B. Consider a simple example in which Y/X is nearly constant

(c) Estimating population totals

i. Estimating $Y = N\overline{Y}$ using $N\bar{y}$ or $N\frac{\bar{y}}{\bar{x}}\overline{X} = \frac{\bar{y}}{\bar{x}}X$

ii. Examples:

A. Estimating total income using total number of books sold

B. Estimating total train ridership using estimates (based on data from previous years)

4. Forecasting using the ratio model

5. Bias and variance of the ratio estimate

(a) Expansion:

$$\begin{aligned}
\frac{v}{u} &= \frac{V(1+\delta_v)}{U(1+\delta_u)} \\
&\approx \frac{V}{U}(1+\delta_v)(1-\delta_u) \\
&= \frac{V}{U}(1+\delta_v-\delta_u+\delta_u\delta_v) \\
&\approx \frac{V}{U}(1+\delta_v-\delta_u)
\end{aligned}$$

(b) Mean:

$$\begin{aligned}
\mathrm{E}(\frac{v}{u}) &\approx \mathrm{E}(\frac{V}{U}(1+\delta_v-\delta_u)) \\
&= \frac{V}{U}(1+\mathrm{E}(\delta_v)-\mathrm{E}(\delta_u)) \\
&= \frac{V}{U}
\end{aligned}$$

So the estimate is *approximately unbiased.*

(c) It is possible to work out a more exact bias expression.

(d) Variance:

$$\begin{aligned}
\mathrm{var}(\frac{v}{u}) &\approx \mathrm{var}(\frac{V}{U}(1+\delta_v-\delta_u)) \\
&= (\frac{V}{U})^2(\mathrm{var}(\delta_v)+\mathrm{var}(\delta_u)-2\mathrm{cov}(\delta_u,\delta_v)) \\
&= (\frac{V}{U})^2(\mathrm{c.v.}(u)^2+\mathrm{c.v.}(v)^2-2\mathrm{c.v.}(u)\mathrm{c.v.}(v)\mathrm{corr}(u,v))
\end{aligned}$$

6. Example: estimating postal volume

7. Regression estimation with b fixed

(a) Bias and variance

i. $\widehat{\overline{Y}} = \bar{y} + b(\overline{X} - \bar{x})$

ii. $\mathrm{E}(\widehat{\overline{Y}}) = \overline{Y} + b(\overline{X} - \overline{X}) = \overline{Y}$. So the estimate is unbiased.

iii. $\mathrm{var}(\widehat{\overline{Y}}) = \mathrm{var}(\bar{y} - b\bar{x}) = \mathrm{var}(\bar{z})$, where

$$z_i = y_i - bx_i$$

iv. So, to estimate the variance, just compute the variance of the n values of z_i

v. For simple random sampling, $\mathrm{var}(\bar{z}) = (1-f)\frac{1}{n}S_Z^2 = (1-f)\frac{1}{n}(s_y^2 + b^2 s_x^2 - 2bs_x s_y \rho_{xy})$

vi. Consider ratio estimation as a special case:

A. $b = \bar{y}/\bar{x}$

B. $z_i = y_i - bx_i$ and $\bar{z} = \bar{y} - b\bar{x}$

C. Variance of ratio estimate of $\overline{Y}$:

$$\mathrm{var}(\widehat{\overline{Y}}) = \mathrm{var}(\bar{z})$$

D. Variance of ratio estimate of $\overline{Y}/\overline{X}$:

$$\begin{aligned}
\mathrm{var}(\frac{\widehat{\overline{Y}}}{\overline{X}}) &= \frac{1}{\overline{X}^2}\mathrm{var}(\widehat{\overline{Y}}) \\
&= \frac{1}{\overline{X}^2}\mathrm{var}(\bar{z})
\end{aligned}$$

E. This is equivalent to the formulas covered in ratio estimation

(b) Optimal b, design effect

i. Optimal b minimizes $\text{var}(\bar{z})$

ii. Under simple random sampling, $\text{var}(\bar{z}) = (1-f)\frac{1}{n}S_Z^2$

iii. Goal is to minimize S_Z^2

iv. Given only the data, we can minimize $s_z^2 = \frac{1}{n-1}\sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - bx_i - (\bar{y} - b\bar{x}))^2$

v. That is, find b to minimize the sum of squares: $\sum_{i=1}^n ((y_i - \bar{y}) - b(x_i - \bar{x}))^2$

vi. The solution is just the least-squares (linear regression) estimate:

$$\begin{aligned} b &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ &= \rho_{xy}\frac{s_y}{s_x} \end{aligned}$$

vii. At the optimal b, we can estimate the variance of the ratio estimate under simple random sampling:

$$\begin{aligned} \text{var}(\widehat{\overline{Y}}) &= \text{var}(\bar{z}) \\ &= (1-f)\frac{1}{n}(s_y^2 + b^2 s_x^2 - 2bs_x s_y \rho_{xy}) \\ &= (1-f)\frac{1}{n}(s_y^2 + \frac{\rho_{xy}^2 s_y^2}{s_x^2} s_x^2 - 2\frac{\rho xy s_y}{s_x} s_x s_y \rho_{xy}) \\ &= (1-f)\frac{1}{n}(s_y^2 + \rho_{xy}^2 s_y^2 - 2\rho_{xy}^2 s_y^2) \\ &= (1-f)\frac{1}{n}s_y^2(1 - \rho_{xy}^2) \end{aligned}$$

viii. Design effect: $\frac{\text{var}(\widehat{\overline{Y}})}{\text{var}(\bar{y})} = 1 - \rho_{xy}^2$

8. Regression estimate with b estimated

   (a) Optimality is only approximate

   (b) Unbiasedness is only approximate

   (c) Regression estimate is OK even if b is poorly chosen

lecture 7: Ratio and regression estimation, Sections 3.3–3.5 of Lohr

1. Example: sampling internet hits

2. Regression forecasting for sample survey inference

(a) Estimate the regression relation of y on x

(b) Imputation of the mean

(c) Uncertainty in the imputations

3. Ratio estimation as a special case of regression estimation

    (a) Goal is to estimate $\overline{Y}$

    (b) Sample mean: estimate is $\bar{y}$

    (c) Ratio estimation: estimate is $\frac{\bar{y}}{\bar{x}}\overline{X}$

    (d) Regression estimation: estimate is $\bar{y} + b(\overline{X} - \bar{x})$

    (e) "b" in regression estimation is the slope of the regression of y on x

    (f) For both ratio and regression estimation, you need to know $\overline{X}$

4. Special cases of regression estimation

    (a) $b = 0$: sample mean

    (b) $b = \frac{\bar{y}}{\bar{x}}$: ratio estimation

5. Advantages over ratio estimation

    (a) Regression estimation is still valid if the X_i's are negative (for example, $Y_i =$ profit of corporation i this year, $X_i =$ profit of corporation i last year

    (b) In regression estimation, b can have any value; in ratio estimate, b must equal $\frac{\bar{y}}{\bar{x}}$

    (c) Why ever do a ratio estimate?

6. Estimation for subsets of the population

    (a) Section 3.3 of Lohr

    (b) Example: average income of males

        i. Let $y_i = \begin{cases} \text{income} & \text{(if respondent } i \text{ is male)} \\ 0 & \text{(otherwise)} \end{cases}$

        ii. Let $x_i = \begin{cases} 1 & \text{(if respondent } i \text{ is male)} \\ 0 & \text{(otherwise)} \end{cases}$

iii. Then average income of males is $\overline{Y}/\overline{X}$, and the ratio estimate is $\bar{y}/\bar{x}$

7. Comparison: sample mean vs. ratio estimate vs. regression estimate

   (a) Simplicity: sometimes sample mean is simplest, sometimes ratio estimation is simplest

   (b) Regression estimate potentially has the minimum variance: illustrate with scatterplots

8. Return to prediction interpretation

9. If *either* the sampling model is correct, *or* the regression model is correct, then the inferences will be reliable.

lecture 8: Stratified sampling, sections 4.1–4.4 of Lohr, Section 4.5 of Groves

1. Example: oversampling in New York City survey

2. Review: understanding the roles of n and N

3. Stratified sampling: definition and notation (W_h and N_h)

   (a) Population analysis (before you take the sample)

      i. Strata $h = 1, \ldots, H$, with $N_1, \ldots, N_H$ units in each stratum

      ii. Population size $N = \sum_{h=1}^{H} N_h$

      iii. Within each stratum h: population stratum mean $\overline{Y}_h$ and population stratum variance S_h^2

      iv. Parameter (population quantity) of interest: $\overline{Y}_W = \sum_{h=1}^{H} W_h \overline{Y}_h$

      v. Sum of weights $\sum_{i=1}^{H} W_h = 1$

      vi. Usually, $W_h = N_h/N$. If so, then $\overline{Y}_W = \overline{Y}$, the population mean.

   (b) Data from the sample

      i. Independent sampling from each stratum

      ii. Sample sizes $n_1, \ldots, n_H$, total sample size $n = \sum_{h=1}^{H} n_h$

      iii. Within each stratum h: sample stratum mean $\bar{y}_h$ and sample stratum variance s_h^2

    iv. Within each stratum, any kind of sampling might be done; for now, we'll assume simple random sampling

4. Interpretation of stratified sampling as a regression estimate

5. Weighted mean estimate

   (a) Estimate $\overline{Y}_W$ by the weighted mean: $\bar{y}_W = \sum_{h=1}^{H} W_h \bar{y}_h$

   (b) The "W" in $\bar{y}_W$ is capitalized because the weights in the weighted mean come from the population analysis

   (c) Key idea: you get a separate estimate in each stratum, then combine them using the weights W_h

6. Why stratify?

   (a) Practicality/cost

      i. Sometimes it's the only way you can collect the data

      ii. Examples

   (b) Bias reduction (for example, suppose you have different nonresponse rates in different strata)

   (c) Variance reduction

      i. Oversampling more important units (do example in class)

      ii. Studying subpopulations of interest (for example, minority groups in a U.S. survey)

7. Relation between W_h and N_h

   (a) Usually $W_h = N_h/N$.

   (b) When does $W_h \neq N_h/N$?

      i. You are interested in a larger superpopulation (for example, age adjustment of death rates)

      ii. N_h's are unknown. Then you must create weights based on estimates

8. Mean and variance of $\bar{y}_W$

   (a) Key assumption: sampling schemes in the different strata are independent

(b) $\mathrm{E}(\bar{y}_W) = \sum_{h=1}^H W_h \mathrm{E}(\bar{y}_h)$

(c) $\mathrm{var}(\bar{y}_W) = \sum_{h=1}^H W_h^2 \mathrm{var}(\bar{y}_h)$

(d) Assume simple random sampling within strata:

   i. $\mathrm{E}(\bar{y}_W) = \sum_{h=1}^H W_h \overline{Y}_h = \overline{Y}_W$ (unbiased estimate)

   ii. Variance $\mathrm{var}(\bar{y}_W) = \sum_{h=1}^H W_h^2 (1 - f_h) \frac{1}{n_h} S_h^2$

   iii. (Sampling fraction in stratum h: $f_h = n_h / N_h$)

   iv. Estimated variance $\widehat{\mathrm{var}}(\bar{y}_W) = \sum_{h=1}^H W_h^2 (1 - f_h) \frac{1}{n_h} S_h^2$

   v. Degrees of freedom in t-interval: somewhere between $\min(n_h - 1)$ and $n - 1$

9. For a unit in stratum h, probability of selection (assuming equal probability of selection of units within the stratum) is n_h / N_h

10. Tasks for stratified sampling (in reverse order):

   (a) Given the strata and the sample size within each stratum, perform the inference (compute $\bar{y}_W$ and $\widehat{\mathrm{var}}(\bar{y}_W)$)

   (b) Given the strata, decide how many to sample within each stratum $(n_1, \ldots, n_H)$

   (c) Set up the strata

   (d) For all of these tasks, we consider practicality/cost, bias, and variance

11. Allocating sample size across strata

   (a) First, the simplest approach: proportional sampling

      i. Simple example:

         A. Employees at a company: 50% have seniority, 50% do not have seniority

         B. Every worker with seniority makes \$30,000; every worker without seniority \$20,000

         C. Simple random sample of size 2: $\overline{y} = 30{,}000$, 35000, or 40,000

         D. Proportional stratified sample (1 from each stratum): $\overline{y} = 35000$ (no variance, perfect estimate)

ii. With proportional sampling, $f_h = n_h/N_h = n/N$, identical for all h

iii. Sampling is equal-probability

iv. Assuming $W_h = N_h/N$ for all h:

A. Then $n_h = nW_h$ for each h.

B.

$$\begin{aligned}
\bar{y}_W &= \sum_{h=1}^{H} W_h \bar{y}_h \\
&= \sum_{h=1}^{H} W_h \frac{1}{n_h}(y_{h1} + \ldots + y_{hn_h}) \\
&= \sum_{h=1}^{H} \frac{1}{n}(y_{h1} + \ldots + y_{hn_h}) \\
&= \frac{1}{n}\sum_{h=1}^{H}(y_{h1} + \ldots + y_{hn_h}) \\
&= \bar{y}
\end{aligned}$$

This is called a "self-weighted" sample, because the weighted estimate is simply the unweighted mean. This makes sense, since with equal-probability sampling, the sample mean is an unbiased estimate of the population mean.

C.

$$\begin{aligned}
\mathrm{var}(\bar{y}_W) &= \sum_{h=1}^{H} W_h^2 \frac{1}{n_h}(1 - f_h) S_h^2 \\
&= \sum_{h=1}^{H} \frac{W_h}{n}(1 - f) S_h^2 \\
&= (1 - f)\frac{1}{n}\sum_{h=1}^{H} W_h S_h^2
\end{aligned}$$

This looks a lot like the variance from simple random sampling, $(1-f)\frac{1}{n}S_Y^2$, with the difference that S_Y^2 is replaced by a weighted average of the within-strata variances.

v. Example (an opinion poll in urban, suburban, and rural areas):

Stratum	N_h	W_h	n_h	$\bar{y}_h$	s_h
urban	5000	0.500	50	0.06 (that is, 3 out of 50 said "yes")	$\sqrt{\frac{50}{49}(0.06)(1-0.06)} = 0.24$
suburban	3010	0.301	30	0.20 (6 out of 30)	$\sqrt{\frac{30}{29}(0.20)(1-0.20)} = 0.41$
rural	1990	0.199	20	0.60 (12 out of 20)	$\sqrt{\frac{20}{19}(0.60)(1-0.40)} = 0.50$

A. Weighted estimate is $\overline{y}_W = 0.500(0.06) + (0.301)(0.20) + (0.199)(0.60) = 0.201$

B. Estimated sd is $\mathrm{sd}(y_W) = \sqrt{(.500)^2(1-\frac{50}{5000})\frac{1}{50}(.24)^2 + (.301)^2(1-\frac{30}{3010})\frac{1}{30}(.41)^2 + (.199)^2(1-\frac{20}{1990})\frac{1}{20}(.50)^2} =$ 0.036

C. Confidence interval is $[0.201 \pm 2(0.036)] = [0.13, 0.27]$

(b) Goal: low variance within strata

(c) Design effect

i. Definition: design effect $= \dfrac{\text{variance of the estimate in your survey}}{\text{variance of the estimate from a SRS of the same sample size}}$

ii. For proportional sampling, design effect $= \frac{\sum_{h=1}^{H} W_h S_h^2}{S_Y^2}$

iii. Recall that $\sum_{h=1}^{H} W_h = 1$ always

iv. In general for stratified sampling, design effect $= \frac{\sum_{h=1}^{H} W_h^2 (1-f_h)\frac{1}{n_h} S_h^2}{(1-f)\frac{1}{n} S_Y^2}$

v. You want the design effect to be less than 1

(d) Optimal allocation

i. For a fixed cost, what design will give the lowest-variance estimate of $\overline{Y}_W$? That is, what should be the within-stratum sample sizes $n_1, \ldots, n_H$?

ii. We'll assume that $W_h = \frac{N_h}{N}$ for all h

iii. Simplest case: cost of sampling is same in each stratum. Then "fixed cost" is the same as "fixed sample size n"

A. Optimal sampling is $n_h \propto W_h S_h$ for all h: set $n_h = \frac{W_h S_h}{W_1 S_1 + \ldots + W_H S_h} n$

B. Proof

C. Special case: if S_h's are equal, then $n_h \propto W_h$: proportional sampling is optimal

iv. More generally: some strata are easier to sample than others. Cost of sampling a unit in stratum h is J_h, and the total cost is $\sum_{h=1}^{H} J_h n_h = C$, a fixed value.

A. Optimal sampling is $n_h \propto W_h S_h / \sqrt{J_h}$ for all h:

B. Proof

v. Other concerns

A. You may need to round up or down to get integer values for n_h

B. You may need to alter the sample sizes if you do 100% sampling in some strata

C. You need at least 2 in each stratum so you can estimate the variance

D. Variances S_h^2 and costs J_h are not always known. If you use bad guesses, then "optimal" sampling can be worse than proportional sampling (and can even be worse than simpler random sampling)

vi. Numerical example

Stratum	$W_h S_h$
1	90.0
2	54.3
3	8.0
4	753.0

A. Suppose $n = 1000$ is the total sample size. Then set $n_1 = \frac{90.0}{90.0+54.3+8.0+753.0}1000$, etc. Then $(n_1, \ldots, n_4) = (99.4, 60.0, 8.8, 831.8)$. After rounding: $(n_1, \ldots, n_4) = (99, 60, 9, 832)$.

B. Now suppose that stratum 4 has only 600 units. Then n_4 must be 600 (not 832). So we have 400 remaining units to assign to strata 1–3. Then set $n_1 = \frac{90.0}{90.0+54.3+8.0}400$, etc. After rounding: $(n_1, n_2, n_3) = (236, 143, 21)$, and $n_4 = 600$.

vii. If costs are unequal: suppose cost per item is \$1 in stratum 1, \$2 in stratum 2, \$3 in stratum 3, and \$4 in stratum 4, and total budget is \$1000.

A. Sample size in stratum h is $n_h \propto W_h S_h / \sqrt{J_h}$, so the total \$ spent in stratum h is $J_h n_h \propto W_h S_h \sqrt{J_h}$.

Stratum	$W_h S_h \sqrt{J_h}$
1	90.0
2	76.8
3	13.9
4	1506.0

B. Set the total \$ value for stratum 1 to be $\frac{90.0}{90.0+76.8+13.9+1506.0}\$1000$, stratum 2 to be $\frac{76.8}{90.0+76.8+13.9+1506.0}\$1000$, etc. That is, \$53.4 for stratum 1, \$45.5 for stratum 2, \$8.2 for stratum 3, \$892.9 for stratum 4.

C. Now divide by the cost and round: $\$53.4/\$1 = 53$ units in stratum 1, $\$45.5/\$2 = 23$ units in stratum 2, $\$8.2/\$3 = 3$ units in stratum 3, $\$892.9/\$4 = 223$ units in stratum 4.

D. Now check: total cost is $53(\$1) + 23(\$2) + 3(\$3) + 223(\$4) = \$1000$. This works. Often, when you do it, it will come out a few dollars over or under (because of rounding), and then you must adjust your budget or else sample one or two more or fewer units.

lecture 9: Stratified sampling, sections 4.5–4.8 of Lohr

1. Example: sampling post offices

2. Setting up the strata (optimality and practical concerns)

    (a) Practical concerns

    i. You have to be able to sample separately within each stratum

    ii. You need to know N_h for each stratum

    iii. Quota sampling

    (b) "Objectivity" is not necessary. Strata can be whatever you want.

    (c) Bias reduction: stratify on variables that might affect the probability of selection

    (d) Variance reduction: S_h should be small for all h (homogeneous strata)

    (e) Example: telephone survey of U.S. households

    i. Why stratify by region?

    ii. Why not stratify by state also?

3. Stratification with 0/1 responses

    (a) Within-stratum sd: $S_h = \sqrt{Y_h(1 - Y_h)}$. Unless your strata are really good, S_h will be close to 0.5, so there's not much variance reduction from stratification.

    (b) If you do stratify, and costs are same for all strata, then proportional sampling is close to optimal.

4. Post-stratification

    (a) Unstratified design, stratified analysis

    (b) How

    i. Gather the data, then define strata $h = 1, \ldots, H$

    ii. Use the weighted estimate, $\bar{y}_W = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$.

    iii. Compare to the sample mean, $\bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h$.

(c) Why

i. Corrects for bias caused by unequal sampling probabilities

ii. Reduces variance, but not as much as stratification (because, in poststratification, you can't control the sample sizes n_h)

5. Measuring several variables

(a) Difficult to get optimality for more than one variable

(b) Use approximately optimal sampling or proportional sampling

lecture 10: Review

1. Ratio and regression estimation
2. Stratified sampling
3. Poststratification
4. Regression modeling of survey data

lecture 11: Questions and answers in surveys, Chapter 7 of Groves

1. Cognitive processes in answering questions
2. Problems in answering survey questions
3. Guidelines for writing good questions
4. Examples

(a) How many X's do you know?

(b) Pre-election polls

5. Class chooses a topic to ask about

lecture 12: Evaluating survey questions, Chapter 8 of Groves

1. Expert reviews and focus groups
2. Radomized experiments
3. Reliability and validity
4. Examples
   (a) National Election Study
   (b) Social Indicators Survey
5. Class designs an experiment to evaluate survey questions

lecture 13: Survey interviewing, Chapter 9 of Groves

1. Interviewer bias and variance
2. Strategies for reducing interviewer bias and variance
3. Examples
   (a) Roaches and rodents
   (b) Social Indicators Survey

lecture 14: Cluster sampling, Sections 5.1–5.2 of Lohr, Section 4.4 of Groves

1. Review: analysis of systematic sampling as stratified sampling
2. Example: fish sampling
3. Cluster sampling with equal cluster sizes
   (a) Why?
      i. Lower cost per element sampled
      ii. Less effort required per element sampled

   iii. Examples: see table on p. 149 of Kish

   iv. In general, cluster sampling has higher variance per element, but the lower cost allows you to sample more units

(b) Notation

   i. For simplicity, we first consider equal cluster sizes

   ii. First stage sampling:

      A. Population of clusters $\alpha = 1, \ldots, A$

      B. Sample of clusters $\alpha = 1, \ldots, a$

      C. Sampling fraction of clusters $f_a = a/A$

   iii. Second stage sampling:

      A. Population: B items within each cluster

      B. Sample: b items within each cluster

      C. Sampling fraction within clusters $f_b = b/B$

   iv. Population means $\overline{Y}_\alpha$ = average of B units in cluster α

   v. Sample means $\overline{Y}_\alpha$ = average of b *sampled* units in cluster α

   vi. Total sampling fraction $f = f_a f_b = \frac{ab}{AB}$

   vii. For equal cluster sizes, the population mean $\overline{Y}$ is $\frac{1}{A}\sum_{\alpha=1}^{A}\overline{Y}_\alpha$, the mean of the cluster means

   viii. For equal sample sizes within clusters, the sample mean $\bar{y}$ is $\frac{1}{a}\sum_{\alpha=1}^{a}\bar{y}_\alpha$, the mean of the cluster means

   ix. Equal-probability sampling: $\bar{y}$ is an unbiased estimate of $\overline{Y}$

(c) How?

   i. You need a sampling frame of clusters

   ii. Clusters must be well-defined (for example, they should not overlap)

   iii. Once you have sampled the clusters, you need sampling frames for each of the *sampled* clusters

4. One-stage cluster sampling

(a) For simplicity, we just consider one-stage cluster sampling today

(b) That is, assume $b = B$: complete sampling within clusters

(c) In this case, $\bar{y}_\alpha = \overline{Y}_\alpha$ for all sampled clusters α

(d) Example: sample classes at random and gather data from all the students in each sampled class

5. Estimation of $\overline{Y}$

   (a) Consider this as a simple random sampling situation

      i. Units are clusters

      ii. Measurements are cluster means, $\overline{Y}_\alpha$

      iii. Population size is A

      iv. Sample size is a

      v. Estimate of $\overline{Y}$ is $\bar{y}$

      vi. Variance is $\text{var}(\bar{y}) = (1-f)\frac{1}{a}S_a^2$, where $S_a^2 = \frac{1}{A-1}\sum_{\alpha=1}^{A}(\overline{Y}_\alpha - \overline{Y})^2$

6. Example:

   (a) $N = 50$ items: the numbers $1, 2, 3, \ldots, 50$, divided into 5 clusters. Cluster 1 is the numbers 1–10, …, cluster 5 is the numbers 41–50.

   (b) $A = 5$, $B = 10$

   (c) Consider 2 options:

      i. Sample 2 clusters, each of size 10: $\text{var}\bar{y} = (1-0.4)\frac{1}{2}S_a^2$

      ii. Simple random sample of $n = 20$: $\text{var}\bar{y} = (1-0.4)\frac{1}{20}S^2$

7. Variance decomposition

   (a) We shall decompose the population variance into within-cluster and between-cluster variance

   (b) This is not necessary for data analysis (as described above, we can use standard SRS methods for that) but is useful for design: deciding how many clusters to sample

(c) Variance decomposition:

$$\begin{aligned}
\text{Total variance} &= \text{Between-cluster variance} + \text{Within-cluster variance} \\
\sigma^2 &= \sigma_a^2 + \sigma_b^2 \\
\frac{1}{AB}\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B}(Y_{\alpha\beta} - \overline{Y})^2 &= \frac{1}{A}\sum_{\alpha=1}^{A}(\overline{Y}_\alpha - \overline{Y})^2 + \frac{1}{A}\sum_{\alpha=1}^{A}\left[\frac{1}{B}\sum_{\beta=1}^{B}(\overline{Y}_{\alpha\beta} - \overline{Y}_\alpha)^2\right] \\
\frac{AB-1}{AB}S^2 &= \frac{A-1}{A}S_a^2 + \frac{B-1}{B}S_b^2
\end{aligned}$$

(d) S_B^2 is the same as S_W^2, the "weighted within-stratum variance" from stratified sampling with proportional allocation. (Here we are considering the special case where all strata have equal sizes.)

(e) In our example of $1, \ldots, 50$ divided into 5 clusters:

i. $\sigma^2 = \frac{1}{50}((1-25.5)^2 + (2-25.5)^2 + \ldots + (50-25.5)^2) = 208.25 = 14.43^2$

ii. $\sigma_a^2 = \frac{1}{5}((5.5-25.5)^2 + (15.5-25.5)^2 + \ldots + (45.5-25.5)^2) = 200 = 14.14^2$

iii. $\sigma_b^2 = \frac{1}{5}\left[(\frac{1}{10}(1-5.5)^2 + \ldots + (10-5.5)^2) + \ldots + (\frac{1}{10}(41-45.5)^2 + \ldots + (50-45.5)^2)\right] = 8.25 = 2.87^2$

iv. $S^2 = 212.5,\ S_a^2 = 250,\ S_b^2 = 9.17$

8. Design effect

(a) Design effect $= \frac{\text{variance under cluster sampling}}{\text{variance under simple random sampling}} = \frac{(1-f)\frac{1}{a}S_a^2}{(1-f)\frac{1}{n}S^2} = B\frac{S_a^2}{S^2}$

(b) Above example:

i. Design effect is $10\frac{250}{212.5} = 11.8$

ii. So the cluster design is about as efficient as a SRS with 1/12 the sample size. That is very inefficient!

i. Homogeneous clusters

A. If clusters are homogeneous, this means that S_b^2 is small (that is, small variance *within* clusters)

B. Then $S_a^2 \approx S^2$, and so the design effect $\approx B$

C. Worst-case scenario for cluster sampling

ii. Randomly created clusters

A. If clusters are assigned at random, then a cluster sample is essentially the same as a SRS, so the design effect ≈ 1

B. Thus, $S_a^2 \approx S^2$

iii. Extreme heterogeneity

A. If clusters are extremely heterogeneous, this means that S_b^2 is large

B. Largest possible value for S_b^2 occurs when $S_a^2 = 0$; that is, the A cluster means are all equal: $\overline{Y}_1 = \overline{Y}_2 = \ldots = \overline{Y}_A$

C. Then the design effect $= 0$ (that is, sampling variance from cluster sampling is 0 because all clusters are identical)

D. Not realistic

iv. In practice, the design effect is usually between 1 and B

A. Design effect $= 1$ implies that sampling a clusters of size b is as good as a SRS of size ab

B. Design effect $= B$ implies that sampling a clusters of size b is only as good as a SRS of size a

v. Intraclass correlation, ρ

A. $\rho = 1 - \frac{N}{N-1}\frac{S_b^2}{S^2}$

B. Compares *within-cluster variance* to *total variance*

C. If clusters are perfectly homogeneous, then $\rho = 1$

D. If clusters are set up at random, then $\rho = 0$

E. It is possible (but highly unusual in practice) for ρ to be negative

9. Interpretation of cluster sampling estimates in terms of missing-data imputation

lecture 15: Cluster sampling, Sections 5.3–5.7 of Lohr

1. Example: industrial sampling

2. 2-stage cluster sampling (different from 2-phase) with equal cluster sizes

   (a) Practical implementation

      i. For simplicity, assume: first stage SRS of clusters, second stage SRS of units within each cluster

      ii. Sample a units, sample b units per cluster

      iii. Sample size $n = ab$, sampling fraction $f = \frac{ab}{AB}$

      iv. "Data" $\bar{y}_1, \ldots, \bar{y}_a$

      v. Estimate is $\bar{y} = \frac{1}{a}\sum_{\alpha=1}^{a} \bar{y}_\alpha$

      vi. To estimate the variance, think of this as a SRS of size a from a population of $A\frac{B}{b}$ batches

      vii. That is, each "batch" is of size b, there are $\frac{B}{b}$ batches per cluster, so there are a total of $A\frac{B}{b}$ batches in the population

      viii. This is not exactly right, because we are actually sampling one "batch" from each sampled cluster—it's not quite a SRS of batches

      ix. But the approximation is reasonable as long as $f_a = \frac{a}{A}$ is small

      x. Estimated variance is $\widehat{\mathrm{var}}(\bar{y}) = (1-f)\frac{1}{a}s_a^2$, where $s_a^2 = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2$

   (b) Theoretical variance expression

      i. Expression for variance (exact variance, not estimated) of cluster sampling: $\mathrm{var}(\bar{y}) = (1 - f_a)\frac{1}{a}S_a^2 + (1-f_b)\frac{1}{ab}S_b^2$

      ii. The first term, $(1-f_a)\frac{1}{a}S_a^2$, is the variance of $\bar{y}$ if we had 100% sampling within each cluster (as discussed in the previous lecture)

      iii. The second term, $(1-f_b)\frac{1}{ab}S_b^2$, is the additional variance because $f_b \neq 1$ (that is, you do not have a complete census within each of the sampled clusters)

      iv. You can use the theoretical expression to make design decisions (how many clusters to sample, how many units to sample within each cluster)

         A. Increasing a decreases both the first and second terms proportionally (and also affects the first term in a minor way through f_a): $\mathrm{var}(\bar{y}) = \frac{1}{a}\left[(1-f_a)S_a^2 + (1-f_b)\frac{1}{b}S_b^2\right]$

B. Increasing b decreases only the second term

3. Variance decomposition

   (a) Estimating S_a^2 and S_b^2

      i. Compute the sample variance of the cluster means: $s_a^2 = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2$

      ii. Compute the average of the within-cluster variances: $s_b^2 = \frac{1}{a}\sum_{\alpha=1}^{a}\frac{1}{b-1}\sum_{\beta=1}^{b}(y_{\alpha\beta} - \bar{y}_\alpha)^2$

      iii. Derive their expectations under cluster sampling:

         A. $\mathrm{E}(s_b^2) = S_b^2$: the within-cluster variances in the sample are unbiased estimates of the within-cluster variances in the population

         B. $\mathrm{E}(s_a^2) = S_a^2 + (1-f_b)\frac{1}{b}S_b^2$: the sample variance of the cluster means tends to be an *over-estimate* of the population variance. This is because the cluster means $\bar{y}_\alpha$ have sampling variability

         C. If $b = B$ (that is, $f_b = 1$), then $\mathrm{E}(s_a^2) = S_a^2$ as discussed in the previous lecture

      iv. Unbiased estimates:

         A. Estimate S_b^2 using $\widehat{S}_b^2 = s_b^2$

         B. Estimate S_a^2 using $\widehat{S}_a^2 = s_a^2 - (1-f_b)\frac{1}{b}s_b^2$

         C. If the estimate of S_a^2 is negative (this is unlikely to happen in practice), then estimate S_a^2 using other information

   (b) Estimating $\mathrm{var}(\bar{y})$

      i. Estimated variance is $\widehat{\mathrm{var}}(\bar{y}) = (1-f_a)\frac{1}{a}\widehat{S}_a^2 + (1-f_b)\frac{1}{ab}\widehat{S}_b^2$

      ii. Plug in the above estimate and clean up the algebra: $\widehat{\mathrm{var}}(\bar{y}) = (1-f_a)\frac{1}{a}s_a^2 + f_a(1-f_b)\frac{1}{ab}s_b^2$

      iii. If f_a is small, this is approximately equal to $(1-f_a)\frac{1}{a}s_a^2$, as discussed earlier

4. Design effect

5. Stratified cluster sampling

   (a) Clusters within strata

i. Within each stratum h, do the complete cluster sampling analysis: this gives you $\bar{y}_h$ and $V_h = \widehat{\text{var}}(\bar{y}_h)$ in each stratum

ii. Compute these based on the data to get numerical values for $\bar{y}_h$ and V_h in each stratum

iii. Now combine as in stratified sampling:

A. Estimate $\overline{Y}_W$ by $\sum_{h=1}^{H} W_h \bar{y}_h$

B. Sampling variance of the estimate is $\sum_{h=1}^{H} W_h^2 V_h$

(b) Why not strata within clusters?

i. Stratifying within clusters would reduce the within-cluster variance, but the main concern in cluster sampling is usually the between-cluster variance

ii. Clusters tend to be homogeneous, so there is not usually much gain in stratifying within clusters

iii. In practice, it is common to have a large number of clusters (for example, schools in New York State) but a small number of strata (for example, public schools, private religious schools, and private non-religious schools), so that the clusters are nested inside the strata

iv. If you happen to have a situation where there is stratified sampling within cluster sampling, you can analyze it by simply replacing the cluster estimates $\bar{y}_\alpha$ by stratified estimates $\bar{y}_{W\,\alpha}$

6. Systematic sampling

(a) Definition

(b) Dealing with extras and blanks

(c) Advantages and disadvantages compared to other methods

(d) Fundamental problem of systematic sampling: variance is not "measurable"

(e) How to deal with the problem (pretend it is stratified sampling)

(f) Difficulties of systematic sampling and remedies

(g) Paired selection

(h) Replicated sampling

7. Cost functions, optimal cluster sampling

   (a) Assume cost is C_a for each sampled cluster and c for each sampled unit

   (b) Simple situation: equal cluster sizes, epsem sampling of a out of A clusters, b out of B units within each sampled cluster

   (c) What are optimal a and b given a fixed total cost:

$$\text{total cost} C_{\text{total}} = C_a a + cab$$

   (d) Goal is to minimize the variance:

$$\text{var}(\bar{y}) = (1 - f_a)\frac{1}{a}S_a^2 + (1 - f_b)\frac{1}{ab}S_b^2$$

   (e) Our answer will depend on S_a^2 and S_b^2

   (f) To do "optimal sampling" in practice, you must use guesses of S_a^2 and S_b^2

   (g) Find the optimal a, b given the total cost C_{total}:

      i. Simplify by ignoring finite population corrections: minimize $\text{var}(\bar{y}) \approx \frac{1}{a}S_a^2 + \frac{1}{ab}S_b^2 = \frac{1}{a}(S_a^2 + \frac{1}{b}S_b^2)$

      ii. Cost constraint is $C_{\text{total}} = C_a a + cab = (C_a + cb)a$, so $\frac{1}{a} = \frac{1}{C_{\text{total}}}(C_a + cb)$

      iii. Substitute for $\frac{1}{a}$ and solve for the b for which variance is minimized:

$$\begin{aligned}
\text{var}(\bar{y}) &\approx \frac{1}{a}S_a^2 + \frac{1}{ab}S_b^2 \\
&= \frac{1}{a}(S_a^2 + \frac{1}{b}S_b^2) \\
&= \frac{1}{C_{\text{total}}}(C_a + cb)(S_a^2 + \frac{1}{b}S_b^2) \\
&= \frac{1}{C_{\text{total}}}(C_a S_a^2 + cS_b^2 + cS_a^2 b + C_a S_b^2 \frac{1}{b})
\end{aligned}$$

      iv. Set the derivative with respect to b equal to 0:

$$\begin{aligned}
0 &= \frac{d}{db}\text{var}(\bar{y}) \\
&\approx \frac{d}{db}\left[\frac{1}{C_{\text{total}}}(C_a S_a^2 + cS_b^2 + cS_a^2 b + C_a S_b^2 \frac{1}{b})\right] \\
&= \frac{1}{C_{\text{total}}}(cS_a^2 - C_a S_b^2 \frac{1}{b^2}) \\
0 &= cS_a^2 - C_a S_b^2 \frac{1}{b^2}
\end{aligned}$$

v. Solve for b:

$$b = \sqrt{\frac{C_a S_b^2}{c S_a^2}}$$

vi. Total cost $C_{\text{total}} = C_a a + cab$ is fixed, so $a = \frac{C_{\text{total}}}{C_a + cb}$

vii. Consider how the optimal design varies as a function of C_a, c, S_a^2, S_b^2

viii. Extreme cases: b must be at least 1 and no more than B

lecture 16: Sampling with unequal probabilities, begin chapter 6 of Lohr

1. Example: adjusting for number of telephone lines and household size in telephone surveys

2. Unequal probabilities of selection

   (a) Single-stage or multi-stage

   (b) Can be an inherent feature of the sampling design

   (c) Can be related to nonresponse

3. Stratification/poststratification

4. Inverse-probability weighting

5. Example: Alcoholics Anonymous survey

6. One-stage cluster sampling with unequal cluster sizes

7. Two-stage cluster sampling with unequal cluster sizes

8. Use of ratio estimation for cluster sampling

9. Mean and variance of the ratio estimate for cluster sampling

lecture 17: Sampling with unequal probabilities, conclude chapter 6 of Lohr, Sections 4.7–4.8 of Groves

1. Control of sample size in cluster sampling

   (a) Why?

   (b) Methods of approximately controlling sample size

2. Example: multistage postal service survey

3. Computing probabilities of selection in multistage sampling

4. Sampling with probability proportional to size

5. Example of a sampling method that is not equal-probability

6. Example of 3 strata and sample size of 1

7. Stratified sampling of clusters

8. Example: traffic exposure of Australian children

9. Estimation with pps sampling

10. Approximate pps sampling using a measure of size

11. Methods of pps sampling of clusters

    (a) Sampling with replacement

    (b) Systematic sampling

    (c) Paired sampling

12. Overview of cluster sampling

lecture 18: Review

1. One-stage cluster sampling

2. Two-stage cluster sampling

3. Sampling with unequal probabilities

lecture 19: Complex surveys, Chapter 7 of Lohr, Chapter 2 of Groves

1. Examples

    (a) Mail pieces and postal service employees

(b) Fragile families study

(c) Sampling Con Edison pipes

2. Assembling design components

3. Sampling weights

4. Estimating distributions

5. Design effects

6. Total survey error

    (a) Survey error and the sampling distribution

        i. Sampling distribution: given your sampling design, the set of all possible samples you could see (along with their probabilities)

        ii. The quantity of interest is θ, the estimate is $\hat{\theta}$

        iii. You only see one value of $\hat{\theta}$, not the whole distribution, but you can estimate important properties of the distribution:

            A. Bias of $\hat{\theta}$: $\mathrm{E}(\hat{\theta}) - \theta$

            B. Variance of $\hat{\theta}$: $\mathrm{var}(\hat{\theta})$

            C. Standard deviation of $\hat{\theta}$ (also called "standard error"): $\mathrm{sd}(\hat{\theta}) = \sqrt{\mathrm{var}(\hat{\theta})}$

            D. The sd of $\hat{\theta}$ is *not* the same as $s_y =$ sd of y_i's

        iv. The *total error*, $\hat{\theta} - \theta$ is unknown (it is a random variable and has a sampling distribution, which depends on the design of the survey)

        v. Mean squared error $= \mathrm{E}((\hat{\theta} - \theta)^2) = \mathrm{bias}^2 + \mathrm{sd}^2$

    (b) Comparing bias and sd:

        i. If $|\mathrm{bias}| > \mathrm{sd}$, then your *first priority* is to reduce bias

        ii. If $\mathrm{sd} > |\mathrm{bias}|$, then your *first priority* is to reduce variance

        iii. Simple hypothetical example:

A. Marketing survey: simple random sample of n respondents from a large population, yes/no responses

B. Bias is expected to be in the range of -10% to 10% (because responses in the survey are not the same as actual purchasing decisions)

C. Sd (ignore finite population correction because large population) is $\sqrt{\frac{1}{n}S_y^2} = \sqrt{\frac{1}{n}\overline{Y}(1-\overline{Y})} \leq 0.5/\sqrt{n}$

D. There is no need for sd to be less than about 5%, so there is no need for n to be more than about $(0.5/0.05)^2 = 100$

E. This conclusion could change if (a) you could reduce the bias, or (b) you are interested in subsets of the population (so that you have to consider other parameters θ and thus other parameters $\hat{\theta}$)

iv. Methods of reducing bias:

A. Be more careful in constructing the frame, collecting accurate responses from all units, etc.

B. Correcting for known biases (for example, by stratification)

C. Look for bias and correct for it (for example, check to see if the proportion of men and women in the sample differs from the population; if so, correct for the possible bias by poststratifying)

v. Methods of reducing variance:

A. Increase the number of primary sampling units in the sample

B. Reduce the variance s_g^2 (see the previous lecture) by improving the design (for example, improved allocation within strata)

C. Reduce the variance s_g^2 using estimation methods (for example, regression estimation or poststratification)

(c) Sources of bias and variance

i. General comments

A. Typically, you can estimate the variance internally using the survey data, but bias is harder to estimate

B. Thus, we try to create unbiased or approximately unbiased estimates using the sampling design

C. Exception: bias of ratio and regression estimates, but this is minor because the bias can be estimated from the survey data alone (and also it is usually small)

D. Sources of error are "variance" if their average effect on $\hat{\theta}$ is 0, "bias" if their average effect is not zero

ii. Sampling bias

A. This is bias that is caused by your sampling method

B. Frame bias: your frame is not the same as your population of interest

C. Example: sampling by kids instead of by families

D. Example: not correcting for blanks in cluster sampling

E. You can correct for these problems by weighting

F. Bias in ratio and regression estimation: these approach 0 in the limit of large sample size

iii. Sampling variance

A. Discussed many times in class (especially in previous lecture)

B. Can be estimated using the survey data

C. Sampling variance approaches 0 in the limit of large sample size

iv. Nonsampling bias

A. This is more of a problem because you *cannot* easily correct for it, and it does *not* approach 0 in the limit of large sample size

B. Undercoverage (or, less important, overcoverage)

C. Nonresponse

D. Observation/measurement error

E. Imprecise measurements

F. Question wording effects

G. These can be studied (for example, randomly assign one question wording to half the respondents and another question wording to the other half)

H. Correcting for these: "adjustments" such as ratio estimation and poststratification

v. Nonsampling variance

A. These are nonsampling effects that have variability

B. Interviewer errors

C. Data coding errors

D. These can be studied (for example, use several interviewers)

(d) Sources of nonsampling error

i. Noncoverage: units that are not on the list

ii. Nonresponse: units that are sampled but do not give you data

A. Unit nonresponse

B. Item nonresponse

C. Example: "not-at-home" in telephone surveys

D. Callbacks

E. Answering machines

F. Refusals

G. Substitutes

lecture 20: Linear regression, Chapters 3 and 4 of Gelman and Hill

1. Interpreting the regression and drawing the regression lines

(a) One predictor

(b) Multiple predictors

(c) Interactions

2. Assumptions and diagnostics

3. Linear transformations

4. Fitting a series of regressions

lecture 21: Logistic regression, Chapter 5 of Gelman and Hill

1. Interpreting the regression and drawing the regression lines

    (a) One predictor

    (b) Multiple predictors

    (c) Interactions

2. Evaluating, checking, and comparing fitted logistic regressions

3. Identifiability and separation

lecture 22: Nonresponse, Sections 8.1–8.4 of Lohr, Chapter 6 of Groves

1. Examples

    (a) NYC Social Indicators Survey

    (b) Survey in homeless shelters

2. Unit and item nonresponse

3. Response rates in actual surveys

4. Survey design and nonresponse

5. Analyzing surveys with weighting adjustments for nonresponse

lecture 23: Correcting for nonresponse, Sections 8.5–8.8 of Lohr, Sections 10.5–10.6 of Groves, Sections 25.1–25.5 of Gelman and Hill

1. Example: missing data and nonresponse in New York City Social Indicators Survey

2. Correction for unit nonresponse

    (a) Example: sample of size 100 has 60 women, 40 men; general population is 52% women, 48% men

    (b) Weighting

    (c) $\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$

    (d) Difficulties

    (e) Example of weights in a national opinion poll: correcting for number of adults in household

3. Correcting for item nonresponse

    (a) Available-case analysis

    (b) Complete-case analysis

    (c) Imputation

        i. Cold deck

        ii. Hot deck

        iii. Multiple imputation

        iv. Examples of good and bad imputations

    (d) Using regression to impute missing data

        i. Example: missing data on incomes in Social Indicators Survey

        ii. Method when only one variable y has missingness

            A. Regress y on predictor variables X: get an estimate $\hat{\beta}$, covariance of estimation V_β, residual s.d. σ

            B. Fix the regression (transformations, interactions, etc.)

            C. Impute the missing data: $\tilde{y} = \tilde{X}\beta +$ noise

            D. Noise is from $\mathrm{N}(0, \sigma^2)$ distribution

            E. Can be done using Stata

        iii. Several variables y with missingness: impute them one at a time and iterate

        iv. Discrete variables: logistic regression

        v. Theoretical understanding: why must we add noise to the imputations

    (e) Some difficulties with regression imputation

    (f) Some tricks

    (g) Partly-missing data

lecture 24: Analysis of surveys collected by others

1. Example: analysis of pre-election polls
2. Design information
3. Stratification and post-stratification weights
4. Estimating population means
5. Estimating population totals
6. More complicated estimands such as regression coefficients

lecture 25: Model-based inference for surveys

1. Placing sampling statistics within a modeling framework
2. Example: measurement of home radon levels
3. Example: missing income values in a survey of New York families
4. Modeling of sample survey data
    - (a) Small-area estimation
        - i. For example: estimating opinions in each state in the U.S. from a national survey
        - ii. Difficulty: small sample sizes in many individual states
        - iii. Consider plot of true vs. predicted values
    - (b) Poststratification into many categories

- i. For example, stratification by sex, ethnicity, age, education
- ii. Similarities to small area estimation:
  - A. Population is divided into many small categories
  - B. Sample sizes within many categories j are too small to get accurate estimates $\bar{y}_j$
- iii. Differences from small area estimation:
  - A. Categories have a natural structure (so you can adjust for one variable at a time)
  - B. In small-area estimation, you are interested in estimates for each cell. In poststratification, you might ultimately be interested only in population means and totals

(c) Regression modeling

- i. Sometimes the goal is to estimate regression coefficients or related summaries such as time trends, not just population means or totals
- ii. Weighted and unweighted regressions
- iii. Weighted regression gives consistent estimate of population regression coefficients
- iv. Best approach is to include, in the model, all variables that affect the probability of selection
  - A. Strata
  - B. Clusters
  - C. Factors that are correlated with nonresponse (for example, sex and age in a telephone survey of people)

(d) Hierarchical analysis of data collected by stratified or cluster sampling

- i. Estimate between-stratum variance and within-stratum variance
- ii. Similarly for clusters

(e) Bayesian inference and shrinkage for small-area estimation and poststratification

- i. Goal is to estimate $\overline{Y}_j$ in small areas (or post-strata) $j = 1, \ldots, J$
- ii. In each area j, sample mean is $\bar{y}_j$, with sampling variance $\mathrm{var}(\bar{y}_j)$
- iii. Estimate $\mathrm{var}(\bar{y}_j)$ using the usual methods (recall from many chapters ago, $\bar{y}_j = \frac{y_j}{n_j}$, which is, in general, a ratio estimate)

iv. Suppose that the values of $\overline{Y}_j$, $j = 1, \ldots, J$, the population means in the small areas, have a variance of τ^2

v. The variance component τ^2 can be estimated using "random effects analysis of variance"

vi. Shrinkage estimate for area j:

$$\begin{aligned} \widehat{\overline{Y}}_j &= \frac{\frac{1}{\mathrm{var}(\bar{y}_j)}\bar{y}_j + \frac{1}{\tau^2}\bar{y}}{\frac{1}{\mathrm{var}(\bar{y}_j)} + \frac{1}{\tau^2}} \\ \mathrm{var}(\widehat{\overline{Y}}_j) &= \frac{1}{\frac{1}{\mathrm{var}(\bar{y}_j)} + \frac{1}{\tau^2}} \end{aligned}$$

vii. Further improvements using the ideas of regression estimation, "hierarchical regression modeling"

5. Imputation of missing data

lecture 26: Looking forward

1. Open problems in survey sampling
2. Combining statistical and practical ideas in survey design and analysis

lecture 27: Review

1. Key ideas of surveys and sampling
2. Simple random sampling
3. Ratio and regression estimation
4. Stratified sampling
5. Survey design and practice
6. Cluster sampling
7. Linear and logistic regression
8. Correction for nonresponse
9. Analysis of surveys collected by others