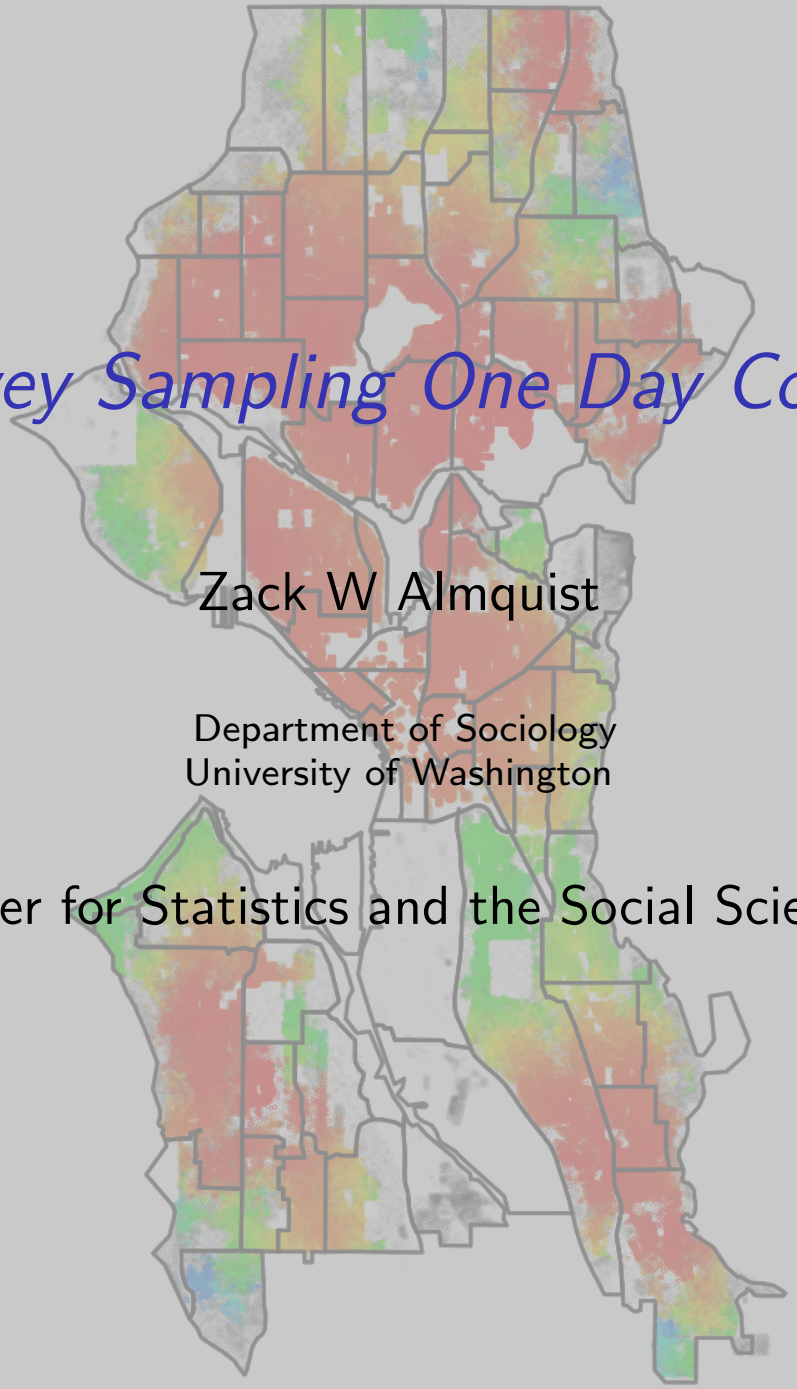
Survey Sampling One Day Course



Zack W Almquist

Department of Sociology
University of Washington

Center for Statistics and the Social Sciences

Regression and Ratio Estimation

- Review of Ratio Estimator and its Properties
- Ratio Estimation with Proportions
- Ratio Estimation Using Weight Adjustments
- Regression Estimation in SRS
- Example: Trees
- Difference Estimation
- Poststratification

Simple Random Sampling

Overview

Goal

- Interest in a function (e.g., mean) from a known population.
  - The number of people who live in Washington.
  - The number of people who own an SUV.
  - The number of people who own an iPhone.
  - The number of people with access to the internet.
  - The average income of an individual who lives in the Seattle.
  - The average age of a person who lives in Minnesota.
  - The percentage of students who passed the state curriculum exams in 4th grade.
  - The proportion of individuals who are registered to vote.
  - The proportion of individuals who are registered to vote who intend to vote.
  - The proportion of individuals who plan to vote Republican.
  - The proportion of individuals who plan to vote Democrat.
  - The proportion of individuals who plan approve of proposition XX.

Overview

Goal

- One option, take a census.

Overview

Goal

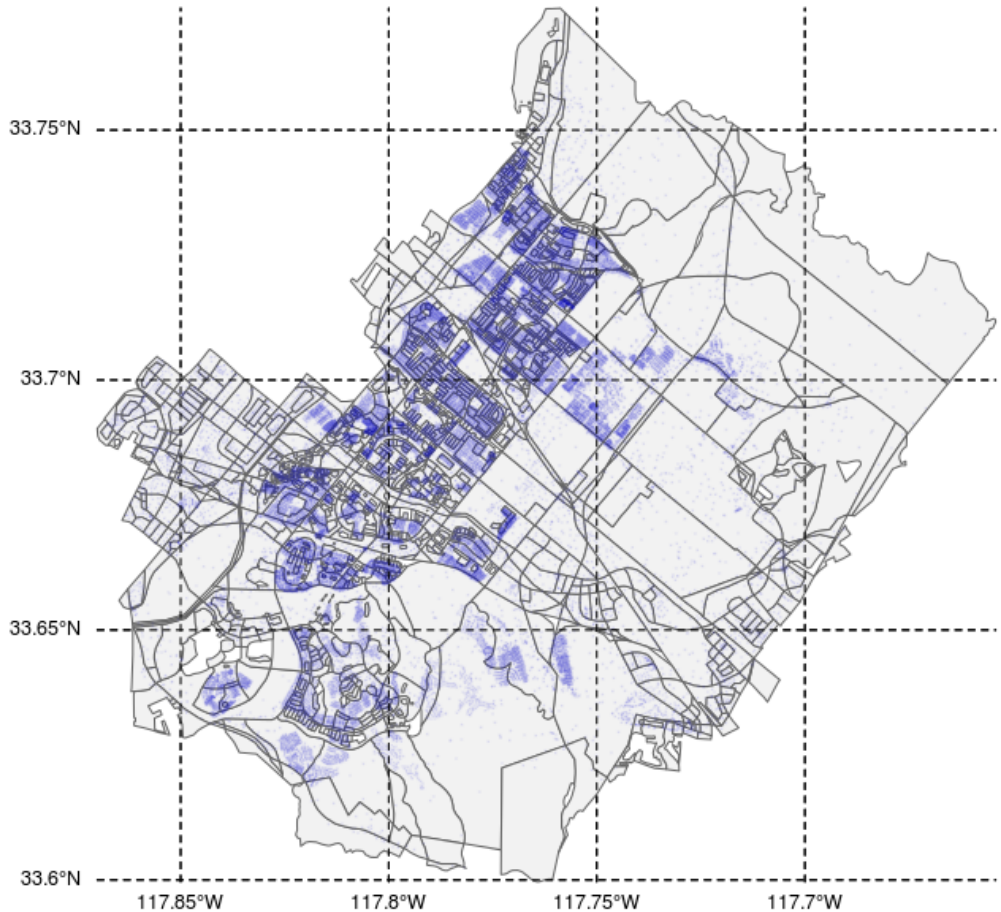
- One option, take a census. **Count Everyone!**

Why sample at all?

- Collecting a sample can be cheaper, quicker and even more accurate than taking a census.
- The design of the sample is much more important than the size.
- Important to balance Sampling vs Nonsampling error.
- Etc.

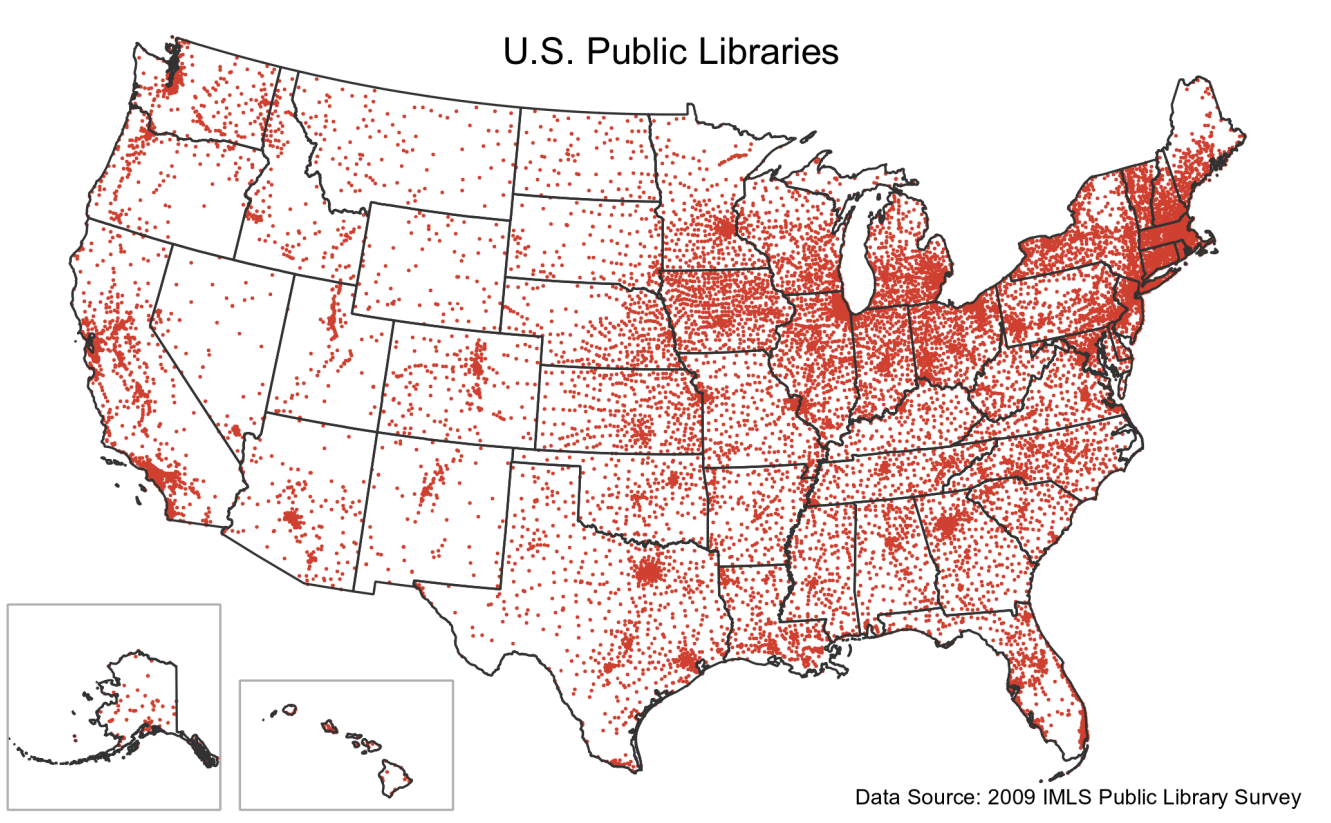
Some Examples of Real Sampling Frames!

HouseHolds in Irvine, CA



Almquist and Butts (2012)

Public Libraries in the US



Overview: Vocabulary

Vocabulary

- **Observation unit**
- **Target population**
- **Sample**
- **Sampled population**
- **Sampling unit**
- **Sampling frame**

Overview: Vocabulary

Vocabulary

- **Observation unit**
- **Target population**
  - **Target population** $= \mathcal{U} =$ population of interest.
- **Sample**
- **Sampled population**
- **Sampling unit**
- **Sampling frame**

Overview: Vocabulary

Vocabulary

- **Observation unit**
- **Target population**
- **Sample**
- **Sampled population**
  - **Sampled population** is the collection of all the possible units we could see in our sample.
- **Sampling unit**
- **Sampling frame**

Overview: Vocabulary

Vocabulary

- **Observation unit**
- **Target population**
- **Sample**
- **Sampled population**
- **Sampling unit**
  - **Sampling unit** is what we can actually sample. We may be interested in individuals but only have a list of households.
- **Sampling frame**

Overview: Vocabulary

Vocabulary

- **Observation unit**
- **Target population**
- **Sample**
- **Sampled population**
- **Sampling unit**
- **Sampling frame**
  - **Sampling frame** is the complete list of sampling units.

Sampling Distribution

Most results in sampling rely on the sampling distribution of a statistic, the distribution of different values of the statistic obtained by the process of taking all possible samples from the population.

- A sampling distribution is an example of a *discrete probability distribution*.

Framework for Probability Sampling

The finite **population**, or **universe**, of N units is denoted by the index set

$$\mathcal{U} = \{1, 2, 3, \ldots, N\}$$

Basic Notation

- $\{y_1, y_2, \ldots, y_N\}$ is a finite population (Notice that this is sometimes referred to as $\mathcal{U}$.
- It contains N units where N is known.
- i is the label identifying a unit.
- The y_i's are the values of the characteristic of interest and are unknown.
- We may select (how?) n units from the population and use them to estimate the population mean.
- $\mathcal{U}$ is a class of 45 students and y_i is the year the mother of the ith student was born.
- Let smp be the labels of the n students selected in the sample. Then the sample mean $= \bar{y}_{smp} = \sum_{i \in smp} y_i / n$ could be a good guess for the population mean $= \sum_{i=1}^{N} y_i / N$.

Simple Random Sampling

- SRS With Replacement (SRSWR)
- SRS Without Replacement (SRS)

SRS

```
s <- sample(1:100, 30, replace = FALSE)
s
```

```
 R >   [1]  1 90 27 72 15 38 95 76 29 58 18 67
 R >  [13] 77 37 26 57 89 50 10 82 41 79 44 56
 R >  [25] 75 33 23 16 97 83
```

SRSWR

```
s <- sample(1:100, 30, replace = TRUE)
s
```

```
 R >   [1] 47 98 46 42 20 94 60 24 93 29 24  6
 R >  [13] 64 36  8 19 59 90 71 52 43 90 59 38
 R >  [25] 97 39 91 62 95 25
```

SRS Weights

The probability that unit i of the population appears in the sample is

$$\pi_i = \frac{n}{N}$$

SRS Weights

The sampling weight for each unit in the sample is

$$w_i = \frac{1}{\pi_i} = \frac{N}{n}$$

SRS Weights

The sampling weight for each unit in the sample is

$$w_i = \frac{1}{\pi_i} = \frac{N}{n}$$

- Each unit in the sample can be thought of as representing $\frac{N}{n}$ units in the population.
  - E.g., A sample of 300 from Minnesota, means that an individual from the sample represents $\frac{5{,}420{,}380}{300} = 18,067.93$ Minnesotans.

Estimates and Variance

Population Quantity	Estimator	Standard Error of Estimator
Population total, $t = \sum_{i=1} y_i$	$\hat{t} = \sum_{i \in S} w_i y_i = N\bar{y}$	$N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$
Population mean, $\bar{y}_{\mathcal{U}} = \frac{t}{N}$	$\frac{\hat{t}}{N}\frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i} = \bar{y}$	$\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$
Population proportion, p	$\hat{p}$	$\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}$

- $\bar{y}_{\mathcal{U}} = \sum_{i=1}^{N} y_i / n$
- $\bar{y} = \sum_{i \in S} y_i / n$
- $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^2 / (N-1)$
- $s^2 = \sum_{i \in S} (y_i - \bar{y})^2 / (n-1)$
- **Key Properties:** Unbiased or biased and S.E.

Examples: SRS

```
readGSS <- function(address, zipFile) {
    require(foreign)  ### for read.dta function
    fileName <- paste(address, zipFile, sep = "")  ##full address of zip file
    zipdir <- tempfile()  ### Create temp file
    dir.create(zipdir)  ### Create a folder in the temp file
    download.file(fileName, destfile = paste(zipdir,
        zipFile, sep = "/"))  ## Download the zip file
    unzip(paste(zipdir, zipFile, sep = "/"),
        exdir = zipdir)  ### Extract the zip file
    files <- list.files(zipdir)  ## Get the name (assumes it is the second file after the zip)
    read.dta(paste(zipdir, files[2], sep = "/"))  ### read in the file and output data.frame
}
GSS2012 <- readGSS("https://gss.norc.org/documents/stata/",
    "2012_stata.zip")
lookup <- function(data, var) {
    out <- names(data)[grep(var, names(data))]
    print(head(data[, out]))
    invisible(out)
}
```

Examples: SRS

```
vot <- lookup(GSS2012, "vot")
```

```
 R >  [1] did not vote ineligible
 R >  [3] voted        voted
 R >  [5] voted        voted
 R >  6 Levels: voted ... No answer
```

```
email <- lookup(GSS2012, "email")
```

```
 R >    emailhr emailmin
 R >  1       4       NA
 R >  2      NA       NA
 R >  3      NA       NA
 R >  4       2       NA
 R >  5       0       30
 R >  6      NA        5
```

Examples: SRS

```
educ <- lookup(GSS2012, "educ")
```

```
 R >    sei10educ spsei10educ pasei10educ
 R >  1      79.4          NA        65.6
 R >  2      72.5          NA          NA
 R >  3      95.6        98.9          NA
 R >  4      93.7        55.4        85.7
 R >  5      54.8          NA        76.9
 R >  6      97.3          NA        81.0
 R >    masei10educ      coneduc educ
 R >  1          NA         <NA>   16
 R >  2        73.3    only some   12
 R >  3        58.8    only some   12
 R >  4          NA    only some   13
 R >  5        37.0 a great deal   16
 R >  6        57.6         <NA>   19
 R >                  inteduc maeduc
 R >  1                  <NA>     16
 R >  2 moderately interested     12
 R >  3 moderately interested     15
 R >  4 not at all interested     16
 R >  5       very interested     12
 R >  6                  <NA>     12
 R >        nateduc   nateducy paeduc sexeduc
 R >  1        <NA> too little     16   favor
 R >  2        <NA> too little     NA    <NA>
 R >  3  too little       <NA>     NA    <NA>
 R >  4        <NA> too little     18   favor
 R >  5        <NA> too little     NA   favor
 R >  6 about right       <NA>     14   favor
 R >    speduc
 R >  1     NA
 R >  2     NA
 R >  3     16
 R >  4     16
 R >  5     NA
 R >  6     NA
```

Examples: SRS

```
vote <- GSS2012[, vot]
Ivote <- as.numeric(vote == "voted")
Ivote <- Ivote[!is.na(Ivote)]

## Population Total
N <- length(Ivote)
N
```

```
 R >  [1] 1948
```

```
## Population Total Voters
t <- sum(Ivote)
t
```

```
 R >  [1] 1304
```

```
## Population Mean/Proportion
p <- mean(Ivote)
p
```

```
 R >  [1] 0.6694045
```

```
set.seed(18744)
s <- sample(1:t, 50, replace = FALSE)

vote_sample <- Ivote[s]
phat <- mean(vote_sample)
phat
```

```
 R >  [1] 0.74
```

Examples: SRS

```
sephat <- sqrt((1 - 50/N) * (phat * (1 -
    phat)/(50 - 1)))
sephat
```

```
 R >  [1] 0.06185262
```

```
t_hat <- N * phat
t_hat
```

```
 R >  [1] 1441.52
```

```
se_t_hat <- N * sephat
se_t_hat
```

```
 R >  [1] 120.4889
```

Confidence Interval

$$\text{estimate} \pm z_{\alpha/2} \text{ SE(estimate)}$$

Confidence Interval

$100(1-\alpha)\%$ CI for the Population Mean

$$\left[\bar{y} - z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}\ ,\ \bar{y} + z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}\right]$$

100(1 − α)% CI for the Population Mean (t-distribution)

$$\left[\bar{y} - t_{\alpha/2,n-1}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}\ ,\ \bar{y} + t_{\alpha/2,n-1}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}\right]$$

- For large samples, $t_{\alpha/2,n-1} \approx z_{\alpha/2}$.
- In smaller samples, using $t_{\alpha/2,n-1}$ instead of $z_{\alpha/2}$ produces wider CI.
- In practice (and most software) one uses t approximation (b/c it is more conservative).

Examples: total and p CI for SRS

```
## Total
t_hat + qt(0.025, 50 - 1) * se_t_hat
```

```
 R >  [1] 1199.388
```

```
t_hat + qt(0.025, 50 - 1, lower.tail = FALSE) *
    se_t_hat
```

```
 R >  [1] 1683.652
```

```
## Proportion
phat + qt(0.025, 50 - 1) * sephat
```

```
 R >  [1] 0.6157025
```

```
phat + qt(0.025, 50 - 1, lower.tail = FALSE) *
    sephat
```

```
 R >  [1] 0.8642975
```

Sample Size Estimation

Sample size estimation

Sample size is never large enough. As n increases, we estimate more interactions, which typically are smaller and have relatively larger standard errors than main effects.
- Andrew Gelman (Professor in Statistics at Columbia University)

Two Core Objectives

- Measure with a precision:
  - Precision analysis
- Assure that the difference is correctly detected
  - Power analysis

Sample size estimation

Precision

Whenever we propose to estimate population parameters, such as, population mean, proportion, or total, we need to estimate with a **specified level of precision**

We like to specify a sample size that is sufficiently large to ensure a high probability that errors of estimation can be limited within desired limits

Sample size estimation

Power

The power of a test is the probability of rejecting the null hypothesis if it is incorrect.

Lohr Algorithm 2.6

1. Ask: "What is expected of the sample, and how much precision do I need?"
2. Find an equation relating the sample size n and your expectations of the sample.
3. Estimate any unknown quantities and solve for n.
   - Use known data
   - Simulate
   - Use prior knowledge to build bounds

* **Lessons**: You will likely realize that the sample you should get is more expensive than you can possible do; carefully go through your assumptions and ask what outcomes, what level of precision, etc you are willing to give up...

Specify the Tolerable Error

$$\Pr(|\bar{y} - \bar{Y}_{\mathcal{U}}| \leq e) = 1 - \alpha$$

Goal

- The investigator must decide on reasonable values for α and e
  - e is called the **margin of error** in many surveys
  - Lohr notes that $e = 0.03$ and $\alpha = 0.05$ in many surveys

Relative precision

Can be achieved by controlling Coefficient of Variation (CV) rather than the absolute error; in this case, if $\bar{Y}_{\mathcal{U}} \neq 0$ the precision may be expressed as

$$\Pr\left(\left|\frac{\bar{y} - \bar{Y}_{\mathcal{U}}}{\bar{Y}_{\mathcal{U}}}\right| \leq r\right) = 1 - \alpha$$

Find an equation

The simplest equation relating the precision and sample size comes from the formal for confidence intervals. To obtain absolute precision e, find a value of n that satisfies

$$e = z_{\alpha/2}\sqrt{(1 - \frac{n}{N})}\frac{S}{\sqrt{n}}$$

SRSWR (Easy case)

$$n_0 = \left(\frac{z_{\alpha/2} S}{e} \right)^2$$

- Obvious if $n_0 > N$ we just a census with $n = N$

SRS and Precision

SRS

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{z^2_{\alpha/2} S^2}{e^2 + \frac{z^2_{\alpha/2} S^2}{N}}$$

Example 2.12

- Many public opinion polls specify using a sample size of about 1,100.
- That number comes from rounding the value of n_0 (from SRSWR) and noting that the population size is so large that the fpc can be ignored (i.e., 1100/300,000,000=0.000003667).
- For large populations, it is the size of the sample, not the proportion of the population that is sampled, that determines the precision.

Example 2.12

- When interested in a proportion, we can use $1/4$ as an upper bound for S^2. For other quantities, S^2 must be estimated or guessed at. Some methods for estimating S^2 include:
  1. Use sample quantities obtained when pretesting your survey (e.g., pilot study)
  2. Use previous studios or data available in the literature.
  3. Make an a guess! (There a better and worse ways to do such a thing)

Systematic Sampling

Systematic Sampling

- **Systematic sampling** is used as a proxy for SRS when
  - No list of the population exists
  - Or when the list is in roughly random order

Systematic Sampling

To obtain a systematic sample, choose a sample size n. If N/n is an integer, let $k = N/n$; otherwise, let k be the next integer after N/n. Then find a random integer R between 1 and k, which determines the sample to be units numbered R, $R + k$, $R + 2k$, ...

For example, to select a systematic sample of 45 students from the list of 45,000 students at a university, the sampling interval k is 1000. Suppose the random integer we choose is 597. Then the students numbered 597, 1597, 2597, ..., 44,957 would be in the sample.

- If the list of students is ordered by randomly generated student identification numbers, we shall probably obtain a sample that will behave much like an SRS.

Systematic Sampling

Issues

- If the population is in some periodic or cyclical order.
  - E.g., If male and female names alternate in the list
- If populations are listed in increasing or decreasing order
- Requires a sampling frame and defining the target population

Design-based and Model-based Frameworks

Overview

- **Design-based** inference: population values are fixed, inference is based on probability distribution of sample selection. Obviously this assumes that we have a probability sample (or "quasi-randomization", where we pretend that we have one)
- **Model-based** inference: survey variables are assumed to come from a statistical model: probability sampling is not the basis for inference, but useful for making the sample selection **ignorable**. (see e.g. Gelman et al., 2003)

Design-Based Approach

Design-Based Approach

Also known as Randomization Theory

- No distributional assumptions are made about the y_i's to ascertain that $\bar{y}$ is unbiased for estimating $\bar{y}_{\mathcal{U}}$
- We treat the y_i's as fixed but unknown numbers
- The RV used in randomization theory inference indicate which population units are in the sample

Design-Based Approach

$$Z_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

Design-Based Approach

The Mean

$$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}$$

Design-Based Approach

RV

The Z_i's are the only random variable because, according to randomization theory, the y_i's are fixed quantities.

Design-Based Approach

RV

When we choose an SRS of n units out of the N units in the population, $\{Z_1, \ldots, Z_N\}$ are identically distributed Bernoulli random variables with

$$\pi_i = \Pr(Z_i = 1) = \Pr(\text{ select unit } i \text{ in the sample}) = \frac{n}{N}$$

and

$$P(Z_i = 0) = 1 - \pi_i = 1 - \frac{n}{N}$$

Proof that $\bar{y}$ is unbiased

$$P(Z_i = 1) = \frac{\#\text{ of samples including unit } i}{\#\text{ number of possible samples}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Proof that $\bar{y}$ is unbiased

Notice that $E[Z_i] = E[Z_i^2] = \frac{n}{N}$.

Proof that $\bar{y}$ is unbiased

$$E[\bar{y}] = E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] = \cdots = \bar{y}_{\mathcal{U}}$$

Variance of $\bar{y}$

$$V(Z_i) = E[Z_i^2] - (E[Z_i])^2 = \cdots = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

for $i \neq j$

$$\begin{aligned} E[Z_i Z_j] &= \Pr(Z_i = 1 \text{ and } Z_j = 1) \\ &= \Pr(Z_j = 1 | Z_i = 1) P(Z_i = 1) \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right) \end{aligned}$$

Variance of $\bar{y}$

Note that b/c the population is finite, the Z_i's are not quite independent, i.e., if we know that unit I is in the sample, we do have a small amount of information about whether unit j is in the sample, reflected in the conditional probability $\Pr(Z_j = 1 | Z_i = 1)$.

Variance of $\bar{y}$

$$
\begin{aligned}
& i \neq j \\
Cov(Z_i, Z_j) &= E[Z_i Z_j] - E[Z_i]E[Z_j] \\
&= \frac{n-1}{N-1}\frac{n}{N} - \left(\frac{n}{N}\right)^2 \\
&= -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right)
\end{aligned}
$$

Variance of $\bar{y}$

The negative covariance of Z_i and Z_j is the source of the fpc!

Variance of $\bar{y}$

$$V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n}$$

Variance of $\bar{y}$

To show that $V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n}$ is unbiased estimator of the variance we need to show that $E[s^2] = S^2$. Remember, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_{\mathcal{U}})^2$.

Variance of $\bar{y}$

$$E\left[\sum_{i \in S}(y_i - \bar{y}_U)^2\right] = \cdots = (n-1)S^2$$

Variance of $\bar{y}$

$$E\left[\frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})\right] = E[s^2] = S^2.$$

Model Based Approach

Model Based Approach

- In the design-based approach the y_i's could be anything! Only the inclusion probabilities (the Z_i's were random!)
- You may be more familiar with the y_i's being random variables?
- For example, say $Y_1, Y_2, \ldots, Y_n$ were independent identically distributed from a $N(\mu, \sigma)$.
- Where estimators/etc were derived from the independence and distributional assumptions.

Model Based Approach

We can extend this approach to sampling by thinking of random variables $Y_1, Y_2, \ldots, Y_N$ generated from some model.

- The actual values for the finite population, $y_1, y_2, \ldots, y_N$, are one realization of the random variables.

Model Based Approach

- The joint probability distribution of $Y_1, Y_2, \ldots, Y_N$ supplies the link between units in the sample and units not in the sample in this **model-based** approach.
- A link that is missing in the randomization approach.
- Here, we sample $\{y_i,\ i \in S\}$ and use these data to predict the unobserved values $\{y_i,\ i \notin S\}$.
- Thus, the problems in finite population sampling may be thought of as prediction problems.

Model Based Approach

In SRS, a simple model is

$Y_1, \ldots, Y_N$ independent with $E_M[Y_j] = \mu$ and $V_M = \sigma^2$.

- M denotes that the expectation is taken over the model.
- μ and σ^2 represent unknown infinite population parameters, not the finite population quantities.
- This model makes assumptions about the observations not in the sample; namely, that they have the same mean and variance as observations that are in the sample.
- We take a sample S and observe the values y_i for $i \in S$. That is, we realizations of the RV Y_i for $i \in S$.
- The other observations in the population $\{y_i,\ i \notin S\}$ are also realizations of random variables, but we do not see those.

total

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in S} y_i + \sum_{i \notin S} y_i$$

and is one possible value that can be taken on by the random variable

$$T = \sum_{i=1}^{N} Y_i = \sum_{i \in S} Y_i + \sum_{i \notin S} Y_i$$

Model Based Approach

total

- We know the values $\{y_i,\ i \in S\}$.
- To estimate t for our sample, we need to predict values for the y_i's not in the sample.
- This is where our model of the common mean μ comes in.
- The least squares estimator of μ from the sample is $\bar{Y}_S = \sum_{i \in S} Y_i / n$.
- That is the best linear unbiased predictor (under the model) of each unobserved RV.

Model Based Approach

total

$$\hat{T} = \sum_{i \in S} Y_i + \sum_{i \notin S} \hat{Y}_i = \sum_{i \in S} Y_i + \frac{N-n}{n} \sum_{i \in S} Y_i = \frac{N}{n} \sum_{i \in S} Y_i$$

total

- The estimator $\hat{T}$ is model-unbiased: if the model is correct, then the average of $\hat{T} - T$ over repeated realizations of the population is

$$E_M[\hat{T} - T] = \frac{N}{n} \sum_{i \in S} E_M[Y_i] - \sum_{i=1}^{N} E_M[Y_i] = 0.$$

Model Based Approach

Notice the difference between finding expectations under the model-based approach and under the design-based approach. In the model-based approach, the Y_i's are the random variables, and the sample has no information for calculating expected values, so we can take the $\sum_{i \in S}$ outside the expected value. In the design-based approach, the random variables are contained in the sample S.

Model Based Approach

Mean Square Error

$$(\hat{t} - t)^2 = \left[\frac{N}{n} \sum_{i \in S} y_i - \sum_{i=1}^{N} y_i \right]^2 .$$

Mean Square Error

$$E_M\left[(\hat{T}-T)^2)\right] = \cdots = N^2(1-\frac{n}{N})\frac{\sigma^2}{n}.$$

- In practice, if the model assumed were adopted, you would estimate σ^2 by the sample variance s^2.

Model Based Approach v Design-Based Approach

- Thus the design-based approach and the model-based approach (in this case) lead to the same estimator of the population total and the same variance estimator. However, this is not guaranteed! If one selects a different model, the estimator could be different.
- The CI is also the same under this model!

Stratified Random Sampling

Stratified Sampling

Why Stratified Sampling

Populations can be extremely heterogenous and it is easy for a single SRS to miss some of this heterogeneity.

Examples:

- We know that on average men earn more than women.
- We know that NYC residents typically pay more for housing than residents of Des Moines.
- That rural residents shop for groceries less often than urban residents.
- That internet usage varies by age.
- That cell only households are more likely to be younger.
- That income varies by state (as does COL).

Stratified Sampling

Goal

- To maximize the representative nature of our sample using known information.

Why we use Stratified Sampling

We want to be protected from the possibility of obtaining a really bad sample.

- When taking an SRS of size 100 from a population of 1000 male and 1000 female students
  - Obtaining a sample with no or very few males is theoretically possible

Why we use Stratified Sampling

We want to be protected from the possibility of obtaining a really bad sample.

- When taking an SRS of size 100 from a population of 1000 male and 1000 female students
  - Obtaining a sample with no or very few males is theoretically possible
  - This is bad!

Why we use Stratified Sampling

We want to be protected from the possibility of obtaining a really bad sample.

- When taking an SRS of size 100 from a population of 1000 male and 1000 female students
  - Obtaining a sample with no or very few males is theoretically possible
- A solution is to take a SRS of 50 for each gender category.

Why we use Stratified Sampling

We may want data of known precision for subgroups of the population.

- These subgroups can easily be strata (with careful definitions).
- Notice that this procedure can result in over sampling one group, e.g.,
  - Two researchers interested in sampling graduates of electrical and mechanical engineering programs at public universities in SoCal stratified their sample by gender.
  - B/c there are many more males in engineering, taking an equal sample in each strata will result in oversampling the female engineers (this might also be done for precision)

Why we use Stratified Sampling

A stratified sample may be more convenient to administer and may result in a lower cost for the survey.

- For example: Sampling frames may be constructed differently in different strata, or different sampling designs or field procedures may be used.
- E.g., A combination of internet based survey instruments, mail and telephone could be used (note this is not without controversy though. . .).

Why we use Stratified Sampling

Stratified sampling often gives more precise (having lower variance) estimates for population means and totals.

- Examples:
  - Persons of different ages tend to have different blood pressures.
  - If studying the concentration of plants in an area, one would stratify by type of terrain.
- Stratification works for lowering the variance b/c the variance within each stratum is often lower than the variance in the whole population. Prior knowledge can be used to save money in the sampling procedure.

Theory of Stratified Sampling

stratified random sampling population quantities:

- y_{hj} = value of the jth unit in the stratum h
- $t_h = \sum_{j=1}^{N_h} y_{hj}$ = population total in stratum h
- $t = \sum_{h=1}^{H} t_h$ = population total
- $\bar{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$ = population mean in stratum h
- $\bar{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}}{N}$ overall population mean
- $S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1}$ = population variance in stratum h

Theory of Stratified Sampling

Corresponding quantities for the sample, using SRS estimators within each stratum are:

- $\bar{y}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj}$
- $\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} = N_h \bar{y}_h$
- $s_h^2 = \sum_{j \in S_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$

Theory of Stratified Sampling

Suppose we only sampled the hth stratum. In effect, we have a population of N_h units and take an SRS of n_h units.

- We could estimate $\bar{y}_{hU}$ by $\bar{y}_h$.
- t_h by $\hat{t}_h = N_h \bar{y}_h$.
- The population total is $t = \sum_{h=1}^{H} t_h$, see we estimate t by

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h$$

- We can estimate $\bar{y}_U$ with:

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Example: Homework Hours

Example: Homework Hours

National Education Longitudinal Study of 1988

`http://nces.ed.gov/surveys/nels88/pdf/QuickGuide.PDF`

Example: Homework Hours

```
## NELS-88 homework is in hours
library(foreign)
imm10 <- read.dta(file = "https://stats.oarc.ucla.edu/stat/stata/examples/mlm_imm/imm10.dta")
head(imm10)
```

```
 R >    schid stuid   ses    meanses homework
 R >  1  7472     3 -0.13 -0.4826087        1
 R >  2  7472     8 -0.39 -0.4826087        0
 R >  3  7472    13 -0.80 -0.4826087        0
 R >  4  7472    17 -0.72 -0.4826087        1
 R >  5  7472    27 -0.74 -0.4826087        2
 R >  6  7472    28 -0.58 -0.4826087        1
 R >    white parented public ratio percmin
 R >  1     1        2      1    19       0
 R >  2     1        2      1    19       0
 R >  3     1        2      1    19       0
 R >  4     1        2      1    19       0
 R >  5     1        2      1    19       0
 R >  6     1        2      1    19       0
 R >    math sex race sctype cstr scsize
 R >  1   48   2    4      1    2      3
 R >  2   48   1    4      1    2      3
 R >  3   53   1    4      1    2      3
 R >  4   42   1    4      1    2      3
 R >  5   43   2    4      1    2      3
 R >  6   57   2    4      1    2      3
 R >    urban region schnum
 R >  1     2      2      1
 R >  2     2      2      1
 R >  3     2      2      1
 R >  4     2      2      1
 R >  5     2      2      1
 R >  6     2      2      1
```

Example: Homework Hours

```
#### Say we are interested in hour
#### stratified on public/private
homework.public <- imm10$homework[imm10$public ==
    1]
homework.private <- imm10$homework[imm10$public ==
    0]

N <- length(imm10$homework)
N1 <- length(homework.public)
N2 <- length(homework.private)

n1 <- ceiling(0.1 * length(homework.public))
n2 <- ceiling(0.1 * length(homework.private))
```

Example: Homework Hours

```
set.seed(98374)
s1 <- sample(homework.public, n1, replace = FALSE)
s2 <- sample(homework.private, n2, replace = FALSE)
```

Example: Homework Hours

```
ybar1 <- mean(s1)
ybar1

 R >  [1] 1.5

ybar2 <- mean(s2)
ybar2

 R >  [1] 4.571429

t.hat.1 <- N1 * ybar1
t.hat.1

 R >  [1] 289.5

t.hat.2 <- N2 * ybar2
t.hat.2

 R >  [1] 306.2857
```

Example: Homework Hours

```
t.str <- t.hat.1 + t.hat.2
t.str
```

```
 R >  [1] 595.7857
```

```
ybar.str <- t.str/N
ybar.str
```

```
 R >  [1] 2.291484
```

Example: Homework Hours

```
v.t.str <- (1 - n1/N1) * (N1^2) * (var(s1)/n1) +
    (1 - n2/N2) * (N2^2) * (var(s2)/n2)
v.t.str
```

```
 R >  [1] 3829.367
```

```
se.t.str <- sqrt(v.t.str)
se.t.str
```

```
 R >  [1] 61.88188
```

```
v.ybar.str <- v.t.str/N^2
v.ybar.str
```

```
 R >  [1] 0.05664744
```

```
se.ybar.str <- sqrt(v.ybar.str)
se.ybar.str
```

```
 R >  [1] 0.2380072
```

Example: Homework Hours

```
## CI
t.str + qt(0.025, 24 - 2) * se.t.str
```

```
 R >  [1] 467.4506
```

```
t.str + qt(0.025, 24 - 2, lower.tail = FALSE) *
    se.t.str
```

```
 R >  [1] 724.1209
```

```
ybar.str + qt(0.025, 24 - 2) * se.ybar.str
```

```
 R >  [1] 1.797887
```

```
ybar.str + qt(0.025, 24 - 2, lower.tail = FALSE) *
    se.ybar.str
```

```
 R >  [1] 2.78508
```

Stratified Random Sampling: Estimators

- If $\hat{t}_h$ is an unbiased estimator or t_h, $h = 1, \ldots, H$, then $\hat{t} = \sum_{i=1}^{H} \hat{t}_h$ is an unbiased estimator for t.
  - Thus if we have an SRS for each stratum, $\hat{t}_h$ will be unbiased and $\hat{t}$ is also unbiased.
- If the stratum samples are independently selected, then:

$$V(\hat{t}_{str}) = V\left(\sum_{h=1}^{H} \hat{t}_h\right) = \sum_{h=1}^{H} V(\hat{t}_h)$$

- If $\hat{V}(\hat{t}_h)$ is unbiased then $\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h)$ is unbiased.
- $\bar{y} = \frac{\hat{t}_{str}}{N}$ is an biased estimator of $\bar{y}_U$ if $\hat{t}$ is an unbiased estimator of t_U and $V(\bar{y}) = \frac{1}{N^2} V(\hat{t}_{str})$ so that $V(\bar{y}) = \frac{1}{N^2} \hat{V}(\hat{t}_{str})$.

- An alternative form for the estimator of $\bar{y}_U$ is given by

$$\bar{y}_{str} = \frac{1}{N}\hat{t}_{str} = \frac{1}{N}\sum_{h=1}^{H}\hat{t}_h = \frac{1}{N}\sum_{h=1}^{H}N_h\bar{y}_h = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{y}_h$$

  - A weighted average of the stratum means (weighted by the proportional stratum size). This indicates that we only need to know the relative stratum sizes, not the actual sizes to estimate the population mean.
- The variance of $\bar{y}_{str}$ may then be expressed as:

$$V(\bar{y}) = V\left(\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{y}_h\right) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 V(\bar{y}_h) \text{ (under independence)}$$

Stratified Random Sampling: Estimators

Note the results just discussed are true for any sampling (probability) plans within each stratum, not just simple random sampling. These general results fall under the heading of *stratified sampling*.

Stratified Random Sampling: Estimators

Stratified random sampling means independent simple random samples (SRSs) taken within each stratum. Under this setting, the stratified estimator of the population mean and total can be derived as follows.

Stratified Random Sampling: Estimators

- Within stratum h: $\hat{t}_h = N_h \bar{y}$, where $\bar{y}_h$ is the sample mean in stratum h.
- $\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h$
-

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N}\right) \frac{S_h^2}{n_h} = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

- $\bar{y}_{str} = \frac{\hat{t}}{N} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{y}_h$
-

$$V(\bar{y}) = \frac{1}{N^2} V(\hat{t}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

- Note that in practice you will use the estimator s_h^2 instead of the population parameter S_h^2.

Example

TVs. . .

Suppose we want to estimate the average number of hours of TV watched in the previous week for all adults in some county. Suppose also that the populace of this county can be grouped naturally into 3 strata (town A, town B, rural) as summarized in the table at the top of the next page.

Statistic	Town A	Town B	Rural
h	1	2	3
N_h	155	62	93
n_h	20	8	12
$\bar{y}_h$	33.90	25.12	19
s_h	5.95	15.24	9.36
$\hat{t}_h$	5254.5	1557.4	1767
$N = 310$			

Example

Statistic	Town A	Town B	Rural
h	1	2	3
N_h	155	62	93
n_h	20	8	12
$\bar{y}_h$	33.90	25.12	19
s_h	5.95	15.24	9.36
$\hat{t}_h$	5254.5	1557.4	1767

$N = 310$

$$\hat{t} = \hat{t}_1 + \hat{t}_2 + \hat{t}_3 = 8578.9$$

$$\bar{y} = \frac{\hat{t}}{N} = \frac{8578.9}{310} = 27.7$$

$$\bar{y}_{str} = \sum_{h=1}^{3} \left(\frac{N_h}{N} \right) \bar{y}_h = 27.7$$

$$\hat{V}(\bar{y}) = \sum_{h=1}^{3} \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_h^2}{n_h} = 1.97$$

$$SE(\bar{y}) = 1.40$$

Example

Statistic	Town A	Town B	Rural
h	1	2	3
N_h	155	62	93
n_h	20	8	12
$\bar{y}_h$	33.90	25.12	19
s_h	5.95	15.24	9.36
$\hat{t}_h$	5254.5	1557.4	1767
$N = 310$			

A 95% CI for $\bar{y}_U$ is given by:

$$\bar{y}_{st} \pm t(\alpha/2, df)SE(\bar{y}_{str})$$
$$27.7 \pm (2.026)(1.4) = (24.76, 30.43)$$

Example

Statistic	Town A	Town B	Rural
h	1	2	3
N_h	155	62	93
n_h	20	8	12
$\bar{y}_h$	33.90	25.12	19
s_h	5.95	15.24	9.36
$\hat{t}_h$	5254.5	1557.4	1767
$N = 310$			

How many degrees of freedom are associated with this t-based critical value? How do we determine these degrees of freedom?

Example

Statistic	Town A	Town B	Rural
h	1	2	3
N_h	155	62	93
n_h	20	8	12
$\bar{y}_h$	33.90	25.12	19
s_h	5.95	15.24	9.36
$\hat{t}_h$	5254.5	1557.4	1767

$N = 310$

$$n - |n_H| = (20 + 8 + 12) - 3 = 37$$

Stratification Principle

- Recall that choosing strata which make the units homogeneous within and heterogeneous between is considered a "good" choice of strata.
- Stratification can often be very effective with just a few strata; more strata lead to diminishing returns with greater effort. Too many strata will usually require more effort to sample and lead to less heterogeneity between strata.
- Stratified random sampling is really nothing more than using a categorical auxiliary variable in the design phase of a study. In the TV example, we assume that where a person lives is associated with the number of hours of TV watched. Here, the auxiliary variable is the stratum (where a person lives).

Stratification Principle

- Ratio and regression estimation are examples of using a continuous auxiliary variable in the estimation phase of a study, after we have collected the data.
- Using a categorical variable in the estimation (rather than the design) phase of a study can be done with post-stratification, discussed later in these notes.
- Note that a continuous variable can be used as an auxiliary variable in the design phase by dividing the range of values into categories.
- Note also that a continuous auxiliary variable could be used as a categorical variable in the design phase of a study by stratification and as a continuous variable in the estimation phase with ratio or regression estimation. The stratification would be to ensure that the sample includes values across the range of the auxiliary variable x which will aid us in deter- mining the appropriate relationship between x and y in ratio or regression estimation.

Allocation in Stratified Random Sampling

In planning a study requiring stratification of the population, an important consideration is how to allocate a total sample size n among the H identified strata. We have discussed four types of allocation in this course so far.

Allocation in Stratified Random Sampling

1. **Equal**: If the strata are presumed to be of roughly equal size, and there is no additional information regarding the variability or distribution of the response in the strata, equal allocation to the strata is probably the best choice: $n_h = \frac{n}{H}$.
2. **Proportional**: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes: $n_h = \left(\frac{N_H}{N}\right) n$.
   - he example where people in three strata were sampled for the # of hours of TV watched is an example of proportional allocation.
   - Proportional allocation is optimal if the the stratum variances are all the same.

3. **Optimum**: The allocation which minimizes the variance and cost of the estimator of the mean (and total) is given by:

$$n = \frac{(C - c_0)\sum_{h=1}^{H} N_h s_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h s_h \sqrt{c_h}}$$

$$n_h = \frac{N_h s_h / \sqrt{c_h}}{\sum_{l=1}^{H} N_l s_l / \sqrt{c_l}} \times n$$

Allocation in Stratified Random Sampling

4. **Optimum (Neyman)**: The allocation which minimizes the variance of the estimator of the mean (and total) is given by:
$n_h = \frac{N_h S_h}{\sum_{k=1}^{H} N_h S_h} n.$
   - Such an allocation rule will minimize $Var(\bar{y}_{str})$ for a given n.
   - Note that the larger the variance S_h^2 is for stratum h, the larger the sample size n_h required. This makes sense intuitively, as populations with higher variability require more sampling effort to attain the same degree of precision as those with lower variability.
   - Note also that the larger the population size N_h of stratum h, the larger the sample size n_h required.
   - For optimum allocation, we need to know or at least be able to make a good guess at the stratum standard deviations, s_h, $h = 1, \ldots, H$ (actually, we only need to know the relative sizes of the standard deviations).
   - Finally, note that if the stratum standard deviations are all equal, the optimum allocation is proportional allocation.

Allocation in Stratified Random Sampling

Recall the estimated mean and corresponding variance for stratified random sampling:

$$\bar{y}_{str} = \sum_{h=1}^{H} \left(\frac{N_H}{N} \right) \bar{y}$$

$$V(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_h^2}{n_h}$$

Allocation in Stratified Random Sampling

Recall we had found from our sample:

Town A	Town B	Rural
$N_1 = 155$	$N_2 = 62$	$N_3 = 93$
$s_1 = 5.946$	$s_2 = 15.24$	$s_3 = 9.36$

Using these sample standard deviations as "guesses" of the true standard deviations, then under optimum allocation, we compute:

$$n_1 = n\left(\frac{(155)(5.946)}{(155)(5.946)+(62)(15.24)+(93)(9.36)}\right) = n(0.337)$$

$$n_2 = n\left(\frac{(62)(15.24)}{(155)(5.946)+(62)(15.24)+(93)(9.36)}\right) = n(0.345)$$

$$n_3 = n - n_1 - n_2 = n(0.318)$$

Allocation in Stratified Random Sampling

Recall we had found from our sample:

Town A	Town B	Rural
$N_1 = 155$	$N_2 = 62$	$N_3 = 93$
$s_1 = 5.946$	$s_2 = 15.24$	$s_3 = 9.36$

- Suppose $n = 100$. Then we might assign $(n_1, n_2, n_3) = (34, 34, 32)$ as optimum allocation. Does this make sense?
- Note that all we really need to know is the relative stratum standard deviations (not the actual values) in optimum allocation. In other words, we only need: $\frac{S_h}{\sum_{h=1}^{H} S_h}$.

Allocation in Stratified Random Sampling: Costs

Cost Considerations

Suppose now that there is some cost associated with the selection of each unit within each stratum. Let c_h = cost of sampling a unit in stratum h. Suppose also that there is some fixed cost c_0 associated with the survey regardless of how many units are sampled. The total cost is C.

Allocation in Stratified Random Sampling: Costs

Under a linear cost function

The goal then is to find $n_1, \ldots, n_H$ subject to the constraint that the total cost is c. Via constrained optimization, the resulting optimum allocation is given by:

$$n = \frac{(C - c_0) \sum_{h=1}^{H} N_h s_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h s_h \sqrt{c_h}}$$

$$n_h = \frac{N_h s_h / \sqrt{c_h}}{\sum_{l=1}^{H} N_l s_l / \sqrt{c_l}} \times n$$

- Note that the higher the cost of sampling c_h in stratum h, the smaller the stratum sample size n_h will be. Again, does this makes sense?
- Do we really need to know any more than the relative costs of sampling in the strata here?

Estimating Total Sample Size in Stratified Random Sampling

The next part of this handout gives formulas for the total sample size required to estimate the population mean $\bar{y}_u$ to within some value d with $100(1-\alpha)\%$ probability with stratified random sampling. If the goal is to estimate the population total t to within d with $100(1-\alpha)\%$ probability, this is equivalent to estimating $\bar{y}_u$ to within d/N . In the formulas given below then, replace d by d/N if d is the allowable difference for the total.

Estimating Total Sample Size in Stratified Random Sampling

The total sample size n depends on the allocation of the sample to the strata. Let w_h be the proportion of the sample which will be allocated to stratum h (the w_h's will sum to 1) so that $n_h = nw_h$. Also, let z be the upper $\alpha/2$ critical point of the standard normal distribution. Then, we want to find n such that:

$$z[Var(\bar{y}_{str})]^{1/2} = d \text{ (margin of error)}$$

where

$$V(\bar{y}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} = \frac{1}{n} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{nw_h}{N_h}\right) \frac{S_h^2}{w_h}$$

Solving this margin of error equation for n leads to:

$$\frac{\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{w_h}}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^{H} N_h S_h^2} = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{w_h}}{\frac{d^2}{z^2} + \frac{1}{N} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h^2}$$

The rightmost expression is useful if you don?t know N but do know the values of N_h/N, the relative stratum sizes. If N is large relative to the sample sizes, we could ignore the second term in the denominator and the formula reduces to:

$$n = \frac{z^2}{d^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{w_h}.$$

This is exactly the formula you would get if you ignored the finite population correction factor (fpc) for each stratum in the formula for the variance of $Var(\bar{y}_{str})$.

Total sample size needed with equal allocation

$$n_h = \frac{n}{H} \text{ and } w_h = \frac{1}{H}$$

$$n = \frac{H\sum_{h=1}^{H} N_h^2 S_h^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^{H} N_h S_h^2} = \frac{H\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h^2}$$

$$n = \frac{Hz^2}{d^2}\left(\frac{N_h}{N}\right)^2 S_h^2 \text{ (No fpc correction).}$$

Total Sample Size

Total sample size needed with proportional allocation

$$n_h = \frac{nN_h}{N} \text{ and } w_h = \frac{N_h}{N}$$

$$n = \frac{N\sum_{h=1}^{H} N_h S_h^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^{H} N_h S_h^2} = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h^2}$$

$$n = \frac{z^2}{d^2}\left(\frac{N_h}{N}\right) S_h^2 \text{ (No fpc correction).}$$

Total sample size needed with Neyman allocation (equal costs)

$$n_h = \frac{nN_hS_h}{\sum_{h=1} HN_hS_h} \text{ and } w_h = \frac{N_hS_h}{\sum_{h=1}^{H} N_hS_h}$$

$$n = \frac{\left(\sum_{h=1}^{H} N_hS_h\right)^2}{\frac{N^2d^2}{z^2} + \sum_{h=1}^{H} N_hS_h^2} = \frac{\left(\sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h\right)^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h^2}$$

$$n = \frac{z^2}{d^2}\left(\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_h\right)^2 \text{ (No fpc correction).}$$

Neyman allocation is equivalent to proportional allocation when the stratum variances (S_h's) are the same.

Total sample size needed with Optimal allocation

In this case, we calculate the total cost c required to achieve the desired level of accuracy, since it is the total cost of the survey which is constrained. Let $c^* = C?c_0$ be the cost of the survey less the fixed cost c_0.

$$n_h = c^* w_h \text{ and } w_h = \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}$$

$$c^* = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right) S_h^2} = \frac{\left(\sum_{h=1}^{H} \left(\frac{N_h}{N}\right) S_h \sqrt{c_h}\right)^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right) S_h^2}$$

$$c^* = \frac{z^2}{d}\left(\sum_{h=1}^{H}\left(\frac{N_h}{N}\right) S_h \sqrt{c_h}\right)^2 \quad \text{(No fpc correction)}$$

Regression and Ratio Estimation

Review of Ratio Estimator and its Properties

Estimators

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x}$$

$$\hat{t}_{yr} = \hat{B} t_x$$

$$\hat{\bar{y}} = \bar{x}_U \times \hat{B}$$

Bias

$$Bias(\hat{\bar{y}}) = -Cov(\hat{B}, \bar{x})$$

$$Bias(\hat{\bar{y}}) \approx \frac{1}{n\bar{x}_U}\left(1 - \frac{n}{N}\right)[BS_x - RS_xS_y]$$

Review of Ratio Estimator and its Properties

The bias of $\hat{\bar{y}}_r$ is small if:

- The sample size n is larger.
- The sampling fraction n/N is large.
- $\bar{X}_U$ is large.
- S_x is small.
- The correlation R is close to 1.
- Note that if all the x's are the same value (and thus $S_x = 0$) then the ratio estimator is the same as the SRS estimator and bias is zero.

Review of Ratio Estimator and its Properties

MSE

$$MSE(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right) \frac{S_y^2 - 2BRS_xS_y + B^2S_x^2}{n}$$

Review of Ratio Estimator and its Properties

The approximated MSE of $\hat{\bar{y}}_r$ will be small when

- The sample size n is larger.
- The sampling fraction n/N is large.
- The deviations $y_i - Bx_i$ are small.
- The correlation R is close to $+1$.

Review of Ratio Estimator and its Properties

If we define $e_i = Y_i - \hat{B}x_i$

Review of Ratio Estimator and its Properties

If we define $e_i = Y_i - \hat{B}x_i$ **(What does this remind us of?)**

Review of Ratio Estimator and its Properties

If we define $e_i = Y_i - \hat{B}x_i$ then we can define $s_e^2 = \frac{1}{n-1}\sum_{i=1}^{N}(Y_i - \hat{B}x_i)^2$ then

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right)\left(\frac{\bar{x}_U}{\bar{x}}\right)^2 \frac{s_e^2}{n}$$

Review of Ratio Estimator and its Properties

Similarly

$$\hat{V}(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}^2}$$

and

$$\hat{V}(\hat{t}_{yr}) = \hat{V}(t_x \hat{B}) = \left(1 - \frac{n}{N}\right) \left(\frac{t_x}{\bar{x}}\right)^2 \frac{s_e^2}{n}$$

Review of Ratio Estimator and its Properties

95% CI

$$\hat{B} \pm 1.96 SE(\hat{B})$$
$$\hat{t}_{yr} \pm 1.96 SE(\hat{t}_{yr})$$
$$\hat{\bar{y}}_r \pm 1.96 SE(\hat{\bar{y}}_r)$$

Example

```
library(lohrData)
data(agsrs)
Bhat <- mean(agsrs[, "ACRES92"])/mean(agsrs[,
    "ACRES87"])
Bhat
```

```
 R >  [1] 0.9865652
```

```
yr <- Bhat * (313343.3)
yr
```

```
 R >  [1] 309133.6
```

```
tyr <- Bhat * (964470625)
tyr
```

```
 R >  [1] 951513191
```

Example

```
e <- agsrs[, "ACRES92"] - Bhat * agsrs[,
    "ACRES87"]
se2 <- var(e)
v.bhat <- (1 - 300/3078) * se2/(300 * mean(agsrs[,
    "ACRES87"])^2)
se.bhat <- sqrt(v.bhat)
se.bhat
```

```
 R >  [1] 0.005750473
```

```
### 95 CI
Bhat + qnorm(0.05/2, lower.tail = TRUE) *
    se.bhat
```

```
 R >  [1] 0.9752945
```

```
Bhat + qnorm(0.05/2, lower.tail = FALSE) *
    se.bhat
```

```
 R >  [1] 0.997836
```

Example

```
var.yr <- (1 - 300/3078) * ((313343.3/mean(agsrs[,
    "ACRES87"]))^2) * se2/300
se.yr <- sqrt(var.yr)
se.yr
```

```
 R >  [1] 1801.872
```

```
### 95 CI
yr + qnorm(0.05/2, lower.tail = TRUE) * se.yr
```

```
 R >  [1] 305602
```

```
yr + qnorm(0.05/2, lower.tail = FALSE) *
    se.yr
```

```
 R >  [1] 312665.2
```

Example

```
var.tyr <- (1 - 300/3078) * ((964470625/mean(agsrs[,
    "ACRES87"]))^2) * se2/300
se.tyr <- sqrt(var.tyr)
se.tyr
```

```
 R >  [1] 5546162
```

```
### 95 CI
tyr + qnorm(0.05/2, lower.tail = TRUE) *
    se.tyr
```

```
 R >  [1] 940642913
```

```
tyr + qnorm(0.05/2, lower.tail = FALSE) *
    se.tyr
```

```
 R >  [1] 962383469
```

Example 2

```
library(combinat)
data <- data.frame(x = c(4, 5, 5, 6, 8, 7,
    7, 5), y = c(1, 2, 4, 4, 7, 7, 7, 8))
sampDist4 <- combn(1:8, 4)

x.mean <- apply(sampDist4, 2, function(x) {
    mean(data$x[x])
})
y.mean <- apply(sampDist4, 2, function(x) {
    mean(data$y[x])
})
Bhat <- y.mean/x.mean
t.srs <- 8 * y.mean
t.yr <- Bhat * sum(data$x)
sampDistEst <- data.frame(SampNum = 1:70,
    Sample = apply(sampDist4, 2, paste, collapse = ","),
    x.mean, y.mean, Bhat, t.srs, t.yr, stringsAsFactors = FALSE)
```

Example 2

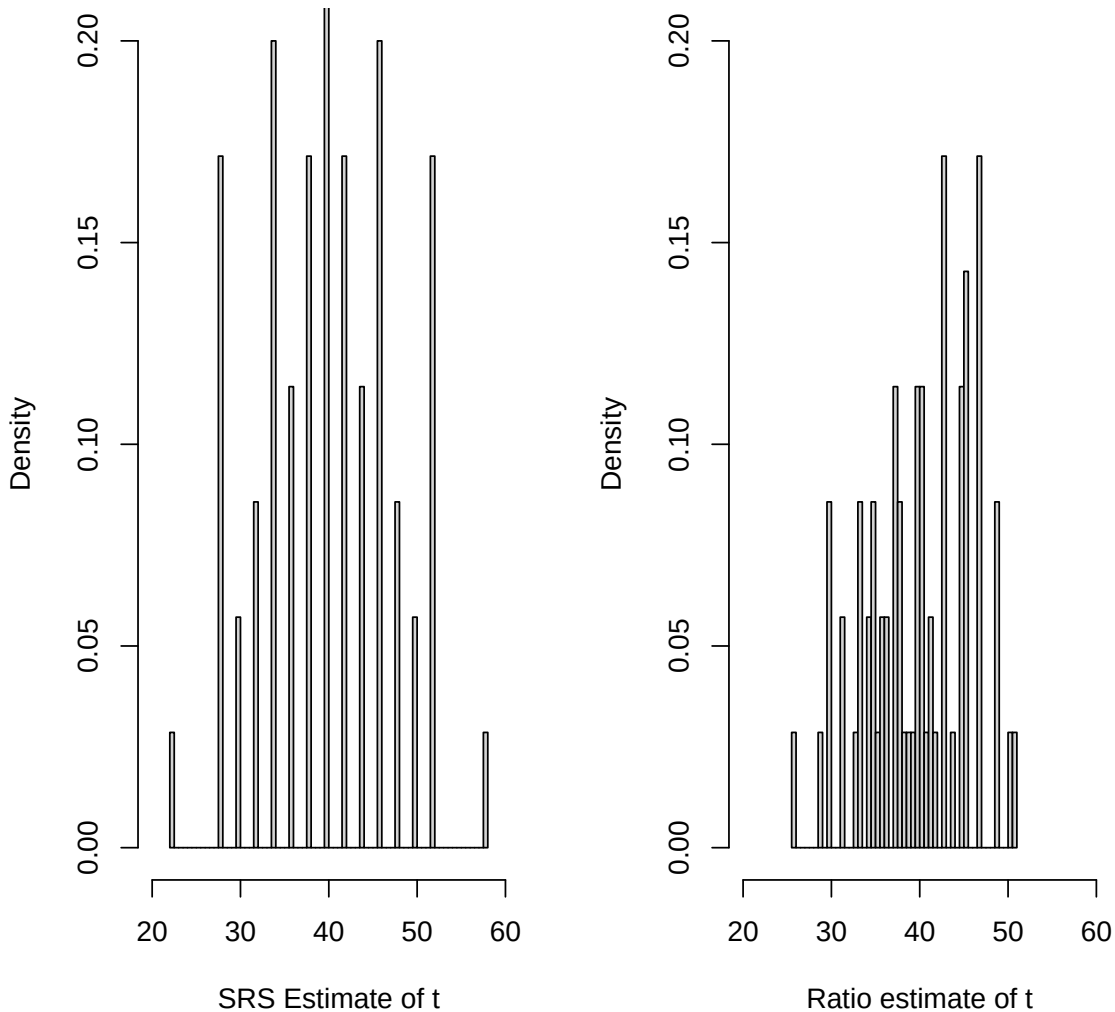
```
head(sampDistEst)
```

```
 R >    SampNum  Sample x.mean y.mean
 R >  1       1 1,2,3,4   5.00   2.75
 R >  2       2 1,2,3,5   5.50   3.50
 R >  3       3 1,2,3,6   5.25   3.50
 R >  4       4 1,2,3,7   5.25   3.50
 R >  5       5 1,2,3,8   4.75   3.75
 R >  6       6 1,2,4,5   5.75   3.50
 R >         Bhat t.srs     t.yr
 R >  1 0.5500000    22 25.85000
 R >  2 0.6363636    28 29.90909
 R >  3 0.6666667    28 31.33333
 R >  4 0.6666667    28 31.33333
 R >  5 0.7894737    30 37.10526
 R >  6 0.6086957    28 28.60870
```

Example 2

```
par(mfrow = c(1, 2))
hist(sampDistEst$t.srs, main = "", xlab = "SRS Estimate of t",
    probability = TRUE, xlim = c(20, 60),
    ylim = c(0, 0.2), breaks = 75)
hist(sampDistEst$t.yr, main = "", xlab = "Ratio estimate of t",
    probability = TRUE, xlim = c(20, 60),
    ylim = c(0, 0.2), breaks = 75)
```

Example 2


Density
0.00
0.05
0.10
0.15
0.20
20
30
40
50
60
SRS Estimate of t
Density
0.00
0.05
0.10
0.15
0.20
20
30
40
50
60
Ratio estimate of t

Ratio Estimation with Proportions

Ratio Estimation with Proportions

Ratio estimation works the exactly same way when the quantity of interest is a proportion!

Ratio Estimation with Proportions

Example: Feral pigs and Vegetation

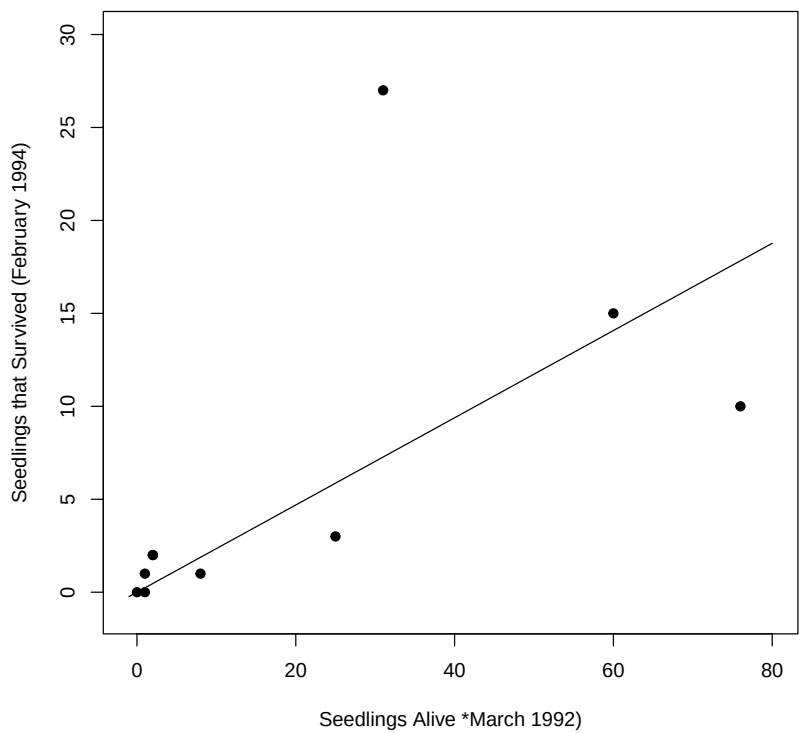
- Peart (1994) collected the data shown (next slide) as part of a study evaluating the effects of feral pig activity and drought on the native vegetation on Santa Cruz Island, CA.
- She counted the number of woody seedlings in pig-protected areas under each of ten sampled oak trees in March 1992, following the drought-ending rains of 1991.
- She put a flag by each seedling, then determined how many were still alive in February 1994.
- One naively calculates the error like it was an SRS when wants to find the sample proportion of the 1992 seedlings that are still alive in 1994.
- This results in SD $\sqrt{(0.2961)(0.7039)/206} = 0.0318$.
- This calculation is incorrect for these data since plots, not individual seedlings, are the sampling units.
- Technically this design is **cluster sample**.

Ratio Estimation with Proportions

```
scIsland <- data.frame(Tree = 1:10, x = c(1,
    0, 8, 2, 76, 60, 25, 2, 1, 31), y = c(0,
    0, 1, 2, 10, 15, 3, 2, 1, 27))

mean(scIsland$x)

 R >  [1] 20.6

sd(scIsland$x)

 R >  [1] 27.47201

mean(scIsland$y)

 R >  [1] 6.1

sd(scIsland$y)

 R >  [1] 8.824839
```

Ratio Estimation with Proportions

```
plot(scIsland$x, scIsland$y, pch = 19, xlab = "Seedlings Alive *March 1992)",
    ylab = "Seedlings that Survived (February 1994)",
    xlim = c(-1, 80), ylim = c(-1, 30))
l <- lm <- lm(y ~ x - 1, data = scIsland)
curve(coef(l)[1] * x, from = -1, to = 80,
    add = TRUE)
```



Ratio Estimation with Proportions

```
## Bhat =phat
Bhat <- mean(scIsland$y)/mean(scIsland$x)
se2 <- var(scIsland$y - Bhat * (scIsland$x))
se.bhat.nfpc <- sqrt((1/(10 * mean(scIsland$x)^2)) *
    se2)
se.bhat.nfpc
```

```
 R >  [1] 0.1152622
```

Ratio Estimation Using Weight Adjustments

Ratio Estimation Using Weight Adjustments

Weights

$$w_i = 1/\pi_i$$

$$\hat{t}_y = \sum_{i \in S} w_i y_i$$

$$\hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \hat{t}_y = \frac{t_x}{\hat{t}_x} \sum_{i \in S} w_i y_i$$

Ratio Estimation Using Weight Adjustments

Weights

We can think of the modification used in ratio estimation as an adjustment to each weight. Define:

$$g_i = \frac{t_x}{\hat{t}_x}$$

$$\hat{t}_{yr} = \sum_{i \in S} w_i g_i y_i$$

Weights

The estimator $\hat{t}_{yr}$ is a weighted sum of observations, with weight $w_i^* = w_i g_i$. Unlike the original weights w_i, however, the adjusted weights w_i^* depend upon values from the sample: If a different sample is taken, the weight adjustment $g_i = t_x / \hat{t}_x$ will be different.

Weights

The weight adjustment g_i **calibrate** the estimates on the x variable. Since

$$\sum_{i \in S} w_i g_i x_i = t_x$$

the adjusted weights force the estimated total for the x variable to equal the known population total t_x. The factors g_i are called the **calibration factors**.

Weights

The variance estimators can be calculated by forming the new variable $u_i = g_i e_i$. Then for an SRS

$$\hat{V}(\bar{u}) = \left(1 - \frac{n}{N}\right) \frac{1}{n(n-1)} \sum_{i \in S} (u_i - \bar{u})^2 = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n} \left(\frac{t_x}{\hat{t}_x}\right)^2 = \hat{V}(\hat{\bar{y}}_r)$$

Similarly $\hat{V}(\hat{t}_u) = \hat{V}(\hat{t}_{yr})$

Ratio Estimation Using Weight Adjustments

```
library(lohrData)
data(agsrs)
g <- 964470625/(mean(agsrs$ACRES87) * 3078)   #t_x/hat_t_x
w <- 3078/300
w.star <- w * g
wgx <- w.star * agsrs$ACRES87
sum(wgx)   #t_x
```

```
 R >  [1] 964470625
```

```
wgy <- w.star * agsrs$ACRES92
sum(wgy)   #hat_t_yr
```

```
 R >  [1] 951513191
```

```
## Note that sum wg \neq N=3078
300 * w.star
```

```
 R >  [1] 3194.101
```

Review of Regression Estimation

The Linear Model: One variable

$$y = B_o + B_1 x$$

Let the estimators for $\hat{B}_0$ and $\hat{B}_1$ be the OLS regression coefficients of the slope and intercept.

$$\hat{B}_1 = \frac{\sum_{i \in S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in S}(x_i - \bar{x})^2} = \frac{r s_y}{s_x}$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

where r is the sample correlation coefficient for x and y.

The Linear Model

- Regression estimation (similar to ratio estimation) uses the correlation between x and y to obtain an estimator for $\bar{y}_U$ with (hopefully) increased precision.
- Suppose we know $\bar{x}_U$ the population mean for the x's. Then the regression estimator of $\bar{y}_U$ is the predicted value of y from the fitted regression equation when $x = \bar{x}_U$.

$$\hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_U = \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x})$$

  - If $\bar{x}$ from the sample is smaller than the population mean $\bar{x}_U$ and x and y are positively correlated, then we would expect $\bar{y}$ to also be smaller then $\bar{y}_U$.
  - The regression estimator adjusts $\bar{y}$ by the quantity $\hat{B}_1(\bar{x}_U - \bar{x})$.

The Linear Model

- Like the ratio estimator, the regression estimator is biased!
- Let B_1 be the least squares regression slope calculated from all the data in the population.

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N}(x_i - \bar{x}_U)^2} = \frac{RS_y}{S_x}$$

The Linear Model

- Bias

The Linear Model

- MSE

The Linear Model

- The SE can be calculated by substituting estimates for the population quantities. We can estimate S_d^2 by using the residuals $e_i = y_i - (\hat{B}_0 + \hat{B}_1 x_i)$. $s_e^2 = \sum_{i \in S} e_i^2/(n-1)$ estimates S_d^2 and

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$

The Linear Model

- In small samples, we may alternatively calculate s_e^2 using the MSE from a regression analysis

$$s_e^2 = \sum_{i \in S} e_i^2/(n-2)$$

.

- Where does this come from?
- To estimate the variance for this formulation

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n}s_y^2(1-r^2)}$$

Example: Trees

- To estimate the number of dead trees in an area, we divide the area into 100 square plots and count the number of dead trees on a photograph of each plot.
- Photo counts can be made quickly, but sometimes a tree is misclassified or not detected.
- So we select an SRS of 25 of the plots for field counts of dead trees (quality control)
- We know that the population mean number of dead trees per plot from the photo count is 11.3.

Example: Trees

```
photo.x <- c(10, 12, 7, 13, 13, 6, 17, 16,
    15, 10, 14, 12, 10, 5, 12, 10, 10, 9,
    6, 11, 7, 9, 11, 10, 10)
photo.y <- c(15, 14, 9, 14, 8, 5, 18, 15,
    13, 15, 11, 15, 12, 8, 13, 9, 11, 12,
    9, 12, 13, 11, 10, 9, 8)
photo <- data.frame(x = photo.x, y = photo.y)
```

Example: Trees

```
desc <- matrix(c(mean(photo$x), mean(photo$y),
    sd(photo$x), sd(photo$y)), byrow = FALSE,
    nc = 2, dimnames = list(c("x", "y"),
        c("Mean", "SD")))
desc
```

```
 R >      Mean       SD
 R >  x 10.60 3.068659
 R >  y 11.56 3.014963
```

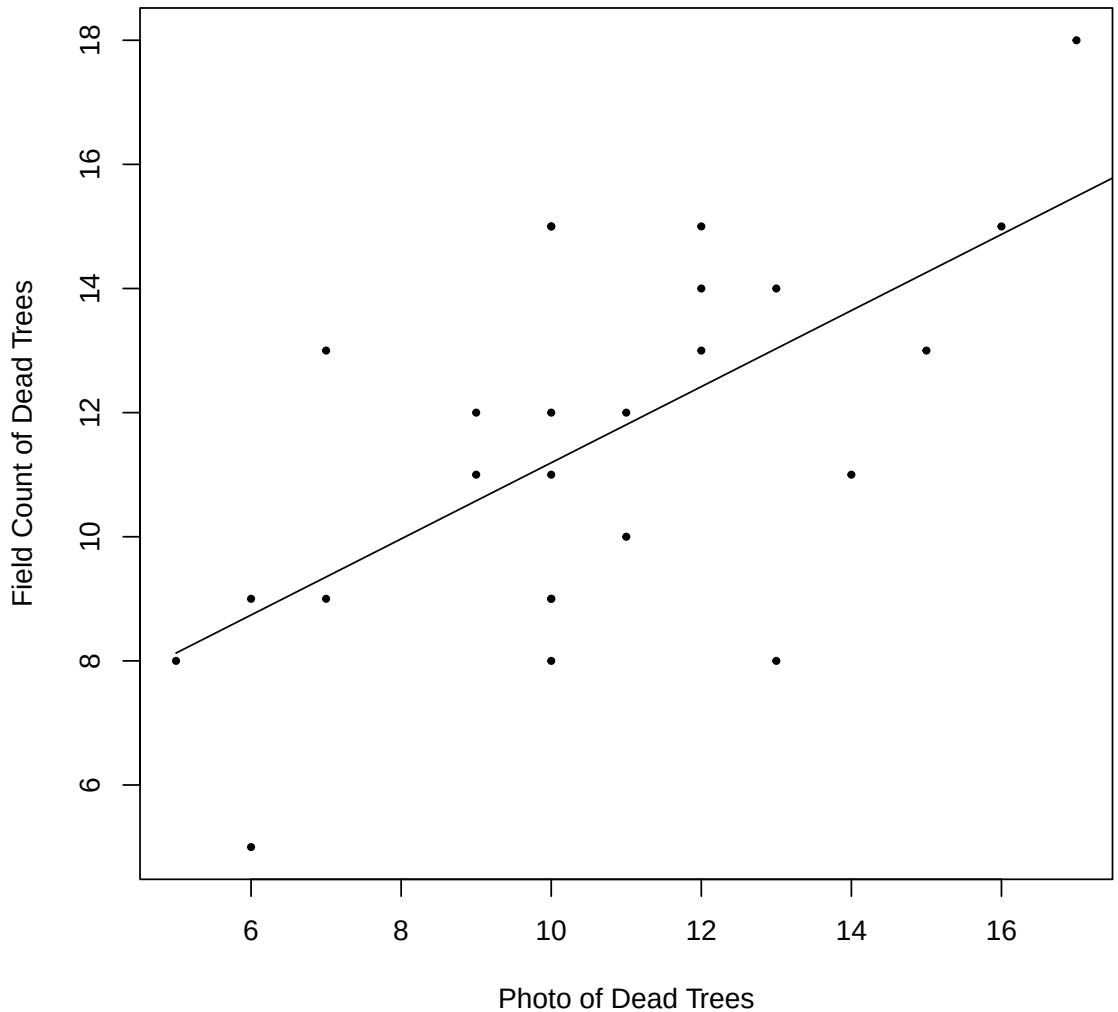
```
cor(photo$x, photo$y)
```

```
 R >  [1] 0.6241967
```

Example: Trees

```
plot(photo$x, photo$y, pch = 19, cex = 0.5,
    xlab = "Photo of Dead Trees", ylab = "Field Count of Dead Trees")
l <- lm(y ~ x, data = photo)
curve(coef(l)[1] + coef(l)[2] * x, from = min(photo),
    to = max(photo), add = TRUE)
```

Example: Trees



Example: Trees

$$\hat{y} = 5.059292 + 0.613274x$$

- $\hat{B}_0 = 5.059292$
- $\hat{B}_1 = 0.613274$
- x and y are positively correlated so that $\bar{x}$ and $\bar{y}$ are also positively correlated.
- Since $\bar{x} < \bar{x}_U$, we expect

$$\hat{B}_1(\bar{x}_U - \bar{x}) = 0.613274(11.3 - 10.6) = 0.43$$

to $\bar{y}$ to compensate.

Example: Trees

$$\hat{y} = 5.059292 + 0.613274x$$

- The regression estimate of the mean is

$$\hat{y} = 5.059292 + 0.61327411.3 = 11.99$$

Example: Trees

```
se.yreg <- sqrt((1 - length(photo$y)/100) *
    (var(photo$y)/25) * (1 - cor(photo$y,
    photo$x)^2))
se.yreg
```

```
 R >  [1] 0.4079831
```

Compare to

```
se.ybar <- sqrt((1 - length(photo$y)/100) *
    (var(photo$y)/25))
se.ybar
```

```
 R >  [1] 0.5222069
```

Example: Trees

We expect the regression estimator to increase the precision in this example b/c the variables photo and field are positively correlated. To estimate the total number of dead trees, use

```
t_reg <- 100 * (coef(l)[1] + coef(l)[2] *
    11.3)
t_reg
```

```
 R >  (Intercept)
 R >     1198.929
```

```
se.treg <- 100 * se.yreg
se.treg
```

```
 R >  [1] 40.79831
```

Difference Estimation

Difference Estimation

Difference Estimation is a special case of regression estimation, used when the investigator "knows" that the slope $B_1 = 1$.

Difference Estimation

- Difference estimation is often recommended in accounting when an SRS is taken.
- A list of accounts receivable consists of the book value for each account – the company's listing of how much is owed on each account.
- In the simplest sampling scheme, the auditor scrutinizes a random sample of the accounts to determine the audited value (actual amount owed).
- The quantities considered

$$y_i = \text{ audited value for company } i$$
$$x_i = \text{ book value for company } i$$

Then, $\bar{y} - \bar{x}$ is the mean difference for the audited accounts.

Difference Estimation

- The estimated total difference is

$$\hat{t}_y - \hat{t}_x = N(\bar{y} - \bar{x})$$

- The estimated audited value for accounts receivable is thus

$$\hat{t}_{ydiff} = t_x - (\hat{t}_y - \hat{t}_x)$$

- The residuals from this model are $e_i = y_i - x_i$. The variance of $\hat{t}_{ydiff}$ is

$$V(\hat{t}_{ydiff}) = V[t_x - (\hat{t}_y - \hat{t}_x] = V(\hat{t}_e)$$

where $\hat{t}_e = (N/n) \sum_{i \in S} e_i$.

Difference Estimation

- Difference estimation works best if the population and sample have a larger fraction of nonzero differences that are roughly equally divided between overstatements and understatements.
- If the sample is large enough so that the sampling distribution of $(\bar{y} - \bar{x})$ is approximately normal.
- Often in the real case of auditing one would want to use a more stable model based estimate.

Poststratification

Poststratification

Motivation

- Earlier in the course we showed the increase in precision that can come from using population data to stratify.
- Stratification is not always a desirable way to use these population data!
- There may be too many potential stratification variables.
- The need for cluster sampling may prevent stratification on individual-level variables.
- Population data may also be available in a form that does not allow for stratification (e.g., random digit dialing).
- Poststratification is a technique for using known population totals for a set of variables to adjust the sampling weights and improve estimation for another set of variables.

Poststratification

Motivation

- All of the techniques we will discuss have the same idea: adjustments are made to the sampling weights so that estimated population totals for the auxiliary variables match the known population totals, making the sample more representative of the population.
- A second benefit is that the estimates are forced to be consistent with the population data.
- This often improves their credibility with people who may not understand the sampling process.

Poststratification

Motivation

- Two core applications of these techniques:
  - Increase precision of estimation.
  - Reduce the bias from nonresponse (especially unit non-response), where a sampled individual refuses to participate or otherwise provides no information for analysis.

Poststratification

Motivation

- Two core applications of these techniques:
  - Increase precision of estimation.
  - Reduce the bias from nonresponse (especially unit non-response), where a sampled individual refuses to participate or otherwise provides no information for analysis.

Generally, the use for non-response is considered the more important of the two in large-scale surveys.

However, this stage is often completed before being handed off to the typical user (e.g., GSS).

Poststratification: Example

Poststratification by Sex

The post-stratified estimate of the proportion of Minnesota residents over the age of 18 who are in favor of public funding of the Viking Stadium, $\hat{p}_{post}$ (where the subscript post denotes ?post-stratified?), is:

$$\begin{aligned}\hat{p}_{post} &= \frac{N_f}{N}\hat{p}_f + \frac{N_m}{N}\hat{p}_m \\ &= \frac{64,398}{123,516}23.5 + \frac{59,118}{123,516}31.1 \\ &= 27.16\end{aligned}$$

Poststratification: Example

Review Stratified Notation

$$y_{hj} = \text{ the value of the } j\text{th unit in stratum } h$$

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{ the population total in stratum } h$$

$$t = \sum_{h=1}^{H} t_h = \text{ the population total}$$

Poststratification: Example

Mean

Remember:

$$\bar{y}_h = \frac{\sum_{j \in S_h} y_{hj}}{n_h}$$

and the post stratified estimate of the overall sample mean will be

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Relationship to Ratio Estimation

- Define $x_{ih} = 1$ if the observation i is in post stratum h and 0 otherwise.
- Let $u_{ih} = y_i \cdot x_{ih}$.
- Then $t_{xh} = \sum_{i=1}^{N} x_{ih} = N_h$
- $t_{uh} = \sum_{i=1}^{N} u_{ih} =$ population total of variable y in poststratum h.
- $\hat{t}_{uh} = \sum_{i \in S} \frac{N}{n} u_{ih}$
- We can then use ratio estimation to obtain:
- $\hat{t}_{uhr} = \frac{t_{xh}}{\hat{t}_{xh}} \hat{t}_{uh} = \frac{N_h}{\hat{N}_h} \hat{t}_{uh} = N_h \bar{y}_h$
- $\hat{t}_{ypost} = \sum_{h=1}^{H} \hat{t}_{uhr} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{t}_{uh} = \sum_{h=1}^{H} N_h \bar{y}_h$
- $\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$
- $\hat{V}(\bar{y}_{post}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{s_h^2}{n}$

Poststratification: R Code Example

```r
set.seed(34567)
pop1 <- rpois(1000, 5)
pop2 <- rpois(5000, 10)
pop <- c(pop1, pop2)
s <- sample(1:length(pop), 300)
y <- pop[s]
#### Post stratification estimator
x1 <- as.numeric(s <= 1000)
x2 <- as.numeric(s > 1000)
```

Poststratification: R Code Example

```
table(x1)

 R >  x1
 R >    0   1
 R >  251  49

table(x2)

 R >  x2
 R >    0   1
 R >   49 251

## Check
sum(x1 + x2) == 300

 R >  [1] TRUE
```

Poststratification: R Code Example

```
### Build u
u1 <- y * x1
u2 <- y * x2

N1 <- 1000
N2 <- 5000

tu1_hat <- (6000/300) * sum(u1)
tu1_hat

 R >  [1] 4760

tu2_hat <- (6000/300) * sum(u2)
tu2_hat

 R >  [1] 51480
```

Poststratification: R Code Example

```
N1_hat <- 6000 * mean(x1)
N1_hat

 R >  [1] 980

N2_hat <- 6000 * mean(x2)
N2_hat

 R >  [1] 5020

t_ypost <- (1000/N1_hat) * tu1_hat + (5000/N2_hat) *
    tu2_hat
t_ypost

 R >  [1] 56132.04

t_srs <- 6000 * mean(y)
t_srs

 R >  [1] 56240
```

Poststratification: R Code Example

```
ybar.post <- (1000/6000) * mean(y[x1 == 1]) +
    (5000/6000) * mean(y[x2 == 1])
ybar.post
```

```
 R >  [1] 9.355341
```

```
t_ypost/6000
```

```
 R >  [1] 9.355341
```

```
y.srs <- mean(y)
y.srs
```

```
 R >  [1] 9.373333
```

Poststratification: R Code Example

```
s1 <- sd(y[x1 == 1])
s2 <- sd(y[x2 == 1])

v.ybar.post <- (1 - 300/6000) * ((1000/6000) *
    s1^2/300 + (5000/6000) * s2^2/300)
v.ybar.post
```

```
 R >  [1] 0.03291677
```