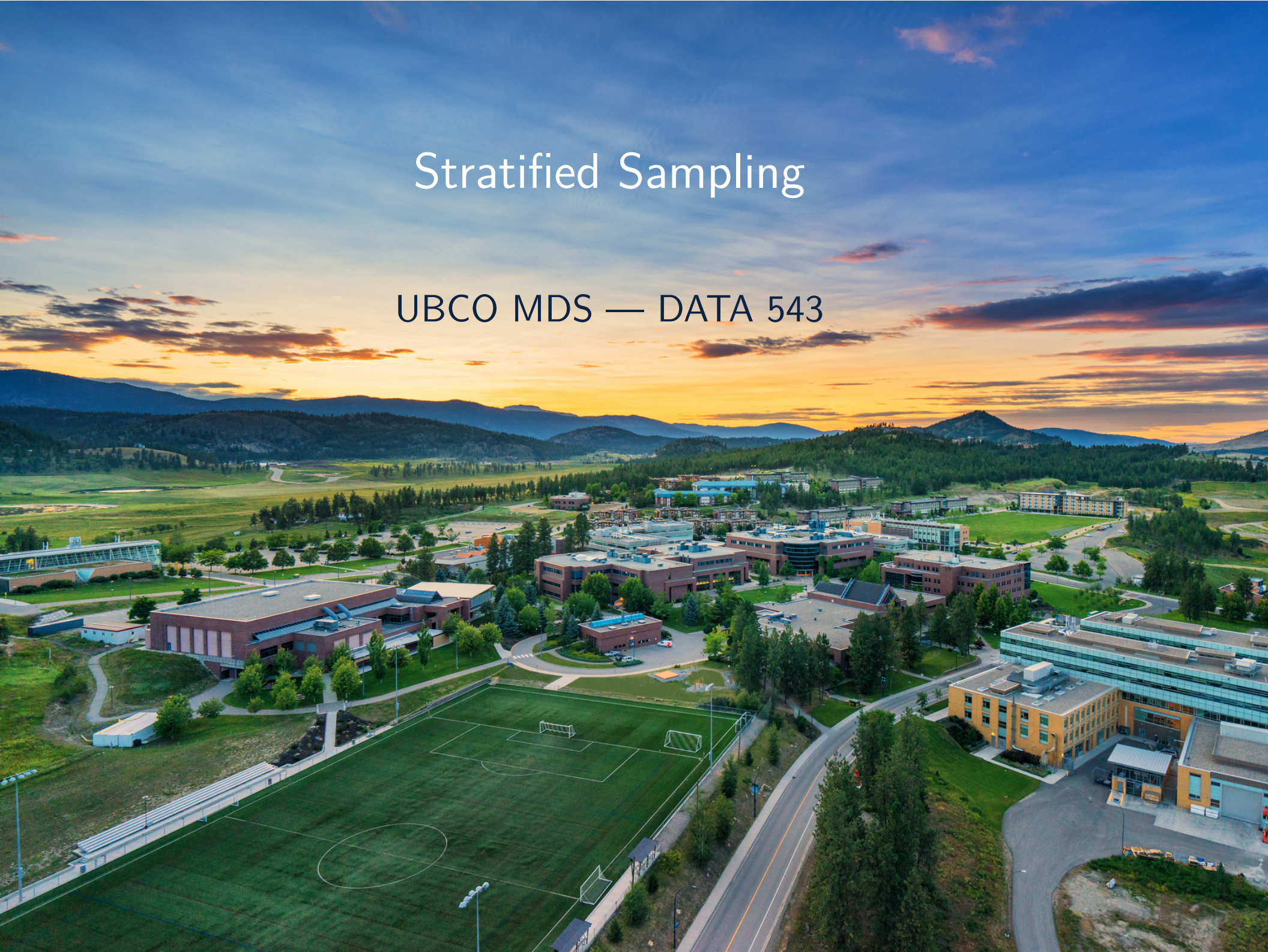
Stratified Sampling

UBCO MDS — DATA 543



Office Hours today (I am around all day) until 17^{00}.

Key concepts for today:

- SRSWOR vs. Stratified sampling
- Stratum vs. sampling weight
- Stratum average and variance (CI)
- Stratum total and proportion (CI)
- Proportional allocation
- Optimal allocation

Stratified Sampling Summary

- The population is divided into H strata.
- Each strata has μ_h and σ_h
- The population stratum h weight is $W_h = \frac{N_h}{N}$
- The number of units selected from strata h is n_h
- Point estimate and estimator variance

$$\hat{\mu}_{\text{str}} = \sum_{h=1}^{H} W_h \hat{\mu}_h, \quad \widehat{Var}(\tilde{\mu}_{\text{str}}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{\sigma}_h^2}{n_h}$$

- Proportional Allocation $n_h \propto W_h$
- Optimal Allocation $n_h \propto W_h \sigma_h$
- Construct strata with small σ_h and large differences among the strata averages.

Stratified sampling is most beneficial when the strata means are different and the variances are different.

Introduction

- Last class we saw estimators and confidence intervals for SRSWOR and SRSWR.
- We saw that SRSWOR is more efficient (less uncertainty/smaller variance) than SRSWR .
- This might have been somewhat intuitive, but how does SRSWOR compare with some of the other probability sampling techniques?
- Today we will explore the following result:

SRSWOR vs. Stratified Sampling

Stratified sampling can be shown to be more efficient that SRSWOR in certain scenarios

Stratified Random Sampling

- Often, we have supplementary information that can help us design our sample.
  - eg, Vancouver, BC residents pay more for housing than Saint John, NB, or that rural residents shop for groceries less frequently than urban residents.
- If the variable we are interested in takes on different mean values in different sub-populations, we may be able to obtain more precise estimates of population quantities by taking a stratified random sample.

Stratified Random Sampling

The basic principle:

By dividing into strata, we group similar units together.

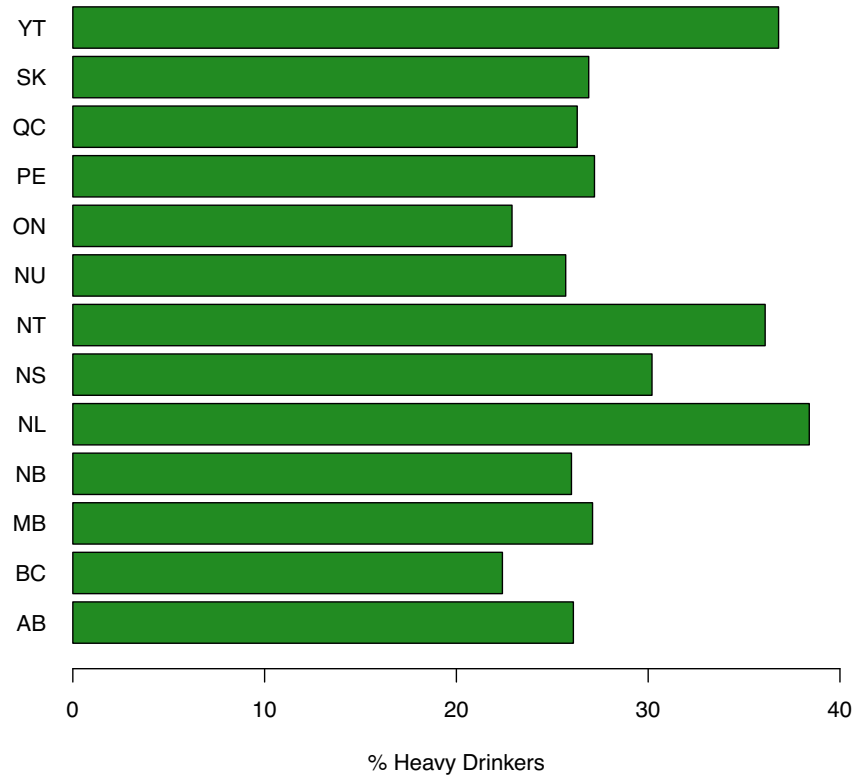
homogeneous units / groups.

This can lead to greater efficiency in our estimates, if we are clever/careful!

- Note that we often have an interest in the strata averages
- So even if the stratified structure does not improve our estimate of interest, we still learn things.
- Often data are arranged in strata 'by default', eg. Provinces of Canada.

Stratified Random Sampling

Alcohol consumption in Canada (data from Stats Canada)



Stratified Examples

Think carefully about your population - strata are often lurking!

- Geographical location: e.g., different provinces, rural vs. urban, etc.
- Human characteristics: e.g., socioeconomic status, field of study, etc.
- Unit characteristics: e.g., large and small fields when studying yield, high schools vs. elementary schools, etc.

Note: it's common for strata to be formed from continuous variates (e.g., income ranges rather than the specific dollar figure).

Stratified Sampling (review)

1. Divide the population in H strata $U_1, U_2, \ldots, U_H$, with strata U_h having N_h units.
   - The strata should be mutually exclusive: the strata do not overlap

$$N_1 + N_2 + \ldots + N_H = N$$

   - The strata should also be collectively exhaustive: no population element can be excluded

$$U_1 \cup U_2 \cup \cdots \cup U_H = U$$

2. Draw an independent probability sample from each strata

$$n_1 + n_2 + \ldots + n_H = n$$

Stratified Random Sampling: Notation

Our variate of interest will be y. For each stratum we write:

y_{hj}	the j^{th} element in the h^{th} stratum
$\mu_h = \frac{1}{N_h}\sum_{j=1}^{N_h} y_{hj}$	the population mean for stratum h
$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)^2$	the population variance for stratum h
$\tau_h = \sum_{j=1}^{N_h} y_{hj}$	the population total for stratum h

Stratified Random Sampling: Notation

add up all the strata.

For the overall population we write:

strata total

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \frac{\tau}{N}$$ the (overall) population mean

$$\sigma^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\mu)^2$$ the (overall) population variance.

$$\tau = \sum_{h=1}^{H}\tau_h$$ the (overall) population total

Stratified Random Sampling: Notation

Treating each stratum as its own population, suppose we only have a sample (rather than a census) from each. Corresponding quantities for the sample using SRSWOR estimators with each stratum are:

$$\tilde{\mu}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj} \tag{1}$$

$$\tilde{\tau}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} = N_h \tilde{\mu}_h \tag{2}$$

$$\tilde{\sigma}_h^2 = \sum_{j \in S_h} \frac{(y_{hj} - \tilde{\mu}_h)^2}{n_h - 1} \tag{3}$$

where S_h denotes the sample of n_h units from strata h.

Stratified Random Sampling: Notation

Then we would estimate μ_h by $\tilde{\mu}_h$ and τ_h by $\tilde{\tau}_h$. The stratified estimators for **population total,** τ and **population mean** μ are given by eq (4) and (5), respectively.

$$\tilde{\tau}_{\text{str}} = \sum_{h=1}^{H} \tilde{\tau}_h = \sum_{h=1}^{H} N_h \tilde{\mu}_h = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj} \tag{4}$$

$$\tilde{\mu}_{\text{str}} = \frac{\tilde{\tau}_{\text{str}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \tilde{\mu}_h = \sum_{h=1}^{H} W_h \tilde{\mu}_h \tag{5}$$

We call $W_h = \frac{N_h}{N}$ the *stratum weights*, and we call $w_{hj} = \frac{N_h}{n_h}$, the *sampling weight*.

Weights

- Hence the stratified sampling estimator $\tilde{\mu}_{\text{str}}$ is a *weighted average* of the sample stratum averages, $\tilde{\mu}_h$.
- The so-called *stratum weights*, $W_h = \frac{N_h}{N}$, represent the proportion of the population units in stratum h.
- The stratified sampling estimator $\tilde{\tau}_{\text{str}}$ is expressed as a weighted sum of the individual sampling units.

Weights

- $w_{hj} = N_h/n_h$, the so-called *sampling weight* for unit j of stratum h can be thought of as the number of units in the population represented by the sample member y_{hj}. For example,
  - Suppose a population has 1600 men and 400 women, and we sample 200 men and 200 women each man in the sample has weight 8: each man represents himself and seven other men not in the sample. Each woman has weight 2: each woman in the sample represents herself and another woman not in the sample.
- The inclusion probability of including unit j of stratum h in the sample is $\pi_{hj} = n_h/N_h$, the *sampling fraction* in stratum h.
- Hence the sampling weight is simply the reciprocal of the inclusion probability.

μ as a Weighted Average

Let's look at the population mean a little more closely:

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}$$

$$= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} \qquad \text{multiplying by 1}$$

$$= \sum_{h=1}^{H} \frac{N_h}{N}\mu_h \qquad \text{where } \mu_h \text{ is the mean for stratum } h$$

$$= \sum_{h=1}^{H} W_h \mu_h \qquad \text{where } W_h = \frac{N_h}{N} \text{ is the } \textit{stratum weight}$$

μ as a Weighted Average

- A stratum weight, $W_h = \frac{N_h}{N}$, is the proportion of the population included in a stratum.
- You can think of this 'weighting' as giving the means in larger strata more 'importance' in the calculation of the overall mean.

For ex, if 35.16 million people live in Canada, (N = 35.16)

- strata 9 is Ontario with 13.6 million people then

$$N_9 = 13.6, \quad \text{and} \quad W_9 = \frac{13.6}{35.16} = 0.39$$

- strata 11 is the Northwest Territories with 41,786 people then

$$N_{11} = 0.042, \quad \text{and} \quad W_{11} = \frac{0.042}{35.16} = 0.0012$$

Stratified Random Sampling Protocol

We first consider using SRSWOR within each stratum.

- We draw n_h units from each strata $h = 1, 2, \ldots H$ forming *stratum samples* $s_1, ..., s_H$.
- The complete sample is then $s = s_1 \cup s_2 \cup ... \cup s_H$.
- The sample size is $n = n_1 + n_2 + ... + n_H$.

For example, if we denoted Alberta as the first stratum, and sampled 100 people from Alberta, we would have $n_1 = 100$ and s_1 would denote that subset of Albertans.

Stratum Average and Variance

For each stratum $h = 1, 2, \ldots, H$ we have:

$$\hat{\mu}_h = \overline{y}_h = \frac{1}{n_h} \sum_{j \in s_h} y_{hj} \qquad \text{sample stratum mean}$$

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \hat{\mu}_h)^2 \qquad \text{sample stratum variance}$$

Since we conducted SRSWOR on each stratum to get our sample, the following results fall from the results of last lecture:

$$E[\tilde{\mu}_h] = \mu_h \qquad \text{and} \qquad Var(\tilde{\mu}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

Properties of Stratified Sampling Estimators

Following from the properties of SRSWOR estimators, both $\tilde{\tau}_{\text{str}}$ and $\tilde{\mu}_{\text{str}}$ are unbiased estimators of τ and μ. The estimator for the population mean using stratified sampling given in (5) is unbiased:

$$E[\tilde{\mu}_{str}] = E\left[\sum_{h=1}^{H} W_h \tilde{\mu}_h\right] = \sum_{h=1}^{H} W_h E[\tilde{\mu}_h] = \sum_{h=1}^{H} W_h \mu_h = \mu$$

Nota Bene (Note well)

N.B: we need to know the *population* stratum weights W_h in order to compute this estimate.

Stratified Estimator Variance

We can use the SRSWOR result for the variance of the estimator:

$$
\begin{aligned}
Var(\tilde{\mu}_{\text{str}}) &= Var\left[\sum_{h=1}^{H} W_h \tilde{\mu}_h\right] \\
&= \sum_{h=1}^{H} W_h^2 \, Var(\tilde{\mu}_h) \\
&= \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}
\end{aligned}
$$

Stratified Estimator Standard Error

We can therefore estimate the variance by

$$\widehat{Var}(\tilde{\mu}_{\text{str}}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{\sigma}_h^2}{n_h}$$

where we have the sample stratum h variance:

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \hat{\mu}_h)^2$$

Note: we might also use the notation $f_h = n_h/N_h$ in analogous fashion to earlier examples.

Properties of Stratified Sampling Estimators

The estimated variance for the unbiased estimators are:

$$\hat{V}(\tilde{\tau}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{\hat{\sigma}_h^2}{n_h} \tag{6}$$

$$\hat{V}(\tilde{\mu}_{\text{str}}) = \frac{1}{N^2} \hat{V}(\tilde{\tau}_{\text{str}}) \tag{7}$$

$$= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{\sigma}_h^2}{n_h}$$

As before, the standard error of an estimator is the square root of the estimated variance.

Confidence intervals for stratified samples

If either

$n_h > 30$ ish.

i) the sample size within each stratum are large, or

ii) the sampling design has a large number of strata

an approximate $100(1-\alpha)\%$ CI for the population mean μ is given by:

$$\hat{\mu}_{\text{str}} \pm z_{\alpha/2}\sqrt{\hat{V}(\tilde{\mu}_{\text{str}})} \tag{8}$$

N.B. some survey software packages use $t_{\alpha/2,n-H}$ in place of $z_{\alpha/2}$.

Stratified Sampling for Proportions

- Since a proportion is just a mean of variables that takes on values 0 and 1, we simply use the results from above for inference about proportions.
- Namely for each stratum we have

$$\hat{\pi}_h = \hat{\mu}_h = \frac{1}{n_h} \sum_{j \in s_h} y_{hj} \qquad \hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} \hat{\pi}_h (1 - \hat{\pi}_h)$$

The overall population estimator for the population parameter π using Stratified Sampling is:

$$\hat{\pi}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\pi}_h \tag{9}$$

Stratified Sampling for Proportions

The estimated variance of the estimator from the previous page is given by:

$$\hat{V}(\hat{\pi}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{\pi}_h(1-\hat{\pi}_h)}{n_h - 1} \tag{10}$$

Estimating the total number of population units having a specified characteristic is similar:

$$\hat{\tau}_{\text{str}} = \sum_{h=1}^{H} N_h \hat{\pi}_h$$

Hence the estimated total is the sum of the estimated totals in each stratum. Similarly $\hat{V}(\hat{\tau}_{\text{str}}) = N^2 \hat{V}(\hat{\pi}_{\text{str}})$

Stratified Sampling: Example

- Consider a student communications survey conducted open to all graduates and undergraduates (31,631 students total) at a particular university.
- One question asked whether students used LinkedIn at least once a week. Answer: Yes / No
- The report provides some results broken down by faculty: do these seem like reasonable strata?

Stratified Sampling: Example

Here are the results for the LinkedIn question (so, e.g., 13% of Applied Health Sciences students said they used LinkedIn at least once a week).

Stratum	N_h	W_h	n_h	n_h/n	$\hat{\pi}_h$
AHS	2,434	0.077	149	0.124	0.13
Arts	6,661	0.211	341	0.284	0.26
Engineering	7,998	0.253	202	0.169	0.34
Environment	2,503	0.079	119	0.099	0.32
Math	6,661	0.211	165	0.138	0.29
Science	5,374	0.170	223	0.186	0.18

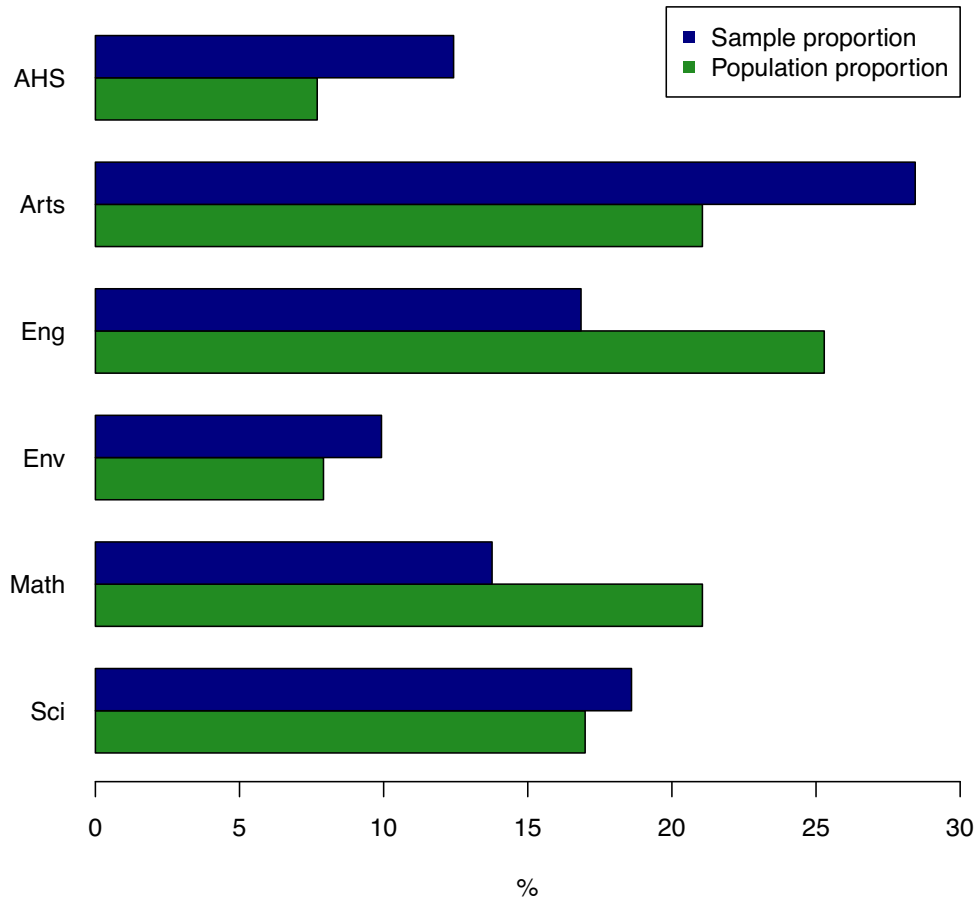
or $\hat{\mu}_h$ and $\hat{\sigma}_h^2$

Stratified sampling is beneficial

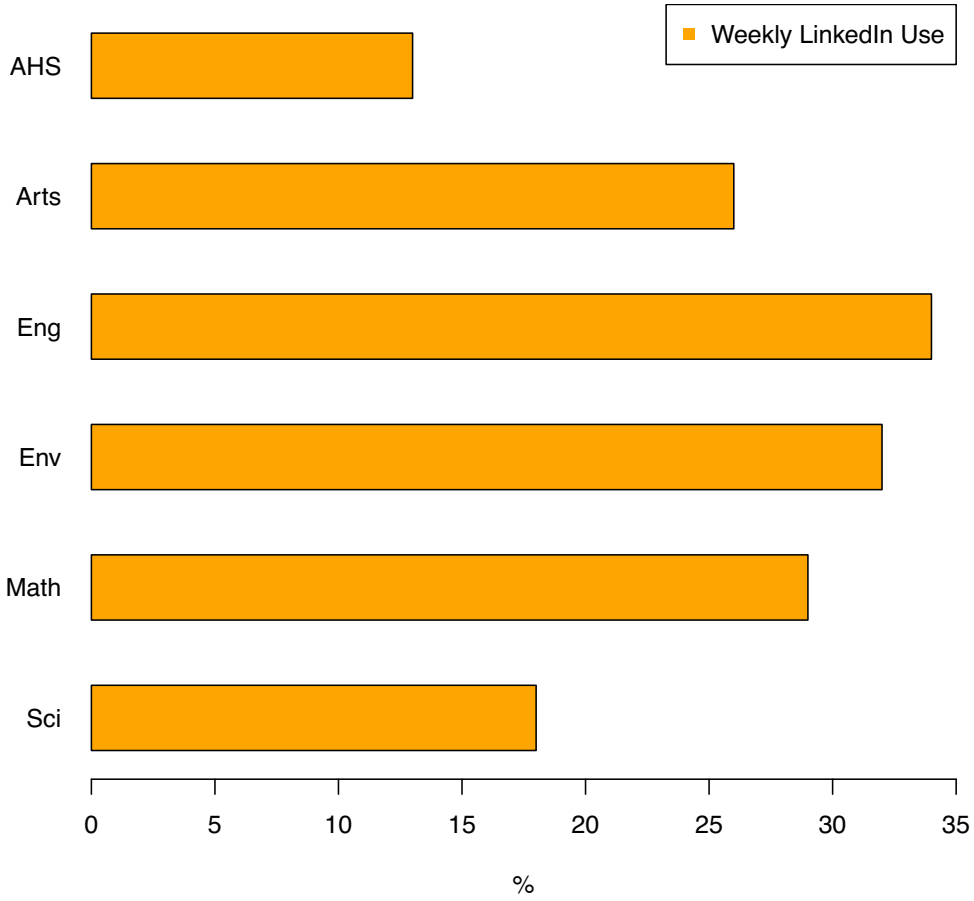
$+$ N $+$ n

We might define $w_h = n_h/n$ to represent the so-called *sample* strata weights.

The sample vs. population strata weights (i.e. $w_h = n_h/n$ vs. $W_h = N_h/N$):



Now let's just look at the survey results:



Stratified Sampling: Example

Some things we might have noticed:

- The sample strata weights don't match up very well with the population strata weights: Engineering and Math seem to be under-sampled while Arts and AHS students seem to be oversampled.

- There's quite a lot of variation in the LinkedIn usage results. In particular, Engineers use it a lot more than AHS students.

Stratum	N_h	W_h	n_h	n_h/n	$\hat{\pi}_h$
AHS	2,434	0.077	149	0.124	0.13
Arts	6,661	0.211	341	0.284	0.26
Engineering	7,998	0.253	202	0.169	0.34
Environment	2,503	0.079	119	0.099	0.32
Math	6,661	0.211	165	0.138	0.29
Science	5,374	0.170	223	0.186	0.18

$W_h \hat{\pi}_h$

Our 'raw' sample average is

Ignoring stratification

$$\hat{\pi} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j\in s_h} y_{hj} = \sum_{h=1}^{H}\frac{n_h}{n}\frac{1}{n_h}\sum_{j\in s_h} y_{hj} = \sum_{h=1}^{H}\frac{n_h}{n}\hat{\pi}_h = 0.253.$$

Our stratified sample estimate from (9):

Takes stratification into account.

$$\hat{\pi}_{\text{str}} = \sum_{h=1}^{H} W_h \hat{\pi}_h = 0.268$$

Stratified Confidence Interval Example

Let's compute a confidence interval for the LinkedIn example!
First let's calculate the estimate of the variance for this estimator given in (10) (rewritten using $W_h = N_h/N$ and $f_h = n_h/N_h$)

Typo: should be $\hat{}$ for consistency.

$$\hat{V}(\hat{\pi}_{\text{str}}) = \sum_{h=1}^{H} (1 - f_h)\,(W_h)^2\,\frac{\hat{\pi}_h(1-\hat{\pi}_h)}{n_h - 1}$$

$$= 0.0001813225$$

confirmed in Lohr (2009):

and

$$\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}. \qquad (3.7)$$

(they use slightly different notation)

I have fixed Lab 2.

The 95% CI is therefore:

$$\hat{\pi}_{\text{str}} \pm z_{\alpha/2}\sqrt{\hat{V}(\tilde{\pi}_{\text{str}})}$$

$$= 0.268 \pm 1.96\sqrt{0.0001813225}$$

$$= [0.2413, 0.2941]$$

"We are 95% confident the interval from 24.13% to 29.41% captures the population percentage of students who use LinkedIn at least once per week."

Lohr (2009) Example 2.5

(Another example)

→ skip to next highlight if running out of time.

The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms[1] in the 50 states. The data contains the 1982, 1987, and 1992 information on:

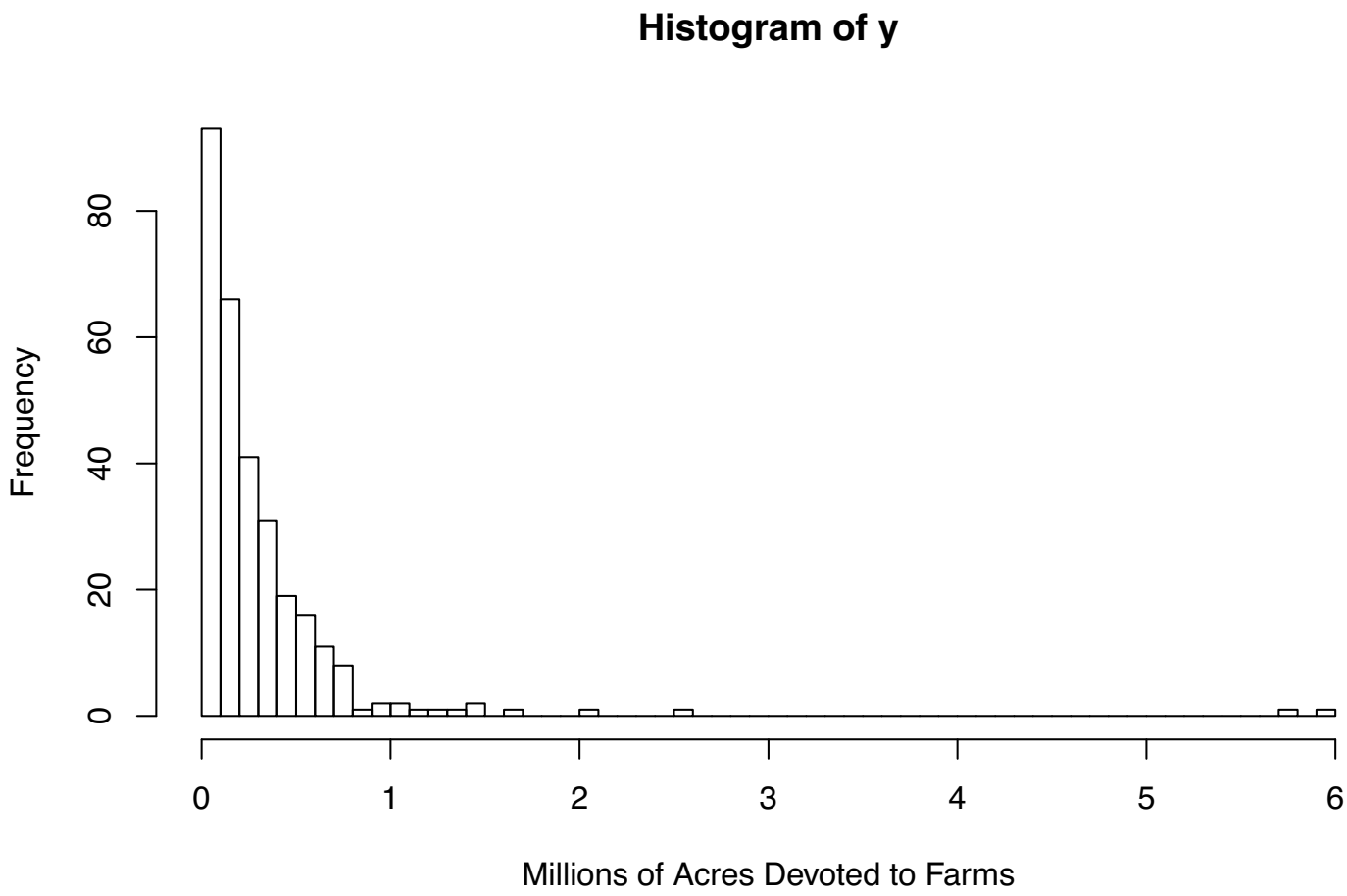
- the number of farms,
- number of farms with fewer/larger than 9/1000 acres, resp.
- the total acreage devoted to farms, farm size, yield of different crops,
- region of country (W = West, NC = North Central, S = South, N = Northeast)
- and a wide variety of other agricultural measures

for the $N = 3078$ counties and county-equivalents in the U.S.

[1] any place from which ≥ 1000 agricultural products were produced and sold

Lohr (2009) Example 2.5

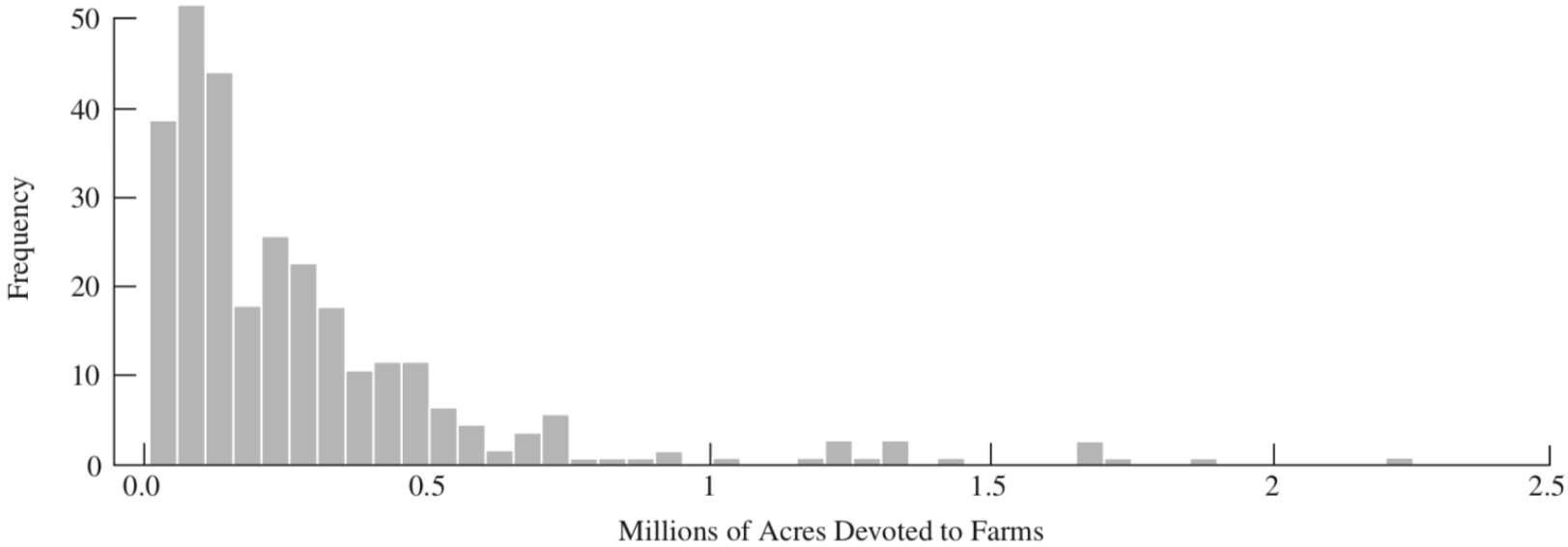
- There is a wide range of values for y_i, the number of acres devoted to farms in county i in 1992, with some states (AK, DE, MA, VT) not being sampled from at all.



Lohr (2009) Example 2.5

- Note that our results aren't identical to Lohr (2009) since their exact sample of 300 was not available for download. Here is the distribution of millions of acres devoted to farms for their SRSWOR:

FIGURE 2.4
Histogram: number of acres devoted to farms in 1992, for an SRS of 300 counties. Note the skewness of the data. Most of the counties have fewer than 500,000 acres in farms; some counties, however, have more than 1.5 million acres in farms.



- Using a SRSWOR, the estimate for the total number of acres devoted to farms in 1992 can be found to be

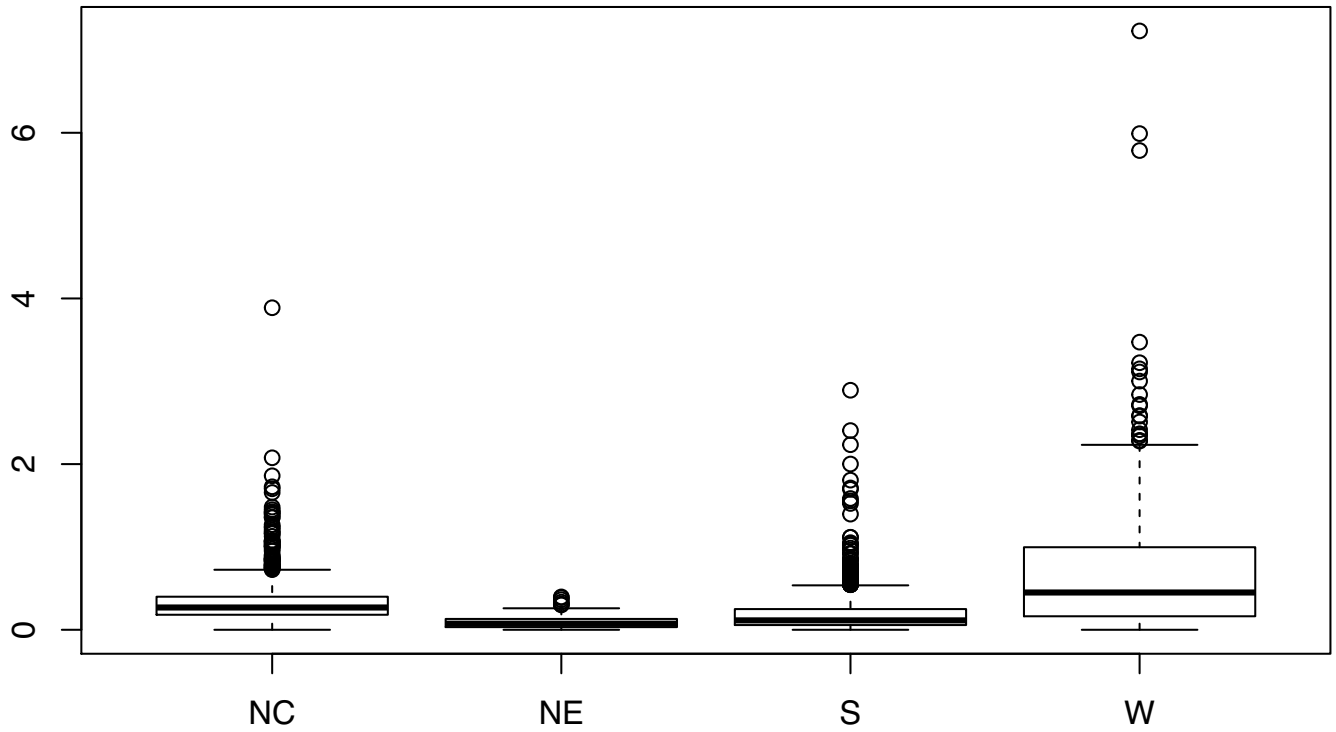
$$\hat{\tau} = N\hat{\mu} = 3078 * \frac{95385175}{300} = 978,651,895$$

- The variance (using the equations from lecture 3) is:

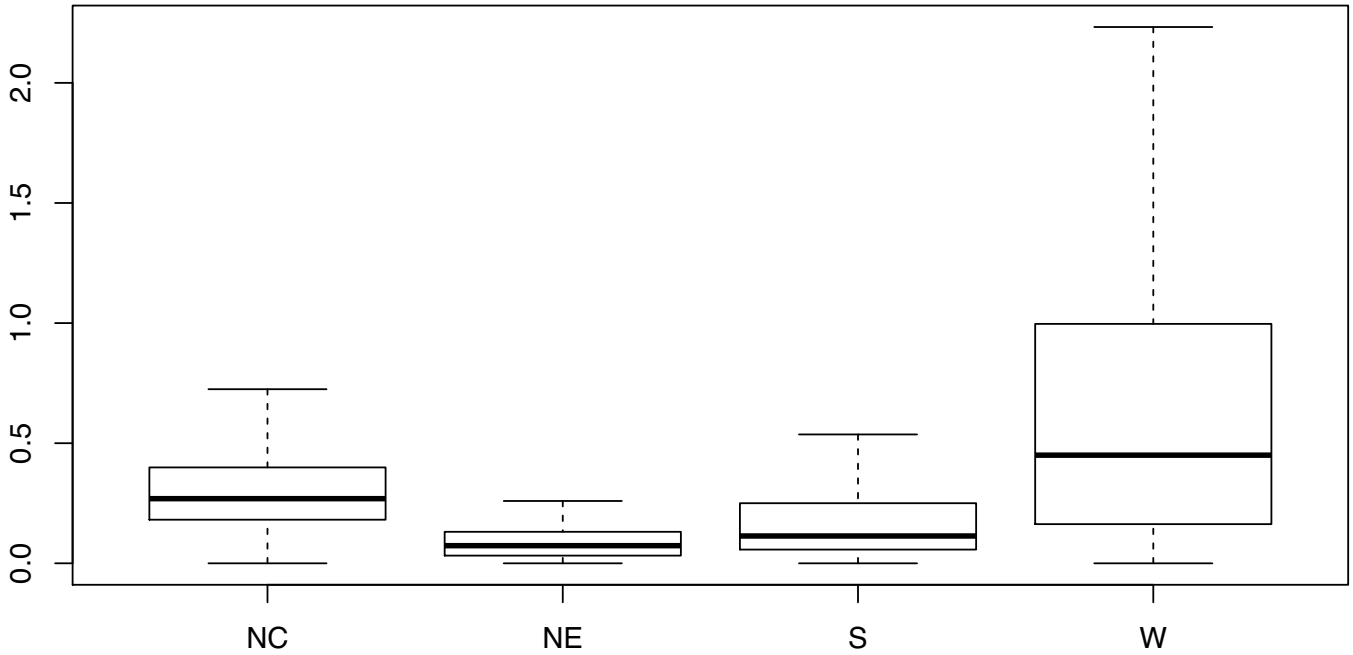
$$\begin{aligned} Var(\tilde{\tau}) &= Var(N\tilde{\mu}) = N^2 Var(\tilde{\mu}) \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{\sigma}^2}{n} \\ &= 3078^2 * \left(1 - \frac{300}{3078}\right) \frac{308,032,130,018}{300} \\ &= 8.779618 \times 10^{15} \\ \implies s.e.(\hat{\tau}) &= 93,699,616 \end{aligned}$$

Lohr (2009) Example 2.5

- You might conjecture that part of the large variability arises because counties in the western United States are larger, and thus tend to have larger values of y, than counties in the eastern United States.



Boxplot without the outliers



Region	Variance
NC	73,605,538,251.16
NE	6,298,874,252.79
S	59,514,700,947.91
W	698,292,372,000.28

Lohr (2009) Example 3.2 (extension to 2.5)

- We could use the four census `regions` of the United States — Northeast, North Central, South, and West — as strata.
- The SRSWOR in Example 2.5 sampled about 10% of the population.
- To be able to compare the results of the stratified sample with the SRSWOR, we also sample about 10% of the counties in each stratum.
- We will later discuss other stratified sampling designs.

Lohr (2009) Example 3.2 (extension to 2.5)

Region	N_h	n_h
North Central	1054	103
NorthEast	220	21
Southern	1382	135
West	422	41

- We take a SRSWOR for each of the four strata (regions).
- The four SRSWORs are selected independently: Knowing which counties are in the sample from the Northeast tells us nothing about which counties are in the sample from the South.

Lohr (2009) Example 3.2 (extension to 2.5)

The data sampled from all four strata are in data file `agstrat.csv` which you can download on the book's website. See lab for a reproduction of these results.

Region	Sample Size	Average	Variance
North Central	103	300,504.16	29,618,183,543
NorthEast	21	97,629.81	7,647,472,708
Southern	135	211,315.04	53,587,487,856
West	41	662,295.51	396,185,950,266

Stratum Total Estimates

Since we took an SRSWOR in each stratum, we can use (2) and to estimate the population quantities for each stratum. We use
For a each stratum $h = 1, 2, \ldots, H$ we can estimate the population quantities such as totals

$$\hat{\tau}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} = N_h * \hat{\mu}_h$$

$$\begin{aligned}
\implies \hat{\tau}_{NC} &= 1054 * 300,504.16 = 316,731,384.64 \\
\hat{\tau}_{NE} &= 220 * 97,629.81 = 21,478,558.2 \\
\hat{\tau}_S &= 1382 * 211,315.04 = 292,037,385.28 \\
\hat{\tau}_W &= 422 * 662,295.51 = 279,488,705.22
\end{aligned}$$

Compared with the truth:

NC	NE	S	W
343,552,309	19,936,172	275,212,084	305,251,153

Lohr (2009) Example 3.2 (extension to 2.5)

Regions	Estimated total of Farm Acres	Estimated variance of total
North Central	316,731,379.73	2.882321×10^{14}
NorthEast	21,478,558.10	1.594316×10^{13}
Southern	292,037,391.42	6.840756×10^{14}
West	279,488,706.15	1.553648×10^{15}

Lohr (2009) Example 3.2 (extension to 2.5)

We can estimate the total number of acres devoted to farming in the United States in 1992 by adding the totals for each stratum; see (4);

$$\begin{aligned}\tilde{\tau}_{\text{str}} = \sum_{h=1}^{H} \tilde{\tau}_h &= 316,731,379.73 + 21,478,558.10 + \\ &\quad 292,037,391.42 + 279,488,706.15 \\ &= 909,736,035\end{aligned}$$

As sampling was done independently in each stratum, the variance of the total is the sum of the variances of the stratum totals; see (6):

$$\hat{V}(\tilde{\tau}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{\hat{\sigma}_h^2}{n_h} = 2.541899 \times 10^{15}$$

$$\implies s.e.(\tilde{\tau}_{\text{str}}) = \sqrt{2.541899 \times 10^{15}} = 50,417,248$$

Lohr (2009) Example 3.2 (extension to 2.5)

- Using Stratified sampling instead of SRSWOR decreased the standard error or $\tilde{\tau}$ from 93,699,616 to 50,417,248!
- In addition, stratified sampling ensures that each region of the United States is represented in the sample, note however, that some states still remain unsampled (AK, DE, NH, RI)
- Note that the estimate using stratified sampling was significantly smaller than that obtained using SRSWOR (909,736,035 vs. 978,651,895, respectively).
- Interestingly, the estimates are off by roughly the same amount:

$$\tau = 943,951,718$$
$$943,951,718 - 909,736,035 = 34,215,683$$
$$943,951,718 - 978,651,895 = -34,700,177$$

Lohr (2009) Example 3.2 (extension to 2.5)

- **Take-home message:** Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate.

- **Implications:** With smaller variance, we would expect that we would need less observations with a stratified sample to obtain the same precision as from an SRSWOR.

- **Food for thought:** We need not sample the same fraction of observations in every stratum. In this example, there is far more variability from county to county in the western region. We can reduce the variance of τ even further by taking a higher sampling fraction in the western region than in the other regions.

Stratified Random Sampling: Extensions

Questions to explore:

- When is stratified sampling more efficient than simple random sampling?
- Is there an optimal way to allocate observations to strata? If so, what is it? i.e. for a fixed sample size n, how to choose $n_1, n_2, \ldots, n_H$.

If we can answer these questions we can (in theory) produce even more precise estimates.

Proportional Allocation

- We can modify $n_1, n_2, \ldots, n_H$ or equivalently the sampling weights $w_h = \frac{n_h}{n}$.
- If the sampling weights are equal to the population weights we have **proportional allocation**.
- Mathematically, this means $\frac{n_h}{n} = \frac{N_h}{N}$, so $\frac{n}{N} = \frac{n_h}{N_h}$ and $n_h \propto N_h$.
  - e.g. if $N = 1000$ and $N_1 = 500$, $N_2 = 300$, and $N_3 = 200$ then a proportional allocation with $n = 100$ would have $n_1 = 50$, $n_2 = 30$, and $n_3 = 20$

Proportional Allocation Variance

In the case of proportional allocation, we have

$$W_h = \frac{N_h}{N} = \frac{n_h}{n} \text{ and } \frac{n_h}{N_h} = \frac{n}{N},$$

so the variance of the stratified estimator becomes

$$\begin{aligned} Var(\tilde{\mu}_{\text{str}}) &= \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} \\ &= \sum_{h=1}^{H} W_h \frac{n_h}{n} \left(1 - \frac{n}{N}\right) \frac{\sigma_h^2}{n_h} \\ &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} W_h \sigma_h^2 \end{aligned}$$

Proportional Allocation vs SRSWOR

We now compare the variance of the stratified sample estimator (via proportional allocation) with that of SRSWOR:

$$Var(\tilde{\mu}_{\text{str}}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} W_h \sigma_h^2$$

$$Var(\tilde{\mu}_{SRSWOR}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\sigma^2$$

and the stratified estimator variance will be lower if

$$\sum_{h=1}^{H} W_h \sigma_h^2 < \sigma^2$$

Proportional Allocation vs SRSWOR

- The left side is the weighted average of within-stratum variances.
- If we form our strata such that these variances are small, we should get a more precise estimate.
- In other words, we gain efficiency if there is greater consistency within strata than compared to the whole population.

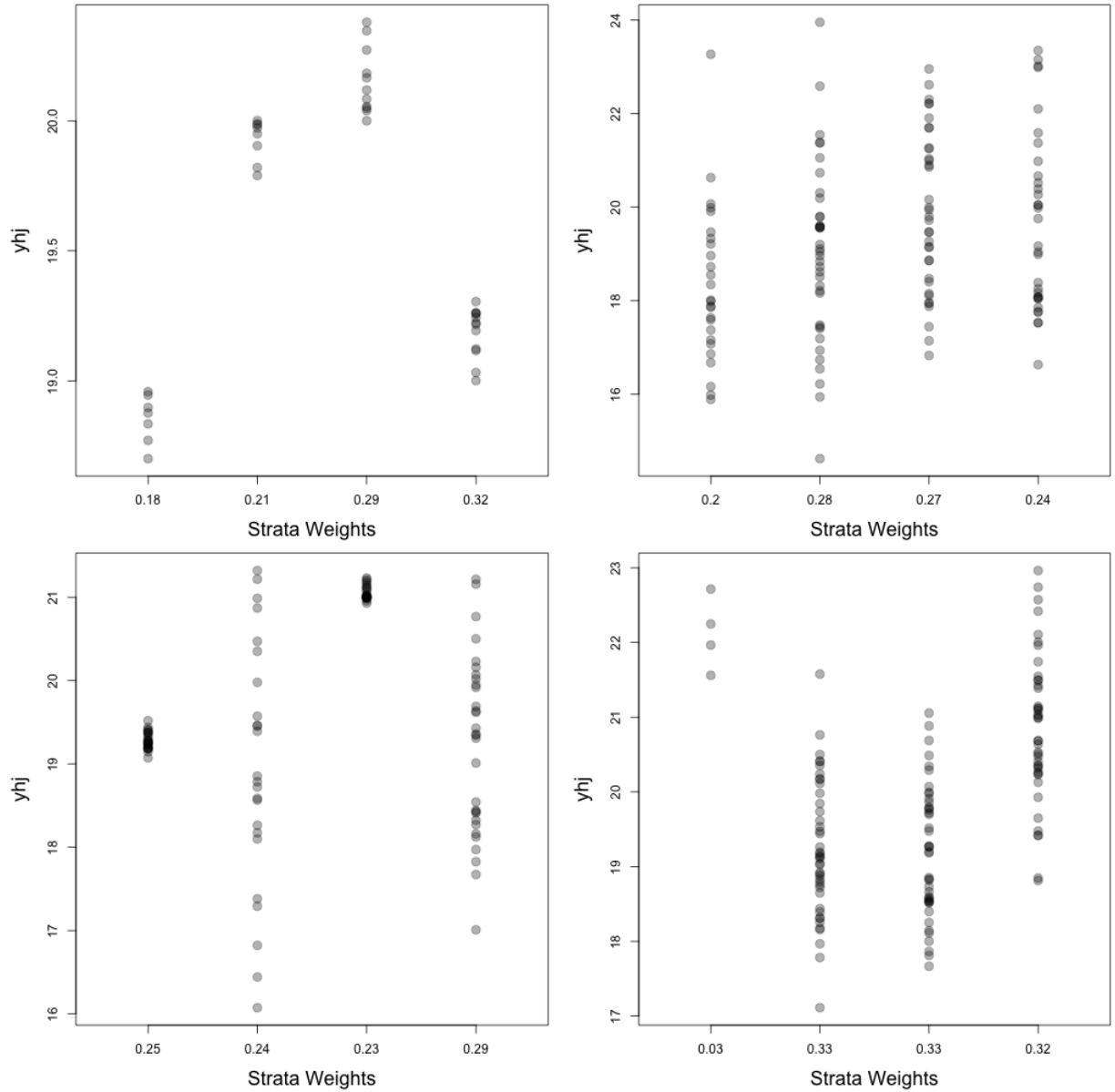
Proportional Allocation vs SRSWOR

- Another way to see this is to decompose the total sum of squares into two components, within and between strata.
- It can be shown that the difference in variance for stratified versus the sample average estimator is proportional to $\sum_h W_h(\mu_h - \mu)^2$.
- This suggests if our strata have different means to one another we'll achieve the greatest gain in efficiency over simple random sampling.

Constructing Strata

- If we form our strata such that the within-stratum variances are small, we get a more precise estimate.
- i.e. we gain efficiency if there is greater consistency within strata than compared to the whole population.
- Key idea: divide our population into strata of similar units (e.g. provinces, or university faculties)

Four Strata Examples



Optimal Allocation

- We may have some control on how to form strata, but this must be done before we collect data.
- We'll likely have more control over how many units we draw from each stratum.
- Suppose we know our sample size will be n.
- How do we choose our stratum sample sizes to minimize the variance of the resulting estimator?
- This is called the **allocation problem**.

Optimal Allocation

- It can be shown using a Lagrange multiplier that the optimal allocation is

$$n_h = \frac{W_h \sigma_h}{W_1 \sigma_1 + ... + W_H \sigma_H} n$$

  or, more simply $n_h \propto$ (is proportional to) $W_h \sigma_h$

- There are therefore two factors that should lead us to draw a larger sample for a given stratum: If that stratum has a
  - higher within-stratum standard deviation or
  - higher stratum weight.

- N.B. if σ_h is the same for each stratum, then proportional allocation is optimal!

Optimal Allocation Example

For optimal allocation we therefore need to know (or estimate) the within-stratum standard deviations (perhaps from a pilot study).

As a quick example, suppose we wanted to pick a sample of size $n = 100$ with the following information:

Stratum	N_h	W_h	σ_h	n_h	$W_h\sigma_h$	$n_h = \frac{W_h\sigma_h}{W_1\sigma_1 + \ldots + W_H\sigma_H} n$
1	250	0.25	5	20	1.25	$n_1 = \frac{1.25}{6.25}(100) = 20$
2	250	0.25	10	40	2.5	$n_2 = \frac{2.5}{6.25}(100) = 40$
3	500	0.50	5	40	2.5	$n_3 = \frac{2.5}{6.25}(100) = 40$
					6.25	

n_h from this

Note: When calculating sample sizes we need to round up to the nearest whole number since we can't sample a fraction of a person/subject.

Forming the Strata

- We've seen how stratified random sampling can lead to more precise estimates (i.e., with shorter confidence intervals).
  - This means if we can form sensible strata we can either improve precision with the same sample size, or
  - attain the same precision with a smaller sample size, compared with other techniques.
- Forming strata can be complex, and often requires subject-level knowledge (i.e., statisticians talking to non-statisticians).
  - If we can't (or are unwilling) to estimate the within-stratum standard deviations, then proportional allocation is a reasonable strategy if we want to improve precision.

Stratified Sampling Summary

- The population is divided into H strata.
- Each strata has μ_h and σ_h
- The population stratum h weight is $W_h = \frac{N_h}{N}$
- The number of units selected from strata h is n_h
- Point estimate and estimator variance

$$\hat{\mu}_{\text{str}} = \sum_{h=1}^{H} W_h \hat{\mu}_h, \quad Var(\tilde{\mu}_{\text{str}}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

- Proportional Allocation $n_h \propto W_h$
- Optimal Allocation $n_h \propto W_h \sigma_h$
- Construct strata with small σ_h and large differences among the strata averages.

Advantages:

- All strata are represented in the sample
- Improves the representativeness of the sample by reducing sampling error
- A stratified survey could be more representative of the population than a survey of simple random sampling or systematic sampling
- It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population
- You can compare between different strata

Disadvantages

- Need names of all population members
- There is difficulty in reaching all selected in the sample
- Stratified sampling is not useful when the population cannot be exhaustively partitioned into disjoint subgroups

References:

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.