3.5 Attribute Proportion Estimation for a Stratified Population

- Suppose the goal is to estimate the population proportion $p = t/N$ where t is the number of units in the population possessing the attribute.
- Consider the indicator function that assigns to each sampling unit the following y_{hj} value:

$$\begin{aligned} y_{hj} &= 1 \quad \text{if unit } i \text{ in stratum } h \text{ possesses the attribute} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

- Let $p_h = \dfrac{1}{N_h}\displaystyle\sum_{j=1}^{N_h} y_{hj}$ = proportion of stratum h units that possess the attribute. Then

$$t = \sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} \qquad \text{and} \qquad p = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H}\frac{N_h}{N}p_h.$$

- Because each p_h is unknown, we can estimate p_h with $\widehat{p}_h = \dfrac{1}{n_h}\displaystyle\sum_{j=1}^{n_h} y_{hj}$ = the sample proportion of units from stratum h that possess that attribute.
- Because each sample proportion $\widehat{p}_h$ is an unbiased estimator of stratum proportion p_h, an unbiased estimator $\widehat{p}_{str}$ of the population proportion p is a weighted average of the $\widehat{p}_h$'s:

$$\widehat{p}_{str} = \qquad\qquad = \sum_{h=1}^{H} W_h\widehat{p}_h \qquad \text{where}$$

- Recall: for SRS attribute sampling, $S^2 = \dfrac{Np(1-p)}{N-1}$. Because we are taking a SRS, $S_h^2 =$
- Recall: for SRS attribute sampling: $V(\widehat{p}) = \left(\dfrac{N-n}{N}\right)\dfrac{S^2}{n} = \left(\dfrac{N-n}{N-1}\right)\dfrac{p(1-p)}{n}$. Thus,

$$V(\widehat{p}_h) = \left(\frac{N_h-n_h}{N_h}\right)\frac{S_h^2}{n_h} = \left(\frac{N_h-n_h}{N_h-1}\right)\frac{p_h(1-p_h)}{n_h}.$$

- By pooling the stratum variances:

$$\begin{aligned} V(\widehat{p}_{str}) &= \sum_{h=1}^{H} W_h^2 V(\widehat{p}_h) = \sum_{h=1}^{H} W_h^2\left(\frac{N_h-n_h}{N_h}\right)\frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{H} W_h^2\left(\frac{N_h-n_h}{N_h-1}\right)\frac{p_h(1-p_h)}{n_h} \qquad (33) \end{aligned}$$

- Recall: for SRS attribute sampling: $s^2 = \dfrac{n\widehat{p}(1-\widehat{p})}{n-1}$. Thus, $s_h^2 =$

- Because S_h^2 is unknown, we use s_h^2 to get the unbiased estimator of the variance in (33):

$$\widehat{V}(\widehat{p}_{str}) \quad = \quad \sum_{h=1}^{H} W_h^2 \left(\frac{N_n - n_h}{N_h}\right) \frac{s_h^2}{n_h} \quad = \quad \sum_{h=1}^{H} W_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\widehat{p}_h(1-\widehat{p}_h)}{n_h - 1} \qquad (34)$$

- To generate a confidence interval for p, calculate

$$\widehat{p}_{str} \pm t^* \sqrt{\widehat{V}(\widehat{p}_{str})} \qquad (35)$$

  where t^* is the upper $\alpha/2$ critical value from the $t(d)$ distribution. In this case, d is Satterthwaite's approximate degrees of freedom d (see equation (32)).

- For larger sample sizes, some people use z^* instead of t^*.

- **Sample Size Determination**: You can use the same sample size formulas defined earlier for estimating $\overline{y}_U$ for a stratified SRS.

Stratification Example using Longleaf Pine Data Data

- This 200 × 200 m study region is located in an old-growth forest in Thomas County, Georgia, USA. The region has been divided into a 20 × 20 grid of 10 m ×10 m quadrats. This data represents the presence=1 or absence=0 (Table 6) of longleaf pine trees located in each quadrat. The population proportion $p = 249/400 = 0.6225$.

- The longleaf pine census data will be stratified into four 10 × 10 strata. The stratum sample sizes are $n_h = 10$ for $h = 1, 2, 3, 4$. The values in parentheses are the sampled quadrats.

Figure 6: The Presence or Absence of Longleaf Pine

1	1	1	1	(1)	1	1	(0)	0	0	(1)	1	0	1	0	1	1	1	0	1
1	1	1	0	1	0	0	0	1	1	1	1	0	1	1	1	0	1	0	1
1	1	1	1	1	1	0	(0)	0	1	1	0	1	1	1	1	1	1	1	(1)
0	1	1	0	0	0	(0)	(0)	(1)	1	1	1	1	(0)	0	0	1	1	1	0
1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	1	1	1	(0)	1	1	1	0	1	1	1	0	0	1	1	(1)
0	0	(1)	1	1	1	1	1	0	1	0	1	0	(1)	1	0	1	(1)	(1)	0
0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1
0	0	1	1	1	1	1	1	1	1	1	1	(0)	0	1	1	0	1	1	0
0	0	(0)	0	1	0	1	1	1	(1)	1	0	1	0	1	0	(1)	1	(1)	0
1	1	1	1	1	0	1	0	1	0	1	1	0	0	1	1	1	1	(1)	1
0	1	1	0	(0)	1	(1)	1	(0)	0	0	1	1	1	0	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1
0	1	(0)	0	1	(1)	0	1	1	1	1	0	0	(1)	1	1	1	0	(1)	1
1	0	0	1	1	0	(0)	0	(1)	1	1	1	1	1	0	0	1	0	(0)	0
0	(1)	0	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	(1)
0	0	0	1	1	1	0	0	0	0	0	1	1	(0)	1	1	1	(1)	1	1
1	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0	(1)	1	1	0
0	0	0	0	0	0	(0)	1	1	1	0	1	0	1	0	(1)	0	1	1	0
1	0	0	0	0	0	0	(0)	1	1	0	(1)	1	0	0	1	0	1	1	1

3.5.1 Using R and SAS to Analyze a Stratified SRS for a Proportion

R Code to Analyze a Stratified SRS for a Proportion

```
source("c:/courses/st446/rcode/confintt.r")

# t-based confidence intervals for SRS in Figure 6

library(survey)

pa <- c(rep(0,6),rep(1,4),rep(0,6),rep(1,4),
rep(0,2),rep(1,8),rep(0,2),rep(1,8))

stratum <- c(rep(1,10),rep(2,10),rep(3,10),rep(4,10))
fpc     <- c(rep(100,10),rep(100,10),rep(100,10),rep(100,10))

strat6dat <- data.frame(pa,fpc,stratum)
strat6dat

strat_design <- svydesign(id=~1, fpc=~fpc, strata=~stratum, data=strat6dat)
strat_design

estp <- svymean(~pa,strat_design)
print(estp,digits=15)
confint.t(estp,degf(strat_design),level=.95)
confint.t(estp,degf(strat_design),level=.95,tails=’lower’)
confint.t(estp,degf(strat_design),level=.95,tails=’upper’)
```

R Output

```
   pa fpc stratum
1   0 100       1           21  0 100       3
2   0 100       1           22  0 100       3
3   0 100       1           23  1 100       3
4   0 100       1           24  1 100       3
5   0 100       1           25  1 100       3
6   0 100       1           26  1 100       3
7   1 100       1           27  1 100       3
8   1 100       1           28  1 100       3
9   1 100       1           29  1 100       3
10  1 100       1           30  1 100       3
11  0 100       2           31  0 100       4
12  0 100       2           32  0 100       4
13  0 100       2           33  1 100       4
14  0 100       2           34  1 100       4
15  0 100       2           35  1 100       4
16  0 100       2           36  1 100       4
17  1 100       2           37  1 100       4
18  1 100       2           38  1 100       4
19  1 100       2           39  1 100       4
20  1 100       2           40  1 100       4

Stratified Independent Sampling design

-----------------------------------------------------------------
mean( pa ) = 0.60000
SE( pa ) = 0.07071
Two-Tailed CI for pa where alpha = 0.05 with 36 df
    2.5 %        97.5 %
  0.45659     0.74341
-----------------------------------------------------------------
```

```
-------------------------------------------------------------------
mean( pa ) = 0.60000
SE( pa ) = 0.07071
One-Tailed (Lower) CI for pa where alpha = 0.05 with 36 df
     5 %        upper
  0.48062     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( pa ) = 0.60000
SE( pa ) = 0.07071
One-Tailed (upper) CI for pa where alpha = 0.05 with 36 df
    lower         95 %
  -infinity     0.71938
-------------------------------------------------------------------
```

SAS Code to Analyze a Stratified SRS for a Proportion

```
data fig5c;
   do Area = 1 to 4;
   do i = 1 to 10;
      input ind @@; output;
   end; end;
datalines;
1 0 1 0 0 0 0 0 1 1
1 0 0 1 1 0 0 0 0 1
1 0 0 1 1 1 1 1 1 1
1 1 0 1 1 1 1 1 0 1
;
data fig5c; set fig5c;

  if Area = 1 then _total_= 100;  *** _total_ = Nh ;
  if Area = 2 then _total_= 100;
  if Area = 3 then _total_= 100;
  if Area = 4 then _total_= 100;

  if Area=1 then W = 100/10;       *** W = Nh / nh ;
  if Area=2 then W = 100/10;
  if Area=3 then W = 100/10;
  if Area=4 then W = 100/10;

  if ind = 0 then pa = 'absent ';
  if ind = 1 then pa = 'present';

title1 'Analysis of Stratified SRS in Figure 5c';
title2 'Estimate Proportion p of Quadrats with Presence of Trees';

proc surveymeans data=fig5c total=fig5c mean clm df;
    stratum Area / list;
    var pa;
    weight W;
run;
```

```
                        Analysis of Stratified SRS in Figure 6
                Estimate Proportion p of Quadrats with Presence of Trees

                              The SURVEYMEANS Procedure

                                    Data Summary

                          Number of Strata                  4
                          Number of Observations           40
                          Sum of Weights                  400

                               Class Level Information

                        Class
                        Variable      Levels    Values
                        pa                 2    absent present

                                 Stratum Information

Stratum          Population  Sampling
 Index    Area        Total      Rate     N Obs  Variable  Level        N
-------------------------------------------------------------------------
   1         1          100     10.0%        10  pa        absent       6
                                                           present      4
   2         2          100     10.0%        10  pa        absent       6
                                                           present      4
   3         3          100     10.0%        10  pa        absent       2
                                                           present      8
   4         4          100     10.0%        10  pa        absent       2
                                                           present      8
-------------------------------------------------------------------------

                                     Statistics

                                          Std Error
Variable  Level      DF         Mean        of Mean     95% CL for Mean
-------------------------------------------------------------------------
pa        absent     36     0.400000       0.070711  0.25659210 0.54340790
          present    36     0.600000       0.070711  0.45659210 0.74340790
-------------------------------------------------------------------------
```

3.6 Quota Sampling

- When the population is stratified based on more than one factor we are using **multifactor stratification**. Typically, a sample is taken within every combination of strata across the factors.

  - A marketing survey was conducted to get information on computer users' use of the internet for shopping. The surveyor suspected there may be differences based on gender (2 levels), household income (6 levels), and age (5 levels). There would be $2 \times 6 \times 5 = 60$ multifactor strata to sample from.

- **Quota sampling** is a form of stratified sampling and typically uses multifactor stratification. Taking a quota sample ensures that data are collected across the population with the belief that doing so will provide a representative sample from the population. It also allows the researcher to generate estimates related to various subgroups.

- So how does quota sampling differ from stratified simple random sampling? In quota sampling, **the within stratum samples may not be random**. Often some element of subjectivity enters into the sampling procedure.

- A typical quota sample is based on:

  1. Defining the multifactor strata.
  2. Determining the stratum sample sizes based on proportional allocation. These are also referred to as the *stratum quotas*. These samples sizes are based on either known or approximate stratum sizes. They can also be based on either known or approximate population proportions associated with each stratum.
  3. Data is collected by predetermined data collection techniques (e.g., phone surveys, mail surveys, personal interviews, etc.) until the stratum quotas are satisfied (that is, until the desired number of responses are collected for each stratum).

- Although taking a quota sample can save a lot of time and money when compared to simple random sampling, the researcher must realize that if quota sampling is used, we cannot be sure that the selection of sampling units would be similar to units collected via simple random sampling.

- Quota sampling is the primary method of sampling used by many commercial data-collection organizations. It is a common method used in political polls and surveys of consumer attitudes regarding products. It is also common in medical studies in which the researchers will select subjects satisfying the requirements for admittance to a study until a desired number is reached.

- In principle, use of the estimation formulas based on random sampling techniques on data from a quota sample violates underlying assumptions. Therefore, we cannot be as confident in the results from quota sampling in comparison to results when a probability sampling method is used.

  - A student organization wants to determine if students favor extending the evening hours that the library remains open. They decided to take a quota sample based on strata related to class standing (freshmen, sophomores, juniors, seniors, graduate students). After deciding on the 5 quotas of 25 students for each of the 5 strata, data was collected at the library on consecutive evenings until all 5 quotas were satisfied. What concerns do you have?
  - You are sampling students who are known to study in the library and not sampling any students who are less likely to study in the library. Therefore, the recorded responses will be biased in support of extending the evening hours that the library remains open.

- Because nonresponse is often ignored in quota sampling, the resulting estimates can be seriously biased.

- However, quota sampling has proven useful if the quotas are designed properly with careful attention paid to when, how, and where the data are collected.

Additional References

Ames, M.H. and Webster J.T. (1991). On estimating approximate degrees of freedom. *The American Statistician* **45** 45-50.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110-114.

3.7 Poststratification

- In some stratified sampling situations, we may not be able to determine from which stratum an observation belongs until it is actually observed.

- For example, a fish population may be stratified by age class, body length, body weight, or sex. A random sample of fish is collected. Then each sampled fish is examined and its stratum is recorded along with the response of interest. Therefore, because the stratum for each fish is determined after sampling, it is not possible to guarantee exact stratum samples sizes n_h for each age class, etc.

- In many studies a simple random sample is taken from the population and then is stratified. This procedure is known as **poststratification** or **post-hoc stratification**.

- In contrast to traditional stratified simple random sampling, the stratum sample sizes $n_1, n_2, \ldots, n_k$ are random variables.

- Bartlett (1988) provides references that discuss poststratification. Thompson (1992) derives an approximation to $V(\widehat{\overline{y}_{U\,str}})$.

Poststratification excerpt from Thompson (1992).

In some situations it may be desired to classify the units of a sample into strata and to use a stratified estimate, even though the sample was selected by simple random, rather than stratified, sampling. For example, a simple random sample of a human population may be stratified on sex after selection of the sample, or a simple random sample of sites in a fishery survey may be poststratified on depth. In contrast to conventional stratified sampling, with poststratification, the stratum sample sizes $n_1, n_2, \ldots, n_L$ are random variables.

With proportional allocation in conventional stratified random sampling, the sample size in stratum h is fixed at $n_h = nN_h/N$ and the variance (Equation 2) simplifies to $\text{var}(\bar{y}_{st}) = [(N - n)/nN)] \sum_{h=1}^{L} (N_h/N)\sigma_h^2$. With poststratification of a simple random sample of n units from the whole population, the sample size n_h in stratum h has expected value nN_h/N, so that the resulting sample tends to approximate proportional allocation. With poststratification the variance of the stratified estimator $\bar{y}_{st} = \sum_{h=1}^{L}(N_h/N)\bar{y}_h$ is approximately

$$\text{var}(\bar{y}_{st}) \approx \frac{N-n}{nN} \sum_{h=1}^{L} \left(\frac{N_h}{N}\right) \sigma_h^2 + \frac{1}{n^2}\left(\frac{N-n}{N-1}\right) \sum_{h=1}^{L} \frac{N-N_h}{N} \sigma_h^2 \qquad (5)$$

and the variance of $\hat{\tau}_{st} = N\bar{y}_{st}$ is $\text{var}(\hat{\tau}_{st}) = N^2 \text{var}(\bar{y}_{st})$. The first term is the variance that would be obtained using a stratified random sampling design with proportional allocation. An additional term is added to the variance with poststratification, due to the random sample sizes.

For a variance estimate with which to construct a confidence interval for the population mean with poststratified data from a simple random sample, it is recommended to use the standard stratified sampling method (Equation 3), rather than substituting the sample variances directly into Equation 5. With poststratification, the standard formula (Equation 3) estimates the conditional variance (given by Equation 2) of $\bar{y}_{st}$ given the sample sizes $n_1, \ldots, n_L$, while Equation 5 is the unconditional variance (and see comments of J.N.K. Rao, 1988, p. 440).

To use poststratification, the relative size N_h/N of each stratum must be known. If the relative stratum sizes are not known, they may be estimated using double sampling (see the later chapter on that topic). Further discussion of poststratification may be found in Cochran (1977), Hansen, Hurwitz, and Madow (1953), Hedayat and Sinha (1991), Kish (1965), Levy and Lemeshow (1991), Singh and Chaudhary (1986), and Sukhatme and Sukhatme (1970). Variance approximations for poststratification vary among the sampling texts. The derivation for the expression given here is given in the final section of the text of this chapter ("Poststratification Variance").