

Stratified Random Sampling (Chapter 11)

This handout introduces the basic ideas and theory behind stratified random sampling estimators, the stratification principle, allocation in stratified random sampling, a number of examples illustrating the method, compares simple and stratified random sampling, and introduces the ideas behind post-stratification.

- Suppose we divide the population into L strata, where the variation within strata is small relative to the variation between strata, in terms of some underlying response variable. We discussed and saw in an earlier handout that this situation minimizes the variability in the stratified random sampling estimator.
- Examples: Landscapes - stratified by habitat characteristics,
People - stratified by characteristics (such as sex, income, etc.).

Notation: N_h = the population size in stratum h , $h = 1, 2, \dots, L$,
 N = $\sum_{h=1}^L N_h$ = the total population size,
 n_h = the sample size in stratum h , $h = 1, 2, \dots, L$,
 n = $\sum_{h=1}^L n_h$ = the total sample size,
 y_{hi} = the i^{th} observation in the h^{th} stratum,
 τ_h = $\sum_{i=1}^{N_h} y_{hi}$ = the total of the observations in stratum h ,
 τ = $\sum_{h=1}^L \tau_h$ = the overall total,
 μ_h = τ_h/N_h = the mean response in stratum h ,
 μ = τ/N = the overall mean response.

Estimating τ and τ_h : Within each stratum, we estimate τ_h by $\hat{\tau}_h = N_h \bar{y}_h$. Then $\hat{\tau}_{st} = \sum_{h=1}^L \hat{\tau}_h$.

- If $\hat{\tau}_h$ is an unbiased estimator of τ_h , $h = 1, \dots, L$, then $\hat{\tau}_{st} = \sum_{h=1}^L \hat{\tau}_h$ is unbiased for τ . Note that we could have a different sampling plan (other than an SRS) in each stratum.
- Also, if the stratum samples are independently selected, then:

$$\underline{\text{Var}(\hat{\tau}_{st})} = \text{Var} \left(\sum_{h=1}^L \hat{\tau}_h \right) = \underline{\sum_{h=1}^L \text{Var}(\hat{\tau}_h)} \quad (\text{due to the independence of the } \hat{\tau}_h \text{'s}).$$

- If $\widehat{\text{Var}}(\hat{\tau}_h)$ is unbiased for $\text{Var}(\hat{\tau}_h)$, then $\widehat{\text{Var}}(\hat{\tau}_{st}) = \sum_{h=1}^L \widehat{\text{Var}}(\hat{\tau}_h)$ is unbiased for $\text{Var}(\hat{\tau}_{st})$.

Estimating μ and μ_h : $\hat{\mu}_{st} = \hat{\tau}_{st}/N$ is an unbiased estimator of μ if $\hat{\tau}$ is unbiased for τ
and $\text{Var}(\hat{\mu}_{st}) = \frac{1}{N^2} \text{Var}(\hat{\tau}_{st})$, so that $\widehat{\text{Var}}(\hat{\mu}_{st}) = \frac{1}{N^2} \widehat{\text{Var}}(\hat{\tau}_{st})$.

- An alternative form for the estimator $\hat{\mu}_{st}$ is given by:

$$\hat{\mu}_{st} = \frac{1}{N} \hat{\tau}_{st} = \frac{1}{N} \sum_{h=1}^L \hat{\tau}_h = \frac{1}{N} \sum_{h=1}^L N_h \hat{\mu}_h = \sum_{h=1}^L \underbrace{\left(\frac{N_h}{N} \right)}_{\text{weights}} \hat{\mu}_h,$$

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 $\hat{\mu}_{st}$ may then be expressed as:

$$\text{Var}(\hat{\mu}_{st}) = \text{Var} \left(\sum_{h=1}^L \left(\frac{N_h}{N} \right) \hat{\mu}_h \right) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \text{Var}(\hat{\mu}_h) \left(\begin{array}{c} \text{under} \\ \text{independence} \end{array} \right).$$

- The results derived above are true for any sampling plans within each stratum, not just simple random sampling. These general results fall under the heading of “stratified sampling.”

Note: “Stratified random sampling” means independent simple random samples (SRS’s) taken within each stratum. Under this setting, the stratified estimator of the population mean and total can be derived as follows.

Within stratum h : $\hat{\tau}_h = N_h \bar{y}_h$ ($\hat{\mu}_h = \bar{y}_h$), where \bar{y}_h is the sample mean in stratum h .

$$\boxed{\hat{\tau}_{st}} = \sum_{h=1}^L \hat{\tau}_h = \boxed{\sum_{h=1}^L N_h \bar{y}_h} \quad (\text{the estimated total from stratified random sampling})$$

$$\begin{aligned} \boxed{\text{Var}(\hat{\tau}_{st})} &= \sum_{h=1}^L \text{Var}(\hat{\tau}_h) = \sum_{h=1}^L N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\sigma_h^2}{n_h} = \boxed{\sum_{h=1}^L N_h (N_h - n_h) \frac{\sigma_h^2}{n_h}} \\ &\Rightarrow \widehat{\text{Var}}(\hat{\tau}_{st}) = \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h}. \end{aligned}$$

$$\boxed{\hat{\mu}_{st}} = \bar{y}_{st} =$$

$$\boxed{\text{Var}(\hat{\mu}_{st})} = \frac{1}{N^2} \text{Var}(\hat{\tau}_{st}) = \boxed{\sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\sigma_h^2}{n_h}}, \quad \text{where } \widehat{\text{Var}}(\hat{\mu}_{st}) \text{ replaces } \sigma_h^2 \text{ with } s_h^2.$$

Example: 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.

Why might we stratify the population in this way?

Statistic	Town A	Town B	Rural	
h	1	2	3	
N_h	155	62	93	($N = 310$)
n_h	20	8	12	(SRS's)
\bar{y}_h	33.90	25.12	19.00	
s_h	5.95	15.24	9.36	
$\hat{\tau}_h$	5254.5	1557.4	1767.0	($N_h \bar{y}_h$)

$$\hat{\tau}_{st} = \hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3 = 8578.9,$$

$$\bar{y}_{st} = \frac{\hat{\tau}_{st}}{N} = \frac{8578.9}{310} = \underline{27.7}$$

Other way:

$$\begin{aligned} \bar{y}_{st} &= \sum_{h=1}^L \left(\frac{N_h}{N} \right) \bar{y}_h \\ &= \frac{155}{310}(33.90) + \frac{62}{310}(25.12) + \frac{93}{310}(19) = 16.95 + 5.024 + 5.7 = \underline{27.7}. \end{aligned}$$

$$\begin{aligned} \widehat{\text{Var}}(\bar{y}_{st}) &= \sum_{h=1}^3 \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h} \\ &= \left(\frac{155}{310} \right)^2 \left(\frac{155 - 20}{155} \right) \frac{5.95^2}{20} + \left(\frac{62}{310} \right)^2 \left(\frac{62 - 8}{62} \right) \frac{15.24^2}{8} + \left(\frac{93}{310} \right)^2 \left(\frac{93 - 12}{93} \right) \frac{9.36^2}{12} \\ &= 0.385 + 1.011 + 0.572 = \underline{1.97} \Rightarrow \underline{\text{SE}(\bar{y}_{st}) = 1.40}. \end{aligned}$$

A 95% confidence interval for μ is given by:

$$\bar{y}_{st} \pm t^* \text{SE}(\bar{y}_{st}) = 27.7 \pm (2.079)(1.40) = 27.7 \pm 2.91 = \underline{(24.79, 30.61)}.$$

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

- We generally do not assume that all the σ_h 's are equal, so a Satterthwaite approximation should be used to get the degrees of freedom associated with t^* . Here, using equation (4) on page 121 of the text, the approximate degrees of freedom are:

$$\text{d.f.} = \frac{\left(\sum_{h=1}^L a_h s_h^2\right)^2}{\sum_{h=1}^L (a_h s_h^2)^2 / (n_h - 1)} = \underline{21.1}, \quad \text{where } a_h = \frac{N_h(N_h - n_h)}{n_h}.$$

- An ultra-conservative choice for the degrees of freedom is to set:

$$\text{d.f.} = \min(n_1 - 1, n_2 - 1, \dots, n_L - 1) = 7.$$

- If all of the stratum sample sizes $n_h \geq 30$, then a z-based critical value can be used.

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). 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 determining 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 L identified strata. This handout discusses three types of allocation, and provides an example & some R code for carrying out estimation in a stratified random sample.

Allocation of a Sample to Strata

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: $n_h = \frac{n}{L}$.

2. Proportional: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes: $n_h = \left(\frac{N_h}{N}\right)n$.

- The 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 (see below).

3. Optimum (Neyman): The allocation which minimizes the variance of the estimator of the mean (and total) is given by: $n_h = \frac{nN_h\sigma_h}{\sum_{k=1}^L N_k\sigma_k}$.

- Such an allocation rule will minimize $\text{Var}(\bar{y}_{st})$ for a given n .
- This allocation can be derived (Section 11.8) by the Lagrange multiplier method: find the values of n_1, \dots, n_L which minimize $\text{Var}(\bar{y}_{st})$ subject to the constraint $n_1 + \dots + n_L = n$.
- Note that the larger the variance σ_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, $\sigma_h, h = 1, \dots, L$ (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.

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

$$\bar{y}_{st} = \sum_{h=1}^L \left(\frac{N_h}{N} \right) \bar{y}_h, \quad \text{Var}(\bar{y}_{st}) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\sigma_h^2}{n_h}.$$

Back to the TV Example

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:

$$n_1 = n \left(\frac{(155)(5.946)}{(155)(5.946) + (62)(15.24) + (93)(9.36)} \right) = \underline{n(.337)},$$

$$n_2 = n \left(\frac{(62)(15.24)}{(155)(5.946) + (62)(15.24) + (93)(9.36)} \right) = \underline{n(.345)}, \quad n_3 = n - n_1 - n_2 = \underline{n(.318)}.$$

- Suppose $n = 100$. Then we might assign $(n_1, n_2, n_3) = (34, 34, 32)$ as the 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: $\frac{\sigma_h}{\sum_{k=1}^L \sigma_k}$.

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 then: $c =$

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

$$n_h = (c - c_0) \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{k=1}^L N_k \sigma_k \sqrt{c_k}}.$$

- Note that the higher the cost of sampling c_h in stratum h , the smaller the stratum sample size n_h will be. Again, 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 μ to within some value d with $100(1 - \alpha)\%$ probability with stratified random sampling. If the goal is to estimate the population total τ to within d with $100(1 - \alpha)\%$ probability, this is equivalent to estimating μ to within d/N . In the formulas given below then, replace d by d/N if d is the allowable difference for the *total*.

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:

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

where:

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} = \frac{1}{n} \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - nw_h}{N_h}\right) \frac{\sigma_h^2}{w_h} \quad \left(\begin{array}{l} \text{using} \\ n_h = nw_h \end{array} \right).$$

Solving this margin of error equation for n leads to:

$$n = \frac{\sum_{h=1}^L \frac{N_h^2 \sigma_h^2}{w_h}}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2} = \frac{\sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{w_h}}{\frac{d^2}{z^2} + \frac{1}{N} \sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h^2}. \quad (*)$$

The rightmost expression in (*) 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:

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

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 $\text{Var}(\bar{y}_{st})$.

The formula in (*) yields the following for the three allocation schemes we have talked about:

1. Total sample size needed with equal allocation: $n_h = \frac{n}{L}$, so $w_h = \frac{1}{L}$ and

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

If the fpc is ignored, this reduces to: $n = \frac{Lz^2}{d^2} \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \sigma_h^2$.

2. Total sample size needed with proportional allocation: $n_h = \frac{nN_h}{N}$ so $w_h = \frac{N_h}{N}$ and

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

If the fpc is ignored, this reduces to: $n = \frac{z^2}{d^2} \sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h^2$.

3. Total sample size needed with optimum allocation (equal costs):

$$n_h = \frac{nN_h \sigma_h}{\sum_{k=1}^L N_k \sigma_k} \text{ so } w_h = \frac{N_h \sigma_h}{\sum_{k=1}^L N_k \sigma_k} \text{ and}$$

$$n = \frac{\left[\sum_{h=1}^L N_h \sigma_h\right]^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2} = \frac{\left[\sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h\right]^2}{\frac{d^2}{z^2} + \frac{1}{N} \sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h^2}.$$

If the fpc is ignored, this reduces to: $n = \frac{z^2}{d^2} \left[\sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h\right]^2$.

- Note: Optimum allocation is equivalent to proportional allocation when the stratum variances (σ_h^2 's) are the same.

4. Total cost with optimum allocation (unequal costs):

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 . Then, from the allocation formula on page 123 of the text (p. 75 of this handout), $n_h = c^* w_h$ where:

$$w_h = \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{k=1}^L N_k \sigma_k \sqrt{c_k}}, \text{ and } c^* = \frac{\left[\sum_{h=1}^L N_h \sigma_h \sqrt{c_h} \right]^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2} = \frac{\left[\sum_{h=1}^L \left(\frac{N_h}{N} \right) \sigma_h \sqrt{c_h} \right]^2}{\frac{d^2}{z^2} + \frac{1}{N} \sum_{h=1}^L \left(\frac{N_h}{N} \right) \sigma_h^2}.$$

If the fpc is ignored, this reduces to:

$$c^* = \frac{z^2}{d^2} \left[\sum_{h=1}^L \left(\frac{N_h}{N} \right) \sigma_h \sqrt{c_h} \right]^2.$$

From here, we could compute $n_h = c^* w_h$, and then ultimately n .

- Each of these allocation methods (in order as presented) gets increasingly optimal, but requires more information.

Example: Survey of TV habits

Estimate the total sample size needed to estimate the mean hours of TV watched in this particular county to within 1 hour with 95% probability.

The R code to answer this question is given below.

```
> s <- c(5.946,15.24,9.36) # vector of estimated stratum standard deviations
> Nh <- c(155,62,93)      # vector of stratum sizes
> N <- sum(Nh)            # total population size
> d <- 1
> z <- qnorm(.975)
```

Equal Allocation

=====

```
> n <- 3*sum(Nh^2*s^2)/(N^2*d^2/z^2+sum(Nh*s^2))
> n
[1] 141.3878
> n/3
[1] 47.12928
```

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

Proportional Allocation

=====

```
> n <- N*sum(Nh*s^2)/(N^2*d^2/z^2+sum(Nh*s^2))
> n
[1] 163.7988
> n*Nh/N
[1] 81.89941 32.75976 49.13965
```

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

Optimal Allocation (Equal Costs)

=====

```
> n <- sum(Nh*s)^2/(N^2*d^2/z^2+sum(Nh*s^2))
> n
[1] 141.224
> n*Nh*s/sum(Nh*s) # sample size by stratum
[1] 47.55452 48.75418 44.91527
```

$$\left(n = \frac{\left[\sum_{h=1}^L N_h \sigma_h \right]^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2} \right)$$

$$\left(n_h = n N_h \sigma_h / \sum_{k=1}^L N_k \sigma_k \right)$$

The total sample sizes required are 142, 164, and 142 for equal, proportional and optimum allocation, respectively. Equal and optimal are so close because the stratum sizes and standard deviations are inversely related, and hence effectively cancel each other.

Now, suppose it costs 2/3 as much to survey an individual in a town as in the rural area. Let the cost of surveying be $c_1 = c_2 = 2$ for strata 1 and 2 (towns A and B) and $c_3 = 3$ for stratum 3 (rural) (the actual cost doesn't matter, only the relative costs). Then the minimum total net cost (less fixed cost) is computed via:

```

> cost <- c(2,2,3)
> totalcost <- (sum(Nh*s*sqrt(cost)))^2/
                (N^2*d^2/z^2 + sum(Nh*s^2))
> totalcost    # total net cost
[1] 324.2689
> (totalcost*Nh*s/sqrt(cost))/ # sample size
   (sum(Nh*s*sqrt(cost)))      #   by stratum
[1] 50.95364 52.23905 39.29450 #   (page 123)

```

$$c^* = c - c_0 = \frac{\left[\sum_{h=1}^L N_h \sigma_h \sqrt{c_h} \right]^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2}$$

$$n_h = \frac{(c - c_0) N_h \sigma_h / \sqrt{c_h}}{\sum_{k=1}^L N_k \sigma_k \sqrt{c_k}}$$

- Note that the sample size allocated for the rural area (≈ 39) is smaller than before, since it costs more to sample units from this area.

Stratified Random Sampling Allocation Example: The following problem comes from Barrett and Nutt, *Survey Sampling in the Environmental Sciences*, COMPRESS, 1979.

Wildland managers want to estimate the total number of caribou in the Nelchina herd located in south-central Alaska by stratified random sampling. The sample unit is a 4-square mile area. A count of caribou is made on each unit selected. Based on a preliminary aerial survey, the area utilized by the herd is divided into strata, and the following rough estimates of standard deviations and costs for surveying sample units in each stratum are:

(a) Determine the total number of sample units and the allocation to each stratum by optimal allocation assuming that it is desired to estimate the total number of caribou to within 5000 caribou with 95% probability. What is the total cost of the survey (assuming no fixed overhead cost)?	Stratum (h)	N_h	σ_h	c_h
	1	400	75	6
	2	30	60	6
	3	61	600	6
	4	18	150	8
	5	70	350	8
	6	120	100	10

- Since we are estimating a population total (instead of a mean) to within d , in the formula given earlier for estimating sample size where we require optimum allocation with unequal costs, we need to replace d by d/N . This simply cancels out the N^2 in the first term in the denominator. The minimum total cost using the formula in this handout is 1945.2. The R output below was used to compute this total cost.

```
> Nh <- c(400,30,61,18,70,120)
> sh <- c(75,60,600,150,350,100)
> ch <- c(6,6,6,8,8,10)
> N <- sum(Nh)
> d <- 5000
> z <- qnorm(.975)
> z
[1] 1.959964
> min.cost <- sum(Nh*sh*sqrt(ch))^2/
              (d^2/z^2 + sum(Nh*sh^2))
> min.cost      # Formula under item 4
[1] 1945.187    # on this handout.
```

$$c^* = \frac{\left[\sum_{h=1}^L N_h \sigma_h \sqrt{c_h} \right]^2}{\frac{N^2 d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2}$$

with $N^2 d^2 \rightarrow N^2 \left(\frac{d}{N}\right)^2 = d^2$

- Using the optimum allocation formula on page 123, the sample sizes are (84.4, 5.1, 102.9, 6.6, 59.7, 26.1) for the 6 strata. However, note that the sample size for stratum 3 is larger than the stratum size; hence, we set $n_3 = N_3 = 61$. The R code to find this optimum allocation for the 6 stratum sample sizes is given on the next page.

```

> nh <- min.cost*(Nh*sh/sqrt(ch))/
      sum(Nh*sh*sqrt(ch))
> nh      # optimum allocation
[1] 84.353469  5.061208 102.911232  6.574702  59.659335  26.135966

```

$$\left(n_h = c^* w_h = c^* \cdot \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{k=1}^L N_k \sigma_k \sqrt{c_k}} \right)$$

- But, the standard deviation of the stratified estimator with sample sizes (84.4, 5.1, 61, 6.6, 59.7, 26.1), using the formula on page 119 of the text, is 3930 with the resulting margin of error being $1.96(3930) = \underline{7704}$. Thus, this allocation will give us an estimate that is only within 7704 with 95% confidence, not 5000 as was desired.

```

> nh[3] <- 61 # Replace n3 by N3.
> z*sqrt(sum(Nh*(Nh-nh)*sh^2/nh)) # Compute z*SE for this allocation (p.119).
[1] 7704.252 # Greater than the target of 5000

```

- Why did this happen? The reason for this problem is that we did not restrict n_h to be less than N_h in the derivation of the optimal allocation, which of course it must be. In the formula for the variance of $\bar{\tau}_{st}$ on page 119, namely:

$$\text{Var}(\bar{\tau}_{st}) = \sum_{h=1}^L N_h (N_h - n_h) \frac{\sigma_h^2}{n_h},$$

if $n_h > N_h$, the contribution of that stratum to the variance is negative. Hence, when we change the contribution to 0 by making $n_h = N_h$, we increase the actual variance above the target variance.

- So, we simply eliminate stratum 3 and calculate the minimum cost necessary to meet the target using the remaining strata. Stratum 3 can be eliminated because if $n_3 = N_3$, it makes no contribution to the variance (i.e.: we take a census in stratum 3). If we eliminate stratum 3, then the minimum total cost to meet the target is 1951.2 for the remaining strata. Repeating the allocation calculation, the resulting allocation is (124.0, 7.4, 61, 9.7, 87.7, 38.4) (including stratum 3's 61 units).

```

> Nh.3 <- Nh[-3] # Eliminates 3rd entry in Nh.
> Nh.3
[1] 400 30 18 70 120
> sh.3 <- sh[-3]; ch.3 <- ch[-3] # Do the same for sh and ch.

# Repeat calculation without stratum 3
> min.cost.3 <- sum(Nh.3*sh.3*sqrt(ch.3))^2/(d^2/z^2 + sum(Nh.3*sh.3^2))
> min.cost.3
[1] 1951.173 # New total cost.

```

```
> nh.3 <- min.cost.3*(Nh.3*sh.3/sqrt(ch.3))/sum(Nh.3*sh.3*sqrt(ch.3))
> nh.3
[1] 123.963066 7.437784 9.661965 87.673384 38.408551
```

nh.3 is the optimal allocation to the remaining strata given that $n_3=61$.

- Now, $n_5 > N_5$. So we set $n_5 = N_5 = 70$ and leave strata 3 and 5 out of the process. The total cost for the four remaining strata is 1456.1 and the allocation is (144.4, 8.7, 61, 11.3, 70, 44.7) (including strata 3 and 5). Rounding these values to the nearest integer gives (144, 9, 61, 11, 70, 45) which meets the criteria ($n_h < N_h$ for all h). The total cost for all strata is 2382.

```
# Set n5 = N5 and calculate z*SE
> nh.3[4] <- 70
> z*sqrt(sum(Nh.3*(Nh.3-nh.3)*sh.3^2/nh.3))
[1] 5624.964
# z*SE > 5000; eliminate strata 3 and 5 and repeat process

> Nh.35 <- Nh[-c(3,5)]      # Eliminates 3rd & 5th entry in Nh.
> sh.35 <- sh[-c(3,5)]
> ch.35 <- ch[-c(3,5)]
> min.cost.35 <- sum(Nh.35*sh.35*sqrt(ch.35))^2/
                (d^2/z^2 + sum(Nh.35*sh.35^2))
> min.cost.35
[1] 1456.104
> nh.35 <- min.cost.35*(Nh.35*sh.35/sqrt(ch.35))/
                sum(Nh.35*sh.35*sqrt(ch.35))
> nh.35
[1] 144.42716 8.66563 11.25698 44.74912
# All sample sizes are now less than strata sizes

> nh.mincost <- c(144,9,61,11,70,45) # Round optimal n's to integers.
> z*sqrt(sum(Nh*(Nh-nh.mincost)*sh^2/nh.mincost))
[1] 5000.685 # z*SE meets its target.
> sum(nh.mincost*ch) # Cost of the optimal minimum cost allocation.
[1] 2382
```

This approach (of dropping strata from subsequent cost calculations) is very heuristic; that is, it does not *guarantee* that the final plan is optimal.

(b) Determine the total number of sample units and the allocation to each stratum by optimal allocation assuming costs are equal. Then calculate the total cost of this survey for the costs given in the table (this would be relevant if you didn't know the costs beforehand and the costs only became apparent after you had done the survey).

- Assuming equal costs, we use the formula for computing sample sizes with optimum allocation and equal costs (given under item 3 earlier in this handout) to compute the total sample size required. Via the R code below, we compute $n = 282.3$.

```
> n <- sum(Nh*sh)^2/(sum(Nh*sh^2)+d^2/z^2)
> n
[1] 282.3435
```

$$\left(n = \frac{\left[\sum_{h=1}^L N_h \sigma_h \right]^2}{\frac{d^2}{z^2} + \sum_{h=1}^L N_h \sigma_h^2} \right)$$

- Using the optimum allocation formula on page 107 $\left(n_h = n \frac{N_h \sigma_h}{\sum_{k=1}^L N_k \sigma_k} \right)$ gives sample sizes of (78.7, 4.7, 96.0, 7.1, 64.3, 31.5). As in part (a), $n_3 > N_3$ so we go through the same process as above, setting $n_3 = N_3$ and recomputing the necessary n for the remaining strata.
- As above, we find that $n_5 > N_5$ at the second stage so we set $n_5 = N_5$ and eliminate stratum 5 as well. The final allocation is (133.2, 8.0, 61, 12.0, 70, 53.3) which we round to (133, 8, 61, 12, 70, 53). The total cost is 2398, only slightly more than the minimum cost allocation.

```
> nh <- n*Nh*sh/sum(Nh*sh)
> nh
[1] 78.720294  4.723218 96.038759  7.084826 64.288240 31.488118
> n.3 <- sum(Nh.3*sh.3)^2/(sum(Nh.3*sh.3^2)+d^2/z^2)
> n.3
[1] 264.6758
> nh.3 <- n.3*Nh.3*sh.3/sum(Nh.3*sh.3)
> nh.3
[1] 111.83483  6.71009 10.06513 91.33178 44.73393
> n.35 <- sum(Nh.35*sh.35)^2/(sum(Nh.35*sh.35^2)+d^2/z^2)
> n.35
[1] 206.5
> nh.35 <- n.35*Nh.35*sh.35/sum(Nh.35*sh.35)
> nh.35
[1] 133.225807  7.993548 11.990323 53.290323
```

```

> nh <- c(133,8,61,12,70,53)
> sum(nh*ch)
[1] 2398

```

(c) Determine the total number of sample units and the allocation to each stratum assuming that proportional allocation will be used. Then calculate the total cost of this survey for the costs given in the table.

- Using proportional allocation, the minimum sample size required to meet the target is 588.1 (from the formula for computing sample sizes with proportional allocation given as item 2 earlier in this handout) and the allocation, after rounding, is (337, 25, 51, 15, 59, 101). The total cost is 4080, as found using the R code below.

```

> n <- N*sum(Nh*sh^2)/(d^2/z^2+sum(Nh*sh^2))
> n
[1] 588.0636
> nh <- n*Nh/N
> nh
[1] 336.51706 25.23878 51.31885 15.14327 58.89049 100.95512
> nh <- c(337,25,51,15,59,101)
> sum(nh*ch)
[1] 4080

```

(d) Summarize your answers to (a) through (c) and discuss your results. How important is knowledge of stratum costs and variances in designing this survey?

The minimum cost and optimal equal-cost allocations are very similar and have almost identical total costs. This occurs because the costs do not vary much across strata compared to how much the strata sizes and standard deviations vary. Proportional allocation results in a cost almost twice as large as the previous two allocations. This occurs because standard deviations vary greatly across strata (by a factor of 10) and proportional allocation ignores this information. For example, it allocates over 300 observations to stratum 1 while the first two allocate fewer than 150. However, this results in little gain in precision because stratum 1's standard deviation is relatively low. In stratum 3, on the other hand, proportional allocation yields a sample size of 51 of the 61 units while the first two sample all 61 units. Though this is a small difference in sample size, it gives a great increase in variance because stratum 3's standard deviation is so high. So, in this problem, knowledge of the true stratum costs is not very important while knowledge of the true stratum standard deviations is.

The table below summarizes the calculations with the different allocation methods considered.

Allocation Type	Total	Total	Allocation					
	n	Cost	n_1	n_2	n_3	n_4	n_5	n_6
Optimal, unequal costs	340	2382	144	9	61	11	70	45
Optimal, equal costs	337	2398	133	8	61	12	70	53
Proportional	588	4080	337	25	51	15	59	101
		N_h	400	30	61	18	70	120
		σ_h	75	60	600	150	350	100
		c_h	6	6	6	8	8	10