Stat 549 Applied Sampling - Overview

The intent of this handout is to provide a brief overview to some of the topics in applied sampling which will be covered this semester.

What is sampling?

The purpose of sampling is to estimate some unknown population parameter from a subset (sample) of the population and to estimate the accuracy of the estimate. An estimate without a measure of accuracy is of limited use.

Definitions:

- Parameter - any numerical characteristic of a population.
- Sampling unit - the basic unit in the population on which we record the variable(s) of interest. We get one value of a variable from each sampling unit.
- Population - the set of all sampling units

Example 1: we want to estimate the average length of grass blades on the oval.
Parameter:
Sampling unit:
Population:

Example 2: we want to estimate the number of grass blades on the oval.
Parameter:
Sampling unit:
Population:

Throughout most of this course, we assume a finite population of size N. The infinite population case can be treated by simply letting N be very large, so the finite population case is the most general setting.

Basic Sampling Designs

1. Census - sampling the entire population. This means that we know the parameter of interest, so that the use of statistics is not necessary.
2. Simple random sample (SRS) (n = sample size) - random selection of n sampling units without replacement.
3. Systematic random sample - selecting every m^{th} member of the population. Where does the "randomness" come in here?

4. Unequal probability sample - any sampling plan where some units have a higher probability of being chosen than others. Does this introduce bias?

    - One type of unequal probability sampling is PPS sampling - sampling with Probability Proportional to Size.

5. Stratified random sampling - sampling where the population is broken into different strata, and an SRS is taken within each stratum.

    - Stratified random sampling can be thought of as a two-stage sampling plan, where the stratification is one stage, and the SRS is the second stage. In fact, any of sampling designs 1-4 above could have been used at the second stage.

6. Cluster sampling - sampling where the population is first broken into clusters, and then an SRS of *clusters* is taken. Typically, every individual within a cluster is then sampled.

    - When is cluster sampling appropriate?

7. Convenience or haphazard sampling

Two-Stage Sampling Plans

Example 3: UM faces the prospect of significant increases in tuition over the next biennium depending on what happens in the legislature. Suppose we want to sample University of Montana first-year students regarding their opinions on the tuition increases.
Population?
Sampling unit?
What are some possible parameters we might be interested in?

There are currently 9 freshman dormitories on campus. So we might take:

- Stage 1: an SRS of 4 dorms
- Stage 2: an SRS of 10 students in each of these 4 dorms

At each stage of a sample, we need to be able to identify the following:

1. What is the sampling unit?
2. What is the population?
3. What is the sampling plan?

Back to Example 3

1st Stage: Primary sampling unit?
Population?
Sampling Plan?

2nd Stage: Secondary sampling unit?
Population?
Sampling Plan?

Example 4: Suppose we want to estimate the number of people who enter the Griz Market in the UC over the 15 weeks of a semester, including weekends. Observers will monitor the entrance and count the number of people entering. For simplicity, suppose we decide that the sampling unit is a day and that we will monitor the entrance on a sample of 21 days during the semester.

- SRS: Randomly select 21 days from the 105 days of the semester.
- Stratified: Stratifying by day of the week, we randomly select 3 Sundays, 3 Mondays, etc. over the 15 week semester.

  1st Stage: Primary sampling unit?
  Population?
  Sampling Plan?
  2nd Stage: Secondary sampling unit?
  Population?
  Sampling Plan?

- Cluster: Use the 15 weeks as clusters of 7 days each. Randomly select 3 weeks and monitor the entrance during every day of each of these weeks.

1st Stage: Primary sampling unit?
Population?
Sampling Plan?
2nd Stage: Secondary sampling unit?
Population?
Sampling Plan?

- There are many other possible designs: we could do a two-stage design with an SRS at each stage: e.g., an SRS of 7 weeks with an SRS of 3 days within each week. If a day is considered too long a period to monitor and we want the sampling unit to be an hour (or a longer period), we could add a third stage where we sample hours within each day. Such designs are common when attempting to estimate visitation to recreation sites, for example. These designs can also be used as the initial stages of a design to select a sample of visitors and the parameter(s) of interest are characteristics of all visitors to a site.
- What are the relative advantages/disadvatantages of the above designs in terms of cost, convenience and accuracy? (These issues will be addressed as we discuss these designs).

Auxiliary Information

Auxiliary information generally consists of variables (other than the response variable) we can measure on the sampling units which may help in the estimation of a parameter of interest for the response variable. If the auxiliary variable (or variables) is related to the response variable, we can sometimes take advantage of the relationship to provide more precise parameter estimates for a given effort. Auxiliary variables can be utilized in either the design phase or in the estimation phase of the sampling scheme. Common uses of auxiliary information are given below.

1. Ratio or regression estimation. We use the relationship between an auxiliary variable (measured on the whole population) and the response variable (measured on a sample) in the estimation phase. Any examples?

2. Double sampling. Like ratio estimation, but we measure the auxiliary variable on an SRS of n sampling units and then measure the response variable on a subsample of these n units.

This is done generally because the response variable may be difficult or expensive to measure, whereas the auxiliary variable may be easier to measure. Any examples?

3. Stratified random sampling. With stratified random sampling, the strata themselves are the auxiliary variable.

    - Here, auxiliary information is incorporated in the design phase of the sample, so that the auxiliary variable affects how we *sample* instead of how we estimate.

4. Ranked set sampling. Here, we take groups of sampling units, visually rank the units within each group according to the size of the response variable, and then randomly choose groups where we sample the unit of rank 1 in the first group chosen, rank 2 in the second group chosen, etc.

5. Spatial sampling. Any sampling method which incorporates spatial proximity information into the sampling plan.

    - The main idea here is that since variables tend to be more alike in value the closer they are to one another spatially, optimal sampling plans purposely avoid sampling units which are spatially "close". Sampling two units close together provides redundant information if there is spatial correlation present.

6. Adaptive sampling. Sampling methods generally used in sampling rare or "difficult to find" species which tend to be spatially clustered. For example, consider sampling to determine how much of a rare species of plant there is in some area.

    - We could take an SRS of say, 1 meter squares, but it's likely that almost all of the squares will have no plants. On the other hand, when we do find a square with the plant, it's likely that adjacent squares, not scheduled for sampling, will have the plant.
    - Adaptive sampling allows you to "adapt" your sampling plan to account for what you actually see, by sampling additional squares around any units where you find the plant. Does this introduce bias?

- Note: In these final two sampling strategies (spatial, adaptive), we embrace two of the most troublesome facts of life in statistics: DEPENDENCE and BIAS!

Examples of Sampling Problems

For the following situations, identify the population, the sampling units, and the sampling plan. If it is a multistage sampling plan, identify the population, sampling unit, and sampling plan at each stage.

1. A researcher has a list of all 4-year colleges and universities in the United States.

   (a) The parameter of interest is the proportion of 4-year schools which offer a degree in education. The names of 50 schools are drawn at random from the list and the proportion of these 50 offering such a degree is computed.

   (b) As in (a), but the list is divided into three groups according to enrollment: less than 2000 students, 2000 to 10000 students, greater than 10000 students. Twenty schools are drawn at random from each group.

   (c) The groups in (b) are each subdivided into two groups: those which offer graduate degrees (in any field) and those which do not. Ten schools are drawn at random from each of these subgroups.

   (d) The parameter of interest is the average age of full-time students in all the schools. Fifty schools are drawn at random and the average age of all students at these 50 schools is computed.

   (e) As in (d), except that 100 students are chosen at random from each of the 50 schools and the average age of these students computed.

2. A researcher wishes to estimate the proportion of bare ground on a 40-acre parcel of land.

   (a) She stands in the "middle" of the parcel and randomly chooses five numbers from 1 to 360. These represent the directions, in degrees from North, of five transects through the middle point which extend to the edges of the parcel. On each transect, she uses a measuring wheel to determine how much of that transect lies in bare ground.

   (b) As in (a), except that for each transect she chooses a random distance from 0 to 5 meters from the center. At five-meter intervals along the transect starting at this point, she centers a circle with radius 0.5 meters. She does this for all the transects. She then determines the proportion of bare ground in all the 0.5-meter circles.

   (c) As in (a), except that she chooses ten points at random along each transect and centers a 0.5-meter circle at each of these points. She then determines the proportion of bare ground in all the 0.5 meter circles.

   (d) Repeat (a), (b), and (c), except that the transects are five parallel North-South lines evenly-spaced across the parcel.

3. A researcher is interested in black bears in a certain geographic region, particularly the size of the bears and the amount of time they spend in various habitats during the summer.

(a) He sets up traps at five locations scattered throughout the region. He continues trapping until ten bears have been caught. He estimates average size characteristics from these ten bears.

(b) As in (a), but he radio collars the bears. Each bear is located once each week during the summer at the same time on the same day of the week. These observations are used to estimate the proportion of time bears spend in various habitats during the summer.

(c) As in (b), but each bear is located at a randomly chosen time on a randomly chosen day during each week.

(d) As in (a), but he radio collars the first five females and the first five males he traps.

4. A bird biologist wants to describe the use of patches for foraging by two species of sparrows. He defines patches based on their discontinuity with the surrounding background. Each year over a 3-year period, 300 patches were selected from an 800x300 m study area by first randomly choosing one of 72 reference grid points on the sampling area, then randomly selecting one of eight principal compass directions, and finally, stretching a 50-meter tape from the grid point in the selected direction. All patches whose canopy area intercepted the tape (omitting those within the first 10 m because of possible trampling around the grid point) were measured. This process was repeated until at least 300 patches had been selected. Each 40-m line transect intercepted 15-20 patches.

5. A researcher is interested in the average size of the ponds in a pothole region. She takes an aerial photograph of the region and places points randomly on the photograph. The pond on or nearest each point is included in the sample. She continues until 50 ponds have been chosen.

6. A geologist is interested in the surface geology of a certain area. He divides the area up with a grid into 20 equal-sized parcels. Within each parcel, he randomly selects 5 points and obtains measurements at each of these points.

7. A fire researcher is interested in estimating the average fuel moisture in the leaves of the bushes in a small area. She randomly selects ten bushes from the area and then randomly selects two branches off of each bush. She strips all the leaves off these two branches to analyze.

8. A sociologist is interested in the sex and age makeup of Missoula bar patrons. He randomly selects five bars and visits them in random order, one on each of five consecutive Friday nights. He observes all people entering the bar from 8 to 12 pm, recording the sex and estimated age of each.

Simple Random Sampling (Chapter 2)

Simple random sampling (SRS) is a sampling design where n units are selected (without replacement) from a population of N units, such that all samples of size n are equally likely to be selected. First, though, we introduce some general terminology and sampling concepts.

Population Notation:

- Finite Population Values: $y_1, \ldots, y_N$
- Population Mean: $\mu = \dfrac{1}{N}\sum_{i=1}^{N} y_i$
- Population Total: $\tau = \sum_{i=1}^{N} y_i$
- Population Median: $M = \text{median}(y_1, \ldots, y_N)$
- Population Variance: $\sigma^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$

Sample Notation

- Sample values: $y_1, \ldots, y_n$, where n = sample size ($n \leq N$).
- Sample mean: $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
- Sample variance: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

Definitions

- An estimator $\widehat{\theta}$ of a population parameter θ is a function of the sample values (e.g., $\overline{y}$ is an estimator of μ).
- The sampling distribution of $\widehat{\theta}$ refers to the distribution of values of $\widehat{\theta}$ for all possible samples of size n from the population. The sampling distribution depends on the sampling plan being used.
- The bias of an estimator $\widehat{\theta}$ of θ is given by:

$$\text{Bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta.$$

- An estimator of θ is unbiased if its bias is 0 for all values of θ, that is, if $\text{E}(\widehat{\theta}) = \theta$.
  Note: Unbiasedness is a property of the estimator, not of the sampling plan.

- The variance of an estimator is $\text{Var}(\widehat{\theta}) = \text{E}\left[\left(\widehat{\theta} - \text{E}(\widehat{\theta})\right)^2\right]$. The variance is a measure of the precision of an estimator. Precision refers to how much the estimator varies from sample to sample, regardless of bias, and is often used to compare estimators when they are unbiased or when the bias is small.

- The mean squared error (MSE) of an estimator is:

$$\text{MSE}(\widehat{\theta}) = \text{E}\left[\left(\widehat{\theta} - \theta\right)^2\right] =$$

$$\text{(accuracy)} =$$

- The MSE incorporates both the bias and the precision of an estimator into a measure of overall accuracy and can be used to compare estimators whether they are unbiased or not.

- Many standard estimators are unbiased or nearly unbiased so the variance of estimators is the most common means of comparison.

Example 1: Consider a population of size $N = 4$, with $y_1 = 6$, $y_2 = 4$, $y_3 = 16$, & $y_4 = 10$.
Population Mean: $\mu = 36/4 = \underline{9}$
Population Variance: $\sigma^2 = \frac{1}{3}\left[(6-9)^2 + (4-9)^2 + (16-9)^2 + (10-9)^2\right] = \frac{1}{3}(84) = \underline{28}$ $(\sigma = 5.29)$.

Consider the following comparison of simple random sampling and stratified random sampling (where a random sample is taken within each stratum) for estimating μ from samples of size 2 from this population.

1. Take an SRS of size $n = 2$: Estimate μ with $\overline{y}$. How many possible samples? What is the sampling distribution of $\overline{y}$?

Sample	$\overline{y}$	Prob.

$$\text{E}(\overline{y}) = \frac{1}{6}(5 + \cdots + 13) = \frac{54}{6} = 9, \text{ so } \overline{y} \text{ is an unbiased estimator of } \mu.$$

$$\text{Var}(\overline{y}) = \frac{1}{6}(5-9)^2 + \cdots + \frac{1}{6}(13-9)^2 = \frac{42}{6} = 7.$$

$$\text{SD}(\overline{y}) = \sqrt{\text{Var}(\overline{y})} = \underline{2.65}.$$

2. Take a stratified random sample

(a) Suppose we have 2 strata: $6, 4|16, 10$ and we take a random sample of size 1 from each stratum. The stratified estimator of the population mean is $\overline{y}_s = \frac{1}{2}(y_1 + y_2)$. How many possible samples? What is the sampling distribution of $\overline{y}_s$?

Sample	$\overline{y}_s$	Prob.

$$\begin{aligned} \mathrm{E}(\overline{y}_s) &= \frac{1}{4}(11+8+10+7) = 9 \text{ (unbiased for } \mu) \\ \mathrm{Var}(\overline{y}_s) &= \frac{1}{4}\left[(11-9)^2 + \cdots + (7-9)^2\right] = 2.5 \\ \mathrm{SD}(\overline{y}_s) &= \sqrt{2.5} = \underline{1.58}. \end{aligned}$$

- Since both estimators (SRS and stratified) are unbiased, but $\overline{y}_s$ has the smaller variance, a stratified random sample is better here.

(b) Suppose instead that the 2 strata were defined as: $4, 16|6, 10$.

Sample	$\overline{y}_s$	Prob.
(4,6)	5	1/4
(4,10)	7	1/4
(16,6)	11	1/4
(16,10)	13	1/4

$\mathrm{E}(\overline{y}_s) = 9$, $\mathrm{Var}(\overline{y}_s) = 10$, and $\mathrm{SD}(\overline{y}_s) = \underline{3.16}$, which is worse than the SRS in this case. Why?

Stratified random sampling is discussed in more detail in Chapter 11, but this example illustrates that stratified sampling can be superior or inferior to simple random sampling.

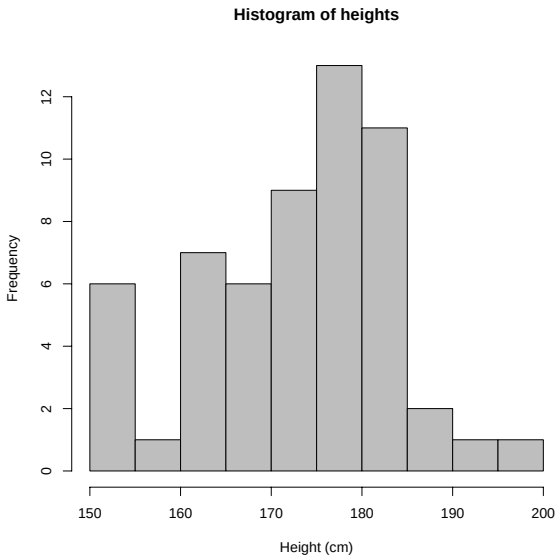
- In the first set of strata, we have low variability within strata (homogeneous within) & high variability between strata (heterogeneous between).
- In the second set of strata, we have high variability within strata (heterogeneous within) & low variability between strata (homogeneous between).
- This example illustrates the importance of choosing strata which are heterogeneous between and homogeneous within. It should also serve as a warning that stratification should not be used in sampling unless there are clear reasons for believing the defined strata are quite different from one another. In other words, don't stratify without a good reason!

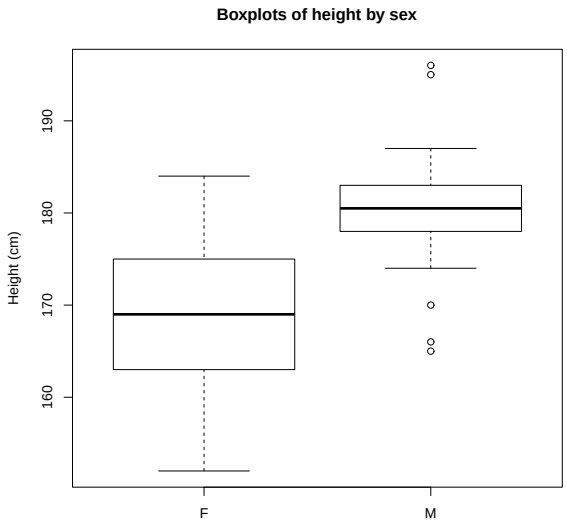
Simulation

In Example 1, with $N = 4$, we were easily able to list all the possible samples of size 2 for both SRS and stratified random sampling. This gave us the exact sampling distributions of $\overline{y}$ and $\overline{y}_s$. For larger populations, where listing all the possible samples of a given size may not be possible, we can use simulation to approximate the sampling distributions.

Example 2: The data file `measurements.csv` contains sex and height (and a couple of other variables) for 57 students (31 females, 26 males) in an introductory statistics class. Suppose we consider these students the population and we want to compare the effectiveness of a simple random sample of 10 students versus a stratified random sample of 5 males and 5 females for estimating the mean height of the whole class. The number of possible SRS's of size 10 is $\binom{57}{10} = 43,183,019,880$ and the number of possible stratified random samples is $\binom{31}{5}\binom{26}{5} = 11,176,745,580$. It's theoretically feasible, but very time-consuming, to have a computer program compute the value of an estimator for every possible sample. It also isn't really necessary since we can estimate the sampling distributions very closely by looking at 10,000 or 100,000 of these samples (randomly selected). This is easily done in R.

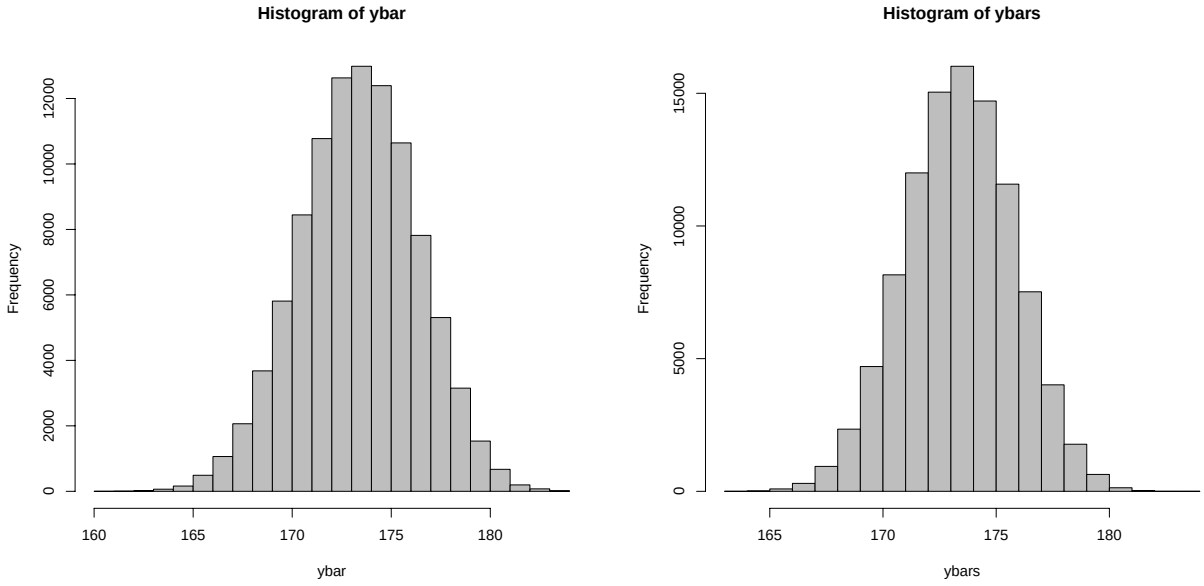
```
> m <- read.csv("measurements.csv")
> names(m)
[1] "Sex"        "FootLength" "HandSpan"   "Height"
> y <- m$Height  # to save typing, let y contain the heights
> mean(y)    # this is the population mean
[1] 173.386
> hist(y,col="gray",xlab="Height (cm)",main="")
> boxplot(Height~Sex,data=m,ylab="Height (cm)")
```






```
# Simulation for SRS of size n=10
> N <- nrow(m)   # population size
> N
[1] 57
> b <- 100000    # number of simulated samples
> ybar <- numeric(b)  # initalize vector to store sample means
> for(k in 1:b){
+   i <- sample(1:N,10)
+   ybar[k] <- mean(y[i])
+ }
> mean(ybar)
[1] 173.3812
> sd(ybar)
[1] 2.973337
> hist(ybar,col="gray")

# Simulation for stratified random sample of size 5 from each sex.
# The stratified estimator ybars is a weighted average of the group means.
> sex <- m$Sex
> table(sex)
sex
 F  M
31 26
> yF <- y[sex=="F"] # vector of female heights
```

```
> yM <- y[sex=="M"] # vector of male heights
> ybars <- numeric(b)
> for(k in 1:b){
+   i <- sample(1:31,5)  # sample from females
+   j <- sample(1:26,5)  # sample from males
+   ybars[k] <- (31/57)*mean(yF[i])+(26/57)*mean(yM[j])
+ }
> mean(ybars)
[1] 173.3786
> sd(ybars)
[1] 2.417199
> hist(ybars,col="gray")
```



Note that the estimated means of both sampling distributions, `mean(ybar)` and `mean(ybars)`, are both very close to the population mean. This is because both $\overline{y}$ and $\overline{y}_s$ are unbiased estimators, as we will see. Therefore, we can compare the estimators by comparing the estimated standard deviations of the sampling distributions. We see that the stratified estimator has a smaller standard deviation than the SRS estimator (2.417 versus 2.973). As we shall also see, the standard deviations of $\overline{y}$ and $\overline{y}_s$ can be derived theoretically so the simulation is not necessary for this purpose (though this is not true for all estimators). What cannot be derived theoretically are the sampling distributions of the estimators which are displayed (as estimated by the simulations) in the histograms. We need to know something about the shape of the sampling distribution of an estimator in order to construct a confidence interval for the parameter. In this example, the distributions of $\overline{y}$ and $\overline{y}_s$ both look roughly normal, a result that follows from the Central Limit Theorem (discussed in Chapter 3).

Back to Simple Random Sampling (N = population size, n = sample size)

What are the properties of the estimators of the population mean and total for SRS's?

Estimating the Mean

- The population mean μ is estimated by $\overline{y}$. What are its expected value $\mathrm{E}(\overline{y})$ and variance $\mathrm{Var}(\overline{y})$?

- The derivation of these results for sampling without replacement from a finite population is somewhat tedious and is covered in Section 2.6 of Thompson.
- If $N >> n$, then the fpc (finite population correction) $= (N-n)/N \approx 1$ and is often omitted. In this case, we revert to the usual formula for the variance of the sample mean (based on sampling with replacement or from an infinite population): $\mathrm{Var}(\overline{y}) = \sigma^2/n$.
- If $N = n$ (as in a census), then $\mathrm{Var}(\overline{y}) =$
- The standard deviation of the sample mean is given by: $\mathrm{SD}(\overline{y}) = \sqrt{\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}}$.
  What is the problem with this as an estimator of variability?

- The population variance σ^2 is unknown, so we need to estimate it. This results in the estimated standard deviation of $\overline{y}$, more commonly referred to as the standard error of $\overline{y}$, given by:

$$\mathrm{SE}(\overline{y}) = \widehat{\mathrm{SD}}(\overline{y}) = \sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Estimating the Total

Estimate $\tau = \sum_{i=1}^{N} y_i$ with $\boxed{\widehat{\tau} = \frac{N}{n}\sum_{i=1}^{n} y_i = N\overline{y}}$.

$\mathrm{E}(\widehat{\tau}) \quad =$

$\mathrm{Var}(\widehat{\tau}) \quad =$

Design-Based vs. Model-Based Sampling: In the derivations above for SRS's and in most of the procedures discussed in this course, we take a design-based or fixed-population approach to sampling. In other words, *no distributional assumptions are placed on the population* in finding the form of the estimators and their variances. In the model-based approach to sampling, the "population" $y_1, \ldots, y_N$ are considered to be random variables, that is, one possible realization of all possible realizations that could have occurred under some model for the population.

For example, suppose I flip a (possibly biased) coin 100 times and record the result of each toss on a slip of paper and place the slips in a box. I ask you, who don't know the results, to estimate the proportion of the 100 flips that came out heads; call this proportion θ. You are allowed to take a sample of slips from which to make your estimate. This is a design-based approach – the 100 flips is the population and the parameter of interest is θ, the proportion of heads in these 100 flips, not the long-run proportion of heads if you kept flipping the coin. If you base your estimate of θ on a random sample of slips of paper, the only uncertainty in your estimate is due to sampling variability. If you could do a census of all100 flips, there there would be no uncertainty in your estimate. If you use the the sample proportion of heads as your estimator, then this estimator will be unbiased for θ and you will be able to generate an unbiased estimate of its variance. These properties depend only on the sample design – they don't depend on whether I really flipped a coin 100 times or simply wrote down 100 0's and 1's in any combination or order I wanted.

In the model based approach, we view these 100 flips as 100 realizations from a random process, and the parameter we are really interested in is not the proportion of heads in these 100 flips, but the true long-term probability of heads, say p. Our model might then be that the 100 flips are 100 independent Bernoulli trials with probability of heads p. Even if I observed all 100 flips, I would still not know p. The difference between the two approaches doesn't really make a difference in how I would estimate θ or p: I would use the proportion of heads in whatever sample I observed. It would make a difference in the standard error of the estimate as the population

is considered infinite in the model based approach. It also has implications for how I sample. If my model is correct – that these 100 flips are 100 independent Bernoulli trials – then it doesn't matter if I observe a random sample of flips or not. If I get to observe 10 flips, then the first 10 are just as good as any other set of 10. I simply view these 10 flips as 10 independent Bernoulli random variables; the randomness (and unbiasedness) in my estimator comes from the model, not the sampling scheme. However, if my model is not correct and these are not independent Bernoulli trials with constant probability of heads (for example, there is dependence between trials, or the probability of heads decreases as I go along), then my estimator based on the first 10 flips might not be very good (biased, large variance) and I won't realize it. In fact, the idea of a single parameter p might not even make sense if it changes during the experiment.

Design-based approaches are valid regardless of how the data were generated, but the scope of inference is confined to the fixed population. The model-based approach allows for inference to a larger population or model, but depends crucially on the appropriateness of the model.

Confidence Intervals (Chapter 3)

An estimate of a parameter is not very useful unless we also we report some measure of the accuracy of the estimate. Usually, this is reported as a standard error of the estimate and/or a confidence interval (CI) for the parameter. The most common way to construct a confidence interval for a parameter θ is as:

$$\text{Estimator} \pm \left(\begin{array}{c} \text{Critical} \\ \text{Value} \end{array} \right) \left(\begin{array}{c} \text{SE of} \\ \text{Estimator} \end{array} \right) = \widehat{\theta} \pm (\text{critical value}) \cdot \text{SE}(\widehat{\theta})$$

where the critical value is usually from either a standard normal distribution (z value) or a t distribution.

- A confidence interval for the population mean μ from an SRS is

$$\overline{y} \pm t \cdot \text{SE}(\overline{y}) \;=\; \overline{y} \pm t \sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}}$$

- A confidence interval for the population total τ from an SRS is

$$\widehat{\tau} \pm t \cdot \text{SE}(\widehat{\tau}) \;=\; N\overline{y} \pm t \sqrt{N(N-n)\frac{s^2}{n}}$$

- Note that the CI for τ can be obtained by simply multiplying the endpoints of the CI for μ by N. Why does this make sense?

Why are these intervals t-based and not standard normal (z)-based?

What assumptions are we are making in the use of these CI's?

1.

2.

If the first of these assumptions is violated, we can appeal to the Central Limit Theorem (CLT) to obtain an approximate $(1-\alpha) \times 100\%$ CI.

Usual Central Limit Theorem (CLT)
If $y_1, \ldots, y_n$ are i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$, then the distribution of $\dfrac{\overline{y} - \mu}{\sigma/\sqrt{n}}$ approaches a N(0,1) distribution as n gets large. Thus, for large n, $\overline{y} \dot{\sim} N(\mu, \sigma^2/n)$.

- This version of the CLT is for an infinite population or sampling with replacement. The result still holds if σ is estimated by the sample standard deviation s (or any consistent estimator of σ).

Finite Population CLT

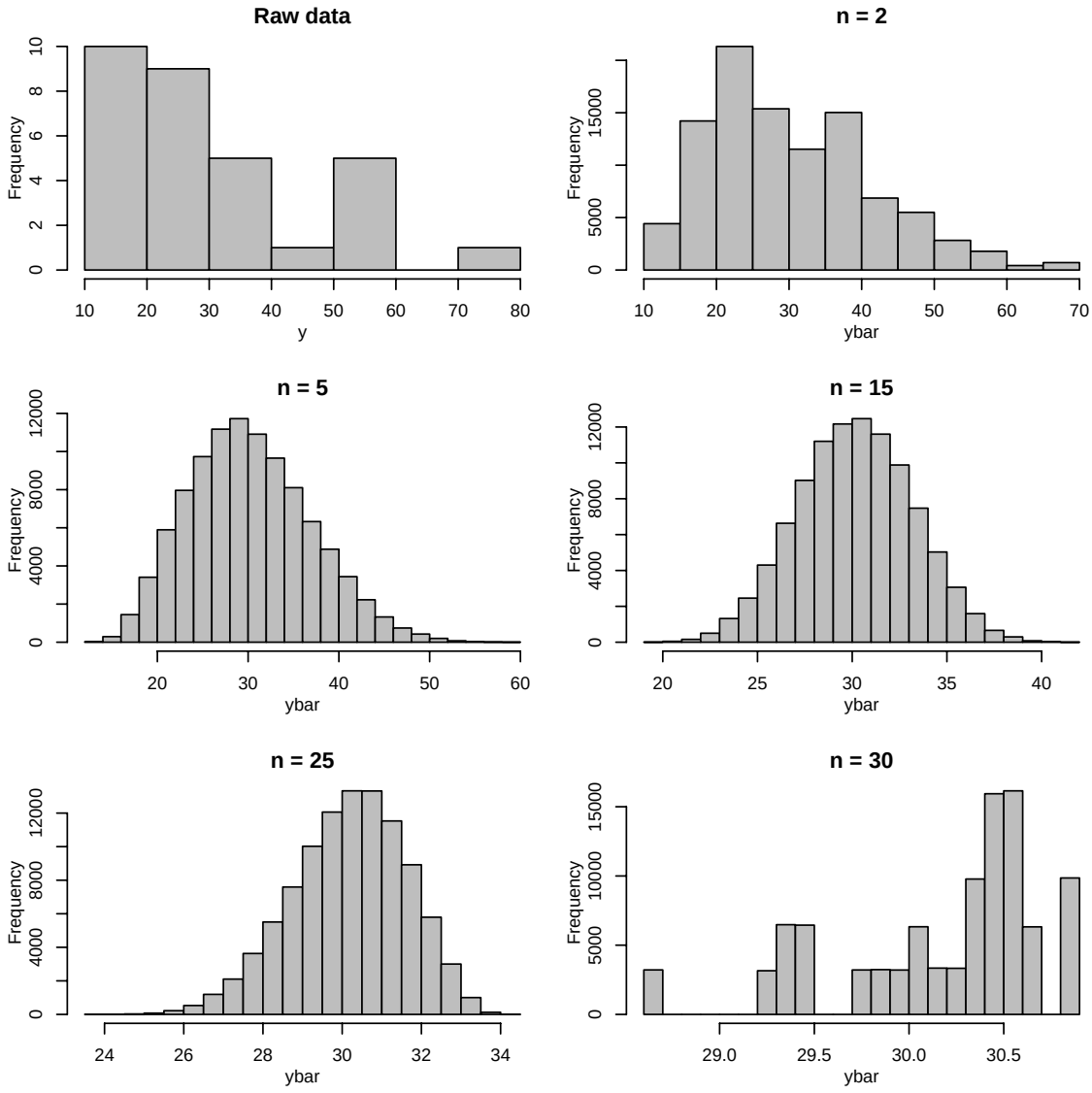
If $y_1, \ldots, y_n$ are an SRS without replacement from a finite population of size N, then

$$\overline{y} \dot{\sim} N\left(\mu, \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}\right) \quad \text{when both } n \text{ and } N-n \text{ are large.}$$

Note that the finite population CLT requires that both n and $N-n$ be large. This is demonstrated in an example on pp. 45-46 of Thompson (not in 2nd ed.) which is replicated here.

Example 1: The data set `trees` is included in R (the command `?trees` gives a description of the data set). The variable of interest is the volume of each of 31 cherry trees. Assume that this is the population and examine the sampling distribution of $\overline{y}$ for sample sizes 2, 5, 15, 25, and 30. Thompson gives the R scripts for carrying out the simulations but does not show the results. The script below draws 100,000 simulated samples (SRS's without replacement) of each size from this population. The histograms of the resulting sample means are given below. The population distribution of the 31 volumes is shown in the first histogram in the panel and is skewed to the right. The sampling distribution of $\overline{y}$ becomes more and more normal as n increase from 2 to 15. However, as n gets closer to the population size, the sampling distribution becomes less normal.

```
> y <- trees$Volume
> # the following command sets graphing parameters for a 3x2 array of plots
> par(mfrow=c(3,2),mar=c(4,3,2,1)+.1,mgp=c(2,1,0))
> hist(y,col="gray",main="Raw data")   # histogram of raw data
> N <- length(y)
> b <- 100000
> for(n in c(2,5,15,25,30)){
+   ybar <- numeric(b)  # initalize vector to store sample means
+     for(k in 1:b){
+       i <- sample(1:N,n)
+       ybar[k] <- mean(y[i])
+     }
+   hist(ybar,col="gray",main=paste("n =",n))
+ }
```



Here's an example to illustrate the calculation of a confidence interval for a population total.

Example 2: Suppose a study is undertaken to estimate the total number of pellet groups for white-tailed deer in a 20-acre field. Sampling was done along 10 randomly located belt transects of dimensions 3'x50', where the number of pellet groups in each belt was counted.

- What is the sampling unit here? What's the variable?

- Suppose the following summary statistics were obtained: $\overline{y} = 5.55$ groups, $s^2 = 14.06$, $s =$

3.75 groups. Let

$$\begin{aligned}\mu &= \text{the mean number of pellet groups per 3'x50' transect for the whole 20 acres}\\ \tau &= \text{the total number of pellet groups in the 20 acres (1 acre = 43560 square feet).}\end{aligned}$$

- Find an estimate and confidence interval for the total number of pellet groups. Computing:

$$\begin{aligned}\overline{y} &= 5.55 \text{ per 150 square feet} \implies \text{there are } 5.55/150 \text{ pellet groups per square foot.}\\ \implies \widehat{\tau} &= \underbrace{\frac{5.55}{150}}_{\begin{pmatrix}\#\text{ groups per}\\ \text{square foot}\end{pmatrix}} \cdot \underbrace{43560}_{\begin{pmatrix}\#\text{square feet}\\ \text{per acre}\end{pmatrix}} \cdot \underbrace{20}_{\#\text{ acres}} = \underline{32234.4 \text{ groups}},\\ &\text{where } N = \frac{43560 \cdot 20}{150} = \underline{5808} \text{ (150 square foot areas in the 20-acre field).}\end{aligned}$$

- Note: This is not an infinite population because the transects have area; there are, however, an infinite number of possible transects.

- The standard error of the sample mean number of groups per transect is:

$$\text{SE}(\overline{y}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}} = \sqrt{\left(1 - \frac{10}{5808}\right)\frac{14.06}{10}} = \underline{1.185}.$$

- This gives 95% confidence intervals for the mean and total as:

$$\begin{aligned}&95\%\text{ CI for } \mu: \quad \overline{y} \pm t_{.025}(9) \cdot \text{SE}(\overline{y}) = 5.55 \pm 2.262(1.185) = \underline{(2.87, 8.23)}.\\ &95\%\text{ CI for } \tau: \quad (2.87 \cdot 5808, 8.23 \cdot 5808) = \underline{(16655, 47813)}.\end{aligned}$$

- Note: we could have calculated the CI for τ directly using

$$\begin{aligned}\widehat{\tau} \pm t \cdot \text{SE}(\widehat{\tau}) &= 32234.4 \pm 2.262 \cdot \sqrt{N(N-n)s^2/n} = 32234.4 \pm 2.262(6880.9)\\ &= 32234.4 \pm 15564.6 = (16670, 47798)\end{aligned}$$

- Note on rounding: notice that the two calculations for the confidence interval for τ gave slightly different answers. This is because of rounding in the intermediate calculations. In general, we should not do such calculations by hand, but should use a program like R and only round the final answer. In this example, we also have to rely on the reported mean and standard deviation, which are probably rounded. If possible, we should use the unrounded values from a direct calculation on the raw data.

  Here is the calculation in R.

```
# Example 2
> n <- 10
> ybar <- 5.55 # use mean(y) if raw data are in vector y
> s <- 3.75  # use sd(y) if you have the raw data in y
> N <- 43560*20/150
> N
[1] 5808
> tauhat <- N*ybar
> tauhat
[1] 32234.4
> fpc <- 1-n/N
> SE.ybar <- sqrt(fpc*s^2/n)
> SE.ybar
[1] 1.184833
> ybar - qt(.975,n-1)*SE.ybar
[1] 2.869722
> ybar + qt(.975,n-1)*SE.ybar
[1] 8.230278
> tauhat - qt(.975,n-1)*N*SE.ybar
[1] 16667.35
> tauhat + qt(.975,n-1)*N*SE.ybar
[1] 47801.45
```

The 95% confidence interval for τ is about 16668 to 47801 pellet groups. Use your judgment in rounding the final answer. There is no need to report extra decimal places which have no practical significance. Even the confidence interval here could probably be reported as "about 17 thousand to 48 thousand pellet groups" without any loss in practical meaning.

Other methods for calculating standard errors and confidence intervals

It's not always possible to derive the theoretical variance of an estimator and/or the sampling distribution may not be normal. In a later chapter, we will look at two methods for dealing with these issues: linearization and bootstrapping.

Sample Size (Chapter 4)

Up to now, we have assumed that the sample size n was known, and have studied properties of various resulting estimators of the population mean or total. Taking a step back, we now consider the more realistic question from a design point of view, namely: How large a sample do we need to attain some desired accuracy for the parameters we wish to estimate?

- One way this is considered is to specify a maximum allowable difference d between the estimate and the true value of the parameter, which is exceeded with some small probability α.
- In mathematical terms, the goal is to find the smallest sample size n which satisfies:

$$\mathrm{P}(|\widehat{\theta} - \theta| > d) < \alpha,$$

for some specified d and α, where θ and $\widehat{\theta}$ denote the population parameter and corresponding estimator respectively.

Sample Size Required to Estimate the Population Mean

Consider the estimation of the population mean. Here, we want:

$$\mathrm{P}(|\overline{y} - \mu| > d) < \alpha.$$

Under simple random sampling, the sampling distribution of $\overline{y}$ is at least approximately normal for large n whether sampling from an infinite or finite population by either the regular Central Limit Theorem or the finite population version (it's exactly normal if sampling from an infinite normal population); that is $\dfrac{\overline{y} - \mu}{\sqrt{\mathrm{Var}(\overline{y})}} \dot{\sim} N(0,1)$.

- Let z be the upper $\alpha/2$ quantile from the standard normal distribution (the upper $\alpha/2$ quantile means the same thing as the $1-\alpha/2$ quantile) . Then, using the approximate normality of $\overline{y}$:

$$\mathrm{P}\left[\left|\frac{\overline{y} - \mu}{\sqrt{\mathrm{Var}(\overline{y})}}\right| > z\right] = \alpha \implies \mathrm{P}\left[\left|\frac{\overline{y} - \mu}{\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}}\right| > z\right] = \alpha$$

$$\implies \mathrm{P}\left[|\overline{y} - \mu| > z\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}\ \right] = \alpha,$$

so that $d = z\sqrt{\dfrac{N-n}{N} \cdot \dfrac{\sigma^2}{n}}$. Solving for n gives:

$$\boxed{n = \frac{1}{\frac{1}{N} + \frac{d^2}{\sigma^2 z^2}} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}}, \quad \text{where } n_0 = \frac{\sigma^2 z^2}{d^2}.$$

- Note that if N is large, the $1/N$ term in the denominator can be ignored, and this formula reduces to: $\boxed{n = n_0}$, which is the sample size requirement calculated in standard statistics textbooks for a given d & α.
- Problem: we don't know σ so we need a preliminary guess. How?

Sample Size Required to Estimate the Population Total: A similar result can be obtained for specifying the accuracy of the estimate for the population *total*, where if:

d = the maximum allowable difference between the population total and its estimate,

α = the probability this difference is larger than d,

then the sample size required to satisfy: $P(|\widehat{\tau} - \tau| > d) < \alpha$ is given by:

$$\boxed{n = \frac{1}{\left(\frac{1}{N} + \frac{d^2}{N^2\sigma^2 z^2}\right)} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}}, \quad \text{where } n_0 = \frac{N^2\sigma^2 z^2}{d^2}.$$

Example 1: Returning to the deer pellet example on page 19 of the notes, suppose that the sample with 10 transects was merely a pilot study to gain information about the variability in the number of pellet groups. In this pilot study, the following summary statistics were calculated:

$$\overline{y} = 5.55 \text{ pellet groups/150 sq.ft}, \quad s^2 = 14.06, \quad N = 5808.$$

Suppose we want to estimate τ (the total number of pellet groups in the 20-acre area) to within 5000, with probability 0.95 ($\alpha = 0.05$). How large a sample is required?

- Since the sample variance from the pilot study, s^2, was a "guess" for the population variance σ^2, we might add something to the sample size determined to be conservative.

Specifying the Relative Accuracy: Instead of specifying a desired difference d as above, the sample size determination problem can equivalently be stated in terms of the relative accuracy with which we would like to estimate some parameter.

- Suppose we want to estimate the population mean to within 10% ($r = 0.10$) with probability 0.95. We want:

$$d = r\mu \implies n_0 = \frac{\sigma^2 z^2}{r^2 \mu^2}.$$

Note that there are two unknown parameters here: σ^2 and μ. Let $\gamma = \sigma/\mu$ (coefficient of variation). Then n_0 can be rewritten as:

$$n_0 = \frac{\sigma^2 z^2}{r^2 \mu^2} = \frac{z^2 \gamma^2}{r^2}.$$

Writing n_0 in this fashion leaves only *one* unknown parameter in computing the sample size, namely γ. Hence, this latter formula can be used in situations where the coefficient of variation can be specified more easily than the mean and variance individually.

- Suppose we want to estimate the population total to within 10% ($r = 0.10$) with probability 0.95. We want:

$$d = r\tau \implies n_0 = \frac{N^2 \sigma^2 z^2}{r^2 \tau^2}.$$

Example 1 (continued): If we had wanted to estimate the population mean number of pellet groups per 150 square feet to within 10% of the mean, we compute:

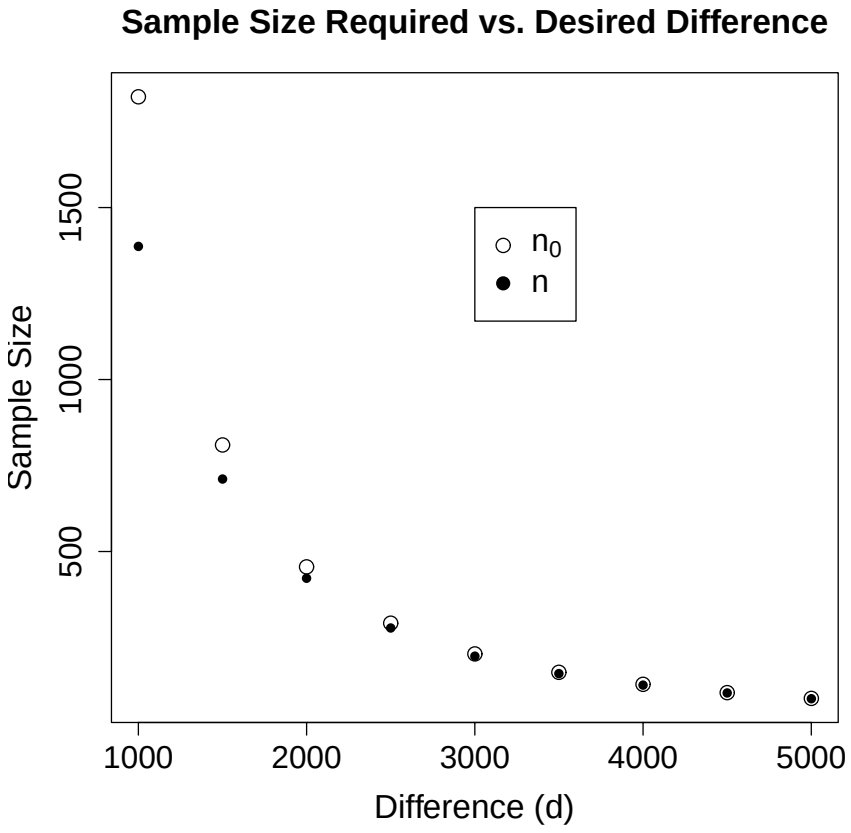
$$n_0 = \frac{\sigma^2 z^2}{r^2 \mu^2} = \frac{(14.06)(1.96)^2}{(.10)^2 (5.55)^2} = \underline{175.35 \text{ transects}}.$$

- With a finite population correction (fpc), $n = 170.2$ transects.

The following R script calculates the required sample size (both with and without the fpc) for various values of d. Note that the fpc doesn't make much difference unless the sample size is up near 10% of the population size.

```
# This is an R script to calculate the sample size needed to
# estimate the total number of deer pellets in a 20-acre field.
# Here, for different choices of the desired detectable
# difference d, the sample size n0 (without the fpc) and n (with
# the fpc) are computed and plotted against the d values.

 N <- 5808                          # Population size.
 s2 <- 14.06                        # Estimated variance.
 z <- qnorm(.975)                   # Standard normal quantile.
 d <- c(1000,1500,2000,2500,3000,# The next two commands are
       3500,4000,4500,5000)         # alternative ways to do the same thing;
 d <- seq(1000,5000,500)            # the "seq" command is very useful here.
 n0 <- N^2*s2*z^2/d^2               # Computes n0, the sample size
                                    # without the fpc.
```

```
n <- 1/(1/n0 + 1/N)             # Computes n, the sample size with the fpc.
plot(d,n0,xlab="Difference (d)",# Plots the sample sizes (w/o fpc) versus
 ylab="Sample size",pch=1,      #   the differences (d) with axis labels,
 cex=1.5)                       #   and plotting character 1 (open circle).
                                #
points(d,n,pch=16)               # Plots the sample sizes (w/ fpc) versus
                                #   the differences (d) in overlay, with
                                #   plotting character 16 (filled circle).
legend(3000,1500,c(expression("n"[0]),"n"), # Puts a legend on the plot at
 pch=c(1,16),cex=1.5)           # (3000,1500) using the plotting characters 1 & 16.
title("Sample Size Required vs. Desired Difference")  # Puts a title on the plot.
```



Estimating Proportions (Chapter 5)

Let $y_i = \left\{ \begin{array}{cc} 1 & \text{if the } i^{th} \text{ unit has some characteristic} \\ 0 & \text{otherwise} \end{array} \right\}$.

Define:

$$\begin{aligned} p &= \frac{1}{N}\sum_{i=1}^{N} y_i = \text{the population proportion of units with the characteristic,} \\ \widehat{p} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \text{the sample proportion,} \\ \sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{N}{N-1}p\,(1-p) = \text{the population variance,} \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{p})^2 = \frac{n}{n-1}\widehat{p}\,(1-\widehat{p}) = \text{the sample variance.} \end{aligned}$$

- Note that p and $\widehat{p}$ are defined exactly as μ and $\overline{y}$ were for these y_i's. So, defined as above, proportions can be represented as means.

The variance and estimated variance of $\widehat{p}$ follow immediately then from the forms of $\text{Var}(\overline{y})$ and $\widehat{\text{Var}}(\overline{y})$ given earlier in the handout:

$$\begin{aligned} \underline{\text{Var}(\widehat{p})} &= \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} = \left(\frac{N-n}{N}\right)\frac{N}{n(N-1)}p\,(1-p) = \underline{\left(\frac{N-n}{N-1}\right)\frac{p\,(1-p)}{n}}, \\ \underline{\widehat{\text{Var}}(\widehat{p})} &= \left(\frac{N-n}{N}\right)\frac{s^2}{n} = \underline{\left(\frac{N-n}{N}\right)\frac{\widehat{p}\,(1-\widehat{p})}{n-1}}. \end{aligned}$$

This estimated variance allows us to construct a confidence interval for p of the form:

$$\widehat{p} \pm z\sqrt{\frac{N-n}{N}\left(\frac{\widehat{p}\,(1-\widehat{p})}{n-1}\right)}.$$

- Is it valid to use such an interval? When?

- Exact confidence limits based on the hypergeometric distribution are given in Section 5.2 of Thompson.

Sample Size Required to Estimate a Population Proportion: To obtain an estimator $\widehat{p}$ within d of the population proportion p with probability $1-\alpha$, the sample size required is:

$$n = \frac{1}{\frac{N-1}{Nn_0} + \frac{1}{N}} \approx \frac{1}{\frac{1}{n_0} + \frac{1}{N}}, \quad \text{where: } n_0 = \frac{z^2 p(1-p)}{d^2}.$$

- Note that these formulas have the same basic form as those for the population mean. As with the mean, if N is sufficiently large, the fpc can be ignored, and n_0 is the desired sample size.
- Just as the analogous computation for the mean required a "guess" of the standard deviation, here we require a "guess" of the population proportion. If no such estimate is available, we could be conservative by setting p to the value which maximizes n. What value of p is that?

- Thompson also provides a section on determining the sample sizes necessary for estimating several proportions simultaneously (Sec. 5.4).

Unequal Probability Sampling (Chapter 6)

Unequal probability sampling is when the units in the population do not all have the same probability of being selected. This handout introduces the Hansen-Hurwitz (H-H) and Horvitz-Thompson (H-T) estimators, examines the properties of both types of estimators for the population total and mean, and compares the two estimators by way of an example.

Sampling with Replacement: The Hansen-Hurwitz (H-H) Estimator (Section 6.1)

Suppose a sample of size n is selected randomly with replacement from a population but that on each draw, unit i has probability p_i of being selected, where $\sum_{i=1}^{N} p_i = 1$. The draw-by-draw probability p_i is called the selection probability for the i^{th} unit. Let y_i be the response variable measured on each unit selected. Note that if a unit is selected more than once, it is used as many times as it is selected. An unbiased estimator of the population total $\tau = \sum_{i=1}^{N} y_i$ is given by:

$$\boxed{\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i}}.$$

An unbiased estimator of the population mean is $\widehat{\mu}_p = (1/N)\widehat{\tau}_p$.

- Dividing by p_i gives higher *weight* to units less likely to be selected.
- What happens to this estimator if $p_i = 1/N,\ i = 1, \ldots, N$, so that each unit has an equal chance of selection?

Example 1: Consider a population of size $N = 3$, with values and corresponding selection probabilities given in the first two columns of the table to the right. Note that the true population total is $\tau = 14$. Consider taking a sample of size 1. The H-H estimates of the total for each of the 3 values (samples) are given in the third column of the table.

Values	Probabilities	$\widehat{\tau}_p$
$y_1 = 3$	$p_1 = .2$	15
$y_2 = 2$	$p_2 = .5$	4
$y_3 = 9$	$p_3 = .3$	30

The expected value of τ_p is: $\mathrm{E}(\widehat{\tau}_p) = .2(15) + .5(4) + .3(30) = 14 = \tau$.

- So, in $\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i}$, each $\frac{y_i}{p_i}$ is unbiased for τ.

Why would you want to select units with unequal probabilities?

- It may be the most convenient way to sample. Recall the example of taking a sample of ponds by selecting random points on a map. If a point lands in a pond then that pond is selected for the sample. It would require a lot more effort to enumerate all the ponds so that an SRS could be selected. See also the farm example below.
- If the response variable is positively correlated with the selection probability, then the Hansen-Hurwitz estimator can have lower variance than the estimator based on an SRS.

Properties of the Hansen-Hurwitz Estimator

Note that since the sampling is with replacement, the y_i's (and therefore the y_i/p_i) are i.i.d. variables.

$$\mathrm{E}(\widehat{\tau}_p) \quad =$$

$$\mathrm{Var}(\widehat{\tau}_p) \quad = \quad \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right] = \left(\frac{1}{n}\right)^2 n\mathrm{Var}(y_1/p_1) = \underline{\frac{1}{n}\sum_{j=1}^{N} p_j\left(\frac{y_j}{p_j}-\tau\right)^2},$$

where τ is unknown. An unbiased estimate of the variance is:

$$\widehat{\mathrm{Var}}(\widehat{\tau}_p) \quad = \quad \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\widehat{\tau}_p\right)^2$$

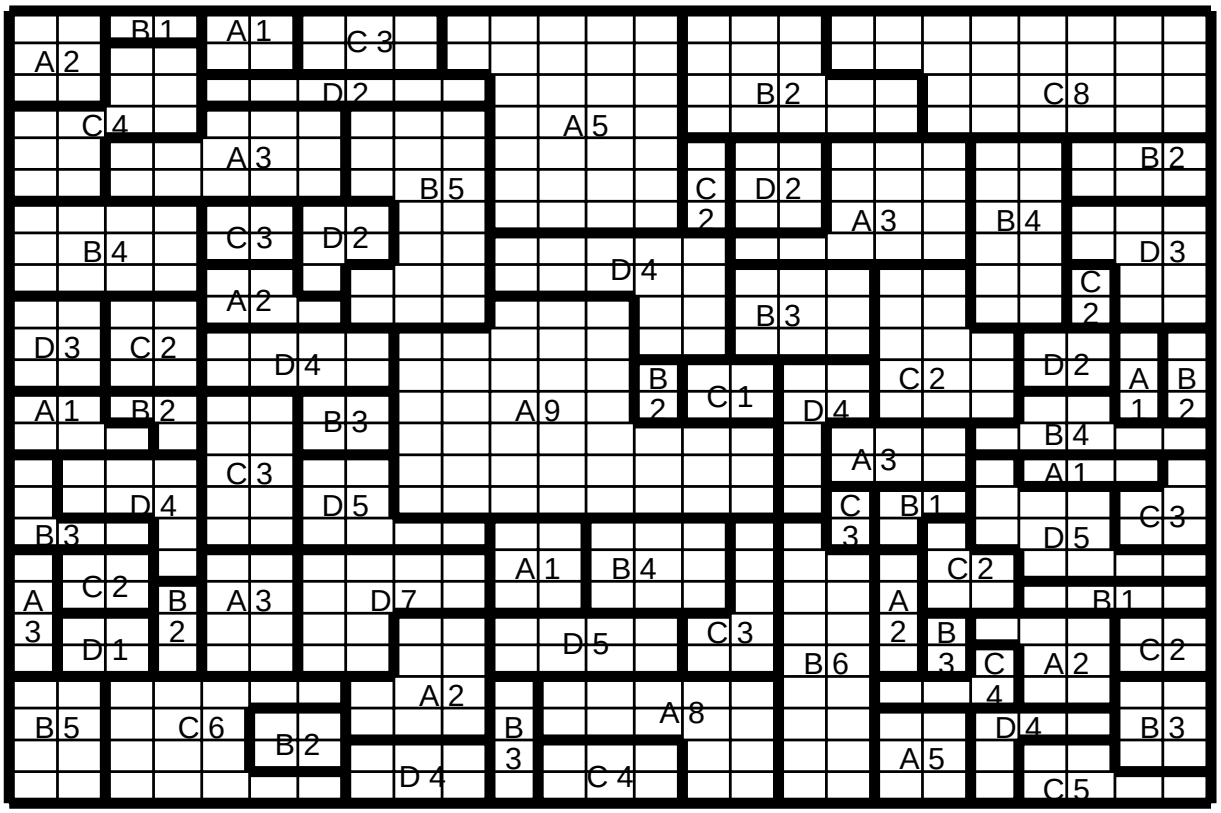
Note that the properties of $\widehat{\mu}_p = (1/N)\widehat{\tau}_p$ follow easily:
$\mathrm{E}(\widehat{\mu}_p) = (1/N)\tau = \mu$ (unbiased), $\mathrm{Var}(\widehat{\mu}_p) = (1/N)^2\mathrm{Var}(\widehat{\tau}_p)$, $\widehat{\mathrm{Var}}(\widehat{\mu}_p) = (1/N)^2\widehat{\mathrm{Var}}(\widehat{\tau}_p)$.

Notes on the Hansen-Hurwitz Estimator

1. We only need p_i for the units in the sample (not the whole population).
2. We need not know N in order to estimate τ.
3. If we let $y_i = 1, i = 1, \ldots, N$, then $\tau = N$ and $\widehat{\tau}_p = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{1}{p_i} = \widehat{N}$ is an estimator of N.
4. If there is low variability between the values of y_j/p_j, then the H-H estimator will have low variance, with the extreme case being when y_j and p_j are exactly proportional to each other. On the other hand, the H-H estimator will have high variance when there is high variability among the values of y_j/p_j.

Example 2: Consider a population of farms of varying sizes and shapes on a 25x25 grid, as given in the figure below. Let x_i = the area of farm i and $A = 625$, the total area of all the farms. One way to select a farm at random from this region (particularly if we don't have a list of the farms) is to randomly select a single square and choose the farm it belongs to. The probability that farm i is selected is $p_i = \dfrac{x_i}{A} = \dfrac{x_i}{625}$.

Farms Map: The letter for each farm represents the "type" of farm (A, B, C or D). The number for each farm represents the number of workers on the farm. The number of farms is $N = 79$ and the total number of workers is $\tau = 249$.



Let y_i = the response variable of interest.

- If $y_i = x_i$, then $\tau = \sum_{i=1}^{N} y_i$ = the total area of all farms. In this example, the total area is known to be $A = 625$, so that $y_i = x_i$ is uninteresting here.
- If $y_i = 1, i = 1, \ldots, N$, then $\tau = \sum_{i=1}^{N} y_i$= the total number of farms (which is more interesting).
- The response variable y_i might also be something like the number of workers for farm i, or the income for farm i, or a 0/1 indicator for a farm of type A.

Consider taking a sample of 5 farms with replacement with probability-proportional-to-size (PPS) and computing:

(i) The estimated number of workers. (ii) The estimated number of farms.

How do we take a random sample of pixels?

Using the R command `sample(1:25,10,replace=T)` gives a vector of 10 random integers from 1 to 25 which we can use as 5 successive pairs of random (x, y) coordinates. The results for one sample are summarized in the table below.

Coordinates	Farm Data	$p_i = \frac{x_i}{A} = \frac{\text{Size of Farm}}{\text{Total Area}}$
8,19	D2	5/625
19,25	C8	28/625
21,21	B4	12/625
15,4	A8	14/625
7,20	A3	13/625

An estimate of the total number of workers (where y_i = the # of workers for farm i) is:

$$\widehat{\tau}_p = \frac{1}{5}\left[\frac{2}{5/625} + \frac{8}{28/625} + \frac{4}{12/625} + \frac{8}{14/625} + \frac{3}{13/625}\right] = \underline{227.66 \text{ workers}}.$$

An estimate of the total number of farms (where $y_i = 1$ for all i) is:

$$\widehat{\tau}_p = \frac{1}{5}\left[\frac{1}{5/625} + \frac{1}{28/625} + \frac{1}{12/625} + \frac{1}{14/625} + \frac{1}{13/625}\right] = \underline{58.42 \text{ farms}}.$$

Class Results	Truth
# Workers:	$\tau = 249$ workers
# Farms:	$N = 79$ farms

For my sample, the estimated variance of the number of workers is:

$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\tau}_p) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{\tau}_p\right)^2 = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\widehat{\tau}_p^2\right] \quad \text{(Computing Formula)} \\
&= \frac{1}{5(5-1)}\left[\left(\frac{2}{5/625} - 227.66\right)^2 + \cdots + \left(\frac{3}{13/625} - 227.66\right)^2\right] = 1350.427 \\
&\Longrightarrow \underline{\widehat{\text{SD}}(\widehat{\tau}_p) = 36.75}.
\end{aligned}$$

- If the $\frac{y_i}{p_i}$ are approximately constant, then $\widehat{\text{Var}}(\widehat{\tau}_p)$ will be small. When might this be true?

How do we estimate:

1. The average number of workers per farm μ?

2. The average size of all the farms, μ? Recall that $p_i = \frac{x_i}{A}$ where x_i = the farm size and A = the total area.

- This implies that we do not actually need the total area A.

Summary of results for PPS sampling

Here are some general results for PPS sampling with replacement. Some of these were illustrated with the farm example. Let:

$$\begin{aligned} N &= \text{the population size,} \\ x_i &= \text{the size of the } i^{th} \text{ unit in the population,} \\ \tau_x &= \sum_{i=1}^{N} x_i = \text{total size of all units in the population,} \\ \mu_x &= \frac{\tau_x}{N} = \text{mean size of the units in the population,} \\ p_i &= \frac{x_i}{\tau_x} = \text{the probability of selecting unit } i \text{ on any one draw,} \\ n &= \text{the sample size.} \end{aligned}$$

1. Suppose N is unknown and we are interested in estimating it. Let $y_i = 1$ for all i in the Hansen-Hurwitz estimator:

$$\widehat{N} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i} = \frac{\tau_x}{n}\sum_{i=1}^{n}\frac{1}{x_i} \text{ (unbiased)}, \quad \widehat{\text{Var}}(\widehat{N}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{1}{p_i} - \widehat{N}\right)$$

2. Suppose the size variable itself is the variable of interest, that is $y_i = x_i$, and we are interested in estimating μ_x or τ_x.

   Estimating μ_x: if N and τ_x are known, then $\mu_x = \tau_x/N$ and there is nothing to estimate. If either N or τ_x (or both) are unknown, then observe that:

$$\widehat{\mu}_x = \frac{\tau_x}{\widehat{N}} = \frac{\tau_x}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

   $\widehat{\mu}_x$ is the harmonic mean of the sizes of the units in the sample. Note that we do not need to know either N or τ_x to estimate μ_x! This would be applicable, for example, to the pond example, where random points are placed on a map to select a sample of ponds. We know only the sizes of the ponds in the sample and not the total area of the ponds nor the number of ponds. There is not a closed form expression for the variance of $\widehat{\mu}_x$ and it must be approximated by either linearization or bootstrapping (discussed in a subsequent section of the notes).

   Estimating τ_x: If N is known (and μ_x is unknown) then we can estimate τ_x by

$$\widehat{\tau}_x = N\widehat{\mu}_x = \frac{N}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

   Since $\mathrm{Var}(\widehat{\tau}_x) = N^2\mathrm{Var}(\widehat{\mu}_x)$, the variance must be estimated by linearization or bootstrapping.

3. Suppose that the variable of interest y_i is not size (e.g., number of workers in the farm example). Let

$$\begin{aligned}\tau_y &= \sum_{i=1}^{N} y_i = \text{the population total of } y \text{ values},\\ \mu_y &= \frac{\tau_y}{N} = \text{the mean } y \text{ value for the population.}\end{aligned}$$

   Suppose we are just interested in estimating τ_y and/or μ_y. Then

   (a) If τ_x is known (which implies that the $p_i = x_i/\tau_x$ are known), then

$$\widehat{\tau}_y = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \text{ (unbiased)}, \quad \widehat{\mathrm{Var}}(\widehat{\tau}_y) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{\tau}_y\right)^2.$$

   If, in addition, N is known, then

$$\widehat{\mu}_y = \frac{\widehat{\tau}_y}{N} \text{ (unbiased)}, \quad \widehat{\mathrm{Var}}(\widehat{\mu}_y) = \frac{1}{N^2}\widehat{\mathrm{Var}}(\widehat{\tau}_y).$$

   (b) If either τ_x or N (or both) is unknown, then observe that

$$\widehat{\mu}_y = \frac{\widehat{\tau}_y}{\widehat{N}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}} = \frac{\sum_{i=1}^{n}\frac{y_i}{x_i}}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{1}$$

Note that we do not need to know either τ_x or N to estimate μ_y. This estimator is a ratio estimator and its variance must be estimated by linearization or bootstrapping.

(c) In part (b), if N is known, then

$$\widehat{\tau}_y = N\widehat{\mu}_y.$$

Since $\text{Var}(\widehat{\tau}_y) = N^2\text{Var}(\widehat{\mu}_y)$, the variance must be estimated by linearization or bootstrapping.

If N and τ_x are unknown, then τ_y cannot be estimated.

Any Probability Sampling Design: The Horvitz-Thompson Estimator (Section 6.2)

The Horvitz-Thompson estimator is a general estimator for a population total, which can be used for any probability sampling plan. This includes both sampling with and without replacement.

- Let π_i be the inclusion probability for the i^{th} unit; this is the probability that the unit is included in the sample (contrast this with the selection probability of the previous section).
- On each unit i, we measure a response y_i, and typically seek to estimate:

$$\tau = \sum_{i=1}^{N} y_i \text{ (population total)} \quad \text{or} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} y_i \text{ (population mean).}$$

Definition: The Horvitz-Thompson (H-T) estimator of τ is given by:

$$\widehat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{\pi_i} \qquad \text{where the sum is taken only over the } v \text{ distinct units in the sample.}$$

- The value v is sometimes referred to as the "effective" sample size.
- The higher the probability of inclusion, π_i, of a unit i to the sample, the less weight the corresponding response y_i is given. In this way the H-T estimator, like the Hansen-Hurwitz estimator, uses probability to weight the responses in estimating the total.
- The primary difference between the H-T and H-H estimator is the fact that the former uses the inclusion probability (π_i) of the units to the sample, whereas the latter uses the probability of selection (p_i) of a unit for a single draw. The H-H estimator is restricted to random sampling with replacement while the H-T estimator can be used in much wider range of sampling plans.

Mean and Variance of the Horvitz-Thompson Estimator

Mean: $\text{E}[\widehat{\tau}_\pi] = \text{E}\left[\sum_{i=1}^{v} \frac{y_i}{\pi_i}\right]$. Now what?

Let $z_i = \left\{ \begin{array}{ll} 1 & \text{if the } i^{th} \text{ unit is in the sample} \\ 0 & \text{otherwise} \end{array} \right\}$, $i = 1, \ldots, N$. Then:

$$\mathrm{E}[z_i] \quad =$$

$$\mathrm{Var}[z_i] \quad =$$

$$\mathrm{Cov}(z_i, z_j) \quad =$$

where π_{ij} is the joint inclusion probability of units i, j.

Returning to the expectation of $\widehat{\tau}_\pi$, we have:

$$\underline{\mathrm{E}\left[\widehat{\tau}_\pi\right]} \quad = \quad \mathrm{E}\left[\sum_{i=1}^{N} z_i \frac{y_i}{\pi_i}\right]$$

$$=$$

$$=$$

Variance:

$$\mathrm{Var}(\widehat{\tau}_\pi) \quad = \quad \mathrm{Var}\left(\sum_{i=1}^{N} z_i \frac{y_i}{\pi_i}\right)$$

$$=$$

$$=$$

$$=$$

- An unbiased estimator of the variance is given by:

$$\widehat{\mathrm{Var}}(\widehat{\tau}_\pi) = \sum_{i=1}^{v} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i\,=\,1}^{v} \sum_{\substack{j\,=\,1 \\ j\,\neq\,i}}^{v} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{y_i y_j}{\pi_{ij}},$$

  where the extra π_i in the denominator of the first term and the π_{ij} in the denominator of the second term can be attributed to the use of v sample units instead of the N population units in the theoretical variance.

Horvitz-Thompson Estimator for the Mean: To estimate the population mean μ, the corresponding Horvitz-Thompson estimator is given by:

$$\widehat{\mu}_\pi = \frac{\widehat{\tau}_\pi}{N}, \text{ with variance } \mathrm{Var}(\widehat{\mu}_\pi) = \frac{1}{N^2}\mathrm{Var}(\widehat{\tau}_\pi).$$

If N is unknown, we can estimate it (let $y_i = 1$ for all i).

The H-T Estimator for SRS without replacement.
Consider taking a simple random sample (SRS), without replacement, of size n from a population of size N. The inclusion and joint inclusion probabilities are:

$\pi_i =$ $\pi_{ij} =$

- Note that the Horvitz-Thompson estimator for SRS without replacement becomes:

$$\widehat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} \frac{y_i}{n/N} = N \cdot \frac{1}{n} \sum_{i=1}^{n} y_i = N\overline{y} = \widehat{\tau},$$

  the usual estimator of the population total τ for an SRS derived earlier.

- Do you think the Horvitz-Thompson variance will be $\mathrm{Var}(\widehat{\tau}) = \sqrt{N(N-n)\sigma^2/n}$ as it was before for SRS?

The H-T Estimator for random sampling with replacement
Reconsider the scenario for the Hansen-Hurwitz estimator, where we sample with replacement from a population such that the probabilities of selection on any given draw are unequal. These probabilities were denoted $p_1, \ldots, p_N$ for a population of size N. The inclusion and joint inclusion probabilities are:

$\pi_i \;=$

$\pi_{ij} \;=$

With these then, the Horvitz-Thompson estimator is: $\widehat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{1-(1-p_i)^n}$, and the H-T variance is a horrendous mess.

- For sampling with replacement, it is generally easier to use the Hansen-Hurwitz estimator.
- Although the Horvitz-Thompson estimator can be used for any probability sampling plan, there is often a simpler way to derive the estimator and its variance than through inclusion probabilities.

Comparison of H-H and H-T Estimators for the Farm Example

Consider again the farm example on page 30 of these notes where we selected a random sample with replacement of size $n = 5$ farms by PPS (probability-proportional to size) sampling. This was carried out by selecting random pixels and including the farms in which these pixels fell. Suppose the goal is to estimate the total number of workers on all the farms. The Hansen-Hurwitz estimator was computed earlier; now we will compute the Horvitz-Thompson estimator for the same sample. The sample is repeated in the table below, along with the relevant components for the H-T estimator.

Coordinates	Data	p_i	$\pi_i = 1-(1-p_i)^n$
8,19	D2	$5/625 = .0080$	.0394
19,25	C8	$28/625 = .0448$	.2048
21,21	B4	$12/625 = .0192$	.0924
15, 4	A8	$14/625 = .0224$	.1071
7,20	A3	$13/625 = .0208$	.0998

As the 5 farms selected here were distinct, the Horvitz-Thompson estimator of the total number of workers, τ_y, is:

$$\begin{aligned} \widehat{\tau}_\pi &= \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \left[\frac{2}{.0394} + \frac{8}{.2048} + \frac{4}{.0924} + \frac{8}{.1071} + \frac{3}{.0998}\right] \\ &= \underline{237.94} \text{ workers.} \end{aligned}$$

- Recall that the Hansen-Hurwitz estimate of the number of workers was 227.66 workers. Since the true total number of workers was $\tau_y = 249$, does this make the Horvitz-Thompson estimator better?

To compute the estimated variance of this estimated total number of workers, we need first to compute the joint inclusion probabilities for each pair of units in the sample. Using the formula derived in class, given as: $\pi_{ij} = \pi_i + \pi_j - (1-(1-p_i-p_j)^n)$, the table below gives the ten π_{ij} values corresponding to the ten pairs of units.

	Unit Number				
Unit #	1	2	3	4	5
1	-	.0066	.0029	.0034	.0032
2	-	-	.0156	.0181	.0169
3	-	-	-	.0081	.0075
4	-	-	-	-	.0087

The estimated variance is then computed as:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\tau}_\pi) &= \sum_{i=1}^{v}\left(\frac{1-\pi_i}{\pi_i^2}\right)y_i^2 + \sum_{i\,=\,1}^{v}\sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^{v}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{y_iy_j}{\pi_{ij}} \\
&= \left[\left(\frac{1-.0394}{.0394^2}\right)2^2 + \cdots + \left(\frac{1-.0998}{.0998^2}\right)3^2\right] \\
&\quad +2\left[\left(\frac{.0066-(.0394)(.2048)}{(.0394)(.2048)}\right)\frac{(2)(8)}{.0066} + \cdots + \left(\frac{.0087-(.1071)(.0998)}{(.1071)(.0998)}\right)\frac{(8)(3)}{.0087}\right] \\
&= 11191.15 - 2(4922.037) = \underline{1347.077},
\end{aligned}$$

giving a standard error of $\widehat{\mathrm{SE}}(\widehat{\tau}_\pi) = \sqrt{1347.077} = \underline{36.70 \text{ workers}}$. This is essentially the same as that found (36.75) with the Hansen-Hurwitz estimator.

Notes: Calculating joint inclusion probabilities can be very difficult for an arbitrary unequal probability design. In addition, the unbiased estimator of the variance of the Horvitz-Thompson estimator can be negative! An alternative estimator which does not require computation of the joint inclusion probabilities and is guaranteed to be non-negative is given in equation (6.8) on p. 70 of Thompson (2nd ed: eq. (8) on p. 54). In addition, other estimators of the variance of the Horvitz-Thompson estimator have been proposed including the Yates-Grundy estimator on p. 71 (p. 55 of 2nd ed.) for designs in which the effective sample size is fixed; the Yates-Grundy estimator is less likely to be negative and has smaller variance than the usual estimator in this situation. Finally, the Horvitz-Thompson estimator can have large variance if the y_i and π_i values are negatively correlated (units with large values of y_i have low probability of inclusion and vice-versa). An alternative estimator of the population mean in that case is the generalized unequal-probability estimator described in Section 6.3.

R Code for computing H-H and H-T estimates and SE's

```
# This is for the farm example: PPS sampling with replacement so both
# Hansen-Hurwitz and Horvitz-Thompson estimators can be used
> n <- 5                              # Sets the sample size
> y <- c(2,8,4,8,3)                   # Sets the vector of y-values
>
```

```
> # Compute H-H estimate and SE
> p <- (1/625)*c(5,28,12,14,13)      # Computes the vector of selection probs.
> tau.p <- (1/n)*sum(y/p)            # Computes the H-H estimate
> var.tau.p <- var(y/p)/n            # Computes the variance of the H-H estimate
> c(tau.p,sqrt(var.tau.p))     # H-H estimate and SE
[1] 227.65568  36.74815

> # Compute H-T estimate and SE
> # Note: all selected farms were distinct, that is v=n
> pi <- 1 - (1-p)^n                  # Computes the vector of inclusion probs.
> tau.pi <- sum(y/pi)                # Computes the H-T estimate

> # Compute the estimated variance of the H-T estimate by computing the two terms
> # (the single sum and the double sum separately)
> var1 <- sum(y^2*(1-pi)/pi^2)     # First term of variance of H-T
> # Second term: the multiplier 2 below is because the pair i,j is the same as j,i
> var2 <- 0
> for(i in 1:(n-1)){
+    for(j in (i+1):n){
+       pi.ij <- pi[i] + pi[j] - (1-(1-p[i]-p[j])^n) #joint inclusion probability
+ var2 <- var2 + 2*(y[i]*y[j]/pi.ij)*(pi.ij - pi[i]*pi[j])/(pi[i]*pi[j])
+    }
+ }
> var.tau.pi <- var1 + var2           # Computes the H-T estimated var.
> c(tau.pi,sqrt(var.tau.pi))          # H-T estimate and SE
[1] 237.93735  36.70255

> # Alternative estimate of variance of H-T estimator which does not
> # require joint inclusion probabilities: eq (8), p. 54; need value of N:
> # will substitute N.hat
> t <- n*y/pi
> N.hat <- sum(1/pi)
> varalt.tau.pi <- (1-n/N.hat)*var(t)/n
> sqrt(varalt.tau.pi)
[1] 36.2092
```

Ratio Estimation (Chapter 7)

This chapter covers the basic idea behind ratio estimation, gives the forms and properties of the relevant estimators, compares ratio estimation to other estimation methods studied, and provides some examples which use ratio estimation.

Reconsider the farm example on page 30 of the notes where we were interested in estimating:

1. the population total τ = the total # of workers, and
2. the population mean μ = average # of workers per farm.

In the last section, we used "size" as an auxiliary variable in the design phase of the study where we used PPS sampling to select farms. This was not only convenient (because we didn't have a list of all the farms from which to draw an SRS), but was advantageous because the number of workers was positively correlated with size. The Hansen-Hurwitz estimator based on a PPS sample has smaller variance than the estimator based on an SRS if there's a strong positive relationship between the size variable and the response variable.

Another way we could use the auxiliary variable "size" is in the estimation stage, after we have collected the data from an SRS. We can do this is through a ratio estimator. Like PPS sampling, a ratio estimator is advantageous only if there is a strong positive relationship between the auxiliary variable and the response variable. Specifically, ratio estimation is optimal when there is a linear relationship through the origin between the two variables.

It's important to note that in order to use an auxiliary variable x in a ratio estimator to estimate τ_y or μ_y, we need to know τ_x, the total value of x for the whole population. Ratio estimation is therefore commonly used when the auxiliary variable is a variable which is easily measured on the whole population while the response variable is harder to measure and is obtained from only an SRS of the population. Some situations where ratio estimation might be beneficial are:

- Let x = the girth of a tree, and y = the volume of the tree
- Let x = the total # of animals on a plot of land, and y = the # of females.
- Let x = total volume of a haul of fish, and y = the number fish in the haul.
- Let x = a visual estimate of the % of some ground cover, and y = the actual % of some ground cover.

Example 1: Consider the second situation above, where the population consists of $N = 20$ plots of land, and we take an SRS of $n = 7$ plots, counting the number of animals and number of females on these 7 plots. In addition, we also count the number of animals on all 20 plots, without knowing

what sex they are because it may be easy to count the number of animals on a plot, but hard to identify which are females. Assume all the plots are equal in size. Primary interest here is in estimating either:

$$\begin{aligned} \mu_y &= \text{the average number of females per plot, or} \\ \tau_y &= \text{the total number of females} = N\mu_y. \end{aligned}$$

Response Variable: y_i = the number of female animals on plot i, $i = 1, \ldots, N$,
Auxiliary Variable: x_i = the total number of animals on plot i, $i = 1, \ldots, N$.

The data for the SRS of size 7 are given to the right.

x_i	y_i	y_i/x_i
10	7	.7
18	12	.67
10	4	.4
12	6	.5
25	19	.76
15	7	.467
10	5	.5

We could estimate μ_y, τ_y without using the auxiliary variable:

$$\begin{aligned} \overline{y} &= \frac{60}{7} = \underline{8.57},\ s = 5.26,\ \widehat{\mathrm{SE}}(\overline{y}) = \sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}} = \underline{1.60}, \\ \widehat{\tau}_y &= N\overline{y} = 20(8.57) = \underline{171.4},\ \mathrm{SE}(\widehat{\tau}_y) = N\mathrm{SE}(\overline{y}) = \underline{32.0}. \end{aligned}$$

- These estimates will later be compared to those obtained through ratio estimation.

To estimate μ_y and τ_y using the auxiliary variable x, we first estimate the overall proportion of females in the population from the sample plots, then apply that estimated proportion to the total counts in all plots to estimate the total number of females in all the plots and the average number of females per plot. These steps are outlined below.

Definition: The population ratio R is:

$$R = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$$

How would we estimate R?

- We might consider calculating the ratio y_i/x_i of females in each sample plot and then taking the average of these ratios. Any problem with this?

- It is better to take the ratio of the means, than the mean of the ratios. This is the same as simply pooling the sampled plots and computing the proportion of females.

Definition: The sample ratio r is:

$$r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\overline{y}}{\overline{x}}.$$

- Expected value of r: $\mathrm{E}(r) = \mathrm{E}\left(\dfrac{\overline{y}}{\overline{x}}\right) \neq \dfrac{\mathrm{E}(\overline{y})}{\mathrm{E}(\overline{x})} = \dfrac{\mu_y}{\mu_x} = R$, so, in general, r is not an unbiased estimator of R. For most cases, however, the bias is small.

- Variance of r: The variance is approximated by:

$$\mathrm{Var}(r) \approx \left(\frac{N-n}{N}\right)\frac{1}{\mu_x^2}\cdot\frac{\sigma_r^2}{n} \text{ where } \sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2,$$

  with an estimated variance given by:

$$\widehat{\mathrm{Var}}(r) = \left(\frac{N-n}{N}\right)\frac{1}{\mu_x^2}\cdot\frac{s_r^2}{n} \text{ where } s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2.$$

  The sample mean $\overline{x}$ can be used in place of μ_x in the above expression if μ_x is not known.

- This variance approximation is based on a linearization of the ratio $\overline{y}/\overline{x}$ by a Taylor series expansion (this technique is discussed in the next section).

- When will this variance be small?

- Normally, since the estimate r of the population ratio R is biased, we would use the $\mathrm{MSE}(r) = \mathrm{Var}(r) + \mathrm{Bias}^2(r)$ to compare the ratio estimator to other estimators. However, the squared bias is generally very small, so it is often ignored.

Recall the table of animal counts given earlier, and consider the augmented table below to compute the estimated ratio and its variance:

x_i (# animals)	y_i (# females)	rx_i	$(y_i - rx_i)^2$
10	7		
18	12		
10	4		
12	6		
25	19		
15	7		
10	5		

- Since r is equal to the sample proportion $\widehat{p} = 60/100 = 0.6$ of females, why can't we compute the standard error based on the SRS formula for $\mathrm{SE}(\widehat{p})$?

$$\mathrm{SE}(\widehat{p}) = \sqrt{\left(\frac{N-n}{N}\right)\frac{\widehat{p}(1-\widehat{p})}{n-1}} = \sqrt{\left(\frac{N-100}{N}\right)\frac{.60(1-.60)}{100-1}} \approx .0416 \text{ (with } N = 350).$$

– This assumes we have taken an SRS of 100 animals, which is not true. What type of sample of animals have we taken?

- In looking at the form of the standard error of the estimator of r, it should be clear that the estimator will be "good" when the $y_i - rx_i$ are "small". If $y_i = rx_i$, then there is a linear relationship between y and x through the origin. This seems reasonable for the animals/females example, as for 0 animals there would be 0 females. And as the number of animals increases, we would expect the number of females to increase linearly with it.
- We can, and should, examine the relationship between x and y for our sample with a scatterplot.

Ratio Estimator of μ_y

- Suppose we are given:

$$\begin{aligned} \overline{x} &= \text{mean \# of animals per plot in the sample} \\ \mu_x &= \text{mean \# of animals per plot in entire population} \\ \overline{y} &= \text{mean number of females per plot in the sample} \end{aligned}$$

- The idea of a ratio estimator is to "adjust" the naive estimator $\overline{y}$ using the relationship between y & x. Recall that $R = \mu_y/\mu_x$ so

$$\mu_y = \left(\frac{\mu_y}{\mu_x}\right)\mu_x = R\mu_x.$$

We replace R by its estimator $r = \overline{y}/\overline{x}$ to give an estimator of the population mean μ_y (mean # of females per plot):

$$\widehat{\mu}_r = r \cdot \mu_x = \overline{y} \cdot \frac{\mu_x}{\overline{x}},$$

with corresponding variance and estimated variance given by:

$$\begin{aligned} \mathrm{Var}(\widehat{\mu}_r) &= \\ &= \\ \widehat{\mathrm{Var}}(\widehat{\mu}_r) &= \left(\frac{N-n}{N}\right)\frac{s_r^2}{n}, \text{ where } s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2. \end{aligned}$$

- Note that we need to know μ_x to use ratio estimation to improve the estimate of μ_y.
- An alternative estimator of variance is given as equation (7.7) on page 95 of Thompson (2nd ed: eq. (7) on p .69). This alternative estimator is more robust to the value of $\overline{x}$ than the variance estimator given above.

- Note that the estimated variance of $\widehat{\mu}_r$ given above has the same form as the estimated variance of $\overline{y}$ (the conventional estimator of the mean), except that s^2 is replaced by s_r^2. This implies that whenever s_r^2 is smaller than s^2, the ratio estimator will be superior to the conventional SRS-based estimator. When will this be true?

Ratio Estimator of τ_y

Since the population total $\tau_y = N\mu_y = NR\mu_x = R\tau_x$, an estimator of the population total (total # of females) is given by:

$$\widehat{\tau}_r \quad = \quad r \cdot \tau_x = \frac{\overline{y}}{\overline{x}}\,\tau_x \;\; \text{(where } \tau_x \text{ is assumed known)},$$

with corresponding variance and estimated variance given by:

$$\text{Var}(\widehat{\tau}_r) = N^2\text{Var}(\widehat{\mu}_r) = N(N-n)\frac{\sigma_r^2}{n}, \;\; \widehat{\text{Var}}(\widehat{\tau}_r) = N(N-n)\frac{s_r^2}{n},$$

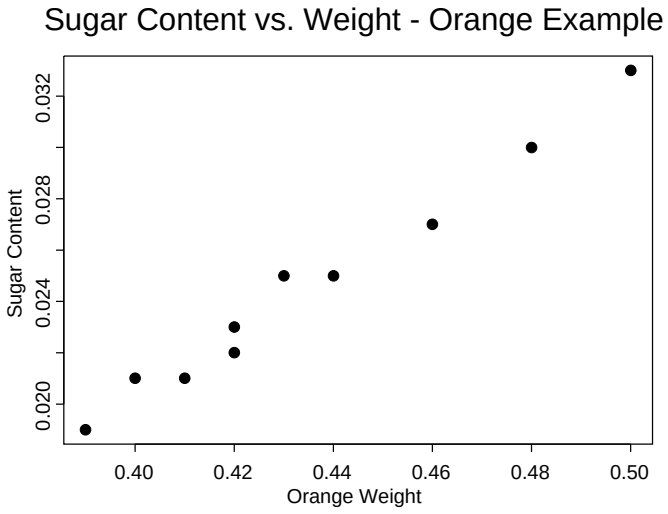
where s_r^2 was given earlier. Suppose $\tau_x = 350$ (total # animals). Then:

$$\begin{aligned}
\widehat{\tau}_r &= r \cdot \tau_x = (.6)(350) = \underline{210 \text{ females}}. \\
\widehat{\text{Var}}(\widehat{\tau}_r) &= N(N-n)\frac{s_r^2}{n} = 20(20-7)\frac{4.813}{7} = 178.77, \;\; \text{so that:} \\
\text{SE}(\widehat{\tau}_r) &= \sqrt{178.77} = \underline{13.37}.
\end{aligned}$$

For comparison, the estimated number of females ignoring the auxiliary information (just based on SRS of 7 plots) was $\widehat{\tau} = 171.4$ with $\text{SE}(\widehat{\tau}) = 32.0$. This illustrates the gains that are possible with ratio estimation when the response and auxiliary variables are linearly related through the origin.

Example 2: In a study to estimate the total sugar content of a truckload of oranges, a random sample of $n = 10$ oranges was juiced and weighed. The data for the 10 oranges are given in the table below and displayed in a plot of sugar content versus weight.

Orange	Sugar Content (in pounds)	Weight of Orange (in pounds)
1	.021	.40
2	.030	.48
3	.025	.43
4	.022	.42
5	.033	.50
6	.027	.46
7	.019	.39
8	.021	.41
9	.023	.42
10	.025	.44
	$\sum_{i=1}^{10} y_i = .246$	$\sum_{i=1}^{10} x_i = 4.35$



The total weight of all the oranges, obtained by first weighing the truck loaded and then unloaded, was found to be 1800 pounds. Estimate τ_y, the total sugar content for the oranges, and place a bound on the error of estimation. In this example, the sugar content of an orange (y) is the response and the weight of an orange (x) is the auxiliary variable.

- Note that if we ignore the auxiliary variable weight here, we cannot estimate the total sugar content τ_y as requested using basic SRS ideas, because we don't know the population size N = total # of oranges (the usual estimator of τ_y is: $\widehat{\tau} = N\overline{y}$, but here we don't know N).
- Here then, is a case where we must use a ratio estimator.
- What do we know?

- The estimated variance of $\widehat{\tau}_r$ is: $\widehat{\text{Var}}(\widehat{\tau}_r) = N(N-n)\dfrac{s_r^2}{n}$. Any problem here? What do we do?

Computing:

$$\begin{aligned} r &= \frac{\overline{y}}{\overline{x}} = \frac{\sum y_i}{\sum x_i} = \frac{.246}{4.35} = \underline{.05655}, \\ s_r^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2 = \frac{1}{9}[(.021 - .05655(.40))^2 + \cdots + (.025 - .05655(.44))^2] \\ &= \underline{(.00241)^2}, \\ \widehat{\mathrm{Var}}(\widehat{\tau}_r) &= \tau_x^2 \widehat{\mathrm{Var}}(r) \approx (1800)^2 \frac{1}{(.435)^2}\frac{(.00241)^2}{10} \\ &= \underline{9.949} \implies \mathrm{SE}(\widehat{\tau}_r) = \underline{3.15 \text{ pounds}}. \end{aligned}$$

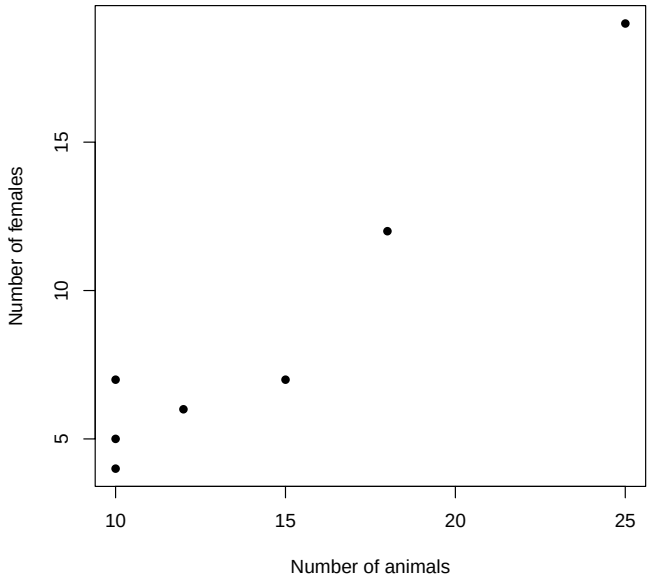
- An approximate 95% confidence interval for the total sugar content τ_y is:

$$\widehat{\tau}_r \pm t_9(.975) \cdot \mathrm{SE}(\widehat{\tau}_r) = 101.79 \pm (2.262)(3.15) = \underline{(94.66, 108.92) \text{ pounds}}.$$

R code for Example 1

```
# Example 1: ratio estimation of number and proportion of females
> x <- c(10,18,10,12,25,15,10)
> y <- c(7,12,4,6,19,7,5)
> n <- 7
> N <- 20
> plot(x,y,pch=16,xlab="Number of animals",ylab="Number of females")

# Estimation of proportion of females
> r <- mean(y)/mean(x)
> r
[1] 0.6
> sr2 <- (1/(n-1))*sum((y-r*x)^2)
> sr2
[1] 4.813333
> SE.r <- sqrt((1-n/N)*sr2/(mean(x)^2*n))  # assumes mu_x not known
> SE.r
[1] 0.04679815
> c(r - qt(.975,n-1)*SE.r,r + qt(.975,n-1)*SE.r)
[1] 0.4854891 0.7145109

# Estimation of total number of females
> tau.x <- 350  # total number of animals on all 20 plots (given)
> tau.hat <- r*tau.x
```

```
> SE.tau <- sqrt(N*(N-n)*sr2/n)
> c(tau.hat,SE.tau)
[1] 210.0000  13.3709
> c(tau.hat - qt(.975,n-1)*SE.tau,tau.hat + qt(.975,n-1)*SE.tau)
[1] 177.2826 242.7174
```



Linearization and Bootstrapping

We can often find exact expressions for the variance of an estimator (for example, the sample mean from an SRS), though the expression may involve unknown parameters, such as the population variance, which must then be estimated to give the standard error of the estimator. However, sometimes we cannot derive an exact expression for the variance of an estimator and we have to rely on other techniques for approximating the variance. Linearization is one such technique. Bootstrapping is another more general technique which also allows direct estimation of confidence intervals when a normal approximation to the sampling distribution of an estimator is not justified.

Linearization

Linearization refers to the method of finding variance approximations for smooth functions of random variables (whose variances and covariances are known) by using first-order Taylor series approximations. For estimators, linearization is important because while the theoretical variance of a mean of an SRS is known, there is no general theoretical expression for the variance of most functions of a mean, such as the inverse of a mean, or the ratio of two means. For example, in the previous chapter on ratio estimation, the variance of the estimator r of the population ratio R was approximated as:

$$\text{Var}(r) \;=\; \text{Var}\left(\frac{\overline{y}}{\overline{x}}\right) \;\approx\; \left(\frac{N-n}{N}\right)\frac{1}{\mu_x^2}\cdot\frac{\sigma_r^2}{n}.$$

This result comes from linearization as there is no exact expression for the variance of a ratio of two random variables in terms of the variances and covariance of the random variables. Its derivation will be shown shortly. First, we show how linearization works for a function of a single random variable.

Taylor Series Expansion: The Taylor series expansion of a function $f(\cdot)$ about a value a is given as:

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2!} + \cdots,$$

where we can often drop the higher order terms to give the approximation:

$$f(x) \approx f(a) + f'(a)(x-a).$$

This approximation will be good if $f(x)$ is a smooth function and x is "near" a.

In sampling, we're interested in approximating the variance of functions of the sample mean $\overline{x}$ from an SRS (or other probability sampling plan for which we have an unbiased estimator of the population mean μ_x). Since $\overline{x}$ will be "near" μ_x with high probability for large n, then a Taylor series expansion of $f(\overline{x})$ about μ_x gives the approximation:

$$f(\overline{x}) \approx f(\mu_x) + f'(\mu_x)(\overline{x}-\mu_x).$$

Taking the variance of both sides yields:

$$\text{Var}(f(\overline{x})) \approx [f'(\mu_x)]^2\text{Var}(\overline{x}).$$

- To approximate the variance of any smooth function of $\overline{x}$, we need only calculate the variance of $\overline{x}$ and the first derivative of the function.

Example 1: Suppose we want to approximate the variance of $\overline{x}^2$. Then $f(x) = x^2$ and $f'(x) = 2x$, so that:

$$\text{Var}(\overline{x}^2) \approx (2\mu_x)^2 \text{Var}(\overline{x}).$$

If $\overline{x}$ is the sample mean from an SRS from a finite population, then

$$\text{Var}(\overline{x}^2) \approx 4\mu_x^2 \left(\frac{N-n}{N}\right) \frac{\sigma_x^2}{n}$$

which can be estimated by

$$\widehat{\text{Var}}(\overline{x}^2) = 4\overline{x}^2 \left(\frac{N-n}{N}\right) \frac{s_x^2}{n}.$$

Example 2: Suppose we want to approximate the variance of $1/\overline{x}$. Then $f(x) = 1/x$ and $f'(x) = -1/x^2$, so that:

$$\text{Var}(1/\overline{x}) \approx \left[-\frac{1}{\mu_x^2}\right]^2 \text{Var}(\overline{x}).$$

Applying this result to an SRS gives

$$\text{Var}(1/\overline{x}) \approx \left(\frac{N-n}{N}\right) \frac{\sigma_x^2}{n\mu_x^4}$$

which can be estimated by substituting $\overline{x}$ for μ_x and s_x^2 for σ_x^2.

Two-Variable Taylor Series Expansion: Suppose now we want to approximate the variance of a function of random variables $\overline{x}$ and $\overline{y}$. A Taylor series expansion of $f(x, y)$ about the point (x_0, y_0) is given by:

$$f(x,y) = f(x_0,y_0) + \left.\frac{\partial f(x,y)}{\partial x}\right|_{(x_0,y_0)} (x - x_0) + \left.\frac{\partial f(x,y)}{\partial y}\right|_{(x_0,y_0)} (y - y_0) + \begin{pmatrix} \text{2nd and higher} \\ \text{order terms} \end{pmatrix}$$

We use the first-order Taylor Series approximation to $f(\overline{x}, \overline{y})$ around the point (μ_x, μ_y) to approximate the variance of $f(\overline{x}, \overline{y})$ as in the following example.

Example 3: Suppose we desire the approximate variance of $f(\overline{x}, \overline{y}) = \dfrac{\overline{y}}{\overline{x}}$. The first-order partial derivatives of $f(x, y) = y/x$ are

$$\frac{\partial f(x,y)}{\partial x} = \frac{-y}{x^2}, \quad \frac{\partial f(x,y)}{\partial y} = \frac{1}{x}.$$

Hence,

$$\begin{aligned} f(\overline{x}, \overline{y}) &= \frac{\overline{y}}{\overline{x}} \approx \frac{\mu_y}{\mu_x} + \frac{-\mu_y}{\mu_x^2}(\overline{x} - \mu_x) + \frac{1}{\mu_x}(\overline{y} - \mu_y) \\ \Longrightarrow \text{Var}\left(\frac{\overline{y}}{\overline{x}}\right) &\approx \frac{\mu_y^2}{\mu_x^4}\text{Var}(\overline{x}) + \frac{1}{\mu_x^2}\text{Var}(\overline{y}) - \frac{2\mu_y}{\mu_x^3}\text{Cov}(\overline{x}, \overline{y}). \end{aligned}$$

(using the fact that the variance of the sum of two random variables is the sum of the variances plus two times the covariance). The approximate variance of the $\overline{y}/\overline{x}$ from an SRS is:

$$\text{Var}\left(\frac{\overline{y}}{\overline{x}}\right) \approx \underbrace{\left(\frac{N-n}{N}\right)}_{\substack{\text{if the fpc} \\ \text{is required}}} \left[\frac{\mu_y^2}{\mu_x^4}\cdot\frac{\sigma_x^2}{n} + \frac{1}{\mu_x^2}\cdot\frac{\sigma_y^2}{n} - \frac{2\mu_y}{\mu_x^3}\cdot\frac{\rho\sigma_x\sigma_y}{n}\right], \tag{2}$$

where $\text{Cov}(\overline{x},\overline{y}) = \dfrac{\text{Cov}(x,y)}{n} = \dfrac{\rho\sigma_x\sigma_y}{n}$.

The corresponding estimated variance of $\overline{y}/\overline{x}$ is:

$$\widehat{\text{Var}}\left(\frac{\overline{y}}{\overline{x}}\right) \approx \left(\frac{N-n}{N}\right)\frac{1}{n}\left[\frac{\overline{y}^2}{\overline{x}^4}s_x^2 + \frac{1}{\overline{x}^2}s_y^2 - \frac{2\overline{y}}{\overline{x}^3}\,\widehat{\rho}\,s_x s_y\right].$$

Some Useful Approximations: The linear approximation via a Taylor series expansion gives the approximate variance for the following three useful functions of sample means $\overline{x}$ and $\overline{y}$.

1. $\text{Var}\left(\dfrac{1}{\overline{x}}\right) \approx \left(\dfrac{1}{\mu_x^4}\right)\text{Var}(\overline{x})$.
2. $\text{Var}\left(\dfrac{\overline{y}}{\overline{x}}\right) \approx \left(\dfrac{\mu_y^2}{\mu_x^4}\right)\text{Var}(\overline{x}) + \left(\dfrac{1}{\mu_x^2}\right)\text{Var}(\overline{y}) - 2\left(\dfrac{\mu_y}{\mu_x^3}\right)\text{Cov}(\overline{x},\overline{y})$.
3. $\text{Var}(\overline{x}\,\overline{y}) \approx \mu_y^2\text{Var}(\overline{x}) + \mu_x^2\text{Var}(\overline{y}) + 2\mu_x\mu_y\text{Cov}(\overline{x},\overline{y})$.
   Note: If $\overline{x}$ and $\overline{y}$ are independent, then an exact expression for $\text{Var}(\overline{x}\,\overline{y})$ can be derived: $\text{Var}(\overline{x}\,\overline{y}) = \mu_y^2\text{Var}(\overline{x}) + \mu_x^2\text{Var}(\overline{y}) + \text{Var}(\overline{x})\text{Var}(\overline{y})$.

To obtain estimates of these variances, simply substitute sample values of the means, variances and correlation.

Some applications:

- Ratio Estimator: Equation (2) above gave the approximate variance of the sample ratio $r = \overline{y}/\overline{x}$ based on linearization. This is algebraically equivalent to the expression given in Thompson in the second to last equation on p. 95, (2nd ed: last equation on p. 69):

  $$\text{Var}(r) \approx \left(\frac{N-n}{N\mu_x^2}\right)\frac{\sigma_r^2}{n}.$$

  Recall that the ratio estimator of the population mean μ_y is $\widehat{\mu}_r = r\mu_x$. Its approximate variance is therefore $\text{Var}(\widehat{\mu}_r) = \mu_x^2\text{Var}(r) = \left(\dfrac{N-n}{N}\right)\dfrac{\sigma_r^2}{n}$ which is estimated by $\widehat{\text{Var}}(\widehat{\mu}_r) = \overline{x}^2\widehat{\text{Var}}(r)$.

- PPS sampling with replacement (Chapter 6, Section 1): The following summarizes the results in the notes starting on page 32, "Summary of results for PPS sampling", and includes estimated variances from the linearization method where needed.

  As before, consider the following notation. Let:

$$\begin{aligned} N &= \text{the population size,} \\ n &= \text{the sample size,} \\ x_i &= \text{the size of the } i^{th} \text{ unit in the population,} \\ \tau_x &= \sum_{i=1}^{N} x_i, \\ p_i &= x_i/\tau_x = \text{the probability of selecting the } i^{th} \text{ unit on each draw,} \\ y_i &= \text{the response variable (variable of interest),} \\ \tau_y &= \sum_{i=1}^{N} y_i, \quad \mu_y = \frac{1}{N}\tau_y. \end{aligned}$$

  The estimates and approximate standard errors, using the Hansen-Hurwitz estimator where possible, of relevant population quantities are given below.

  1. Estimating N:

$$\widehat{N} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i} \text{ (unbiased)}, \quad \widehat{\text{Var}}(\widehat{N}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{1}{p_i} - \widehat{N}\right)^2 \text{ (Hansen-Hurwitz)}$$

  2. Estimating μ_x and τ_x:

$$\widehat{\mu}_x = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \text{ (biased)}, \quad \widehat{\text{Var}}(\widehat{\mu}_x) = \left(\frac{1}{\overline{v}^4}\right)\frac{s_v^2}{n} \text{ (linearization)}$$

     where $v_i = \dfrac{1}{x_i}$, and $\overline{v}$ and s_v^2 are the sample mean and variance of the v_i's. Note that we can calculate the variance of the denominator of $\widehat{\mu}_x$ because it is the mean of the $1/x_i$.

$$\widehat{\tau}_x = \frac{N}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \text{ (biased)}, \quad \widehat{\text{Var}}(\widehat{\tau}_x) = N^2\widehat{\text{Var}}(\widehat{\mu}_x) = \left(\frac{N^2}{\overline{v}^4}\right)\frac{s_v^2}{n} \text{ (linearization)}$$

  3. Estimating τ_y and μ_y:

$$\widehat{\tau}_y = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \text{ (unbiased)}, \quad \widehat{\text{Var}}(\widehat{\tau}_y) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{\tau}_y\right)^2 = \frac{s_w^2}{n} \text{ (Hansen-Hurwitz)}$$

where $w_i = \dfrac{y_i}{p_i}$ and s_w^2 is the sample variance of the w_i's (note that $\widehat{\tau}_y = \overline{w}$).

Estimating μ_y, N known:

$$\widehat{\mu}_y = \frac{\widehat{\tau}_y}{N} \text{ (unbiased)}, \quad \widehat{\mathrm{Var}}(\widehat{\mu}_y) = \frac{1}{N^2}\widehat{\mathrm{Var}}(\widehat{\tau}_y) \text{ (Hansen-Hurwitz)}$$

Estimating μ_y, N unknown:

$$\begin{aligned}
\widehat{\mu}_y &= \frac{\widehat{\tau}_y}{\widehat{N}} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{y_i}{p_i}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{1}{p_i}} = \frac{\displaystyle\sum_{i=1}^{n}\frac{y_i}{x_i}}{\displaystyle\sum_{i=1}^{n}\frac{1}{x_i}} \text{ (biased)} \\
\widehat{\mathrm{Var}}(\widehat{\mu}_y) &= \left(\frac{\overline{t}^2}{\overline{v}^4}\right)\frac{s_v^2}{n} + \left(\frac{1}{\overline{v}^2}\right)\frac{s_t^2}{n} - 2\left(\frac{\overline{t}}{\overline{v}^3}\right)\frac{\widehat{\rho}_{t,v}s_t s_v}{n} \text{ (linearization)}
\end{aligned}$$

where $t_i = \dfrac{y_i}{x_i}$, $v_i = \dfrac{1}{x_i}$, and $\overline{t}, \overline{v}, s_t^2, \& s_v^2$ are the sample means and variances, and $\widehat{\rho}_{t,v}$ is the sample correlation between the t_i's and the v_i's.

- Other Applications: Suppose we want to estimate $R = \mu_y/\mu_x$ and we already have independent estimates $\widehat{\mu}_x$ and $\widehat{\mu}_y$ of the means available (from two different studies, for example) along with estimated variances $\widehat{\mathrm{Var}}(\widehat{\mu}_x)$ and $\widehat{\mathrm{Var}}(\widehat{\mu}_y)$ (i.e., the standard errors squared). It doesn't matter what sampling plans or sample sizes were used to generate these independent estimates as long as valid standard errors can be calculated. Then, we can estimate R by $\widehat{R} = \widehat{\mu}_y/\widehat{\mu}_x$ and, by the linearization approximation, estimate $\mathrm{Var}(\widehat{R})$ by:

$$\widehat{\mathrm{Var}}(\widehat{R}) = \left(\frac{\widehat{\mu}_y^2}{\widehat{\mu}_x^4}\right)\widehat{\mathrm{Var}}(\widehat{\mu}_x) + \left(\frac{1}{\widehat{\mu}_x^2}\right)\widehat{\mathrm{Var}}(\widehat{\mu}_y).$$

  - Note that the covariance term drops out because these are assumed to be independent estimates.

Example 4: Suppose it is desired to estimate the average expenditure per day for visitors to Yellowstone National Park. One study estimated the average expenditures per trip per person, but did not obtain trip length information. The estimate was \$240 with a standard error of \$60. Another study estimated the average length of a trip but did not gather expenditure data. The estimate was 2.3 days with a standard error of 0.5 days. With these two studies then, the estimated expenditure per day is $240/2.3 = \$104.3$ per person. The estimated variance of the estimate is

$$\left(\frac{240^2}{(2.3)^4}\right)(.5)^2 + \left(\frac{1}{(2.3)^2}\right)(60)^2 = 1195.1,$$

so the estimated standard error is $\sqrt{1195.1} = \$34.60$.

If we were interested in estimating the product of two means for which we already had independent estimates of the individual means, we could use the exact formula for $\text{Var}(xy)$ in formula 3 on page 50 of the notes (under "Some Useful Approximations").

Confidence Intervals The approximate standard errors generated by linearization can be used to generate normal-based confidence intervals of the form $\widehat{\theta} \pm z\text{SE}(\widehat{\theta})$. This assumes the sampling distribution of the estimator is approximately normal. Normality can be justified asymptotically by the delta method, which states that a function of asymptotically normal random variables is also asymptotically normal under some fairly general conditions on the function. This is, of course, only an asymptotic result, and the finite n distribution of the estimator may not be approximately normal. Bootstrapping is one way to address this issue.

Bootstrapping

Suppose $y_1, \ldots, y_n$ is a random sample from some large population (so that the fpc can be ignored for now), and we wish to estimate some population parameter θ. If $\widehat{\theta}$ is an estimate of θ with some standard error $\text{SE}(\widehat{\theta})$, then under certain conditions a confidence interval for θ can be formed as:

$$\widehat{\theta} \pm t\text{SE}(\widehat{\theta}) \quad (z \text{ may sometimes be used instead of } t).$$

The validity of this CI depends on three basic assumptions:

1. $\text{SE}(\widehat{\theta})$ is a good estimate of the standard deviation of the sampling distribution of $\widehat{\theta}$.
2. $\widehat{\theta}$ is unbiased or nearly unbiased for θ.
3. The sampling distribution of $\widehat{\theta}$ is approximately normal.

The method of bootstrapping can address all of these issues:

1. It can provide an estimate of $\text{SE}(\widehat{\theta})$ when no theoretical closed-form expression for $\text{SE}(\widehat{\theta})$ exists, or provide an alternative if we are uncertain about the accuracy of an existing estimate. For example, an approximate standard error based on linearization may not be very accurate. Bootstrapping allows us to check this. Can you think of some examples where the SE cannot be estimated or must be approximated?

2. It can provide an estimate of the bias of $\widehat{\theta}$ as an estimator of θ.
3. It can provide information on the shape of the sampling distribution of $\widehat{\theta}$. This can be used to calculate improved confidence intervals over the normal-based ones if the sampling distribution is not normal.

So what is bootstrapping? What we will describe in these notes is the most basic (and most common) type of bootstrap: the nonparametric bootstrap. The theory behind the nonparametric

bootstrap can best be explained with a simple example.

Suppose we have an SRS and wish to estimate the standard error of the sample median, say m, as an estimate of the population median, say M. Unlike the sample mean, whose variance depends only on the population variance σ^2, which can easily be estimated from the sample, $\text{Var}(m)$ depends on the exact population distribution, which we don't know.

The idea behind the bootstrap: if we knew the y-values for the entire population $(y_1, \ldots, y_N)$, then we could estimate the sampling distribution of m by simulation. How?

1.

2.

3.

We could estimate $\text{SE}(m)$ from the standard deviation of the several thousand medians.

Unfortunately, we do not know the y-values for the population; we only know the y-values for the sample. The nonparametric bootstrap says to assume that the population looks exactly like our sample, only many times larger. This is really our best nonparametric guess as to what the population looks like.

For example, if our SRS of y-values ($n = 5$) is:

$$4 \quad 8 \quad 2 \quad 10 \quad 12$$

then we assume that 20% of the population y-values are 4, 20% are 8, etc. In this way, this "bootstrap population" represents our best guess as to what the actual population looks like.

To perform bootstrapping, we simulate drawing samples of size $n = 5$ from this "bootstrap population." If N is large relative to n, this is equivalent to drawing random samples of size 5 with replacement from the original sample of size 5. Since we are sampling with replacement, we will not necessarily get the same sample every time. For example, issuing the R command `sample(c(4,8,2,10,12),5,replace=T)` three times gave the following three samples:

$$8 \quad 2 \quad 12 \quad 12 \quad 4 \qquad\qquad 10 \quad 2 \quad 2 \quad 8 \quad 12 \qquad\qquad 4 \quad 2 \quad 4 \quad 4 \quad 12$$

Finally, generate a large number of bootstrap samples (say 10000) and calculate the sample median for each sample. We can use this set of 10000 values as follows.

1. The standard deviation of the 10000 sample medians is an estimate of $\text{SE}(m)$.

2. We can estimate the bias of m by the following reasoning. Recall that the bias of an estimator $\widehat{\theta}$ is given by: $\text{Bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta$. The sample median of our original sample is the population median of our "bootstrap population" from which we are generating bootstrap samples. If the mean of the 10000 bootstrap sample medians is very different from the bootstrap population median (the original sample median), then this suggests that the sample median is biased for estimating the population median for a population which looks like our sample. An estimate of the bias is then given by $\overline{m}_B - m$, where $\overline{m}_B$ is the mean of the 10000 bootstrap sample medians. If $\overline{m}_B$ is much bigger than m, this suggests that the sample median tends to overestimate the population median (for these data) and therefore m is likely bigger than M, the population median (and vice-versa). We generally don't adjust the parameter estimate even if the estimator appears to be biased, but we can adjust for bias in the confidence interval for the parameter (see below).

3. We can also look at a histogram of the 10000 bootstrap medians to give us an estimate of what the sampling distribution of the sample median looks like for this sample size. This will help us in deciding whether a normal based confidence interval is appropriate.

The more bootstrap samples you take, the more accurate your estimates of the SE and bias. Since it's generally very fast to do many thousands of bootstrap replications, there is usually no reason not to do 10,000 or even 100,000. In addition, a large number of replications tare needed to construct the nonparametric bootstrap confidence intervals discussed below.

Bootstrapping in R

Suppose as discussed above that an SRS of size 5 yielded the y-values 4,8,2,10,12, and we would like to estimate SE(m) by bootstrapping. There is a command in R called `boot` that will do this automatically, but to understand how to use it, it's useful to see how we might write our own program to do bootstrapping in R. To generate one bootstrap sample, we could do as follows:

```
> y <- c(4,8,2,10,12)
> y
[1]  4  8  2 10 12
> i <- sample(5,5,replace=T)
> i
[1] 3 3 1 2 2
> y[i]
[1] 2 2 4 8 8
> median(y[i])
[1] 4
```

We would want to store the value of the median for the bootstrap sample (the `4`) and then repeat the process (with a loop) several thousand times, getting a new vector `i` each time we called `sample`.

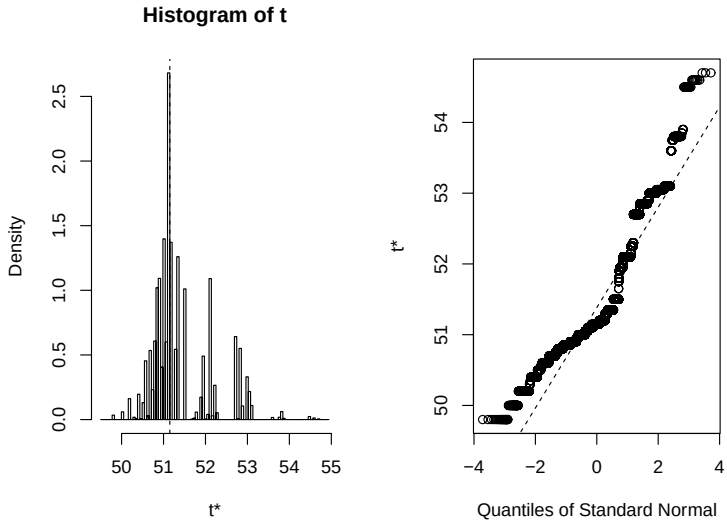
Notice that rather than sampling with replacement from the data vector (e.g., `sample(y,5,T)`), I sampled from the integers 1 to n and then got the data values for the bootstrap sample by using the expression `y[i]`. There is no difference between the two ways of doing it, but the latter illustrates how the `boot` command does it and will help in understanding how to use the command `boot`.

The `boot` command is available only in a separate package also called `boot` which is included with the base distribution of R. The boot package must be loaded before the `boot` command can be used. The package can be loaded by either the command `library(boot)` or through the menus by selecting **Packages...Load package...**. A package only needs to be loaded once at the beginning of each session.

Bootstrapping the median for a sample of size 5 is not very interesting (there are only 5 possible values of the median for any bootstrap sample) so we'll demonstrate the use of the `boot` command with a larger data set. The `boot` command is explained further at the end of this handout. For now, we'll focus on the output.

Bootstrap Example 1: Actual net weights were measured for 18 packages of M&M's labeled "49.3 grams". Assume these 18 packages represent an SRS from a very large shipment (that is, assume the population is essentially infinite). Suppose we are interested in estimating the median net weight and calculating a confidence interval.

```
> library(boot)
> y <- c(47.9, 48.5, 49.8, 49.8, 50.2, 50.6, 50.8, 50.9, 51.1, 51.2, 51.5,
+      52.7, 53.0, 53.1, 54.5, 54.7, 55.8, 55.9)
> bmed <- function(x,i) median(x[i])
> b <- boot(y,bmed,100000)
> b

ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = y, statistic = bmed, R = 1e+05)
Bootstrap Statistics :
    original    bias    std. error
t1*    51.15 0.2306955   0.7055834

> plot(b)
```



This output shows that the sample median is $m = 51.15$ grams, the estimated bias is 0.231 grams, and the bootstrap estimated standard error is $\text{SE}_B(m) = 0.7056$ grams. The `plot` command applied to the stored result of a boot command gives two plots of the bootstrap values: a histogram and a normal probability plot. These represent the estimated sampling distribution of the statistic we're interested in. The plots show that the sampling distribution of m is multimodal and not normal.

Bootstrap Confidence Intervals

Numerous ways have been proposed to generate confidence intervals through bootstrapping. I will discuss only three pretty common ones – Standard, Percentile, and BCa. The R command `boot.ci(b)` (where `b` is the stored result of the `boot` command) generates the latter two (plus three others). It does not generate the Standard one, but it is easily done by hand. The default confidence level in `boot.ci` is 0.95 but this can be changed with the `conf=` option. As you move down the list, the methods become more general (and more complicated) in that they try to adjust for non-normality and bias.

1. Standard: The simplest way to generate a confidence interval is simply to use the bootstrap SE – call it $\text{SE}_b(\widehat{\theta})$ – in the usual normal-based formula:

$$\widehat{\theta} \pm z\text{SE}_b(\widehat{\theta}).$$

   This is reasonable if $\widehat{\theta}$ is approximately unbiased and its sampling distribution is approximately normal (which we can assess from the bootstrap output as outlined above).

   In the M&M data, the 95% confidence interval for M by this method is $51.15 \pm 1.96(0.7055834) = (49.77, 52.53)$ grams.

2. Percentile: This interval uses the $\alpha/2$ and $1-\alpha/2$ quantiles of the 10000 (or whatever number) bootstrap estimates. It's given by the `boot.ci` command, but is easily computed also by the command `quantile(b$t,c(.025,.975))` for a 95% confidence interval (where `b` is the saved result of the `boot` command). This method implicitly assumes the sampling distribution of $\widehat{\theta}$ is symmetric, but not necessarily normal, and centered at θ.

```
> # Percentile CI
> quantile(b$t,c(.025,.975))
 2.5% 97.5%
50.40 53.05
```

3. BCa: The `BCa` interval attempts to adjust for both bias and non-normality of the sampling distribution. It is the most commonly used non-normal confidence interval method. You'll need to use the `boot.ci` command to compute it since it's somewhat complicated.

```
> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100000 bootstrap replicates
CALL :
boot.ci(boot.out = b)

Intervals :
Level      Normal              Basic
95%   (49.54, 52.30 )   (49.25, 51.90 )

Level     Percentile            BCa
95%   (50.40, 53.05 )   (50.40, 53.00 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(b) : bootstrap variances needed for studentized intervals
```

Note that the Percentile and BCa intervals are pretty similar. They would be preferred to the Standard interval because the sampling distribution is not normal. In general, just reporting the bootstrap SE is all you may want sometimes, but if you are going to report a confidence interval, it would be preferable to report the Percentile or BCa interval which take better advantage of the information the bootstrap provides about the sampling distribution of the estimator.

Assumptions in Bootstrapping

In the bootstrap procedure outlined above, it is important to note the following assumptions and related issues:

1. It assumes an infinite population since it uses sampling with replacement from the observed sample. This is not a problem if N is large relative to n. However, if N is small, it is unclear what to do! One possibility is to calculate the usual bootstrap estimate of $\text{Var}(\widehat{\theta})$ and then multiply the result by the finite population correction, $(N-n)/N$. This only works for the normal-based confidence intervals, though. If $k = N/n$ is an integer, it has been suggested to repeat each observation in the sample k times to create a "bootstrap population" of size N. A bootstrap sample would then be a random sample of size n taken without replacement from the bootstrap population. Other methods for finite populations are discussed in Davison nd Hinkley (1997), Section 3.7.

2. It assumes simple random sampling. Since the bootstrap samples are generated by simple random sampling, it is estimating the sampling distribution of $\widehat{\theta}$ for simple random sampling. If the original sample was a stratified random sample, for example, then we want to estimate the sampling distribution of $\widehat{\theta}$ for a *stratified random sample*. Therefore, we would sample with replacement within each stratum. If the original sample was a multistage sample, then we want to mimic that procedure in the bootstrap sample. It is sometimes very tricky to figure out how to do this correctly and is still an active area of research for many types of sampling.

Bootstrapping for Unequal Probability Sampling

It turns out that the bootstrap, as outlined above (random sampling with replacement from the observed sample), is valid for one other situation: unequal probability sampling with replacement (where the draws are therefore independent), as discussed in Section 6.1 of Thompson.

- Even though the original sample was generated by sampling with unequal probabilities, the bootstrap is performed by sampling with replacement from the original sample with *equal* probabilities. Why? Because the units with high p_i's are more likely to be in the original sample, so high p_i units already tend to be over-represented in the sample.

- So, although it might seem that the bootstrap should require unequal probability sampling from the original sample to simulate exactly how the original sample was drawn, this is not the case here. This demonstrates how tricky bootstrapping can be for more complicated sampling plans.

Bootstrap Examples Using R

Here are two additional examples that demonstrate how to use the `boot` command in R and what output is available. The first example demonstrates how to bootstrap a ratio of two means. The second example demonstrates bootstrapping for unequal probability sampling with replacement.

Bootstrap Example 2: Estimating a ratio from an SRS. This is Example 2 from page 44 of the

notes. In a study to estimate the total sugar content of a truckload of oranges, a random sample of $n = 10$ oranges was juiced and weighed. x is the weight (lbs) and y is the sugar content (lbs). The goal is to estimate the ratio, $R = \tau_y/\tau_x$ for the shipment.

```
> x <- c(.4,.48,.43,.42,.5,.46,.39,.41,.42,.44)
> y <- c(.021,.03,.025,.022,.033,.027,.019,.021,.023,.025)

# Estimate the ratio: pounds of sugar per pound of oranges
> r <- sum(y)/sum(x)
> r
[1] 0.05655172

# Calculate the approximate SE derived by linearization
# (p. 95 of Thompson, 2nd ed: p. 70)
> sr2 <- (1/9)*sum((y-r*x)^2)
> sr2
[1] 5.810649e-006
> se.r <- sqrt(sr2/(10*mean(x)^2))
> se.r
[1] 0.001752359
```

$$\left(s_r^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - r x_i)^2 \right)$$

$$\left(\mathrm{SE}(r) \approx \sqrt{\frac{1}{\mu_x^2} \cdot \frac{s_r^2}{n}} \right)$$

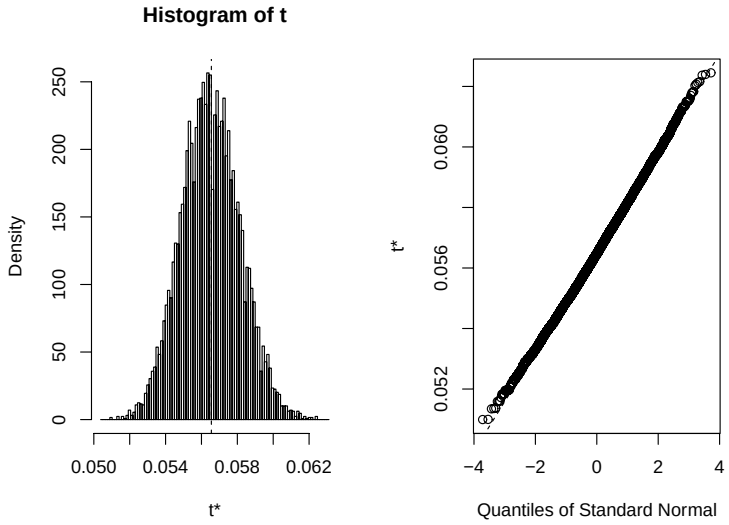
```
# Calculate a 95% confidence interval for R, the population ratio
> c(r - qt(.975,9)*se.r,r + qt(.975,9)*se.r)
[1] 0.05258761 0.06051584
```

A 95% CI is $(.0526, .0605)$ pounds of sugar per pound of oranges for the whole shipment.

Now use the bootstrap to obtain an alternative estimate of the SE and alternative confidence intervals and to examine the bias and sampling distribution of r. In order to use the `boot` function, we need to create a data frame.

```
> library(boot)
> orange <- data.frame(x,y) # x and y defined above
> orange
     x     y
1  0.40 0.021
2  0.48 0.030
3  0.43 0.025
4  0.42 0.022
5  0.50 0.033
6  0.46 0.027
7  0.39 0.019
```

```
8  0.41 0.021
9  0.42 0.023
10 0.44 0.025

> b <- boot(orange,function(a,i) sum(a$y[i])/sum(a$x[i]),100000)
> b
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = orange, statistic = function(a, i) sum(a$y[i])/sum(a$x[i]),
    R = 1e+05)
Bootstrap Statistics :
      original        bias    std. error
t1* 0.05655172 -3.737428e-05 0.001658623

> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100000 bootstrap replicates

CALL :
boot.ci(boot.out = b)

Intervals :
Level      Normal              Basic
95%   ( 0.0533,  0.0598 )   ( 0.0533,  0.0598 )

Level     Percentile            BCa
95%   ( 0.0533,  0.0598 )   ( 0.0536,  0.0601 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(b) : bootstrap variances needed for studentized intervals

> plot(b)
```



We can see that the estimated bias is less than .1% of the estimate of r, meaning the bias is likely negligible. We also see that the bootstrap estimate of the SE is only slightly smaller than the approximate formula via linearization gives. The sampling distribution looks approximately normal and centered at the observed r (which is represented by the dotted line). Note that the Percentile and BCa confidence intervals are very similar to each other and are not too much different than the symmetric interval based on the linearization estimate of the SE.

To estimate the total amount of sugar in the shipment, we multiply r by 1800, the total weight of the shipment. The estimated SE is also multiplied by 1800 (whether from the formula or the bootstrap). A 95% confidence interval for the total amount of sugar is then obtained by simply multiplying the endpoints of any of the 95% CI's for r by 1800.

Bootstrap Example 3. Unequal Probability Sampling. Reconsider the farm example on page 30 of the notes. Farms are sampled with replacement with probability proportional to size. The goal is to estimate the average size of all of the farms. The estimator is the harmonic mean of the sizes of the farms in the sample with approximate variance given by linearization (see item 2 on p. 51 of the notes). As discussed above, bootstrapping can be used for sampling with replacement with unequal probabilities. No adjustment is necessary for the unequal probability sampling since the observed sample already reflects this (units with higher probabilities of selection are likely overrepresented in the sample, so for the bootstrap, one samples at random with replacement with equal probabilities of selection).

A sample of 15 farms was selected by PPS sampling. Here are their sizes.

```
> size <- c(8,4,28,4,3,8,6,8,8,28,4,12,6,9,30)
```

```
> mean(size)
[1] 11.06667
```

The mean size of the farms is 11.07 but this is a biased estimate of the mean size of all the farms because of the PPS sampling. A nearly unbiased estimate is the harmonic mean, calculated as:

```
> v <- 1/size
> 1/mean(v)
[1] 6.769341
```

$$\left(\widehat{\mu}_x = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} v_i} = \frac{1}{\overline{v}},\ v_i = \frac{1}{x_i} \right)$$

The harmonic mean is 6.77; this is the estimate of the mean size of all the farms (the true mean size is actually $\mu = 625/79 = 7.91$ units). Now, compute the estimated SE via linearization:

```
> mv <- mean(v)
> se.hm <- sqrt((1/mv^4)*var(v)/15)
> se.hm
[1] 1.059578
```

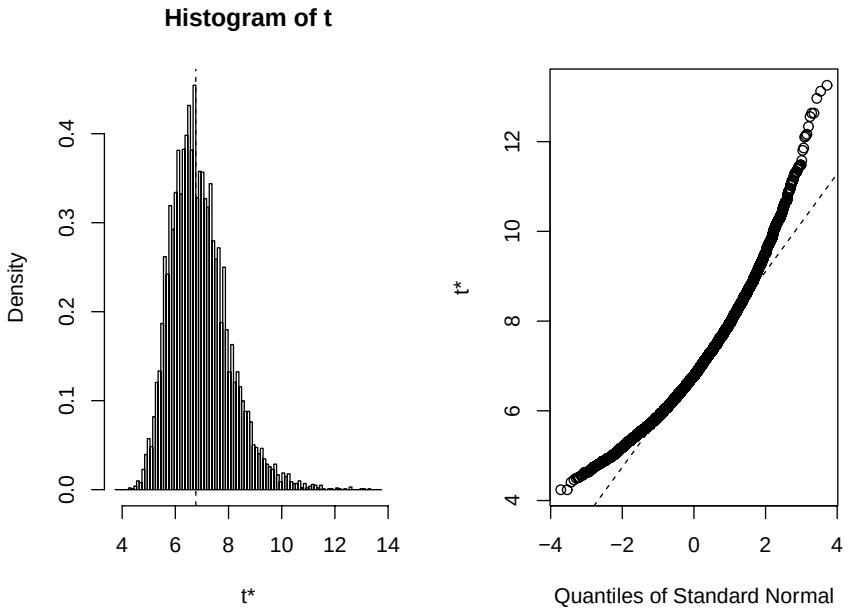
$$\left(\mathrm{SE}(\widehat{\mu}_x) = \sqrt{\frac{1}{\overline{v}^4} \cdot \frac{s_v^2}{n}} \right)$$

The SE is 1.06. A 95% CI is thus $(4.69, 8.85)$:

```
> 1/mv - qnorm(.975)*se.hm
[1] 4.692605
> 1/mv + qnorm(.975)*se.hm
[1] 8.846077
```

We can also bootstrap this estimate:

```
> bsize <- boot(size,function(x,i) 1/mean(1/x[i]),10000)
> bsize
Call:
boot(data = size, statistic = function(x, i) 1/mean(1/x[i]),
    R = 1e+05)
Bootstrap Statistics :
    original    bias    std. error
t1* 6.769341 0.1560974    1.100858

> plot(bsize)
```



The bootstrap SE is slightly higher than that obtained by the Taylor series approximation. However, the estimator appears to be slightly biases; it tends to overestimate the true mean size (though the bias is still on the order of only about 2%). In addition, the sampling distribution appears to be slightly skewed to the right; this is confirmed by the normal probability plot. Therefore, using the BCa confidence interval of 5.11 to 9.26 would be best. Note that it's not symmetric about the point estimate of 6.77.

```
> boot.ci(bsize)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS Based on 10000
bootstrap replicates
CALL : boot.ci(boot.out = bsize)
Intervals : Level      Normal              Basic
95%   ( 4.461,  8.761 )   ( 4.119,  8.361 )

Level     Percentile            BCa
95%   ( 5.178,  9.419 )   ( 5.109,  9.256 )
Calculations and Intervals on Original Scale Warning message: In
boot.ci(bsize) : bootstrap variances needed for studentized
intervals
```

Using the boot command in R

The format of the `boot` command in R is best illustrated by looking at the three examples from

this handout:

```
# Example 1: bootstrapping a median
> y <- c(47.9, 48.5, 49.8, 49.8, 50.2, 50.6, 50.8, 50.9, 51.1, 51.2, 51.5,
+      52.7, 53.0, 53.1, 54.5, 54.7, 55.8, 55.9)
> bmed <- function(x,i) median(x[i])
> b <- boot(y,bmed,100000)

# Example 2: bootstrapping a ratio
> x <- c(.4,.48,.43,.42,.5,.46,.39,.41,.42,.44)
> y <- c(.021,.03,.025,.022,.033,.027,.019,.021,.023,.025)
> orange <- data.frame(x,y)
> b <- boot(orange,function(a,i) sum(a$y[i])/sum(a$x[i]),100000)

# Example 3: bootstrapping N-hat for a PPS sample:
> size <- c(8,4,28,4,3,8,6,8,8,28,4,12,6,9,30)
> bsize <- boot(size,function(x,i) 1/mean(1/x[i]),100000)
```

The first argument to the `boot` command is the name of the data set which is being bootstrapped. The data set can be a vector (as in Examples 1 and 3), a matrix, or a data frame (as in Example 2). The third argument is the number of bootstrap replications. The second argument is a function which computes the statistic of interest from the data. This function can either be defined separately (as in Example 1) or within the call to boot (as in Examples 2 and 3).

- The function you define must have two arguments: the first (you can call it `x`) must represent the original data set and the second argument (call it `i`) must represent the set of indices which indicates which rows of the original data set are in a particular bootstrap sample.
- The arguments to any function are internal to the function, meaning that it does not matter what you call the arguments; they're just placeholders for the values that will be passed to it when the boot command uses it. In Example 2, I called the first argument `a` only to avoid confusion with the variable called `x`.
- In Example 1, `median(x[i])` computes the median of the values in `x` indexed by `i`. In Example 2, the original data is a data frame with columns named `x` and `y` so the argument `a` is also a data frame with columns named `x` and `y`. Hence, `sum(a$y[i])/sum(a$x[i])` computes the estimated ratio for the bootstrap sample indicated by `i`.
- The boot command in R will use your function to compute the value of your statistic for each of the bootstrap samples it generates. If you want to see all 100000 values, they're stored in `b$t` where `b` is the name you used to store the bootstrap result.
- Remember that before you can use the `boot` command, you must issue the command `library(boot)` or load the library using **Packages...Load package...**

References on bootstrapping

There are numerous references on bootstrapping both online and in the library. Here are a few books.

1. Efron and Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993.

2. Dixon, Philip, "The Bootstrap and Jackknife: Describing the Precision of Ecological Indices," Chapter 13 in *Design and Analysis of Ecological Experiments*, S. Scheiner and J. Gurevitch, eds., Chapman and Hall, 1993.

3. Davison, A.C. and Hinckley, D.V., "Bootstrap Methods and their Applications," Cambridge University Press, 1997.

Regression Estimation

Recall that the method of ratio estimation is appropriate when the response variable y is linearly related to some auxiliary variable x, and the value of $y = 0$ when $x = 0$. Sometimes, there is a linear relationship between the response y and an auxiliary variable x such that when $x = 0$, the value of y does not equal zero. In such cases, the method of regression estimation can be employed. Regression estimation requires population information on x, either the population mean μ_x or total τ_x.

- For example, if y and x are both positive variables but are negatively associated, such as weight of a car (x) and mpg (y), then the relationship could be linear, but does not go through the origin.
- As another example, suppose y = the sale value of a home in Missoula county, and x = the appraisal value of the home in the previous tax year. Although we certainly expect y & x to be related (perhaps linearly), we have no information about the relationship when x is near zero, and no reason to believe the relationship, if linear, should be forced to go through the origin.
- Regression estimation can be applied to more situations than just simple linear regression. It can accommodate more than one auxiliary variable or higher order relationships such as quadratic ones. We will only consider simple linear regression estimation here; extensions to multiple linear regression models are straightforward.

The Linear Regression Estimator

Notation:

$$\begin{aligned} y_i &= \text{response variable on the } i^{th} \text{ unit,} \\ x_i &= \text{auxiliary variable on the } i^{th} \text{ unit,} \\ \mu_y, \tau_y &= \text{mean and total of the y-values (book just uses } \mu, \tau), \\ \mu_x, \tau_x &= \text{mean and total of the x-values.} \end{aligned}$$

The population total and mean for the y-values are given by:

τ_y =

μ_y =

- The regression estimator for μ_y is: $\boxed{\widehat{\mu}_L = a + b\mu_x}$, where, b and a are the least squares estimators of the linear regression of y on x:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ (slope)}, \quad a = \overline{y} - b\overline{x} \text{ (intercept)},$$

- The estimator may also be written as $\boxed{\widehat{\mu}_L = a + b\mu_x = \overline{y} - b\overline{x} + b\mu_x = \overline{y} + b(\mu_x - \overline{x})}$

- Via the linearization method, the variance of $\widehat{\mu}_L$ is approximated by:

$$\text{Var}(\widehat{\mu}_L) \approx \frac{(N-n)}{Nn(N-1)} \sum_{i=1}^{N} (y_i - A - Bx_i)^2, \text{ where:}$$

$$B = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2}, \quad A = \mu_y - B\mu_x,$$

  the population slope and intercept.

- Since A and B are unknown population parameters, the estimated variance of $\widehat{\mu}_L$ is given by:

$$\widehat{\text{Var}}(\widehat{\mu}_L) = \frac{(N-n)}{Nn(n-2)} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

- A normal-based confidence interval for μ_y can be obtained as usual, by appealing to the finite population CLT, although the resulting CI's have been shown to be somewhat conservative (Scott & Wu, 1981).

Regression Estimation of the Total τ_y:

$$\begin{aligned} \widehat{\tau}_L &= N\widehat{\mu}_L = N(a + b\mu_x) = N\overline{y} + b(N\mu_x - N\overline{x}) = N\overline{y} + b(\tau_x - N\overline{x}), \\ \text{Var}(\widehat{\tau}_L) &= N^2 \cdot \text{Var}(\widehat{\mu}_L), \quad \widehat{\text{Var}}(\widehat{\tau}_L) = N^2 \cdot \widehat{\text{Var}}(\widehat{\mu}_L). \end{aligned}$$

Ratio and Regression Estimation in R

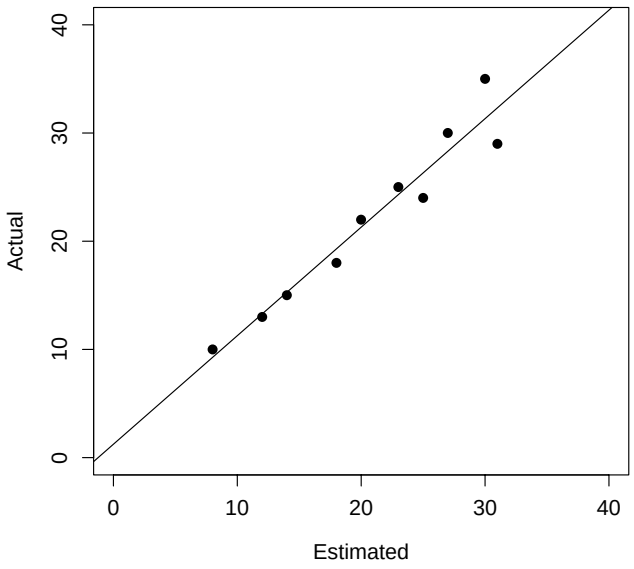
To illustrate the use of R in performing ratio and regression estimation, consider the following example.

Example 1: An investigator wishes to estimate the total number of trees on a 250-acre plantation. She divides the plantation into 1000 1/4-acre plots. She has aerial photographs from which she can easily estimate the total number of trees on each plot. She counts the actual number of trees on an SRS of 10 plots in order to calibrate her estimates from the photographs. Based on the aerial photographs, she estimates there are a total of 23,100 trees on the whole plantation or 23.1 per plot. For the SRS of ten plots, she finds the following:

Plot	1	2	3	4	5	6	7	8	9	10
Actual # Trees	25	15	22	24	13	18	35	30	10	29
Photo Estimate	23	14	20	25	12	18	30	27	8	31

A scatterplot of the data is below. Does it appear that a ratio estimate is appropriate or should a regression estimate be used (or neither)?

```
> x <- c(23,14,20,25,12,18,30,27,8,31)
> y <- c(25,15,22,24,13,18,35,30,10,29)
> plot(x,y,xlim=c(0,40),ylim=c(0,40),pch=16,xlab="Estimated",ylab="Actual",
  cex=1.2,cex.lab=1.2,cex.axis=1.2)
> reg <- lsfit(x,y)  # least-squares fit
> abline(reg)  # add least squares line to plot
```



Treating the photo estimate as an auxiliary variable, suppose we first use a ratio estimate to estimate the total number of trees τ_y. First, estimate R, the ratio of the total actual number of trees to the photo estimate. The SE of r is also estimated, although we really are not interested in R itself.

```
> r <- mean(y)/mean(x)
> r
[1] 1.0625
> sr2 <- (1/9)*sum((y-r*x)^2)
> sr2
[1] 4.225694
> sqrt((990/1000)*sr2/(10*mean(x)^2))   # if the mean of x were unknown
[1] 0.03109591
```

$$\left(s_r^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - r x_i)^2 \right)$$

$$\left(\mathrm{SE}(r) = \sqrt{\left(\frac{N-n}{N} \right) \frac{1}{\overline{x}^2} \cdot \frac{s_r^2}{n}} \right)$$

```
> se.r <- sqrt((990/1000)*sr2/(10*23.1^2)) # Use this since mux = 23.1 is known
> se.r
[1] 0.02799978
```

$$\left(\mathrm{SE}(r) = \sqrt{\left(\frac{N-n}{N}\right)\frac{1}{\mu_x^2}\cdot\frac{s_r^2}{n}}\right)$$

To compute a 99% confidence interval for r:

```
> c(r - qt(.995,9)*se.r,r + qt(.995,9)*se.r)
[1] 0.9715053 1.1534947
```

To estimate μ_y, the average number of trees per plot:

```
> muhat <- r*(23.1)
> muhat
[1] 24.54375
> se.muhat <- sqrt((990/1000)*sr2/10)
> se.muhat
[1] 0.646795
```

$$(\widehat{\mu}_y = r\cdot\mu_x)$$

$$\left(\mathrm{SE}(\widehat{\mu}_y) = \sqrt{\frac{N-n}{N}\cdot\frac{s_r^2}{n}}\right)$$

To estimate τ_y, the total number of trees:

```
> tauhat <- r*23100
> tauhat
[1] 24543.75
> se.tauhat <- 1000*se.muhat
> se.tauhat
[1] 646.795
```

$$(\widehat{\tau}_y = N\cdot\widehat{\mu}_y)$$

$$(\mathrm{SE}(\widehat{\tau}_y) = N\cdot\mathrm{SE}(\widehat{\mu}_y))$$

A 99% CI for τ_y:

```
> c(tauhat - qt(.995,9)*se.tauhat,tauhat + qt(.995,9)*se.tauhat)
[1] 22441.77 26645.73
```

Now, consider obtaining a regression estimate of μ_y and τ_y. First, perform a linear regression of y on x:

```
> reg <- lsfit(x,y)
> reg$coef
 Intercept         X
  1.239003 1.002933
> reg$residual
 [1]  0.6935484 -0.2800587  0.7023460 -2.3123167 -0.2741935 -1.2917889  3.6730205
 [8]  1.6818182  0.7375367 -3.3299120
> muhat <- mean(y)+reg$coef[2]*(23.1-mean(x))
> muhat
        X
 24.40674
```

$$(\widehat{\mu}_y = \overline{y} + b(\mu_x - \overline{x}))$$

```
> se.muhat <- sqrt((990/1000)*(1/(10*8))*sum(reg$residual^2))
> se.muhat
[1] 0.6683408
```

$$\left(\mathrm{SE}(\widehat{\mu}_y) = \sqrt{\left(\frac{N-n}{N}\right) \cdot \frac{1}{n(n-2)} \sum_{i=1}^{n} \widehat{e}_i^2} \right)$$

To estimate τ_y:

```
> tauhat <- 1000*muhat
> tauhat
        X
 24406.74
> se.tauhat <- 1000*se.muhat
> se.tauhat
[1] 668.3408
```

A 99% confidence interval for τ_y is given by:

```
> c(tauhat - qt(.995,8)*se.tauhat,tauhat + qt(.995,8)*se.tauhat)
       X        X
22164.20 26649.29
```

Note that the regression and ratio estimates are very similar, but the standard error for the regression estimate is higher than for the ratio estimate. This is caused by the fact that the least squares line goes almost through the origin and that estimating an extra parameter for the regression estimate has not helped much. Based on the scatterplot and the fact that a linear relationship through the origin seems very plausible,, we might want to stick with the ratio estimate.

Some R Functions for Ratio and Regression Estimation:
Functions for performing ratio and regression estimation were written and can be found on the course web page under the names "ratio.R" and "regress.R" as script files. To use these functions, simply run the script files (which places their definitions into the memory for R), and then call them, as illustrated below.

To illustrate the use of these functions (both given below), consider Example 1, the tree counting example. Recall that there are 1000 plots with the mean number of trees μ_x estimated from aerial photos as 23.1. Ten random plots are ground-truthed.

```
> x <- c(23,14,20,25,12,18,30,27,8,31)   # estimated no. of trees from photo
> y <- c(25,15,22,24,13,18,35,30,10,29)  # actual number of trees
> ratio.est(x,y,23.1,1000)
r= 1.0625   SE= 0.02799978
mu-hat= 24.54375   SE= 0.646795
tau-hat= 24543.75   SE= 646.795
```

A regression estimate of the mean number of trees per plot, μ_y, can be found as:

```
> regr.est(x,y,23.1,1000)
mu-hat= 24.40674   SE= 0.6683408
tau-hat= 24406.74   SE= 668.3408
```

The functions:

```
ratio.est <- function(x, y, mux = NA, N = NA) {
 # estimate of a ratio and ratio estimate of population mean and total.
 # x is auxiliary variable, y is response, mux is population mean
 # of x (xbar is used if no value is given),
 # N is population size (assumed infinite if no value given),
    if(length(x) != length(y)) stop("x and y must be same length")
    n <- length(x)
    fpc <- 1
    if(!is.na(N))
        fpc <- (N - n)/N
    r <- sum(y)/sum(x)
    sr2 <- (1/(n - 1)) * sum((y - r * x)^2)
    if(is.na(mux)) mx <- mean(x) else mx <- mux
    cat("r=", r, "  SE=", sqrt((fpc * sr2)/(mx^2 * n)), "\n")
    if(!is.na(mux))
        cat("mu-hat=", r * mux, "  SE=", sqrt((fpc * sr2)/n), "\n")
    if(!is.na(N) & !is.na(mux))
        cat("tau-hat=", N*r * mux, "  SE=", N*sqrt((fpc * sr2)/n), "\n")
}

regr.est <- function(x, y, mux, N = NA) {
 # regression estimator of a population mean and total.
 # x is auxiliary variable, y is response, mux is population mean
 # of x, N is population size (assumed infinite if no value given).
    if(length(x) != length(y)) stop("x and y must be same length")
    n <- length(x)
    fpc <- 1
    if(!is.na(N))
        fpc <- (N - n)/N
    ab <- lsfit(x, y)
    cat("mu-hat=", ab$coef[1] + ab$coef[2] * mux, "  SE=", sqrt((fpc *
            sum(ab$residual^2))/(n * (n -2))), "\n")
    if(!is.na(N))
    cat("tau-hat=", N*(ab$coef[1] + ab$coef[2] * mux), "  SE=", N*sqrt((fpc *
```

```
            sum(ab$residual^2))/(n * (n -2))), "\n")
}
```

Ratio and Regression Estimation - Some Further Notes

1. Unequal Probability Sampling

   Ratio and regression estimators can be applied in the situation of unequal probability sampling (see sections 7.5 and 8.2 of Thompson).

   - The generalized ratio estimator of the population total τ_y with unequal probability sampling is

     $$\widehat{\tau}_G = \frac{\widehat{\tau}_y}{\widehat{\tau}_x}\tau_x$$

     where $\widehat{\tau}_y$ and $\widehat{\tau}_x$ are the Horvitz-Thompson estimators of τ_y and τ_x, respectively. As a ratio of two unbiased estimators, this estimator is not unbiased. The estimated variance is given by eq. 7.10 on p. 103 of Thompson (2nd ed: eq. 10 on p. 78).

   - The generalized ratio estimator can also be used to derive an estimator of μ_y with unequal probability sampling when N is unknown by letting x be equal to 1 for all units:

     $$\widehat{\mu}_G = \frac{\widehat{\tau}_\pi}{\widehat{N}} = \frac{\sum_{i=1}^{\nu} y_i/\pi_i}{\sum_{i=1}^{\nu} 1/\pi_i}$$

     where ν is the number of distinct units in the sample. We already looked at this estimator for PPS sampling in the notes on "Unequal Probability Sampling" (eq. (1) on p. 33) using Hansen-Hurwitz estimators for the numerator and denominator, so this is a generalization to any unequal probability sampling plan. Thompson notes on pp. 103-104 (2nd ed: p. 78) that this estimator is sometimes a useful alternative even when N is known if there is not a linear relationship between the inclusion probabilities and the y values; $\widehat{\mu}_G$ may have a smaller variance than the estimator $\widehat{\tau}_\pi/N$, as for the elephant example he discusses. The analogous estimator of τ_y is then

     $$\widehat{\tau}_G = N\widehat{\mu}_G = N\frac{\widehat{\tau}_\pi}{\widehat{N}} = N\frac{\sum_{i=1}^{\nu} y_i/\pi_i}{\sum_{i=1}^{\nu} 1/\pi_i}.$$

     Note that we "adjust" the estimator of τ_y by the ratio $N/\widehat{N}$ with the idea being that $\widehat{N}$ will tend to overestimate N when $\widehat{\tau}_\pi$ overestimates τ_y and vice-versa.

   - These same ideas can be extended to regression estimation yielding generalized regression estimators of μ_y and τ_y (section 8.2 of Thompson).

2. Design and Model Approaches to Sampling

   The ratio and regression estimators are not design-unbiased. What does that mean? That means if we view the population of y values $y_1, \ldots, y_N$ as fixed with $\mu_y = \sum_{i=1}^{N} y_i/N$, then it is not necessarily true that $\mathrm{E}(\widehat{\mu}_r) = \mu_y$ or that $\mathrm{E}(\widehat{\mu}_L) = \mu_y$ (we have to say "not necessarily true" because it may be true for specific cases). This is shown by Thompson for the ratio estimator in an example in section 7.2. The bias is usually small if there is a linear relationship between the x and y variables.

However, the ratio and regression estimators are model-unbiased if we assume the right model for the population. What is the model-based approach?

- In the model-based approach to sampling, we consider the y values in the population to be random variables denoted by $Y_1, \ldots, Y_N$. That is, they are only one possible realization of a process that generated the population. Therefore, the population total $\tau_y = \sum_{i=1}^{N} Y_i$ and the population mean $\mu_y = \tau_y/N$ are also random variables (this is very important to remember). What it means, therefore, for an estimator of the population total to be model-unbiased, is if the expected value of the estimator is equal to the expected value of the population total (and analogously for an estimator of the population mean). Since the population total and mean are not fixed quantities but random variables under the model-based approach, we might more accurately say that our estimators are predictors of these random quantities.
- As an example, suppose we have an auxiliary variable x. One way to model y is to consider the x values $x_1, \ldots, x_N$ to be fixed, but the y values to be random variables:

$$Y_i = \beta x_i + \epsilon_i$$

where the ϵ_i's are independent random variables with $\mathrm{E}(\epsilon_i) = 0$. This is the linear regression model without an intercept. Under this model,

$$\mathrm{E}(Y_i) = \mathrm{E}(\beta x_i + \epsilon_i) = \beta x_i + \mathrm{E}(\epsilon_i) = \beta x_i.$$

Now consider the ratio estimator of μ_y:

$$\widehat{\mu}_r = r\mu_x = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}\,\mu_x$$

(note that we have replaced the fixed values $y_1, \ldots, y_n$ by the random variables $Y_1, \ldots, Y_n$ in the model-based approach). Then, regardless of how our sample is chosen:

$$\begin{aligned}\mathrm{E}(\widehat{\mu}_r) &= \mathrm{E}\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}\,\mu_x\right) = \frac{\mathrm{E}(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} x_i}\,\mu_x \text{ (since the } x_i\text{'s and } \mu_x \text{ are fixed)} \\ &= \frac{\sum_{i=1}^{n} \mathrm{E}(Y_i)}{\sum_{i=1}^{n} x_i}\,\mu_x = \frac{\sum_{i=1}^{n} \beta x_i}{\sum_{i=1}^{n} x_i}\,\mu_x = \beta\mu_x.\end{aligned}$$

Note also that the expected value of the population mean is

$$\mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}(Y_i) = \frac{1}{N}\sum_{i=1}^{N} \beta x_i = \beta\mu_x.$$

Since $\mathrm{E}(\widehat{\mu}_r) = \mathrm{E}(\tau_y)$, $\widehat{\mu}_r$ is model-unbiased for predicting the population mean under the model above.

- Note that the randomness in the model-based approach comes from the model for the Y_i's and not from how the sample was chosen. Therefore, model-unbiasedness holds regardless of how the sample was chosen, randomly or not. However, it depends crucially on the model assumed. In the example here, this means not only that $\mathrm{E}(Y_i) = \beta x_i$ but also that the ϵ_i are independent, regardless of how the units were chosen for the sample.
- In the design-based approach, all the randomness comes from how the sample was chosen since $y_1, \ldots, y_N$ are fixed. Therefore, design-unbiasedness of an estimator depends entirely on how the sample was chosen (e.g., SRS or whatever sampling plan is assumed).
- Extension to the regression estimator: model-unbiasedness of the regression estimator holds under the linear regression model with an intercept:

  $$Y_i = \alpha + \beta x_i + \epsilon_i$$

  where the ϵ_i's are independent random variables and $\mathrm{E}(\epsilon_i) = 0$.
- Implications for sampling design: if we truly believe in the regression model with independent errors, then we can select our sample any way we like and still preserve model-unbiasedness. In particular, we could select units in such a way to minimize the variance of our estimators. For the linear regression model without an intercept, regression theory tells us to choose the units with the largest x-values. For the linear regression model with an intercept, it tells us to choose the units with the smallest and largest x-values. Unfortunately, these designs do not allow us to assess the appropriateness of the model and could lead to large errors if the model turns out to be wrong. A compromise would be to choose units with a wide range of x-values, perhaps by stratified random sampling, stratifying by x. Most independent observers would be skeptical of a selection process without some element of randomness in it.

3. Least-squares estimators in the model-based approaches

The regression estimators of α and β in the linear regression model in Chapter 8 are the usual least-squares estimators for linear regression. However, the ratio estimator of β in the regression model without an intercept (Chapter 7) is $\overline{y}/\overline{x} = \sum_{i=1}^n y_i / \sum_{i=1}^n x_i$ which is not the least squares estimator. The least squares estimator is

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \tag{3}$$

Why is that and why don't we use the least-squares estimator in the ratio estimator?

- First, we can show that the least-squares estimator of β in the regression model without an intercept also gives model-unbiased estimators of the population mean and total, so neither the ratio or least-squares estimator has an advantage in that respect.

- The model unbiasedness of the ratio and regression estimators (as well as the least squares estimator for the model without intercept) of the population mean and total depends only on the assumption that the ϵ_i's are independent with mean 0. We do not need to assume that they have identical distributions. As Thompson discusses on p. 109 (2nd ed: p. 84), the ratio estimator of Chapter 7 is the best linear unbiased estimator (BLUE) of β if we assume that $\text{Var}(\epsilon_i)$ is proportional to x_i while the least squares estimator of equation (3) above is the BLUE if we assume that the variances are all equal. Variance proportional to x is actually a quite reasonable assumption in many ratio estimation problems, or at least more reasonable than constant variance. For example, in the trees example on p. 68 of these notes (where a photo estimate of the number of tress in a plot is used as the auxiliary variable), it seems reasonable to assume that that the larger the number of trees in a plot, the more the actual number might vary. That is, if we had a set of plots all with photo estimates of 10 trees, we would expect the actual numbers to vary less than for a set of plots all with photo estimates of 50 trees. Therefore, the usual ratio estimator of Chapter 7 is a reasonable default estimator.

Stratified Random Sampling (Chapter 11)

This chapter 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 the Chapter 2 notes that this situation minimizes the variability in the stratified random sampling estimator.

- Examples: Landscapes - stratified by habitat characteristics,
  People - stratified by sex, income, etc.

Notation:

$$\begin{aligned}
N_h &= \text{the population size in stratum } h,\ h = 1, 2, \ldots, L, \\
N &= \sum_{h=1}^{L} N_h = \text{the total population size}, \\
n_h &= \text{the sample size in stratum } h,\ h = 1, 2, \ldots, L, \\
n &= \sum_{h=1}^{L} n_h = \text{the total sample size}, \\
y_{hi} &= \text{the } i^{th} \text{ observation in the } h^{th} \text{ stratum}, \\
\tau_h &= \sum_{i=1}^{N_h} y_{hi} = \text{the total of the observations in stratum } h, \\
\tau &= \sum_{h=1}^{L} \tau_h = \text{the overall total}, \\
\mu_h &= \tau_h / N_h = \text{the mean response in stratum } h, \\
\mu &= \tau / N = \text{the overall mean response}.
\end{aligned}$$

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

- If $\widehat{\tau}_h$ is an unbiased estimator of τ_h, $h = 1, \ldots, L$, then $\widehat{\tau}_{st} = \sum_{h=1}^{L} \widehat{\tau}_h$ 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:

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

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

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

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

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

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

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

- The results derived above are true with any sampling plan 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: $\widehat{\tau}_h = N_h\overline{y}_h$ ($\widehat{\mu}_h = \overline{y}_h$), where $\overline{y}_h$ is the sample mean in stratum h.

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

$$\boxed{\text{Var}(\widehat{\tau}_{st})} = \sum_{h=1}^{L}\text{Var}(\widehat{\tau}_h) = \sum_{h=1}^{L} N_h^2\left(\frac{N_h - n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} = \boxed{\sum_{h=1}^{L} N_h(N_h - n_h)\frac{\sigma_h^2}{n_h}}$$

$$\Rightarrow \quad \boxed{\widehat{\text{Var}}(\widehat{\tau}_{st}) = \textstyle\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}.}$$

$$\boxed{\widehat{\mu}_{st}} = \overline{y}_{st} =$$

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

Example 1: Survey of TV Habits. 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.

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)
$\overline{y}_h$	33.90	25.12	19.00	
s_h	5.95	15.24	9.36	
$\widehat{\tau}_h$	5254.5	1557.4	1767.0	$(N_h\overline{y}_h)$

$$\begin{aligned}\widehat{\tau}_{st} &= \widehat{\tau}_1 + \widehat{\tau}_2 + \widehat{\tau}_3 = \underline{8578.9},\\ \overline{y}_{st} &= \frac{\widehat{\tau}_{st}}{N} = \frac{8578.9}{310} = \underline{27.7}\end{aligned}$$

Other way:

$$\begin{aligned}\overline{y}_{st} &= \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\overline{y}_h\\ &= \frac{155}{310}(33.90) + \frac{62}{310}(25.12) + \frac{93}{310}(19) = .5(33.90) + .2(25.12) + .3(19) = \underline{27.7}.\end{aligned}$$

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

A 95% confidence interval for μ is given by:

$$\overline{y}_{st} \pm t\,\mathrm{SE}(\overline{y}_{st}) = 27.7 \pm (2.079)(1.40) = 27.7 \pm 2.91 = \underline{(24.79, 30.61)}.$$

- 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 (11.4) on p. 145 of Thompson (2nd ed: eq. (4) on p. 121), the approximate degrees of freedom are:

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

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

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

- 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 section 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: $\boxed{n_h = \frac{n}{L}}$.
2. Proportional: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes: $\boxed{n_h = \left(\frac{N_h}{N}\right) n}$.
   - 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:
$$\boxed{n_h = \frac{n N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k}.} \tag{4}$$
   - Such an allocation rule will minimize $\text{Var}(\overline{y}_{st})$ for a given n.
   - This allocation can be derived (Section 11.8 of Thompson) by the Lagrange multiplier method: find the values of $n_1, \ldots, n_L$ which minimize $\text{Var}(\overline{y}_{st})$ subject to the constraint $n_1 + \ldots + n_L = n$.
   - 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, σ_h, $h = 1, \ldots, L$ (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:

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

Example 1 (continued): Survey of TV Habits.
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 $n_1, \ldots, n_L$ subject to the constraint that the total cost is c. Via constrained optimization, the resulting optimum allocation is given by:

$$n_h = (c - c_0)\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{k=1}^{L} N_k\sigma_k\sqrt{c_k}}. \tag{5}$$

- 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 section 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:

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

where:

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

Solving this margin of error equation for n leads to:

$$n = \frac{\displaystyle\sum_{h=1}^{L}\frac{N_h^2\sigma_h^2}{w_h}}{\dfrac{N^2d^2}{z^2} + \displaystyle\sum_{h=1}^{L}N_h\sigma_h^2} = \frac{\displaystyle\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\frac{\sigma_h^2}{w_h}}{\dfrac{d^2}{z^2} + \dfrac{1}{N}\displaystyle\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\sigma_h^2}. \tag{6}$$

The rightmost expression in (6) 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 $\text{Var}(\overline{y}_{st})$.

The formula in (6) yields the following for the three allocation schemes we have talked about:

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

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

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

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

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

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

3. **Total sample size needed with optimum allocation (equal costs)**:
   $n_h = \dfrac{nN_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k}$ so $w_h = \dfrac{N_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k}$ and

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

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

   - 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 in equation (5) on p. 83 of these notes (and on page 147 of Thompson; 2nd ed: p. 123), $n_h = c^* w_h$ where:

$$w_h = \frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{k=1}^{L} N_k\sigma_k\sqrt{c_k}}, \text{ and } c^* = \frac{\left[\sum_{h=1}^{L} N_h\sigma_h\sqrt{c_h}\right]^2}{\frac{N^2d^2}{z^2} + \sum_{h=1}^{L} N_h\sigma_h^2} = \frac{\left[\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\sigma_h\sqrt{c_h}\right]^2}{\frac{d^2}{z^2} + \frac{1}{N}\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\sigma_h^2}. \quad (7)$$

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 1 (continued): 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                 # total sample size
[1] 141.3878
> n/3               # sample size for each stratum
[1] 47.12928
```

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

```
# Proportional Allocation
# =======================
> n <- N*sum(Nh*s^2)/(N^2*d^2/z^2+sum(Nh*s^2))
> n                 # total sample size
[1] 163.7988
> n*Nh/N            # sample size by stratum
[1] 81.89941 32.75976 49.13965
```

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

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

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

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

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)
```

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

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

- 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.

Example 2: Allocation for Stratified Random Sampling: 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 equation (7) on p. 86 of these notes, 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 this equation 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.
```

$$\left(\begin{array}{c} c^* = \dfrac{\left[\sum_{h=1}^{L} N_h \sigma_h \sqrt{c_h}\right]^2}{\dfrac{N^2 d^2}{z^2} + \sum_{h=1}^{L} N_h \sigma_h^2} \\ \text{with } N^2 d^2 \to N^2 \left(\frac{d}{N}\right)^2 = d^2 \end{array} \right)$$

- Using the optimum allocation formula (5) on page 83 of these notes (also on p. 147 of Thompson; 2nd ed: p. 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 143 of Thompson (2nd ed: page 119), is $\underline{3930}$ 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.
[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 $\overline{\tau}_{st}$ on p. 142 of Thompson (2nd ed: p. 119), namely:

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

  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 n3=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 formula (7) on p. 85 of the notes to compute the total sample size required, remembering to replace d by d/N. 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 (4) on p. 82 of the notes $\left(n_h = n \frac{N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k} \right)$ 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 n	Total Cost	Allocation 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

Stratified Random Sampling for Proportions

To illustrate allocation in stratified random sampling for estimating a proportion, we'll consider an example below. First, recall from the material on estimating proportions (Chap. 5 of Thompson and page 26 of these notes), the variance of the sample proportion $\widehat{p}$ based on an SRS of size n is

$$\text{Var}(\widehat{p}) = \left(\frac{N-n}{N}\right)\frac{p\,(1-p)}{n}$$

where p is the population proportion. The variance is estimated by

$$\widehat{\text{Var}}(\widehat{p}) = \left(\frac{N-n}{N}\right)\frac{\widehat{p}\,(1-\widehat{p})}{n-1}.$$

Now, suppose we have a stratified random sample from which we are going to estimate a population proportion. If p_h is the true proportion in stratum h and, as before, N_h and n_h are the stratum size and the sample size, then the stratified estimator of p with L strata is

$$\widehat{p}_{st} = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\widehat{p}_h$$

with variance

$\text{Var}(\widehat{p}_{st}) =$

The variance would be estimated by

$\widehat{\text{Var}}(\widehat{p}_{st}) =$

Example 3: Suppose that a large area is divided into 1000 quadrats in three strata. In stratum 1, there are 600 quadrats and it is guessed that about 50% of these contain plants of species X. In stratum 2, there are about 370 quadrats and it is guessed that about 30% contain species X. In stratum 3, there are 30 quadrats and it is guessed that about 10% contain species X. A survey is to be conducted of $n = 100$ quadrats to estimate the proportion of all 1000 quadrats that contain species X. Compare the accuracy of simple random sampling, stratified sampling with proportional allocation, and stratified sampling with optimal allocation.

Based on the guessed percentages, the proportion of all quadrats containing species X is about $p = (600 \times 0.5 + 370 \times 0.3 + 30 \times 0.1)/1000 = 0.414$. The standard deviation of the the sample proportion $\widehat{p}$ based on an SRS of 100 plots would then be estimated to be:

$\mathrm{SD}(\widehat{p}) =$

For stratified random sampling with proportional allocation, the sample sizes will be:

$n_1 =$ $n_2 =$ $n_3 =$

The standard deviation of the stratified estimator (based on the guessed proportions) is then

$\mathrm{SD}(\widehat{p}_{st}) =$

The optimum equal-cost allocation is

$$n_h = \frac{nN_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k},\ h = 1, 2, \ldots, L$$

where σ_h is the standard deviation of the 0-1 values in stratum h. From Chapter 5, this is equal to

$$\sigma_h = \sqrt{\frac{N_h}{N_h - 1}p_h(1 - p_h)}.$$

Calculating these quantities for the example (using the guessed proportions):

$\sigma_1 =$

$\sigma_2 =$

$\sigma_3 =$

The denominator in the allocation formula is therefore

$\sum_{k=1}^{3} N_k \sigma_k =$

Hence, the optimal allocation is

$n_1 =$ $n_2 =$ $n_3 =$

Rounding the sample sizes to integers, the standard deviation of the stratified estimator is then

$\text{Var}(\widehat{p}_{st}) =$

Conclusions? Is there an advantage to stratified sampling over SRS? Is there an advantage to optimal allocation over proportional?

Some remarks:

1. If the cost to sample a unit is the same for all strata, then the gain from stratified random sampling is small unless the proportions vary greatly across strata.
2. Optimum allocation is little better than proportional allocation unless some strata proportions are outside the range 0.1 to 0.9. This is because the standard deviation of 0-1 data is proportional to $\sqrt{p(1-p)}$ which varies by less than a factor of 2 over the range p =0.1 to 0.9.

More on Stratified Random Sampling

This section has some additional information on stratified sampling, compares stratified random sampling to simple random sampling under the setting of proportional allocation, and examines what is known as post-stratification.

Comparison of stratified sampling to SRS

- Recall that at the beginning of the semester in a small example with a population of size 4, we looked at properties of the mean estimator under both simple random and stratified random sampling. We found that if the strata were heterogeneous between and homogeneous within, then the stratified mean had smaller variance than the SRS-based mean. (We compared the variances since both estimators are unbiased for the population mean.)
- However, we also saw that if the strata are poorly defined, then an SRS can give a smaller variance than a stratified random sample for estimating a mean. In addition, a poor allocation can also make stratified sampling worse (for example, allocating most of the observations to the smallest and least variable strata). Hence, in general, one method will not always be superior to the other; it depends on the degree of heterogeneity between the strata defined and the allocation used.

It is possible to make a more formal comparison of SRS with stratified random sampling in the special case of proportional allocation. First, think about the parallels between the data for an ANOVA, where we are comparing responses across groups, to the data from a stratified random sample where we obtain data from separate groups (strata). Our goal is different; in the latter case, we know (or suspect) there are differences between the strata and our primary interest is not necessarily in that comparison. We're exploiting the stratum differences to estimate a population mean or total. However, the data have the same structure as in an ANOVA. Recall from ANOVA that we can partition the total variability in the whole sample (sum-of squares total or SST) into two parts: the variability between the group means (sum-of-squares between or SSB) and the pooled variability within groups (sum-of-squares within or SSW). An analogous decomposition can be made for finite populations:

$$\begin{aligned}
\text{SSB} &= \sum_{h=1}^{L}\sum_{i=1}^{N_h}(\mu_h-\mu)^2 = \sum_{h=1}^{L}N_h(\mu_h-\mu)^2 \\
\text{SSW} &= \sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\mu_h)^2 = \sum_{h=1}^{L}(N_h-1)\sigma_h^2 \\
\text{SST} &= \sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2 = (N-1)\sigma^2
\end{aligned}$$

where it can be shown that SST = SSB + SSW. Recall the forms of the variances for the means in

both SRS and stratified random sampling:

$$\begin{aligned}
\text{Var}(\overline{y}) &= \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} = \left(\frac{N-n}{N}\right)\frac{\text{SST}}{n(N-1)} = \frac{1-f}{n(N-1)}\ \text{(SSW+SSB)} \\
\text{Var}(\overline{y}_{st}) &= \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h-n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} = \frac{1}{N^2}\sum_{h=1}^{L}N_h(N_h-n_h)\frac{\sigma_h^2}{n_h}
\end{aligned}$$

where $f = n/N$. Now assume $n = n_1 + \ldots + n_L$ (same total sample size) and also assume proportional allocation. Then $f = \dfrac{n_h}{N_h} = \dfrac{n}{N}$ and so using the substitution $n_h = fN_h$,

$$\begin{aligned}
\text{Var}(\overline{y}_{st}) &= \frac{1}{N^2}\sum_{h=1}^{L}N_h(N_h-n_h)\frac{\sigma_h^2}{n_h} = \frac{1}{fN^2}\sum_{h=1}^{L}(N_h-fN_h)\sigma_h^2 \\
&= \frac{1-f}{nN}\sum_{h=1}^{L}N_h\sigma_h^2 = \frac{1-f}{nN}\left(\text{SSW}+\sum_{h=1}^{L}\sigma_h^2\right)
\end{aligned}$$

Comparing this last expression to $\text{Var}(\overline{y})$ above, we can show that $\text{Var}(\overline{y}) < \text{Var}(\overline{y}_{st})$ only if

$$\text{SSB} = \sum_{h=1}^{L}N_h(\mu_h-\mu)^2 \ < \ \sum_{h=1}^{L}\left(1-\frac{N_h}{N}\right)\sigma_h^2.$$

Since the left-hand side is large when the strata population sizes are large and/or when the stratum means are very different, it turns out that this condition is rarely satisfied in practice except if the strata sizes are exceptionally small and their means are nearly identical. Hence, stratification with proportional allocation can rarely hurt; obviously, optimal allocation will help even more (if the relative variances can be accurately estimated). The tradeoff is the extra effort needed for stratified sampling and the extra information needed (we need to know the stratum sizes and be able to sample within each stratum).

Stratification when stratum means are not different: Generally, we think of stratification as being beneficial when the stratum means are very different. However, it can be beneficial when the stratum means are not different, but the stratum variances are and we use optimal allocation. Then, we sample more heavily in the strata with large variances which can reduce the variance of the estimator of the population mean as compared to an SRS of the same total size.

Post-Stratification: Post-stratification occurs when we take an SRS from the whole population, then decide afterwards to stratify by some auxiliary variable. For example, suppose we take a simple random sample (SRS) of adults in Missoula and estimate the proportion of adults who favor a state sales tax. After the fact, we might then decide to stratify by some variable such as age group.

- So we take an SRS of size n from the whole population, and then *note* that there are n_1 from stratum 1, ..., n_L from stratum L. So, the stratum sizes are not fixed ahead of time and are hence viewed as random.

- With SRS, we would estimate μ with $\overline{y}$. However, if we know the relative stratum sizes (N_h/N), then we should adjust this estimate for the proportions actually sampled. The post-stratification estimate is: $\overline{y}_{st} = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right) \overline{y}_h$. Note that this is just a weighted average of stratum means based on the relative stratum sizes.

- Under proportional allocation (the usual stratification situation), the variance of this post-stratification estimate of the mean μ is just the usual $\text{Var}(\overline{y}_{st})$ plus an additional variance component due to the variability in the stratum sizes. The approximate variance is given below (from equation (11.5) on page 148 of Thompson; 2nd ed: eq. (5) on p. 124).

$$\begin{aligned} \text{Var}(\overline{y}_{pst}) &\approx \frac{N-n}{nN} \sum_{h=1}^{L} \left(\frac{N_h}{N}\right) \sigma_h^2 + \frac{1}{n^2} \left(\frac{N-n}{N-1}\right) \sum_{h=1}^{L} \frac{N-N_h}{N} \sigma_h^2 \\ &= \text{Var}(\overline{y}_{st}) + \frac{1}{n^2} \left(\frac{N-n}{N-1}\right) \sum_{h=1}^{L} \frac{N-N_h}{N} \sigma_h^2 \ \text{ (using } n_h = nN_h/N). \end{aligned}$$

- To estimate $\text{Var}(\overline{y}_{st})$ for post-stratification, the recommendation is to ignore the second term above and use the same estimate as one would if the sample were pre-stratified. Thus, the recommendation is to carry out an analysis for a post-stratified sample just as if it were a pre-stratified sample with fixed stratum sample sizes.

- Post-stratification is like ratio or regression estimation but with a categorical auxiliary variable. We are taking advantage of the relationship of the response to an auxiliary variable in the estimation phase. It can be useful when we know the relative stratum sizes at the start, but it is inconvenient or impossible to sample separately within each stratum.

Cluster Sampling & Systematic Sampling (Chap. 12)

Recall that cluster sampling is where we first divide the population into "clusters," then select a simple random sample (SRS) of these clusters, and sample every unit within the selected clusters. This is a two-stage sampling plan, where we employ an SRS of clusters at the first stage and a census within these selected clusters at the second stage.

Systematic sampling is where we first select a random starting point in the population, and then sample every m^{th} unit beginning at that starting point. This also is a two-stage sampling plan, where we employ an SRS of size 1 from the list of potential starting points, and then census the sampling units at multiples of m units from the initial unit.

- Cluster sampling and systematic random sampling, as defined, are special cases of two-stage random sampling plans which will be discussed in the next chapter. The population is first partitioned into mutually exclusive groups, known as primary units, each of which contains a number of sampling units, known as secondary units. Selection occurs on the *primary* units, and then *every* secondary unit within a selected primary unit is sampled.
- This should be clear for cluster sampling. With systematic sampling, the primary units are groups of observations m units apart in the population list. For example, if we want every 5th member of a population, then there are 5 primary units corresponding to these sets of secondary units: 1,6,11,...; 2,7,12,...; 3,8,13,...; 4,9,14,...; and 5,10,15,... In systematic sampling, we generally take an SRS of one of these primary units and then examine every secondary unit in the primary unit selected.

Main Idea: The important point for these two sampling plans is that whenever a primary unit is selected, all secondary units within are sampled. In truth then, the primary units are the sampling units in a cluster or systematic sample, even though measurements or observations are actually made on the secondary units.

Thompson lists three special considerations for these two sampling plans which warrant further discussion and separate consideration (top of page 159; 2nd ed: pages 129 & 131).

1. In cluster sampling, the size of the cluster may serve as auxiliary information that may be used either in selecting clusters with unequal probabilities (PPS sampling) or in forming ratio estimators.
2. The size and shape of clusters may affect efficiency.
3. In systematic sampling, it is not uncommon to have a sample size of one; that is, a single primary unit.

After defining the relevant notation for these sampling plans, each of the special considerations above will be addressed by way of examples.

Notation: Consider taking a cluster sample from some population with response variable y. We let:

$$\begin{aligned}
N &= \text{the number of primary units (clusters) in the population,}\\
n &= \text{the number of primary units (clusters) in the sample,}\\
M_i &= \text{the \# of secondary units in the } i^{th} \text{ primary unit,}\\
M &= \text{the total \# of secondary units in the population} = \sum_{i=1}^{N} M_i\\
y_{ij} &= \text{the } y\text{-value of the } j^{th} \text{ secondary unit in the } i^{th} \text{ primary unit,}\\
y_i &= \sum_{j=1}^{M_i} y_{ij} = \text{the total of the } y\text{'s in the } i^{th} \text{ primary unit (cluster totals),}\\
\tau &= \sum_{i=1}^{N} y_i = \text{the total of the } y\text{'s in all the units,}\\
\mu &= \frac{\tau}{M} = \text{the mean of the } y\text{'s per secondary unit,}\\
\mu_1 &= \frac{\tau}{N} = \text{the mean of the } y\text{'s per primary unit.}
\end{aligned}$$

Example 1: A sociologist wants to estimate the average per capita income in a certain small city. As no list of resident adults is available, she decides that each of the city blocks will be considered one cluster. The clusters are numbered on a city map from 1 to 415, and the experimenter decides she has enough time and money to sample $n = 25$ clusters where every household will be interviewed within the clusters (blocks) chosen. The data on the next page give the number of residents and the total income for each of the 25 blocks sampled. [Problem taken from Scheaffer, Mendenhall, & Ott, *Elementary Survey Sampling*, page 248.] Given that $M = 2500$ residents, use these data to estimate the average per capita income in the city.

Notation:

$$\begin{aligned}
N &= 415 \text{ blocks, } n = 25 \text{ blocks, } M = 2500 \text{ residents,}\\
M_i &= \text{the number of residents in the } i^{th} \text{ block,}\\
M &= \text{the total number of residents in all 415 blocks,}\\
y_{ij} &= \text{the income of the } j^{th} \text{ resident in the } i^{th} \text{ block,}\\
y_i &= \text{the total income of all residents on the } i^{th} \text{ block,}\\
\tau &= \text{the total income of all residents in the city,}\\
\mu &= \text{the mean income per resident,}\\
\mu_1 &= \text{the mean income per block.}
\end{aligned}$$

Cluster i	Number of Residents, M_i	Total Income per Cluster, y_i	Cluster i	Number of Residents, M_i	Total Income per Cluster, y_i
1	8	\$192,000	14	10	\$98,000
2	12	\$242,000	15	9	\$106,000
3	4	\$84,000	16	3	\$100,000
4	5	\$130,000	17	6	\$64,000
5	6	\$104,000	18	5	\$44,000
6	6	\$80,000	19	5	\$90,000
7	7	\$150,000	20	4	\$74,000
8	5	\$130,000	21	6	\$102,000
9	8	\$90,000	22	8	\$60,000
10	3	\$100,000	23	7	\$78,000
11	2	\$170,000	24	3	\$94,000
12	6	\$86,000	25	8	\$82,000
13	5	\$108,000			
				$\sum_{i=1}^{25} M_i = 151$	$\sum_{i=1}^{25} y_i = \$2,658,000$

There are two basic ways to approach estimation of τ and μ:

- Unbiased estimation: treat the sample as an SRS of clusters (blocks in this example), each with response y_i and use the SRS formulas from Chap. 2. We ignore M_i, the cluster sizes. We can estimate τ and μ_1 (mean per cluster) in this way. If we know M, we can also estimate μ.
- Ratio estimation: If the primary unit total y_i is correlated with the cluster size M_i (such that we expect $y_i = 0$ when $M_i = 0$), then ratio estimators of τ and μ may be advantageous. Ratio estimators are biased but can have substantially smaller MSE than the unbiased estimators if there's a strong relationship between the cluster sizes and the cluster totals.

Unbiased estimation

- Estimate μ_1, the mean total income per block, by $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{25}[2658000] = \underline{\$106{,}320}$, with standard error
  $\text{SE}(\overline{y}) = \sqrt{\left(\frac{N-n}{N}\right)\frac{s_u^2}{n}} = \sqrt{\left(\frac{415-25}{415}\right)\frac{1898226667}{25}} = \underline{\$8,447}$,
  where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{25}(y_i - \overline{y})^2$ = the sample variance of the cluster totals (y_i's).

- Estimate τ, the total income for the whole city by $N\overline{y} = 415(106320) =$ \$44,122,800, with standard error

$$\text{SE}(\widehat{\tau}) = \sqrt{N(N-n)\frac{s_u^2}{n}} = \sqrt{415(415-25)\frac{1898226667}{25}} = \$3,505,584.$$

- If we know M (= 2500 here), we can estimate μ (average income per resident) by

$$\begin{aligned} \widehat{\mu} &= \frac{\widehat{\tau}}{M} = \frac{N}{M}\overline{y} = \frac{44122800}{2500} = \$17,649, \\ \text{SE}(\widehat{\mu}) &= \text{SE}\left(\frac{\widehat{\tau}}{M}\right) = \frac{1}{M}\text{SE}(\widehat{\tau}) = \frac{1}{2500}(3505584) = \$1,402. \end{aligned}$$

- To form confidence intervals with the above estimators, we would multiply the SE by a t value with $n-1$ degrees of freedom (24 df in this example).

Ratio Estimation

With cluster sampling, we have an auxiliary variable M_i (the number of residents on each block), so we may be able to take advantage of this to improve the SRS-based estimates of τ and μ by using ratio estimators (Chapter 7).

- Estimate μ by the sample ratio r:

$$\widehat{\mu}_r = r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = \frac{\text{Sample total income}}{\text{Sample total \# of residents}}$$

  For this example, $\widehat{\mu}_r = 2658000/151 =$\$17,603. The standard error is

$$\text{SE}(\widehat{\mu}_r) = \sqrt{\left(\frac{N-n}{N\mu_x^2}\right)\frac{s_r^2}{n}}, \text{ where } s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rM_i)^2$$

  where $\mu_x = \overline{M} = M/N$ is the the mean cluster size for the population. If $\overline{M}$ is unknown, then we can substitute $\overline{m}$= mean cluster size for the sample. For this example, where $\overline{M} = 2500/415 = 6.24$, we get SE($\widehat{\mu}_r$)=\$1,621 (see R code below for calculation).

- If we know M, then the ratio estimator of the population total τ is $\widehat{\tau}_r = Mr$ with $\text{SE}(\widehat{\tau}_r) = M\,\text{SE}(\widehat{\mu}_r)$. Noting that $\mu_x = M/N$ in the expression for SE($\widehat{\mu}_r$), it follows that

$$\text{SE}(\widehat{\tau}_r) = \sqrt{N(N-n)\frac{s_r^2}{n}}.$$

  In our example, $\widehat{\tau}_r = 2500(1763) =$\$44,006,623(total income for the city) with $\text{SE}(\widehat{\tau}_r) = 2500(1621) =$ \$4,053,522.

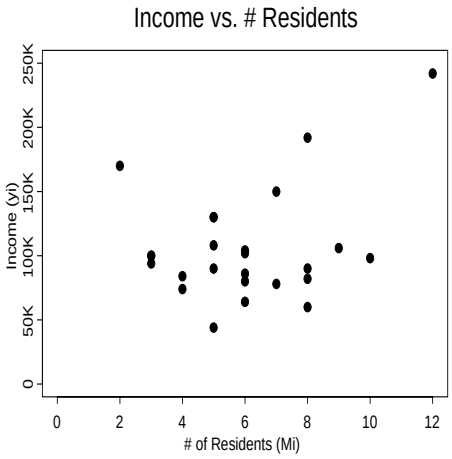
- We can also estimate μ_1, the mean per cluster (block), by the ratio estimator $\widehat{\mu}_{1r} = \widehat{\tau}_r/N$ with $\text{SE}(\widehat{\mu}_{1r}) = \text{SE}(\widehat{\tau}_r)/N$. For our example, $\widehat{\mu}_{1r} = 44,006,000/415 =$\$106,040 with $\text{SE}(\widehat{\mu}_{1r}) = 4053522/415 =$\$9,768.

- The variances of the ratio estimators are based on linearization approximations. We can also use bootstrapping to obtain standard errors, as we showed in the notes on bootstrapping.

- As with the unbiased estimators, we would form confidence intervals by multiplying the SE's by a t value with $n-1$ degrees of freedom.

- Ratio estimators are biased. The bias is typically negligible, and so we compare the variances of these estimators to the variances of the unbiased estimators. We can choose whichever one gives smaller variance.

Note that the ratio estimates are very similar to the unbiased estimates for the income example, but that the standard errors are higher. Why do you think this happened?



R Code for Example 1: Cluster Sampling of Incomes

```
> N <- 415; n <- 25; fpc <- (N-n)/N; M <- 2500
> Mi <- c(8,12,4,5,6,6,7,5,8,3,2,6,5,10,9,3,6,5,5,4,6,8,7,3,8)
> y <- 1000*c(192,242,84,130,104,80,150,130,90,100,170,86,
              108,98,106,100,64,44,90,74,102,60,78,94,82)
> # Unbiased Estimators for Cluster Sampling
> # ========================================
> ybar1 <- mean(y)
> ybar1            # Estimated average income per block
[1] 106320
> su2 <- var(y)
> se.ybar1 <- sqrt(fpc*su2/n)
> se.ybar1         # SE of average income per block
[1] 8447.19
```

```
> tauhat <- N*ybar1
> tauhat            # Estimated total income in city
[1] 44122800
> se.tauhat <- N*se.ybar1
> se.tauhat      # SE of estimated total income
[1] 3505584
> muhat <- tauhat/M
> muhat     # Estimated average income per resident
[1] 17649.12
> se.muhat <- se.tauhat/M
> se.muhat      # SE of average income per resident
[1] 1402.234

> # Ratio Estimators with Cluster Sampling
> # ====================================
> r <- sum(y)/sum(Mi)
> r   # Sample ratio = estimated average income per resident
[1] 17602.65
> mux <- M/N
> sr2 <- (1/(n-1))*sum((y-r*Mi)^2)
> se.r <- sqrt(fpc*sr2/(n*mux^2))
> se.r                    # SE of average income per resident
[1] 1621.409              # (assuming M = 2500 is KNOWN)
> xbar <- mean(Mi)
> se.r2 <- sqrt(fpc*sr2/(n*xbar^2))
> se.r2       # SE of average income per resident
[1] 1617.140  #   (if M were UNKNOWN)
> tauhat <- M*r
> tauhat            # Estimated total income in city
[1] 44006623
> se.tauhat <- M*se.r
> se.tauhat           # SE of estimated total income
[1] 4053522
> muhat1 <- tauhat/N
> muhat1     # Estimated income per block
[1] 106040.1
> se.muhat1 <- se.tauhat/N
> se.muhat1    # SE of estimated income per block
[1] 9767.524
```

In summary, for cluster sampling, we have two basic options:

1. Work only with the primary units (clusters) and use the unbiased estimators.
2. Use ratio estimation to make use of the relationship between cluster size and cluster totals, if such a relationship exists. In many examples, we would expect such a relationship, particularly if the clusters vary greatly in size. Note that if the clusters are all the same size, then the unbiased and ratio estimators are identical.

The true (theoretical) variances of the unbiased estimators all depend ultimately on $\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_1)^2$, the variability between cluster totals, while the true variances of the ratio estimators depend on $\sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - M_i\mu)^2$, the variability between cluster totals, adjusting for cluster size.

PPS Sampling of Clusters

Suppose in cluster sampling that the primary units are drawn with replacement with selection probabilities proportional to the sizes of the primary units (i.e., larger clusters are more likely to be selected than smaller clusters). In the income example, sampling of blocks would be with probabilities proportional to the number of residents on the blocks.

Recall that for PPS sampling with replacement, an unbiased estimator of the population total is given by the Hansen-Hurwitz estimator:

$$\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{M}{n}\sum_{i=1}^{n}\frac{y_i}{M_i}, \text{ where: } p_i = \frac{M_i}{M}$$

with variance given by:

$$\begin{aligned}
\text{Var}(\widehat{\tau}_p) &= \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - \tau\right)^2 = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left(\frac{y_i}{M_i/M} - \tau\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left[M\left(\frac{y_i}{M_i} - \mu\right)\right]^2 = \frac{M}{n}\sum_{i=1}^{N} M_i(\overline{y}_i - \mu)^2 \\
&= \frac{M}{n}\sum_{i=1}^{N}\frac{1}{M_i}(y_i - \mu M_i)^2
\end{aligned}$$

where $\overline{y}_i = y_i/M_i$ is the average per secondary unit in cluster i (e.g., average income per resident in block i).

- An unbiased estimator of $\text{Var}(\widehat{\tau}_p)$ given above is

$$\widehat{\text{Var}}(\widehat{\tau}_p) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}(\overline{y}_i - \widehat{\mu}_p)^2, \text{ where } \widehat{\mu}_p = \widehat{\tau}_p/M.$$

- The theoretical variance of the PPS estimator is roughly the same as for the ratio estimator, but the PPS estimator is unbiased. Both have low variance when the cluster total is proportional to cluster size. In the income example, this would be when block income is proportional to block size.
- A Horvitz-Thompson estimator based on the selection probabilities $p_i = M_i/M$ can also be developed, as indicated on page 161 of Thompson (2nd ed: p. 134).

Systematic Sampling

Recall that systematic sampling is the special case of cluster sampling where each cluster is determined through some random starting point, and a single cluster is chosen.

Example 2: Suppose we are estimating visitation to a particular site and will count the total number of visitors every fifth day for 30 days, starting at a random day from 1 to 5. This gives the following five clusters:

- Every sampling unit is in exactly one cluster.
- In systematic sampling, experimenters generally select ONE cluster from this group. This gives a sample size of 1, which prohibits any type of inference. Why?

  Note: in Example 2, the population (number of days) is finite. If we sample at points in time (for example, recording the temperature every 10 minutes from 8am to 8pm, starting at a random time from 0 to 10 minutes after 8am), then the population size is infinite.
- There are two basic outlooks one takes toward this "problem":
  1. Take more than one systematic sample (replication) and use the unbiased or ratio estimators as outlined above for cluster sampling (identical if the clusters are all the same size, as they often are in systematic sampling).
  2. Assume the variation from one systematic sample to another is not greater than (and probably less than) the variation from one SRS to another. Then use SRS formulas, assuming that they are conservative. In other words, we assume a systematic sample is likely to be more "representative" of the population (and give more accurate estimates of population parameters) than an SRS. This is by far the most common approach in practice. Systematic samples over space (e.g., evenly space plots along a transect) or time

are generally treated as SRS's. So are items or people selected systematically from a list.

In the following example, a systematic sample is treated as an SRS, but has an additional wrinkle.

Example 3: Consider sampling via line transects from an irregularly-shaped area. Interest is in estimating the proportion of the area which has some attribute (say, bare ground).

- Suppose we take a systematic random sample of 10 parallel transects, where the initial starting point is randomly chosen along a baseline, and the resulting 10 transects are evenly spaced along the baseline.
- Because the area is irregular in shape, the transects will be of different lengths.
- Although we really have a cluster sample of size 1, we will assume that the 10 transects are representative of the area and treat them as an SRS of size 10. However, we could also select multiple random starting points and do several systematic samples, perhaps with fewer transects per sample.
- So, the sampling unit here is a transect. What is the population?

How might we estimate the proportion of the area (using these 10 transects) with the desired attribute (bare ground)? Three ways are considered:

1. Simple Average of Transect Proportions: We could measure the proportion of each transect with the attribute, and average the 10 resulting proportions. Problem?

2. Ratio Estimator: Suppose we let

$$\begin{aligned} y_i &= \text{the length of transect } i \text{ with the attribute,} \\ x_i &= \text{the total length of transect } i. \end{aligned}$$

Then a ratio estimate of the proportion of the area with the attribute is:

$$r = \frac{\sum_{i=1}^{10} y_i}{\sum_{i=1}^{10} x_i}, \text{ with } \widehat{\text{Var}}(r) = \underbrace{\left(\frac{N-n}{N}\right)}_{=1 \text{ here}} \frac{1}{\mu_x^2} s_r^2, \text{ where: } s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - r x_i)^2.$$

- As a side note, we might estimate μ_x with $\overline{x}$ (the average length of the 10 transects), but we can actually determine μ_x if we know the area A of the region and the length b of the baseline – then $\mu_x = A/b$.
- This is a perfectly valid way of estimating the proportion of the area with the attribute, although, as with all ratio estimators, the estimator is biased.

3. Unbiased Estimator: Consider only the lengths (y_i's) of bare ground on the sampled transects and not their lengths.

   - We can determine the true mean length of a transect μ_x since we know the total area and the length along the baseline of the region. Then an unbiased estimator of the proportion of the area which is bare ground is

   $$\frac{\overline{y}}{\mu_x} = \frac{\text{average length of the attribute among the transects}}{\text{true mean length of a transect}}.$$

   Since $\text{Var}(\overline{y}/\mu_x) = (1/\mu_x^2)\text{Var}(\overline{y})$, the SE of this estimator is $(1/\mu_x)s/\sqrt{n}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.
   - If there is large variation in transect lengths, ratio estimation will likely do better than the unbiased estimator.

Basic Principle of Cluster and Systematic Sampling: Because a census is taken on all secondary units within a primary unit, the within-primary-unit variance plays no role in the variances of population means or totals. This explains why the ideal case for cluster or systematic sampling is that when there is large variability within primary units relative to the variability between primary units, because this large within-primary-unit variability will have no effect on the variance of estimators!

Hence, we want the primary units to be "mini-populations"; that is, we want them to be representative of the population as a whole. This will minimize differences between primary units while maximizing differences within.

Multistage Sampling (Chapter 13)

Multistage sampling refers to sampling plans where the sampling is carried out in stages using smaller and smaller sampling units at each stage. In a two-stage sampling design, a sample of primary units is selected and then a sample of secondary units is selected within each primary unit. This handout outlines the development of estimators under the general setting of two-stage sampling, considers the allocation question under the setting of equal sized primary and secondary units, and briefly examines three-stage sampling.

The simplest version of two-stage sampling is to use simple random sampling at each stage – an SRS of primary units, and an SRS of secondary units within each selected primary unit. The primary units do not need to be the same size and you do not need to select the same number of secondary units within each primary unit.

- Stratified random sampling and cluster sampling can be viewed as special cases of two-stage sampling. A stratified random sample is a census of the primary units (the strata) followed by an SRS of the secondary units within each primary unit. A cluster sample is an SRS of the primary units (the clusters) followed by a census of the secondary units within each selected primary unit.

- We can use any probability sampling plan at each stage of a multistage plan and the plans can be different at each stage. The formulas developed below are only for an SRS at each stage. It's possible to derive formulas for other situations.

Example 1: In order to estimate the condition of highways under its jurisdiction and the cost of urgent repairs, the state Department of Transportation selected a number of "highway miles" in two stages. In the first stage, a number of highways were selected at random and without replacement from the list of all highways maintained by the Department. In the second stage, a number of one-mile segments were selected at random and without replacement from the total length of each selected highway; for example, if the length of highway 101 is 73 miles, it is seen as consisting of 73 one-mile segments ("highway miles"), from which a number are selected at random. Highway engineers then visit the selected segments, inspect the pavement condition, rate the condition of the segment, and estimate the cost of urgently needed repairs.

For the purpose of this problem, assume there are 352 highways in the state, with a total length of 28,950 miles. A simple random sample of five highways was selected without replacement. From each selected highway, approximately 10% of its one-mile segments were then selected. The inspection results are presented in the table below.

Highway Number	Length (miles)	Selected One-Mile Segments	Number Rated Excellent	Cost of Urgent Repairs (in $1,000)
155	85	10	2	90
489	120	15	1	110
283	47	5	0	60
698	98	10	0	100
311	34	5	1	30

For example, Highway 155 has a length of 85 miles. Ten of its 85 one-mile segments were selected and inspected. Two of these segments were rated Excellent. The total cost of urgent repairs on the 10 selected segments was $90,000.

(a) Estimate the proportion and number of state highway miles that are in Excellent condition.
(b) Estimate the average cost per highway mile and the total cost of urgently needed repairs.

- First, why would a two-stage sampling plan be adopted for this highway problem in the first place? Why not an SRS?

- Multistage samples are used primarily for cost or feasibility (practicality) reasons. For example, to select an SRS of households in the U.S. would be extremely difficult because no list of all households exists. However, we could proceed in stages: an SRS of counties in the U.S., an SRS of "blocks" within each county, and an SRS of households within each block. You would then only need to have a list of households within each block that was selected. Two-stage sampling also has the flexibility to sample more intensely in primary units which are larger or more variable. The disadvantage of two-stage sampling is that the variance of the resulting estimators are likely to be larger than for an SRS of the same total number of secondary units. This may well be more than offset by the cost efficiency of two-stage sampling.

- Note that a two-stage sample can never be better than a cluster sample with the same number of primary units selected because a census within each primary unit is the best you can do.

Notation for Two-Stage Sampling:

$$
\begin{aligned}
N &= \text{the number of primary units in the population,}\\
n &= \text{the number of primary units in the sample,}\\
M_i &= \text{the number of secondary units in the } i^{th} \text{ primary unit,}\\
M &= \sum_{i=1}^{N} M_i = \text{ total number of secondary units in the population,}\\
m_i &= \text{the size of the sample in the } i^{th} \text{ primary unit,}\\
y_{ij} &= \text{the response of the } j^{th} \text{ secondary unit within the } i^{th} \text{ primary unit,}\\
y_i &= \sum_{j=1}^{M_i} y_{ij} = \text{the total in the } i^{th} \text{ primary unit,}\\
\mu_i &= \frac{1}{M_i}\, y_i = \text{the mean response per secondary unit in the } i^{th} \text{ primary unit,}\\
\overline{y}_i &= \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} = \text{ the sample mean response per secondary unit in the } i^{th} \text{ primary unit}
\end{aligned}
$$

In two-stage problems, we are generally interested in:

$$
\begin{aligned}
\tau &= \sum_{i=1}^{N} y_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \text{ (total of the } y\text{-values for the population),}\\
\mu &= \tau/M \text{ (mean per secondary unit in the population),}\\
\mu_1 &= \tau/N \text{ (mean per primary unit in the population).}
\end{aligned}
$$

- For stratified random sampling: $n = N$ (census of primary units).
- For cluster sampling: $m_i = M_i$ (census of secondary units).

As with cluster sampling, there are two main methods of estimating population parameters of interest: unbiased estimation and ratio estimation. As in cluster sampling, ratio estimation uses cluster size as an auxiliary variable and is usually better if there is a relationship between cluster size (M_i) and cluster total (y_i).

Unbiased Estimation of the Population Total τ and Mean μ.

- First, an unbiased estimator of the total in the i^{th} cluster (y_i) is

  $\widehat{y}_i =$

- With these estimated cluster totals, estimators for the population total τ, population mean per secondary unit μ, and population mean per primary unit μ_1 are given by:

$$
\widehat{\tau} = \frac{N}{n}\sum_{i=1}^{n} M_i\overline{y}_i, \qquad \widehat{\mu} = \frac{\widehat{\tau}}{M} = \frac{N}{nM}\sum_{i=1}^{n} M_i\overline{y}_i, \qquad \widehat{\mu}_1 = \frac{\widehat{\tau}}{N} = \frac{1}{n}\sum_{i=1}^{n} M_i\overline{y}_i.
$$

- Note: we need to know M, the total number of secondary units in the population, to estimate μ with the unbiased estimator. The ratio estimator, discussed next, can be used when M is not known.

R code to compute the estimated total for the highway example follows:

```
> N <- 352; n <- 5; M <- 28950
> Mi <- c(85,120,47,98,34)  # total no. segments on the highways sampled
> mi <- c(10,15,5,10,5)     # no. of segments sampled
> yi <- c(2,1,0,0,1)        # no. of excellent segments

> # Unbiased estimation of total number of segments rated Excellent
> # ============================================================
> yhati <- (Mi/mi)*yi
> yhati   # estimated no. excellent segments on each highway
[1] 17.0  8.0  0.0  0.0  6.8

> tauhat <- (N/n)*sum(yhati) # estimated total no. excellent
> tauhat
[1] 2238.72
```

So an unbiased estimate of the total number of highway segments rated as Excellent is $\widehat{\tau} = 2238.7$. What is $\widehat{\mu}$?

It can be shown that in two-stage sampling, the variance of the estimator of μ is

$$\text{Var}(\widehat{\mu}) = \frac{N^2}{M^2}\left(\frac{N-n}{N}\right)\frac{\sigma_u^2}{n} + \frac{N}{nM^2}\sum_{i=1}^{N} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{\sigma_i^2}{m_i},$$

where

$$\begin{aligned}
\sigma_u^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_1)^2 \text{ (where } \mu_1 = \tau/N \text{ is the population mean per primary unit),} \\
&= \text{variance between primary unit totals,} \\
\sigma_i^2 &= \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \mu_i)^2 = \text{variance between secondary units within } i^{th} \text{ primary unit.}
\end{aligned}$$

Note that $\text{Var}(\widehat{\tau}) = M^2\text{Var}(\widehat{\mu})$ and $\text{Var}(\widehat{\mu}_1) = \text{Var}(\widehat{\tau})/N^2 = (M^2/N^2)\text{Var}(\widehat{\mu})$.

- If $n = N$, then the first term = 0, and the second term = $\text{Var}(\widehat{\mu})$ for stratified random sampling.
- If $m_i = M_i$, then the second term = 0, and the first term = $\text{Var}(\widehat{\mu})$ for cluster sampling.
- The estimated variance of $\widehat{\mu}$ is

$$\widehat{\text{Var}}(\widehat{\mu}) = \frac{N^2}{M^2}\left(\frac{N-n}{N}\right)\frac{s_u^2}{n} + \frac{N}{nM^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

  where

$$\begin{aligned} s_u^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \widehat{\mu}_1)^2, \quad s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2, \\ \widehat{y}_i &= M_i\overline{y}_i, \quad \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i. \end{aligned}$$

  There are two levels of approximation in s_u^2: we use n for N and $\widehat{y}_i$ for y_i (primary unit total).

- Similarly, $\widehat{\text{Var}}(\widehat{\tau}) = M^2\widehat{\text{Var}}(\widehat{\mu})$ (since $\widehat{\tau} = M\widehat{\mu}$) and $\widehat{\text{Var}}(\widehat{\mu}_1)$.
- In the highway example, we are counting the number of 1-mile segments rated as Excellent; hence, we have binary data ($y_{ij} = 0$ or 1). So the mean in this example is the *proportion* of one-mile segments rated as Excellent. Here then, the within-primary unit sample variance is:

$$s_i^2 = \frac{m_i}{m_i - 1}\,\widehat{p}_i(1 - \widehat{p}_i) \text{ (the binomial variance).}$$

So, to finish answering part (a) of the highway problem (where we already estimated the total number of segments in Excellent condition to be $\widehat{\tau} = 2238.7$), the estimated proportion of highway miles in Excellent condition, as well as standard errors for both this proportion and the total are given via the R code on the next page.

```
> su2 <- var(yhati)
> su2
[1] 49.248
> pi <- yi/mi     # Proportion of segments rated excellent on each highway
> pi
[1] 0.20000000 0.06666667 0.00000000 0.00000000 0.20000000
> si2 <- (mi/(mi-1))*pi*(1-pi)  # Estimated variance within each primary unit
> si2
[1] 0.17777778 0.06666667 0.00000000 0.00000000 0.20000000
> var1 <- (N*(N-n)*su2)/n  # Term 1 of variance
> var2 <- (N/n)*sum((Mi*(Mi-mi)*si2)/mi)  # Term 2 of variance
> c(var1,var2)
[1] 1203069.54   14697.64
> var.tauhat <- var1 + var2
> SE.tauhat <- sqrt(var.tauhat)        # SE of estimate of total
> SE.tauhat
[1] 1103.525
> c(tauhat-qt(.975,n-1)*SE.tauhat,tauhat+qt(.975,n-1)*SE.tauhat) # 95% CI
[1] -825.1563 5302.5963

> phat <- tauhat/M
> phat                    # estimate of proportion Excellent
[1] 0.07733057
> SE.phat <- SE.tauhat/M      # SE of estimate of proportion
> SE.phat
[1] 0.0381183
> c(phat-qt(.975,n-1)*SE.phat,phat+qt(.975,n-1)*SE.phat) # 95% CI
[1] -0.02850281  0.18316395
```

$$\left(s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \widehat{\mu}_1)^2\right)$$

$$(\widehat{p}_i = y_i/m_i)$$

$$\left(s_i^2 = \frac{m_i}{m_i - 1}\widehat{p}_i(1-\widehat{p}_i),\ i = 1,2,3,4,5\right)$$

$$\left(\widehat{\mathrm{Var}}(\widehat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}\right)$$

$$\left(\widehat{p} = \frac{\widehat{\tau}}{M} = \widehat{\mu}\right)$$

$$\left(\mathrm{SE}(\widehat{p}) = \frac{1}{M}\mathrm{SE}(\widehat{\tau}) = \frac{1}{M}\sqrt{\mathrm{Var}(\widehat{\tau})} = \sqrt{\frac{\widehat{\mathrm{Var}}(\widehat{\tau})}{M^2}}\right)$$

Note that the confidence interval extends below 0. Since the estimated proportions within each highway are near 0, our sample sizes are too small to assume a normal sampling distribution for $\widehat{p}$. We might consider bootstrapping.

Ratio Estimation in Two-Stage Sampling

If the sizes of the primary units (highways) are linearly related (through the origin) with the values of the response (number rated as excellent, or cost of urgent repairs), a ratio estimator may provide a better estimator of the population total or mean.

- The estimators are given by: $\widehat{\mu}_r = \widehat{r} = \dfrac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i}$, $\widehat{\tau}_r =$

  We'll use $\widehat{\mu}_r$ and $\widehat{r}$ interchangeably (Thompson uses $\widehat{r}$).

- The variance of $\widehat{\mu}_r$ is given by:

$$\mathrm{Var}(\widehat{\mu}_r) = \underbrace{\frac{N^2}{M^2}\left(\frac{N-n}{N}\right)\frac{1}{n}\left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i - M_i\mu)^2\right)}_{\substack{\text{variability between primary units (highways)} \\ \text{after adjusting for highway length}}} + \underbrace{\frac{N}{nM^2}\sum_{i=1}^{N} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}}_{\substack{\text{within primary unit variability} \\ \text{(same as earlier)}}}.$$

- If M_i (the size of primary unit i) is related to y_i (total for primary unit i), then the first term in the above equation should be small. This is the situation where ratio estimation should be used. The estimated variance of $\widehat{\tau}_r$ is given at the bottom of page 175 of Thompson (2nd ed: p. 147). Dividing this expression by M^2 gives estimated variance of $\widehat{\mu}_r$ since $\widehat{\mu}_r = \widehat{\tau}_r / M$:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\mu}_r) &= \frac{N(N-n)}{M^2 n(n-1)}\sum_{i=1}^{n}(\widehat{y}_i - M_i\widehat{r})^2 + \frac{N}{M^2 n}\sum_{i=1}^{n} M_i(M_1 - m_i)\frac{s_i^2}{m_i} \\
&= \left(\frac{N-n}{N}\right)\left(\frac{1}{\overline{M}^2 n}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - M_i\widehat{r})^2\right] + \frac{1}{N\overline{M}^2 n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}
\end{aligned}$$

  where, if necessary, the population mean primary unit size $\overline{M} = M/N$ can be estimated by the mean primary unit size for the sample, $(1/n)\sum_{i=1}^{n} M_i$. Note that as N becomes very large, the second term goes to 0.

- The estimators and corresponding SE's for ratio estimation are computed via R on the next page.

```
> # Ratio Estimation of total segments and proportion rated Excellent
> # ================================================================
> rhat <- sum(yhati)/sum(Mi)
> rhat        # estimate of the proportion
[1] 0.0828125
> tauhat.r <- M*rhat
> tauhat.r      # estimate of the total
[1] 2397.422
> sr2 <- (1/(n-1))*sum((yhati - Mi*rhat)^2)
> sr2
[1] 49.96548
> var.tauhat.r <- (N*(N-n)*sr2)/n + (N/n)*sum((Mi*(Mi-mi)*si2)/mi)
> sqrt(var.tauhat.r)     # SE of estimate of total
[1] 1111.438
> sqrt(var.tauhat.r/M^2)   # SE of estimate of proportion
[1] 0.03839164
```

$$\left(\widehat{r} = \widehat{\mu}_r = \frac{\sum \widehat{y}_i}{\sum M_i} = \frac{\sum M_i \overline{y}_i}{\sum M_i}\right)$$

$$(\widehat{\tau}_r = M \cdot \widehat{\mu}_r)$$

$$\left(s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - M_i\widehat{\mu}_r)^2\right)$$

$$\left(\widehat{\mathrm{Var}}(\widehat{\tau}_r) = N(N-n)\cdot\frac{s_r^2}{n} + \left(\frac{N}{n}\right)\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}\right)$$

- $\mathrm{SE}(\widehat{\mu}_r) = .038$, which is about the same as with the earlier unbiased estimator (.038), as there was no real relationship between the highway length and the number of Excellent segments on the highway.
- If appropriate, a ratio estimator will generally be better if there is a lot of variation in the highway lengths.

Part (b) of Example 1: Estimate the average cost per highway mile and the total cost of urgently needed repairs. Here, we have $N = 352$ total highways, $n = 5$ sampled highways, and $M = 28950$ total 1-mile segments, and:

$$\begin{aligned}
M_i &= \text{the number of segments for highway } i,\\
m_i &= \text{the number of sampled segments for highway } i,\\
y_i &= \text{the cost of repairs for highway } i\\
&\quad \text{(This was the \# of segments rated Excellent for highway } i \text{ earlier)}\\
\overline{y}_i &= \text{the average cost per segment for highway } i,\\
M_i\overline{y}_i &= \widehat{y}_i = \text{the estimated total cost for highway } i.
\end{aligned}$$

Again, we can use either the unbiased estimators of the mean and total or the ratio estimators. The ratio estimators would be expected to be better if there's a linear relationship (through the origin) between the length of a highway and the estimated total repair costs, or, equivalently, if there's little variation between the average repair costs per mile on the different highways.

There's an important piece of information missing in the data table on p. 111 which we need to calculate the SE's of our estimates for the cost data. Can you see what it is?

In the R analysis below, some values are assumed for the missing information, but we'll also see that what values we assume makes little difference in the SE's.

```
> N <- 352; n <- 5; M <- 28950
> Mi <- c(85,120,47,98,34)  # total no. segments on the highways sampled
> mi <- c(10,15,5,10,5)     # no. of segments sampled
> yi <- c(90,110,60,100,30) # total cost on sampled segments
> si <- c(3.1,3.5,4.8,2.9,2.5)
> si2 <- si^2

> # Unbiased estimation of total cost of repairs and mean cost per segment
> # ============================================================
> yhati <- (Mi/mi)*yi
> yhati   # estimated total cost on each highway
[1] 765 880 564 980 204

> tauhat <- (N/n)*sum(yhati) # estimated total cost
> tauhat
[1] 238867.2

> su2 <- var(yhati)
> su2
[1] 94311.8
> var1 <- (N*(N-n)*su2)/n  # Term 1 of variance of tauhat
> var2 <- (N/n)*sum((Mi*(Mi-mi)*si2)/mi) # Term 2 of variance of tauhat
> c(var1,var2)
[1] 2303924100    2393449
> var.tauhat <- var1 + var2
> SE.tauhat <- sqrt(var.tauhat)        # SE of estimate of total
> SE.tauhat
[1] 48024.14
> c(tauhat-qt(.975,n-1)*SE.tauhat,tauhat+qt(.975,n-1)*SE.tauhat) # 95% CI
[1] 105530.8 372203.6

> muhat <- tauhat/M
> muhat                  # estimate of mean cost per segment
```

```
[1] 8.251026
> SE.muhat <- SE.tauhat/M      # SE of estimate of proportion
> SE.muhat
[1] 1.658865
> c(muhat-qt(.975,n-1)*SE.muhat,muhat+qt(.975,n-1)*SE.muhat) # 95% CI
[1]  3.645279 12.856773


> # Ratio Estimation of total cost and mean cost per segment
> # ======================================================
> rhat <- sum(yhati)/sum(Mi)
> tauhat.r <- M*rhat
> tauhat.r      # estimate of the total cost
[1] 255800.4
> sr2 <- (1/(n-1))*sum((yhati - Mi*rhat)^2)
> sr2
[1] 19283.25
> var.tauhat.r <- (N*(N-n)*sr2)/n + (N/n)*sum((Mi*(Mi-mi)*si2)/mi)
> sqrt(var.tauhat.r)     # SE of estimate of total cost
[1] 21759.14

> rhat # estimated mean cost per segment
[1] 8.835938
> sqrt(var.tauhat.r/M^2) # SE of estimated mean cost per segment
[1] 0.751611
```

- We were not given the standard deviations of the costs for each sampled highway and we were not given the individual data values (the cost for each of the sampled sections on a highway) from which to compute them. However, we can also see that the within highway variability contributed little to the estimated variance of our estimators. If we had assumed a standard deviation of \$10,000 on each highway (very high, considering that the average costs ranged from \$6,000 to \$12,000), the SE of the unbiased estimate of the total cost would have increased from \$48,024 to only \$48,214 (and would have decreased to \$47,999 if all the standard deviations were 0). This points out something important in two-stage designs: it is generally the variability between the primary units and the number of primary units sampled that determines the accuracy of the estimators.
- Note that for the cost data, the ratio estimator decreased the SE's of the estimates of the total and mean by over half. This was because there was a relationship between the lengths of the highways and the estimated total cost of repairs.

Comparison of $\text{Var}(\widehat{\tau})$ and $\widehat{\text{Var}}(\widehat{\tau})$ for the unbiased estimator.

Recall:

$$\begin{aligned}\text{Var}(\widehat{\tau}) &= N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i} \\ \widehat{\text{Var}}(\widehat{\tau}) &= N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}.\end{aligned}$$

- It was mentioned earlier that $\widehat{\text{Var}}(\widehat{\tau})$ is an unbiased estimator of $\text{Var}(\widehat{\tau})$.
- Although s_i^2 is an unbiased estimator of σ_i^2, the second term in the expression for $\widehat{\text{Var}}(\widehat{\tau})$ above, $\frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}$, is not an unbiased estimator of the second term in $\text{Var}(\widehat{\tau})$, but is an unbiased estimator of $\sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$ without the N/n constant.
- So the 2nd piece in $\widehat{\text{Var}}(\widehat{\tau})$ underestimates the corresponding 2nd piece in $\text{Var}(\widehat{\tau})$.
  And the 1st piece in $\widehat{\text{Var}}(\widehat{\tau})$ overestimates the corresponding 1st piece in $\text{Var}(\widehat{\tau})$. Why?
- It is easy to show with an example that s_u^2 cannot, in general, be an unbiased estimator of σ_u^2 and that (in this example) we must have $\text{E}(s_u^2) > \sigma_u^2$. How? (Think of an extreme case.)

  So s_u^2 overestimates σ_u^2 because it includes both the variability between primary units and the variability within primary units.

Allocation in Two-Stage Sampling

A practical question in developing a two-stage sampling plan is how to allocate resources to the sampling of primary units versus secondary units. Here, we consider the special case of:

1. Equal-sized primary units: $M_1 = \cdots = M_N = \overline{M}$.
2. Equal-sized samples within primary units: $m_1 = \cdots m_n = m$.

The total sample size then is mn, and $\overline{M}N = M$ where:

$$\begin{aligned}\overline{M} &= \text{the number of secondary units per primary unit,} \\ N &= \text{the total number of primary units,} \\ M &= \text{the total number of secondary units.}\end{aligned}$$

- Under the allocation assumptions of equal-sized primary units and equal-sized samples within primary units, the unbiased and ratio estimators of the total are identical, where:

$$\begin{aligned}\widehat{\tau}_r &= M\widehat{\mu}_r = M \cdot \frac{\sum_{i=1}^n M_i\overline{y}_i}{\sum_{i=1}^n M_i} = \frac{M}{n\overline{M}}\sum_{i=1}^n M_i\overline{y}_i = \frac{N}{n}\sum_{i=1}^n M_i\overline{y}_i = \frac{N}{n}\overline{M}\sum_{i=1}^n \overline{y}_i \\ &= M\left(\frac{1}{n}\sum_{i=1}^n \overline{y}_i\right) = M\overline{y} = \widehat{\tau}.\end{aligned}$$

  So, we take the average of all responses and multiply it by the number of secondary units.

- Also: $\widehat{\mu} = \dfrac{\widehat{\tau}}{M} = \overline{y}$.

Working with the variance of $\widehat{\mu}$:

$$\begin{aligned}\boxed{\mathrm{Var}(\widehat{\mu})} &= \frac{\mathrm{Var}(\widehat{\tau})}{M^2} = \frac{N(N-n)}{M^2}\frac{\sigma_u^2}{n} + \frac{N}{M^2 n}\sum_{i=1}^N \overline{M}(\overline{M}-m)\frac{\sigma_i^2}{m} \\ &\quad \left[\begin{array}{c} \text{Note: } \sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i-\mu_1)^2 = \frac{1}{N-1}\sum_{i=1}^N (\overline{M}\mu_i - \overline{M}\mu)^2 \\ = \frac{\overline{M}^2}{N-1}\sum_{i=1}^N (\mu_i-\mu)^2 = \underline{\overline{M}^2\sigma_b^2}, \\ \text{where: } \sigma_b^2 = \text{ variability } \underline{\text{between}} \text{ primary units} \end{array}\right] \\ &= \frac{N(N-n)}{\overline{M}^2 N^2}\cdot\frac{\overline{M}^2\sigma_b^2}{n} + \frac{N\overline{M}(\overline{M}-m)}{\overline{M}^2 N^2 nm}\sum_{i=1}^N \sigma_i^2 \\ &= \left(\frac{N-n}{N}\right)\frac{\sigma_b^2}{n} + \left(\frac{\overline{M}-m}{\overline{M}}\right)\frac{1}{nm}\left(\frac{1}{N}\sum_{i=1}^N \sigma_i^2\right) \\ &= \boxed{(1-f_1)\frac{\sigma_b^2}{n} + (1-f_2)\frac{\sigma_w^2}{mn}},\end{aligned}$$

where: σ_w^2 = the average <u>within</u> primary unit variability, and f_1 and f_2 are the sampling fractions at the first and second stages, respectively.

- <u>Note</u>: If we increase m (the # of secondary units), we can drive the 2nd term in the variance above to zero, but this will have no effect on the 1st term.

<u>Goal</u>: W̄e want to find those values of n and m which minimize the $\mathrm{Var}(\widehat{\mu})$ subject to some restriction on how many samples we can take. First, we will simply assume that the <u>total sample size nm is fixed</u>; this assumes (usually unrealistically) that it takes the same amount of effort, for example, to sample 2 secondary units from each of 10 randomly selected primary units as it does to sample 10 secondary units from each of 2 randomly selected primary units.

First, note that: $\boxed{\mathrm{Var}(\widehat{\mu})} = \left(1-\dfrac{n}{N}\right)\dfrac{\sigma_b^2}{n} + \left(1-\dfrac{\overline{m}}{\overline{M}}\right)\dfrac{\sigma_w^2}{nm} = \dfrac{\sigma_b^2}{n} - \dfrac{\sigma_b^2}{N} + \dfrac{\sigma_w^2}{nm} - \dfrac{\sigma_w^2}{n\overline{M}}$,

where the middle two terms are fixed for any choice of n, m ($nm = c$).
So, we want to minimize $\dfrac{1}{n}\left(\sigma_b^2 - \dfrac{\sigma_w^2}{\overline{M}}\right)$ with respect to n.

- If $\sigma_b^2 > \frac{\sigma_w^2}{\overline{M}}$, then we should:

- If $\sigma_b^2 < \frac{\sigma_w^2}{\overline{M}}$, then we should:

- If the primary unit size, $\overline{M}$, is large (it usually is), σ_b^2 will be larger than $\sigma_w^2/\overline{M}$.
- All of this ignores any differences in cost.

Cost Considerations in Allocation: Suppose we let $C = c_0 + c_1 n + c_2 nm$ where:

c_0 = the startup cost,
c_1 = the cost of selecting a primary unit (travel, time, etc.),
c_2 = the cost of sampling a secondary unit once we've selected a primary unit.

Suppose we fix the total cost C. Then, the optimal allocation for m (that which minimizes $\text{Var}(\widehat{\mu})$ for fixed C) is:

$$m_{opt} = \sqrt{\frac{c_1 \sigma_w^2}{c_2\left(\sigma_b^2 - \sigma_w^2/\overline{M}\right)}}.$$

- Note that this optimal choice for m does not depend in any way on the total cost C.
- If c_1 increases relative to c_2, it makes sense that m_{opt} will increase, because the cost of sampling primary units increases.
- Often, if $\overline{M}$ is large, then $\frac{\sigma_w^2}{\overline{M}} \approx 0$, and $m_{opt} \approx \sqrt{\frac{c_1\sigma_w^2}{c_2\sigma_b^2}}$. In this case, we need only know the relative costs and relative variabilities.

Back to the Highway Example: Suppose it takes 1/2 hour to actually sample a 1-mile segment (c_2). It might be much more costly to select a primary unit, and suppose we guess: $\frac{c_1}{c_2} \approx 25$.

- Suppose we have preliminary data (say on 4 highways with 5 segments each) and we conduct an analysis of variance (ANOVA) to estimate the two variance components σ_b^2 and σ_w^2, given below:

$$\begin{aligned} \sigma_b^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(\mu_i - \mu)^2 \\ \sigma_w^2 &= \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2, \text{ where: } \sigma_i^2 = \frac{1}{\overline{M}-1}\sum_{j=1}^{\overline{M}}(y_{ij} - \mu_i)^2. \end{aligned}$$

- Recall that $s_b^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2$ overestimates σ_b^2.

- And $s_w^2 = \dfrac{1}{n}\sum_{i=1}^{n} s_i^2$ is an <u>unbiased</u> estimate of σ_w^2 (since s_i^2 is unbiased for σ_i^2).

- Conducting an ANOVA on the y_{ij}'s with the primary units (highways) as factors yields the partitioning:

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \overline{y})^2}_{} = \underbrace{m\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2}_{} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2}_{}.$$

- This gives the following ANOVA table:

Source of of Variance	Degrees of Freedom	Sums of Squares	Mean Squares	E(MS)
Between				
Within				
Total				

$$\mathrm{E(MSW)} = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} s_i^2\right] = \underline{\sigma_w^2}.$$

- We want to use the expected mean squares above to estimate σ_b^2. How?

R Code to Estimate σ_b^2 and σ_w^2 via ANOVA: Reconsider the highway repair example with hypothetical data on repair costs.

```
> hwy <- rep(c("A","B","C","D"),rep(5,4))
> hwy
[1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
[16] "D" "D" "D" "D" "D"
> repcost <- c(3,6,7,9,4,12,8,14,9,10,6,8,10,7,10,5,4,8,6,6)
> repcost
[1]   3  6  7  9  4 12  8 14  9 10  6  8 10  7 10  5  4  8  6  6

> d <- data.frame(hwy,repcost)
> a <- aov(repcost~hwy,data=d)  # Conducts an ANOVA of repair costs
> summary(a)                    #   on the highway ID
            Df Sum Sq Mean Sq F value   Pr(>F)
 hwy         3 79.200  26.400  6.2485 0.005184 **
 Residuals  16 67.600   4.225
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> sigb2 <- (1/5)*(26.4 - 4.225)                    (σ_b^2 ≈ (1/m)(s_b^2 - s_w^2))
> sigb2
[1] 4.435                  # Estimate of sigma-b-sq
> 4.225/90
[1] 0.04694444             # Estimate of sigma-w-sq / Mbar
```

- Since $\sigma_w^2/\overline{M} = 0.0469$ is small relative to σ_b^2, and hence effectively negligible, then the approximate optimal allocation for m is given by:

$$m_{opt} = \sqrt{\frac{c_1\sigma_w^2}{c_2(\sigma_b^2 - \sigma_w^2/\overline{M})}} \approx \sqrt{\frac{c_1\sigma_w^2}{c_2\sigma_b^2}} = \sqrt{\frac{c_1(4.225)}{c_2(4.435)}}.$$

- We had guessed that $\dfrac{c_1}{c_2} = 25$. Then $m_{opt} = 4.8$, so we might use 5 one-mile segments per highway.

- Had we guessed that $\dfrac{c_1}{c_2} = 10$, then $m_{opt} = 3.09$ and we would have used 3 one-mile segments per highway.

- The value of n is now determined by the overall budget (or cost). Recall that the total cost was given by: $C = c_0 + c_1 n + c_2 nm$

$$\Longrightarrow \;\; C = c_0 + c_1 n + 5c_2 n \;\; \Longrightarrow \;\; \boxed{n = \frac{C - c_0}{c_1 + 5c_2}}.$$

PPS Sampling in Two-Stage Problems

As with cluster sampling, Hansen-Hurwitz estimation can be employed in two-stage sampling if primary units are selected with probability proportional to size with replacement. For details of the forms of the resulting unbiased estimators, see Section 13.2 of Thompson.

Horvitz-Thompson Estimator in Two-Stage Sampling: Recall that the Horvitz-Thompson estimator can be applied in virtually any sampling problem. Recall also that this estimator depends on the inclusion probabilities for the units in the population. How do we find $\pi_{(ij)}$, the inclusion probability for the j^{th} secondary unit in the i^{th} primary unit? (The parentheses around ij are to make it clear that it's not a joint inclusion probability).

$\pi_{(ij)} =$

With these inclusion probabilities, the Horvitz-Thompson estimator of the population total is:

$$\widehat{\tau}_\pi = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{y_{ij}}{\pi_{(ij)}} =$$

which is what we found earlier as the unbiased estimator.

Three-Stage Sampling: Suppose now that we sample:

n out of N primary units
m out of $\overline{M}$ secondary units (3 Stages)
t out of $\overline{T}$ tertiary units

For this three-stage sampling plan, the variance of the estimated mean is:

$$\mathrm{Var}(\widehat{\mu}) = \underbrace{(1-f_1)\frac{\sigma_1^2}{n}} + \underbrace{(1-f_2)\frac{\sigma_2^2}{mn}} + \underbrace{(1-f_3)\frac{\sigma_3^2}{nmt}} \quad \left(f_1 = \frac{n}{N},\ f_2 = \frac{m}{\overline{M}},\ f_3 = \frac{t}{\overline{T}}\right)$$

$$\Rightarrow \widehat{\mathrm{Var}}(\widehat{\mu}) = (1-f_1)\frac{s_1^2}{n} + f_1(1-f_2)\frac{s_2^2}{nm} + f_1f_2(1-f_3)\frac{s_3^2}{nmt}.$$

- Again, s_1^2 overestimates σ_1^2, s_2^2 overestimates σ_2^2, and s_3^2 underestimates σ_3^2, but $\widehat{\mathrm{Var}}(\widehat{\mu})$ is unbiased for $\mathrm{Var}(\widehat{\mu})$.

Double Sampling (Chapter 14)

To this point, we have considered a number of sampling or estimation methods (ratio and regression estimation, stratified sampling, e.g.) whereby auxiliary information was used to estimate a population mean or total. In all of these cases, it was assumed that we knew population information on the auxiliary variable.

Double sampling is a sampling method which makes use of auxiliary data where the auxiliary information is obtained through sampling. More precisely, we first take a sample of units strictly to obtain auxiliary information, and then take a second sample where the variable(s) of interest are observed. It will often be the case that this second sample is a *subsample* of the preliminary sample used to acquire auxiliary information. Two common situations where double sampling is employed to use auxiliary information to improve the estimate of some response variable are outlined below.

1. If the variable of interest is "expensive" to measure, but a related variable is much "cheaper," we might first sample many of the sampling units and measure the "cheaper" (auxiliary) variable, and only measure the response variable on a subsample (or smaller sample).

    - A common example of this scenario is any case where a visual estimate of some response variable can be made much more quickly (cheaply) than measuring the response outright. For example, if we want to estimate the number of leaves in some area, it could be quite time-consuming to count the number of leaves in a number of 18x18 inch quadrats, whereas a visual estimate of the number of leaves in such an area is relatively simple to make. If a definite relationship between the visual and actual number of leaves in a quadrat can be established, we could make efficient use of the visual estimates (even if they are highly biased) to improve an estimate of the total number of leaves through double sampling.

2. A common problem in many surveys (as discussed in this class) is that of potential bias from nonresponse. Double sampling can be used with stratification principles to adjust for nonresponse in surveys by taking a second sample of the nonrespondents.

Example 1: Suppose it is desired to estimate the total biomass of vegetation and average biomass (gm/m^2) on a 1000 m^2 area. A systematic sample of twenty 1 m^2 plots is selected and a visual estimate of the total grams of biomass is recorded. Five of the twenty plots are then randomly selected and the total biomass is carefully determined on these 5 plots. The visual and actual measurements of total biomass are given on the next page.

Visual Only

60	60	200	0	100	80	20	150	100	80	60	60	20	150	60

Visual and Actual

Visual:	20	80	150	40	80
Actual:	14	62	155	36	71

Ratio Estimation in Double Sampling: Let

$$\begin{aligned} n' &= \text{the total \# of units observed,} \\ n &= \text{the \# of units in the second sample, where both } x \text{ and } y \text{ are observed.} \end{aligned}$$

In the biomass example, $n' = 20$ and $n = 5$.

Suppose we want to estimate $\tau = \sum_{i=1}^{N} y_i$ (total biomass in the 1000 m^2 area).

- First, compute the sample ratio: $r = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$ for the second (smaller) sample.
- Recall that with ratio estimation, we assumed we knew τ_x, the population total for x, and we then estimated the total for y by $\widehat{\tau}_r = r\tau_x$. But here, we need to estimate τ_x (since we only have a sample of the x-values). How?

- With this estimate for τ_x then, a ratio estimate of τ is given by:

$$\widehat{\tau}_r = r\widehat{\tau}_x, \text{ where via linearization, the variance of } \widehat{\tau}_r \text{ is:}$$

$$\text{Var}(\widehat{\tau}_r) = N(N-n')\frac{\sigma^2}{n'} + N^2\left(\frac{n'-n}{n'}\right)\frac{\sigma_r^2}{n},$$

$$\text{where } \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2, \;\; \sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2, \;\; \mu = \tau/N \text{ and } R = \tau/\tau_x.$$

- With ratio estimation in Chapter 7, we had $n' = N$, so that the first term of the variance expression above was just zero. In that case, $\text{Var}(\widehat{\tau}_r) = N^2\left(\frac{N-n}{N}\right)\frac{\sigma_r^2}{n}$ as in Chapter 7.

- The estimated variance is given by:

$$\widehat{\text{Var}}(\widehat{\tau}_r) = N(N-n')\frac{s^2}{n'} + N^2\left(\frac{n'-n}{n'}\right)\frac{s_r^2}{n}, \text{ where:}$$

$$s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2, \; s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \text{ (usual sample variance).}$$

- The value of s_r^2 will be small (as before) if the relationship between the visual and actual estimates is linear and goes through the origin.

- The ratio estimate of the mean response and corresponding standard error are given by:

$$\widehat{\mu}_r = \frac{\widehat{\tau}_r}{N}, \; \text{SE}(\widehat{\mu}_r) = \frac{\text{SE}(\widehat{\tau}_r)}{N}.$$

Back to Example 1 (Biomass Example): To estimate the total biomass in the 1000 m^2 area or the mean biomass, the following R code was used.

```
> x <- c(20,80,150,40,80,60,60,200,0,100,80,20,150,100,80,60,
         60,20,150,60)
> y <- c(14,62,155,36,71)
> N <- 1000                    # Population size
> np <- 20                     # Initial sample size: n'
> n <- 5                       # Subsample size
> x1 <- x[1:5]                 # x-values for the subsample
```

$$\left(r = \left(\sum_{i=1}^{n} y_i\right) / \left(\sum_{i=1}^{n} x_i\right)\right)$$

```
> r <- sum(y)/sum(x1)
> r
[1] 0.9135135                  # Estimate of actual / visual
```

$$\left(\widehat{\tau}_x = \frac{N}{n'}\sum_{i=1}^{n'} x_i\right)$$

```
> tau.hat.x <- (N/np)*sum(x)
> tau.hat.x
[1] 78500                      # Estimate of tau.x for all 1000 plots
```

$$(\widehat{\tau}_r = r\widehat{\tau}_x)$$

```
> tau.hat.r <- r*tau.hat.x
> tau.hat.r
[1] 71710.81                   # Estimate of total biomass
```

```
> sr2 <- (1/(n-1))*sum((y-r*x1)^2)
> var.tau.hat.r <- (N*(N-np)*var(y))/
           np + N^2*((np-n)/np)*sr2/n
> sqrt(var.tau.hat.r)
[1] 12613.57                  # SE of tau.hat.r

> mu.hat.r <- tau.hat.r/N
> mu.hat.r
[1] 71.71081                  # Estimate of biomass per plot (gm/m sq)
> sqrt(var.tau.hat.r)/N
[1] 12.61357                  # SE of mu.hat.r
```

$$\left(s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2\right)$$

$$\left(\widehat{\mathrm{Var}}(\widehat{\tau}_r) = N(N-n')\frac{s^2}{n'} + N^2\left(\frac{n'-n}{n'}\right)\frac{s_r^2}{n}\right)$$

$$(\widehat{\mu}_r = \widehat{\tau}_r/N)$$

$$\left(\mathrm{SE}(\widehat{\mu}_r) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\tau}_r)}/N\right)$$

To investigate how much improvement these estimators with double sampling give, consider obtaining estimates of total biomass and biomass per plot based just on the y-values (i.e., an SRS of size $n = 5$).

```
> # Unbiased estimators not using visual estimate
> N*mean(y)              # Estimate of total biomass (SRS)
[1] 67600
> mean(y)                # Estimate of mean biomass (SRS)
[1] 67.6
> sqrt(N*(N-n)*(1/n)*var(y))   # SE of estimated total biomass (SRS)
[1] 23974.4
> sqrt(((N-n)/N)*(1/n)*var(y))     # SE of estimated mean biomass (SRS)
[1] 23.9744
```

$$(\widehat{\tau} = N\overline{y})$$

$$(\widehat{\mu} = \overline{y})$$

$$\left(\mathrm{SE}(\widehat{\tau}) = \sqrt{N(N-n)s^2/n}\right)$$

$$\left(\mathrm{SE}(\widehat{\mu}) = \sqrt{\frac{N-n}{N}\frac{s^2}{n}}\right)$$

Allocation in Double Sampling for Ratio Estimation

It was mentioned at the outset that one major reason for using double sampling is because auxiliary information may be cheaper to measure than the main variable of interest. Hence, the cost of sampling at the first stage (auxiliary info) as compared to the second stage (both auxiliary and response info) is very important in deciding how to allocate sample units at the two stages.

Suppose the total cost of sampling is fixed at C, and let $\boxed{C = c'n' + cn}$, where:

$$\begin{aligned} c' &= \text{cost of observing } x \text{ on one unit (visual estimate),} \\ c &= \text{cost of observing } y \text{ on one unit (actual estimate).} \end{aligned}$$

For a fixed cost (C, c', c), we want to find the optimal values of n, n', that is, those values that minimize the variance of the mean or total estimator. These optimal values can be shown to satisfy:

$$\frac{n}{n'} = \sqrt{\frac{c'}{c}\left(\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}\right)}.$$

- s_r^2 is approximately unbiased for σ_r^2 and s^2 is unbiased for σ^2. So we could use s_r^2 and s^2 from a preliminary or prior study as "guesses" of these standard deviations to answer the allocation question.

- In the biomass example, we found $s_r^2 = 117.2$ and $s^2 = 2888.3$. Hence, there was much more variation in the y's than between the y's and x's. This in turn makes $\left(\frac{s_r^2}{s^2 - s_r^2}\right)$ small. Is this typical? When?

- Using $\frac{s_r^2}{s^2 - s_r^2} = 0.0423$ as an estimate of the ratio $\frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}$, we can compute allocations for a variety of cost ratios:

c'/c	0.1	0.2	0.9	1.0
n/n'	.065	.091	.195	.206

- So, for example, if the cost ratio is $c'/c = 0.1$ (10 times more expensive to measure the actual biomass than to give a visual estimate), then the subsample (of the y's) should be about 6.5% the size of the visual sample (of the x's) (i.e., we would measure y on roughly every 15^{th} unit.)

Regression Estimation for Double Sampling

Suppose there is a linear realtionship between y and x which does not go through the origin. Then regression estimation will be more appropriate than ratio estimation.

- The regression estimate of the total in Chapter 8 was given as: $\widehat{\tau}_L = a + b\tau_x$ where a and b are the least squares estimators for the regression of y on x. With double sampling, however, τ_x is unknown and must be estimated:

- The resulting estimated variance of $\widehat{\tau}_L$ is:

$$\widehat{\mathrm{Var}}(\widehat{\tau}_L) = N(N - n')\frac{s^2}{n'} + N^2\left(\frac{n' - n}{n'}\right)\frac{1}{n(n-2)}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Double Sampling for Stratification

Recall that to take a stratified random sample, we need to know which strata the individual sampling units belong to before the sampling is done, as it is with this information that the sampling is actually performed. In cases where the stratum identifications are not known, but the relative population stratum sizes (N_h/N) are known, poststratification techniques can be applied to obtain desired estimators of the population mean or total, as outlined in Section 11.6 of Thompson. If, however, the relative population stratum sizes are not known, then they need to be estimated. One way to accomplish this is to take an initial large sample used to classify the units into strata and to estimate the relative stratum sizes, and then take a second stratified subsample from this first sample to measure the variable of interest.

Before describing how double sampling is used here, recall the following notation for stratified random sampling. Let:

$$\begin{aligned} N &= \text{population size,} \\ N_h &= \text{the size of stratum } h \text{ in the population,} \\ W_h &= N_h/N = \text{the proportion of the population in stratum } h. \end{aligned}$$

Double sampling in this stratified framework is conducted as follows in two steps:

1. Select an SRS of size n' and identify the stratum to which each of these observations belongs. Let n'_h be the number of observations in this "easy-to-take" sample from stratum h, $h = 1, \ldots, L$.
   - Let $w_h = \dfrac{n'_h}{n'}$ = the proportion of the sample in stratum h.
2. Take a stratified random sample from the n' units in the initial sample.
   - Let n_h = the number of units sampled from stratum h.

- Normally, the stratified random sample estimates the population mean μ by $\overline{y} = \displaystyle\sum_{h=1}^{L} \left(\frac{N_h}{N}\right) \overline{y}_h$. The difference here is that we are estimating N_h/N with n'_h/n' based on the initial sample. This gives the double sampling stratified estimator of the mean:

$$\overline{y}_d = \sum_{h=1}^{L} \left(\frac{n'_h}{n'}\right) \overline{y}_h, \quad \text{where } \overline{y}_h = \text{ the sample mean in the } h^{th} \text{ stratum.}$$

- The variance and estimated variance of $\overline{y}_d$ are given by:

$$\mathrm{Var}(\overline{y}_d) = \left(\frac{N-n'}{N}\right)\frac{\sigma^2}{n'} + \underbrace{\sum_{h=1}^{L}\frac{W_h\sigma_h^2}{n'}\left(\frac{n'_h}{n_h}-1\right)}_{\substack{\text{extra source of variability} \\ \text{from the secondary sample}}} \quad , \text{ estimated by:}$$

$$\widehat{\mathrm{Var}}(\overline{y}_d) = \underbrace{\left(\frac{N-1}{N}\right)\sum_{h=1}^{L}\left(\frac{n'_h-1}{n'-1}-\frac{n_h-1}{N-1}\right)w_h\frac{s_h^2}{n_h}}_{} + \underbrace{\left[\frac{N-n'}{N(n'-1)}\right]\sum_{h=1}^{L}w_h(\overline{y}_h-\overline{y}_d)^2}_{}$$

Notes:

- Suppose we take an SRS of 100 plots, where we plan to subsample 10 of these plots to measure the variable of interest. Normally, we would sample the "subsample plots" during the course of sampling the original 100 plots for efficiency reasons (i.e., the sampling would be done all at once). However, for a stratified random sample in this setting, we don't know the strata identifications until we sample the 100 plots.
- An alternative to this double sampling for stratification procedure might be to take a systematic sample at the first stage and at the second stage. This might be better at guaranteeing coverage of the area of interest. At the second stage (to mimic stratified random sampling), we could systematically select sites in each of the strata proportionally.
- In general, we may want to sample more units than dictated by the proportional stratum sizes to guarantee that at least 2 samples from each stratum are taken. Why?

Example 2: Using double sampling for stratification to handle nonresponse in a survey. Consider a population of $N = 400$ individuals on which we want to conduct a mail survey. The parameter of interest in this study is the proportion of people who respond with a "Yes" to a particular question. Call this population proportion p.

- Suppose an initial SRS of size $n' = 120$ people is taken where 30 people respond and the other 90 do not. On the basis of this sample, we will view the respondents as one stratum and the nonrespondents as a second stratum (call them strata 1 & 2 respectively). We view them as separate strata to allow for the (likely) possibility that the nonrespondents might have responded differently than the respondents, had they responded. So $L = 2$ here.

- In terms of the stratification notation, we have:

$$n_1' \quad = \quad 30 \text{ respondents, and } \quad n_2' = 90 \text{ nonrepondents.}$$

We then estimate the population stratum proportions by $w_1 = \dfrac{n_1'}{n'} = 0.25$ and $w_2 = \dfrac{n_2'}{n'} = 0.75$.

- Suppose 20 of the 30 respondents answered "Yes" to the question. Our initial estimate of the proportion who responded "Yes" is thus: $\overline{y}_1 = 20/30 = 0.667$.

- To try to obtain more information about the population of nonrespondents, suppose we now take an SRS of 25 nonrespondents from the 90 in our original sample. As the preliminary survey was a mail survey (which often has a low response rate), we might try a phone survey or even a face-to-face interview at this second stage. Suppose in this intensive follow-up, 20 of the 25 nonrespondents sampled respond and 4 answer "Yes" to the original question.

- Viewing this two-stage sample as a double sample, we have the following:

Stratum	Initial Sample Size	Secondary Sample Size	Stratum Mean
Respondents (1)	$n_1' = 30$	$n_1 = 30$	$\overline{y}_1 = 0.667$
Nonrespondents (2)	$n_2' = 90$	$n_2 = 25$	$\overline{y}_2 = 4/20 = 0.2$

Note that since we have already obtained a response from the 30 people in stratum 1, there is no need to survey them again.

- The estimated stratum mean for the nonrespondents stratum, namely $\overline{y}_2 = 0.2$ assumes that the 20 people actually sampled are representative of the 25 we attempted to sample.

- The resulting estimate of the proportion answering "Yes" to the question is:

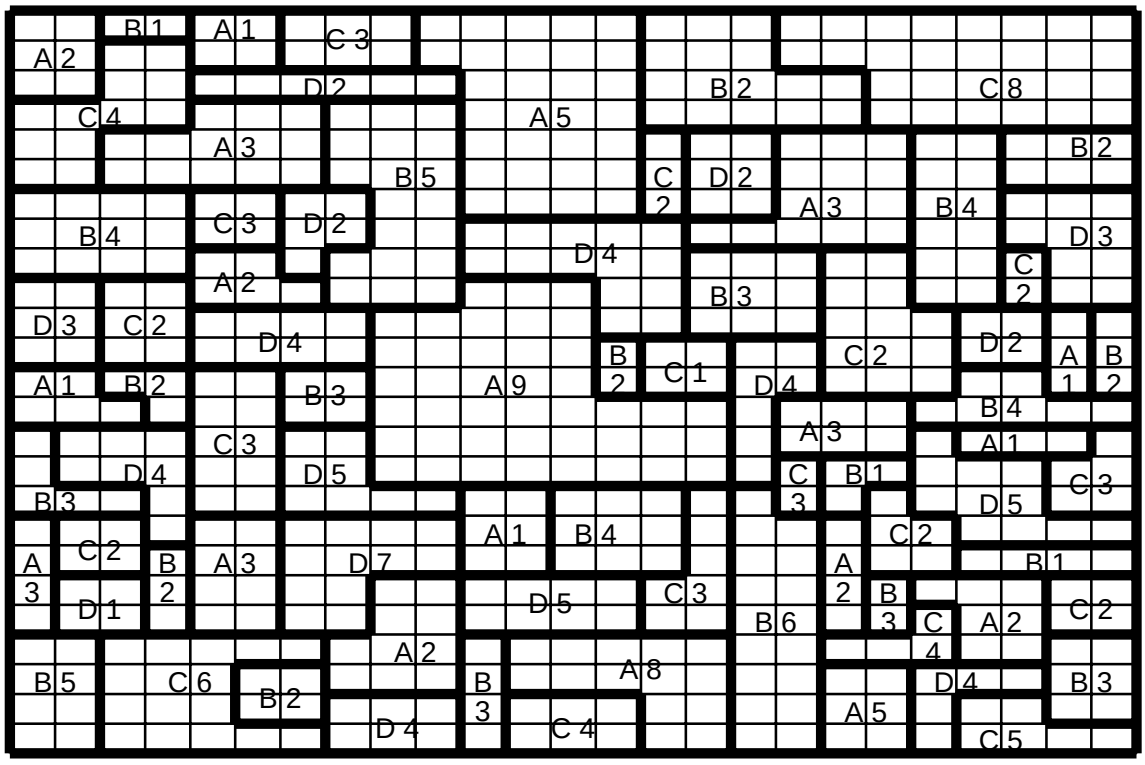
$$\widehat{p}_d = \overline{y}_d = \sum_{h=1}^{2} w_h \overline{y}_h = (.25)(.667) + (.75)(.2) = \underline{0.317},$$

a significant decrease from the estimate of 0.667 which ignored the nonrespondents.

Line-Intercept Sampling (Chapter 19)

Most of the sampling methods discussed thus far have consisted of choosing a sample from a well-defined sampling frame according to some probabilistic mechanism. Consider now taking a sample in some area where there are stationary "objects" of interest (e.g., ponds, shrubs, patches of a certain habitat type, wolf tracks in the snow) and the objective is to estimate the total number or density of the objects or the average or total value of some characteristic of the objects. Since we don't have a list of the objects, even conceptually, there is no way to take an SRS or related type of sample. One way to sample the objects is to randomly choose "lines" (transects) through the area, and sample all units which are intercepted by these lines; hence the term *line-intercept sampling.*

Example 1: Reconsider the farms example, on page 30 in Chapter 6 of the notes. The 25x25 grid map below represents the boundaries of farms in a certain area. The letters for each farm represent the "type" of farm. The numbers for each farm represent the number of workers on the farm. There are $N = 79$ total farms and $\tau_y = 249$ total workers.



The goal is to estimate the total number of farms, total number of workers per farm, the mean size of a farm, or the mean number of workers per farm. Recall that we sampled farms with probability proportional to size (PPS) by picking random points on the grid. We looked at Hansen-Hurwitz and Horvitz-Thompson estimators of the population parameters.

- For line-intercept sampling, consider taking line transects perpendicular to the base of the farms area. Since the length along this base is 25 units, we randomly select a number between 0 and 25, and sample the line drawn vertically from the chosen point.
- For each vertical transect chosen, we sample all farms intersected by the transect and record the variables of interest (the total number of farms will be estimated by letting $y_i = 1$).

We'll use this example to illustrate the calculations for line-intercept sampling later in this handout using R. First, we'll derive the estimators for line-intercept sampling.

Notation for Line-Intercept Sampling:
Let

$$
\begin{aligned}
K &= \text{the number of distinct objects in the population (\# of farms)} \\
y_k &= \text{the response variable for the } k^{th} \text{ object, } k = 1, \ldots, K \\
\tau &= \sum_{k=1}^{K} y_k = \text{the population total (total \# farms, or total \# workers)} \\
A &= \text{the total area of the study region (= 625 for the farms example)} \\
b &= \text{the baseline width of the study region (= 25 units for the farms example),} \\
w_k &= \text{the width along the baseline of the } k^{th} \text{ object}, k = 1, \ldots, K, \\
D &= \tau/A = \text{the density of the response variable per unit area (\# workers per unit}^2) \\
p_k &= \frac{w_k}{b} = \text{the probability the } k^{th} \text{ unit is intersected by a randomly chosen line transect} \\
&\quad \text{(selection probability)}, k = 1, \ldots, K, \\
\pi_k &= \text{the probability that the } k^{th} \text{ object is included in a sample of } n \text{ line transects} \\
&\quad \text{(inclusion probability)}, k = 1, \ldots, K, \\
\pi_{kh} &= \text{the joint inclusion probability of the } k^{th} \text{ and } h^{th} \text{ objects in a random sample of } n \\
&\quad \text{line transects, } k, h = 1, \ldots, K.
\end{aligned}
$$

Note that we use K instead of N to represent the total number of farms. This is because N represents the total number of sampling units in the population; when we sampled farms with PPS sampling, a farm was the sampling unit. In line-intercept sampling, the transect is the sampling unit and N, the total number of transects in the population, is assumed infinite.

Thompson presents two basic ways with line-intercept sampling to estimate population totals and means and calculate associated standard errors. For estimating a population total, these two ways are:

1. Separate transects estimator: calculate the Horvitz-Thompson estimate for each of the n transects separately, then average these n estimates to get an overall estimate of the population total.

2. Horvitz-Thompson estimator: calculate the Horvitz-Thompson estimate based on the distinct objects intercepted by the whole set of n transects.

These two approaches are discussed in more detail below.

Separate Transects Estimator

Use the information attained from a single line transect (say the i^{th} transect) to estimate the population total τ. Let C_i be the set of all objects in the region intersected by the i^{th} transect and let y_k be the response for the k^{th} object in the population ($k = 1, \ldots, K$). Then an unbiased estimator of the population total is given by

$$v_i = \sum_{k \in C_i} \frac{y_k}{p_k}.$$

This is the Horvitz-Thompson estimator for a single transect. We cannot calculate the Hansen-Hurwitz estimator because individual objects are not selected randomly with replacement.

- For n such vertical line transects, we obtain n estimates of τ, given by $v_1, v_2, \ldots, v_n$, where these estimates are independent and identically distributed. Hence, an unbiased estimator of τ and corresponding variance based on all n transects are given by:

$$\begin{aligned} \widehat{\tau}_p &= \frac{1}{n}\sum_{i=1}^{n} v_i \ \text{(the average of the } v_i\text{'s)}, \\ \text{Var}(\widehat{\tau}_p) &= \frac{1}{n}\text{Var}(v_i) = \frac{\sigma_v^2}{n} \ \text{(where } \sigma_v^2 \text{ is the variance of } v_i). \end{aligned}$$

- Using the sample variance of the v_i's as an estimate of σ_v^2, an unbiased estimate of the variance of $\widehat{\tau}_p$ is given by:

$$\widehat{\text{Var}}(\widehat{\tau}_p) = \frac{s_v^2}{n} \ \text{ where: } \ s_v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \widehat{\tau}_p)^2.$$

- Note that there is no finite population correction (fpc) here. Why not?

- Note that with this estimator, a given object might be intersected by more than one transect and hence be included in the sample more than once.

Horvitz-Thompson Estimator

Recall that the Horvitz-Thompson estimator gives a general way of acquiring an unbiased estimator

of a population total where distinct selected units are used but once in the development of the estimator. For line-intercept sampling, we use the whole set of objects intercepted by at least one of the n transects. Letting v be the number of distinct objects intercepted, recall that the general form of the Horvitz-Thompson estimator of the total and corresponding variance and estimated variance are given by:

$$\begin{aligned}
\widehat{\tau}_\pi &= \sum_{k=1}^{v} \frac{y_k}{\pi_k} \text{ (where } v = \text{the \# of distinct objects selected)}, \\
\text{Var}(\widehat{\tau}_\pi) &= \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i}\right) y_i^2 + \sum_{i=1}^{N} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_i y_j, \\
\widehat{\text{Var}}(\widehat{\tau}_\pi) &= \sum_{i=1}^{v} \left(\frac{1-\pi_i}{\pi_i^2}\right) y_i^2 + \sum_{i=1}^{v} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) \frac{y_i y_j}{\pi_{ij}},
\end{aligned}$$

where the π_i's and π_{ij}'s are the inclusion and joint inclusion probabilities.

- For example, if y_k = the number of workers on farm k, and v = the total number of distinct farms selected in the line transect samples, then an unbiased (H-T) estimate of the total number of workers in the farm region is given by:

$$\widehat{\tau}_\pi = \sum_{k=1}^{v} \frac{y_k}{\pi_k}.$$

To use the Horvitz-Thompson estimator for line-intercept sampling, we need to compute the inclusion and joint inclusion probabilities for all objects intercepted (sampled) by the n vertical transects. How do we do this?

π_k =

=

π_{kh} =

=

=

- One favorable aspect of the Horvitz-Thompson estimator over the separate transect estimator (as previously discussed) is the fact that it only relies on distinct units from the sample. The

major drawback is the difficulty with which the joint inclusion probabilities are computed. In addition, the estimated variance of a Horvitz-Thompson estimator is not guaranteed to be positive, while the separate transects variance estimator is.

Estimating the Density using the Total

With either of the two estimators for the population total just developed, an estimate of the density per unit area (that is, the mean number of objects per unit area) and corresponding variance are given by:

$$\widehat{D} = \frac{\widehat{\tau}}{A}, \quad \mathrm{Var}(\widehat{D}) =$$

where A = the total area of the study region.

- If the study region is rectangular with base of width b and constant transect length l, then $A = b\,l$. Note that the estimators for the total and the density do not rely on the transect length in any way.
- If the study region is irregularly-shaped, so that the transect lengths may vary from transect to transect, the estimators given are still unbiased estimators of the population quantities. However, if the lengths of the transects vary greatly, we might make use of any relationship between the response variable and the transect length via ratio or regression estimation to improve the estimators. This will be discussed later.

Example 2: (this is Example 1 on pp. 277-280 (2nd ed: pp. 247-250) of Thompson): Researchers were interested in estimating the abundance of wolverines in a certain region. Aircraft were flown over selected transects looking for tracks in the snow. Each set of tracks encountered along the transect was mapped. The variable of interest in the study was y_k = the number of wolverines associated with the k^{th} set of tracks. A map of the survey area is shown on p. 278 of Thompson (2nd ed: p. 248). The region had a baseline width of 36 miles and was 20 miles in length from the base.

Four systematic samples of 3 transects each were taken with random starting points in the first 12 miles given by A1, B1, C1, and D1 in the map. Each systematic sample of three transects was viewed as a single long transect of 60 miles, as would be the case if one stacked the three 12 mile sections vertically, so that the region was 12 miles wide by 60 miles long. The 12 selected transects intersected $v = 4$ distinct sets of tracks. Let

$$\begin{aligned}
y_k &= \text{the number of wolverines on the } k^{th} \text{ set of tracks } (k = 1, 2, 3, 4), \\
w_k &= \text{the width of the projection of the } k^{th} \text{ set of tracks onto the base of the region,} \\
p_k &= w_k/12 = \text{the probability of intersecting the } k^{th} \text{ set of tracks for a given transect,} \\
\pi_k &= 1 - (1 - p_k)^4 = \text{the inclusion probability of the } k^{th} \text{ set of tracks in the sample.}
\end{aligned}$$

The following table summarizes the information on the tracks:

k = Track #	y_k	w_k (miles)	$p_k = w_k/12$	π_k
1	1	5.25	.4375	.90
2	2	7.50	.6250	.98
3	2	2.40	.2000	.59
4	1	7.05	.5875	.97

Separate Transects Estimate of τ = Total # of Wolverines

The first transect (A1, A2, and A3) intersects the 1st, 2nd, and 4th sets of tracks, giving the following estimate of the total:

$$v_1 = \frac{y_1}{p_1} + \frac{y_2}{p_2} + \frac{y_4}{p_4} = \frac{1}{.4375} + \frac{2}{.6250} + \frac{1}{.5875} = 7.1878 \text{ wolverines},$$

and similarly, $v_2 = 7.1878$, $v_3 = 11.7021$, & $v_4 = 11.7021$. Hence, the estimated total number of wolverines in the region, and corresponding standard error are given by:

$$\begin{aligned} \widehat{\tau}_p &= \frac{1}{4}\sum_{i=1}^{4} v_i = \frac{1}{4}[7.1878 + 7.1878 + 11.7021 + 11.7021] = \underline{9.44 \text{ wolverines}}, \\ \widehat{\text{SE}}(\widehat{\tau}_p) &= \sqrt{\frac{s_v^2}{n}} = \sqrt{\frac{1}{4-1}\sum_{i=1}^{4}(v_i - \widehat{\tau}_p)^2/n} = \sqrt{\frac{6.7930}{4}} = \sqrt{1.698} = \underline{1.30 \text{ wolverines}}. \end{aligned}$$

Horvitz-Thompson Estimate of τ = Total # of Wolverines

Using the inclusion probabilities in the above table, the estimated total number of wolverines in the region is:

$$\widehat{\tau}_\pi = \sum_{k=1}^{v} \frac{y_k}{\pi_k} = \left[\frac{1}{.90} + \frac{2}{.98} + \frac{2}{.59} + \frac{1}{.97}\right] = \underline{7.57 \text{ wolverines}}.$$

The joint inclusion probabilities are computed on page 279 of Thompson (2nd ed: p. 250), with the standard error found to be 2.30 wolverines, which is nearly twice that of the separate transects estimate.

Example 1 continued (Farms Example): Recall the scenario given earlier where line transects are used to sample farms. R will be used to estimate population parameters of interest under two settings. First, we consider estimation using a single transect, and second we consider estimation using three transects. The reason for looking at a single transect first is to illustrate clearly the computation of the joint inclusion probabilities, and to examine the gains in terms of the SE's in taking more than one transect.

```
> runif(1,0,25)
[1] 10.78054         # Select x-coordinate for transect

# Record the widths of the 6 farms intersected
# ============================================
> wk <- c(1,4,2,8,5,5)
> pk <- wk/25        # Prob. of inclusion for a single transect
> pk
[1] 0.04 0.16 0.08 0.32 0.20 0.20
```

$$\left(p_k = \pi_k = \frac{w_k}{b}\right)$$

```
# Record overlap of the farms for computation of joint inclusion probabilities
# ============================================================================
> wkh <- matrix(0,nrow=6,ncol=6)
> wkh[1,2] <- 1; wkh[1,3] <- 1; wkh[1,4] <- 1; wkh[1,5] <- 1
> wkh[1,6] <- 1; wkh[2,3] <- 2; wkh[2,4] <- 4; wkh[2,5] <- 4
> wkh[2,6] <- 4; wkh[3,4] <- 2; wkh[3,5] <- 2; wkh[3,6] <- 2
> wkh[4,5] <- 5; wkh[4,6] <- 5; wkh[5,6] <- 4
> pkh <- wkh/25
> pkh
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0 0.04 0.04 0.04 0.04 0.04
[2,]    0 0.00 0.08 0.16 0.16 0.16
[3,]    0 0.00 0.00 0.08 0.08 0.08
[4,]    0 0.00 0.00 0.00 0.20 0.20
[5,]    0 0.00 0.00 0.00 0.00 0.16
[6,]    0 0.00 0.00 0.00 0.00 0.00
```

$$\left(\pi_{kh} = \frac{w_{kh}}{b}\right)$$

```
# Estimate the total number of farms (N).  Here, yi=1 for all farms
# ================================================================
> yk <- rep(1,6)
> yk
[1] 1 1 1 1 1 1
> tau.hat <- sum(yk/pk)
> tau.hat            # H-T Estimate of
[1] 56.875           #   N  (N=79 here)
```

$$\left(\widehat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{\pi_i}\right)$$

```
# Estimate the variance of the estimated total number of farms (Eq. 6, p.54)
# ============================================================================
> c1 <- sum((1/pk^2 - 1/pk)*yk^2)
> c2 <- 0
> for (k in 1:5){
   for (h in (k+1):6){
    c2 <- c2 + 2*((1/(pk[k]*pk[h]) - 1/pkh[k,h])*yk[k]*yk[h])
   }
  }
> var.tau.hat <- c1 + c2
> sqrt(var.tau.hat)  # Estimated SE of tau.hat
[1] 52.51562

# Estimate the total number of workers
# ====================================
> yk <- c(3,5,1,9,4,5)
> tau.hat <- sum(yk/pk)
> tau.hat            # Estimated number of workers
[1] 191.875
> c1 <- sum((1/pk^2 - 1/pk)*yk^2)
> c2 <- 0
> for (k in 1:5){
   for (h in (k+1):6){
    c2 <- c2 + 2*((1/(pk[k]*pk[h]) - 1/pkh[k,h])*yk[k]*yk[h])
   }
  }
> var.tau.hat <- c1 + c2
> sqrt(var.tau.hat)  # Estimated SE of tau.hat
[1] 172.0585
```

$$\left(c_1 = \sum_{k=1}^{v} \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k} \right) y_k^2 \right)$$

$$\left(c_2 = 2 \sum_{k=1}^{v} \sum_{h>k} \left(\frac{1}{\pi_k \pi_h} - \frac{1}{\pi_{kh}} \right) y_k y_h \right)$$

Now, suppose three transects are selected at random. We will use the transect selected above and just select two more below.

```
> runif(2,0,25)
[1] 7.020874 4.758928   # Two more randomly selected x-coordinates

# Widths and Selection Probabilities for all 3 Transects
# ======================================================
> w1 <- c(1,4,2,8,5,5)
```

```
> w2 <- c(3,3,4,4,2,4,3,2,6,3)
> w3 <- c(5,2,2,4,3,2,5,6,2)
> p1 <- w1/25; p2 <- w2/25; p3 <- w3/25

# Estimate the total number of farms (separate transects)
#========================================
> y1 <- rep(1,6)
> y2 <- rep(1,10)
> y3 <- rep(1,9)
> v1 <- sum(y1/p1)
> v2 <- sum(y2/p2)
> v3 <- sum(y3/p3)
> c(v1,v2,v3)
[1] 56.875 81.250 78.750
> mean(c(v1,v2,v3))          # Estimated total number
[1] 72.29167                 #   of farms
> sqrt(var(c(v1,v2,v3))/3)
[1] 7.742043                 # SE of the estimate

# Estimate the total number of workers (separate transects)
#=========================================
> y1 <- c(3,5,1,9,4,5)
> y2 <- c(4,2,7,5,3,4,5,2,2,3)
> y3 <- c(6,3,3,4,2,3,3,2,1)
> v1 <- sum(y1/p1)
> v2 <- sum(y2/p2)
> v3 <- sum(y3/p3)
> c(v1,v2,v3)
[1] 191.875 287.500 220.000
> mean(c(v1,v2,v3))
[1] 233.125                  # Estimated total number of workers
> sqrt(var(c(v1,v2,v3))/3)
[1] 28.3739                  # SE of the estimate
```

$$\left(\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} v_i\right)$$

$$\left(\mathrm{SE}(\widehat{\tau}_p) = \sqrt{\frac{s_v^2}{n}}\right)$$

```
# Horvitz-Thompson Estimates: Use only DISTINCT farms
# ===================================================
> w1 <- c(1,4,2,8,5,5)
> w2 <- c(3,3,4,4,2,4,3,2,6,3)
> w3 <- c(5,2,2,3,2,5,2)
> p1 <- w1/25
> p2 <- w2/25
> p3 <- w3/25

# Estimate the total number of farms (H-T)
# ========================================
> y1 <- rep(1,6)
> y2 <- rep(1,10)
> y3 <- rep(1,7)
> p <- c(p1,p2,p3)
> y <- c(y1,y2,y3)
> sum(y/(1-(1-p)^3))
[1] 77.26119
```

$$\left(\widehat{\tau}_{\pi} = \sum_{k=1}^{v} \frac{y_k}{\pi_k} = \sum_{k=1}^{v} \frac{y_k}{1-(1-p_k)^n} \right)$$

```
# Estimate the total number of workers (H-T)
# ==========================================
> y1 <- c(3,5,1,9,4,5)
> y2 <- c(4,2,7,5,3,4,5,2,2,3)
> y3 <- c(6,3,3,2,3,3,1)
> p <- c(p1,p2,p3)
> y <- c(y1,y2,y3)
> sum(y/(1-(1-p)^3))
[1] 253.6919
```

The standard errors for the Horvitz-Thompson estimates were not calculated here, as the code becomes increasingly lengthy with more than one transect.

Some Extensions of Line Intercept Sampling

Assume we have a random sample of n transects and that we will use the separate transects estimator $\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} v_i$ (p. 276 of Thompson, 2nd ed: p. 246) to estimate the population total τ of a variable of interest. We will estimate its variance by $\widehat{\text{Var}}(\widehat{\tau}_p) = s_v^2/n$. That is, we will calculate the Horvitz-Thompson estimate for each transect separately and then average the estimates to get an overall estimate.

Recall that w_k is the width of the k^{th} object on a transect and so its probability of inclusion for a single transect is $p_k = w_k/b$, where b is the width of the baseline. Recall also that C_i is the collection of sampled objects on the i^{th} transect. The w_k's should also have a subscript i indicating which transect the object is on, but Thompson has chosen to suppress the additional subscript with the understanding that the w_k's are different for each transect.

There are two basic extensions of line intercept sampling as discussed in Thompson which are given here. The first considers the case of estimating the mean response per object (as opposed to per unit area). The second considers the situation of variable lengths in the transects.

Extensions:

1. Estimating μ, the Per Object Mean: While Thompson considers estimating τ, the total of the y-values for the population, and $D = \tau/A$, the density as number per unit area, he does not consider estimating the mean value of y per object, $\mu = \tau/K$, where K is the number of objects in the population.

   - For example, suppose we wanted to estimate the mean height of shrubs sampled via this method. If K is known, the estimation is straightforward:
   $$\widehat{\mu} = \frac{\widehat{\tau}}{K}, \quad \widehat{\text{Var}}(\widehat{\mu}) = \frac{\widehat{\text{Var}}(\widehat{\tau})}{K^2}.$$
   - However, if K is not known (as would usually be the case), then K must also be estimated and a ratio of means estimator used. First, to estimate K using the method in equation (19.3) on page 276 $\left(\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n} v_i \text{ where } v_i = \sum_{k\epsilon C_i} \frac{y_k}{p_k}\right)$, we first estimate K for each transect separately using a Horvitz-Thompson estimator (with $y_k = 1$ for all objects) and then average the individual transect estimates of K:
   $$\widehat{K}_i = \sum_{k\in C_i} \frac{1}{w_k/b} = b\sum_{k\in C_i} \frac{1}{w_k}, \text{ where then: } \widehat{K} = \frac{1}{n}\sum_{i=1}^{n} \widehat{K}_i = \frac{b}{n}\sum_{i=1}^{n}\sum_{k\in C_i} \frac{1}{w_k}.$$
   The estimate of the population total τ is:
   $$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{k\in C_i} \frac{y_k}{w_k/b} = \frac{b}{n}\sum_{i=1}^{n}\sum_{k\in C_i} \frac{y_k}{w_k}.$$

Combining the estimates of K and τ, a ratio estimator of μ is:

$$\widehat{\mu} = \frac{\widehat{\tau}}{\widehat{K}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{k\in C_i}\frac{y_k}{w_k}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{k\in C_i}\frac{1}{w_k}} = \frac{\frac{1}{n}\sum_{i=1}^{n} t_i}{\frac{1}{n}\sum_{i=1}^{n} u_i} = \frac{\overline{t}}{\overline{u}},$$

a ratio of two means, where $t_i = \sum_{k\in C_i}\frac{y_k}{w_k}$ and $u_i = \sum_{k\in C_i}\frac{1}{w_k}$.

- To estimate the variance, we can use linearization (see the notes on linearization, part 3, for the case where N is unknown - page 42 in your notes) or bootstrapping of the pairs. The linearization approximation to the variance is:

$$\widehat{\mathrm{Var}}(\widehat{\mu}) \approx \left(\frac{\overline{t}^2}{\overline{u}^4}\right)\frac{s_u^2}{n} + \left(\frac{1}{\overline{u}^2}\right)\frac{s_t^2}{n} - 2\left(\frac{\overline{t}}{\overline{u}^3}\right)\frac{\widehat{\rho}_{t,u}s_t s_u}{n},$$

where s_t^2 and s_u^2 are the sample variances and $\widehat{\rho}_{t,u}$ is the sample correlation between the t_i's and the u_i's. Note that b is not needed in either formula.

2. Transects of Variable Length: If the region is irregularly shaped, then the transects may vary in length. Let $L_1, \ldots, L_n$ denote the lengths of the n randomly chosen transects. The L_i are then random variables whose values are unknown until the transects are selected.

   - Let $\mathrm{E}(L)$ be the *expected* length of a random transect. Then $\mathrm{E}(L) = A/b$ where A is the total area of the region and b is the length of the baseline.
   - The separate transects and Horvitz-Thompson estimators in Section 19.1 of Thompson (and on 136 and p. 136 of the notes) are still unbiased even if transect lengths differ and we can, if we choose, ignore the differing lengths of the transects.
   - However, we might want to consider alternative estimators for two reasons:

     (i) Estimators which account for the differing transect lengths might be better (i.e., have smaller variance).

     (ii) We might not know the total area A for an irregularly shaped region and might want alternative estimators which do not require knowledge of the total area.

   Each of these reasons is explored below.

(i) Accounting for Different Transect Lengths: Consider estimation of τ, the population total. The unbiased estimator of τ (eq.(19.3) on p. 276 of Thompson; 2nd ed: eq. (3) on p. 246) is constructed by computing the Horvitz-Thompson (H-T) estimator of τ on each transect and then averaging the estimates. The H-T estimator on each transect is:

$$v_i = \sum_{k\in C_i}\frac{y_k}{p_k} = b\sum_{k\in C_i}\frac{y_k}{w_k} = bt_i, \quad \text{where } t_i = \sum_{k\in C_i}\frac{y_k}{w_k}.$$

- It seems clear that v_i based on longer transects will tend to be larger than those based on shorter transects. Therefore, we could use transect length as an auxiliary variable if we knew $\mathrm{E}(L)$, the average transect length for the whole area. If we view the transect as the sampling unit and v_i as the response, then we want to use ratio estimation to estimate the mean value of v for the whole population of possible transects. Note that the mean value of v for the population of all possible transects is the population total τ of the variable of interest. That is why we use the ratio estimator for a mean in the formulas on p. 94 of Thompson (2nd ed: p. 68). From those formulas, with y_i representing v_i, x_i representing L_i, and $\mu_x = \mathrm{E}(L)$, the ratio estimate of τ is:

$$\widehat{\tau}_r = \left(\frac{\sum_{i=1}^{n} v_i}{\sum_{i=1}^{n} L_i} \right) \mathrm{E}(L) = \frac{b\,\mathrm{E}(L) \sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} L_i} = \frac{A \sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} L_i}. \tag{8}$$

- The variance of $\widehat{\tau}_r$ can be approximated by equation (7.5) on page 95 of Thompson (2nd ed: eq. (5) on p. 69):

$$\widehat{\mathrm{Var}}(\widehat{\tau}_r) = \frac{s_r^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (v_i - r L_i)^2,$$

where $r = \sum_{i=1}^{n} v_i / \sum_{i=1}^{n} L_i$ and where the finite population correction (fpc) has been ignored because the population of possible transects is assumed to be infinite.

(ii) Estimating D when A is Unknown: Consider estimation of $D = \tau/A$, the population density per unit area. Based on equation (8) above, an estimate of D is:

$$\widehat{D}_r = \frac{\widehat{\tau}_r}{A} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} L_i}, \quad \text{where } t_i = \sum_{k \in C_i} \frac{y_k}{w_k}.$$

Note that $\widehat{D}_r$ does not depend on knowing A or $\mathrm{E}(L)$.

- If each object intersected by a transect represents one animal, then the density of animals is estimated by setting $y_k = 1$. However, if each object can represent more than one animal (such as the wolverine track example on p. 277 of Thompson; 2nd ed.: p. 247) then the y_k are the number of animals associated with each object. If objects encountered are bushes and y_k is the weight of berries on a bush, then D is the average weight of berries per unit area.

- The estimated variance of $\widehat{D}_r$ is (see the formulas at the top of p. 70 of the text):

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{D}_r) &= \frac{\widehat{\mathrm{Var}}(\widehat{\tau}_r)}{A^2} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{v_i - rL_i}{A} \right)^2 \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{t_i - L_i \sum_{j=1}^{n} t_j / \sum_{j=1}^{n} L_j}{E(L)} \right)^2 \\
&= \frac{1}{n\mathrm{E}(L)^2} \left[\frac{1}{n-1} \sum_{i=1}^{n} (t_i - L_i \widehat{D}_r)^2 \right]
\end{aligned}$$

  since $v_i = bt_i$ and $A = b\mathrm{E}(L)$. If $\mathrm{E}(L)$ is not known, it can be estimated by $\overline{L} = \frac{1}{n} \sum_{i=1}^{n} L_i$, the sample mean transect length.

- Note that the ratio estimate of D based on a single transect of length L_i would be

$$\widehat{D}_i = \frac{\frac{bt_i}{L_i}\mathrm{E}(L)}{A} = \frac{t_i}{L_i}$$

  so that

$$\widehat{D}_r = \frac{\sum_{i=1}^{n} L_i \widehat{D}_i}{\sum_{i=1}^{n} L_i}$$

  is a weighted average of the ratio estimates of density from the individual transects.

Transect Sampling Experiment

Suppose the desks in this classroom are randomly dispersed, and we use transect sampling to estimate the proportion of the room covered by desks, as well as the total number of desks in the room. Note that both of these quantities can be computed exactly.

- To estimate these population quantities via line intercept sampling, the front wall is considered the base of the classroom, and four transects are randomly selected. Along these line transects, we count the number of desks intersected, and need to measure the inclusion and joint inclusion probabilities of all desks encountered. How might we do this?
- The length of the baseline is 306 inches (=777 cm). Four random transects will be selected (we could do a systematic sample with interval $306/4 = 76$ inches and random starting point from 0 to 76 and treat it as an SRS). Breaking into groups, we want to record the following for each desk intersected by a transect (The transects are each of length 274 inches = 696 cm):
  1. The vertical length of intersection with the desk (i.e.: how much of the surface of the desk is intersected along the transect?)
  2. The horizontal width of the desk to the left of the transect.
  3. The horizontal width of the desk to the right of the transect.
- The resulting data are recorded in the table below:

Transect 1			Transect 2			Transect 3			Transect 4		
Length	Left Width	Right Width	Length	Left Width	Right Width	Length	Left Width	Right Width	Length	Left Width	Right Width

- The total number of desks in the room can be estimated in two ways:
  1. Treating the desks encountered as one big sample, the total number of desks can be estimated with a Horvitz-Thompson estimator, through computation of the inclusion and joint inclusion probabilities.
  2. Treating the transects separately, a Horvitz-Thompson estimate v_i can be computed for each of the four transects separately and then averaged to give $\widehat{\tau}$.

- There are two ways we could estimate the total area covered by the desks:

  1. Let y_i = the length of transect i which intersects a desk; then a ratio estimator of the area covered by desks is: $\widehat{\tau}_r = A \cdot \sum y_i / \sum L_i$, where A = the area of the region (the room in this case). The estimated proportion of the area covered by desks is $\sum y_i / \sum L_i$.
  2. Rather than just measuring the length of each transect which falls on each desk, we could measure the area y_k of each intercepted desk. Then we could use either a separate transects estimate or Horvitz-Thompson estimate of the total area covered by desks. The separate transects estimate is

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{4} v_i, \quad \text{where: } v_i = \sum_{k \in C_i} \frac{y_k}{w_k / b}$$

     If the area of each desk is the same, say y, then this simplifies to $\widehat{\tau} = y \sum_{k \in C_i} \frac{1}{w_k/b} = y\widehat{K}$, where $\widehat{K}$ is the estimated total number of desks. If we know the area A of the room, then an alternative estimate of the proportion of the room covered is $\widehat{\tau}/A$; if the desks are all the same size, this equals $y\widehat{K}/A$.

Detectability and Sampling (Chapter 16)

To this point, all sampling methods considered have assumed that the variable of interest is measured without error and that the only source of variation is natural variation between the observed sampling units. Particularly in situations where we count the number of some species in some subplot of a large region, it is not always the case that detection of all such species is "perfect." In fact, with many elusive animal species (birds, fish, bears, etc.), detectability is far from perfect, and we need to account for the probability of detection of such species in estimating species population totals or means.

Consider some region within which we want to estimate the total number of objects, or the mean number of objects per unit area. To carry out this type of estimation under imperfect detectability, we introduce the following notation:

τ = the actual total # of objects in the whole region,

A = the area of the region,

D = τ/A = the actual density of objects per unit area in the region,

y = the observed # of objects in the region under imperfect detectability,
[Note that y here is a random variable]

p = the probability of detection of any object (assumed equal for all objects).

- Assume also that the detections are independent of one another (i.e., the fact that one object is detected has no bearing on whether or not any other object is detected). Is this reasonable?

- What possible values could the random variable y have?

- Under the assumption of independent detections, and equal probabilities of detection on each object, the distribution of y is: _____

  with mean: _____ and variance: _____.

We will consider four situations: a survey of a whole region with known detectability, a survey fo a whole region with estimated detectability, a survey of an SRS of plots with known detectability, and a survey of an SRS of plots with estimated detectability. The goal is to estimate τ and/or D.

1. Survey of whole region with known detectability:
   Assume we detect y animals in the region where the probability of detection p is known.

Since $\mathrm{E}(y) = \tau p$ and the observed y is an unbiased estimate of $\mathrm{E}(y)$, then y/p is an unbiased estimator of τ. Then,

- the estimated total and corresponding variance are given by:

$$\boxed{\widehat{\tau} = \frac{y}{p}} \text{ and } \boxed{\mathrm{Var}(\widehat{\tau})} = \mathrm{Var}\left(\frac{y}{p}\right) = \frac{\tau p(1-p)}{p^2} = \boxed{\frac{\tau(1-p)}{p}}.$$

- the estimated variance of $\widehat{\tau}$ is given by: $\widehat{\mathrm{Var}}(\widehat{\tau}) = \dfrac{\widehat{\tau}(1-p)}{p}$.
- the estimated density of objects per unit area in the region is:

$$\widehat{D} = \frac{\widehat{\tau}}{A} = \frac{y}{pA}, \;\; \mathrm{Var}(\widehat{D}) = \frac{\tau(1-p)}{A^2 p}, \;\; \widehat{\mathrm{Var}}(\widehat{D}) = \frac{\widehat{\tau}(1-p)}{A^2 p}.$$

Example 1 (Problem 1 on p. 227 of Thompson; 2nd ed: p. 197): In an aerial survey in Alaska, 82 moose were detected. Intensive independent studies determined the probability of detection to be 0.89. Estimate the total number of moose in the study region and estimate the variance of that estimate.

Here, $y = 82$ moose, and $p = 0.89$, so we estimate:

$$\begin{aligned} \widehat{\tau} &= \frac{y}{p} = \frac{82}{.89} = 92.135 \approx \underline{92.1 \text{ moose}}, \text{ with standard error:} \\ \widehat{\mathrm{SE}}(\widehat{\tau}) &= \sqrt{\frac{\widehat{\tau}(1-p)}{p}} = \sqrt{\frac{92.135(1-.89)}{.89}} = \sqrt{11.3874} \approx \underline{3.37 \text{ moose}}. \end{aligned}$$

- We assume that p is somehow known here, but, in practice, it must be estimated. Using the SE formula above ignores the uncertainty in the estimate of p.
- If p is estimated within the same study, by, for example, ground-truthing the aerial estimates on a subset of plots in the study area, then $\widehat{\tau}$ above is the same as the ratio estimator (because p would be the reciprocal of the ratio r of actual to visual; see Chap. 7) and the standard error should be estimated by the formula for ratio estimation and not the formula above.
- If the estimate of p comes from another study independent of the current one and if there is a standard error associated with the estimate, we should use the methods described below for estimated detectability.
- Methods of estimating detectability include mark-recapture methods, radio-collaring methods, distance-based methods, and regression-based methods.
- Whatever was done, it is important to recognize that if p comes from outside the current study, then it is considered an independent estimate of the detectability.

2. Survey of whole region with estimated detectability

Suppose now that instead of assuming the detectability p is known, p is in fact estimated independently by $\widehat{p}$ with some variance $\text{Var}(\widehat{p})$. Then τ is estimated through a ratio estimator given by:

$$\begin{aligned}
\widehat{\tau} &= \frac{y}{\widehat{p}}, \text{ with variance given by:} \\
\text{Var}(\widehat{\tau}) &\approx \left(\frac{\mu_y^2}{\mu_{\widehat{p}}^4}\right)\text{Var}(\widehat{p}) + \frac{1}{\mu_{\widehat{p}}^2}\text{Var}(y) \text{ (linearization method)} \\
&= \left(\frac{\tau^2 p^2}{p^4}\right)\text{Var}(\widehat{p}) + \frac{1}{p^2}\text{Var}(y) = \frac{1}{p^2}\left[\text{Var}(y) + \tau^2\text{Var}(\widehat{p})\right] \\
&= \frac{1}{p^2}\left[\tau p(1-p) + \tau^2\text{Var}(\widehat{p})\right] \\
&= \underbrace{\tau\left(\frac{1-p}{p}\right)}_{\substack{\text{variation due to} \\ \text{imperfect detection}}} + \underbrace{\frac{\tau^2}{p^2}\text{Var}(\widehat{p})}_{\text{variation in } \widehat{p}}.
\end{aligned}$$

- The estimated variance comes from the linearization approximation formula #2 on page 50. There is no covariance term in the variance approximation as it is assumed that the current survey is independent of the one used to estimate p.
- The estimated approximate variance (taking into account the effect of estimated detectability) is

$$\widehat{\text{Var}}(\widehat{\tau}) = \widehat{\tau}\left(\frac{1-\widehat{p}}{\widehat{p}}\right) + \frac{\widehat{\tau}^2}{\widehat{p}^2}\widehat{\text{Var}}(\widehat{p}).$$

Back to Example 1 (Moose): Suppose in addition to being told that $y = 82$ moose were detected and that the detectability was estimated to be $\widehat{p} = .89$, we are told that $\widehat{\text{SE}}(\widehat{p}) = .05$. Then

$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\tau}) &= \frac{92.135(1-.89)}{.89} + \frac{92.135^2}{.89^2}(.05)^2 = 11.3874 + 26.7923 = 38.180 \\
&\Longrightarrow \widehat{\text{SE}}(\widehat{\tau}) = \underline{6.18}.
\end{aligned}$$

- Note that this SE is almost double what it was in the earlier calculation (3.37). Hence, taking into account the variation in $\widehat{p}$ demonstrates how badly we underestimated the SE initially, and how important it is to take this extra source of variation into account.
- It will usually be the case (and certainly should be the case!) that an estimate $\widehat{p}$ of p includes an estimate of the variability of $\widehat{p}$. Unfortunately, independent estimates taken

from other papers are often treated as "truth" without any consideration of the variability underlying such estimates.

3. Survey of SRS of plots with known detectability

Suppose we take an SRS (without replacement) of size n from a population of N units. We might consider units to be plots within some region, where animals within a selected plot are detected with constant probability p, independently. Let:

$$\begin{aligned} Y_i &= \text{the actual number of objects (animals) in unit } i,\ i = 1, \ldots, N, \\ y_i &= \text{the observed number of objects in unit } i, \end{aligned}$$

so that $y_i \sim \text{Bin}(Y_i, p)$, where we assume for now that p is known. The goal, as before, is to estimate the population total number of objects $\tau = \sum_{i=1}^{N} Y_i$.

- We know from earlier that for a given sampled unit i:

$$\widehat{Y}_i = \frac{y_i}{p}, \ \ \text{Var}(\widehat{Y}_i) = \frac{y_i(1-p)}{p}, \ \ \text{where: } E(\widehat{Y}_i) = Y_i.$$

  Then an unbiased estimate of the population total τ is:

$$\boxed{\widehat{\tau}} = \frac{N}{n}\sum_{i=1}^{n}\widehat{Y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{y_i}{p} = \boxed{N\frac{\overline{y}}{p}}, \ \ \text{where: } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- The variance of $\widehat{\tau}$ (derived later in these notes and in Sec. 16.7 of Thompson) is

$$\text{Var}(\widehat{\tau}) = N^2\left[\underbrace{\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}}_{} + \underbrace{\left(\frac{1-p}{p}\right)\frac{\mu}{n}}_{}\right],$$

  where: $\mu = \dfrac{\tau}{N}$ and $\sigma^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2$ is the natural variability in the population units.

- An unbiased estimator of $\text{Var}(\widehat{\tau})$ is given by

$$\widehat{\text{Var}}(\widehat{\tau}) = \frac{N^2}{p^2}\left[\left(\frac{N-n}{N}\right)\frac{s^2}{n} + \left(\frac{1-p}{N}\right)\overline{y}\right]$$

  where $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is the sample variance of the observed counts.

  Note: s^2 does not estimate σ^2 (as was the case earlier with two-stage sampling). s^2

underestimates σ^2. Variability in observed counts tends to be less than the variability in the true counts because the imperfect detectability causes observed counts to be lower and have lower variability than the true counts. The quantity s^2/p^2 estimates σ^2.

4. Worst Case: Survey of SRS of plots with estimated detectability:
   Suppose now we take an SRS of n units from a population of N units, where p is unknown and is estimated independently by $\widehat{p}$ with variance $\text{Var}(\widehat{p})$. As before for p unknown, the population total τ is estimated via a ratio estimator:

$$\widehat{\tau} = \frac{N\overline{y}}{\widehat{p}}, \quad \text{with approximate variance given by:}$$

$$\text{Var}(\widehat{\tau}) \approx N^2 \left[\underbrace{\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}}_{\substack{\text{variability} \\ \text{due to SRS}}} + \underbrace{\left(\frac{1-p}{p}\right)\frac{\mu}{n}}_{\substack{\text{variability due to} \\ \text{imperfect detectability}}} + \underbrace{\frac{\mu^2}{p^2}\text{Var}(\widehat{p})}_{\substack{\text{variability due to} \\ \text{estimating } p}} \right] \left(\begin{array}{c} \text{Delta} \\ \text{Method} \end{array} \right)$$

   This variance is estimated by:

$$\widehat{\text{Var}}(\widehat{\tau}) = \frac{N^2}{\widehat{p}^2}\left[\left(\frac{N-n}{N}\right)\frac{s^2}{n} + \left(\frac{1-\widehat{p}}{N}\right)\overline{y} + \frac{\overline{y}^2}{\widehat{p}^2}\widehat{\text{Var}}(\widehat{p})\right].$$

   - There are 3 variance components here to account for all 3 levels of estimation (SRS, Imperfect Detectability, Estimation of p).

Example 2 (Problem 4, p. 227 of Thompson; 2nd ed: p. 197): Suppose an SRS of $n = 5$ plots is selected from a study area of $N = 100$ plots and that the numbers of animals detected in the five plots are 10, 7, 0, 0, and 5, but that the probability of detection for any animal in a selected plot is $p = .80$. Estimate the total number of animals in the study region and estimate the variance of the estimator.

Doing the calculation in R:

```
> N <- 100
> n <- 5
> y <- c(10,7,0,0,5)
> p <- .8
> N*mean(y)/p    # Estimate of the total number
[1] 550          #   of animals in the region
```

$\left(\widehat{\tau} = N\frac{\overline{y}}{p}\right)$

The estimated variability will be computed in three ways: assuming no error in estimating p (detectability known), assuming $\text{SE}(\widehat{p}) = .05$, and assuming $\text{SE}(\widehat{p}) = .2$, for the sake of comparison.

```
# If Detectability Known
# ======================================
> var.srs <- (N^2/p^2)*((N-n)/N)*var(y)/n
> var.srs
[1] 57296.9      # This is the major contribution to the variability.
> var.det <- (N^2/p^2)*((1-p)/N)*mean(y)
> var.det
[1] 137.5        # Variability due to detectability is minor.
> sqrt(var.srs + var.det)
[1] 239.6547     # SE of tau-hat
```

$$\left(\widehat{\mathrm{Var}}_{\mathrm{SRS}}(\widehat{\tau}) = \frac{N^2}{p^2}\left(\frac{N-n}{N}\right)\frac{s^2}{n}\right)$$

$$\left(\widehat{\mathrm{Var}}_{\mathrm{ID}}(\widehat{\tau}) = \frac{N^2}{p^2}\left(\frac{1-p}{N}\right)\overline{y}\right)$$

$$\left(\mathrm{SE}(\widehat{\tau}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\tau})} = \sqrt{\widehat{\mathrm{Var}}_{\mathrm{SRS}}(\widehat{\tau}) + \widehat{\mathrm{Var}}_{\mathrm{ID}}(\widehat{\tau})}\right)$$

```
# If Detectability Estimated with SE of .05
# ========================================
> var.srs <- (N^2/p^2)*((N-n)/N)*var(y)/n
> var.det <- (N^2/p^2)*(((1-p)/N)*mean(y) + (mean(y)^2/p^2)*.05^2)
> sqrt(var.srs + var.det)
[1] 242.1074     # SE of tau-hat

# If Detectability Estimated with SE of .2
# ========================================
> var.srs <- (N^2/p^2)*((N-n)/N)*var(y)/n
> var.det <- (N^2/p^2)*(((1-p)/N)*mean(y) + (mean(y)^2/p^2)*.2^2)
> sqrt(var.srs + var.det)
[1] 276.2981    # SE of tau-hat
```

- Note that the estimated SE with detectability estimated with a SE of .05 is not much different than having no error in the estimation of p (i.e., assuming p is known).
- The estimated SE with detectability estimated with a SE of 0.2 is somewhat larger than when p is assumed known, as the third piece of the variance above is now larger relative to the first component. However, an SE of 0.2 in estimating a proportion is a huge amount of uncertainty; one would expect smaller SE's in practice.
- The bottom line in this problem is that getting a very accurate estimate of p is less important here than getting a larger sample size n.

Derivation of Variance Expressions

The derivation of the variances of the estimators of τ under the various scenarios described above illustrates some useful techniques: linearization (described earlier in the notes) and two common results on iterated expectations (sometimes called the laws of total expectation and total variance). These latter two results are:

1. $\mathrm{E}(Y) = \mathrm{E}\left[\mathrm{E}\left[Y|X\right]\right]$.

2. $\mathrm{Var}(Y) = \underbrace{\mathrm{E}\left[\mathrm{Var}(Y|X)\right]}_{\substack{\text{var. within } y \text{ at each} \\ x \text{ averaged over all } x}} + \underbrace{\mathrm{Var}\left[\mathrm{E}(Y|X)\right]}_{\substack{\text{var. due to differences} \\ \text{in the } \mu_{Y|X}\text{'s}}}$.

The derivation of the variance expressions for $\widehat{\tau}$ in the order in which they were considered above follows.

1. Known detectability over a whole region (Section 16.1)
   The expression for $\mathrm{Var}(\widehat{\tau})$ (eq. (16.4) on p. 216 of Thompson; 2nd ed: eq. (4) on p. 186) follows directly from the variance of a binomial random variable. The unbiasedness of $\widehat{\tau}$ follows from the expected value of a binomial random variable.

2. Estimated detectability over a whole region (Section 16.3)
   The estimated population total is $\widehat{\tau} = y/\widehat{p}$ where y is a binomial(τ, p) random variable, and $\widehat{p}$ is a random variable with $\mathrm{E}(\widehat{p}) = p$ (at least approximately) and variance $\mathrm{Var}(\widehat{p})$. We then use the linearization approximation for the variance of the ratio of two random variables:

$$\mathrm{Var}\left(\frac{Y}{X}\right) \approx \left(\frac{\mu_Y^2}{\mu_X^4}\right)\sigma_X^2 + \left(\frac{1}{\mu_X^2}\right)\sigma_Y^2 - 2\left(\frac{\mu_Y}{\mu_X^3}\right)\rho\sigma_X\sigma_Y.$$

   where y plays the role of Y and $\widehat{p}$ plays the role of X. Also, $\rho = 0$ since y and $\widehat{p}$ are assumed to be independent. Substituting $\mathrm{E}(y) = \tau p$ and $\mathrm{Var}(y) = \tau p(1-p)$ (since y is a binomial random variable), and $\mathrm{E}(\widehat{p}) = p$ (we assume that $\widehat{p}$ is at least approximately unbaised), gives eq. (16.7) on p. 218 of Thomspn (2nd ed: eq. (7) on p. 188):

$$\mathrm{Var}(\widehat{\tau}) \approx \tau\left(\frac{1-p}{p}\right) + \frac{\tau^2}{p^2}\mathrm{Var}(\widehat{p}).$$

   As a ratio estimator, $\widehat{\tau}$ is not unbiased in this situation.

3. Known detectability with simple random sampling (Section 16.4)
   This derivation is outlined in Section 16.7. Based on an SRS of n plots from N plots, where the detection probability p is known, the estimator for the population total was given as:

$$\widehat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\widehat{Y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{y_i}{p}, \text{ with variance } \mathrm{Var}(\widehat{\tau}) = N^2\left[\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} + \left(\frac{1-p}{p}\right)\frac{\mu}{n}\right].$$

   Note that the estimate $\widehat{\tau}$ obtained depends on which n plots are chosen. It's easy to compute the expectation and variance of $\widehat{\tau}$ given the particular set S of n plots chosen for the SRS. So we condition on S and write:

$$\mathrm{E}(\widehat{\tau}|S) = \mathrm{E}\left(\frac{N}{np}\sum_{i\in S} y_i \mid S\right) = \frac{N}{np}\sum_{i\in S}\mathrm{E}(y_i|S) = \frac{N}{np}\sum_{i\in S} pY_i = \frac{N}{n}\sum_{i\in S} Y_i$$

because y_i is binomial(Y_i, p) where Y_i is the actual number of individuals in unit i. Now we take the expectation of the above expression over all possible SRS's S. Since this is just the expected value of N times the sample mean from an SRS, we know by results in chapter 2 (for a finite population) that the expected value is N times the population mean. In other words, the unconditional expected value of $\widehat{\tau}$ is

$$\mathrm{E}(\widehat{\tau}) = \mathrm{E}[\mathrm{E}(\widehat{\tau}|S)] = \mathrm{E}\left[\frac{N}{n}\sum_{i\in S} Y_i\right] = N\mathrm{E}(\overline{Y}) = N\mu = N\left(\frac{\tau}{N}\right) = \tau.$$

Hence, $\widehat{\tau}$ is an unbiased estimator of τ.

To obtain the variance of $\widehat{\tau}$, we again condition on S. First, note that

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}|S) &= \mathrm{Var}\left(\frac{N}{np}\sum_{i\in S} y_i \mid S\right) = \frac{N^2}{n^2p^2}\sum_{i\in S}\mathrm{Var}(y_i|S) \\
&= \frac{N^2}{n^2p^2}\sum_{i\in S} Y_i p(1-p) \\
&= \frac{N^2}{n^2}\left(\frac{1-p}{p}\right)\sum_{i\in S} Y_i.
\end{aligned}$$

The variance of $\widehat{\tau}$ is then computed as

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}) &= \mathrm{E}\left[\mathrm{Var}(\widehat{\tau}|S)\right] + \mathrm{Var}\left[\mathrm{E}(\widehat{\tau}|S)\right] \\
&= \mathrm{E}\left[\frac{N^2}{n^2}\left(\frac{1-p}{p}\right)\sum_{i\in S} Y_i\right] + \mathrm{Var}\left[\frac{N}{n}\sum_{i\in S} Y_i\right] \\
&= \frac{N^2}{n^2}\left(\frac{1-p}{p}\right)\cdot n\mu + N^2\mathrm{Var}(\overline{Y}) \ \ (\text{since } \mu = \tau/N) \\
&= \frac{N^2(1-p)\mu}{pn} + N^2\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} = \underline{N^2\left[\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} + \left(\frac{1-p}{p}\right)\frac{\mu}{n}\right]}.
\end{aligned}$$

This is the equation at the top of p. 223 of Thompson (2nd ed: in the middle of p. 193).

4. Estimated detectability with simple random sampling (Section 16.5)
The derivation of the variance of $\widehat{\tau}$ in this case is outlined on p. 224 (2nd ed: p. 194); the complete derivation will be left as a homework exercise.

Distance Sampling (Chapter 17)

Consider sampling with transects to estimate abundance or density of some kind of object (an animal or a plant, for example) in a region. In moving along a transect (whether on the ground or aerially), the observer sees objects of interest at various distances from the transect. If only objects on the transect (line-intercept) or within some small fixed distance of the transect (strip transects) are recorded, then valuable information is lost if other objects are detected off the transect or outside of the strip. However, if we record all objects, we need to be aware that objects farther from the transect are less likely to be detected and we must take this into account if we are to accurately estimate abundance or density. Therefore, one should also record the perpendicular distance of each object from the transect. Distance sampling methods use the detection distances to (hopefully) generate accurate measures of abundance and density.

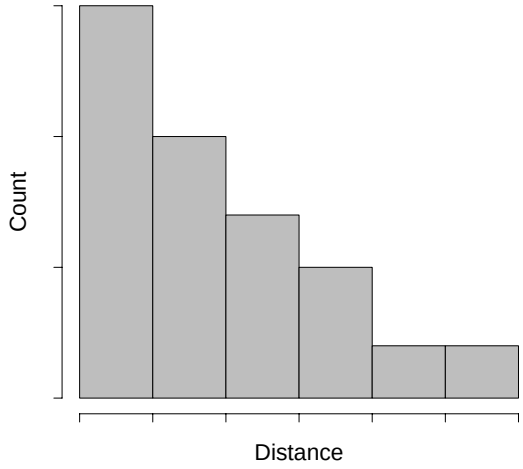
- The primary goal in distance sampling is as it was for line-intercept sampling: to estimate a population total τ or density D per unit area. For example, we might want to estimate the total number of a certain species of bird in some region, or the density per hectare.

- In most transect sampling problems, it will be the case that more objects are detected near the line transect than far away from it. This does not mean that there are actually more objects near the line, but that objects near the line are more easily *detected*. We need to estimate a *detectability function* to describe the decrease in the detection probability as a function of the distance from the line.

- The reliability of distance sampling depends on the accuracy of the distance measurements, particularly for objects near the transect. If we walk off the transect to measure the perpendicular distance from the transect, we might see other objects even further from the transect that we did not detect from the transect itself. For this reason, it is recommended that the observer remain on the transect as much as possible so as not to bias the sampling. With two observers, we might designate one as the "detecter" and one as the "measurer" to avoid this problem.

- We could either measure directly the perpendicular distance from the transect to the object or we could record the sighting distance and angle from the transect (from which we can calculate perpendicular distance)

- Ideally, a histogram of the # of objects versus the distance from the transect will look like the histogram below. If we could then model this count-distance relationship, all objects detected from the transect could be incorporated into an estimator of the population total or density.



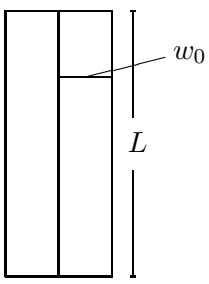
- In this handout (following the treatment in Chapter 17 of Thompson), we will take two basic approaches to estimating τ and D:

  1. Narrow-Strip Method: Suppose that the detectability is perfect within some distance w of the line transect. Then knowing the area of the rectangular strip at distance w around the line, and counting the number of objects in this strip, we can estimate the density of the objects per some unit area.
  2. Detectability Function Method: Instead of only using objects in some narrow strip of "perfect detectability," we might instead model the detectability function as a function of distance from the line. This will be considered using both parametric and nonparametric methods at the end of the next section.

1. Narrow Strip Method

Suppose that the detectability within some distance w_0 from the line is perfect. Typically, the value of w_0 will be determined from the data. Let:

$$\begin{aligned} y &= \text{the total number of detections (at whatever distance),} \\ y_0 &= \text{the total number of detections within } w_0, \\ L &= \text{the length of the transect.} \end{aligned}$$

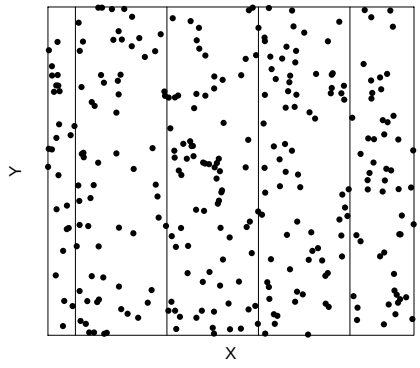


Since the area of the strip is $2Lw_0$, then estimates of the density D and total number of objects in the region τ are given by:

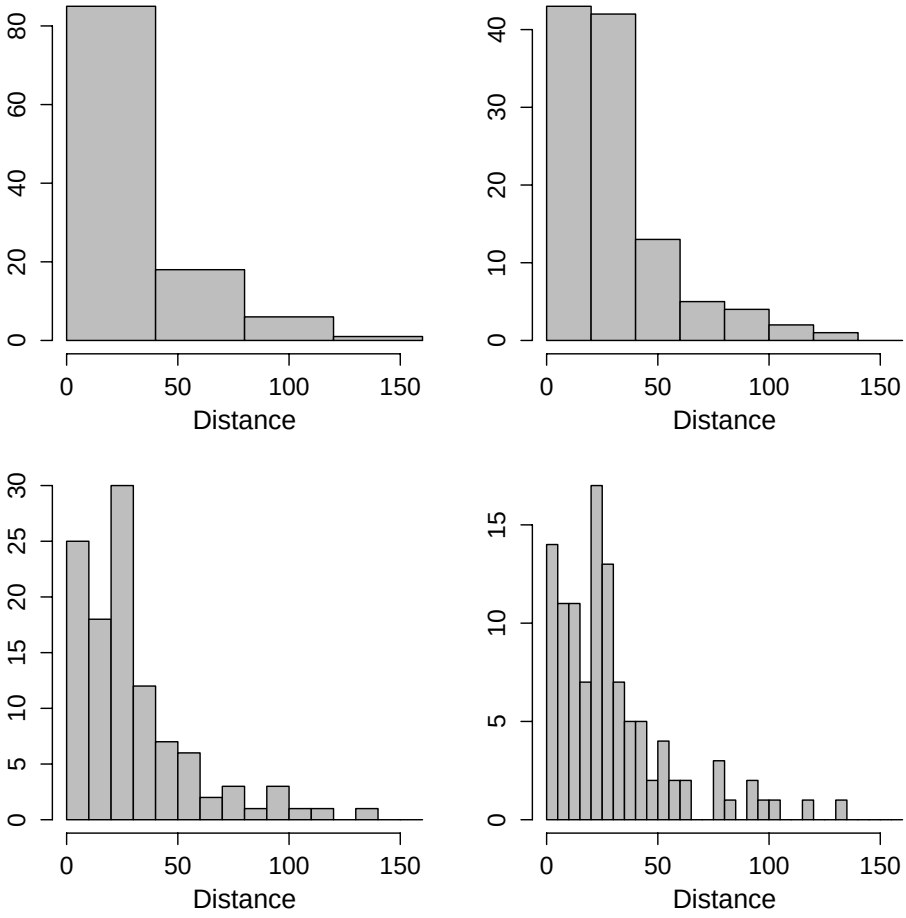
$$\widehat{D} = \frac{y_0}{2Lw_0}, \;\; \widehat{\tau} = A\widehat{D} = \frac{Ay_0}{2Lw_0} \;\; \left(= \frac{By_0}{2w_0} \text{ if } A = BL, \text{ a rectangle} \right),$$

where A is the area of the region and B is the baseline width of the region.

Example 1: Suppose in a rectangular region with baseline width $B = 100$, distance sampling is conducted using a single transect for a certain plant species. In walking along the transect, all plants observed were counted and the distance of each plant (in feet) from the line was measured.



- This is a nice scenario; unfortunately, I do not have real data to support it. Hence, this plant sampling situation was simulated by first randomly placing 300 objects in a 1000x1000 square area. Consider the units here to be feet.
- A systematic sample of four transects was selected.
- A detectability function was specified. Objects were then independently "detected" or "not detected" as Bernoulli trials, with the probability of detection for each object determined by its perpendicular distance from a transect through the detectability function. The distances to all objects detected from a transect were recorded.
- There are no theoretical difficulties in detecting the same object more than once (with 2 different transects), although this might be troublesome from a practical perspective.
- A total of 110 plants (out of the 300 total) were detected on the four transects, and the distances of these plants from their corresponding transect are given in the sequences of histograms shown below. Plotting four histograms of different interval widths gives a comprehensive view of the distribution of distances, and allows us to estimate the width w_0 within which the detectability is roughly "perfect."



- Based on these frequency histograms, what might you estimate the width w_0 to be? Using this estimate, calculate the narrow-strip estimate of the density:

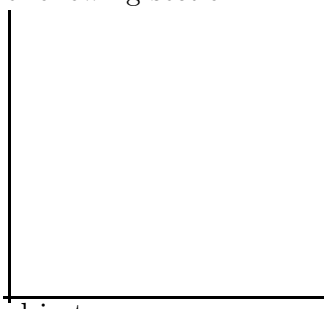
- A few problems with this narrow-strip method are:

  1. It only uses objects within a given small width w_0, not all the objects actually observed.
  2. Detectability may be poor even at small distances, so the method may not be applicable in such situations.

2. Detectability Function Method

The objective in this section is to derive Equation (17.4) on p. 235 of Thompson (2nd ed: eq. (4) on p. 205), which gives an estimator for the density D based on an estimated detectability function. Methods for estimating the detectability function will be discussed in the following section.

The detectability function $g(x)$ is the probability of detecting an object at distance x from the transect; we assume that $g(0) = 1$. A typical detectability function is shown to the right.



- Given some $g(x)$, what is the theoretical distribution of the distances to objects detected? Let $f(x)$ be the probability density function (pdf) of the distances to the detected objects.
- If the objects were uniformly distributed over the entire region (not clumped), we would expect the theoretical pdf $f(x)$ of detection distances to have the same shape as $g(x)$, but normalized to have area 1. Hence, the pdf of the detection distances is:

$$f(x) = \frac{g(x)}{\int_0^\infty g(x)dx} \tag{9}$$

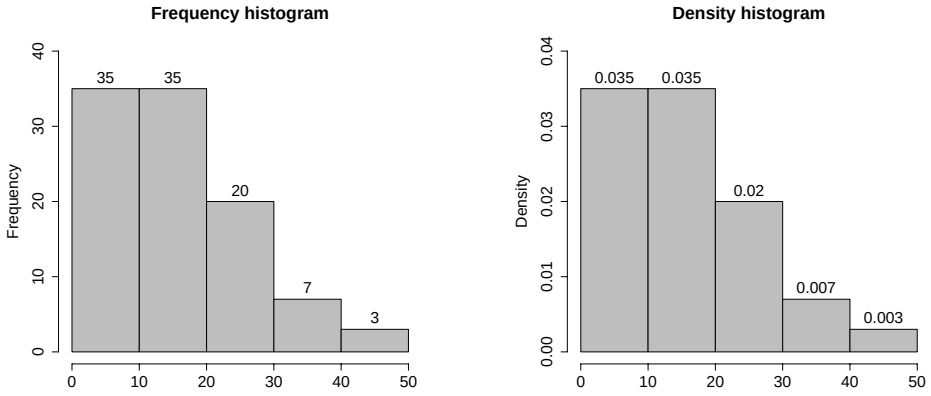
where we note that

$$f(0) = \frac{1}{\int_0^\infty g(x)dx}. \tag{10}$$

- Note that $f(x)$ is just $g(x)$ scaled to area 1 (hence $f(x)$ is a *probability density function*). One way to estimate $f(x)$ is to construct a density histogram of the observed detection distances. A density histogram is constructed by letting the height of each bar of the histogram be:

$$\widehat{f}(x) = \frac{\text{the number of observations in the interval}}{(\text{the total number of observations}) \times (\text{the interval width})}.$$

This makes the area under the density histogram equal to 1. $\widehat{f}(x)$ is then an estimate of $f(x)$.



- In the narrow-strip method described earlier, we assume perfect detectability within some distance w_0 of the transect of length L. In other words, the detection function $g(x)$ and therefore the pdf $f(x)$ of distances to detected objects are "flat" between 0 and w_0. Letting y_0 be the number of detections within the distance w_0, and noting that the area of the strip of assumed perfect detectability is $2w_0L$, an estimate of the density of objects under this method is:

$$\widehat{D} = \frac{y_0}{2w_0L}.$$

- The value of w_0 can be chosen based on a histogram of detection distances (as was done earlier), but this is clearly very subjective. It might be worth trying a range of widths in a given problem to discern how sensitive the density estimate is to this width.

- Note that if we construct a probability histogram of the detection distances where the first bar covers the interval from 0 to w_0, then the height of the bar over the first interval is: $\dfrac{y_0}{yw_0}$ where y is the total number of detections.

- Using this expression for the height of this first bar, the form of the density estimator for the narrow-strip method may be rewritten as:

$$\boxed{\widehat{D}} = \frac{y_0}{2w_0L} = \frac{\left(\dfrac{y_0}{yw_0}\right)y}{2L} = \boxed{\frac{\widehat{f}(0)y}{2L}} \tag{11}$$

because $\widehat{f}(0)$ is the height of the histogram bar over the first interval.

- This all suggests that an estimate of the density could be $\widehat{D} = \dfrac{\widehat{f}(0)y}{2L}$ where $\widehat{f}(0)$ is *any* estimate of $f(0)$. We will investigate some alternative methods (both parametric and non-parametric) for estimating $f(0)$ to simply "by eye" using a histogram. We will also not be constrained to the assumption that there is perfect detectability in some narrow strip about the transect, only that there is perfect detectability on the transect.

- In this density estimate, there are two sources of variability:

  1. The detection function $g(x)$ is unknown.
  2. Even if the detection function were known, there is still the variability in what is detected.

Back to Example 1: If we think the width of perfect detection is $w_0 = 30$, then under the narrow strip method (recalling that $L = 4000$ is the total length of the 4 transects):

$$\widehat{f}(0) = \frac{y_0}{yw_0} = \frac{73}{110 \cdot 30} = .022 \implies \widehat{D} = \frac{y\left(\frac{y_0}{yw_0}\right)}{2L} = \frac{110\left(\frac{73}{110 \cdot 30}\right)}{2 \cdot 4000} = \frac{73}{240000} = \underline{.0003042}.$$

Effective Half-Width of a Detection Function: A second motivation for the form of the density estimator given in equation (11) is outlined below using the notion of the *effective half-width* of a detection function $g(x)$. To understand what is meant by this effective half-width, imagine a strip transect with perfect detectability out to some distance w, with no objects beyond this distance detected. This corresponds to a detection function $g^*(x)$ given by:

$$g^*(x) = \left\{ \begin{array}{ll} 1 & x \leq w \\ 0 & x > w \end{array} \right\}.$$

- The value of w such that the expected number of detections would be the same with $g^*(x)$ as with $g(x)$ is the effective half-width of a transect with detection function $g(x)$. How do we calculate w?

- We calculate the expected number of detections for $g(x)$, then calculate the expected number of detections for $g^*(x)$ (with arbitrary w), set them equal, and solve for w. The expected numbers of detections will be the same for the two detection functions if the *average* detectability is the same.

- For simplicity, assume that there is a maximum detection distance $w_{\max}$ so that $g(x)$ is 0 for $x > w_{\max}$. The average detection probability for $g(x)$ over the interval 0 to $w_{\max}$ is the average height of $g(x)$ over 0 to $w_{\max}$, which is the area under $g(x)$ divided by $w_{\max}$, i.e., $\dfrac{\int_0^{w_{\max}} g(x)dx}{w_{\max}}$. This is illustrated with a detection function on the plot to the right.

- The average detectability for $g^*(x)$ over the interval 0 to $w_{\max}$ is $w/w_{\max}$ (again, area divided by width). The average detectability for $g(x)$ is equal to the average detectability for $g^*(x)$ when $w = \int_0^{w_{\max}} g(x)dx$. More generally, the effective half-width of the detection function $g(x)$ is

$$w = \int_0^{\infty} g(x)dx. \tag{12}$$

This generalizes the formula to detection functions which are positive for all x.

- It follows from equation (10) given earlier (p. 162), that the effective half-width can also be expressed as $w = \dfrac{1}{f(0)}$. Now, what is the expected number of objects y which will be observed on a transect of length L with a detection function $g(x)$ with effective half-width w when the average density of objects is D per unit area?

- Since the effective half-width is w, the expected number of detections is the total number of objects expected to be in a strip of width $2w$ and length L, which is $D(2wL)$. Hence, $\mathrm{E}(y) = D(2wL)$. Solving for D:

$$D = \frac{\mathrm{E}(y)}{2wL} = \frac{\mathrm{E}(y)f(0)}{2L}.$$

- We don't know $\mathrm{E}(y)$ since we don't know D. But, we observe a value of y for a single transect. So, if we were to estimate D from a single transect, it would make sense to simply use y as an estimate of $\mathrm{E}(y)$. If we also don't know $f(0)$ (because it depends on the true detectability function which is generally unknown), then we also substitute an estimate of $f(0)$, say $\widehat{f}(0)$. Then our estimate of D is:

$\boxed{\widehat{D} = \dfrac{y\widehat{f}(0)}{2L}}$, which is equation (17.4) on page 235 of Thompson (2nd ed: eq. (4) on p. 205).

Methods of Estimating the Detectability Function $f(x)$

We will discuss both parametric and nonparametric estimation methods for estimating the detectability function in the next section. In addition, we will also address estimating the variance of the estimator of density.

Assumptions. Four assumptions are essential for reliable estimation of density using line-transect methods:

1. Lines are placed according to some randomized design (note: it is not necessary that the *objects* be randomly distributed, but it is critical that the lines be placed randomly with respect to the distribution of the objects).

2. Objects directly on the line are always detected, that is, $g(0) = 1$.

3. Objects are detected at their initial location, prior to any movement in response to the observer.

4. Distances are measured accurately, particularly for objects close to the line.

It is sometimes assumed that whether any individual object is detected is independent of whether any other object is detected. This assumption can be avoided by taking the transect to be the

sampling unit when estimating the variance of the estimator. This is the approach that Thompson takes and will be discussed in the next section.

There are a couple of other things that are not assumptions, but that are important in order for distance methods to work well:

- The detection function has a "shoulder"; that is, the probability of detection remains at or close to one near the line.
- There are enough detections to estimate the detection function well. It has been suggested that at least 60-80 detections are needed.

References:

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., & Thomas, L. (2001). *Introduction to Distance Sampling*, Oxford University Press, Oxford.
Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., & Thomas, L., ed. (2004). *Advanced Distance Sampling*, Oxford University Press, Oxford.

An earlier version of the first book is available as a pdf file online at
http://www.colostate.edu/depts/coopunit/download.html

Program **Distance** is a Windows-based computer package for the design and analysis of distance sampling surveys of wildlife populations. It is available for download at no cost at http://www.ruwpa.st-and.ac.uk/distance/.

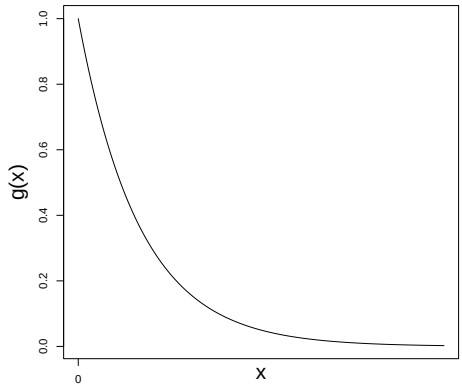
Methods of Estimating the Detectability Function

1. Parametric Methods

The term "parametric" here simply means we are assuming that the detectability can be expressed as a function of the distance depending on one or more parameters. In the context of detectability, a parametric approach means that we are going to assume a certain shape for the detectability function, such as quadratic or exponential. We include unknown parameters in the general formula for such a curve so that it can be scaled to different data sets. It turns out that if we assume a certain parametric shape for the detectability function, then that implies a particular parametric family of distributions for the theoretical distribution of observed detection distances. However, that is a consequence of our assumptions about the shape of $g(x)$.

We will consider two types of parametric models here: the exponential and half-normal models. Other model types may be chosen, and would lead to different detectability functions, and ultimately different estimates of the density D.

Exponential Fit: Suppose we were to assume that the detectability function $g(x)$ is an exponentially decreasing function so that it has the shape shown to the right (recall that we always assume $g(0) = 1$).



- A general equation which describes such a curve, and which has the property $g(0) = 1$, is:

$$g(x) = \exp(-\beta x), \quad \text{for } x \geq 0,$$

  where β is any number greater than 0. The value of the parameter β would change, depending on the scale for the x-axis.

- By equation (12) on p. 164 of the notes, we know that the effective half-width w for $g(x)$ is:

$$w = \int_0^\infty g(x)dx = \int_0^\infty e^{-\beta x}dx = -\frac{1}{\beta}e^{-\beta x}\Big|_0^\infty = \frac{1}{\beta}.$$

  Therefore, we can rewrite the exponential detection function as $g(x) = \exp(-x/w)$. It is convenient for interpretation to parameterize $g(x)$ in terms of w.

- Now, how do we estimate the unknown parameter w from a particular set of data $x_1, \ldots, x_n$ (that is, a set of detection distances)?

Using equation (9) on p. 162 of the notes, the resulting theoretical probability density function (pdf) $f(x)$ of detection distances assuming an exponential detection function is:

$$f(x) = \frac{g(x)}{\int_0^\infty g(x)dx} = \frac{\exp(-x/w)}{w},\ x \geq 0.$$

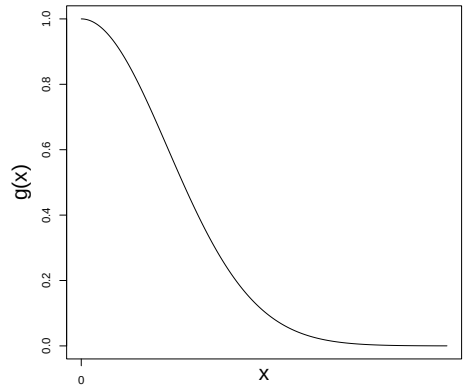
This is the pdf of an exponential distribution with mean w.

- The maximum likelihood estimator of the mean of an exponential distribution is the sample mean. Therefore, the maximum likelihood estimator $\widehat{w}$ of w is the sample mean $\overline{x} = \sum_{i=1}^n x_i/n$ of the observed detection distances $x_1, \ldots, x_n$. And hence, an estimate of $f(0)$ is given by $\widehat{f}(0) = 1/\widehat{w} = 1/\overline{x}$. Substituting this into the density equation given earlier for estimating D, the estimated density of objects is:

$$\boxed{\widehat{D}} = \frac{y\widehat{f}(0)}{2L} = \boxed{\frac{y}{2\overline{x}L}}.$$

Example 1 (continued): For Example 1 on p. 160 of the notes, the mean distance to the 110 detected objects was $\overline{x} = 29.68447$. Therefore, $\widehat{D} = \dfrac{110}{2(29.68447)(4000)} = .0004632$ and the estimated number of objects in the whole area is $\widehat{\tau} = 1000^2\widehat{D} = 463.2$. This estimate is quite a bit higher than the narrow strip estimate for these data and quite a bit higher than the actual value of $\tau = 300$; the graphs near the end of this handout make it clear why the exponential overestimates τ.

Half-Normal Fit: It might be reasonable to assume that detectability does not drop off so quickly near the transect as with the exponential detectability function, but that it decreases more slowly at first. The upper half of a normal density, shown to the right, has such a shape, where detectability is initially somewhat flat, and then decreases more quickly as a function of distance.



- The usual normal pdf with mean μ and variance σ^2 ($X \sim N(\mu, \sigma^2)$) is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ -\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0.$$

For the shape above, the right half of this normal density with $\mu = 0$ could be used as the detection function. This gives:

$$g(x) = \exp\left\{-\frac{x^2}{2\sigma^2}\right\},\ 0 \leq x < \infty,\ \text{ or more simply: } \underline{g(x) = e^{-\beta x^2},\ x \geq 0},$$

where $\beta > 0$ is a parameter which allows this detectability function to be scaled to different distances.

- The effective half-width is $w = \int_0^\infty e^{-\beta x^2} dx$. Using the substitution $t = \beta x^2$,

$$w = \int_0^\infty e^{-\beta x^2} dx = \frac{1}{2\beta^{1/2}} \int_0^\infty e^{-t} t^{-1/2} dt = \frac{1}{2\beta^{1/2}} \Gamma\left(\frac{1}{2}\right) = \sqrt{\frac{\pi}{4\beta}}$$

where $\Gamma(\cdot)$ is the gamma function, recalling that $\Gamma(1/2) = \sqrt{\pi}$. Writing $g(x)$ in terms of w, we have

$$g(x) = \exp\left\{-\pi x^2/4w^2\right\} \implies f(x) = \frac{g(x)}{w} = \frac{\exp\{-\pi x^2/4w^2\}}{w},\ x \geq 0.$$

- The resulting maximum likelihood estimator of w based on this half-normal p.d.f. is $\widehat{w} = \sqrt{\frac{\pi}{2y} \sum_{i=1}^{y} x_i^2}$. Since $\widehat{f}(0) = 1/\widehat{w}$, this gives

$$\boxed{\widehat{D}} = \frac{y\widehat{f}(0)}{2L} = \boxed{\frac{y}{2\widehat{w}L}}.$$

- Example 1 (continued): For Example 1 on p. 160 of the notes, the mean squared distance is $\frac{1}{y} \sum_{i=1}^{y} x_i^2 = 1563.955$ and so $\widehat{w} = \sqrt{\frac{\pi}{2}(1563.955)} = 49.56465$. Hence, the estimated density per square unit is $\widehat{D} = \frac{110}{2(49.56465)(4000)} = 0.0002774$ and the estimated total number of objects in the region is $\widehat{\tau} = 1000^2(.0002774) = 277.4$.

Comments:

1. The major advantage to these parametric detectability functions is the fact that *all* of the data values are being used to estimate $f(0)$, which is clearly the key component in estimating the density D. Earlier in this handout, we estimated $f(0)$ by simply eyeballing the distance histogram near 0.

2. Clearly, the eventual estimate of $f(0)$ and the subsequent density estimate $\widehat{D}$ will differ depending on which detectability function gets used. Hence, the estimate of $f(0)$ is only as good as the detectability function used. This is true for any situation where a parametric model is used to describe some relationship.

3. Although the theory for the exponential detectability function is fairly simple, this function is not commonly used in practice as distributions of detection distances rarely seem to follow this shape.

2. Nonparametric Methods

The term "nonparametric" means that no functional form is being assumed for the detectability function $g(x)$. In this way, we avoid having to determine a function for the apparent shape of the detectability function, and instead use nonparametric smoothing techniques to estimate the probability density function $f(x)$ of the detection distances. Although there are a number of nonparametric techniques available for fitting $f(x)$, only the kernel density method is considered here. Approximating the density by Fourier series (essentially a mixture of cosine curves with different periods) is discussed in Thompson on p. 239 (2nd ed: p. 209). Program Distance (see references on p. 166 of the notes) includes Fourier series and other methods.

Kernel Density Estimation

To understand kernel density estimation, recall that the Narrow Strip method could be viewed as being based on a crude nonparametric method for estimating $f(0)$. It used the empirical density estimate of $f(x)$ derived from a histogram with interval width w_0. The empirical density estimate of $f(x)$ was

$$\widehat{f}(x) = \frac{\text{the number of observations in the interval } x \text{ belongs to}}{(\text{the total number of observations}) \times (\text{the interval width})}.$$

This did not give a smooth estimate of $f(x)$, however; it gave a function that is flat across bins with jumps at the boundaries of the bins. We could smooth this density estimate out somewhat by making a simple change: let

$$\widehat{f}(x) = \frac{\text{the number of observations in the interval } \textit{centered} \text{ at } x}{(\text{the total number of observations}) \times (\text{the interval width})}.$$

That is, we estimate $f(x)$ by counting the number of observations in the interval of width w_0 centered at x. We can think of this process as taking an interval of width w_0 and sliding it along the x-axis. This is a kernel density estimator as we'll shortly see. To switch into the notation of kernel density estimation, let h be the half-width of the interval that we slide along the x-axis; that is the interval centered at x is $[x-h, x+h]$. We can then write the density estimate of $f(x)$ as

$$\widehat{f}_h(x) = \frac{1}{2hy} \, \#\{x_i \in [x-h, x+h]\}$$

where the subscript h on f indicates the dependence on the choice of h.

This density estimator gives weight 1 to every observation within distance h of x and weight 0 to observations farther away. It seems intuitively appealing to consider using a smoother weighting function which gives steadily more weight to observations the closer they are to x. In order to do that in a organized framework, let's rewrite the simple density estimator we already have using the indicator function I where $I(A) = 1$ if A is true and is 0 otherwise. Then

$$\begin{aligned} \widehat{f}_h(x) &= \frac{1}{2hy} \sum_{i=1}^{y} I(x-h \le x_i \le x+h) = \frac{1}{2hy} \sum_{i=1}^{y} I\left(-1 \le \frac{x-x_i}{h} \le 1\right) \\ &= \frac{1}{hy} \sum_{i=1}^{y} K\left(\frac{x-x_i}{h}\right) \end{aligned} \tag{13}$$

where

$$K(u) = \frac{1}{2}\, I(|u| \leq 1).$$

K is called the *kernel function*. This particular kernel function is called the *uniform* or *rectangular* kernel function. We've taken the parameter h out of the kernel function and put it into the argument to make it clear that, no matter what the value of h, the weighting function has the same basic shape.

Now we're ready to choose other weighting functions. A valid weighting function is $K\left(\frac{x - x_i}{h}\right)$ where $K(u)$ is any kernel function. A kernel function is any nonnegative real-valued function $K(u)$ defined on $(-\infty, \infty)$ such that

$$\int_{-\infty}^{\infty} K(u)du = 1.$$

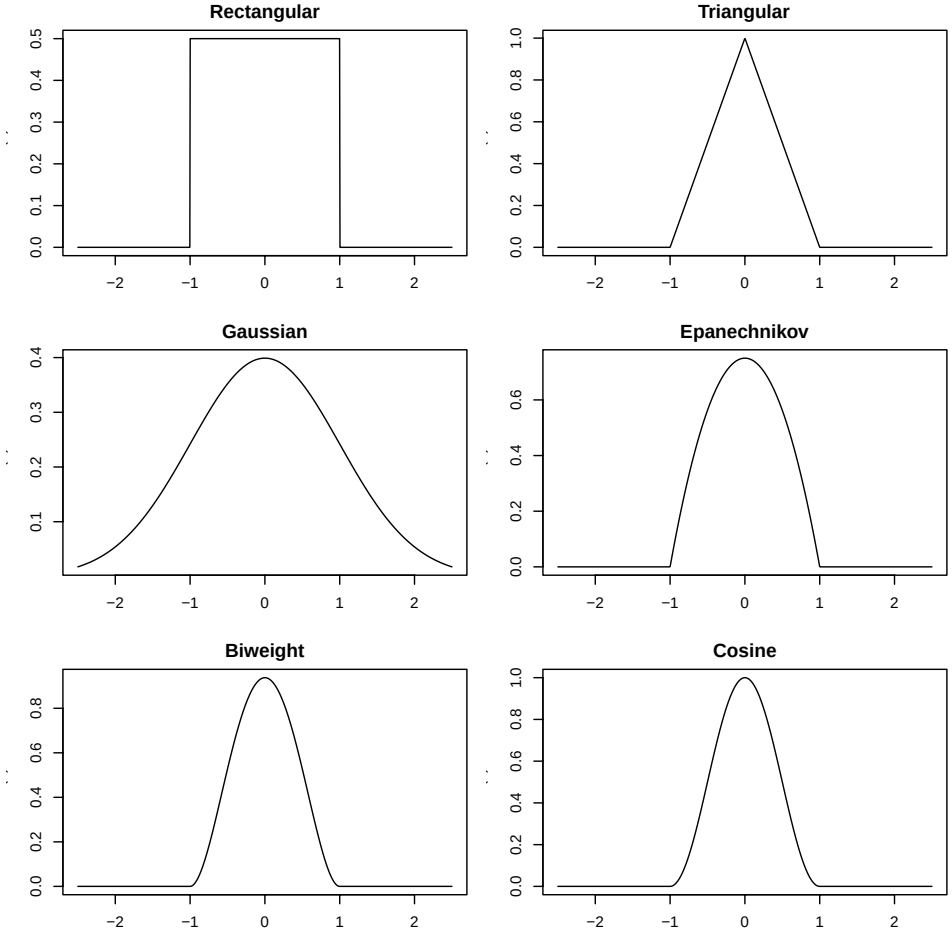
(Hey, does that definition sound familiar?) Equation (13) above is a density estimate if K is a kernel function. To verify that (13) gives a valid density estimate, we must verify that $\widehat{f}_h(x) \geq 0$ for all x and also that $\int_{-\infty}^{\infty} \widehat{f}_h(x)\, dx = 1$. The former is obvious because $K(u) \geq 0$. To verify the latter, note that

$$\begin{aligned} \int_{-\infty}^{\infty} \widehat{f}_h(x)dx &= \int_{-\infty}^{\infty} \frac{1}{hy} \sum_{i=1}^{y} K\left(\frac{x - x_i}{h}\right)\, dx \\ &= \end{aligned}$$

Common Kernel Functions: Some common kernel functions are given in the table below. The scaling of all these functions occurs through selection of the *window width* h (also called the *bandwidth*) of a kernel density estimator. Selection of h will be discussed below. Note that all the kernel functions below except the Gaussian (normal) give zero weight to observations which are farther than h from x. The Gaussian gives positive weight to all observations, but the weights decrease quickly as x_i moves away from x and are essentially zero for observations farther than about $2h$ from x.

Name	$K(u)$
Rectangular	$\frac{1}{2}I(\|u\| \leq 1)$
Triangular	$(1 - \|u\|)I(\|u\| \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Epanechnikov	$\frac{3}{4}(1 - u^2)I(\|u\| \leq 1)$
Biweight	$\frac{15}{16}(1 - u^2)^2 I(\|u\| \leq 1)$
Cosine	$\frac{\pi}{4} \cos(\frac{\pi}{2}u)I(\|u\| \leq 1)$

Plots of these kernel functions are below. There's not much difference among most of them and the choice of a kernel function matters less than the choice of the bandwidth h. The exception is the rectangular kernel which is not often used (for the reason, see the example below). The Gaussian kernel is a common choice and the one we will use. The fact that the Gaussian kernel is more "spread out" than the other kernels is of no consequence since we can control the spread through the choice of the bandwidth h.



In R, the function **density** computes kernel density estimates for a set of data using any of the kernel functions in the table above. To plot the kernel density estimate for a vector of data values x, use

```
> plot(density(x))
```

The default is to use the Gaussian kernel with bandwidth determined by equation (17.12) on p. 238 of Thompson (2nd ed: eq. (12) on p. 208):

$$h = 0.9 \min(s, Q/1.34) n^{-1/5}$$

where s is the sample standard deviation of the data values, Q is the IQR of the data values (the 75th percentile minus the 25th percentile), and n is the number of data values. To change these, use, for example,

```
> plot(density(x,kernel="epanechnikov",bw=5))
```

A density plot can be added to a histogram by using the "lines" command:

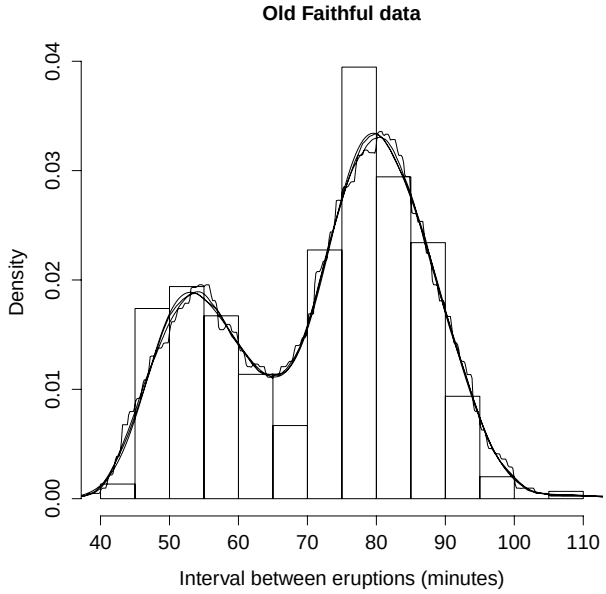
```
> hist(x)
> lines(density(x))
```

Example 2: The data frame `geyser` (available in R in the MASS package) gives the interval between eruptions (called "waiting") and the duration (in minutes) for 299 eruptions of Old Faithful geyser. We'll use only the intervals between eruptions. First, let's compare several of the density estimators using the default bandwidths:

```
> library("MASS")
> x <- geyser$waiting
> # freq=F option in hist makes density histogram
> hist(x,freq=F,ylab="Density",xlab="Interval between eruptions (minutes)",
+ main="Old Faithful data",cex.lab=1.3,cex.axis=1.3,cex.main=1.3)
> lines(density(x,kernel="rectangular"))
> lines(density(x,kernel="gaussian"))
> lines(density(x,kernel="triangular"))
> lines(density(x,kernel="epanechnikov"))
```

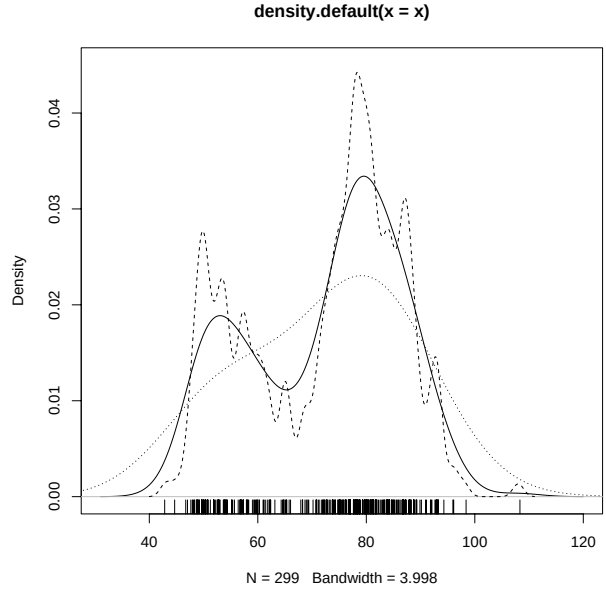
The really squiggly line is the rectangular kernel estimate; the others are much smoother (though with the same overall shape) and more appealing. The others are all very similar to each other so we'll restrict attention to the Gaussian kernel from now on.

The histogram clutters the visual comparison of the smoothers so from now on, we'll plot the data with a **rug plot**, illustrated below. First, let's see how the bandwidth affects the Gaussian smoother. In the R commands below, the `lty` option on the plot command specifies different line



types for plotting. The command `jitter(x)` is used for plotting data with lots of repeat values (as there are in the geyser data); it adds a small random amount to each value so that they are separated on the plot (the value 2 specifies a random amount twice as large as the default).

```
> plot(density(x),ylim=c(0,.045))   # default bandwidth
> lines(density(x,bw=1),lty=2)  # lty is line type
> lines(density(x,bw=10),lty=3)
> rug(jitter(x,2))
```



We can see that the default bandwidth works pretty well: smooth, but not so smooth that it fails to capture the bimodality of the distribution of waiting times. It's a good idea to compare several bandwidths before choosing a density estimate; the default will not always be the most appropriate. By the way, the default bandwidth is a value estimated from the data. Thompson gives the formula in equation (17.12) on p. 238 (2nd ed.: eq. (12) on p. 208).

We can obtain the value of the estimated density at any x by storing the results of the density function without plotting it and then using the `approx` function to do a linear interpolation:

```
> d <- density(x)
> approx(d$x,d$y,c(60,80,90,110))
$x [1]  60  80  90 110
$y [1] 0.0141942256 0.0333562639 0.0172889356 0.0003000962
```

Kernel Density Estimation for Distance Data

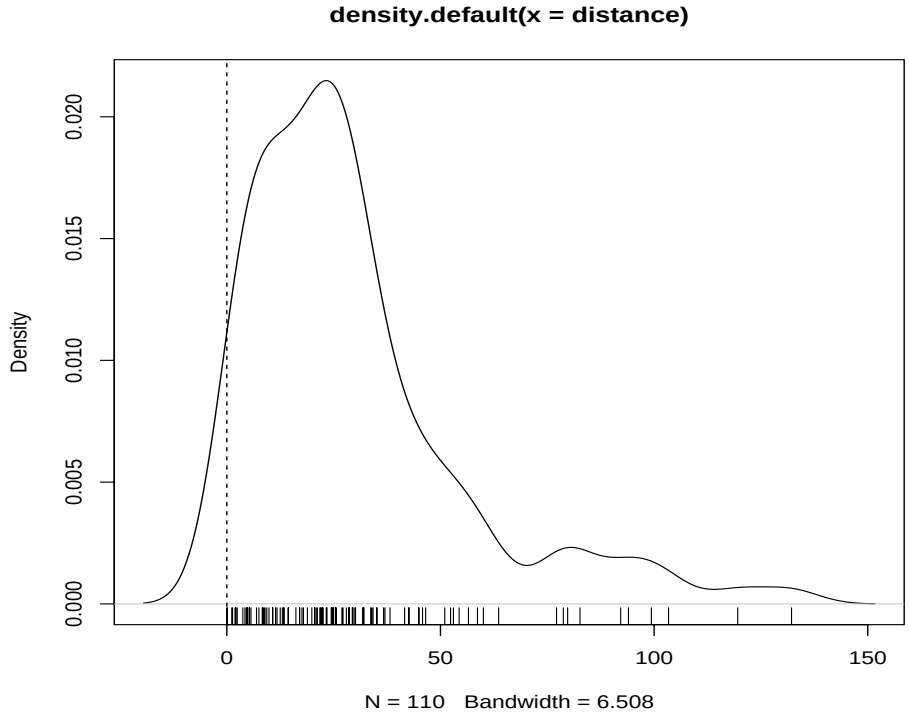
The reason for this discussion on kernel density estimation is because we want to use it on distance data to estimate $f(0)$. However, there are two problem with kernel density estimation at the boundaries of data which are constrained to be greater than 0 (or constrained to be above or below any values). These are:

- You can get positive estimated density for negative x values because the window for x values just below 0 will include positive x values.
- The estimated density at 0 often won't look "right."

The first of these isn't a concern for distance data since we're only interested at the estimated density at $x = 0$, but the second is a major concern. As an example, look at the kernel density estimate for the simulated data from Example 1 on page 160 of the notes.

```
> plot(density(distance))   # Gaussian kernel with default bandwidth
> rug(distance)
> abline(v=0,lty=2) # plots vertical dotted line at x=0
```

Do you see any problem with the estimated density at $x = 0$?



The kernel density estimate of the detection distances decreases as as the distance decreases to 0, implying that estimated detectability is lower near the line than farther away. This isn't consistent with what we saw in the histograms on page 160 of the notes. Why does this happen?

In the dotplot above, think about what set of points is included in an interval of width $2h$ (that is, a bandwidth of h) around the point $x = h$ and around the point $x = 0$. Since the interval centered at h includes all the points in the interval centered at 0 plus the points in the interval from h to $2h$, it is not surprising that $\widehat{f}(0) < \widehat{f}(h)$. Since distances must be ≥ 0, the interval around $x = 0$ is really only of width h rather than $2h$. Therefore, in the formula for the kernel density estimate for $f(0)$, it makes sense that we should divide by $h/2$ rather than h; that is we should use

$$\widehat{f}(0) = \frac{1}{(h/2)y} \sum_{i=1}^{y} K\left(\frac{0 - x_i}{h}\right) = \frac{2}{hy} \sum_{i=1}^{y} K\left(\frac{x_i}{h}\right) \tag{14}$$

(noting that $K(u) = K(-u)$ by the symmetry of the kernel). This is equation (17.10) on p. 238 of Thompson (2nd ed: eq. (10) on p. 208). For the Gaussian kernel, this gives

$$\widehat{f}(0) = \frac{2}{hy} \sum_{i=1}^{y} \frac{1}{\sqrt{2\pi}} e^{-(1/2)(x_i/h)^2}.$$

This gives an adjusted estimator of $f(0)$ for distance data, but it doesn't give an adjusted estimator of $f(x)$ for $x > 0$. The following argument gives an adjusted estimator of $f(x)$ for all $x > 0$ and, in addition, justifies the estimate of $f(0)$ in equation (14).

A Kernel Density Estimator for Distance Data

Consider the *signed* detection distances; call them $x_1^*, x_2^*, \ldots, x_y^*$ where $x_i^* = -x_i$ if the i^{th} objected was detected to the left of the transect and $x_i^* = x_i$ if the object was detected to the right of the transect. Now, we could use kernel density estimation on the signed distances to estimate $f^*(x), -\infty < x < \infty$, the pdf of the signed detection distances. We would not have the boundary problem at $x = 0$ that we had with the unsigned detection distances. Once we had an estimate $\widehat{f^*}(x)$ of $f^*(x)$, we could obtain an estimate of the pdf of the unsigned detection distances as follows:

$$\widehat{f}(x) = \widehat{f^*}(x) + \widehat{f^*}(-x).$$

Intuitively, this says that the "probability" of obtaining an unsigned detection distance of x is the "probability" of obtaining a detection distance of x to the left of the transect plus the "probability" of obtaining a detection distance of x to the right of the transect.

However, rather than estimating $f^*(x)$ from the observed signed detection distances, consider the following. We assumed the detection function is the same on both sides of the transect. That's why we didn't worry about which side of the transect an object was detected on. Therefore, we're assuming $f^*(x)$ is symmetric about 0. It therefore seems reasonable to require the estimate of $f^*(x)$ to be symmetric around 0. One way to do that is to take the unsigned detection distances and create two copies of each: one with a positive sign and one with a negative sign. Thus we end up with $2y$ signed distances: $x_1, \ldots, x_y; -x_1, \ldots, -x_y$. We then use kernel density estimation to estimate $f^*(x)$ from these $2y$ data values. Because the $2y$ values are symmetric around 0, the kernel density estimate of $f^*(x)$ will also be symmetric around 0. From equation (13) on p. **??** of the notes, the kernel density estimate is

$$\widehat{f^*}(x) = \frac{1}{2yh}\left[\sum_{i=1}^{y} K\left(\frac{x - x_i}{h}\right) + \sum_{i=1}^{y} K\left(\frac{x + x_i}{h}\right)\right].$$

Notice that $\widehat{f^*}(x) = \widehat{f^*}(-x)$. Therefore, the estimate of the pdf of the unsigned distances is

$$\widehat{f}(x) = \widehat{f^*}(x) + \widehat{f^*}(-x) = 2\widehat{f^*}(x) = \frac{1}{yh}\left[\sum_{i=1}^{y} K\left(\frac{x - x_i}{h}\right) + \sum_{i=1}^{y} K\left(\frac{x + x_i}{h}\right)\right]. \qquad (15)$$

In particular,

$$\widehat{f}(0) = \frac{1}{yh}\left[\sum_{i=1}^{y} K\left(\frac{-x_i}{h}\right) + \sum_{i=1}^{y} K\left(\frac{x_i}{h}\right)\right] = \frac{2}{yh}\sum_{i=1}^{y} K\left(\frac{x_i}{h}\right)$$

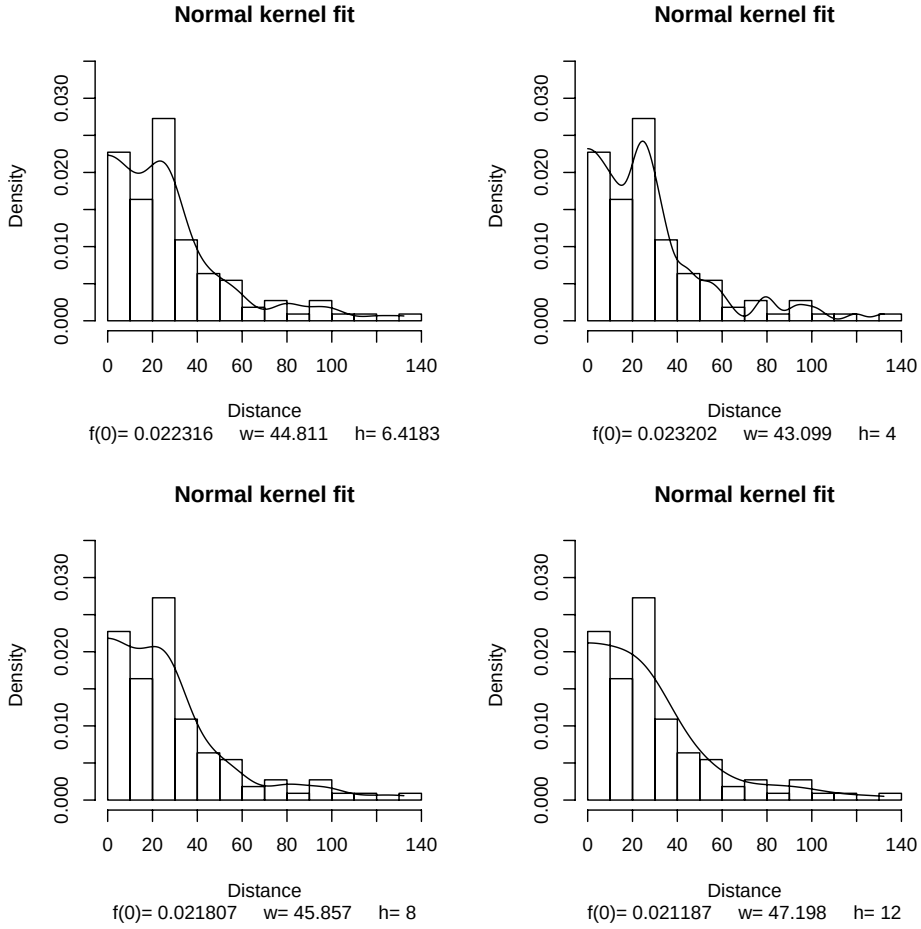
by the symmetry of $K(u)$ about 0. This is precisely the estimate in equation (14) and is the same as equation (17.10) on p. 238 of Thompson (2nd ed: eq.(10) on p. 208). Equation (15) gives an estimate of $f(x)$ for all $x \geq 0$.

- **Note: equation (17.9) on p. 238 of Thompson (2nd ed: eq.(9) on p. 208) is incorrect; it should be replaced by equation (15) above. The error in equation (17.9) makes no difference for $x = 0$ so equation (17.10) is correct.**

- The kernel density estimate in equation (15) cannot be obtained directly from the `density` function in R. At the end of these notes is a function I have written, `distance.fit`, for obtaining this estimate (along with the exponential and half-normal estimates).

Example 1 (continued): Returning to the simulated data from p. 160 of the notes, the plot below shows the Gaussian (normal) kernel density estimates for these data for the default bandwidth (first plot) and three other bandwidths.

The estimated density with the kernel density estimate with default bandwidth is $\widehat{D} = \widehat{f}(0)y/2L = (.022316)(110)/(2 \cdot 4000) = 0.0003068$ objects per square unit and $\widehat{\tau} = 1000^2\widehat{D} = 306.8$ objects total. The estimated total number of objects for the other bandwidths are: 319.0 ($h = 4$), 299.8 ($h = 8$), and 291.3 ($h = 12$). The bandwidth doesn't make a tremendous amount of difference in the estimates.



- Estimates of density and associated standard errors should be based on a sample of more than one transect: "Variance estimates based on a sample of several transects are to be preferred to 'analytical' estimates based on observations within a single transect - a recommendation emphasized by a number of authors..." (p. 240 of Thompson; 2nd ed.: p. 210).

- Suppose we have an SRS or a systematic sample of n transects of possibly differing lengths, $L_1, \ldots, L_n$. Let y_i be the number of detections on the i^{th} transect, so $y = y_1 + \ldots + y_n$. For multiple transects, there are two possible estimators of density. Both are based on the idea of computing an estimate of density, $\widehat{D}_i$, from the detections on each transect separately and then combining those estimates. The two estimators are:

  1. The unbiased estimator on p. 242 of Thompson (2nd ed: p. 212). This is simply the average of the estimates of density from the individual transects: $\widehat{D} = (1/n)\sum_{i=1}^{n} \widehat{D}_i$.
  2. The ratio estimator on p. 243 of Thompson (2nd ed: p. 213) which is a weighted average based on the transect lengths:
  $$\widehat{D}_r = \frac{\sum_{i=1}^{n} L_i \widehat{D_i}}{\sum_{i=1}^{n} L_i},$$

  The ratio estimator is preferred. It is equivalent to the unbiased estimator for a rectangular region where all the transects are the same length and is preferred when the region is irregular with unequal length transects because it takes transect length into account. In the general formula for a ratio estimator in Section 7.1 ($r = \overline{y}/\overline{x}$), the role of y_i is played by $L_i\widehat{D}_i$ and the role of x_i is played by L_i.

- The estimated variance for a ratio (by linearization) is on p. 95 of Thompson (2nd ed: p. 70). In this application, the population size N of possible transects is considered infinite so the estimated variance is:
$$\widehat{\mathrm{Var}}(\widehat{D}_r) = \frac{1}{\mu_L^2 n(n-1)} \sum_{i=1}^{n} \left(L_i\widehat{D}_i - L_i\widehat{D}_r \right)^2 \tag{16}$$
where $\mu_L = E(L) = A/B$ is the expected length of a transect (A is the area of the region and B is the length of the baseline). If μ_L is unknown, it can be replaced by the average length of the sample transects $\overline{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$ which gives the adjusted estimator of variance (top of p. 96 of Thompson; 2nd ed.: second formula on p. 70). If the transects are all the same length, then the formula for the estimated variance reduces to that for a mean: $\widehat{\mathrm{Var}}(\widehat{D}) = (1/n)s^2$ where s^2 is the sample variance of $\widehat{D}_1, \ldots, \widehat{D}_n$.

- The estimator of variance in equation (16) assumes that the $\widehat{D}_i$'s are independent estimates of the true density. This will be the case if, for example, $\widehat{D}_i = \dfrac{y_i \widehat{f}_i(0)}{2L_i}$ where $\widehat{f}_i(0)$ is an estimate of $f(0)$ based *only* on the distances recorded for transect i. That is, one would not pool the distance data from all the transects to estimate $f(0)$. This might be reasonable if detectability varied widely from transect to transect because of high variability in terrain or vegetation which happened to coincide with the direction of the transects.

- However, in most cases, one would probably want to pool the distance data to obtain a more accurate estimate of $f(0)$. In this case, $\widehat{D}_i = \dfrac{y_i \widehat{f}(0)}{2L_i}$ and the $\widehat{D}_i$'s are *not* independent because $\widehat{f}(0)$ is based on the pooled data. In this case,

$$\widehat{D}_r = \frac{\sum_{i=1}^{n} L_i \widehat{D}_i}{\sum_{i=1}^{n} L_i} = \frac{\sum_{i=1}^{n} \dfrac{y_i \widehat{f}(0)}{2}}{\sum_{i=1}^{n} L_i} = \frac{\sum_{i=1}^{n} y_i}{2\sum_{i=1}^{n} L_i} \widehat{f}(0) \tag{17}$$

  which is the formula Thompson gives on page 243 (2nd ed: p. 213).

- If one ignored the lack of independence in the $\widehat{D}_i$'s, then the adjusted estimator of variance would be the one given by Thompson at the top of page 214, which he calls $\widehat{\mathrm{Var}}_1(\widehat{D}_r)$. However, this variance estimate is not appropriate because of the lack of independence.

- This is a situation for which a variance estimator like the bootstrap is needed. To bootstrap the standard error, you would take random samples, with replacement, of the n transects. For each bootstrap sample, you would first pool the distance data (if a transect was selected twice in the bootstrap sample, then all distances for that transect would be entered twice) and estimate $f(0)$ by the same method by which you estimated $f(0)$ for the original sample (e.g., normal kernel density method with default bandwidth). Then you would compute the estimator in equation (17) above. The standard error of the estimate based on the original sample would be the standard deviation of the $\widehat{D}_r$'s computed for the bootstrap samples.

Analysis of Distance Data using R

Two functions were written in R, called **distance.fit** and **distance.boot** to help analyze distance data from line transects. The function **distance.fit** computes the estimates of $f(0)$ for three different methods: exponential, half-normal and normal kernel density estimate. You can optionally specify the window width h for the kernel method; the default is to use the value from equation (12) on page 208 of Thompson. This function also draws a plot showing each of the curves overlaid on a histogram of the distances. You can specify the approximate number of bars for the histogram using the optional argument **nclass** in **distance.fit**; the default is to let R do it.

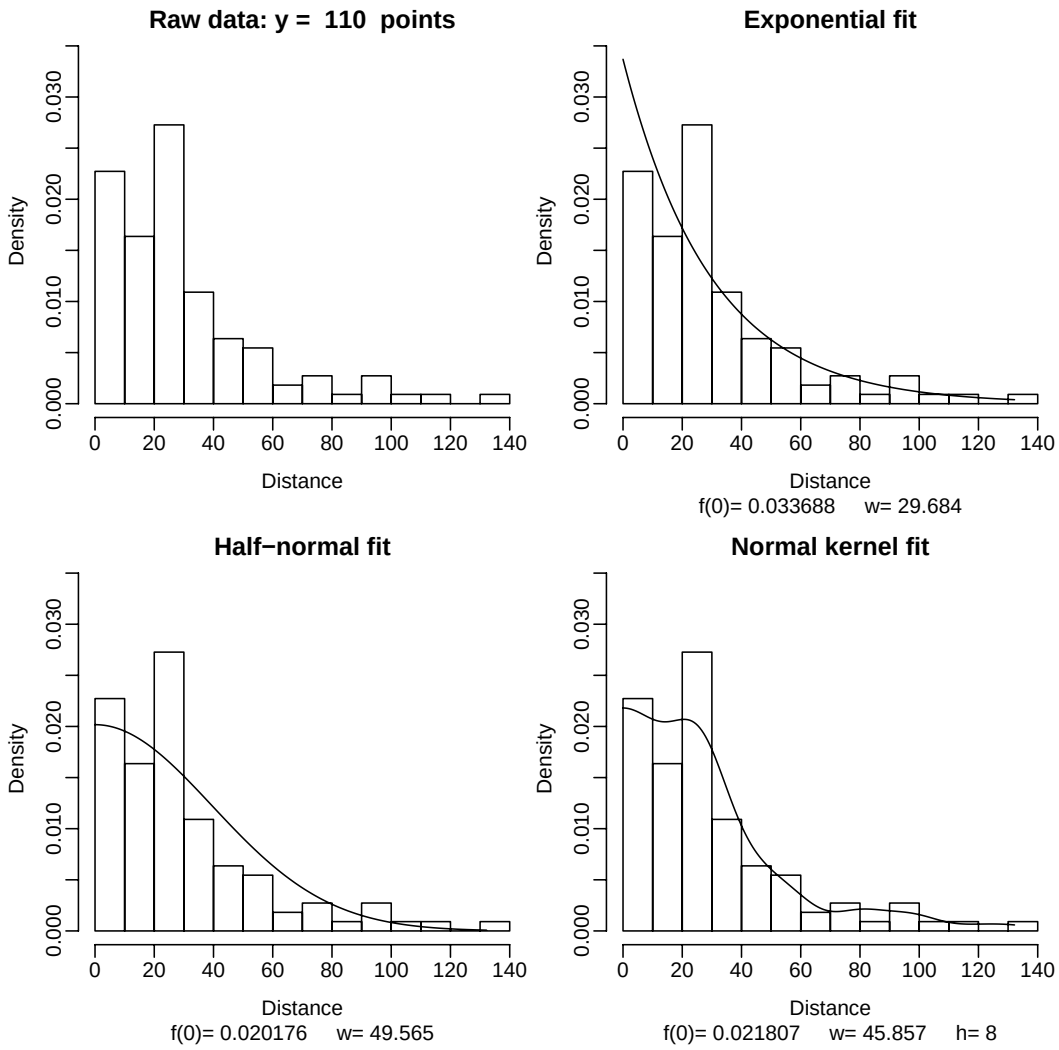
The function **distance.boot** computes the ratio estimate of density for a set of transects and

a bootstrap estimate of the standard error. The distance data must be given to this function in a special form known in R as a "list." This is required because, in order to carry out the bootstrap, **distance.boot** must have the distance data separately for each transect since it is the transects which are sampled.

Example 1 (continued): Back to the example on p. 160 of the notes. The $y = 110$ detection distances came from four different transects with $y_1 = 38, y_2 = 26, y_3 = 25$, and $y_4 = 21$. These data have been entered into a two-column text file, DistanceData.txt, with 110 rows. The first column is the transect number (1 through 4) and the second column is the distance. We could also have entered the distances directly into four separate vectors in R.

```
> DistanceData <- read.table("DistanceData.txt",col.names=c("Transect","Distance"))
> x <- DistanceData$Distance
> transect <- DistanceData$Transect
> distance.fit(x) # Does parametric and normal kernel fits; plot shown below
No. of distances= 110
  Exponential: f(0)_hat= 0.03368765  w_hat= 29.68447
  Half-normal: f(0)_hat= 0.02017567  w_hat= 49.56465
  Kernel: f(0)_hat= 0.02231597  w_hat= 44.81095   Window width: h= 6.418256
> distance.fit(x, nclass=15, h=8) # Change defaults; plot not shown
No. of distances= 110
  Exponential: f(0)_hat= 0.03368765  w_hat= 29.68447
  Half-normal: f(0)_hat= 0.02017567  w_hat= 49.56465
  Kernel: f(0)_hat= 0.02180684  w_hat= 45.85717   Window width: h= 8
```

Based on the plots, it appears that a half-normal detection function or normal kernel function might be appropriate.



We can now compute the ratio estimate of density based on one of these models. Since the transect lengths are all the same (1000 units) for these data, the ratio and unbiased estimates are the same, but we'll illustrate the ratio estimate calculation anyway. For example, for the half-normal:

```
> yi <- tapply(x,transect,length) # number of detections by transect
> yi
 1  2  3  4
38 26 25 21
> Li <- rep(1000,4)   # transect lengths
> Di <- (yi*0.02017567)/(2*Li)  # estimated density from each transect
> Di
           1            2            3            4
0.0003833377 0.0002622837 0.0002521959 0.0002118445
```

```
> sum(Li*Di)/sum(Li) # ratio estimate of density
[1] 0.0002774155
> mean(Di)  # unbiased estimate (same as ratio because transects are same length)
[1] 0.0002774155
```

The function **distance.boot** will compute the ratio estimate and also calculate a bootstrap standard error. To use this function, we need to create a list of the detection distances by transect.

```
> x1 <- x[transect==1]
> x2 <- x[transect==2]
> x3 <- x[transect==3]
> x4 <- x[transect==4]
> x0 <- list(x1,x2,x3,x4)
```

In the function **distance.boot**, the first argument is the "list" of vectors of distances, the second argument is the vector of transect lengths, the third argument is the number of bootstrap replications desired, and the fourth argument is the method: "exp", "hnorm" or "kernel" (the quotation marks are necessary). If method "kernel" is chosen, then you can specify the bandwidth h; if it's not specified, the function will use the default method in eq. (19.12) on p. 238 of Thompson (2nd ed: eq. (12) on p. 208) with a new bandwidth selected for each bootstrap sample. The function will print (but not store) the numerical summary. The values of all the bootstrap estimates of density will be stored in whatever object you choose to store the results in (b1, b2 and b3 below), if you care to look at their distribution.

```
> b1 <- distance.boot(x0,Li,10000,"hnorm")
Ratio estimator: Dhat= 0.0002774154   SE (from formula, p.214)=3.69505e-05
Bootstrap SE= 3.16217e-05 ,  Estimated bias=8.207024e-07
> b2 <- distance.boot(x0,Li,10000,"kernel")
Ratio estimator: Dhat= 0.0003068446  SE (from formula, p.214)=4.08703e-05
Bootstrap SE= 3.52933e-05 ,  Estimated bias= -9.702187e-08
> b3 <- distance.boot(x0,Li,10000,"kernel",h=10)
Ratio estimator: Dhat= 0.0002942853  SE (from formula, p.214)=3.91975e-05
Bootstrap SE= 2.50177e-05 ,  Estimated bias= 8.812689e-08
```

The estimate of density using the half-normal is .0002774 objects per square unit with a bootstrap SE of .0000316 or, equivalently, an estimate of 277.4 objects total with an SE of 31.6. Using the kernel estimate with default bandwidth, the estimated total number of objects is 306.8 with a bootstrap SE of 35.3.

The functions **distance.fit** and **distance.boot** are given below and are available on the course web page in the R script file DistanceFunctions.R.

```
distance.fit <- function(x, nclass = NA, h = NA) {
# Computes estimate of f(x) (pdf of observed distances for transect data)
# by three methods: "exp","hnorm","kernel" (pp. 206-8 of Thompson) and shows
# graphs of fits for all three to histogram of the raw data.
# "exp" is exponential, "hnorm" is half-normal, and "kernel" is kernel
# estimate with normal kernel.
# x is vector of observed distances,
# nclass is number of bars in histogram (may be omitted),
# h is the window width for the normal kernel method.
# If h is not given then formula (12), p. 208, is
# used to calculate a window width from the data.
       n <- 200        # no. of points at which to plot curves
       oldpar <- par(mfrow=c(2,2),mgp=c(2,0.5,0),mar=c(4,3,2 0)+0.1)
       y <- length(x)
       xx <- seq(0, max(x), length = n)
       what.exp <- mean(x) # effective half-width for exponential
       f.exp <- exp(-xx/what.exp)/what.exp   # exponential fit
       what.hn <- sqrt((pi/2)*mean(x^2))  # effective half-width for half-normal
       f.hn <- exp((-pi*xx^2)/(4*what.hn^2))/what.hn  # half-normal fit
       a <- min(sd(x),median(x)/1.34)
       if(is.na(h))
               h <- (0.9*a)/y^0.2    # default bandwidth
       f.ker <- rep(0, n)
       for(i in 1:n) f.ker[i] <- (1/(y*h))*(1/sqrt(2*pi))*sum(exp(-0.5*
                      ((xx[i]-x)/h)^2)+exp(-0.5*((xx[i]+x)/h)^2))

     what.ker <- 1/f.ker[1]
       if(is.na(nclass)) {
               ymax <- max(c(f.exp,f.hn,f.ker,hist(x,plot=F)$density))
               hist(x+0.0001*max(x),col=0,freq=F,ylim=c(0, ymax),
                    main = paste("Raw data: y = ",y," points"),xlab=
                    "Distance")
               hist(x+0.0001*max(x),col=0,freq=F,ylim=c(0,ymax),
                    main="Exponential fit",sub=paste("f(0)=",round(1/
                    what.exp,6),"    w=",round(what.exp,3)),xlab="Distance")
               lines(xx, f.exp)
               hist(x+0.0001*max(x),col=0,freq=F,ylim=c(0,ymax),
```

```
            main="Half-normal fit",sub=paste("f(0)=",round(1/
            what.hn,6),"    w=",round(what.hn,3)),xlab =
            "Distance")
        lines(xx, f.hn)
        hist(x+0.0001*max(x),col=0,freq=F,ylim=c(0, ymax),
            main="Normal kernel fit",sub=paste("f(0)=",round(1/
            what.ker,6),"    w=",round(what.ker, 3),"    h=",
            round(h,4)),xlab="Distance")
        lines(xx,f.ker)
}
else {
    ymax <- max(c(f.exp,f.hn,f.ker,hist(x,plot=F,breaks=nclass)$density))
    hist(x+0.0001*max(x),col=0,breaks=nclass,freq = F,
        ylim = c(0,ymax),main=paste("Raw data: y = ",y,
        " points"),xlab = "Distance")
    hist(x+0.0001*max(x),col = 0,breaks = nclass,freq = F,
        ylim = c(0,ymax),main="Exponential fit",sub=paste(
        "f(0)=",round(1/what.exp, 6), "    w=",round(what.exp,
        3)),xlab="Distance")
    lines(xx, f.exp)
    hist(x + 0.0001 * max(x), col = 0, breaks = nclass, freq = F,
        ylim = c(0, ymax), main = "Half-normal fit", sub = paste(
        "f(0)=", round(1/what.hn, 6), "    w=", round(what.hn,
        3)), xlab = "Distance")
    lines(xx, f.hn)
    hist(x + 0.0001 * max(x), col = 0, breaks = nclass, freq = F,
        ylim = c(0, ymax), main = "Normal kernel fit", sub =
        paste("f(0)=", round(1/what.ker, 6), "    w=", round(
        what.ker, 3), "    h=", round(h, 4)), xlab = "Distance")
    lines(xx, f.ker)
}
cat("No. of distances=", y)
cat("\n  Exponential: f(0)_hat=", round(1/what.exp, 8), " w_hat=",
    round(what.exp, 8))
cat("\n  Half-normal: f(0)_hat=", round(1/what.hn, 8), " w_hat=",
    round(what.hn, 8))
cat("\n  Kernel: f(0)_hat=", round(1/what.ker, 8), " w_hat=",
    round(what.ker, 8), "  Window width: h=", round(h, 8))
cat("\n")
par(oldpar)
```

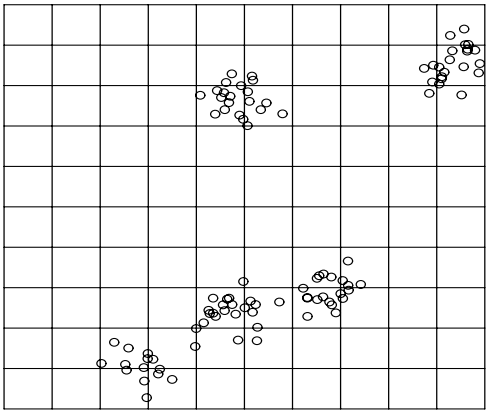
```
        invisible()
}

distance.boot <- function(x, l, m, method, h = NA) {
# Compute ratio estimator of density for transect data (p. 213) along
# with SE from formula on p. 214 and bootstrap SE.
# x is a list of n vectors of distances at which objects were
# observed on n random transects and l is a vector of
# lengths of the transects.
# m is the number of bootstrap replications. method is the
# method used to estimate the effective half-width:
# "exp" means exponential detectability function,
# "hnorm" means half-normal and "kernel" means kernel method with
# normal kernel. h is the window width for the normal kernel method;
# if h is not given then eq. (12),p. 208 of Thompson is used.
        n <- length(x)
        if(n < 2)
                stop("x must be a list of length >= 2")
        if(length(l) != n)
                stop("length of l must be number of transects")
        if(is.na(match(method, c("exp", "hnorm", "kernel"))))
                stop("method not valid")
        yi <- unlist(lapply(x, length)) # vector of no. of detections
# First compute density estimate from original data
        ux <- unlist(x)
        if(method == "exp")
                f0 <- 1/mean(ux)
        else if(method == "hnorm")
                f0 <- 1/sqrt((pi/2) * mean(ux^2))
        else {
                if(is.na(h))
                  h <- (0.9 * min(sqrt(var(ux)), median(ux)/1.34))/
                       length(ux)^0.2
                f0 <- (2/(h * sqrt(2 * pi))) * mean(exp(-0.5 * (ux/h)^2))
        }
        Dhat <- (length(ux) * f0)/(2 * sum(l))
        se.Dhat <- sqrt(sum(((yi * f0)/2 - Dhat * l)^2)/(n * (n - 1) *
                   mean(l)^2))
        cat("Ratio estimator: Dhat=", round(Dhat, 10),
            "     SE (from formula, p.214)=", round(se.Dhat, 10),
```

```
        "\n") # Bootstrap
    dhat <- rep(0, m)
    for(i in 1:m) {
            z <- sample(n, size = n, replace = T)
            xz <- unlist(x[z])
            if(method == "exp")
                    f0 <- 1/mean(xz)
            else if(method == "hnorm")
                    f0 <- 1/sqrt((pi/2) * mean(xz^2))
            else {
                    if(is.na(h))
                      h <- (0.9 * min(sqrt(var(xz)), median(xz)/1.34))/i
                           length(xz)^0.2
                    f0 <- (2/(h * sqrt(2 * pi))) * mean(exp(-0.5 *
                          (xz/h)^2))
            }
            dhat[i] <- (f0 * length(xz))/(2 * sum(l[z]))
    }
    cat("Bootstrap SE=", round(sqrt(var(dhat)), 10),
        ",  Estimated bias=",
        mean(dhat) - Dhat, "\n")
    dhat
}
```

Adaptive Cluster Sampling (Chapter 24)

Suppose it is desired to estimate the abundance or density of objects (such as plants or animals) which are clustered, like those in Figure 1 below. One sampling plan would be to divide the area into equal sized plots with a grid, take an SRS of plots, and estimate the total number of objects from the counts on those plots. It's quite likely that most (or perhaps even all) of the sampled plots would have no objects. It's also likely that, because of the clustered distribution, if the observer finds objects on one plot, there is a high likelihood of finding objects on neighboring plots. Could it improve estimation of the total if a researcher were allowed to search neighboring plots of plots that contain objects? This is the reasoning behind adaptive sampling.

The following example shows a square region divided into 100 plots with 100 objects clustered in the region. The plot on the lower right shows the number of objects on each plot.



									4
				3	2			3	13
				10	6			1	1
				1		4	4		
				12	6	9	3		
		7	3	1	1				
		3	3						

Figure 1. $\tau = 100$ objects distributed on a 10×10 grid with $N = 100$ plots.

The basic idea in adaptive sampling is to start with an initial sample of plots (the plot is the sampling unit) and set some condition for the response variable (the number of objects in the plot). If a plot satisfies the condition, then all plots in a neighborhood of the initial plot are searched. If any of the plots in the neighborhood of the original plot satisfies the condition, then all plots in the neighborhoods of those plots are searched, and so on, until no more plots that satisfy the condition are found.

In Figure 1, suppose we define the neighborhood of a plot to consist of the four plots directly above, below, to the left and to the right of the initial plot. Suppose also that the condition is that the number of objects is greater than or equal to 1; that is, the neighborhood of a plot will be searched if the plot has any objects in it. Given these definitions of neighborhood and the condition under which a neighborhood of a plot will be searched, suppose that in Figure 2 below, the three plots with an "X" in them were the three randomly chosen initial plots. Then all points with an

asterisk in Figure 2 would also be included in the sample.

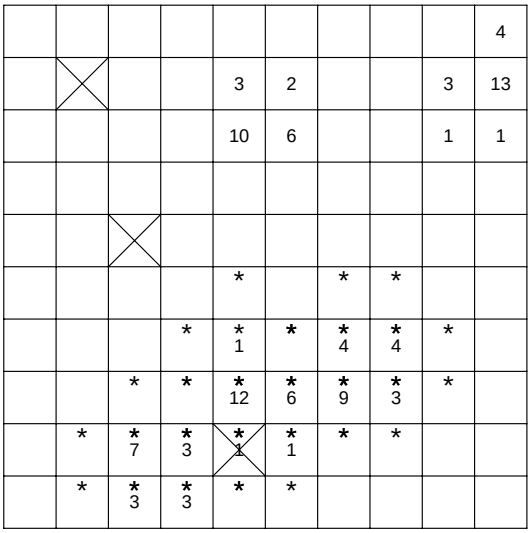


Figure 2. An adaptively selected sample.

The sample mean number of objects per plot is likely to overestimate the mean for the population because we've purposely included extra plots which are more likely to have objects in them. Can we construct unbiased estimators of the mean per plot and estimate the variance of the estimators? The answer is yes.

To understand precisely how adaptive cluster sampling works and how the corresponding population quantities (total, mean) are estimated based on such a design, consider first the following notation and terminology. Let

$$\begin{aligned} N &= \text{total number of plots in the population,} \\ y_i &= \text{the number of objects in the } i^{th} \text{ plot,} i = 1, \ldots, n, \\ \tau &= \sum_{i=1}^{N} y_i = \text{the total number of objects,} \\ \mu &= (1/N)\tau = \text{mean number of objects per plot.} \end{aligned}$$

Some terminology:

- **Neighborhood**: The neighborhood A_i of plot i is the set of plots that will be added to the sample if plot i satisfies some condition. In the example above, we defined the neighborhood of a plot to be the four plots above, below, and to the right and left of the initial plot. It

could also be all eight plots that surround the initial plot. However, one wants to be careful of making the neighborhood too big as the sample can grow quickly. The only requirement on the definition of a neighborhood is that if plot i belongs to the neighborhood of plot j, then plot j belongs to the neighborhood of plot i. So, for example, the neighborhood could not consist of just the two plots above and to the right of the initial plot, but it could consist of the two plots above and below the initial plot.

- **Condition of Interest**: Let C be a set of values in the range of the variable of interest. If y_i is in C, then plot i is said to satisfy the condition and all plots in its neighborhood are added to the sample. C is set by the researcher; generally, C is the set of values greater or equal to some constant. In our example, we used $C = \{y : y \geq 1\}$; that is, the plot satisfies the condition if it contains at least one object.

- **Cluster**: A cluster is the set of all plots that would be included in the sample if plot i were chosen initially. The cluster consists of only plot i if plot i doesn't satisfy the condition. If plot i satisfies the condition then the cluster includes plot i plus plot i's neighborhood, plus the neighborhoods of any of plot i's neighbors that satisfy the condition, etc. In Figure 2, the 28 points with asterisks are a cluster. This cluster would be selected if any of the non-zero plots within it were selected in the initial sample.

- **Network**: A network is a subset of a cluster, consisting of all units in the cluster that satisfy the condition. In Figure 2, the 13 plots with asterisks that also have at least one object in them constitute a network. A network has the property that if any unit in the network is chosen in the initial sample, then all units in the network will be included.

- **Edge Unit**: An edge unit is a unit which does *not* satisfy the condition, but is in the neighborhood of a plot that does. In Figure 2, the 15 plots with asterisks that do not satisfy the condition are all edge units. Selection of an edge unit will not result in the selection of any other plots in a cluster. A plot can be an edge unit of more than one cluster or network.

It's convenient to think of a plot that does not satisfy the condition as a network of size 1. Then, the networks form a *partition* of the plots in the grid (a separation of them into disjoint subsets whose union is the whole set of plots). The networks for the example are delineated by the bold lines in Figure 3. There are 81 networks, only three of which are bigger than one plot.

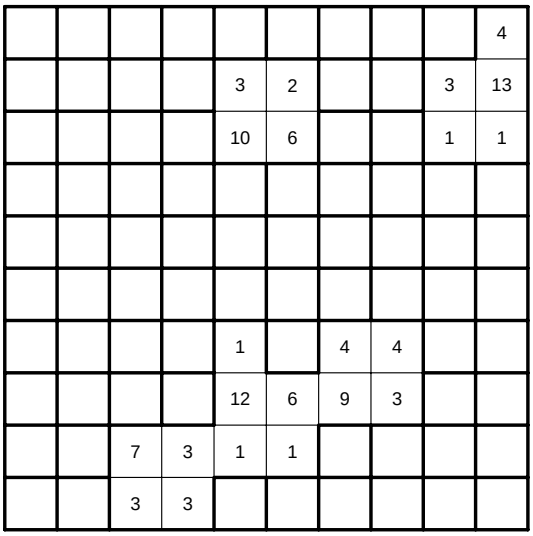


Figure 3. Bold lines delineate the networks for the example in Figure 1.

Estimation of density and abundance

The sampling protocol is that an initial random sample of n plots will be selected, either with or without replacement. All plots in the neighborhoods of the initial plots satisfying the condition will be added, and so on. We'll describe two estimators, each of which can be adapted to with or without replacement sampling of the initial plots. One is a Hansen-Hurwitz estimator based on draw-by-draw selection probabilities. The second is a Horvitz-Thompson estimator based on inclusion probabilities. Both account for the fact that the bigger the network that a plot belongs to, the higher the probability of inclusion.

There are two key elements to the estimation. One is to consider the network the sampling unit. That is because selection of a plot selects all plots in the network that the initial plot is part of (of course, the initial plot may be a network of size 1). It is possible to determine the inclusion probability for a network. The second key element is not to include in the estimator any edge units that were not part of the initial sample. Those edge units are part of the sample, but play no role in the estimator. The reason they are not included is that it is not possible to calculate an inclusion probability for an edge unit. Since we don't search the neighborhood of an edge unit, we can't tell if it is an edge unit of another network not in the sample. If it were, its inclusion probability would depend on the size of that network.

Hansen-Hurwitz estimator

Suppose the initial random sample of n plots is selected with replacement. The Hansen-Hurwitz

estimator of a population total is

$$\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

where p_i is the selection probability for the i^{th} unit on any draw. The Hansen-Hurwitz estimator is unbiased. The theoretical variance and an unbiased estimator of it are

$$\mathrm{Var}(\widehat{\tau}_p) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - \tau\right)^2, \qquad \widehat{\mathrm{Var}}(\widehat{\tau}_p) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{\tau}_p\right)^2.$$

Recall that we are considering the network the sampling unit. Each selection of an initial plot is equivalent to the selection of the network to which that plot belongs. Then, in the formula above, y_i represents the total number of objects in the network selected. The notation now can become a little confusing because we have to decide whether we should change notation and index everything by network number rather than plot number. However, since we already have everything indexed by plot number, let's keep it that way, even though the wording gets a little awkward. When we select plot i, we're really selecting the whole network that plot i belongs to. We'll want to define variables that indicate how many plots are in that network and what the total number of objects (the total of the y-values) is for that network. So we'll define

$$\begin{aligned}
\Psi_i &= \text{the indices of the plots that belong to the network that plot } i \text{ belongs to,}\\
m_i &= \text{the number of plots in network } \Psi_i,\\
\tau_i &= \sum_{j\in\Psi_i} y_j = \text{ the total number of objects in network } \Psi_i\\
w_i &= \frac{\tau_i}{m_i} = \text{ the mean number of objects per plot in network } \Psi_i.
\end{aligned}$$

We have to remember that the index i refers to plot i, not network i. For example, w_i is the same for all plots that belong to the same network – it's the mean number of objects per plot in the network that plot i belongs to. All the notation above is the same as in Thompson Sec. 24.2 (pp. 323-4, 2nd ed: pp. 293-4) except that he doesn't define τ_i.

The selection probability on a single draw for the network that plot i belongs to is simply $p_i = m_i/N$ because the network that plot i belongs to will be selected if any of the m_i plots in the network is selected. Hence, the Hansen-Hurwitz estimator of τ for an initial random sample of n plots with replacement is

$$\widehat{\tau}_p = \frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i}{m_i/N} = \frac{N}{n}\sum_{i=1}^{n}\frac{\tau_i}{m_i} = \frac{N}{n}\sum_{i=1}^{n} w_i.$$

The Hansen-Hurwitz estimator of the population mean number of objects per plot is

$$\widehat{\mu}_1 = \frac{\widehat{\tau}_p}{N} = \frac{1}{n}\sum_{i=1}^{n} w_i.$$

($\widehat{\mu}_1$ is the notation Thompson uses). Remember that if a network is selected more than once (either because the same initial plot was selected or because different plots in the network were selected in

the initial sample), then it appears multiple times in the estimator.

The form of the Hansen-Hurwitz estimator is that of a simple sample mean (of the w_i values for the plots selected) for a random sample with replacement. That would suggest that the variance and estimated variance of $\widehat{\mu}_1$ are simply

$$\mathrm{Var}(\widehat{\mu}_1) = \frac{\sigma^2}{n} = \frac{1}{nN}\sum_{i=1}^{N}(w_i - \mu)^2, \qquad \widehat{\mathrm{Var}}(\widehat{\mu}_1) = \frac{s^2}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(w_i - \widehat{\mu}_1)^2$$

where σ^2 is the population variance and s^2 the sample variance of the w_i values. We can derive these formulas from the general formulas for the theoretical and estimated variances of a Hansen-Hurwitz estimator. But when we apply the formula for the theoretical variance we have to remember that the sum is taken over all distinct sampling units in the population. Since the sampling unit is a network, we must sum over all the networks, say K, in the population. We'll be a little sloppy in notation in the following derivation and let m_k, τ_k, and w_k denote the number of plots, the total number of objects, and the mean number of objects per plot for the k^{th} *network* in the population. Then

$$\begin{aligned}
\mathrm{Var}(\widehat{\mu}_1) &= \mathrm{Var}(\widehat{\tau}_p/N) = \left(\frac{1}{N^2}\right)\mathrm{Var}(\widehat{\tau}_p) = \frac{1}{N^2 n}\sum_{k=1}^{K} p_i \left(\frac{y_k}{p_k} - \tau\right)^2 \\
&= \frac{1}{N^2 n}\sum_{k=1}^{K}\left(\frac{m_k}{N}\right)\left(\frac{\tau_k}{m_k/N} - \tau\right)^2 = \frac{1}{N^2 n}\sum_{k=1}^{K}\left(\frac{m_k}{N}\right)(Nw_k - N\mu)^2 \\
&= \frac{1}{nN}\sum_{k=1}^{K} m_k(w_k - \mu)^2.
\end{aligned}$$

Now we just have to recognize that $\displaystyle\sum_{k=1}^{K} m_k(w_k - \mu)^2 = \sum_{i=1}^{N}(w_i - \mu)^2$. Do you see why?

The formula for $\widehat{\mathrm{Var}}(\widehat{\mu}_1)$ follows similarly, but in this formula, we sum over all sampling units in the sample, even if there are repeats. Therefore, we can still use the index i.

Numerical Example

Suppose the initial sample of size $n = 3$ was chosen with replacement and was as illustrated in Figure 2 and the 27 additional plots with asterisks were therefore added to the sample. Numbering the initial plots from top to bottom, we have:

$$m_1 = 1, m_2 = 1, m_3 = 13; \qquad \tau_1 = 0, \tau_2 = 0, \tau_3 = 57; \qquad w_1 = 0, w_2 = 0, w_3 = 57/13.$$

Therefore,

$$\widehat{\mu}_1 = (0 + 0 + 57/13)/3 = 57/39 \approx 1.46$$

and

$$\widehat{\text{Var}}(\widehat{\mu}_1) = \frac{1}{3(2)}[(0 - 57/39)^2 + (0 - 57/39)^2 + (57/13 - 57/39)^2] \approx 2.136$$

and $\text{SE}(\widehat{\mu}_1) = \sqrt{2.136} \approx 1.46$. The estimated total number of objects in the whole grid is

$$N\widehat{\mu}_1 = (100)(57/39) \approx 146.2$$

with standard error

$$N \text{ SE}(\widehat{\mu}_1) = 100(1.462) \approx 146.2.$$

Recall that the population mean and total are $\mu = 1$ and $\tau = 100$. The standard error of the estimates is relatively large in this example; we'll investigate later in this handout whether adaptive sampling is helpful for this population. Note also that if two of the initial plots had ended up in the same network, then we would have included that network twice in the estimator.

Returning to $\widehat{\mu}_1 = (1/n)\sum_{i=1}^{n} w_i$, suppose we were to replace the value of y_i for every plot in the population by w_i, the mean for all plots in that plot's network. For our example, the population would look like Figure 4.

0	0	0	0	0	0	0	0	0	4.4
0	0	0	0	5.25	5.25	0	0	4.4	4.4
0	0	0	0	5.25	5.25	0	0	4.4	4.4
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	4.38	0	4.38	4.38	0	0
0	0	0	0	4.38	4.38	4.38	4.38	0	0
0	0	4.38	4.38	4.38	4.38	0	0	0	0
0	0	4.38	4.38	0	0	0	0	0	0

Figure 4. The value of w_i is displayed in each plot for the population in Figure 1.

The values of τ and μ are exactly the same for this population as for the original population. The Hansen-Hurwitz estimator can then be viewed as the mean of a random sample with replacement from the population in Figure 4. This way of looking at the Hansen-Hurwitz estimator also suggests when adaptive cluster sampling might be better than random sampling. The population in Figure 4 has lower variance among the plots than the original population in Figure 1. Therefore, the sample mean count from n plots from the population in Figure 4 will have lower variance than the sample mean count from n plots from the population in Figure 1. How much lower depends on the amount of variability within networks. The more variability in the plot counts within networks, the bigger the difference. However, the advantage of adaptive sampling has to be weighed against the extra cost involved because in order to observe the value of w_i for a plot, you must sample the whole network to which that plot belongs. The initial sample size n does not reflect that. The actual sample size, say n^*, for an adaptive plan is a random variable. Later, when we compare SRS to adaptive sampling, we'll use the *expected* sample size for the adaptive plan.

Viewing the Hansen-Hurwitz estimator for an adaptive plan as simply the mean of a random sample with replacement of n plots from the population in Figure 4 makes it clear that we could also draw the initial sample of n plots without replacement, that is, an SRS of size n. In that case, the estimator $\widehat{\mu}_1$ would be the same but the variance would have the finite population correction:

$$\mathrm{Var}(\widehat{\mu}_1) = \left(\frac{N-n}{N}\right)\frac{1}{n(N-1)}\sum_{i=1}^{N}(w_i-\mu)^2, \qquad \widehat{\mathrm{Var}}(\widehat{\mu}_1) = \left(\frac{N-n}{N}\right)\frac{1}{n(n-1)}\sum_{i=1}^{n}(w_i-\mu)^2.$$

Numerical Example

In the earlier example from Figure 2, we calculated the standard error of the Hansen-Hurwitz estimates as if the initial sample was drawn with replacement. If it were drawn without replacement, then the SE of the estimated mean would have been $\mathrm{SE}(\widehat{\mu}_1) = \sqrt{((100-3)/100)2.136} = 1.439$ and of the estimated total as 143.9 (as opposed to 1.462 and 146.2).

Horvitz-Thompson estimator

We could also use a Horvitz-Thompson estimator of the population mean and total. The Horvitz-Thompson estimator is based on the distinct plots sampled or, equivalently, on the distinct networks sampled. As with Hansen-Hurwitz, edge units are not used in the estimator unless they were part of the initial sample. Whether the initial sample is with or without replacement, the same network can be sampled more than once but we only include it in the estimator once.

The general formulas for the Horvitz-Thompson estimator and its variance and estimated variance are given in Sec. 6.2 of Thompson. To apply these formulas to adaptive cluster sampling, it is easier to index quantities by network rather than plot, so we introduce some new notation,

following Thompson (p. 326; 2nd ed: p. 296). Let

$$\begin{aligned} x_k &= \text{the number of plots in network } k, k = 1, \ldots, K, \\ \alpha_k &= \text{probability of inclusion of network } k, \\ y_k^* &= \text{total number of objects in network } k. \end{aligned}$$

These quantities are analogous to m_i, π_i' (defined on p. 325 of Thompson; 2nd ed: p. 295) and τ_i which were indexed by plot number. With this notation, the Horvitz-Thompson estimator of the population total is

$$\widehat{\tau}_\pi = \sum_{k=1}^{v} \frac{y_k^*}{\alpha_k}$$

with variance

$$\mathrm{Var}(\widehat{\tau}_\pi) = \sum_{k=1}^{K} \left(\frac{1-\alpha_k}{\alpha_k} \right) y_k^{*2} + \sum_{k=1}^{K} \sum_{h \neq k} \left(\frac{\alpha_{kh} - \alpha_k \alpha_h}{\alpha_k \alpha_h} \right) y_k^* y_h^*$$

and estimated variance

$$\widehat{\mathrm{Var}}(\widehat{\tau}_\pi) = \sum_{k=1}^{v} \left(\frac{1-\alpha_k}{\alpha_k^2} \right) y_k^{*2} + \sum_{k=1}^{v} \sum_{h \neq k} \left(\frac{\alpha_{kh} - \alpha_k \alpha_h}{\alpha_k \alpha_h} \right) \frac{y_k^* y_h^*}{\alpha_{kh}}$$

where v is the number of distinct networks in the sample and α_{kh} is the joint inclusion probability of networks k and h, $k \neq h$ (that is, the probability that both networks are included).

To calculate α_k, first suppose the initial sample of n plots is taken without replacement. A network is included in the sample only if the n initial plots include at least one plot in the network. This is one minus the probability that the n initial plots don't include any plots from the k^{th} network. Therefore, the probability of inclusion of the k^{th} network is

$$\alpha_k = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}}.$$

To calculate α_{kh}, recall that for any two events A and B,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B).$$

Here, A represents the event that network k is included and B the event that network h is included in the sample. The probability $\mathrm{P}(A \cup B)$ that at least one of the networks k and h is included in the sample is one minus the probability that neither is included:

$$\mathrm{P}(A \cup B) = 1 - \binom{N - x_k - x_h}{n} \Big/ \binom{N}{n}.$$

Also, $\mathrm{P}(A) = \alpha_k$ and $\mathrm{P}(B) = \alpha_h$. Hence,

$$\begin{aligned} \alpha_{kh} &= \mathrm{P}(A \cap B) = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}} + 1 - \frac{\binom{N-x_h}{n}}{\binom{N}{n}} - \left[1 - \frac{\binom{N-x_k-x_h}{n}}{\binom{N}{n}} \right] \\ &= 1 - \frac{\binom{N-x_k}{n} + \binom{N-x_h}{n} - \binom{N-x_k-x_h}{n}}{\binom{N}{n}}. \end{aligned}$$

The Horvitz-Thompson estimator of the population mean is

$$\widehat{\mu}_2 = \frac{\widehat{\tau}_\pi}{N} = \frac{1}{N}\sum_{k=1}^{v}\frac{y_k^*}{\alpha_k}$$

with variance $\mathrm{Var}(\widehat{\mu}_2) = (1/N^2)\mathrm{Var}(\widehat{\tau}_\pi)$ and estimated variance $\widehat{\mathrm{Var}}(\widehat{\mu}_2) = (1/N^2)\widehat{\mathrm{Var}}(\widehat{\tau}_\pi)$.

If the initial sample were chosen with replacement, then the inclusion probabilities would change:

$$\begin{aligned}\alpha_k &= 1-(1-x_k/N)^n \\ \alpha_{kh} &= 1-[(1-x_k/N)^n+(1-x_h/N)^n-(1-(x_k+x_h)/N)^n]\end{aligned}$$

Numerical Example

For the sample in Figure 2, the three networks associated with the initial sample of $n = 3$ plots are all distinct. We'll assume the three initial plots were chosen without replacement. Then, numbering the three networks as before,

$$x_1 = 1, x_2 = 1, x_3 = 13; \quad y_1^* = 0, y_2^* = 0, y_3^* = 57$$

and the inclusion probabilities are

$$\begin{aligned}\alpha_1 &= 1-\binom{100-1}{3}\Big/\binom{100}{3} = 0.03 \\ \alpha_2 &= 1-\binom{100-1}{3}\Big/\binom{100}{3} = 0.03 \\ \alpha_3 &= 1-\binom{100-13}{3}\Big/\binom{100}{3} = 0.344496 \\ \alpha_{12} &= 1-\left[\binom{100-1}{3}+\binom{100-1}{3}-\binom{100-1-1}{3}\Big/\binom{100}{3}\right] = 0.0006060606 \\ \alpha_{13} &= 1-\left[\binom{100-1}{3}+\binom{100-13}{3}-\binom{100-1-13}{3}\Big/\binom{100}{3}\right] = 0.007396413 \\ \alpha_{23} &= \alpha_{13} = 0.007396413\end{aligned}$$

This yields an estimated total number of objects of

$$\widehat{\tau}_\pi = \frac{0}{.03}+\frac{0}{.03}+\frac{57}{.344496} = 165.5$$

with SE = 134.0. The estimated mean number of objects per plot is therefore $\widehat{\mu}_2 = 1.65$ with SE = 1.34. The R script below illustrates this computation.

```
N <- 100; n <- 3; v <- 3
x <- c(1,1,13) y <- c(0,0,57)
alphak <- 1 - choose(N-x,3)/choose(N,n)   #inclusion probabilities
```

```
alphakh <- matrix(0,nrow=v,ncol=v)  # joint inclusion probabilities
for(k in 1:(v-1)){
   for(h in (k+1):v){
       alphakh[k,h] <- alphakh[h,k] <-
       1-(choose(N-x[k],n)+choose(N-x[h],n)-choose(N-x[k]-x[h],n))/
       choose(N,n)
   }
}
tauhat <- sum(y/alphak)
muhat.2 <- tauhat/N
term1 <- sum((1-alphak)*y^2/alphak^2)  # first term of variance of tauhat
term2 <- 0    # second term of variance
for(k in 1:(v-1)){
   for(h in (k+1):v){
       term2 <- term2 + 2*((alphakh[k,h]-alphak[k]*alphak[h])/
       (alphak[k]*alphak[h]))*y[k]*y[h]/alphakh[k,h]
   }
}
var.tauhat <- term1 + term2
var.muhat.2 <- (1/N^2)*var.tauhat
c(tauhat,sqrt(var.tauhat))   # estimate of total and SE
c(muhat.2,sqrt(var.muhat.2))   # estimate of mean and SE
```

The output is

```
> c(tauhat,sqrt(var.tauhat))   # estimate of total and SE
[1] 165.4591 133.9610
> c(muhat.2, sqrt(var.muhat.2))# estimate of mean and SE
[1] 1.654591 1.339610
```

The output is

```
> c(tauhat,sqrt(var.tauhat))   # estimate of total and SE
[1] 165.4591 133.9610
> c(muhat.2, sqrt(var.muhat.2))# estimate of mean and SE
[1] 1.654591 1.339610
```

Comparison of adaptive sampling with SRS for two examples

To compare an adaptive sampling scheme to an SRS, we need to account for the total sample size in the adaptive scheme, v, which is a random variable. v includes all plots that end up in the sample,

including the n initial plots and all neighbors of plots satisfying the condition. This includes edge units, which are sampled but not included in the estimate. Although we can't know v in advance, we can compute its expected value in a known population.

To compute $\mathrm{E}(v)$, let the random variable X_i be 1 if the i^{th} plot is included in the sample and 0 otherwise. Then $v = X_1 + X_2 + \ldots + X_N$ and $\mathrm{E}(v) = \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_N)$. To compute $\mathrm{E}(X_i)$, note that the probability that plot i is included in the sample is one minus the probability that none of the plots in its network nor the plots in networks of which plot i is an edge unit are in the initial sample; that is,

$$\pi_i = \mathrm{P}(X_i = 1) = \frac{\binom{N-m_i-a_i}{n}}{\binom{N}{n}}$$

where m_i is the number of units in the network to which plot i belongs and a_i is the number of plots in networks for which plot i is an edge unit (note that if plot i satisfies the condition, then $a_i = 0$, while if it doesn't satisfy the condition, then $m_i = 1$). Since X_i is a Bernoulli random variable, $\mathrm{E}(X_i) = \pi_i$, so

$$\mathrm{E}(v) = \pi_1 + \ldots + \pi_N.$$

If we know the population arrangement, as for the example in Figure 1, then we can calculate the theoretical variances of $\widehat{\mu}_1$ and $\widehat{\mu}_2$ and the expected sample size $\mathrm{E}(v)$. We can compare the theoretical variances to the variance of the mean from an SRS of the same expected sample size as the adaptive scheme. In actuality, it is probably more efficient to sample a given total number of plots with an adaptive scheme than with an SRS, because with the adaptive scheme, at least some of the plots are likely to be neighbors of each other and the travel time will be less.

Thompson, in Table 24.1 (p. 329; 2nd ed: p. 299), compares the variances of $\widehat{\mu}_1$, $\widehat{\mu}_2$ and the mean from an SRS of size $\mathrm{E}(v)$ for the population in Figure 24.1 for various initial sample sizes. He denotes the mean from the SRS by $\widehat{\mu}_0^*$. For his example, he finds that the H-T estimator $\widehat{\mu}_2$ has smaller variance than $\widehat{\mu}_1$ and that $\widehat{\mu}_2$ has smaller variance than $\widehat{\mu}_0^*$ for $n \geq 2$. The differences grow larger as the initial sample size n grows larger.

The table below gives a similar summary for the population in Figure 1 of these notes. The initial sample is assumed to be an SRS (that is, without replacement) of n plots. The pattern is exactly the same as for Thompson's example: $\widehat{\mu}_2$ (the H-T estimator) has smaller variance then $\widehat{\mu}_1$ (the H-H estimator) and the difference becomes larger as n gets larger. $\widehat{\mu}_2$ is worse than $\widehat{\mu}_0^*$ (the SRS estimator) for smaller sample sizes but becomes better at initial samples of $n \geq 30$.

n	E(v)	Var($\widehat{\mu}_1$)	Var($\widehat{\mu}_2$)	Var($\widehat{\mu}_0^*$)
1	5.50	3.5697	3.5697	1.0864
2	10.47	1.7668	1.7006	0.5409
3	14.96	1.1659	1.0792	0.3594
5	22.73	0.6851	0.5853	0.2150
10	36.88	0.3245	0.2243	0.1082
20	52.98	0.1442	0.0639	0.0561
30	62.45	0.0841	0.0242	0.0380
50	75.30	0.0361	0.0041	0.0207

We discovered earlier in examining the Hansen-Hurwitz estimator that the more variation within clusters, the better adaptive sampling would do relative to SRS. That may explain why adaptive sampling helps more for Thompson's example than the one presented here.

In both examples, the Horvitz-Thompson estimator had smaller variance than the Hansen-Hurwitz estimator for all sample sizes. The difference was large for large sample sizes. This certainly suggests that the Horvitz-Thompson estimator $\widehat{\mu}_2$ is preferred.

Extensions

The examples in these notes and in Thompson are about estimating the number and density of some type of object (plant or animal, for example). However, the response variable does not have to be a discrete count. It could be the biomass of a particular plant on each plot, for example; the condition for searching the neighborhood of a plot could then be framed in terms of the biomass. But there's another possibility. The response variable could be biomass but the condition could still be based on the number of plants found. That is, there's no reason that the condition can't be based on a different variable than the variable of interest. The condition variable must be recorded for every plot surveyed. The response variable must be recorded for every plot that is included in the estimator, i.e., every plot that either meets the condition or is in the initial sample. The response variable does not need to be recorded for edge units that were not part of the initial sample. The selection or inclusion probabilities are computed from the condition variable, while the means and totals needed are computed from the response variable.

Strip Adaptive Cluster Sampling (Chapter 25)

To utilize strip adaptive cluster sampling, suppose we have some study area divided into square plots. The plots are partitioned into equal-sized strips. These are the primary sampling units. We take a random sample of strips and search every plot within each strip. If any plot within these strips satisfies some condition (for example, contains one or more objects of interest), we also sample the plots in the neighborhood of the original plot. If any of those additional plots satisfies the condition of interest, then we sample the plots in their neighborhoods, etc. Plots, in this scenario, are the secondary sampling units.

The analysis of the data from an adaptive strip plan is similar to that for adaptive cluster sampling. Consider the following notation and terminology. Let:

$$\begin{aligned} N &= \text{the number of primary units (strips) in the population,} \\ M &= \text{the number of secondary units in each primary unit,} \\ u_{ij} &= \text{the } j^{th} \text{ plot in the } i^{th} \text{ primary unit. There are } MN \text{ plots.} \\ y_{ij} &= \text{the measurement taken on unit } u_{ij}. \end{aligned}$$

The goal is to estimate the mean $\mu = \dfrac{1}{MN}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$, or the total $\tau = NM\mu$. If we were sampling birds, μ might represent the mean number of birds per plot, and τ the total number of birds in the region.

Terminology

The terms *neighborhood, condition of interest, network,* and *edge unit* have the same meaning as in adaptive cluster sampling. These are all defined in terms of the secondary units, the plots; the strips are only used to select the initial plots to sample.

To illustrate these concepts, consider the example region shown on the next page. Here, we have a 20x20 square plot region where we might use a random sample of five vertical strips to estimate the number of objects in the region. This example is from pages 340 and 351 of Thompson (2nd ed: pp. 310 and 321), and we will revisit it later in this handout to estimate the total number of objects in the region (which is $\tau = 326$). The numbers in the boxes are the numbers of objects in the plots (blank indicates 0).

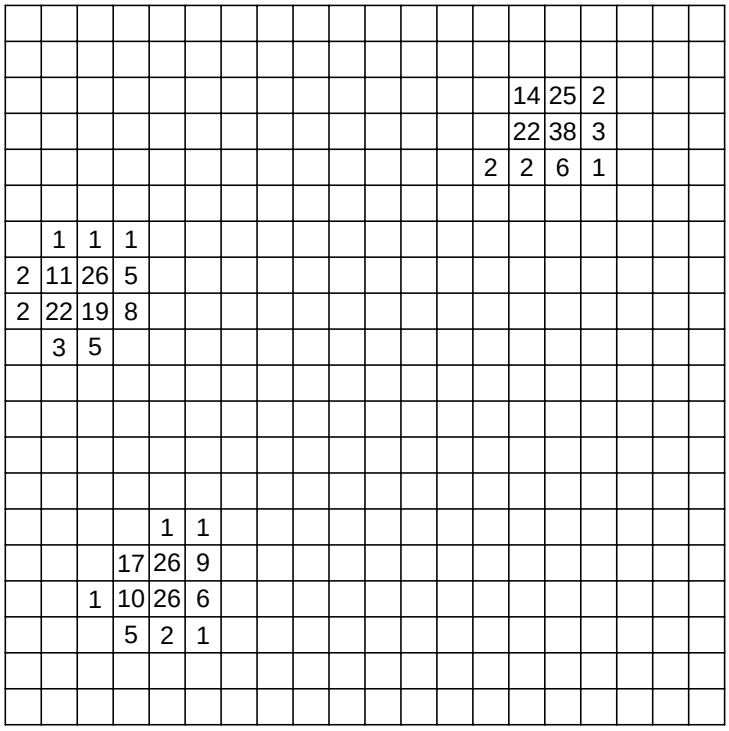


Figure 1: The example from Thompson, Chap. 25.

Estimation of Abundance and Density

As in adaptive cluster sampling, one key is to view the network as a sampling unit. The grid is partitioned into K networks just as for adaptive cluster sampling. A plot not satisfying the condition is considered to be a network of size 1. The fact that strip sampling is used to select the initial set of plots does not affect how networks are defined. That depends only on the definition of the neighborhood of a plot and the condition that a plot must satisfy in order for its neighborhood to be added to the sample. For the example illustrated in Figure 1, we'll use the same definition of neighborhood and the same criterion as in the adaptive cluster sampling example – the neighborhood of a plot consists of the four plots above, below and to the right and left of the plot, and the condition for searching the neighborhood is that there is at least one object in the plot.

We will look at three ways of estimating the population mean number of objects per unit. The first is just based on the initial strips selected, ignoring plots adaptively added to the sample. This is done only for comparison. The second and third use what are called partial selection probabilities and partial inclusion probabilities, respectively. These are analogous to the "separate transects"

and Horvitz-Thompson estimators for line-intercept sampling. In the separate transects approach, an unbiased estimate of the population density is computed for each strip and then the estimates are averaged over the n strips. The second approach uses a Horvitz-Thompson estimator based on all the strips together. As with cluster adaptive sampling, any edge units are ignored in the estimation unless they were part of the initial sample of strips.

1. Using a Simple Random Sample of Primary Units (No Adaptive Component):

   Let: $Y_i = \sum_{j=1}^{M} y_{ij}$ = the total of the y-values in the i^{th} primary unit (strip). Then an estimate of the mean response per secondary unit (plot) is given by:

$$\begin{aligned} \widehat{\mu}_0 &= \frac{1}{Mn}\sum_{i=1}^{n} Y_i, \text{ with an estimated variance given by:} \\ \widehat{\mathrm{Var}}(\widehat{\mu}_0) &= \frac{N-n}{M^2Nn}s_0^2, \text{ where } s_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M\widehat{\mu}_0)^2, \end{aligned}$$

   the usual sample variance of the strip totals.

2. Separate transects estimator (using partial selection probabilities)

   Thompson refers to this as the estimator based on "partial selection probabilities" (p. 344; 2nd ed: p. 314). We assume an SRS of n strips from the population of N possible strips is selected. Consider a single strip. The single-draw selection probability for a network k is the probability that a single random strip would intersect the network. This depends on the orientation of the strips. In Figure 1, the probability that a random vertical strip intersects a network is the horizontal "width" of the network (in terms of the number of plots) divided by the total number of vertical strips. For each of the three non-zero networks in Figure 1, the selection probability is $4/20 = .20$. The selection probability for a network of size 1 is $1/20 = .05$. In general, we let

$$\begin{aligned} x_k &= \text{the number of strips which intersect the } k^{th} \text{ network}, k = 1, \ldots, K, \\ y_k &= \text{the total of the } y \text{ values for the } k^{th} \text{ network (denoted } y_k^* \text{ in Chapter 24).} \\ p_k &= x_k/N = \text{the single draw selection probability for the } k^{th} \text{ network.} \end{aligned}$$

   Note: contrast this definition of x_k with that for adaptive cluster sampling where x_k was the number of plots in the k^{th} network.

   Before we proceed, we need to recognize that a single strip will intersect more than one network (unless the entire strip lies within one network). Hence, adaptive strip sampling is not a random sample with replacement of individual networks as adaptive cluster sampling

was. It is a random sample with replacement of a set of networks, the set consisting of all networks intersected by a strip. Therefore, we need some more definitions. Let

$$I_{ik} = \begin{cases} 1, & \text{if the } i^{th} \text{ strip intersects the } k^{th} \text{ network;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the sum of the y-values for all networks intersected by strip i is

$$\sum_{k=1}^{K} y_k I_{ik}.$$

Therefore, the Hansen-Hurwitz estimator (equivalent to Horvitz-Thomspson for a single strip) of the mean number of objects per plot based on the networks intersected by strip i is

$$w_i = \frac{1}{NM} \sum_{k=1}^{K} \frac{y_k I_{ik}}{p_k} = \frac{1}{NM} \sum_{k=1}^{K} \frac{y_k I_{ik}}{x_k/N} = \frac{1}{M} \sum_{k=1}^{K} \frac{y_k I_{ik}}{x_k}$$

(this is equation (25.1), p. 345 of Thompson; 2nd ed: eq. (1), p. 315). An unbiased estimate of the population density from the SRS of n strips is

$$\widehat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} w_i \quad \text{(equation (25.2), p. 345 in Thompson; 2nd ed: eq. (2), p. 315)}$$

with estimated variance

$$\widehat{\text{Var}}(\widehat{\mu}_1) = \left(\frac{N-n}{N} \right) \frac{s_w^2}{n}, \text{ where } s_w^2 = \frac{1}{n-1} \sum_{i=1}^{n} (w_i - \widehat{\mu}_1)^2.$$

3. Horvitz-Thompson estimator (using partial inclusion probabilities)

This estimator is based on the unique networks intersected by the SRS of n strips. Again, networks of size 1 (that is, plots which don't meet the condition) are included only if they are part of the initial sample (i.e., on one of the n initially selected strips).

Let α_k be the probability that network k is included in the final sample. This is the probability that one or more of the n strips (primary units) in the initial sample intersect network k. Recalling that x_k is the number of strips in the population of N strips which intersect network k, then

$$\alpha_k = 1 - \frac{\begin{pmatrix} N - x_k \\ n \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}.$$

The probability that networks k and j ($k \neq j$) are both included in the sample is

$$\alpha_{kj} = 1 - \frac{\begin{pmatrix} N - x_k \\ n \end{pmatrix} + \begin{pmatrix} N - x_j \\ n \end{pmatrix} - \begin{pmatrix} N - x_k - x_j + x_{kj} \\ n \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

where x_{kj} is the number of strips in the population which intersect both networks k and j. Now let

$$z_k = \begin{cases} 1, & \text{if one or more strips in the sample intersect network } k; \\ 0, & \text{otherwise.} \end{cases}$$

With these definitions of inclusion and joint inclusion probabilities, we have the following Horvitz-Thompson estimator and corresponding estimated variance for the mean per unit:

$$\widehat{\mu}_2 = \frac{1}{MN}\sum_{k=1}^{K}\frac{y_k z_k}{\alpha_k}, \quad \widehat{\mathrm{Var}}(\widehat{\mu}_2) = \frac{1}{M^2N^2}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{y_k y_j z_k z_j}{\alpha_{kj}}\left(\frac{\alpha_{kj}}{\alpha_k\alpha_j} - 1\right), \text{ where } \alpha_{kk} = \alpha_k.$$

Returning to Thompson's Animal Example

Recall the 20x20 square plot region where there are some number of animals in each unit (plot) of the region. It is clear (in viewing the region) that there are 3 non-zero networks in the region. Of course, we do not know this prior to sampling.

Suppose five strip samples are taken at random from the base of the region in columns 3,6,11,12, and 19. The column #3 strip intersects two networks, the column #6 strip intersects one network, and none of the other three strips intersect any networks. Using these five strip samples, each of the three methods discussed in this handout will be used to estimate the population mean number of animals per plot.

1. Using an SRS: There were 52, 17, 0, 0, and 0 animals respectively in the five strips. Hence, the mean number of animals per plot is estimated by:

$$\widehat{\mu}_0 = \frac{1}{Mn}\sum_{i=1}^{n} Y_i = \frac{1}{(20)(5)}(52+17) = \frac{69}{100} = \boxed{0.69}, \text{ with:}$$

$$\mathrm{SE}(\widehat{\mu}_0) = \sqrt{\frac{N-n}{M^2Nn}s_0^2} = \boxed{0.437}.$$

```
> n <- 5
> N <- 20
> M <- 20
> fpc <- (N-n)/N
> Y = c(52,17,0,0,0)
> mu.hat.0 <- mean(Y)/M
> SE.mu.hat.0 <- sqrt(fpc*var(Y)/n)/M
> c(mu.hat.0,SE.mu.hat.0)
[1] 0.6900000 0.4374071
```

2. Separate Transects Estimator Using Partial Selection Probabilities: Since only the first two strips intersected non-zero networks (those with animals), we need only compute w_1 & w_2 as defined earlier. The values of $w_3, w_4, \& w_5$ are all zero. Before computing these, note that there are 106 and 105 animals in the two clusters which are intersected. Using these:

$$\begin{aligned} w_1 &= \frac{1}{M}\sum_{k=1}^{K}\frac{y_k I_{1k}}{x_k} = \frac{1}{20}\left[\frac{106}{4}+\frac{105}{4}\right] = \underline{2.6375}, \\ w_2 &= \frac{1}{M}\sum_{k=1}^{K}\frac{y_k I_{2k}}{x_k} = \frac{1}{20}\left[\frac{105}{4}\right] = \underline{1.3125}. \end{aligned}$$

$$\text{This gives: } \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{5}\left[2.6375+1.3125+0+0+0\right] = \boxed{0.790}, \text{ with:}$$

$$\text{SE}(\widehat{\mu}_1) = \sqrt{\frac{N-n}{Nn}s_w^2} = \boxed{0.457}.$$

```
> w = (1/M)*c(106/4+105/4,105/4,0,0,0)
> mu.hat.1 <- mean(w)
> SE.mu.hat.1 <- sqrt(fpc*var(w)/n)
> c(mu.hat.1,SE.mu.hat.1)
[1] 0.790000 0.456559
```

3. Horvitz-Thompson Estimator Using Partial Inclusion Probabilities: The estimated inclusion probabilities and joint inclusion probability for the two intersected networks (both of width 4) are:

$$\alpha_k = 1 - \frac{\binom{N-x_k}{n}}{\binom{N}{n}} = 1 - \frac{\binom{20-4}{5}}{\binom{20}{5}} = \underline{0.7183}, \; k = 1, 2.$$

$$\begin{aligned} \alpha_{12} &= 1 - \frac{\binom{N-x_1}{n} + \binom{N-x_2}{n} - \binom{N-x_1-x_2+x_{12}}{n}}{\binom{N}{n}} \\ &= 1 - \frac{\binom{20-4}{5} + \binom{20-4}{5} - \binom{20-4-4+2}{5}}{\binom{20}{5}} = \underline{0.5657}. \end{aligned}$$

With these then, the estimated mean number of animals per unit is given by:

$$\widehat{\mu}_2 = \frac{1}{MN}\sum_{k=1}^{K}\frac{y_k z_k}{\alpha_k} = \frac{1}{400}\left[\frac{106}{.7183}+\frac{105}{.7183}\right] = \boxed{0.7344}, \text{ with:}$$

$$\text{SE}(\widehat{\mu}_2) = \sqrt{\frac{1}{M^2N^2}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{y_k y_j z_k z_j}{\alpha_{kj}}\left(\frac{\alpha_{kj}}{\alpha_k\alpha_j}-1\right)} = \boxed{0.316}.$$

```
> # Horvitz-Thompson estimator (partial inclusion probabilities)
> y1 <- 106; y2 <- 105
> a1 <- a2 <- 1-choose(16,5)/choose(20,5)
> a1
[1] 0.7182663
> a12 <- 1 - (choose(16,5)+choose(16,5)-choose(14,5))/choose(20,5)
> a12
[1] 0.5656605
> mu.hat.2 <- (y1/a1+y2/a2)/(M*N)
> SE.mu.hat.2 <- (1/(M*N))*sqrt((y1^2/a1)*(1/a1-1)+
+ (y2^2/a2)*(1/a2-1)+2*(y1*y2/a12)*(a12/(a1*a2)-1))
> c(mu.hat.2,SE.mu.hat.2)
[1] 0.7344073 0.3157506
```

- Note that $\alpha_{kk} = \alpha_k$ for $k = 1, 2$ since $x_{kk} = x_k$. Also note that the $k = 1, j = 2$ and $k = 2, j = 1$ terms in the double summation are the same; that's why there's a coefficient of 2 in the R code for that term.

- The estimated variance of the Horvitz-Thompson estimator is less than that of the separate transects estimator. Thompson's simulation results in Table 25.2 (p. 350; 2nd ed: p. 320) show that the Horvitz-Thompson estimator has smaller variance than the separate transects estimator for this problem. This is analogous to what was found for adaptive cluster sampling (Chap. 24) where $\widehat{\mu}_2$ had smaller variance than $\widehat{\mu}_1$. This certainly suggests that the Horvitz-Thompson estimator is preferred, although the calculations are more involved.