Introduction

Survey Sampling
Statistics 4234/5234
Fall 2018

September 4, 2018

Sample Surveys

Example: Ann Landers (1976) asked readers of her column to respond to the question: "If you had it to do over again, would you have children?" About 70% of the readers who responded said "No." She received over 10,000 responses, 80% of those from women.

Exercise: Describe the target population, sampling frame, sampling unit, and observation unit; and discuss any possible sources of selection bias or inaccuracy of responses.

Requirements of a Good Sample

A perfect sample would be a "scaled-down" version of the population, mirroring every characteristic of the whole population.

A good sample will be **representative** in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.

Some definitions:

- **Observation unit** An object with an associated numerical or categorical value of interest; the basic unit of observation.

  In studying human populations, observation units are often individuals.

- **Target population** The complete collection of units we want to study.

  Ideally we would know the associated value for every unit in the target population.

- **Sample** A subset of a population; the set of units for which we do observe a measurement of that value.

- **Sampled population** The collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.

- **Sampling unit** A unit that can be selected for a sample.

- **Sampling frame** A list, map, or other specification of sampling units in the population from which a sample may be selected.

Ideally the sampled population will be identical to the target population, but this is rarely the case.

Insert Figure 1.1 here

Selection Bias

A good sample will be as free from selection bias as possible.

Selection bias occurs when some part of the target population is not in the sampled population, or more generally, when some population units are sampled at a different rate than intended by the investigator.

Examples of selection bias:

- Sampling units that are easiest to select or most likely to respond, called **convenience sampling**

- Purposively selecting a "representative" sample, called **judgment sampling**

- Failure to include all of the target population in the sampling frame, called **undercoverage**

- Including population units in the sampling frame that are not in the target population, called **overcoverage**

- Failure to obtain responses from all of the chosen sample, called **nonresponse**

Measurement Error

When a response in the survey differs from the true value, **measurement error** has occurred.

Measurement bias occurs when the response has a tendency to differ from the true value in one direction.

Questionnaire Design

The most important step in writing a questionnaire is to decide what you want to find out. Write down the goals of your survey, and be precise.

- *Always test your questions before taking the survey.*
- *Keep it simple and clear.*
- *Use specific questions instead of general ones, if possible.*
- *Relate your questions to the concept of interest.*
- *Decide whether to use open or closed questions.*
- *Report the actual questions asked.*

- *Avoid questions that prompt or motivate the respondent to say what you would like to hear.*

- *Consider the social desirability of responses to questions, and write questions that elicit honest responses.*

- *Avoid double negatives.*

- *Use forced-choice, rather than agree/disagree questions.*

- *Ask only one concept per question.*

- *Pay attention to question order effects.*

Sampling and Nonsampling errors

Sampling error is the error that results from taking one sample instead of examining the whole population.

Sampling errors are usually reported in probabilistic terms.

Nonsampling error refers to any errors that cannot be attributed to sample-to-sample variability.

Selection bias and measurement error are examples of nonsampling error.

Why sample at all?

There are three main justifications for using sampling:

- Sampling can provide reliable information at far less cost than a census.

- Data can be collected more quickly, so estimates can be published in a timely fashion.

- Estimates based on sample surveys are often more accurate than those based on a census, because investigators can be more careful when collecting data.

Probability Review

Survey Sampling
Statistics 4234/5234
Fall 2018

September 6, 2018

Example 1: Flip a coin 3 times, the sample space is

$$\Omega = \{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}$$

Definition: The *sample space* for a random experiment is the set of possible outcomes.

Assume Ω is finite, $\Omega = \{\omega_1, \ldots, \omega_k\}$.

Associated with each outcome ω_i is a probability p_i satisfying

$$p_i \geq 0 \text{ for } i = 1, \ldots, k \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1$$

Definition: A collection of outcomes (any subset of the sample space) is called an *event*. The probability of an event is the sum of the probabilities of the outcomes that make up that event.

Example 1: $p_i = \frac{1}{8}$ for $i = 1, \ldots, 8$. Define the event A to be "exactly two heads" so $A = \{HHT,\ HTH,\ THH\}$ and $P(A) = \frac{3}{8}$.

Simple random sampling with replacement (SRSwR)

N balls in urn, sample one ball n times, replacing the ball in the urn between each draw.

There are N^n posible ordered samples (permutations), each equally likely.

Example 2: Population of $N = 5$ units, sample $n = 2$ with replacement. The probability that unit 5 is in the sample is

$$P\left(\{15, 25, 35, 45, 51, 52, 53, 54, 55\}\right) = \frac{9}{25} = 0.36$$

Simple random sampling without replacement (SRS)

N balls in urn, draw n of them at random.

There are

$$N \times (N-1) \times \cdots \times (N-n+1) = \frac{N!}{(N-n)!}$$

possible ordered samples (permutations).

Ignoring the order we have

$$\frac{N!}{n!(N-n)!} = \left(\begin{array}{c} N \\ n \end{array} \right)$$

possible samples, equally likely.

Example 2: With $N = 5$ and $n = 2$ there are $\binom{5}{2} = 10$ possible samples. The probability that unit 5 is in our sample is

$$P(\{15, 25, 35, 45\}) = \frac{4}{10} = 0.40$$

Example 3: An urn has 5 black balls and 3 red ones, we will draw 4 at random. Then

$$P(\text{no red}) = \frac{5 \times 4 \times 3 \times 2}{8 \times 7 \times 6 \times 5} = \frac{\binom{5}{4}\binom{3}{0}}{\binom{8}{4}} = \frac{1}{14}$$

and

$$P(\text{one red}) = \frac{\binom{5}{3}\binom{3}{1}}{\binom{8}{4}} = \frac{3}{7} \quad \text{and} \quad P(\text{two reds}) = \frac{\binom{5}{2}\binom{3}{2}}{\binom{8}{4}} = \frac{3}{7}$$

Random variables

A *random variable* assigns a numeric value to each outcome,

$$X : \Omega \to \mathbb{R}$$

The set of possible values and their probabilities is the *probability distribution* of the random variable.

Example 3: Urn contains 5 black and 3 red balls, pick 4 at random, let $X =$ the number of reds.

x	0	1	2	3
$P(X = x)$	1/14	6/14	6/14	1/14

Definition: The *expected value* of the random variable X is

$$E(X) = \sum_x xP(X = x)$$

Example 3: $E(X) = 1.5$

Proposition: For any random variable X and function $g(\cdot)$, the mean of the random variable $g(X)$ is

$$E[g(X)] = \sum_x g(x)P(X = x)$$

Definition: The *variance* of the random variable X is

$$V(X) = E\left[(X - EX)^2\right]$$

Example 3: $V(X) = 0.5357$

Proposition: An alternative expression for the variance is

$$V(X) = E(X^2) - (EX)^2$$

Joint distributions

Example 4: Let the random variables (X, Y) have the joint probability distribution given by the following table of $P(X = x, Y = y)$.

	y		
x	1	2	3
1	1/6	1/6	1/6
2	1/12	0	1/12
3	0	1/3	0

Proposition: Given the pair of random variables (X, Y), and a function $g(\cdot, \cdot)$, the expected value of the random variable $g(X, Y)$ is

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P(X = x, Y = y)$$

Example 4: Find $E(X)$ and $E(Y)$ and $E(XY)$.

OK.

$$E(X) = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{3}\right) = \frac{11}{6}$$

and $E(Y) = 2$, by inspection, and

$$E(XY) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{6}\right) + 6\left(\frac{5}{12}\right) = \frac{11}{3}$$

Definition: The *covariance* between the random variables X and Y is

$$\text{Cov}(X, Y) = E\left[(X - EX)(Y - EY)\right]$$

and the *correlation* is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X) \cdot V(Y)}}$$

Proposition: $\text{Cov}(X, Y) = E(XY) - (EX)(EY)$.

Definition: The random variables X and Y are *independent* if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all x and y.

Proposition: If X and Y are independent then $\text{Cov}(X, Y) = 0$.

The converse of this propostion is not true (see Example 4).

Conditional probability

Definition: The *conditional probability* of B given A, where $P(A) > 0$, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Proposition: If the events $A_1, \ldots, A_k$ form a partition of the sample space, that is if $A_1 \cup \cdots \cup A_k = \Omega$ and $A_i \cap A_j = \varnothing$ for all $i \neq j$, then

$$P(B) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

Example 5: Suppose we have three urns, $U_1 = \{4, 6, 7, 9\}$ and $U_2 = \{6, 8\}$ and $U_3 = \{5\}$. We first randomly select an urn, then randomly select a number from that urn.

Then

$$P(4) = P(U_1)P(4|U_1) + 0 + 0 = \left(\frac{1}{3}\right)\left(\frac{1}{4}\right) = \frac{1}{12}$$

and

$$P(5) = 0 + 0 + P(U_3)P(5|U_3) = \frac{1}{3}(1) = \frac{1}{3}$$

and

$$\begin{aligned} P(6) &= P(U_1)P(6|U_1) + P(U_2)P(6|U_2) \\ &= \left(\frac{1}{3}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{4} \end{aligned}$$

Conditional expectation

Given discrete random variables X and Y, and a number x with $P(X = x) > 0$, the *conditional distribution* of Y given $X = x$ is defined by

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

The *conditional expected value* of Y given $X = x$ is

$$E(Y|X = x) = \sum_y yP(Y = y|X = x)$$

and the *conditional variance* of Y given $X = x$ is

$$V(Y|X = x) = E\left\{[Y - E(Y|X = x)]^2 |X = x\right\}$$

For each value of x with $P(X = x) > 0$ we can define the functions

$$g(x) = E(Y|X = x) \quad \text{and} \quad h(x) = V(Y|X = x)$$

and thus have defined the random variables

$$g(X) = E(Y|X) \quad \text{and} \quad h(X) = V(Y|X)$$

which have some interesting properties:

1. $E(Y) = E[E(Y|X)]$

2. $V(Y) = V[E(Y|X)] + E[V(Y|X)]$

Example 5: Define the random variables X and Y so that X indicates the urn selected (1, 2 or 3), and Y equals the number drawn from that urn. It is straightforward to show that Y is distributed as

y	4	5	6	7	8	9
$P(Y = y)$	1/12	4/12	3/12	1/12	2/12	1/12

and compute

$$E(Y) = 6.17 \quad \text{and} \quad V(Y) = 2.14$$

The conditional distributions $P(Y = y|X = x)$ for $x = 1, 2, 3$ are given in the following table, from which are easily computed the conditional means and variances $E(Y|X = x)$ and $V(Y|X = x)$:

x	4	5	6	7	8	9	E	V
1	1/4	0	1/4	1/4	0	1/4	6.50	3.25
2	0	0	1/2	0	1/2	0	7.00	1.00
3	0	1	0	0	0	0	1.00	0.00

Then we have

$$E[E(Y|X)] = \sum_{x=1}^{3} E(Y|X = x)P(X = x)$$
$$= \frac{1}{3}(6.5 + 7.0 + 5.0) = 6.17$$

confirming Property 1 above.

Also

$$E[V(Y|X)] = \sum_{x=1}^{3} V(Y|X=x)P(X=x)$$
$$= \frac{1}{3}(3.25 + 1.00 + 0.00) = 1.4167$$

and

$$V[E(Y|X)] = \sum_{x=1}^{3} [E(Y|X=x)]^2 P(X=x) - (EY)^2$$
$$= \frac{1}{3}(6.5^2 + 7.0^2 + 5.0^2) - 6.17^2 = 0.7222$$

and thus

$$V[E(Y|X)] + E[V(Y|X)] = 2.14$$

confirming Property 2.

Probability Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 11, 2018

Here (and throughout most of the course) we will assume: the sampling frame population is the target population (no under-coverage or overcoverage); no nonresponse or missing data; and no measurement error.

For most of the course we assume there is no *nonsampling error*, so that we can focus our attention on the study of *sampling error*.

In a **probability sample** from a population of N units, each of the 2^N possible samples has a known selection probability, and thus each unit in the population has a known inclusion probability.

Types of probability samples

1. **Simple random sample** without replacement (SRS).

   Imagine an urn containing N balls, well mixed. Select n of them at random.

   Let

$$Z_i = \begin{cases} 1 & \text{unit } i \text{ in sample} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, N$$

   The probability of any particular sample, that is, any collection $(z_1, \ldots, z_N)$ of 0's and 1's, is

$$p(z_1, \ldots, z_N) = \begin{cases} \binom{N}{n}^{-1} & \sum_{i=1}^{N} z_i = n \\ 0 & \text{otherwise} \end{cases}$$

2. **Stratified random sample**

   The population is divided into subgroups, called *strata*. A separate, independent SRS is taken within each *stratum*.

3. **Cluster sampling**

   The population is divided into subgroups, called *clusters*. Then a SRS of clusters is drawn.

   - In *one-stage cluster sampling*, we take a complete census in the selected clusters;
   - in *two-stage cluster sampling* we take separate SRSs in the selected clusters.

4. **Systematic sampling**

   Given a list of the units, choose a starting point at random, then sample that unit and every kth unit after it.

Example 1: Suppose the population is $\{1, 2, \ldots, 100\}$

- Suppose 10 strata: $\{1, \ldots, 10\}, \{11, \ldots, 20\} \ldots, \{91, \ldots, 100\}$.

- Suppose 20 clusters: $\{1, \ldots, 5\}, \{6, \ldots, 10\} \ldots, \{96, \ldots, 100\}$.

1. In a SRS of $n = 20$, any of the $\binom{100}{20}$ possible samples has the same probability of being the sample selected.

2. Stratified sample: Pick 2 units from stratum.

   There are $\binom{10}{2}^{10}$ possible samples, all equally likely.

3. Cluster sample: Take a SRS of 4 of the 20 clusters.

   There are $\binom{20}{4}$ possible samples, all equally likely.

4. Systematic sample: Pick one of $\{1, 2, 3, 4, 5\}$ at random, sample that unit and every 5th unit thereafter.

   There are 5 possible samples, equally likely.

Framework for probability sampling

The population is $\mathcal{U} = \{1, 2, \ldots, N\}$.

Consider a probability sampling method with m different possible samples, denote

$$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$$

Each $\mathcal{S}_j$ is a subset of $\mathcal{U}$, and

$$P(\mathcal{S}_1) + \cdots + P(\mathcal{S}_m) = 1$$

For each unit i let

$$\pi_i = P(\text{unit } i \text{ in sample}) = \sum_{j: i \in \mathcal{S}_j} P(\mathcal{S}_j)$$

Suppose that associated with the ith unit of the population is a numeric value y_i.

Sometimes we will write that the population is $\{y_1, y_2, \ldots, y_N\}$.

Suppose we want to estimate the population mean

$$\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i$$

using the sample mean

$$\bar{y}_{\mathcal{S}} = \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} y_i$$

Definition: The **sampling distribution** of the statistic $\bar{y}$ is specified by the possible values $\bar{y}$ can take, and the probabilities it takes those values.

For a finite population this is a discrete distribution!

$$P(\bar{y} = k) = \sum_{\mathcal{S}:\bar{y}_{\mathcal{S}}=k} P(\mathcal{S})$$

Definition: The **expected value** of $\bar{y}$ is the mean of this sampling distribution,

$$E(\bar{y}) = \sum_{\mathcal{S}} \bar{y}_{\mathcal{S}} P(\mathcal{S}) = \sum_{k} k P(\bar{y} = k)$$

Example 2: Consider the following population of $N = 6$ units

i	1	2	3	4	5	6
y_i	9	12	14	15	12	12

The population mean is $\bar{y}_U = 12.33$.

Consider a sampling scheme of three possible samples of size $n = 4$, with $\mathcal{S}_1 = \{1, 2, 3, 4\}$ and $\mathcal{S}_2 = \{2, 3, 4, 5\}$ and $\mathcal{S}_3 = \{3, 4, 5, 6\}$, and suppose that $P(\mathcal{S}_j) = 1/3$ for $j = 1, 2, 3$.

The inclusion probabilities are

$$\pi_1 = \pi_6 = \frac{1}{3} \quad \text{and} \quad \pi_2 = \pi_5 = \frac{2}{3} \quad \text{and} \quad \pi_3 = \pi_4 = 1$$

The possible values of the sample mean are

$$\bar{y}_{\mathcal{S}_1} = 12.50 \quad \text{and} \quad \bar{y}_{\mathcal{S}_2} = \bar{y}_{\mathcal{S}_3} = 13.25$$

The expected value of the sample mean is

$$E(\bar{y}) = 12.50\left(\tfrac{1}{3}\right) + 13.25\left(\tfrac{2}{3}\right) = 13.00$$

Definition: The **estimation bias** of $\bar{y}$ as an estimator of $\bar{y}_U$ is

$$\text{Bias}(\bar{y}) = E(\bar{y}) - \bar{y}_U$$

Example 2: $\text{Bias}(\bar{y}) = 0.67$.

Definition: The **variance** of the estimator $\bar{y}$ is

$$V(\bar{y}) = E\left\{[\bar{y} - E(\bar{y})]^2\right\} = \sum_{\mathcal{S}} [\bar{y}_{\mathcal{S}} - E(\bar{y})]^2 \, P(\mathcal{S})$$

and the **mean squared error** is

$$\text{MSE}(\bar{y}) = E\left[(\bar{y} - \bar{y}_U)^2\right] = \sum_{\mathcal{S}} [\bar{y}_{\mathcal{S}} - \bar{y}_U)]^2 \, P(\mathcal{S})$$

Note that

$$\begin{aligned}\text{MSE}(\bar{y}) &= E\left[(\bar{y}-\bar{y}_U)^2\right] \\ &= E\left\{[\bar{y}-E(\bar{y})+E(\bar{y})-\bar{y}_U]^2\right\} \\ &= E\left\{[\bar{y}-E(\bar{y})]^2\right\}+[E(\bar{y})-\bar{y}_U]^2 \\ &= V(\bar{y})+[\text{Bias}(\bar{y})]^2\end{aligned}$$

The cross-product term is zero since $E[\bar{y}-\bar{y}_U]=0$

Example 2:

$$V(\bar{y}) = \tfrac{1}{3}(12.50-13.00)^2+\tfrac{2}{3}(13.25-13.00)^2 = 0.125$$

and

$$\begin{aligned}\text{MSE}(\bar{y}) &= \tfrac{1}{3}(12.50-12.33)^2+\tfrac{2}{3}(13.25-12.33)^2 \\ &= 0.5694 \\ &= 0.125+(0.67)^2\end{aligned}$$

Simple Random Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 13, 2018

Consider a population that has N units. We will take a random sample consisting of n draws from this population.

In **simple random sampling with replacement** (SRSwR) we take n independent samples of size 1.

In **simple random sampling without replacement** (SRS), there are $\binom{N}{n}$ possible samples, and each is equally likely. Thus the probability of any particular set of n units is

$$P(\mathcal{S}) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

Under SRS the probability that the ith unit is included in the sample is

$$\pi_i = \frac{\text{number of samples including unit } i}{\text{total number of possible samples}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Denote the population values by $\{y_1, y_2, \ldots, y_N\}$.

The population mean is

$$\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i$$

and the population variance is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2$$

and the population standard deviation is $S = \sqrt{S^2}$.

Suppose we take a simple random sample of size n and compute the sample mean

$$\bar{y}_{\mathcal{S}} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$$

Proposition: $E(\bar{y}) = \bar{y}_U$ and $V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$.

Proof: Define the random variables

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

for $i = 1, \ldots, N$.

Then

$$E(Z_i) = E(Z_i^2) = P(Z_i = 1) = \pi_i$$

and

$$V(Z_i) = E(Z_i^2) - (EZ_i)^2 = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i)$$

and thus

$$E(Z_i) = \frac{n}{N} \quad \text{and} \quad V(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

Also, for $j \neq i$ we have

$$E(Z_i Z_j) = P(Z_i Z_j = 1) = P(\text{both } i \text{ and } j \text{ in sample})$$

$$= \frac{\text{number samples include units } i \text{ and } j \text{ both}}{\text{total number of possible samples}}$$

$$= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

and thus

$$\begin{aligned}
\text{Cov}(Z_i, Z_j) &= E(Z_i Z_j) - (EZ_i)(EZ_j) \\
&= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\
&= -\frac{1}{N-1}\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)
\end{aligned}$$

Thus we have

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i\in\mathcal{S}} y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{n}\frac{n}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$$

and

$$
\begin{aligned}
V(\bar{y}) &= V\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + 2\sum_{i=1}^{N-1}\sum_{j=1+1}^{N} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right] \\
&= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{2}{N-1}\sum_{i=1}^{N-1}\sum_{j=1+1}^{N} y_i y_j\right] \\
&= \frac{1}{nN}\left(1-\frac{n}{N}\right)\left(\frac{1}{N-1}\right) N \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 \quad \text{see Sec 2.8} \\
&= \frac{S^2}{n}\left(1-\frac{n}{N}\right)
\end{aligned}
$$

The factor $\left(1 - \frac{n}{N}\right)$ is called the **finite population correction**.

- If $n = N$ then $V(\bar{y}) = 0$.

- If $n = 1$ then $V(\bar{y}) = \frac{S^2}{1}\left(1 - \frac{1}{n}\right) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$

Of course in practice the population variance is unknown, and thus estimated by the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2$$

The estimated variance of $\bar{y}$ is

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

and the standard error is

$$\text{SE}(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

Proposition: $E(s^2) = S^2$

Proof: (See Section 2.8)

$$
\begin{aligned}
E\left[\sum_{i\in\mathcal{S}}(y_i-\bar{y})^2\right] &= E\left\{\sum_{i\in\mathcal{S}}[(y_i-\bar{y}_U)-(\bar{y}-\bar{y}_U)]^2\right\} \\
&= E\left[\sum_{i=1}^{N} Z_i(y_i-\bar{y}_U)^2\right] - nE\left[(\bar{y}-\bar{y}_U)^2\right] \\
&= \frac{n}{N}(N-1)\sum_{i=1}^{N}(y_i-\bar{y}_U)^2 - n\frac{S^2}{n}\left(1-\frac{n}{N}\right) \\
&= \frac{n}{N}(N-1)S^2 - S^2\left(\frac{N-n}{N}\right) = (n-1)S^2
\end{aligned}
$$

Two final points about estimation under SRS

1. Want to estimate $t = \sum_{i=1}^{N} y_i = N\bar{y}_U$?

   Use the estimator $\widehat{t} = N\bar{y}$. Follows immediately from the work above that

$$E(\widehat{t}) = NE(\bar{y}) = N\bar{y}_U = t$$

   and

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

   and so the standard error of the unbiased estimator $\widehat{t}$ is

$$\mathrm{SE}(\widehat{t}) = N\,\mathrm{SE}(\bar{y}) = N\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

2. If we want to estimate the proportion of a population which possess some trait of interest, we let

$$y_i = \begin{cases} 1 & \text{unit } i \text{ has that trait} \\ 0 & \text{otherwise} \end{cases}$$

and proceed as above.

In this situation we will often employ specialized notation: the population proportion is commonly denoted by $\bar{y}_U = p$, and the population variance reduces to

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{N}{N-1}p(1-p)$$

We estimate the population mean (proportion) by the sample mean (proportion) $\bar{y} = \widehat{p}$, and find

$$V(\widehat{p}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

The sample variance reduces to

$$s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \widehat{p})^2 = \frac{n}{n-1}\widehat{p}(1-\widehat{p})$$

Then

$$\widehat{V}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n-1}\left(1 - \frac{n}{N}\right) \quad \text{and} \quad \mathrm{SE}(\widehat{p}) = \sqrt{\widehat{V}(\widehat{p})}$$

Sampling weights (Sec 2.4)

Define the **sampling weight** of unit i to be the reciprocal of the inclusion probability

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

We interpret w_i as the number of population units represented by unit i.

In SRS $w_i = 1/\pi_i = N/n$ for each i. Thus each unit in the sample represents N/n units, itself plus $N/n - 1$ of the unsampled units.

Definition: A sampling design for which every unit has the same sampling weight is called a *self-weighting* sample.

Thus SRS is a self-weighting method.

Also for SRS, we can write our estimates of t and $\bar{y}_U$ as

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

and

$$\bar{y} = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i}$$

Sampling weights will become useful later in the course, when we consider sampling schemes with unequal selection probabilities.

Confidence Intervals

Survey Sampling
Statistics 4234/5234
Fall 2018

September 18, 2018

Confidence intervals (Sec 2.5)

Consider the population $\mathcal{U} = \{1, 2, \ldots, N\}$ with numerical values $\{y_1, y_2, \ldots, y_N\}$. As usual we denote the population mean and variance by

$$\bar{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{and} \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$$

We will take a random sample according to the sampling scheme $P(\mathcal{S})$, then estimate the population parameter θ by the point estimate $\widehat{\theta}_{\mathcal{S}}$, and the interval estimate $\text{CI}(\mathcal{S})$.

Definition: The interval estimator CI is a $100(1-\alpha)\%$ **confidence interval for** θ if

$$P(\theta \in \text{CI}) = \sum_{\mathcal{S}} P(\mathcal{S})\, I_{\{\theta \in \text{CI}(\mathcal{S})\}} \geq 1 - \alpha$$

Example: Consider the population of $N = 5$ units, with numerical values $\{20, 4, 10, 2, 12\}$. Suppose we wish to estimate the parameter $\theta = \bar{y}_U = 9.6$ using SRS of size $n = 2$ and

$$\mathsf{CI}(\mathcal{S}) \quad \Leftarrow \quad \bar{y}_{\mathcal{S}} \pm 2\mathsf{SE}_{\mathcal{S}}(\bar{y}_{\mathcal{S}})$$

where

$$\mathsf{SE}(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

Is this a 95% confidence interval? 85%? 75%?

Consider all 10 possible samples; compute the sample mean and standard deviation, and resulting confidence interval, for each.

You will find that the interval covers the value 9.6 for 8 of the 10 possible samples. Thus the interval defined above gives a $100(1-\alpha)\%$ confidence interval for and $\alpha \geq .20$.

In sampling from infinite populations (i.e., in every statistics course other than this one), we learned that if the sample size is sufficiently large, for many situations (for example maximum likelihood estimation) we have

$$\widehat{\theta} \dot{\sim} \text{Normal}\left[\theta,\ V\left(\widehat{\theta}\right)\right]$$

and thus

$$\widehat{\theta} \pm z_{\alpha/2}\text{SE}\left(\widehat{\theta}\right)$$

gives an *approximate* $100(1-\alpha)\%$ confidence interval for θ.

There exists a simliar result for inference about the population mean (or total) in finite populations.

Propostion: If N and n and $N-n$ are sufficiently large, under simple random sampling,

$$\frac{\bar{y}-\bar{y}_U}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \dot{\sim} \text{Normal}\,(0,\ 1)$$

and thus an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is (under SRS) given by

$$\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

i.e.,

$$\bar{y} \pm z_{\alpha/2}\text{SE}\,(\bar{y})$$

A useful tool for assessing the appropriateness of the normal approximation in a particular problem is the *bootstrap*: Take repeated samples *with replacement* from your original sample; the distribution of the resulting sample means approximates the sampling distribution of $\bar{y}$.

Sample size calculations (Sec 2.6)

Suppose we want our estimate $\bar{y}$, based on a SRS of size n to be within e of the population mean $\bar{y}_U$, with probability at least $1-\alpha$, where e and α are specified. How large must n be?

Well, we require that

$$P\left(|\bar{y} - \bar{y}_U| \leq e\right) \geq 1 - \alpha$$

and thus

$$P\left(\left|\frac{\bar{y}-\bar{y}_U}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\right| \leq \frac{e}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\right) \geq 1-\alpha$$

and thus

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

We require that

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

We can solve this in two stages.

First let

$$n_0 = \frac{n}{1 - \frac{n}{N}}$$

and solve

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

Then

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

gives the necessary sample size for the desired precision and confidence.

Effectively, n_0 is the required sample size ignoring the fpc, and thus recommends the sample size required under simple random sampling with replacement. Since without replacement is more efficient, it will always be the case that $n \leq n_0$.

Example: The population size is $N = 500$; based on a pilot study, we estimate the population SD is about 0.85.

We want to estimate the population mean to within ± 0.30 with probability at least .95.

Then $e = 0.30$ and $\alpha = .05$, so $z_{\alpha/2} = 1.96$; it is generally a good idea to build some conservatism into the standard deviation, we'll take $S = 1.20$ (no particular reason for this value, just what I choose).

Take

$$n_0 = \left(\frac{1.96 \times 1.20}{0.30}\right)^2 = 61.47$$

and

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{61.47}{1 + 61.47/500} = 54.74$$

so take a SRS of size $n = 55$.

In the case of estimating a population proportion, we have

$$S^2 = \frac{N}{N-1} p(1-p) \approx p(1-p) \leq 0.25$$

In some situations this may be overly conservative. If, for example, we are know that the population proportion is at least 0.65 and at most 0.85, we'd use

$$S^2 = (.65)(.35) = 0.2275$$

instead. Why?

Stratified Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 20, 2018

What is stratified sampling?

If the variable we are interested in takes on different mean values in different subpopulations, we may be able to obtain more precise estimates of population quantities by taking a **stratified** random sample.

We divide the population into H subpopulations, called **strata**. The strata do not overlap, and they constitute the whole population so that each sampling unit belongs to exactly one **stratum**.

We draw an independent probability sample from each stratum, then pool the information to obtain overall population estimates.

We use stratified sampling for one or more of the following reasons:

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

   Example: Population of size $N = 2000$ consists of 1000 male and 1000 female students. The gender mix in a SRS of size $n = 100$ is likely to be close to 50-50, but there's about a 5% chance the split is 60-40 or worse.

```
Pop <- c(rep(1,1000), rep(0,1000))
male <- rep(NA, 1e5)
for(j in 1:1e5){ male[j] <- sum(sample(Pop, 100)) }
mean(male <= 40 | male >= 60)
```

   Eliminate this possibility by taking independent SRSs of 50 males and 50 females.

2. We may want data of known precision for subgroups of the population; these subgroups should be the strata.

   Example: Population of size $N = 2000$ consists of 1800 male and 200 female graduates. If a quantity of interest is the difference in average salary, we should sample a higher fraction of female graduates than male graduates to obtain comparable precision for the two groups.

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

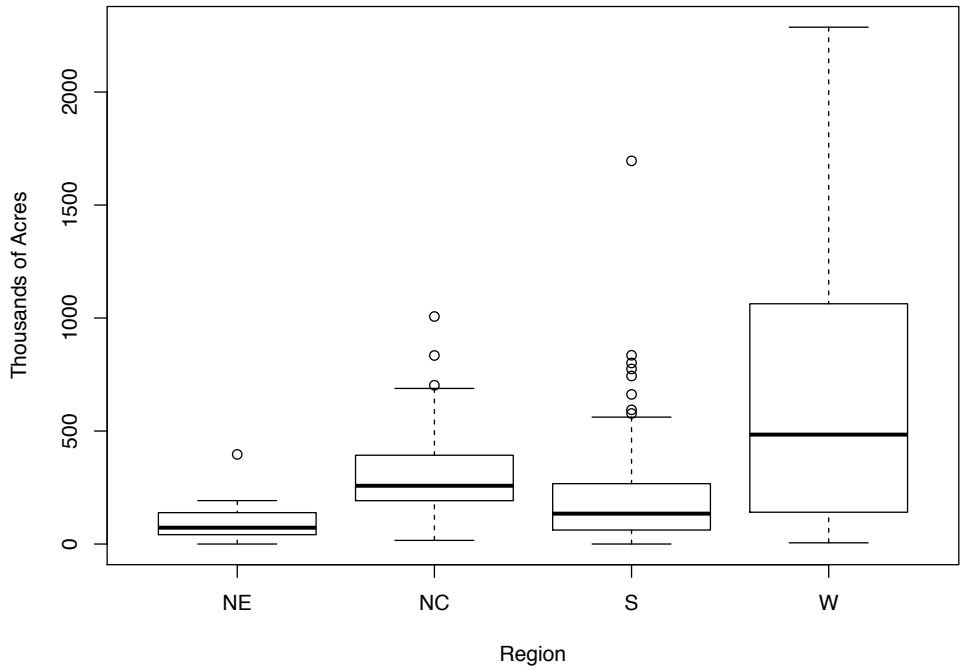
4. Stratified sampling often gives more precise (having lower variance) estimates for population means and totals.

Stratification works for lowering the variance because the variance within each stratum is often lower than the variance in the whole population.

Example: We wish to estimate the total acreage devoted to farms in the United States in 1992. There are $N = 3078$ counties and county-equivalents in the country; we obtain the number of acres devoted to farms in $n = 300$ counties.

Use the four census regions — Northeast, North Central, South, and West — as strata. The data for a stratified random sample using proportional allocation are:

Region	Counties	Sample size	Average	Std Dev
Northeast	220	21	97,630	87,450
North Central	1054	103	300,503	172,099
South	1382	135	211,315	231,490
West	422	41	662,296	629,433



Stratified random sampling

We divide the population of N sampling units into H strata, with N_h sampling units in stratum h; for stratified sampling to work we must known the values of $N_1, N_2, \ldots, N_H$ and must have

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

where N is the total number of units in the entire population.

In **stratified random sampling** we independently take an SRS from each stratum, so that n_h observations are randomly selected from the N_h population units in stratum h; the total sample size is $n = n_1 + n_2 + \cdots + n_H$.

Notation for stratification: population quantities

y_{hj} = value of jth unit in stratum h

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

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

$$\bar{y}_{hU} = \frac{1}{N_H} \sum_{j=1}^{N_h} y_{hj} = \text{population mean in stratum } h$$

$$\bar{y}_U = \frac{t}{N} = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \text{overall population mean}$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(y_{hj} - \bar{y}_{hU} \right)^2 = \text{population variance in stratum } h$$

Notation for stratification: sample quantities

Define $\mathcal{S}_h$ to be the set of n_h units in the SRS for stratum h.

$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$$

$$\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = N_h \bar{y}_h$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{j \in \mathcal{S}_h} \left(y_{hj} - \bar{y}_h \right)^2$$

Estimation in stratified random sampling

We estimate the population total $t = \sum_{h=1}^{H} t_h$ by

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

and the population mean $\bar{y}_U$ by

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

Properties of estimators

- **Unbiasedness** We have $E(\bar{y}_h) = \bar{y}_{hU}$ in each stratum and thus

$$E\left(\bar{y}_{\text{strat}}\right) = E\left(\frac{N_h}{N}\bar{y}_h\right) = \sum_{h=1}^{H}\frac{N_h}{N}E(\bar{y}_h) = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_{hU} = \bar{y}_U$$

- **Variance of the estimators**

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

and

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

- **Standard errors for stratified samples** We obtain unbiased estimators of the variances by substituting sample estimators s_h^2 for the population parameters S_h^2:

$$\widehat{V}\left(\widehat{t}_{\text{strat}}\right) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

and

$$\widehat{V}\left(\bar{y}_{\text{strat}}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n}\left(1 - \frac{n_h}{N_h}\right)$$

Then

$$\text{SE}\left(\widehat{t}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\widehat{t}_{\text{strat}}\right)} \quad \text{and} \quad \text{SE}\left(\bar{y}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\bar{y}_{\text{strat}}\right)}$$

- **Confidence intervals for stratified samples** If the sample sizes within each stratum are sufficiently large (or the sampling design has a very large number of strata), and an approximate $100(1-\alpha)\%$ confidence interval for the population mean $\bar{y}_U$ is

$$\bar{y}_{\text{strat}} \pm z_{\alpha/2}\text{SE}\left(\bar{y}_{\text{strat}}\right)$$

Example: For the farms data (working with thousands of acres) we have

$$\begin{aligned}\widehat{t}_{\text{strat}} &= 220(98) + 1054(301) + 1382(211) + 422(662) \\ &= 909{,}736\end{aligned}$$

and the estimated variance is

$$\begin{aligned}\widehat{V}\left(\widehat{t}_{\text{strat}}\right) &= 220^2\frac{87.45^2}{21}\left(1-\frac{21}{220}\right)+1054^2\frac{172.1^2}{103}\left(1-\frac{103}{1054}\right)\\ &\quad+1382^2\frac{231.5^2}{135}\left(1-\frac{135}{1382}\right)+422^2\frac{629.4^2}{41}\left(1-\frac{41}{422}\right)\\ &= 50{,}417^2\end{aligned}$$

Converting now to millions of acres we obtain

$$909.736 \pm 1.96\,(50.41725) \;\Rightarrow\; [810.9,\ 1008.6]$$

and we are 95% confident that in 1992 there were somewhere between 811 and 1009 million acres of farmland in the United States.

Stratified sampling for proportions

To make inference about a population proportion based on stratified random sampling, proceed as above with

$$\widehat{y}_h = \widehat{p}_h \quad \text{and} \quad s_h^2 = \frac{n_h}{n_h - 1}\widehat{p}_h(1 - \widehat{p}_h)$$

Then

$$\widehat{p}_{\text{strat}} = \sum_{h=1}^{H} \frac{N_h}{N}\widehat{p}_h$$

and

$$\widehat{V}\left(\widehat{p}_{\text{strat}}\right) = \sum_{h=1}^{H} \left(\frac{N_H}{N}\right)^2 \frac{\widehat{p}_h(1 - \widehat{p}_h)}{n_h - 1}\left(1 - \frac{n_h}{N_h}\right)$$

and

$$\text{SE}\left(\widehat{p}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\widehat{p}_{\text{strat}}\right)}$$

Estimate the total number of population units having a specified characteristic by

$$\widehat{t}_{\mathsf{strat}} = \sum_{h=1}^{H} N_h \widehat{p}_h$$

Then

$$\widehat{V}\left(\widehat{t}_{\mathsf{strat}}\right) = N^2 \widehat{V}\left(\widehat{p}_{\mathsf{strat}}\right)$$

Of course

$$\mathsf{SE}\left(\widehat{t}_{\mathsf{strat}}\right) = \sqrt{\widehat{V}\left(\widehat{t}_{\mathsf{strat}}\right)} = N\mathsf{SE}\left(\widehat{p}_{\mathsf{strat}}\right)$$

Stratified Sampling, part 2

Survey Sampling
Statistics 4234/5234
Fall 2018

September 25, 2018

Example: Chapter 3 Exercise 7

Consider the population of $N = 807$ college faculty members, stratified into $H = 4$ academic units. Let

y_{hj} = publications by faculty member j of academic unit h

The goal is to estimate the total number of publications by the entire college faculty, and also the proportion of faculty with no publications

The data consist of a stratifed random sample, summarized here

Stratum	N_h	n_h	$\bar{y}_h$	s_h	0's
Biological Sciences	102	7	3.14	2.61	1
Physical Sciences	310	19	2.11	2.87	10
Social Sciences	217	13	1.23	2.09	9
Humanities	178	11	0.45	0.93	8

We estimate $t = t_1 + t_2 + t_3 + t_4$ by

$$\begin{aligned}\widehat{t}_{\text{strat}} &= \sum_{h=1}^{H} \widehat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h \\ &= 102(3.14) + 310(2.11) + 217(1.23) + 178(0.45) \\ &= 1321.2\end{aligned}$$

Thus

$$\bar{y}_{\text{strat}} = \frac{\widehat{t}_{\text{strat}}}{N} = \frac{1321.2}{807} = 1.64$$

For standard errors we find

$$\begin{aligned}\widehat{V}\left(\widehat{t}_{\text{strat}}\right) &= \sum_{h=1}^{H} N_h^2 \widehat{V}\left(\widehat{t}_h\right) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \\ &= 102^2 \frac{2.61^2}{7}\left(1-\frac{7}{102}\right) + \cdots + 178^2 \frac{0.45^2}{11}\left(1-\frac{11}{178}\right) \\ &= 65{,}611\end{aligned}$$

and thus

$$\text{SE}\left(\widehat{t}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\widehat{t}_{\text{strat}}\right)} = 256.15$$

Also

$$\text{SE}\left(\bar{y}_{\text{strat}}\right) = \text{SE}\left(\frac{\widehat{t}_{\text{strat}}}{N}\right) = \frac{1}{N}\text{SE}\left(\widehat{t}_{\text{strat}}\right) = \frac{256.15}{807} = 0.32$$

We estimate that this college faculty produced a total of 1.321 published works, the standard error of this estimate is 256 publications.

Equivalently, we estimate that the average publications per faculty member at this college was 1.64; the standard error of our estimate is 0.32.

Treating the data as an SRS would have given us $\bar{y} = 1.66$ and $\text{SE}(\bar{y}) = 0.33$.

We now take up the estimation of the proportion of faculty members who had no publications.

$$\begin{aligned}
\widehat{p}_{\text{strat}} &= \sum_{h=1}^{H} \frac{N_h}{N} \widehat{p}_h \\
&= \frac{1}{807}\left[102\left(\frac{1}{7}\right) + 310\left(\frac{10}{19}\right) + 217\left(\frac{9}{13}\right) + 178\left(\frac{8}{11}\right)\right] \\
&= 0.57
\end{aligned}$$

For standard error we obtain

$$\begin{aligned}\widehat{V}\left(\widehat{p}_{\text{strat}}\right) &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \\ &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{p_h(1-p_h)}{n_h-1}\left(1-\frac{n_h}{N_h}\right) \\ &= 0.0658^2\end{aligned}$$

We estimate that 57% of the faculty had no publications; the standard error of this estimate is 6.6%.

Treating the data as a SRS, we'd have gotten an estimate of 56%, with a standard error of 6.9%.

Sampling weights (sections 2.4 and 3.3)

First consider the population $\{y_1, y_2, \ldots, y_N\}$.

Recall the *inclusion probability* for the ith unit is

$$\pi_i = P(\text{unit } i \text{ included in sample})$$

Define the **sampling weight** of unit i, for a particular sampling plan, by

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

The sampling weight w_i can be interpreted as the number of population units represented by unit i (if unit i is included in the sample).

1. Special case: Simple random sampling (SRS)

   Under SRS,

$$\pi_i = \frac{n}{N} \quad \text{and} \quad w_i = \frac{N}{n}$$

   Each unit in the sample represents itself plus $N/n - 1$ of the unsampled units.

   Also, for SRS,

$$\sum_{i \in \mathcal{S}} w_i = \sum_{i \in \mathcal{S}} \frac{N}{n} = n\left(\frac{N}{n}\right) = N$$

and thus

$$\widehat{t}_{\mathsf{SRS}} = N\bar{y} = \frac{N}{n}\sum_{i\in\mathcal{S}} y_i = \sum_{i\in\mathcal{S}} w_i y_i$$

and

$$\bar{y} = \frac{\widehat{t}}{N} = \frac{\sum_{i\in\mathcal{S}} w_i y_i}{\sum_{i\in\mathcal{S}} w_i}$$

Definition: A sampling plan in which every unit has the same sampling weight is called a **self-weighting** sample.

Proposition: SRS is self-weighting.

2. Special case: stratified random sampling (section 3.3)

   Now the population is

$$\left\{y_{hj} : j = 1, \ldots, N_h; h = 1, \ldots, H\right\}$$

   Under stratified random sampling the inclusion probabilities are

$$\pi_{hj} = \frac{n_h}{N_h}$$

   and the sampling weight for unit j of stratum h is

$$w_{hj} = \frac{1}{\pi_{hj}} = \frac{N_h}{n_h}$$

Again, the sum of the sampling weights of sampled units, for any set of samples $\mathcal{S}_1, \ldots, \mathcal{S}_H$, gives the number of units in the population

$$\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} \frac{N_h}{n_h} = \sum_{h=1}^{H} N_h = N$$

And again, we find that the estimators of the population total and population mean satisfy

$$\widehat{t}_{\text{strat}} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}$$

and

$$\bar{y}_{\text{strat}} = \frac{\widehat{t}_{\text{strat}}}{N} = \frac{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj}}$$

Example: In an SRS of $n = 50$ from a population of size $N = 807$, each sampled unit represents

$$\frac{807}{50} = 16.14 \text{ units}$$

In the stratified random sample we find

Stratum	N_h	n_h	w_{hj}
Biological Sciences	102	7	14.6
Physical Sciences	310	19	16.3
Social Sciences	217	13	16.7
Humanities	178	11	16.2

Thus each sampled social science professor represents 16.7 professors, whereas each sampled biology professor represents only 14.6.

Stratified Sampling, part 3

Survey Sampling
Statistics 4234/5234
Fall 2018

September 27, 2018

Sampling: Design and Analysis, second edition; by Sharon L. Lohr

(Sections 3.4–3.5)

Survey design includes methods for controlling nonsampling as well as sampling error; today we discuss features that affect the sampling error.

Simple random sampling involves one design feature: the sample size.

For stratified random sampling, we need to determine what the strata should be, then decide how many observations to sample in each stratum.

We'll attack these questions in the reverse order.

Allocating observations to strata: proportional allocation

If you are taking a stratified sample in order to ensure that the sample reflects the population with respect to the stratification variable, and you would like your sample to be a miniature version of the population, you should use proportional allocation.

In **proportional allocation**

- the number of sampled units in each stratum is proportional to the size of the stratum, $n_h \propto N_h$;

- the inclusion probability $\pi_{hj} = n_h/N_h$ is the same $(= n/N)$ for all strata.

Example: Population of 2400 men and 1600 women, sample 10% of the population; proportional allocation would mean sampling 240 men and 160 women.

Under proportional allocation, the probability that an individual will be selected is n/N, same as for SRS, but many of the "bad" samples that could occur with SRS are no longer possible.

Example: Under SRS, each unit in the sample represents 10 people in the population. In stratified sampling with proportional allocation, each man in the sample represents 10 men in the population, and each woman represents 10 women in the population.

When the strata are large enough, the variance of $\bar{y}_{\text{strat}}$ under proportional allocation is usually less than the variance of the sample mean from an SRS with the same number of observations.

This is true no matter how silly the stratification scheme may seem.

Proposition: For estimating a population mean or total, stratification with proportional allocation will give smaller variance than SRS *unless*

$$\sum_{h=1}^{H} N_h \left(\bar{y}_{hU} - \bar{y}_U\right)^2 < \sum_{h=1}^{H} \left(1 - \frac{N_H}{N}\right) S_h^2$$

This rarely happens when the N_h are large.

In general, the variance of the estimator of t from a stratified sample with proportional allocation will be smaller than the variance of the estimator of t from an SRS with the same number of observations.

The more unequal the stratum means $\bar{y}_{hU}$, the more precision you will gain by stratifying and using proportional allocation.

If the variances S_h^2 are more or less equal across all the strata, proportional allocation is probably the best allocation for increasing precision.

In cases where the S_h^2 vary greatly, **optimal allocation** can result in smaller costs.

Allocating observations to strata: optimal allocation

In practice, when we are sampling units of different sizes, the larger units are likely to be more variable than the smaller units, and we should sample them with a higher sampling fraction.

Example (accounting): Use recorded book amount to stratify a population of loans; stratum 1 is loans over \$1 million, stratum 2 is loans between \$500,000 and \$999,999, etc; S_h^2 will be much larger in the strata with large loan amounts, so optimal allocation should prescribe a higher sampling fraction for those strata. An error in the recorded amount of a \$3 million dollar loan is much more important for the bank to know about that that of a \$3000 loan! Maybe even set $n_1 = N_1$ in stratum 1.

The objective in optimal allocation is to gain the most information for the least cost. We want to minimize

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

subject to a constraint on the total cost of the sample, given by

$$c_0 + \sum_{h=1}^{H} c_h n_h \leq C$$

where C denotes the maximum total cost allowed, c_0 represents overhead costs, and c_h is the cost of taking an observation in stratum h.

Using Lagrange multipliers, the solution is to take

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}}$$

In optimal allocation, we sample heavily within a stratum if

- the stratum accounts for a large part of the population;
- the variance within the stratum is large (sample more heavily to compensate for the heterogeneity);
- sampling in the stratum is expensive.

If all variances and costs are equal, proportional allocation is the same as optimal allocation. If we know the variances within each stratum and they differ, optimal allocation gives a smaller variance for the estimator of $\bar{y}_U$ than proportional allocation.

Neyman allocation is a special case of optimal allocation, used when the costs in the strata (but not the variances) are approximately equal; under Neyman allocation

$$n_h \propto N_h S_h$$

If the variances S_h^2 are specified correctly, Neyman allocation will give an estimator with smaller variance than proportional allocation.

When the stratum variances S_h^2 are approximately known, Neyman allocation gives higher precision than proportional allocation. If the information about the stratum variances is of poor quality, however, disproportional allocation can result in a higher variance than simple random sampling. Proportional allocation, on the other hand, almost always has smaller variance than simple random sampling.

Determining sample sizes

Summing up, we have

- optimal allocation $n_h \propto N_h S_h / \sqrt{c_h}$
- Neyman allocation $n_h \propto N_h S_h$
- proportional allocation $n_h \propto N_h$

The different methods of allocating observations to strata give the relative sample sizes n_h/n. After strata are constructed and observations allocated to strata, the total sample size required to achieve a prespecified margin of error can be determined:

Take

$$n = V \cdot \left(\frac{z_{\alpha/2}}{e}\right)^2$$

where

$$V^* = \frac{1}{N^2} \sum_{h=1}^{H} \frac{n}{n_h} N_h^2 S_h^2$$

Defining strata

Stratification is most efficient when the stratum means differ widely — ideally we would stratify by the values of y.

Although, if we had this information we would not need to do a survey at all!

Instead we try to find some variable closely related to y, and stratify on that.

The number of strata you choose depends on many factors such as the difficulty in constructing a sampling frame with stratifying information, and the cost of stratifying.

A general rule to keep in mind is: The more information, the more strata you should use. Thus, you should use an SRS when little prior information about the target population is available.

You can often collect preliminary data that can be used to stratify your design.

In a survey with more precise information, you will want to use more strata — many surveys are stratified to the point where only two sampling units are observed in each stratum.