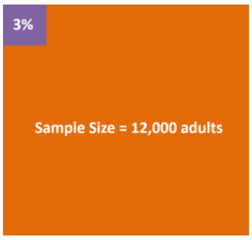
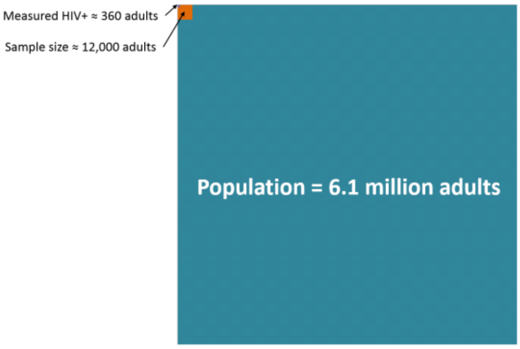
Methods: Simple Random Sampling

Topics: Introduction to Simple Random Sampling

The Rwanda 2010 Demographic and Health Survey reported that 3.0% of adults age 15-49 were HIV-positive in 2010. This was based on blood samples collected from about 12,000 people in a population of 10.5 million, of whom about 6.1 million were adults. The sample size of 12,000 is only 1/5th of 1% of the 6.1 million total population. How is it possible to draw conclusions about such a large group of people when relatively few were interviewed? This is the beauty of sampling!




Measured HIV+ ≈ 360 adults
Sample size ≈ 12,000 adults
Population = 6.1 million adults

1. Sampling Theory

Two core principles of sampling theory are (1) we have an infinite (or very large) population, and (2) each person has an equal (or known) chance of being sampled. We will come back to these ideas in this lecture and future lectures.

Why should we believe the results of samples? How do we know that the information collected in the survey accurately represents the population? And how do we know the survey is reliable? Let us review these concepts of accuracy and reproducibility.

2. Accuracy – depends on biases

One way to understand accuracy is to define its opposite: bias. There are several ways that a survey can be biased.

Accuracy: How close the measurement is to reality

Factors effecting accuracy:

- Selection bias
- Measurement bias
- Recall bias

Selection bias occurs when the types of people in the sample are not representative of the population. For example, a sample of only men would not provide information about the whole population because half of the population (women) are not represented in the sample.

Measurement error (or **measurement bias**) happens when there is faulty equipment, such as a broken blood pressure cuff that always records blood pressure higher than it actually is.

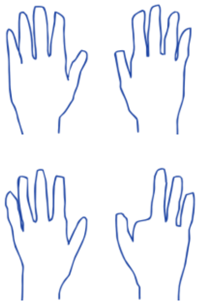
When respondents are unable or unwilling to respond correctly, we are concerned about potential **recall bias**. For example, if the survey question is about first sexual encounter, and respondents range in age from 25 to 50, we would be concerned that respondents who do not remember important details about their first sexual encounter are systematically older, and therefore respondents who provide data are not necessarily representative of that population of interest.

3. Precision – depends on variability

Precision: How reproducible is the measurement

Factors effecting precision:

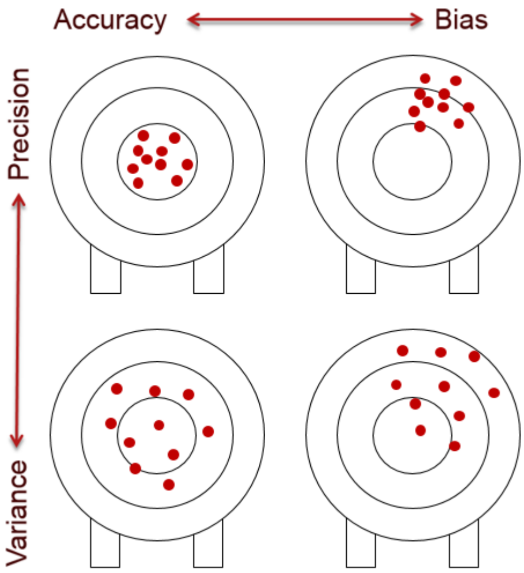
- Variability in the population
- Sample design



Precision refers to how reproducible results are, and it depends on how variable the outcome is in the population. Here is a silly example: how variable is the total number of fingers that the average person has in the population? Of course there is slight variations in the number of fingers that people have due to accidents and birth abnormalities, but generally, most people have 10 fingers, so even a very small sample of people could produce a very precise estimate of average number of fingers in the population. This is because there is so little variability in the population. We would need a much larger sample to estimate the average number of siblings per person because there is much greater variability in family size than fingers.

Precision is also a function of sample size, and of the way that the sample was selected. Some sampling designs yield more precise estimates than others. For example, simple random samples (which are discussed in this video) are more precise than multi-stage cluster samples (which are discussed in a future video).

4. Visualizing Accuracy and Reliability

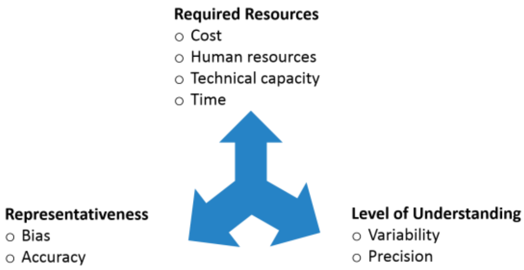


Bias and accuracy are opposites. We can think of accurate statistics as being centered on a bullseye (left), while biased statistics are off target (right). An estimate for the total population would likely be on target if it included a representative number of men and women; and it would likely be off target, or biased, if the sample were only comprised of men.

Precision and variance are opposites. We would have high precision if we repeated a study in the same population at the same point in time and came up with similar results. Precise estimates are similar estimates (top). We would have high variance if we repeated the study in the same population but produced different results (bottom). For example, larger samples result in more precise estimates than smaller samples.

So shouldn't we always aim for unbiased, low variance estimates?

5. Resources

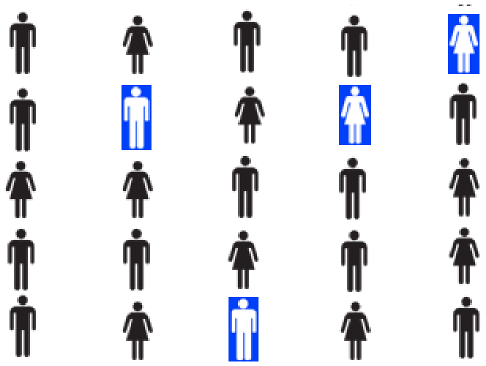


There are three factors that we consider when designing a sample survey. First, the amount of bias that we are willing to accept. Second, the amount of variability that we are willing to accept. And third, the resources required to achieve the target accuracy and precision. Generally, to achieve more accuracy and precision, we need a larger sample and additional people, time, and technical capacity. During survey planning, statisticians and project managers must negotiate a sample design and sample size that satisfies resource availability.

6. Randomization to achieve representivity

In our example of selection bias, we understood that a sample of only men could not be used to extrapolate results about the entire population because no women where included. Ideally, samples are representative of the population. Though how would we know if a sample were representative of the population? We would have to take a census of the population first to be able to sample in a representative way, and that of course, is impractical, and usually impossible. Instead, we achieve representivity by randomizing.

7. Simple Random Sampling (SRS)



Simple random sampling is when we have a full list of everyone in the population, and we randomly choose individuals from the list. The simple random sampling approach ensures that every person in the population has the same probability of being selected. Most sample size calculators, and simple statistics and analyses assume simple random sampling.

8. Other SRS Methods

Variants on the Simple Random Sampling method include consecutive sampling whereby the researcher chooses a random spot in the order of individuals, and samples everyone until the target sample size is met. Another simple random sampling method is called systematic (segment) sampling whereby the researcher chooses a random spot in the order of individuals, and samples every kth individual until the target sample size is met. For these methods to be valid, the order of individuals must be random. For example, a line of passengers boarding an airplane is not random; first class passengers, families, and people with disabilities are at the front of the line. But the order in which people enter a shop or clinic is usually random.

9. Estimating proportions from SRS

We are working with binary outcomes in this course, so let us look at a binary variable in the village.dta dataset and review a few statistical concepts.

Here is a scenario: In a village, we want to estimate the percent of adults who have a bank account. We sample 40 adults at random and ask them if they have a bank account. We code yes as 1, and no as 0.

How do we calculate a **prevalence estimate** (proportion) from binary data? We sum the observations with the trait of interest – in this case, the number of adults that had a bank account – and divide it by the total number of sampled observations, n.

p = prevalence in the total population
Y_j = indicator that observation j has trait of interest
n = sample size

Estimator: $\hat{p} = \frac{\sum Y_j}{n}$

Variance: $\widehat{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$

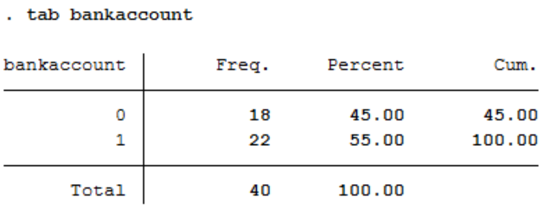
$$\text{Estimator: } \hat{p} = \frac{\sum Y_j}{n}$$

$$\text{Variance: } \widehat{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$$

Open the following dataset in Stata:
village.dta

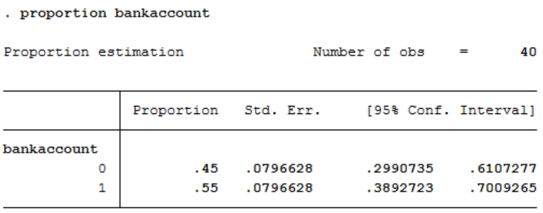
Code: Prevalence estimate

```
tab bankaccount
```



```
. tab bankaccount

bankaccount |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         18       45.00       45.00
          1 |         22       55.00      100.00
------------+-----------------------------------
      Total |         40      100.00
```



```
. proportion bankaccount

Proportion estimation             Number of obs   =         40

--------------------------------------------------------------
             |  Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
bankaccount  |
           0 |        .45    .0796628      .2990735    .6107277
           1 |        .55    .0796628      .3892723    .7009265
--------------------------------------------------------------
```

The variance around this prevalence estimate is equal to $\frac{\hat{p}(1-\hat{p})}{n-1}$. We see that variance is linked to how common the trait is in the population ($\hat{p}$) and the sample size (n).

If the symbology I am using is knew to you, the “hat” over p indicates that it is an estimate, and not a known, measured prevalence in the real population. “Sigma” is a Greek letter used in mathematics to mean “the sum of”. And subscripts like *j* and *i* are used to indicate a calculation “with each” observation.

Let us go back to our example, and in Stata estimate the prevalence of bank account ownership among our simple random sample of adults using the tabulate or proportion statements. The tabulate statement, which we call tab for shorthand, calculates the frequencies and proportions of the sample with the outcome. The proportion statement calculates the percent and variance for a simple random sample study design.

We see that 55% of households in this sample have a bank account. If we could repeat the study, drawing 100 different random samples of 40 people from the population, we would expect the true proportion of households with a bank account to be between 39% and 70% in 95 of those samples. We call this the **95% confidence interval**, which represents the variance, or uncertainty, around our estimate due to small sample size, complex sample design, and lower prevalence in the population.

Assumptions of Simple Analyses

1. Sample is randomly selected
2. Everyone has an equal probability of selection
3. The population is large (we only sample a small fraction of the population), or we sample with replacement

10. Analyses of Simple Random Samples

When performing basic analyses of simple random samples, we typically assume that (1) the sample is randomly selected, (2) everyone has an equal probability of selection, and (3) that the population is very large (so that we only sample a small fraction of the total population), or that we sample with replacement.

What if the population is small? What if we sample a large fraction of the population? Or what if individuals had unequal probability of selection? Watch the videos on finite population correction and sampling weights to learn more about these scenarios, and the videos about complex sampling for more information about how to analyze these data.

Acknowledgement: This presentation was based off of learning materials developed by Dr. Bethany Hedt-Gauthier.