Review of Some Basic Statistical Concepts and the Horvitz-Thompson Estimator

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Reading Assignment:
Scheaffer, Mendenhall, Ott, & Gerow, Chapter 3
Goals for this Lecture

• Compare and contrast classical statistical assumptions to survey data requirements
  – Infinite vs. finite populations
• Estimators for infinite and finite populations
  – Particularly Horvitz-Thompson estimator
• Review:
  – Sampling distributions
  – Central Limit Theorem
  – Margin of error
Purpose of Survey Analysis: Statistical Inference

• Values calculated from survey data (i.e., means and standard deviations) are *statistics*

• Statistics are estimates of the true values of population values (or *parameters*)
  – They’re unlikely to correspond exactly to the values had the entire population been surveyed

• Whole point of a survey is to use the sample data to *infer* back to the entire population

✓ Can be relatively easy to very complicated depending on sampling design
Classical Statistical Assumptions vs. Survey Practice / Requirements

• Classic statistical methods assume:
  – Population is of infinite size (or so large as to be essentially infinite)
  – Sample size is a small fraction of the population
  – Sample is drawn from the population via SRS

• In surveys:
  – Population always finite (though may be huge)
  – Sample could be sizeable fraction of the population
    • “Sizeable” is roughly > 5%
  – Sampling may be complex
Summarizing Population Information: Infinite Population Case

- Consider a population of integers that are equally likely to be 0, 1, ..., 8, 9
- That is, $p(0) = p(1) = \cdots = p(8) = p(9) = \frac{1}{10}$
- The distribution (probability mass function) can be depicted as:

![Probability Mass Function Diagram](attachment:image.png)
Summarizing Population Information: Infinite Population Case

- Summarize the population using the expected value ("mean") and the variance:

\[ \mu = E(Y) = \sum_y y p(y) \]

\[ \sigma^2 = \text{Var}(Y) = E(Y - \mu)^2 = \sum_y (y - \mu)^2 p(y) \]

- For the example, the mean is

\[ \mu = \sum_{y=0}^{9} y p(y) = \frac{1}{10} \sum_{y=0}^{9} y = \frac{1}{10} (45) = 4.5 \]
Summarizing Population Information: Infinite Population Case

- And \( \sigma^2 = \sum_{y=0}^{9} (y - \mu)^2 p(y) \)
  \[
  = \frac{1}{10} \sum_{y=0}^{9} (y - 4.5)^2 \\
  = \frac{1}{10} \left[ (0 - 4.5)^2 + \cdots + (9 - 4.5)^2 \right] \\
  = \frac{1}{10} \left[ 82.5 \right] = 8.25
  \]

- So, \( \sigma = \sqrt{8.25} = 2.9 \)
Estimating Population Information: Infinite Population Case

- But, we only observe a sample from the population: \( y_1, \ldots, y_n \)
- Estimate \( \mu \) with \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) and \( \sigma^2 \) with
  \[
  s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
  \]
- Why these? They have good statistical properties, such as they’re unbiased:
  \[
  E(\bar{Y}) = \mu \quad \text{and} \quad E(S^2) = \sigma^2
  \]
Estimating Population Information: Infinite Population Case

• Also we can derive the standard error of the mean:

\[ \text{s.e.}(\bar{Y}) = \sqrt{\frac{\text{Var}(Y)}{n}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \]

• And we can estimate the standard error of the mean with

\[ \hat{\text{s.e.}}(\bar{Y}) = \sqrt{\frac{\text{Var}(Y)}{n}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \]

• These are important quantities for inference
Key Idea

• Probability distributions are **models of reality**
  – They assume that the population is so large, and the sample is so small with respect to the population, that *each draw of an observation into the sample has no effect on the probability of drawing the next and future observations*
  – So we can ignore issues like whether the observations are drawn with or without replacement

• When the population is finite and sampling is without replacement then this is no longer true
Summarizing Population Information: Finite Population Case

• In your statistics classes, everything was based on the infinite population case

• In surveys, populations can be finite: \( \{u_1, \ldots, u_N\} \)

• Consider the situation where you will choose \( n \) elements out of the \( N \) with probabilities \( \{\delta_1, \ldots, \delta_n\} \), perhaps different on each draw

• How to estimate the population total \( \tau = \sum_{i=1}^{N} u_i \)?

• An unbiased estimator is

\[
\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i}
\]
Summarizing Population Information: Finite Population Case

• To illustrate, imagine you know all the \( y \) values (all positive), and thus the total \( \tau \)
  
  – Choose any \( n \) items each with probability \( \delta_i = y_i / \tau \)
  
  • This is probability sampling according to size

  – Then

  \[
  \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\tau} = \tau
  \]

  – And every estimate is perfect!

• But there’s no point in sampling and estimating if you already know all the values
  
  – So, optimal sampling probabilities not possible
Summarizing Population Information: Finite Population Case

- Now, a sampling with replacement case (as an example – it’s not what you’d really do)
  - Choose \( n \) items with probability \( \delta_i = 1/N \) each
  - Then

\[
E(\hat{t}) = E\left(\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i}\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{E(y_i)}{1/N} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mu}{1/N}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{N} \sum_{j=1}^{N} u_j = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} u_j = \frac{1}{n} \times n \times \tau = \tau
\]
Summarizing Population Information: Finite Population Case

• For example, consider the population $\{u_1, u_2, u_3, u_4\} = \{1, 2, 3, 4\}$ where
  - $\Pr(\text{pick } 1) = 0.1 = \delta_1$
  - $\Pr(\text{pick } 2) = 0.1 = \delta_2$
  - $\Pr(\text{pick } 3) = 0.4 = \delta_3$
  - $\Pr(\text{pick } 4) = 0.4 = \delta_4$

• Note the population total is $\tau = 1 + 2 + 3 + 4 = 10$

• Now, imagine we are going to randomly choose two elements from the population again with replacement
Summarizing Population Information: Finite Population Case

• Then if we happen to choose a 1 and a 2, our estimate of the total is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i} = \frac{1}{2} \left( \frac{1}{0.1} + \frac{2}{0.1} \right) = \frac{1}{2} (10 + 20) = 15$$

• Also, the variance of the total is estimated as

$$\hat{\text{Var}}(\hat{\tau}) = \frac{1}{n} \left[ \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{y_i}{\delta_i} - \hat{\tau} \right)^2 \right] = \frac{1}{2} \left[ \frac{1}{2-1} \sum_{i=1}^{2} \left( \frac{y_i}{\delta_i} - \hat{\tau} \right)^2 \right]$$

$$= \frac{1}{2} \left[ \left( \frac{1}{0.1} - 15 \right)^2 + \left( \frac{2}{0.1} - 15 \right)^2 \right] = \frac{1}{2} \left[ 25 + 25 \right] = 25$$
Summarizing Population Information: Finite Population Case

- Table gives all possible outcomes

- From this, we see
  \[ E(\hat{\tau}) = 15(0.02) + \frac{35}{4}(0.08) + \cdots + 10(0.16) = 10 \]
  - Unbiased!

- Also,
  \[ \text{Var}(\hat{\tau}) = \left(15 - 10\right)^2 (0.02) + \left(\frac{35}{4} - 10\right)^2 (0.08) + \cdots + (10 - 10)^2 (0.16) = 6.25 \]

✓ Under sampling with replacement, the estimator is unbiased for any choice of \( \delta \)s
Summarizing Population Information: Finite Population Case

• What about sampling without replacement?
  – That’s what most surveys do

• Define the $\delta_i$ as the average probability the $i^{th}$ observation is selected: $\delta_i = \pi_i / n$

\[
\hat{t} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\pi_i / n} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}
\]

• Often it’s expressed as a weight, $w_i = 1 / \pi_i$, so

\[
\hat{t} = \sum_{i=1}^{n} w_i y_i
\]
Estimator Still Unbiased

\[ E\left(\hat{\tau}_e\right) = \frac{6 + 8 + 10 + 10 + 12 + 14}{6} = 10 \]

\[ E\left(\hat{\tau}_u\right) = 0.0222 \times 12.2748 + \cdots + 0.5333 \times 9.2652 = 10 \]
Horvitz-Thompson Estimators

• Generally referred to as the Horvitz-Thompson estimator
  – To estimate mean: \( \hat{\mu} = \frac{\hat{\tau}}{N} = \frac{1}{N} \sum_{i=1}^{n} w_i y_i = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{\pi_i} y_i \)

• Estimator is particularly useful in complex sampling where \( \pi_i \) is the probability a respondent is selected from sampling frame
  – Probability can vary by each respondent depending on the sampling scheme
  – Probability can also incorporate the probability of nonresponse
  – \( w_i \) has a nice interpretation we will discuss later
Other Important Concepts: Remember Sampling Distributions

- Abstract from people and surveys to random variables and their distributions
- **Sampling distribution** is the probability distribution of a sample statistic

\[
\sigma_{\bar{X}} = \sigma / \sqrt{n}
\]
Remember: Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean $\mu$ and standard deviation $\sigma$

- For large sample size $n$, the distribution of the sample mean has approximately a normal distribution
  - with mean $\mu$, and
  - standard error $\sigma/\sqrt{n}$

- The larger the value of $n$, the better the approximation
Example: Sums of Dice Rolls

Why do we care about dice? Translate this into the probability of response on a six-point Likert scale…
What Does “Margin of Error” Mean?

• **Margin of error** is just half the width of a 95 percent confidence interval

• Common survey terminology
  – Convention is a 3% margin of error
  – Means a 95% CI is the survey result +/- 3%

• To achieve a desired margin of error, must have the right sample size ($n$)
  – **Power calculations** are done by statisticians to figure out the required sample size to achieve a particular margin of error
What We Have Just Covered

• Compared and contrasted classical statistical assumptions to survey data requirements
  – Infinite vs. finite populations

• Estimators for infinite and finite populations
  – Particularly Horvitz-Thompson estimator

• Briefly reviewed:
  – Sampling distributions
  – Central Limit Theorem
  – Margin of error