Basic Statistical Inference for Survey Data

Professor Ron Fricker
Naval Postgraduate School
Monterey, California
Goals for this Lecture

• Review of descriptive statistics
• Review of basic statistical inference
  – Point estimation
  – Sampling distributions and the standard error
  – Confidence intervals for the mean
  – Hypothesis tests for the mean
• Compare and contrast classical statistical assumptions to survey data requirements
• Discuss how to adapt methods to survey data with basic sample designs
Two Roles of Statistics

- **Descriptive**: Describing a sample or population
  - Numerical: (mean, variance, mode)
  - Graphical: (histogram, boxplot)

- **Inferential**: Using a sample to *infer* facts about a population
  - Estimating (e.g., estimating the average starting salary of those with systems engineering Master’s degrees)
  - Testing theories (e.g., evaluating whether a Master’s degree increases income)
  - Building models (e.g., modeling the relationship of how an advanced degree increases income)
A Descriptive Statistics Question: *What was the average survey response to question 7?*
An Inferential Question: *Given the sample, what can we say about the average response to question 7 for the population?*
Lots of Descriptive Statistics

- **Numerical:**
  - Measures of location
    - Mean, median, trimmed mean, percentiles
  - Measures of variability
    - Variance, standard deviation, range, inter-quartile range
  - Measures for categorical data
    - Mode, proportions

- **Graphical**
  - Continuous: Histograms, boxplots, scatterplots
  - Categorical: Bar charts, pie charts
Continuous Data: Sample Mean, Variance, and Standard Deviation

- Sample average or sample mean is a measure of location or central tendency:
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

- Sample variance is a measure of variability
  \[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

- Standard deviation is the square root of the variance
  \[ s = \sqrt{s^2} \]
Statistical Inference

- Sample mean, $\bar{x}$, and sample variance, $s^2$, are statistics calculated from data.
- Sample statistics used to estimate true value of population called estimators.
- **Point estimation**: estimate a population statistic with a sample statistic.
- **Interval estimation**: estimate a population statistic with an interval—incorporates uncertainty in the sample statistic.
- **Hypothesis tests**: test theories about the population based on evidence in the sample data.
• Basic 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 very large)
  – Sample could be sizeable fraction of the population
    • “Sizeable” is roughly > 5%
  – Sampling may be complex
Point Estimation (1)

• Example: Use sample mean or proportion to estimate population mean or proportion

• Using SRS or a self-weighting sampling scheme, usual estimators for the mean calculated in all stat software packages are generally fine
  – Assuming no other adjustments are necessary
    • E.g., nonresponse, poststratification, etc

• Except under SRS, usual point estimates for standard deviation almost always wrong
Point Estimation (2)

• Naïve analyses just present sample statistics for the means and/or proportions
  – Perhaps some intuitive sense that the sample statistics are a measure of the population
  – But often don’t account for sample design

• However, when using point estimates, no information about sample uncertainty provided
  – If you did another survey, how much might its results differ from the current results?

• Also, even for mean, if sample design not self-weighting, need to adjust software estimators
Sampling Distributions

- Abstract from people and surveys to random variables and their distributions

Population

Sample

Random Sample

Different sample, different value

Sample Statistic
Sampling Distributions

• **Sampling distribution** is the probability distribution of a sample statistic

\[ \sigma_{\bar{X}} = \sigma / \sqrt{5} \]

Distribution individual obs with standard deviation \( \sigma \)
Demonstrating Randomness

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html
Simulating Sampling Distributions

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html
Central Limit Theorem (CLT) for the Sample Mean

• 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 deviation $\frac{\sigma}{\sqrt{n}}$.
• The larger the value of $n$, the better the approximation.
Example: Sums of Dice Rolls

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
</tr>
<tr>
<td>20</td>
<td>300</td>
</tr>
<tr>
<td>30</td>
<td>400</td>
</tr>
<tr>
<td>40</td>
<td>500</td>
</tr>
<tr>
<td>50</td>
<td>600</td>
</tr>
<tr>
<td>60</td>
<td>700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>27</td>
</tr>
<tr>
<td>15</td>
<td>29</td>
</tr>
<tr>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>17</td>
<td>33</td>
</tr>
<tr>
<td>18</td>
<td>35</td>
</tr>
<tr>
<td>19</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>

1 roll -> 2 rolls -> 5 rolls -> 10 rolls
Demonstrating Sampling Distributions and the CLT

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html
Interval Estimation for $\mu$

- Best estimate for $\mu$ is $\bar{X}$
- But $\bar{X}$ will never be exactly $\mu$
  - Further, there is no way to tell how far off
- BUT can estimate $\mu$’s location with an interval and be right some of the time
  - Narrow intervals: higher chance of being wrong
  - Wide intervals: less chance of being wrong, but also less useful
- AND with confidence intervals (CIs) can define the probability the interval “covers” $\mu$!
Confidence Intervals: Main Idea

- Based on the normal distribution, we know $\bar{X}$ is within 2 s.e.s of $\mu$ 95% of the time.
- Alternatively, $\mu$ is within 2 s.e.s of $\bar{X}$ 95% of the time.

![Graph showing confidence intervals and distributions.](image)
A Simulation

95% Confidence Intervals for mean = 50, sd = 10, n = 5

intervals not including population mean: 2
Another Simulation

95% Confidence Intervals for mean = 50, sd = 10, n = 95

intervals not including population mean: 10
• For $\bar{X}$ from sample of size $n$ from a population with mean $\mu$, 

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

has a $t$ distribution with $n-1$ “degrees of freedom”
  – Precisely if population has normal distribution
  – Approximately for sample mean via CLT

• Use the $t$ distribution to build a CI for the mean:

$$\Pr\left(-t_{\alpha/2, n-1} < T < t_{\alpha/2, n-1}\right) = 1 - \alpha$$
Review: the $t$ Distribution

$Z = \text{number of SE's from the mean}$

- $T3$
- $T10$
- $T100$
Confidence Interval for $\mu$ (2)

• Flip the probability statement around to get a confidence interval:

1. $\Pr\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1 - \alpha$

2. $\Pr\left(-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha$

3. $\Pr\left(\frac{\bar{X} - t_{\alpha/2,n-1}}{s/\sqrt{n}} < \mu < \frac{\bar{X} + t_{\alpha/2,n-1}}{s/\sqrt{n}}\right) = 1 - \alpha$
Example: Constructing a 95% Confidence Interval for $\mu$

- Choose the confidence level: $1 - \alpha$
- Remember the degrees of freedom ($\nu$) = $n - 1$
- Find $t_{\alpha/2, n-1}$
  - Example: if $\alpha = 0.05$, df=7 then $t_{0.025, 7} = 2.365$
- Calculate $\bar{X}$ and $s / \sqrt{n}$

Then

$$\Pr\left( \bar{X} - 2.365 \frac{s}{\sqrt{n}} < \mu < \bar{X} + 2.365 \frac{s}{\sqrt{n}} \right) = 0.95$$
Hypothesis Tests

- Basic idea is to test a hypothesis / theory on empirical evidence from a sample
  - E.g., “The fraction of new students aware of the school discrimination policy is less than 75%.”
  - Does the data support or refute the assertion?

If we assume this is true, how likely are we to see this (or something more extreme)?

\[ \hat{p} = 0.832 \]

This is the probability. If it’s small, we don’t believe our assumption.
One-Sample, Two-sided $t$-Test

- **Hypothesis:**
  - $H_0$: $\mu = \mu_0$
  - $H_a$: $\mu \neq \mu_0$

- **Standardized test statistic:**
  $$ t = \frac{\bar{X} - \mu_0}{\sqrt{s^2 / n}} $$

- **$p$-value:** $\Pr(T<-t \text{ and } T>t) = \Pr(|T|>t)$, where $T$ follows a $t$ distribution with $n-1$ degrees of freedom

- **Reject $H_0$ if $p < \alpha$, where $\alpha$ is the predetermined significance level**
One-Sample, One-sided $t$-Tests

• Hypotheses:
  \[ H_0: \mu = \mu_0 \quad \text{or} \quad H_0: \mu = \mu_0 \]
  \[ H_a: \mu < \mu_0 \quad \text{or} \quad H_a: \mu > \mu_0 \]

• Standardized test statistic:
  \[ t = \frac{\bar{X} - \mu_0}{\sqrt{s^2 / n}} \]

• $p$-value = $\Pr(T < t)$ or $p$-value = $\Pr(T > t)$, depending on $H_a$, where $T$ follows a $t$ distribution with $n-1$ degrees of freedom

• Reject $H_0$ if $p < \alpha$
Applying Continuous Methods to Binary Survey Questions

• In surveys, often have binary questions, where desire to infer proportion of population in one category or the other

• Code binary question responses as 1/0 variable and for large \( n \) appeal to the CLT
  – Confidence interval for the mean is a CI on the proportion of “1”s
  – T-test for the mean is a hypothesis test on the proportion of “1”s
Applying Continuous Methods to Likert Scale Survey Data

- Likert scale data is inherently categorical
- If willing to make assumption that “distance” between categories is equal, then can code with integers and appeal to CLT

- Strongly agree → 1
- Agree → 2
- Neutral → 3
- Disagree → 4
- Strongly disagree → 5
Adjusting Standard Errors (for Basic Survey Sample Designs)

- Sample > 5% of population: finite population correction
  - Multiply the standard error by $\sqrt{(N - n) / N}$
  - E.g. $s.e. (\bar{x}) = \sqrt{(N - n) / N} \times s / \sqrt{n}$

- Stratified sample, weighted sum of the strata variances:
  $s.e. (\bar{x}) = \sqrt{\sum_{h=1}^{H} \left( N_h / N \right) \text{Var}(\bar{x}_h)}$
What We Have Just Reviewed

• Review of descriptive statistics
• Review of basic statistical inference
  – Point estimation
  – Sampling distributions and the standard error
  – Confidence intervals for the mean
  – Hypothesis tests for the mean
• Compare and contrast classical statistical assumptions to survey data requirements
• Discuss how to adapt methods to survey data with basic sample designs