Introduction to Survey Analysis

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Reading Assignment: None
Goals for this Lecture

• Introduction to analysis for surveys
  – Assuming simple random sampling (SRS)
  – With and without finite population correction (fpc)

• Basic methods
  – Confidence intervals
    • Binary questions
    • Simultaneous CIs for Likert scale responses
  – Hypothesis tests
Important Question: What Methods Appropriate for Survey Data Analysis?

• Sub-question: When can we use sample mean or proportion to estimate population mean or proportion?
  – Answer: Depends on the response scale
• For this lecture, we will start by assuming SRS and sample size less than 5 percent of the population
  – So standard methods apply
  – Then we’ll add in corrections for fpc
Remember the Central Limit Theorem

• Analysts often a bit confused by the (typically) discrete nature of survey data
  – Most methods in “Stat 101” classes based on Normal distribution, which is not discrete

• But remember the Central Limit Theorem:
  Sampling distributions of sums and averages of \( iid \) data are approximately normally distributed

• So, under simple random sampling (SRS), with large enough samples, Likert scale and other types of discrete survey data \textit{may not be a problem}
Naïve Analyses

• 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

• Further, “point estimates” provide no information about sample uncertainty
  – If you did another survey, how much might its results differ from the current results?
  – Leaves it up to the reader to guess at the precision
  – not a good idea
Illustrative Example

- Consider a binary yes/no question in survey with a population of $N=100,000$ people:
  - For a sample of $n_1=4$ people, if 3 say “yes” our point estimate is $\hat{p}_1 = 3/4 = 0.75$
  - For a sample of $n_2=500$ people, if 375 say “yes” our point estimate is $\hat{p}_2 = 375/500 = 0.75$

- But clearly uncertainty of $\hat{p}_1$ is greater than $\hat{p}_2$:

  \[
  s.e.(\hat{p}_1) = \sqrt{\frac{0.75(1-0.75)}{4}} = 0.217 \quad \text{VS.} \quad s.e.(\hat{p}_2) = \sqrt{\frac{0.75(1-0.75)}{500}} = 0.019
  \]
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
So, for Binary Questions

• The $(1-\alpha)100\%$ confidence interval is

$$\left( \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \right)$$

where

$$\hat{p} = \frac{\# \text{ of successes/agrees/yeses}}{\# \text{ of respondents}}$$

• Similarly, one can do a hypothesis test, just treat $\hat{p}$ as the mean and the standard error is

$$\sqrt{\hat{p}(1-\hat{p})/n}$$
When the FPC Applies

• The $(1-\alpha)100\%$ confidence interval is

\[
\left( \hat{p} - z_{\alpha/2} \sqrt{\frac{1}{N} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \left( \frac{1}{n} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \right) } , \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{1}{N} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \left( \frac{1}{n} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \right) } \right)
\]

where \( \hat{p} = \frac{\text{# of successes/agrees/yeses}}{\text{# of respondents}} \)

• Similarly, one can do a hypothesis test, just treat \( \hat{p} \) as the mean and the standard error is

\[
\sqrt{\frac{1}{N} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \left( \frac{1}{n} \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \right) }
\]
Example: CI for Fraction of DL Students Who Have Received a Library Briefing

- Point estimate: \( \hat{p} = \frac{408}{654} = 0.624 \) or 62.4%
- Standard error (with fpc):
  \[
  s.e.(\hat{p}) = \sqrt{1 - \frac{654}{1031}} \left( \frac{0.624(1-0.624)}{654} \right) = 0.011
  \]
- (Approximate) 95% CI:
  \[
  (0.624 - 1.96 \times 0.011, 0.624 + 1.96 \times 0.011) = (0.602, 0.646)
  \]
- So, we’re 95% confident between 60.2% and 64.6% of DL students briefed by library
- Major assumption: Respondents are random sample of all DL students
Solution Using the R survey Package

```r
> library(survey)
> briefing.data <- data.frame(briefing.ind=c(rep(1,408),rep(0,246)),N=rep(1031,654))
> dim(briefing.data)
[1] 654 2
> head(briefing.data)
  briefing.ind   N
  1          1 1031
  2          1 1031
  3          1 1031
  4          1 1031
  5          1 1031
  6          1 1031
> summary(briefing.data)
  briefing.ind   N
     Min.   :0.0000 Min. :1031
     1st Qu.:0.0000 1st Qu.:1031
     Median :1.0000 Median :1031
     Mean  :0.6239 Mean  :1031
     3rd Qu.:1.0000 3rd Qu.:1031
     Max.  :1.0000 Max. :1031
> briefing.design <- svydesign(~1,fpc=~N,data=briefing.data)
> svy.mean(~briefing.ind,briefing.design)

mean       SE
briefing.ind 0.62385 0.0115
```
Example Continued: Hypothesis Testing

• Test the hypothesis that the proportion of DL students in the population who have received a briefing is greater than 60%

• We want to know

\[ p\text{-value} = \Pr(\bar{X} > 0.624 | \mu = 0.60, s = 0.011) \]

• In R, we have

\[
\begin{align*}
&> \text{pnorm}(0.62385, 0.6, 0.0115, \text{lower.tail}=\text{F}) \\
&[1] \ 0.01904369
\end{align*}
\]

• \( p\text{-value} < \alpha = 0.05 \), so we can conclude that the true proportion is greater than 60%
CIs For Likert Scale Questions with More than Two Levels

- Need to calculate simultaneous confidence intervals for \( k \) Likert scale levels

- Use

\[
\frac{n_i + B / 2}{n + B} \pm \sqrt{\frac{B^2 / 4 + B n_i \left(1 - n_i / n\right)}{(n + B)^2}}
\]

where

- \( n_i \) = number of respondents choosing the \( i \)th level

- \( n = \sum_{i=1}^{k} n_i \)

- \( B \) is the upper \( (\alpha/k)100^{th} \) percentile of the \( \chi^2 \) distribution with 1 degree of freedom

Example from Library Survey

• Survey results:

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Library Research is a Crucial Part of My NPS DL Studies”</td>
<td>95 (15.0%)</td>
<td>218 (34.4%)</td>
<td>196 (31.0%)</td>
<td>86 (13.6%)</td>
<td>27 (4.3%)</td>
<td>11 (1.7%)</td>
</tr>
</tbody>
</table>

• For a 95% CI, \( \alpha = 0.05 \), so with \( k = 6 \):

\[
> \text{qchisq}(0.05/6,1,\text{lower.tail}=\text{FALSE})
\]

[1] 6.960401

• Then for the percent saying strongly agree

\[
\frac{95 + 6.96 / 2}{633 + 6.96} \pm \sqrt{\frac{6.96^2 / 4 + 6.96 \times 95 \times (1 - 95 / 633)}{(633 + 6.96)^2}} = 0.154 \pm 0.037
\]

2/22/13

* Data from 2011 survey of DL students for NPS Library
Graphically
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
An Example of Inference for the Mean of a Likert Scale

- Consider the new student survey conducted by a previous class.
- Question 1 asked, “In general how do you rate the satisfaction or dissatisfaction with the IN-PROCESSING procedures at NPS?”
  - Response scale was a 5-point Likert from Very Satisfied (5) to Very Dissatisfied (1).
- We might be interested in the mean response for the population.
  - Is it near neutral, in the positive range, or the negative?

* Data from 2008 survey of NPS new students
Example, continued

- A confidence interval is appropriate to address this question
  - Assume the respondents were a SRS from a population of size $N=183$

- Data:

```r
> summary(NewStdSvyEx)

Sex                  | Country              | Race          |
----------------------|----------------------|---------------|
F: 18                | Other Country        | :52           |
M:152                | Singapore            | Asian American/Pacific Islander: 6 |
                      | United States of America:118 | Black/African American: 10 |
                      |                      | Hispanic/Latinos: 8 |
                      |                      | Unknown: 9     |
                      |                      | White: 85      |

Military.Branch      | Mil.Rank | Q1  |
----------------------|----------|-----|
CIV: 2               | :52      | :1.000 |
USA: 19              | Junior:16| Median: 4.000 |
USAF: 3              | Mid:61   | Mean: 3.821 |
USMC: 8              | Senior:39| 3rd Qu.: 4.000 |
USN: 86              |          | Max.: 5.000 |
                      |          | NA's: 8.000 |
```

* Data from 2008 survey of NPS new students
Example, continued

• To calculate:  
\[ \bar{x} \pm z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right)} \]

• The 95% confidence interval:

```r
> fpc <- 1 - 170/183
> fpc
[1] 0.07103825
> x.bar <- mean(NewStdSvyEx$Q1, na.rm=T)
> x.bar
[1] 3.820988
> s.e. <- sqrt(fpc*var(NewStdSvyEx$Q1, na.rm=T)/162)
> s.e.
[1] 0.01745072
> x.bar - 1.96 * s.e.
[1] 3.786784
> x.bar + 1.96 * s.e.
[1] 3.855191
```

* Data from 2008 survey of NPS new students
Example, continued

- Using the R survey package:
  
  ```r
  > NewStdSvyEx <- NewStdSvy
  > NewStdSvyEx <- data.frame(NewStdSvy,N=rep(183,170))
  > NewStdSvy.design <- svydesign(id=~1,fpc=~N,data=NewStdSvyEx)
  > svymean(~Q1,NewStdSvy.design,na.rm=T)
  
  mean  SE
  Q1 3.821 0.0174
  ```

- With 95% confidence, the mean response for the population is in the range (3.79, 3.86)
  - Average person in the population is on the satisfied side of the Likert scale
Example, continued

• Perhaps we want to test the hypothesis that the population mean response is less than 4
  – I.e., the average person is not quite “satisfied” (i.e., a 4) on the Likert scale
  – First, assume a SRS from a large population

• Then

```r
> t.test(NewStdSvyEx$Q1, mu=4, alternative="less")

One Sample t-test

data:  NewStdSvyEx$Q1
t = -2.7341, df = 161, p-value = 0.003478
alternative hypothesis: true mean is less than 4
95 percent confidence interval:
   -Inf 3.929306
sample estimates:
mean of x
 3.820988
```

* Data from 2008 survey of NPS new students
Now, Accounting for the FPC

• Note that we’ve already proven the alternative hypothesis that the mean is less than 4
  – Using the fpc means the standard errors will be smaller, so we’ll just reject more strongly

• But, if we want the exact \( p \)-value anyway:

\[
\text{> pt((3.821-4)/0.01745,161,lower.tail=}T) \\
\text{[1] 1.292626e-19}
\]

• It’s really, really small: We can be sure that, while the population feels positively about NPS check-in, the average person is not quite “satisfied” as defined on the Likert scale

* Data from 2008 survey of NPS new students
To Summarize

• To analyze mean responses when (1) SRS used, (2) fpc does not apply, and (3) coding scale appropriate, proceed with “Stat 101” analytical tools
  – With fpc, just need to adjust the standard errors
  – With complex sampling, more complicated analytical methods or software required

• Can also use categorical data methods, such as chi-squared tests
  – More on this in upcoming lecture
What We Have Covered

• Introduction to analysis for surveys
  – Assuming simple random sampling (SRS)
  – With and without finite population correction (fpc)

• Basic methods
  – Confidence intervals
    • Binary questions
    • Simultaneous CIs for Likert scale responses
  – Hypothesis tests