Business Statistics

Lecture 6: Polls and Sampling
Goals for this Lecture

- Surveys (polls)
- Types of sampling
- Bias vs. variance
- Power calculations
- Confidence intervals for proportions

✓ Don’t worry about the discussion in Business Statistics about binomial and hypergeometric distributions
An Aside: Demonstrating Randomness

This is a demonstration showing that statistics are random variables too.

(www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html)
What is a Poll?

• A poll (or survey) is a:
  • systematic method for gathering information
  • from a sample of people
  • for the purposes of constructing quantitative descriptors
  • of the attributes of the larger population of which the people are members
• It’s data collection from people for inference!
Proper Sampling is Critical

• If we are to use a sample to infer something about a population, we need to:
  • Have some assurance that the sample is representative of the population (i.e., minimize the chance of bias)
  • Be able to quantify how far off our sample statistic could be from the population statistic (sampling error)
• Using a random sample is protection against (unknowingly) selecting a biased sample
• Classical statistics is all about quantifying uncertainty (i.e., sampling error) and using that information to determine statistical significance
Why Sample?

- E.g., why conduct:
  - Nielson survey a sample of US television viewers
  - Clinical trial of how a drug affects a sample of individuals in the trial
  - Test a sample of “widgets” from a factory production line

- Rather than:
  - Evaluate TV viewing preferences for every individual in the US
  - Test how a drug affects every person in a population
  - Do a quality control check of every “widget” produced

✔ Collecting data for whole populations can be expensive and/or impossible
Major Sampling Categories

• **Random sample**: Choose participants or units randomly
  • Key idea: Surveyor or pollster chooses who can take the survey
  • Statistical inference possible

• **Convenience sample**: Choose sample in some other way—often up to the respondent to choose to participate
  • Examples:
    • 900 number and other call-in polls
    • Internet and e-mail surveys (usually)
    • Shopper and visitor surveys
  • Statistical inference **not** possible
Types of Random Sampling

- **Simple random sample (SRS):** any two samples of the same size are equally likely to be selected.
- **Some other random sampling methods:**
  - **Stratified sampling**
    - Divide population into nonoverlapping, homogeneous groups and then draw a SRS from each group.
  - **Cluster sampling**
    - Data naturally occurs in clusters.
    - Use SRS to select clusters.
Sampling Using Randomization

• “Flip a coin” to decide who gets included
  • Most biases are caused by survey designers taking a convenience sample of units they can get their hands on
  • Units selected for convenience may not look like the units in the larger population
  • Units randomly selected from the population will, on average, look like the population

• Randomization turns bias into variability
  • Just taking larger samples will not fix bias
Bias and Variance

• Poorly constructed data collection methods suffer from bias
  • **Bias**: There is something systematically wrong with the study
  • If you did many studies and averaged the results would you get the right answer?

• Variance is a feature of all methods
  • If you did the study again you’d get a different answer
  • How precise are your estimates?
  • What is the standard error of the sample statistic?
**Low bias**

**Low variance**

**High bias**

**High variance**

**Observed sample statistic** (sample mean)

**True population value** (pop. mean)

**Other sample statistics you could have observed**
Cures for Bias and Variance

• Possibility of bias minimized by randomization
  • When studying people, can still have bias even after randomization

• Variance is reduced when you add more data
  • Confidence intervals for population mean:

\[
\left( \bar{x} - t \frac{s}{\sqrt{n}}, \bar{x} + t \frac{s}{\sqrt{n}} \right)
\]

- \(s\) determined by variability in data
- \(n\) chosen to make the interval small enough for your practical purposes
- \(t\) determined by desired confidence level \(t \approx 2\) gives 95% confidence
Steps in a Survey (1)

• Define population and sampling unit
  • **Sampling unit**
    • Basic unit of analysis
    • E.g., per capita income or household income?

• Construct **sampling frame**
  • Big list containing “almost” everyone in the population
    • **Frame coverage bias** happens when the sampling frame misses important members of the population
    • Useful sampling frame: US census
Steps in a Survey (2)

• Select sample
  • For an SRS, every unit should have an equal probability of appearing
  • Size bias: “bigger” units are more likely to be represented than “smaller” units

• Collect data
  • Non-response bias: units that do not answer your questions look different than those that do
  • Selection bias: units with a particular trait strong opinions are favored
    • Strong opinion
    • Access to survey mode (telephone, internet, etc.)
Steps in a Survey (3)

- Analyze data
  - Were response patterns as expected?
  - Were there errors in completing the survey?
  - Sensitivity bias: answers to questions of salary/sex/other social taboos might not be truthful

- Report results.
  - Reporting bias:
    - Everything gives you cancer
    - Nobody reports the results from studies that show no link between X and cancer
Ex: 1936 US Presidential Election

• Literary Digest Poll, FDR vs Alfred Landon
  • Sampling frame taken from telephone numbers and club membership lists
    • Only one house in 4 had a phone
    • Rich tend to join clubs more than poor
    • Sampling (frame coverage) bias
  • Sent questionnaires to 10 million people
    • Only 2.4 million people responded
    • Nonresponse bias
Results of the Poll

- LD Prediction: FDR 43%, Landon 57%
- Election Result: FDR 62%, Landon 38%
- Largest error ever made by a major poll
  - Literary Digest had correctly forecast every election since 1916
  - Literary Digest went bankrupt soon after 1936
- The Gallup organization (just getting started in 1936)
  - Guessed the election more or less correctly
  - Guessed LD’s result using much less data
Ex: Dewey Defeats Truman

- Even a survey chosen on purpose to be representative can suffer from bias
  - 1948 Gallup Poll (others did similar things)
  - Quota sampling: Survey crews interview people matching particular characteristics
- Interview 13 people:
  - 7 men, 6 women
  - Of the men, 3 under 40, 4 over 40
  - etc...
- What’s wrong with quota sampling?
Who Would You Interview?

- Republican
- Democrat

• Survey interviewers showed republican bias
CIs for Proportions

- Observe something with a **binary** outcome
  - Vote for candidate “A” or “B”
  - Agree or disagree with survey question
- Probability individual in population has characteristic is $p$
  - Take a sample of size $n$
  - Then $X$ has a **binomial** distribution
- How to calculate a CI for $p$?

![Diagram]

Population
Proportion $p$ with characteristic

Random sample of size $n$

With characteristic
Observe $x$ with characteristic

Without characteristic
Observe $n-x$ without characteristic
CI for Population Proportion $p$

- Estimate $p$ with $\hat{p} = x/n$ then
- When $np > 5$ and $n(1-p) > 5$, binomial can be approximated by the normal
- Calculate a $100(1-\alpha)\%$ CI as

\[
\hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

- E.g., a 95% confidence interval is

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]
Example from *Business Stats*

- From a sample of $n=200$, estimate $\hat{p} = 0.62$
  - I.e., 124 respondents agreed to a question
- So,
  
  \[
  \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{0.62(1-0.62)/200} = 0.03432
  \]
- Thus,
  
  \[
  \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.62 \pm 1.96 \times 0.03432 \\
  = [55.27\%, 68.73\%]
  \]
What Does “Margin of Error” Mean?

• **Margin of error** \((E)\) is just half the width of a 95 percent confidence interval
  • In a poll, it’s \(\approx 2\sqrt{\hat{p}(1-\hat{p})/n}\)
  • Previous example \(E=6.7\%\)

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

• To achieve a desired margin of error, must have the right sample size \((n)\)
  • Called power calculations
Power Calculations for $p$?

- Start with margin of error for a confidence interval:
  \[ E = 2\sqrt{\hat{p}(1 - \hat{p})/n} \]

- Estimate using worst case: $\hat{p} = 1/2$

- For 95% CI, solve for $n$ to get required sample size: $n = 1/E^2$

- Example: Want a 3% margin of error
  - $n = 1/0.03^2 = 1,111.1$
  - So, sample 1,112 people
Additional Details

• Previous calculations assume
  • Very large population
  • Simple random sampling
• If sample is large fraction of population (> 5%), should also use finite population correction
  • See textbook
• If sampling scheme is not SRS, power calculations more complicated
  • See a statistician
Case: Survey1.jmp

• Hotel sampled 1,124 guests present on given day
  • Asked, “Do you plan to return?”
  • 97% response rate! (hotel had to work hard to get this rate)

• Can they believe their results?

<table>
<thead>
<tr>
<th>Moments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.630</td>
</tr>
<tr>
<td>Std Dev</td>
<td>0.483</td>
</tr>
<tr>
<td>Std Error Mean</td>
<td>0.014</td>
</tr>
<tr>
<td>Upper 95% Mean</td>
<td>0.658</td>
</tr>
<tr>
<td>Lower 95% Mean</td>
<td>0.602</td>
</tr>
<tr>
<td>N</td>
<td>1124.000</td>
</tr>
<tr>
<td>Sum Weights</td>
<td>1124.000</td>
</tr>
</tbody>
</table>
CI Calculations

• We have that out of 1,124 people 63% said they plan to return
  • I.e., \( n=1,124 \) and \( \hat{p} = 0.63 \)
• So,
  \[
  \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.63(1-0.63)}{1124}} = 0.0144
  \]
• Thus,
  \[
  \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.63 \pm 1.96 \times 0.0144
  \]
  \[
  = [60.2\%, 65.8\%]
  \]
• Same as JMP!
But, a Problem…

- People who stay longer say they are more likely to return.
- Size bias: People who stay longer are more likely to be included in the sample.
- Sample proportion for “WillReturn” suffers from upward bias.
- Lesson: Do the right randomization!
What We Covered in this Class

- Surveys
  - Random selection ensures survey is representative
  - Randomized surveys can be generalized to population
- Types of sampling
- Bias vs. variance
- Power calculations
- Confidence intervals for proportions