Math 161 Lecture 2

Glen Pugh

September 8, 2007

Lecture 2

Chapter 2: Samples, Good and Bad

Samples and Sampling

Overview

- We wish to obtain information about a population by choosing a sample and gathering information from the individuals which make up the sample.
- Expectation (or hope!): The sample is representative of the entire population.
- Choose sample with care: Use chance!

First: How not to Sample

- Convenience Sampling: Sample consists of individuals which are easiest to reach.
- Problem: Sample not representative of population of interest.
- Example: Ontario welfare reform in 1995– reduced benefits resulted in reduced welfare ranks. A 1996 telephone poll of people who had gone off welfare found that 62% of these people listed finding a job as their reason for no longer collecting welfare. Aha! claimed the government, 62% of those previously on welfare are now working thanks to the reforms.

First: How not to Sample

- Voluntary Response Sample: Sample consists of individuals who elect to respond to a question/study.
- Problem: People with strong feelings about the issue tend to respond– sample not representative of population of interest.
- Examples:
  - RateMyProfessors.com
  - B.C. Treaty Referendum of 2002

Bias in Sampling

- Convenience Sampling and Voluntary Response Sampling introduce bias into a statistical study.
- Bias: The systematic favoring of one outcome over another.
- Example: Evaluate a professor by asking all failing students if they were satisfied with the professor's teaching.

Avoiding Bias: Simple Random Samples

- Use chance to select a sample from the population of interest
- Simple Random Sample (or SRS): An SRS of size n is a sample of n individuals chosen from a population in such a way that every sample of size n has the same chance of being selected.
- Selecting an SRS: To select an SRS of size n:
  - Label each individual in the population with a unique number.
  - Randomly draw n numbers from the collection of numbers representing the population.
  - The n individuals with labels matching the numbers drawn form the sample.

Selecting an SRS: Table of Random Digits

- Table of Random Digits: A table made up of digits from 0 to 9, each entry in the list equally likely to be any of the digits 0 through 9, each entry independent of the others.
- See Table A on page 550 of text
- Example: Choose an SRS of size 5 from the class using line 122 of Table A.

Selecting an SRS: Computer Software

- More practical (and modern) method of choosing an SRS: Use computer software.
- Some web-based random number generators:
  - `http://bcs.whfreeman.com/scc6e` and click on 'Statistical Applets', then click on 'Simple Random Sample' (Java required).
  - `http://www.randomizer.org/form.htm`
- Example: Choose an SRS of size 5 from the class.

Chapter 3: What Do Samples Tell Us?

Putting our SRS to Work

- We use an SRS as a snapshot of the entire population.
- We wish to use facts about our SRS to make estimates about the entire population.

Putting our SRS to Work: Example

- Example (from first day): What proportion (or percentage) of Americans believe there is solid evidence that the earth is warming?
- Take an SRS of 1708 adult Americans and ask them the question.
- We find that 1315 of the 1708 people surveyed believe the earth is warming.
- So we *estimate* that

$$\frac{1315}{1708} = 0.769 = 77\%$$

of the entire population believes the earth is warming.

Parameters and Statistics

- Need terminology to distinguish between the quantities used to describe the entire population and those used to describe the sample.
- Parameter: Number (usually unknown) which describes the entire population.
- Statistic: Number (known) which describes a sample.

Parameters and Statistics: Example

- Example:
  - Let $p =$ true (but unknown) proportion of Americans who believe there is solid evidence that the earth is warming.
  - p is a parameter.
  - Let $\hat{p} =$ proportion of our sample of 1708 adult Americans who believe there is solid evidence that the earth is warming.
  - $\hat{p}$ is a sample statistic. ($\hat{p}$ is read as "p hat").
- Notice the notation: The "hat" denotes the statistic used to estimate the corresponding parameter. For example, if c is some parameter, then $\hat{c}$ would be the corresponding statistic used to estimate c.

Variability

- We expect that an SRS from a population avoids bias and gives a statistic which is a good estimate for the corresponding population parameter.
- If we take a second SRS and calculate the same statistic, it is unlikely that its value will be exactly the same as the first, but we expect it to be close.
- If we continue this, taking many SRS's and calculate statistics, will the results all be close or quite spread out?

Variability Demo from Text

Simplified Example: Suppose we have a population of 300,000,000 of which half 150,000,000 are males, other half females.

- Let $p =$ male proportion of the population, so $p = 1/2 = 0.5$
- Repeatedly take SRS's of size $n = 100$ and for each sample determine $\hat{p} =$ male proportion of the sample.
- Plot the results on a histogram:

<See Figure 3.1 on p.36 of text>

Variability Demo from Text

Same Example, only this time we take SRS's of size $n = 2527$, and again $\hat{p} =$ male proportion of the sample:

<See Figure 3.2 on p.36 of text>

Conclusions

- Notice in both examples most values of $\hat{p}$ are centered around the true value of $p = 0.5$
- The $\hat{p}$ values from the $n = 100$ sample size are more spread out compared to those of the $n = 2527$ sample size.
- In general, increasing the sample size reduces the variability.

Bias and Variability in a Nutshell

- Bias: Consistent, repeated deviation of the sample statistic from the population parameter in the same direction when we take many samples.
- To reduce bias use random sampling.
- Variability: How spread out the values of the sample statistics are when many samples are taken.
- To reduce variability of an SRS use a larger sample size.

Bias and Variability in Pictures

<See Figure 3.2 on p.36 of text>

Confidence Statements

- Back to our Example: Recall our example which stated that 77% of adult Americans believe there is solid evidence that the earth is warming.
- We were told that the survey was "accurate to within 3% nineteen times out of twenty".
- This is called a confidence statement.
- What does this mean?

Confidence Statement: Translation

- First, $\frac{19}{20} = 0.95 = 95\%$, so "nineteen times out of twenty" means 95%.
- "the survey was accurate to within 3% nineteen times out of twenty"

  means

  "if we repeated this survey over and over, 95% of the samples would give a sample statistic within 3 percentage points of the true (unknown) value of the population parameter"
- The 3% in this case is called the margin of error, and depends on the sample size.
- The 95% is called the confidence level and tells us how much we can trust the result of a single sample.

Margin of Error

- Increasing the sample size reduces the margin of error.
- How does sample size translate into margin of error?
- Rule of thumb: If estimating a population proportion p using a sample proportion $\hat{p}$ from a SRS of size n, and assuming a 95% confidence level, the margin of error is approximately $1/\sqrt{n}$.
- Example: In our earth warming example, the sample size was 1708, so we expect the margin of error to be

$$\frac{1}{\sqrt{1708}} = 0.024 = 2.4\%$$

  Not exactly 3%, but close!

Margin of Error: Example

Once again, consider the study of the proportion of Americans who believe there is solid evidence that the earth is warming. Assuming a 95% confidence level, suppose we wish the margin of error to be only 1%. How large a sample is required?