Lecture 6: sampling distribution & confidence intervals

What we will do today

- Sampling distribution
- Chapter 5
  - Bias and efficiency
  - Confidence intervals
    - For quantitative variables (e.g., income, test score, number of siblings)
    - For 0/1 categorical variables (e.g., did you vote for Romney, did you graduate from college)

Sampling Distribution

Sampling Distribution

- A sampling distribution (of sample means) is a relative frequency distribution where each observation is a sample mean)
  - Imagine we draw n (e.g., 1000) random samples
  - For each sample, we record the sample mean
  - We create a frequency distribution of sample means
    - X-axis=value; Y= number of times a particular sample mean (e.g., $\bar{y} = 1050$ is observed)
- A sampling distribution can be created for any sample statistic (e.g., mean, median, a regression coefficient)

Sampling distribution pictures

- Draw three pictures one OneNote
  - Population (we don’t see this)
  - Sample (we see this for one sample)
  - Sampling distribution (we don’t see this)
- Show Applet:
  - Very useful web application
    - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Show applet for normal population distribution
  - Show applet for skewed population distribution

Mean of sample means = population mean

- If we take random samples the value of the each sample mean, $\bar{y}$, fluctuates around the population mean, μ
- if the sample mean is found repeatedly for a large number of samples, then:
  - the overestimates of the population mean would tend to counterbalance the underestimates
  - the mean of all sample means, $\bar{y}_{\bar{y}}$, would be equal to the population mean, μ
    - $\bar{y}_{\bar{y}} = \mu$
  - Draw a picture of population distribution (label mean, std dev) over sampling distribution

Sampling Distribution

- The sampling distribution shows how the value of a sample statistic varies from sample to sample
  - For example, each Presidential Election Poll represents a single sample mean from a single random sample
  - If values from each individual poll are close to one another, then we have more faith that the sample mean from one poll is close to population mean.
  - If values from each poll are far apart, then we wouldn’t put too much faith in the sample mean from any single poll
  - Draw picture of two sampling distributions on OneNote
    - One w/ a more narrow sampling distribution than the other

Standard Error

- Standard deviation (for population)
  - Population standard deviation, σ, of a variable, y, is the average distance of an observation from the population mean, μ
- Standard error (for sampling distribution)
  - Average distance of a single sample mean, $\bar{y}$, from the mean of the sample means, $\bar{y}_{\bar{y}}$
  - Standard error, $\sigma_{\bar{y}}$, is the standard deviation of the sampling distribution.
- Draw pictures:
  - population distribution (show mean, std dev); sampling distribution (show mean, std err)

Standard Error

- Standard error, $\sigma_{\bar{y}}$
  - Average distance of a single sample mean, $\bar{y}$, from the mean of the sample means, $\bar{y}_{\bar{y}}$
  - $\sigma_{\bar{y}} = \frac{std\ dev}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$
- Ex: $\mu = 100;\ \sigma = 23; n = 100;\ \sigma_{\bar{y}} = \frac{23}{\sqrt{100}} = 2.3$
  - On average, each sample mean is 2.3 away from the population mean
- Note that standard error, $\sigma_{\bar{y}}$, is a population parameter because it depends on population standard deviation, σ
  - i.e., we usually don't know it; we will learn a sample version

Standard Error and election polls

- Why is standard error important?
  - Standard error tells us how much statistics derived from a sample are likely to diverge from population parameters
- Sample mean, $\bar{y}$, is best estimate of the population mean, μ, pct of people who will vote for Obama (e.g., $\bar{y} = 51\%$)
- Standard error provides an indication of how far away each sample mean is likely to be from the population mean
  - Standard error=10%: On average, the sample mean from each poll is likely to be 10% away from population mean
  - Standard error=2%: On average, the sample mean from each poll is likely to be 2% away from population mean
- Do we want standard error to be large or small? Why?

Properties of Standard Error

- $\sigma_{\bar{y}} = \frac{std\ dev}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$
- Standard error decreases as size of your sample increases
  - If you have a large sample size, the sample mean is likely to be close to population mean.
    - When sample means are close to close to population mean, then standard error is small
  - Example, mean income in US, with sample size of 10 versus sample size of 200
  - Show in Applet: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

- Standard error increases when standard deviation, σ, increases (i.e., $\uparrow variability\ \rightarrow \uparrow std.error$)

Central Limit Theorem

- **Central Limit Theorem**
  - For random sampling with a large sample size n, the sampling distribution of the sample mean , $\bar{y}$, is approximately normally distributed
  - Restated: no matter what the distribution of the variable, the sampling distribution will have a normal distribution
  - What is a “large” sample size
    - Agresti: n>=30 (approximately)
- Show in Applet
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Change sample sizes; use skewed distribution

In class exercise

- Play with the "applet"
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Note: "distribution of means" is sampling distribution
- (1) Choose "normal distribution";
  - click on "animated" several times to get several random samples; click on "5" to get five random samples at once; click on "1,000" to get 1,000 random samples
    - Watch how the sampling distribution changes as you add more sample means
- (2) Choose "skewed" distribution instead of "normal"
  - Repeat above exercises in (1)
- (3) Choose "skewed" distribution, select "N=25" instead of "N=5"
  - Repeat above exercises in (1)
  - How does shape of sampling distribution differ from (2)? What does this have to do with the central limit theorem

Khan Academy on Sampling Distributions

- http://www.khanacademy.org/math/statistics/v/sampling-distribution-of-the-sample-mean

Shape of distribution: 3 distributions

- Three distributions
  – Population distribution
  – Sample data distribution
  – Sampling distributions
- Draw pictures for three types of variables (assume sample size > 30)
  – Normal distribution
  – Skewed distribution
  – A “proportion” variable (i.e., a 0/1 variable such as vote for Obama or Romney)
- What is shape of sampling distribution?

Chapter 5
Statistical Inference: Estimation

Point and Interval Estimation

- Parameter
  - A summary of the population; usually unknown
- Estimates (sometimes called statistics)
  - A summary of the sample; used to make predictions about the population
  - Point estimate
    - A single number that is the best guess for the parameter (e.g., Obama approval = 46%)
  - Interval estimate
    - Interval around the point estimate, within which the parameter value is believed to fall (e.g., we are 95% sure that Obama’s approval rating lies somewhere between 44% and 48%)

Estimator vs. point estimate

- Don’t worry about this, but just for clarification:
- Estimator
  - refers to the type of statistic used to for estimating parameter (e.g., sample mean, sample median)
- Point estimate
  - Refers to the actual value of the estimator in a specific example (e.g., sample mean income is $34,000)

Properties of good estimators: Unbiased

- If an estimator is unbiased, the mean of the sampling distribution equals the parameter value
  - the overestimates would tend to counterbalance the underestimates
- For example, if the parameter is the population mean, μ, and the estimator is the sample mean, $\bar{y}$:
  - The sample mean is unbiased if the mean of sample means, $\bar{y}_{\bar{y}}$, is equal to the population mean, μ
- Biased
  - A biased estimator tends to underestimate or overestimate the value of a parameter
  - Bias often occurs because of non-random sampling or non-random missing variables
- Draw picture
  - Two population distributions, w/ biased and unbiased sampling distributions

Properties of good estimators

- Efficient estimator
  - An efficient estimator is an estimator with a low standard error
  - An efficient estimator falls closer, on average, than other estimators to the parameter.
  - The more efficient your estimator (lower standard error) the closer your estimates (e.g., sample mean $\bar{y}$) are likely to be to the parameter value (e.g., population mean μ)
  - Estimates become more precise
  - Draw picture
    - Two sampling distributions; one w/ smaller standard error than the other; smaller standard error means that each sample mean is likely to be closer to the population mean than the sampling distribution w/ larger standard error

Properties of good estimators

- What are some reasons we want estimators (e.g., sample mean) to be unbiased (mean of sampling distribution=parameter value)?
- What are some reasons we want estimators to be efficient (low standard error)?
- How do we know if the estimator is biased?
  - (think about reading empirical literature)

Interval Estimates
(e.g., Confidence Intervals)

Confidence intervals

- Method of teaching
  - Define confidence interval
  - Confidence intervals for “means” (e.g., income, test score, etc.)
    - Explain conceptually with pictures (most important)
    - Show how to calculate using formulas
  - Confidence intervals for “proportions” (variables that take on two values; e.g., vote for Obama or Romney, attend college or not)

Define confidence intervals

- Confidence interval:
  - A confidence interval for a parameter is an interval of numbers within which parameter is believed to fall (e.g., we are 95% sure that Obama’s approval rating is between 44% and 48%)
  - Has the form: point estimate $\pm$ margin of error
  - The “confidence level” (e.g., 95%, 99%) is the probability that the confidence interval contains the parameter.

Define confidence intervals

- Example
  - We are interested in population mean of "number of hours per week on internet"
  - The sample mean is the best guess of the population mean
    - Sample mean=10
  - Confidence interval says, we are 95% sure that population mean number of hours per week on the internet is between these two numbers
  - Confidence interval is the sample mean $\pm$ "some margin of error"
  - We are 95% (or 99%) sure that the population mean number of hours per week on internet is 10 ± 2
    - Alternatively, we are 95% sure that population mean number of hours on internet is between 8 and 12 hours.
  - Draw picture on OneNote

CIs: Conceptual Understanding

- Why do we need Confidence Intervals (CIs) at all? Why not just say, “the population mean is probably pretty close to the sample mean”?

- What describes how sample means vary from sample to sample?

CIs: Conceptual Understanding

- The key to a conceptual understanding of confidence intervals is integrating your understanding of these things:
  - normal distributions (and standard normal distributions)
  - sampling distributions
  - standard deviation
  - standard error
  - z-scores

CIs: Conceptual Understanding

- Sampling distribution describes how sample mean varies from sample to sample
  - By the central limit theorem, we know that sampling distributions are *normally distributed* (as long as sample size is sufficiently large)
- Standard error is a measure of average distance between sample mean and population mean
  - Standard error is the *standard deviation* of the sampling distribution
  - Remember that z-score table represents number of standard deviations from the mean
- Given that the sampling distribution is *normally distributed* and we can estimate its *standard deviation*, we can use Z-score table to find probability of observing a sample mean Z standard deviations (e.g., 2) away from population mean

CIs: Conceptual Understanding

- Draw a picture of the sampling distribution
- What percentage of observations fall within Z standard deviations of the population mean?
  - 1 std dev; 2 std dev
  - 95%; 99%

CIs: Conceptual Understanding

- We know that 95% of sample means fall within 1.96 standard deviations of the population mean
  - Equivalently, if we picked a single random sample, there is a 95% chance that the sample mean would be within 1.96 standard deviations of the population mean
- We know that 99% of sample means fall within 2.58 standard deviations of the population mean
  - Equivalently, if we picked a single random sample, there is a 99% chance that the sample mean would be within 2.58 standard deviations of the population mean

CIs: Conceptual Understanding

- The problem is, in empirical research we usually don’t know the sampling distribution
- Usually we only get to see one sample
  - That sample might be a “good draw” (sample mean is close to population mean)
  - That sample might be a “bad draw” (sample mean is far from population mean)
- We don’t know how close our sample mean is to the population mean, but theorems developed by statisticians can give us a sense

CIs: Conceptual Understanding

- What we see and what we don’t see
- Draw three pictures
  - Population
  - Sample distribution
  - Sampling distribution

CIs: Conceptual Understanding

- To calculate a 95% confidence interval:
- We take a random sample
- We imagine that the sample we take is one sample out of an infinite number of samples in the sampling distribution
- We are 95% sure that the true population mean is within 1.96 standard deviations of the sample mean that we found
- Draw picture

CIs: Conceptual Understanding

- But our 95% confidence interval might not contain the population mean….
  - If we take 100 random samples, the 95% confidence interval will not contain the population mean in about 5 of those samples
- Show picture

Calculating Confidence Interval for means

Calculating CI for means

- Formula for 95% CI
  - $\bar{X} \pm (some\ margin\ of\ error)$
  - $\bar{X} \pm 1.96 * se$
    - se= estimated standard error
- Why 1.96?
  - Assuming normal distribution, 95% of sample means fall 1.96 standard errors from population mean (from Z-score table)
- Estimated standard error, se
  - $se = \frac{s}{\sqrt{n}}$; s= sample standard deviation
  - Where $s = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}}$; usually this is given

Calculating CI for means

- Formula for 95% CI
  - $\bar{X} \pm 1.96 * se$
    - se= estimated standard error
  - $se = \frac{s}{\sqrt{n}}$; s= sample standard deviation
- Example: mean weekly hours on internet
  - $\bar{X}$=9.6; s=6; n=144
  - $\bar{X} \pm 1.96 * se = 9.6 \pm 1.96 * \frac{6}{\sqrt{144}} = 9.6 \pm 1.96 * \frac{6}{12}$
  - $= 9.6 \pm 1.96 * .5 = 9.6 \pm .98$
  - We are 95% confident that the population mean number of hours spent on the internet per week lies somewhere between 8.62 and 10.58

General formula for confidence interval

- Formula for 95% CI
  - $\bar{X} \pm (some\ margin\ of\ error)$
  - $\bar{X} \pm Z * se$
    - se= estimated standard error
    - Z= z-score (from Z-score table) associated with the desired level of confidence
- We will typically be interested in three different confidence intervals
  - 90% CI
  - 95% CI
  - 99% CI
- What are the Z-scores associated with each confidence level?

Calculating CI for means

- Formula for CI
  - $\bar{X} \pm Z * se$
  - $se = \frac{s}{\sqrt{n}}$; s= sample standard deviation
- Example: calculate 99% CI for mean weekly hours on internet (same example as befor)
  - $\bar{X}$=9.6; s=6; n=144; z=2.58 (from Z-score table)
  - $\bar{X} \pm 2.58 * se = 9.6 \pm 2.58 * \frac{6}{\sqrt{144}} = 9.6 \pm 2.58 * \frac{6}{12}$
  - $= 9.6 \pm 2.58 * .5 = 9.6 \pm 1.29$
  - We are 95% confident that the population mean number of hours spent on the internet per week lies somewhere between 8.31 and 10.89

Should you choose higher or lower CIs?

- Should you choose a 90% CI? A 95% CI? A 99% CI?
- What is the benefit of higher confidence levels?
- What is the downside of higher confidence levels?
- What confidence level should you choose?

Standard err vs. estimated standard err

- 95% CI: $\bar{X} \pm 1.96 * se$
- Why estimated (i.e., sample) standard error rather than population standard error?
  - Standard error of the sample mean, $\sigma_{\bar{y}}$
    - $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$,
    - $where\ \sigma$ is the population standard deviation
  - Estimated standard error, se
    - $se = \frac{s}{\sqrt{n}}$; s= sample standard deviation

Z-score table vs. t-score table

- Agresti uses t-score table; I think it is OK to use Z-score table
- Why use a t-score table
  - We don't know population standard dev, so we don't know exact standard error;
  - substituting sample standard error for pop std error introduces extra error in our measurements (so we want to make our CIs a little bit wider), especially when sample size is small;
  - to account for increased error, we use the t-score distribution which has fatter tails (larger proportion of obs are far away from mean)
  - T-score distribution approaches z-score distribution as sample size increases, because sample standard error becomes an increasingly precise estimate of population standard error
- I think it is OK to use Z-score table
  - In practice, Stata will do all the work for you (will account for sample size) so I don't think it is worthwhile learning about another distribution
  - Also, sample sizes usually greater than 100

In Class Exercises (answer on next page)

- Mean income of UofA higher ed graduates = \$54,000; sample size=100; sample standard deviation= 12,580
  - What is 95% CI?
  - What is 99% CI?
- Formula for CI
  - $\bar{X} \pm Z * se,$
  - $se = \frac{s}{\sqrt{n}}; \mathrm{s} = \text{sample standard dev}$

In Class Exercises: answers

- mean=$54,000; n =100; sample std dev= 12,580
  - Se= 12,580/sqrt(100)= 1258
  - 95% CI:
    - Z-score=1.96
    - CI: 54000 $\pm$ 1.96*1258
    - CI: 54000 $\pm$ 2466
    - We are 95% certain that the population mean salary for UofA higher ed graduates lies somewhere between $51,534 and $56,466
  - What is 99% CI?
    - Z-score=2.58
    - CI: 54000 $\pm$ 2.58*1258
    - CI: 54000 ± 3246
    - We are 99% certain that the population mean salary for UofA higher ed graduates lies somewhere between $50,754 and $57,246

Confidence intervals for proportions

Means vs. Proportion (for this book)

- Mean
  - Refers to a quantitative variable (income, number of siblings, number of years married, etc.)
- Proportion
  - Refers to a categorical variable with two categories (will you vote for Obama?; did you graduate from college? Are you male? Are you white?)
  - These are often called “0/1 variables” where, for example, voting for Obama=1 and not voting for Obama=0; graduating from college=1 and not graduating from college=0.
- Statistical methods for means differ from statistical methods for proportions

Notation for proportions

- Population proportion (we usually don't know)
  - $\pi = population\ proportion$ ("pi")
- Sample proportion (we know)
  - $\hat{\pi} = sample\ proportion$ ("pi hat")
- Confidence interval
  - we use the sample proportion, $\hat{\pi}$, to make a confidence interval for the population proportion, π
  - e.g., we are 95% sure that the population proportion of people who prefer Obama is between .49 and .53

Show Proportion in Stata

- IPEDS dataset of colleges and universities
- Variable called “public”: is the institution private or public
  - 0= private; 1=public
- Show histogram of population distribution
- Show frequency distribution (tabulate) and mean (summarize)
  - Note that pct of orgs that are public (i.e., public=1) is equal to the mean

CIs for proportions: Conceptual Understanding

- Variable called "Obama": Do you plan to vote for Obama
  - 0= No; 1= Yes
- Show three pictures
  - Population distribution (unknown)
  - Sample distribution (known for one sample)
  - Sampling distribution (unknown)

CIs for proportions: Conceptual Understanding

- Problem:
  - We have one sample and the sample proportion for that sample could be far away from population proportion
- Solution: We think of our sample as being randomly chosen from the sampling distribution
  - 95% of sample proportions (from the sampling distribution) will be within 1.96 standard deviations of the population proportion (show picture)
  - Equivalently, if we select a random sample and calculate the sample proportion, there is a 95% chance that the population proportion will be within 1.96 standard deviations of the sample proportion (show picture)

Calculating confidence intervals

Calculating CI for proportions

- 95% Confidence interval (CI)
  - What we want: 95% CI for the population proportion of people who say they will vote for Obama
  - 95% CI = $\hat{\pi}$ $\pm$ some margin of error
  - $\hat{\pi} \pm 1.96 * se$
  - Where $\hat{\pi}$ = sample proportion
    - In a sample of 1,000 universities, with 300 that are public, the sample proportion of public universities is =300/1,000=.3
  - se= estimated standard error
- General Confidence interval (CI)
  - $\hat{\pi} \pm z * se$
  - Where z=z=score associated with desired confidence level
    - Question: where can we find the z-scores associated with each CI?

Calculating sample std. err. For proportions

- General Confidence interval (CI)
  - $\hat{\pi} \pm z * \boldsymbol{se}$
- Population parameters
  - Standard deviation, σ, of the probability distribution
    - $\sigma = \sqrt{\pi(1-\pi)}$
  - Standard error of sample proportion, $\sigma_{\hat{\pi}}$
    - $\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}$
    - but $\sigma_{\hat{\pi}}$ uses π, which is an unknown population parameter
- Sample Statistic
  - Sample standard error of the sample proportion, se
    - $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

Calculating CI for proportions

- Confidence interval (CI)
  - $\hat{\pi} \pm z * se$, *where:*
  - $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
  - z=z-score of desired confidence level
    - Z=1.645 for 90% CI; Z=1.96 for 95% CI; Z=2.58 for 99% CI
- Recommended steps when calculating CI for proportions
  - First, calculate $\hat{\pi}$ = (# of “successes”)/n
  - Second, calculate se
  - Third, calculate confidence interval

Calculating CI for proportions, Example

- 200 people sampled; 110 say they will vote for Obama; find 95% CI
  - $\hat{\pi} = \frac{110}{200} = .55;$
  - $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{.55(1-.55)}{200}} = \sqrt{\frac{.2475}{200}} = \sqrt{.0012375} = .03518$
- Confidence interval (CI)
  - $\hat{\pi} \ \pm z * se = .55 \pm 1.96 * .03518 = .55 \pm .069$
  - We are 95% sure that the pop proportion of people who will vote for Obama lies somewhere between .481 and .619

Confidence Interval Mechanics

- Do we think a confidence interval of .481 to .619 is good enough when trying to predict the proportion of people who will vote for Obama?
- What are two ways we get “more narrow” confidence intervals?

Properties of Confidence Intervals

- Width of confidence interval decreases as sample size increases
- Width of confidence interval increases as desired confidence level increases
- Sample size considerations
  - Z-distribution is for “large” sample sizes
  - To use z-distribution to calculate CI, you sample should have at least 15 observations in each category
    - e.g., proportion vegetarian; sample must have at least 15 vegetarians and 15 non-vegetarians to use z-score table

In Class Exercise (answer on next pg)

- $\hat{\pi} \pm z * se$; same as: $\hat{\pi} \pm z * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
- Proportion on Facebook; sample=193; 90 people on facebook
  - What is 95% CI? What is 99% CI?
- Proportion on Facebook; sample=976; 423 people on Facebook
  - What is 95% CI? What is 99% CI?

In Class Exercise: Answers

- $\hat{\pi} \pm z * se;$
- sample=193; 90 people on facebook
  - $\hat{\pi}$=0.466321244; se= 0.035909049
  - 95% CI: 0.466321244 +- 0.070381736
  - 99% CI: 0.466321244 +- 0.092645346
- sample=976; 423 people on Facebook
  - $\hat{\pi}$= 0.433401639; se= 0.015862003
  - 95% CI: 0.433401639 +- 0.031089526
  - 99% CI: 0.433401639 +- 0.040923967

Assumptions and Assumption Violations

Assumptions and Assumption Violations

- Assumptions for confidence interval of a mean
  - (1) sample is a random sample from population
  - (2) population distribution of variable is normal
- “Robust”
  - A statistical method is robust with respect to a particular assumption, when it performs adequately even when that assumption is violated
  - Statisticians have shown that CI for a mean is robust against violations of normal population assumption, especially when sample size > 30

Assumptions and Assumption Violations

- Why robust to normal population assumption?
- **Central limit theorem**:
  - when sample size is large, the sampling distribution of the sample mean, $\bar{y}$, is approximately normal, even if the population distribution of the variable is not normal
  - How large is large enough?
    - If population distribution is normal then sampling distribution is normal for any sample size
    - If population distribution is not normal, sample size of about 30 is sufficient
  - Show in Applet

Assumptions and Assumption Violations

- Confidence interval for a mean is not robust to violations of random sampling (i.e., if you take a non-random sample from the population, you cannot make good predictions about the population)