Notes 7: Confidence Intervals

Julio Garín
Department of Economics

Statistics for Economics

Spring 2012

Introduction

- Why we conduct surveys?
- There are many population characteristics that we would like to obtain estimates for.
  - Examples?
  - Issues?
- What we know about $\bar{x}$ and $\bar{p}$?
- Can we obtain a "correct" answer?
  - Therefore, we will have a *margin of error*.
- The purpose of this chapter is to recognize inherent uncertainty about $\bar{x}$ and $\bar{p}$ and construct *confidence intervals*.
  - What is a "confidence interval"?
  - What is "sampling variation"?

Why Would I Care?

"(NYT 08/04/2010) – As President Obama heads home to Chicago to celebrate his 49th birthday with friends, about one-quarter of the American public still believes that he was not born in the United States, according to a CNN poll released Wednesday.

Republicans in particular doubt the president. Nationwide, they were nearly three times more likely than Democrats to say the president was not a natural born citizen. Forty-one percent of Republican respondents believe Mr. Obama was either probably or definitely born in another country, compared with just 15 percent of Democrats who held the same opinion. About 3 in 10 independents agree with Republicans.

Over all, 42 percent of those surveyed nationwide think Mr. Obama was definitely born in the United States, 29 percent say he was probably born here, 16 percent say he probably was not and 11 percent say he definitely was not.

The CNN/Opinion Research Corporation poll was conducted by telephone July 16 to July 21 with 1,018 adults across the country and has a **margin of sampling error** of plus or minus 3 percentage points."

Confidence Intervals for the Mean.

Setting Up the Problem

- Suppose you want a sample from a population to obtain information about a variable X.
  - Underlying characteristics of the population.
  - What information do you obtain from the sample?
- How do we know how the distribution of $\bar{x}$ and $\bar{p}$ looks like?
- What we would like to know?
- What are the "likely" values of μ given what we know about $\bar{x}$?
- We need to define "likely".

Constructing an Interval

- Standard assumption in economic statistics: the confidence interval specifies the likely values of μ that represent 95% of the distribution, centered on μ.
- From there, you can choose any "confidence" level.
- Trade-off between confidence and precision.
- What is the chance that μ falls outside the interval?
- We need to exploit the CLT!
- What is the question that we are trying to answer?

Case I: σ Known

- Basically, this is the problem:

$$\mu \in [\bar{x} - a, \bar{x} + a]$$

- However, $\bar{x}$ is a random variable, so, what are we really evaluating?
- Recalling some properties of a standard normal:
  - $P(z \leq b) = \Phi(b)$
  - What symmetry implies?
- Now suppose we ask the question

$$P(-b \leq z \leq b) = 0.95$$

  - Given the symmetry of the normal distribution, what this implies?
  - What is b?

Case I: σ Known

- Suppose that we have a more 'general' problem:

$$P(\mu - a \leq \bar{x} \leq \mu + a) = 0.95$$

- We need to transform the equation in something we can solve.
- Form for a 95% confidence interval when σ known:

$$\left[\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\right]$$

- What $1.96\frac{\sigma}{\sqrt{n}}$ represents?
- Remember: this assumes σ is known.

Examples

In a sample of 121 women the mean height is estimated to be 64 inches. It is known that $\sigma = 3$.

1. Construct a 95% confidence interval.

IQ scores are scaled to have a mean of 100 and a standard deviation of 16. Suppose that in a sample of 42 people $\bar{x} = 103$.

1. Construct a 95% confidence interval.

Other Confidence Intervals

- We picked 95%, but we could have specified any other value.
- Advantages and disadvantages of large confidence intervals.
- General procedure:
  - Define $1 - \alpha$ as the confidence level or confidence coefficient.
  - α is the probability the true value does not equal the values in the CI.
  - Define $z_{\alpha/2}$ as the value of the standard normal such that $P(z > z_{\alpha/2}) = \frac{\alpha}{2}$.

The interval estimate of a population mean with σ known is

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $(1 - \alpha)$ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution.

Examples (Variations on a Theme...)

In a sample of 121 women the mean height is estimated to be 64 inches. It is known that $\sigma = 3$.

1. Construct a 99% confidence interval.
2. Construct a 90% confidence interval.

Case II: σ Unknown

- You could say "Julio, these calculations are all fine and dandy, you are awesome, I wish I could take this class next semester just to listen to your Shakespearean English again, and, by the way, I love your green shirt, but, what if you do not know σ?"
- What if we do not know σ?
- You should have demonstrated that s^2 is an unbiased estimate of σ^2.
- Remember our problem:

$$P(\mu - a \leq \bar{x} \leq \mu + a) = 0.95$$

- How should we proceed from there?
- Easy: CLT.
- What are the implications of dividing $\bar{x}$ over s?

$\frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a student's t-distribution with degrees of freedom equal to $n-1$.

A Little bit of Intuition

- Usually we construct

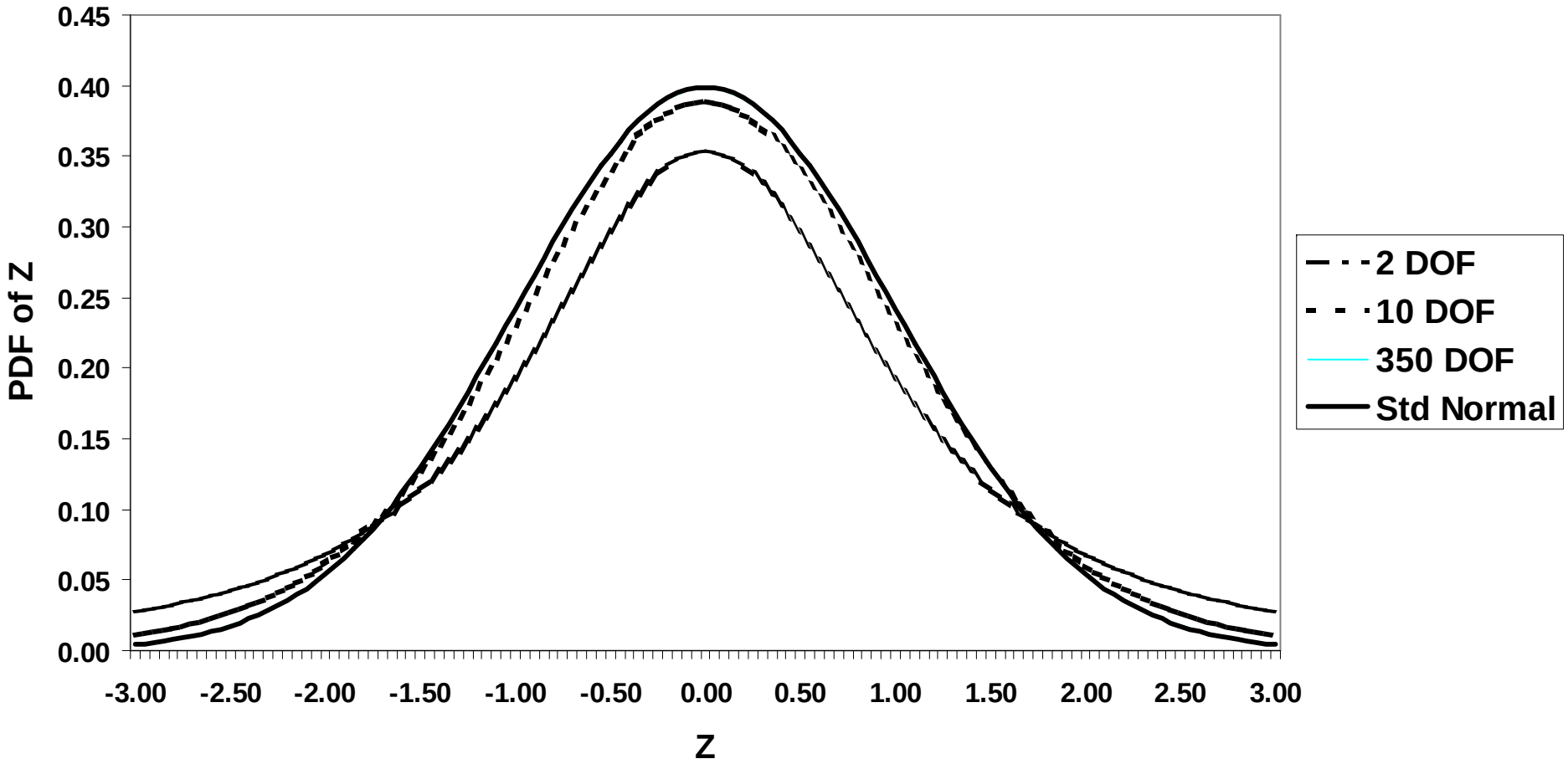
$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

  - What are the implications for the distribution of $\bar{x}$?
- Now we have

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

  - This generates a student's t-distribution.
    - It is a symmetric unimodal distribution.
    - Mean and median are zero.
    - Similar to a normal but with fatter 'tails'.
    - It is a function of the "degrees of freedom" (dof).
    - As dof increase, it looks more like a normal distribution.

t-distribution



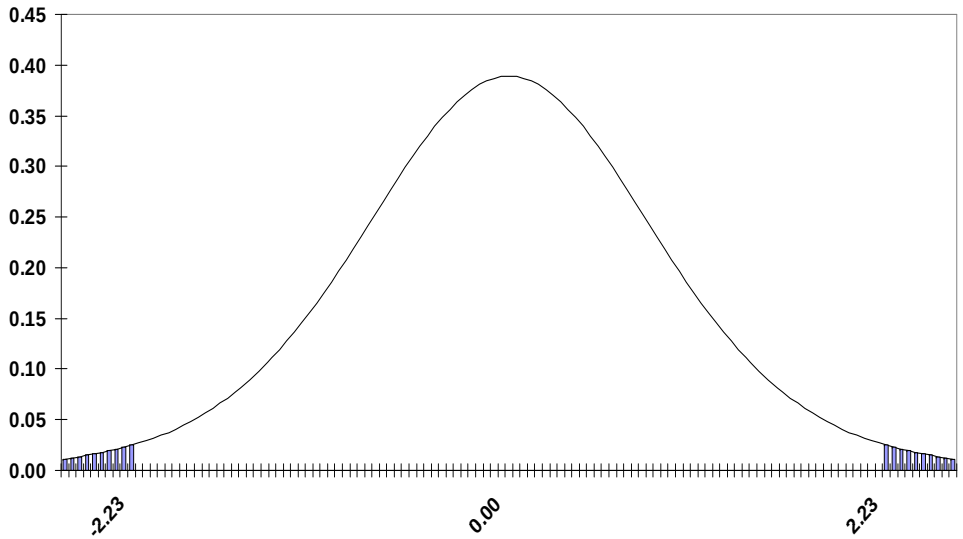
William Sealy Gosset

How to Use the t-distribution?

- As before, we will exploit the symmetry.
- Define the confidence level.
- Now check $t_{\alpha/2,\nu}$ that is the value of the t-distribution where only $\alpha/2$ percent of the distribution lies above the value and $\alpha/2$ percent lies below $-t_{\alpha/2,\nu}$.

t-distribution With 10 DOF and α=0.05

area in right hand tail of distribution					
dof	0.25	0.10	0.05	0.025	0.01
1	1.00	3.08	6.31	12.7	31.8
2	0.82	1.89	2.92	4.30	6.96
3	0.76	1.64	2.35	3.18	4.54
4	0.74	1.53	2.13	2.78	3.75
5	0.73	1.48	2.02	2.57	3.36
6	0.72	1.44	1.94	2.45	3.14
7	0.71	1.41	1.89	2.36	3.00
8	0.71	1.40	1.86	2.31	2.90
9	0.70	1.38	1.83	2.26	2.82
10	0.70	1.37	1.81	2.23	2.76
20	0.69	1.33	1.72	2.09	2.53
30	0.68	1.31	1.70	2.04	2.46
40	0.68	1.30	1.68	2.02	2.42
60	0.68	1.30	1.67	2.00	2.39
120	0.68	1.29	1.66	1.98	2.36



How to Use the t-distribution?

The interval estimate of a population mean with σ unknown is

$$\bar{x} \pm t_{\alpha/2,\nu} \frac{s}{\sqrt{n}}$$

where $(1-\alpha)$ is the confidence coefficient and $t_{\alpha/2,\nu}$ is the t value providing an area of $\alpha/2$ in the upper tail of the student's t-distribution with ν degrees of freedom.

- Remember when σ was known:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- We have that, for small samples:

$$t_{\alpha/2,\nu} > z_{\alpha/2}$$

- Why is that?
  - It is the cost of our ignorance!

Examples

In a sample of 121 women the mean height is estimated to be 64 inches and $s = 3$.

1. Construct a 95% confidence interval.
2. Compare with the case in which σ was known.

In a random sample of 10's students scores for a very cool Stats class the mean score is 80.1 with $s = 11.4$.

1. Construct a 90% confidence interval.
2. Construct a 95% confidence interval.
3. Construct a 99% confidence interval.

A Case Study

Nabisco, the maker of Chips Ahoy! cookies challenged students across the nation to confirm the claim that there are at least 1,000 chocolate chips in every 18-ounce bag. They offered $25,000 in scholarship if students could verified the claim. Nabisco defines a chocolate chip as: *any distinct piece of chocolate that is baked into or on top of the cookie dough regardless of whether or not it is 100% whole.*
An introductory statistics class from the United States Air Force Academy designed and carried out the following experiment:

- Friends and family of the cadets sent 275 bags of Chips Ahoy! cookies from all over the country.
- The class randomly selected 42 bags from the 275 bags.
- Cookies from each of the 42 bags were dissolved in water to separate the chips, and then the number of chips counted for each bag.

They obtained a mean of 1,261 with a standard deviation of 117.6

1. Construct a 95% confidence interval.
2. What can you say about Nabisco's claim?

Confidence Intervals for the Population Proportion.

Case III: Population Proportion

In a sample of 29 ND graduate students, 20.68% say they smoke (6 students).

- How would you go about constructing a 95% confidence interval?
  - The variable is a discrete outcome, isn't?
  - Therefore, the response can be thought of as an outcome from a binomial process: let X_i be the response of student i and use dummy variables.
  - How can we obtain the true underlying fraction of people that smoke?
  - Since we do not know p, we need to estimate of $\sigma_{\bar{p}}$
- So...what would be the 95% confidence interval?

The interval estimate of a population proportion is

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where $(1-\alpha)$ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution.

Margin of Error and Sample Size.

Margin of Error: Example I

"(NYT 08/04/2010) – As President Obama heads home to Chicago to celebrate his 49th birthday with friends, about one-quarter of the American public still believes that he was not born in the United States, according to a CNN poll released Wednesday.

Overall, 42 percent of those surveyed nationwide think Mr. Obama was definitely born in the United States, 29 percent say he was probably born here, 16 percent say he probably was not and 11 percent say he definitely was not.

The CNN/Opinion Research Corporation poll was conducted by telephone July 16 to July 21 with 1,018 adults across the country and has a margin of sampling error of plus or minus 3 percentage points."

1. Where do they get 'plus or minus 3 percentage points'?

Margin of Error: Example II

"(NYT 10/29/2010) – The Florida governor's race looks like a tossup heading into Election Day. A Quinnipiac University poll released Thursday shows the Democrat Alex

Sink and the Republican Rick Scott locked in a tight race, with Ms. Sink supported by 45 percent of likely voters and Mr. Scott by 41 percent. The poll was conducted Oct.

18 to 24 among 784 likely voters and has a margin of sampling error of plus or minus 3.5 percentage points."

1. Where do they get 'plus or minus 3.5 percentage points'?

Determining the Sample Size for a Desired Margin of Error

- We could have two cases:
  1. CI for the mean with σ known.
  2. CI for the population proportion.
- *Les doigts dans le nez*.
- Remember:
  - For the population mean with σ known:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

  - For the population proportion:

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{p^*(1-p^*)}{n}}$$

- You need to specify a level of error and and just solve for n!

Examples

How Large a sample should be selected to provide a 95% confidence interval with a margin of error of 10? Assume that the population standard deviation is 40.

At 95% confidence, how large a sample should be taken to obtain a margin of error of 0.03 for the estimation of a population proportion? Assume past data are not available for developing a planning value for p^*.

Confidence Intervals for Means From Two Independent Samples.

Nature of the Problem

- We will use similar techniques as before.
- What is the motivation?
  - We have $\bar{x}_1$ and $\bar{x}_2$.
  - Both are r.v. So the difference will be a r.v..
- Once we know the standard error of the difference, we can use the same techniques from the last two classes.

Set Up of the Problem

- Start by collecting data from two independent samples.
  - n_1 observations from sample 1.
  - n_2 observations from sample 2.
- Calculate their sample mean.
  - $\bar{x}_1$= sample mean for sample 1.
  - $\bar{x}_2$= sample mean for sample 1.
- How do we measure the differences in the two samples?
  - Create a new variable: $\delta = \bar{x}_1 - \bar{x}_2$

Problem Set Up

- What we want is $d = \mu_1 - \mu_2$.
  - The true difference in expected values.
- What we can estimate is: $\Delta = \bar{x}_1 - \bar{x}_2$.
  - The difference in mean across samples.
- The difference in means across samples is an unbiased estimate of the true difference in expected values across samples.
- What do we know about Δ?
  - Mean?
  - Variance?

A New Concept: Pooled Variance

Under the assumption that the two population standard deviations are equal ($\sigma_1 = \sigma_2 = \sigma_p$), the two sample standard deviation are combined to provide the following *pooled sample variance*:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

- What it represents?
- Where this came from?
  - Remember when we went over linear combinations of random variables?
- What would be the estimated variance?

$$\sigma_\Delta^2 = s_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]$$

- Let that value be se_Δ.

Constructing a Confidence Interval

- As before should start by

$$d = \Delta \pm a$$

- Then, standardize the probability statement:

$$P(d - a \leq \Delta \leq d + a) = 0.95$$

- It can be shown that:

$$\frac{\Delta - d}{se_d}$$

has a student's t-distribution with degrees of freedom equal $n_1 + n_2 - 2$.

- Therefore, the 95% confidence interval is

$$d = \Delta \pm a = \pm t_{0.025, n_1+n_2-2} s_p \sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Example: Who Pays Your Bills?

In a survey of college students with credit cards, 491 reported paying the balance themselves, while parent paid the balance for 769. Balance was higher for those whose parents paid their credit cards.

$$n_1 = 769 \qquad \bar{x}_1 = 1,366 \qquad s_1 = 600$$

$$n_2 = 491 \qquad \bar{x}_2 = 968 \qquad s_2 = 540$$

1. Construct a 95% CI for the difference in balances.
2. Interpret findings.