

STA 291 Summer 2008

Lecture 9

STA 291 - Lecture 9

1

Review: Point Estimator

- A point estimator of a parameter is a sample statistic that predicts the value of that parameter
- A good estimator is
 - **Unbiased**: Centered around the true parameter
 - **Consistent**: Gets closer to the true parameter as the sample size gets larger
 - **Efficient**: Has a standard error that is as small as possible

STA 291 - Lecture 9

2

Review: Confidence Interval

- The sampling distribution of the sample mean \bar{X} has mean m and standard error

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- If n is large enough, then the sampling distribution of \bar{X} is approximately normal/bell-shaped (Central Limit Theorem)

STA 291 - Lecture 9

3

Confidence Interval: Derivation

- With 95% probability, the sample mean falls in the interval between

$$m - 1.96 \frac{s}{\sqrt{n}} \text{ and } m + 1.96 \frac{s}{\sqrt{n}}$$

(m = population mean, s = population standard deviation)

- This is equivalent to: With 95% probability, the following **random** interval contains the population mean

$$\bar{X} - 1.96 \frac{s}{\sqrt{n}} \text{ and } \bar{X} + 1.96 \frac{s}{\sqrt{n}}$$

STA 291 - Lecture 9

4

Confidence Interval: Practical Formula

- In practice, the population standard deviation is unknown and needs to be estimated by the sample standard deviation s
- A large sample 95% confidence interval for the population mean m is

$$\bar{X} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

- where \bar{X} is the sample mean and
- s is the sample standard deviation

STA 291 - Lecture 9

5

Confidence Interval

- Example: Find and interpret the 95% confidence interval for the population mean, if the sample mean is 70 and the sample standard deviation is 10, based on a sample of size

1. $n = 25$
2. $n = 100$

- 1.) $70 \pm 3.98 = [66.02, 73.98]$
- 2.) $70 \pm 1.96 = [68.04, 71.96]$

STA 291 - Lecture 9

6

Confidence Interval: Interpretation

- If we repeatedly took random samples using the same method, then, in the long run, in 95% of the cases, the confidence interval will cover the true unknown parameter
- For one given sample, we do not know whether the confidence interval covers the true parameter
- The **95% probability** only refers to the **method** that we use, but not to the individual sample/interval

Confidence Interval: Interpretation

- Wrong statement (referring to *one concrete* interval):
"With 95% probability, the population proportion is in the interval from 0.4 to 0.6"
- To avoid misleading use of the word "probability", we say:
"We are 95% confident that the true population proportion is in this interval"

Confidence Intervals

- In general, a large sample confidence interval for the mean μ has the form

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}}$$

- Where z is chosen such that the probability under a normal curve within z standard deviations equals the confidence coefficient

Different Confidence Coefficients

- We can use Table B3 to construct confidence intervals for other confidence coefficients
- For example, there 99% probability of a normal distribution is within 2.58 standard deviations of the mean
($z=2.58$, tail probability = 0.005)
- A 99% confidence interval for μ is

$$\bar{X} \pm 2.58 \cdot \frac{s}{\sqrt{n}}$$

Error Probability

- The error probability (α) is the probability that a confidence interval does **not** contain the population parameter
- For a 95% confidence interval, the error probability $\alpha=0.05$
- $\alpha = 1 - \text{confidence level}$ or
confidence level = $1 - \alpha$
- The error probability is the probability that the sample mean \bar{X} falls more than z standard errors from μ (in both directions)
- The confidence interval uses the z -value corresponding to a one-sided tail probability of $\alpha/2$

Different Confidence Coefficients

Confidence Coefficient	α	$\alpha/2$	z
90%	0.1		
95%			1.96
98%			
99%			2.58
			3
			4

Facts About Confidence Intervals I

- The width of a confidence interval
 - Increases as the confidence coefficient increases
 - Increases as the error probability decreases
 - Increases as the standard error increases
 - Decreases as the sample size increases

Facts About Confidence Intervals II

- If you calculate a 95% confidence interval, say from 10 to 14, there is **no probability associated** with the true unknown parameter being in the interval or not
- The true parameter is either in the interval from 10 to 14, or not – we just don't know it
- The 95% refers to the method: If you repeatedly calculate confidence intervals with the same method, then 95% of them will contain the true parameter

Choice of Sample Size

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm B$$

- So far, we have calculated confidence intervals starting with z , s , n
- These three numbers determine the precision B of the confidence interval
- Now we reverse the equation:
 - We specify a desired precision B (bound ; margin of error)
 - Given z and s , we can find the minimal sample size needed for this precision

Choice of Sample Size

- Let's start with the confidence equations that include the unknown population standard deviations

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm B$$

- Mathematically, we need to solve the above equation for n
- The result is

$$n = s^2 \cdot \left(\frac{z}{B} \right)^2$$

Example

- For a random sample of 100 UK employees, the mean distance to work is 3.3 miles and the standard deviation is 2.0 miles
- Find and interpret a 90% confidence interval for the mean residential distance from work of all UK employees
- About how large a sample would have been adequate if we merely needed to estimate the mean to within 1.0, with 90% confidence?

Confidence Interval for a Proportion

- Examples:
 - Proportion of international students at UK
 - Proportion of registered voters who will vote for candidate A in the presidential election
 - Proportion of Kentucky families with income below the poverty level

Confidence Interval for a Proportion

- The sample proportion \hat{p} is an unbiased and efficient point estimator of the population proportion p
- The proportion is a special case of the mean
- Therefore, we can use the formula for the confidence interval for the mean also for proportions
- We only need to replace \bar{x} by \hat{p} and s by a different estimator of the standard deviation

Confidence Interval for a Proportion

- A large sample confidence interval for the population proportion p has the form

$$\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Where \hat{p} is the sample proportion

Example

- In a recent telephone survey (conducted in mid-October), people were asked whether they have seen a ghost or felt its presence.
- Of 1013 adults interviewed, 230 answered *yes*, and 783 answered *no*.
- Find the point estimate of the *population proportion* of adults who would answer *yes*
- Construct and interpret a 95% confidence interval for the population proportion.
- Can you conclude that fewer than half of the population has seen a ghost or felt its presence?

Choice of Sample Size

- Given a desired margin of error B and a confidence coefficient with corresponding z -score,
- The minimal sample size needed to guarantee that precision of the confidence statement is

$$n = \hat{p} \cdot (1 - \hat{p}) \left(\frac{z}{B} \right)^2$$

- Using this formula requires guessing \hat{p} before taking the sample, or taking the safe but conservative approach of setting

$$\hat{p} = 0.5 \text{ which results in } \hat{p} \cdot (1 - \hat{p}) = 0.25$$

- This is like a "worst case scenario" because the product can not exceed 0.25

Example

- To estimate the proportion of traffic deaths in Florida last year that were alcohol related, determine the necessary sample size for the estimate to be accurate to within 0.06 with probability 0.90.
- Based on results of a previous study, we expect the proportion to be about 0.30.

Typical Question

- Based on a sample of $n=1000$ people, a 95% confidence interval for the population proportion of people voting for candidate A is calculated. It turns out to be from 67% to 73%.
- What does "95% confidence" mean?
- The confidence interval (0.67, 0.73) either does or does not contain the population proportion. We don't know whether it does.
- We are 95% confident that the true population proportion is between 67% and 73%.
- That is, if we repeatedly selected random samples of the same size and each time constructed a 95% confidence interval, then in the long run about 95% of the intervals would contain the true, unknown population proportion.

Multiple Choice Question
Which of the following statements are true?

- "95% confidence" means that
 1. 95% of the true population parameters are in the confidence interval
 2. If we were to repeat the procedure of sampling and calculating confidence intervals from the same population, then 95% of the times our confidence interval will contain the true population parameter
 3. If we were to repeat the procedure of sampling and calculating confidence intervals from the same population, then 95% of the population parameters are going to be in every calculated interval

Multiple Choice Question
Which of the following statements are true?

- "If we calculate a specific confidence interval based on a sample (say the interval turns out to be from 2.6 to 4.6), then
 1. The true population parameter is in this interval with 95% probability
 2. We do not know whether the true population parameter is in this interval or not
 3. 95% of the time, the interval will be from 2.6 to 4.6.

Confidence Interval for Population Mean

- A large sample (1-alpha) confidence interval for the mean μ has the form

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}}$$

- Where z is chosen such that there is (1-alpha) probability under a normal curve within z standard deviations from the mean

**Confidence Interval for Population Mean
Choice of Sample Size**

- The confidence interval for the population mean has the form
sample mean plus/minus margin of error

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm B$$

- Given z and s , and a desired precision B (bound ; margin of error)
- the minimal sample size needed for this precision is

$$n = s^2 \cdot \left(\frac{z}{B}\right)^2$$

Summary: Confidence Intervals for Population Mean and Population Proportion

- Confidence Interval Formulae

$$\bar{X} \pm z \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm B \quad \hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm B$$

- Minimal sample size formulae

$$n = s^2 \cdot \left(\frac{z}{B}\right)^2 \quad n = \hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{z}{B}\right)^2$$

- Replace \hat{p} by 0.5 if no reasonable initial guess of \hat{p} is possible.

Student T adjustment

- In practice, the population standard deviation is (almost always) unknown and needs to be estimated by the sample standard deviation s
- A 95% confidence interval for the population mean is μ

$$\bar{X} \pm ?? \cdot \frac{s}{\sqrt{n}}$$

- where \bar{X} is the sample mean and
- s is the sample standard deviation

"Student" t - adjustment

- If sigma is unknown, we may replace it by s (sample SD) but the value Z (for example 1.96) needs adjustment to take into account of extra variability introduced by s
- There is another table to look up: t-table or another applet
- http://www.socr.ucla.edu/Applets.dir/Normal_T_ChI2_F_Tables.htm

Degrees of freedom, n-1

- Student t - table with infinite degrees of freedom is same as Normal table
- When degrees of freedom is over 200, the difference is very small

- The number ?? to replace 1.96 can be found from table 4 on page B-9
- We need two piece of info:
 - (1) what was the number if σ were known? [Here it is 1.96]
 - (2) sample size n. (n-1) called degrees of freedom.
- (1) tells us which column to look, (2) tells us which row to look

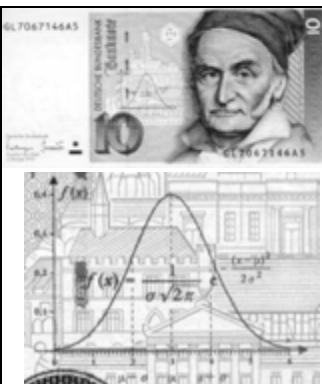
Confidence Coefficients

- In general, a confidence interval for the mean μ has the form

$$\bar{X} \pm z \cdot \frac{S}{\sqrt{n}} \quad \text{or} \quad \bar{X} \pm t \cdot \frac{S}{\sqrt{n}}$$

- Error probability: $\alpha = 1 - \text{confidence coefficient}$
- The confidence interval uses the z-value corresponding to a tail probability of $\alpha/2$

Gauss and "Student"



11.1 Significance Tests

- A significance test checks whether data agrees with a hypothesis
- A hypothesis is a statement about a characteristic of a variable or a collection of variables
- If the data is very unreasonable under the hypothesis, then we will reject the hypothesis
- Usually, we try to find evidence **against** the hypothesis

Logical Procedure

1. State a hypothesis that you would like to find evidence against
2. Get data and calculate a statistic (for example: sample mean)
3. The hypothesis (for example: population mean equals 5) determines the sampling distribution of our statistic
4. If the calculated value in 2. is very unreasonable given 3., then we conclude that the hypothesis was wrong

STA 291 - Lecture 9

37

Example

- Somebody makes the claim that "50% of all UK students wear sandals to class if it is sunny and at least 70 degrees"
- You don't believe it, so one of those days, you take a random sample of 10 students, and find that only 2 out of these 10 students actually wear sandals
- How unlikely is this under the hypothesis?
- The sampling distribution helps us quantify the unlikeliness in terms of a probability (p -value)

STA 291 - Lecture 9

38

Significance Test

- A **significance test** is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis
- Data that fall far from the predicted values provide **evidence against the hypothesis**

STA 291 - Lecture 9

39

Elements of a Significance Test

- Assumptions
- Hypotheses
- Test Statistic
- P-value
- Conclusion

STA 291 - Lecture 9

40

Assumptions

- What type of data do we have?
 - Qualitative or quantitative?
 - Different types of data require different test procedures
- What is the population distribution?
 - Is it normal? Symmetric?
 - Some tests require normal population distributions
- Which sampling method has been used?
 - We usually assume simple random sampling
- What is the sample size?
 - Some methods require a minimum sample size (like $n=25$)

STA 291 - Lecture 9

41

Assumptions in the Example

- What type of data do we have?
 - Qualitative with two categories: Either "wearing sandals" (1) or "not wearing sandals" (0)
- What is the population distribution?
 - Discrete, taking the two values 0 and 1
- Which sampling method has been used?
 - We assume simple random sampling
- What is the sample size?
 - $n=10$

STA 291 - Lecture 9

42

Hypotheses

- The **null hypothesis (H_0)** is the hypothesis that we test (and try to find evidence against)
- The name null hypothesis refers to the fact that it often (not always) is a hypothesis of “no effect” (no effect of a medical treatment, no difference in characteristics of countries, etc.)
- The **alternative hypothesis (H_1)** is a hypothesis that contradicts the null hypothesis
- When we reject the null hypothesis, the alternative hypothesis is judged acceptable
- Often, the alternative hypothesis is the actual research hypothesis that we would like to “prove” by finding evidence against the null hypothesis (proof by contradiction)

Hypotheses in the Example

- **Null hypothesis (H_0):**
50% of all UK students wear sandals to class if it is sunny and at least 70 degrees
 H_0 : Population proportion = 0.5
- **Alternative hypothesis (H_1):**
The proportion of UK students wearing sandals is different from 0.5

Test Statistic

- The **test statistic** is a statistic that is calculated from the sample data
- Often, the test statistic involves a point estimator of the parameter about which the hypothesis is stated
- For example, the test statistic may involve the sample mean or sample proportion if the hypothesis is about the population mean or population proportion

Test Statistic in the Example

- **Test statistic:**
Sample proportion,
 $2/10=0.2$

p -Value

- How unusual is the observed test statistic when the null hypothesis is assumed true?
- The **p -value** is the probability, assuming that H_0 is true, that the test statistic takes values at least as contradictory to H_0 as the value actually observed
- The smaller the p -value, the more strongly the data contradict H_0

p -Value in the Example

- The sampling distribution for the sample proportion when the true population proportion is 0.5 is (similar to Binomial)

.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.001	.01	.04	.12	.21	.25	.21	.12	.04	.01	.001

- At least as contradictory as the observed “2” are all the proportions .0, .1, .2, .8, .9, 1.0 that are at least as far away from 0.5 as 0.2

p -Value in the Example (contd.)

- We obtain the p -value by adding up the respective probabilities

.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.001	.01	.04	.12	.21	.25	.21	.12	.04	.01	.001

- $0.001+0.01+0.04+0.04+0.01+0.001=0.1=10\%$
- If truly 50% of all the UK students wear sandals, then the chance is 10% that a sample is at least as extreme as “2 out of 10”

STA 291 - Lecture 9

49

p -Value in the Example (contd.)

- What would be the p -value if the sample proportion was 0.1?
- What if the sample proportion was 1?

STA 291 - Lecture 9

50

Conclusion

- Sometimes, in addition to reporting the p -value, a formal decision is made about rejecting or not rejecting the null hypothesis
- Most studies require small p -values like $p < 0.05$ or $p < 0.01$ as significant evidence against the null hypothesis
- “The results are significant at the 5% level”

STA 291 - Lecture 9

51

Conclusion in the Example

- We have calculated a p -value of $0.1=10\%$
- This is not significant at the 5% level
- So, we cannot reject the null hypothesis (at the 5% level)
- So, do we believe the claim that the proportion of UK students wearing sandals is truly 50%?

STA 291 - Lecture 9

52