STA 321 Spring 2014

Lecture 13 *Tuesday, March 11th*

- **Confidence Intervals**

Two Types of Estimators

- Point Estimate
  - A single number that is the best guess for the parameter
  - For example, the sample mean is usually at good guess for the population mean
- Interval Estimate
  - A range of numbers around the point estimate
  - To give an idea about the precision of the estimator
  - For example, “the proportion of people voting for candidate A is between 67% and 73%”

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

Unbiased

- An estimator is unbiased if its sampling distribution is centered around the true parameter
- For example, we know that the mean of the sampling distribution of $\overline{Y}$ equals mu, which is the true population mean
- So, $\overline{Y}$ is an unbiased estimator of mu

Unbiased

- However, for any particular sample, the sample mean $\overline{Y}$ may be smaller or greater than the population mean
- "Unbiased" means that there is no systematic under- or overestimation
- If you repeatedly took samples, then the average of the sample means would converge to the population mean

Biased

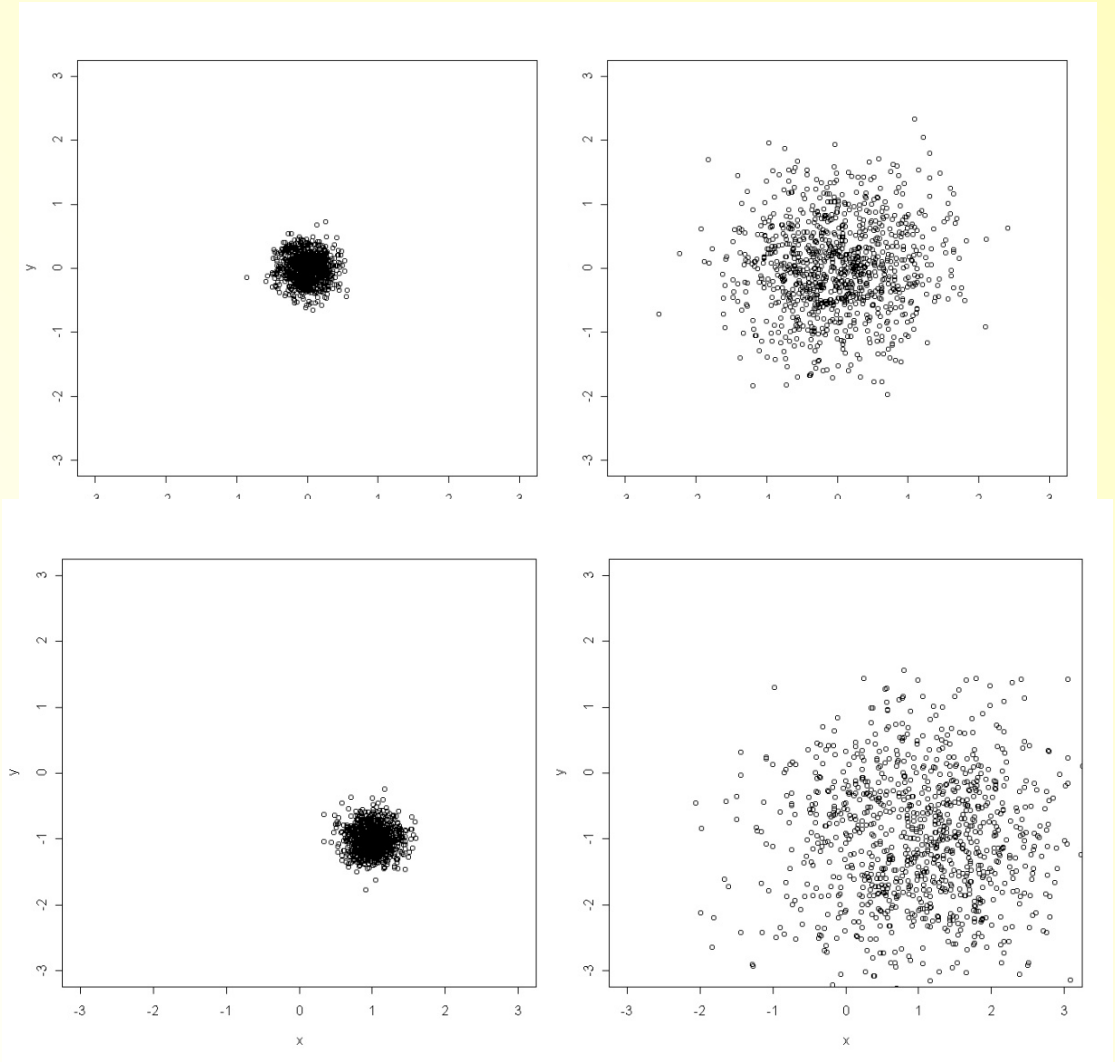
- A biased estimator systematically under- or overestimates the population parameter
- The definition of sample variance and sample standard deviation uses *n-1* instead of *n*, because this makes the variance estimator unbiased
- With *n* in the denominator, it would systematically underestimate the variance

Efficiency

- An estimator is efficient if its standard error is small compared to other estimators
- Such an estimator has high precision
- A good estimator has ***small standard error and small bias*** (or no bias at all)

- The following pictures represent different estimators with different bias and efficiency
- Assume that the true population parameter is the point (0,0) in the middle of the picture

Bias and Efficiency



Note that even an unbiased and efficient estimator does not always hit exactly the population parameter.

But in the long run, it is the best estimator.

Point Estimators of the Mean, Median, and Standard Deviation

- The sample mean is unbiased, consistent, and sometimes relatively efficient
- It is the most efficient estimator when the population distribution is normal (can be proved mathematically)
- The sample median is more efficient for many skewed and "heavy-tailed" distributions
- The sample variance is unbiased when we use *n-1* in the denominator
- It is also consistent (and in some situations relatively efficient)

Example: Three Estimators

- Suppose we want to estimate the proportion of UK students voting for candidate A
- We take a random sample of size $n=100$
- The sample is denoted $Y_1, Y_2, \ldots, Y_n$, where $Y_i=1$ if the ith student in the sample votes for A, $Y_i=0$ otherwise

Example: Three Estimators

- Estimator 1 = the sample mean (sample proportion)
- Estimator 2 = the answer from the first student in the sample (Y_1)
- Estimator 3 = 0.3
- Which estimator is unbiased?
- Which estimator is consistent?
- Which estimator has high precision (small standard error)?

Confidence Interval

- An inferential statement about a parameter should always provide the probable accuracy of the estimate
- How close is the estimate likely to fall to the true parameter value?
- Within 1 unit? 2 units? 10 units?
- This can be determined using the sampling distribution of the estimator/ sample statistic
- In particular, we need the standard error to make a statement about accuracy of the estimator

Confidence Interval

- Example: If the sample size is at least n=25, then with 95% probability, the sample mean falls between

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

(μ = population mean,

σ = population standard deviation)

Confidence Interval

- A confidence interval for a parameter is a range of numbers within which the true parameter likely falls
- The probability that the confidence interval contains the true parameter is called the confidence coefficient
- The confidence coefficient is a chosen number close to (but smaller than) 1, usually 0.95 or 0.99

Confidence Interval

- The sampling distribution of the sample mean $\bar{Y}$ has mean μ and standard error

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

- If *n* is large enough, then the sampling distribution of $\bar{Y}$ is approximately normal/ bell-shaped

  (Central Limit Theorem)

Confidence Interval

- To calculate the confidence interval, we usually use the Central Limit Theorem (unless the variable already has a normal population distribution)
- Therefore, we typically need sample sizes of at least about $n=25$
- Also, we need a z-score that is determined by the confidence coefficient
- Let's choose 0.95, then $z=1.96$

Confidence Interval

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

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

(μ = population mean, σ = population standard deviation)

- Whenever the sample mean falls within 1.96 standard errors from the population mean, the following interval contains the population mean

$$\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}$$

Confidence Interval

- So, the ***random*** interval between

$$\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}$$

contains the population mean
with 95% probability

- This is a confidence statement, and the interval is called a 95% confidence interval
- In practice, the population standard deviation is unknown and has to be replaced by its unbiased estimator, the sample standard deviation *s*

Sampling Distribution of the Sample Proportion

- ***Center:*** p
- ***Spread:*** $\sqrt{\frac{p(1-p)}{n}}$

  also called the *standard error* of $\hat{p}$

- ***Shape:*** As the sample size increases, the distribution of the $\hat{p}$'s gets closer and closer to the normal.

Sampling Distribution of the Sample Mean

- ***Center:*** μ.
- ***Spread:*** $\frac{\sigma}{\sqrt{n}}$

  also called the *standard error* of $\bar{Y}$

- ***Shape:*** As the sample size increases, the distribution of the $\bar{Y}$'s gets closer and closer to a normal curve.

Confidence Interval

- A large sample 95% confidence interval for the population mean μ is

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

  - where $\bar{Y}$ is the sample mean and
  - *s* is the sample standard deviation

Confidence Interval: Interpretation

- "Probability" means that "in the long run, 95% of these intervals would contain the parameter"
- 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

Confidence Interval: Interpretation

- To avoid the sometimes misleading word "*probability*", we say:

  "We are 95% *confident* that the true population mean is in this interval"

http://www.amstat.org/publications/jse/v6n3/applets/ConfidenceInterval.html

Confidence Interval

- If we change the confidence coefficient from 0.95 to 0.99, the confidence interval changes
- Increasing the probability that the interval contains the true parameter requires increasing the length of the interval
- In order to achieve 100% probability to cover the true parameter, we would have to take the whole range of possible parameter values, but that would not be informative
- Tradeoff between precision and coverage probability
- *More coverage probability = less precision*

Example

- Find and interpret the 95% confidence interval for the population mean, if the sample mean is 50 and the sample standard deviation is 15, based on a sample of size
  1. *n = 25*
  2. *n = 100*

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

Example: Sleep

- How much do Americans sleep each night?
- A random sample of 1120 Americans (15y and older) had a mean amount of sleep per night of 7.67 hours, and the standard deviation was 1.2 hours.
- Construct and interpret a 95% confidence interval for the mean amount of sleep per night of Americans 15 years of age and older.

Different Confidence Coefficients

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

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

- 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 the online tools to construct confidence intervals for other confidence coefficients
- For example, for every normal distribution, there is 99% probability within 2.58 standard deviations of the mean

  (*z=2.58, tail probability =0.005)*
- A 99% confidence interval for μ is

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

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 α=0.05
- α = 1 - confidence coefficient or
- confidence coefficient = 1 – α
- The error probability is the probability that the sample mean $\overline{Y}$ falls more than *z* standard errors from μ (in both directions)

Different Confidence Coefficients

Confidence Coefficient	α	z
90%	0.1	
95%		1.96
98%		
99%		2.59
		3
		4

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 $\bar{Y}$ 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 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

Confidence Interval for a Proportion

- Where does the new standard error come from?
- The sample proportion $\hat{p}$ has a sampling distribution with mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$
- Of course, in practice, we don't know the population proportion p

  (otherwise, we would not need a confidence interval for it)
- So, we replace p by $\hat{p}$ in the formula for the standard error

Confidence Intervals for Mean and Proportion

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

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

Example: Relaxing

- In a recent Gallup poll, 455 of 1011 randomly selected adults aged 18 years and older said that they had too little time for relaxing or doing nothing.
- Construct and interpret a 95% confidence interval for the population proportion.
- Construct a 99% confidence interval and compare.

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.

Calculating *z*-Scores

1. **z-Score for an individual observation**
   - You need to know *Y*, *mu*, and sigma to calculate *z*

$$z = \frac{Y - \mu}{\sigma}$$

2. **z-Score for a sample mean**
   - You need to know *Y-bar*, *mu*, *sigma*, and *n* to calculate *z*

$$z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

3. **z-Score for a sample proportion**
   - You need to know *p-hat*, *p*, and *n* to calculate *z*

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

Confidence Intervals for Mean and Proportion

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

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

Confidence Interval: Interpretation

- For a 95% confidence interval:
- 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*** refers to the ***method*** that we use, but not to the individual sample

Example: Sleep

- How much do Americans sleep each night?
- A random sample of 1120 Americans (15y and older) had a mean amount of sleep per night of 7.67 hours, and the standard deviation was 1.2 hours.
- Construct and interpret a 95% confidence interval for the mean amount of sleep per night of Americans 15 years of age and older.

Choice of Sample Size

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

- So far, we have calculated confidence intervals starting with z, s, n or $z, \hat{p}, n$
- 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, or z and $\hat{p}$, we can find the minimal sample size needed for this precision by solving the above equations for n

Choice of Sample Size

- The results are

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

- However, in practice, we usually don't know s or $\hat{p}$ before taking the sample
- **In the second formula (only there!)**, we can take the safe but conservative approach of setting p-hat=0.5, resulting in $n = 0.25 \cdot \left(\frac{z}{B}\right)^2$
- This is like a "worst case scenario" because the product can never exceed 0.25

Example: Sleep

- We would like to find out how much Americans sleep each night?
- Our last random sample had a sample standard deviation of 1.2 hours.
- How large a random sample do we need to obtain a 95% confidence interval with precision plus/ minus 15 minutes?
- “Predicting the population average to within 15 minutes, with 95% confidence”

Example: Relaxing

- We would like to find out which (population) proportion of adults thinks that they have too little time for relaxing or doing nothing.
- The last poll had a proportion of 45% saying this.
- How many people do we need in our sample if we want to predict the population proportion to within five percentage points, with 90% confidence?
- What if we don't know about the results from the last poll?

Summary: CIs for the Mean

- Large sample confidence interval for the mean μ has the form

$$\left[\bar{Y} - z\frac{s}{\sqrt{n}}, \bar{Y} + z\frac{s}{\sqrt{n}}\right]$$

- If given a bound on the margin of error, *B*, and asked for a minimum *n* to achieve that bound:

$$n = \left\lceil s^2\left(\frac{z^2}{B^2}\right)\right\rceil$$, where $\lceil\ \rceil$ means "round up".