Statistics and Quantitative Analysis U4320

Segment 6: Confidence Intervals

Prof. Sharyn O Halloran

URL: http://www.columbia.edu/itc/sipa/U4320y-003/

Review

- Population and Sample Estimates:

	Population	Sample
Mean	$m = \frac{\sum_{i=1}^{N} X_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Variance	$s^2 = \frac{\sum_{i=1}^{N} (X_i - m)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

  - The mean defines central tendency distribution.
  - The variance defines dispersion of the distribution.

Review: Sampling

- When we sample from a population, our sample should be representative of the underlying population.
  - That is, our sample should be unbiased.
  - A simple random sample is selected in such a way that each member of the population has the same chance of being included in the sample.
  - Sampling variability is the variance of sample estimates around population parameters
    - This is inherent in the sampling process.

Review: Sampling (cont.)

- Two sources of sampling variability:
  - Sampling error occurs by chance
    - It is simply the difference between the value of a sample statistic and the value of the corresponding population parameter.
    - Sampling Error = $\bar{x} - m$
  - Non-sample Errors
    - Errors that occur in the collection, recording, and tabulation of that data.
      - Using non-random samples in polling
      - Over-sampling one class or group
      - Under-sampling other class or groups

Review: Sampling (cont.)

- Example:
  - Consider a population of five employees 'salaries:

take a random sample

Individual	Salary in Thousand $
1	17
2	24
3	35
4	35
5	43
Mean	30.8

record sample data

Individual	Salary in Thousand $
2	17
3	35
5	43
Mean	31.66666667

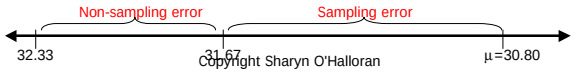
Sampling Error

Oops! Typed 37 instead of 35!

Individual	Salary in Thousand $
2	17
3	37
5	43
Mean	32.33333333

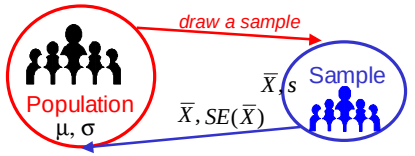
Nonsampling Error

- Sampling error = $x - m = 31.67\text{-}30.80 = \$.87$ thousand
- Non-Sampling error = $x - m = 32.33\text{-}30.80 = \$1.53$ thousand
  $= \$1.53 - \$.87 = \$.66$ thousand



Review: Central Limit Theorem (cont.)

- If a simple random sample is taken from any population with mean μ and standard deviation σ,

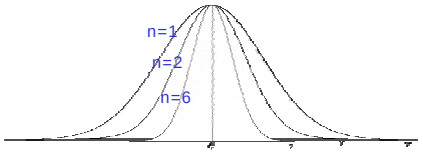


make inference about how good an estimate $\bar{X}$ is of m.

- As n increases, the sampling distribution of $\bar{X}$ tends toward the true population mean μ.

Review: Central Limit Theorem (cont.)

- The Central Limit Theorem states
  - that the sampling distribution of the sample means will be normally distributed with:
    - $\bar{X} \sim N\left(m, \frac{s}{\sqrt{n}}\right)$
    - Sample Mean $E(\bar{X})$ = μ; and
    - Standard Error of the sampling process $SE(\bar{X}) = \frac{s}{\sqrt{n}}$



Review: Central Limit Theorem (cont.)

- Implications
  - From the Central Limit Theorem we are able to show that even if the population is not normally distributed, but the sample size is large,
    - the sampling distribution of $\bar{X}$ can be approximated by a normal distribution.
  - This allows us to use the standard normal tables to make inferences about the population from our sample estimates.

Review: Inference

- To make inferences about the population from a given sample, though, we make one correction:
  - Instead of dividing by the standard deviation σ, we divide by the standard error of the sampling process: $SE = \frac{s}{\sqrt{n}}$
  - We can then standardize by converting observed values to z-values: $Z = \frac{X - m}{SE}$
  - And then use the standard normal table find the probability of events.

Review: Inference

- Think of this process as one of changing hats
  - We first put on our statistician hat to study distributions in the abstract
    - We start with some given distribution with mean μ and standard deviation σ
    - We then discover that the mean of a sample of size n will be distributed N(μ, σ/√n)
  - This is like a controlled experiment; we get to choose the initial distribution ourselves

Review: Inference

- Using this result, we now put on our practitioner's hat.
  - As a researcher, we have some data, but no idea what is the real parent distribution.
- Say we have a sample of size n from a distribution with standard deviation σ
  - This sample happens to have mean $\overline{X}$
  - Then our best guess is that the parent distribution has mean $\overline{X}$ as well.

Confidence Interval

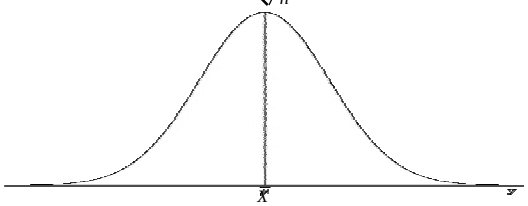
- Motivation
  - We now want to develop tools that allow us to determine how *confident* we are of the that our sample estimates are representative of the underlying population.
    - We know that, on average, $\overline{X}$ is equal to μ.
    - We want some way to express how confident we are that a given $\overline{X}$ is near the actual μ of the population.
  - We do this by constructing a confidence interval, which is some range around $\overline{X}$ that most probably contains μ.

Confidence Interval (cont)

- Definitions
  - A confidence interval is constructed around a point estimate (e.g., $\bar{X}$), and it is stated that this interval is likely to contain the corresponding population parameter (e.g., μ).
    - Two components:
      - The standard error is a measure of how much error there is in the sampling process.
      - The level of confidence attached to the interval.
  - The confidence level associated with a confidence interval states how much confidence we have that the interval contains the true population parameter.
    - The confidence level is denoted by (1-α)100%
      - Common values are 90%, 95% and 99%
      - Corresponding α-levels are .10, .05, and .01.

Confidence Interval: σ known(cont.)

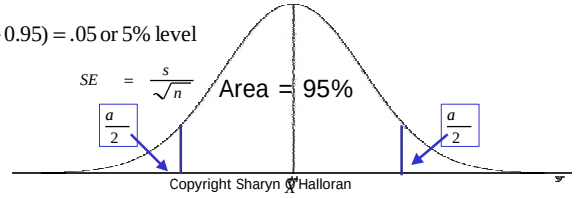
- Constructing a 95% Confidence Interval
  - Graph
    - First, we know from the central limit theorem that the sample mean $\bar{X}$ is distributed normally, with mean μ and standard error $SE = \frac{s}{\sqrt{n}}$



Confidence Interval (cont.)

- Second, we determine how *confident* we want to be in our estimate of μ.
  - Defining how confident you want to be is called the α-level.
    - A 95% confidence interval has an associated α-level of .05.
  - We find a range under the curve with area of 0.95.
    - If we are concerned with both higher and lower values, then the relevant range will have α/2 probability in each tail.

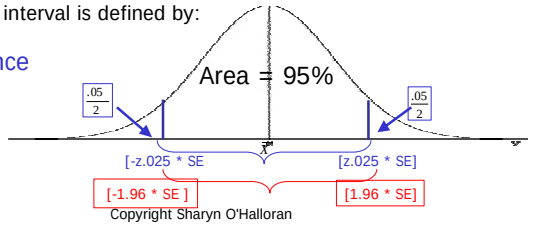
$a = (1 - 0.95) = .05$ or 5% level



Confidence Interval (cont.)

- Third, we find an interval around $\bar{X}$ that contains 95% of the area under the curve
  - The actual interval is [1.96*SE] on either side of the sample mean.
  - We then know that 95% of the time, this interval will contain μ.
  - This interval is defined by:

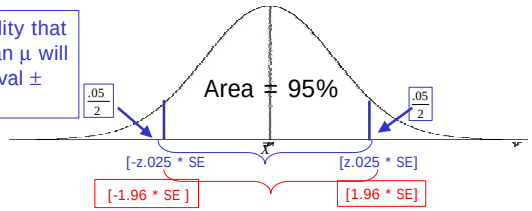
95% confidence interval



Confidence Interval (cont.)

- How should we interpret confidence Intervals…?

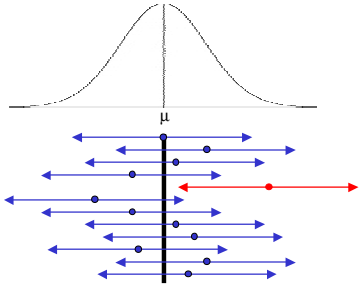
What's the probability that the population mean μ will fall within the interval ± 1.96 * SE?



- Now, let's take this interval of size [-1.96 * SE , 1.96 * SE] and use it as a measuring rod

Confidence Interval (cont.)

- …Think of it as a game of horseshoes



Say the true sampling distribution has mean μ and standard deviation $\sigma/\sqrt{n}$

Then 95% of the time the conf. interval $\overline{X} \pm 1.96*SE$ generated will contain μ

The larger the interval, the less certain we are of our estimates.

Confidence Interval (cont.)

- In general:
  - We know from the 68-95-99.7 rule that a 95% confidence interval will be about 2 standard deviations on either side of $\overline{X}$.
    - To be precise, from the z-table, we find the z-value associated with a .025 probability is 1.96.
    - If we take a random sample of size n from the population,
      - 95% of the time the population mean will be within the range:

$$\overline{X} - (Z_{.025} * \frac{s}{\sqrt{n}}) < m < \overline{X} + (Z_{.025} * \frac{s}{\sqrt{n}}) \Rightarrow$$
$$m = \pm Z a_{/2} * SE$$

Confidence Interval (cont.)

- Example: Calculating a 95% confidence interval
  - Say we sample 180 people and see how many times they ate at a fast-food restaurant in a given week.
    - Sample size n=180
    - The sample has a mean of 0.82, and
    - The population standard deviation σ is 0.48.
  - Calculate the 95% confidence interval for these data.

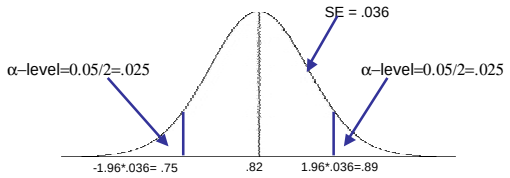
Confidence Interval (cont.)

- Answer:

$$\text{Step 1: Calculate SE} = \frac{.48}{\sqrt{180}} = 0.036$$

$$\text{Step 2: Calculate Margin of Error} = z_{.025} * SE = 1.96 * 0.036 = 0.071$$

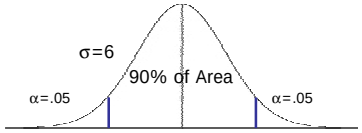
$$\text{Step 3: Calculate Confidence Interval} = .82 \pm .07, \text{ or } [.75 < m < .89]$$



Confidence Interval (cont.)

- Example 2: Calculating a 90% confidence interval
  - A random sample of 16 observations was drawn from a normal population with
    - Standard deviation, σ = 6, and
    - Sample mean μ = 25.
  - Find a 90% (α=.10) confidence interval for the population mean.

10% of the area lies outside the confidence interval



Confidence Interval (cont.)

- First, find $Z_{.10/2}$ in the standard normal tables:

$$Z_{.05} = 1.64$$

- Second, calculate the 90% confidence interval

$$\mu = \bar{X} \pm Z_{.05} * \sigma/\sqrt{n}$$

$$\mu = 25 \pm 1.64 * 6/\sqrt{16}$$

$$SE = 1.5$$

$$\mu = 25 \pm 1.64 * 1.5 = 25 \pm 2.46$$

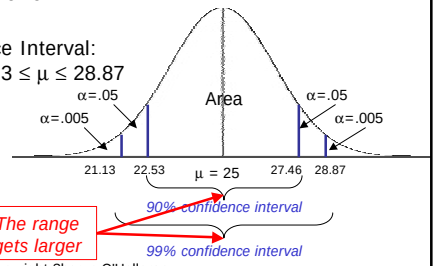
$$22.53 < \mu < 27.46$$

- 90% of the time, the mean lies within this range.

Confidence Interval (cont.)

- What if we wanted to be 99% of the time sure that the mean falls with in the interval?
  - Select α-level: $Z_{.005} = 2.58$
  - Calculate margin of error:
    2.58 *1.5 = 3.87
  - Calculate Confidence Interval:
    25 ± 3.87 or 21.13 ≤ μ ≤ 28.87

What happens when we move from a 90% to a 99% confidence interval?



The range gets larger

Confidence Interval(cont.)

- Why 95%?
  - It is standard to accept that our estimate will be wrong 1 out of 20 times.
  - We could reduce the possibility of error, of course, by making the interval larger.
  - Increasing the interval, however, makes our estimates less precise.
    - That is, the margin of error increases

      $[z_{\alpha\text{-level}} * SE]$
    - Trade off precision for the probability that the true mean lies in a given range.

Confidence Interval: σ Unknown

- Confidence Intervals when σ is unknown
  - We have been calculating confidence intervals assuming that we know the population standard deviation σ.
  - Of course, in most cases, we are not only uncertain of the mean μ, but also of the underlying variance of the parent population.
  - When this is the case, we must estimate σ.
    - The best estimate of σ is the sample standard deviation s:

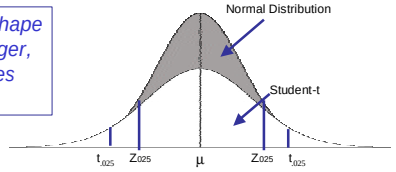
$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

    - This introduces a new source of error that must be taken into account.

Confidence Interval: σ Unknown (cont.)

- Characteristics of a Student-t distribution
  - Shape the student t-distribution

The t-distribution changes shape as the sample size gets larger, and in the limit it becomes identical to the normal.



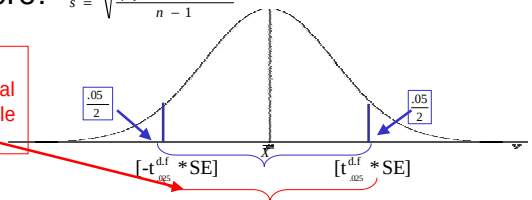
  - When to use t-distribution
    - σ is unknown
    - Sample size n is small (n<30)

Confidence Interval: σ Unknown (cont.)

- Constructing Confidence Intervals using t-Distribution
  - 95% confidence interval is: $\bar{X} \pm t_{.025} \frac{s}{\sqrt{n}}$.

    Where: $s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$

The size of the confidence interval changes as sample size changes.



Confidence Interval:

σ Unknown (cont.)

- Using t-tables
  - Given a sample size n, what is the critical value to get 95% of the area under the curve?
    - Step 1: Find Degrees of Freedom
      - Degrees of freedom is the amount of information used to calculate the standard deviation, s.
      - We denote it as d.f. = n-1
    - Step 2: Look up in the t-table
      - Now we go down the side of the table to the degrees of freedom and across to the appropriate t-value.
      - That's the cutoff value that gives you area of .025 in each tail, leaving 95% under the middle of the curve.
  - Application:
    - Suppose we have sample size n=15 and $t_{.025}$.
    - What is the critical value? 2.13

Confidence Interval:

σ Unknown (cont.)

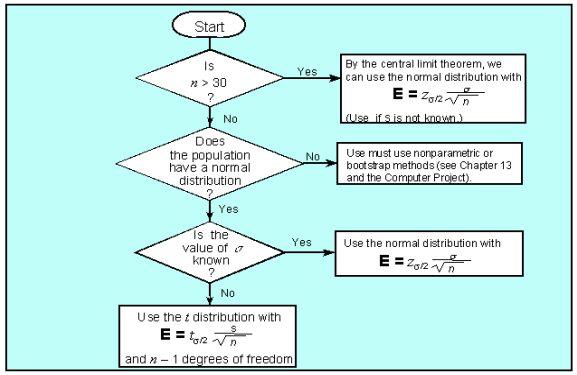
- Comparison to the normal distribution
  - As d.f. gets large the shape of the curve tends toward a normal distribution.
    - As n get larger, $t_{.025}$ gets closer and closer to 1.96 and with infinite degrees of freedom, it equals 1.96.
    - As the sample size grows, the difference between the t and the normal distribution disappears.
  - Look back at the standard normal tables...

Confidence Interval:

σ Unknown (cont.)

What table do I use?

Figure 6-6 Choosing between Normal and *t* Distribution



Confidence Interval:

σ Unknown (cont.)

- Example:
  - Four students had grades on a test of 64, 66, 89, 77. Calculate a 95% confidence interval for the class average.

$$\text{Mean } \bar{X} = \frac{(64+66+89+77)}{4} = 74$$

$$\text{Sample Variance } s^2 = \frac{(64-74)^2+(66-74)^2+(89-74)^2+(77-74)^2}{3}$$

$$= 132.7$$

$$\text{Sample Standard Deviation } s = \sqrt{132.7} = 11.52$$

Confidence Interval:

σ Unknown (cont.)

- Answer:
  - Calculate Margin of Error:

$$SE = \frac{s}{\sqrt{n}} = \frac{\sqrt{132.7}}{\sqrt{4}} = 5.76$$

d.f. = 3

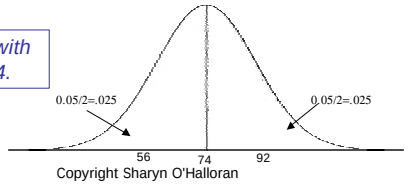
$t_{.025}$ = 3.18

$t_{.025} * SE = 3.18 * 5.76 = 18$

  - Calculate confidence interval:

$$m = \pm 18 = 56 < m < 92$$

Not very precise with a sample of size 4.



Confidence Interval:

Differences of Means

- We can use these same techniques to address a number of different questions.
  - For example, we may wish to determine if two populations (e.g., men and women) have the *same* mean (e.g., salary).
  - Other examples:
    - How two sections of the same class did on an exam.
    - The comparative effectiveness of two drugs in treating the same disease.

Confidence Interval:

Differences of Means (cont.)

- Population Variance Known (σ-known)
  - We are interested in estimating the value ($\mu_1 - \mu_2$) by the sample means, using $(\bar{X}_1 - \bar{X}_2)$.
    - Take samples of the size n_1 and n_2 from the two populations.
    - Estimate the differences in two population means.
    - To tell how accurate these estimates are, we can construct the familiar confidence interval around their difference:

$$(m_1 - m_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This is just the standard error

  - This holds if the sample size is large and we know both σ_1 and σ_2.

Confidence Interval:

Differences of Means(cont.)

- Population Variance Unknown (σ-unknown)
  - If, as usual, we do not know σ_1 and σ_2, then we use the sample standard deviations instead.
  - When the variances of populations are not equal ($s_1 \neq s_2$):

$$(m_1 - m_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - Example:
    - Test scores of two classes where one is from an inner city school and the other is from an affluent suburb.

Confidence Interval:

Pooled Sample Variances(cont.)

- Pooled Sample Variances, $s_1 = s_2$ (σ^2 is unknown)
  - If both samples come from the same population (e.g., test scores for two classes in the same school), we can assume that they have the same population variance, s_p^2.
    - where $s_p^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$
    - 95% Confidence Interval

$$(m_1 - m_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{.025} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (m_1 - m_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{.025} * \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

    - The degrees of freedom are $(n_1-1) + (n_2-1)$, or (n_1+n_2-2).

Confidence Interval:

Pooled Sample Variances(cont.)

- Example:
  - Two classes from the same school take a test. Calculate the 95% confidence interval for the difference between the two class means.

Observation	Class1	Class 2
1	64	56
2	66	71
3	89	53
4	77	
Sum	296	180
Mean	74	60

Confidence Interval:

Pooled Sample Variances(cont.)

- Answer
  - Step 1: Calculate sample estimates

$$\bar{X}_1 = 74;\ \bar{X}_2 = 60$$
$$\bar{X}_1 - \bar{X}_2 = 14$$
$$n_1 = 4;\ n_2 = 3$$

$$s_p^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

$$s_p^2 = \frac{\left[(64-74)^2 + (66-74)^2 + (89-74)^2 + (77-74)^2\right] + \left[(56-60)^2 + (71-60)^2 + (53-60)^2\right]}{(4-1)+(3-1)} \Rightarrow$$

$$s_p^2 = (398 + 186)/(3+2) = 117 \Rightarrow$$

$$s_p = 10.8.$$

Confidence Interval:

Pooled Sample Variances(cont.)

- Step 2: Calculate standard error

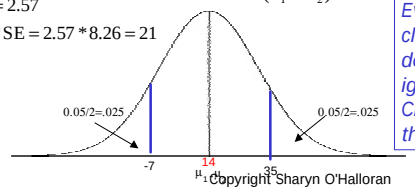
$$SE = s_p \bullet \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 10.8 \bullet \sqrt{\frac{1}{4} + \frac{1}{3}} = 8.26$$

- Step 3: Calculate 95% Confidence Interval

$$(m_1 - m_2) = (\bar{X}_1 - \bar{X}_2) \pm 21 = 14 \pm 21$$

$$-7 \le (m_1 - m_2) \le 35.$$

$$d.f. = 5$$
$$t_{.025} = 2.57$$
$$t_{.025} * SE = 2.57 * 8.26 = 21$$



Even though the first class looks like they are doing better, we can t ignore the possibility that Class 2 is doing better than Class 1.

Confidence Interval:

Matched Samples

- Matched Samples
  - Definition
    - Matched samples are ones where you take a single individual and measure him or her at two different points and then calculate the difference.
  - Advantage
    - One advantage of matched samples is that it reduces the variance because it allows the experimenter to control for many other variables which may influence the outcome.

Confidence Interval:

Matched Samples(cont.)

- Calculating a 95% Confidence Interval
  - For each individual we can calculate their difference D from one time to the next.
    - We then use these D's as the data set to estimate Δ, the population difference.
    - The sample mean of the differences will be denoted $\overline{D}$.
  - The standard error will just be: $SE = \frac{s_D}{\sqrt{n}}$
  - Use the t-distribution to construct 95% confidence interval: $\Delta = \overline{D} \pm t_{.025} * \frac{s_D}{\sqrt{n}}$

Confidence Interval:

Matched Samples(cont.)

- Example:

Student	X1 (Fall)	X2 (Spring)	D = X1-X2
Trimble	64	57	7
Wilde	66	57	9
Giannos	89	73	16
Ames	77	65	12
Sum	296	252	44
Mean	74	63	11

d.f. = n-1 = 3

$t_{.025} = 3.18$

$s_D^2 = \frac{(7-11)^2+(9-11)^2+(16-11)+(12-11)^2}{3} = \frac{46}{3} = 15.233 \Rightarrow$ $SE = \frac{s_D}{\sqrt{4}} = \frac{3.91}{2} = 1.96$

$s_D = 3.91$

- 95% Confidence Interval $\Delta = \overline{D} \pm t_{.025} * \frac{S_D}{\sqrt{n}} \Rightarrow$

$t_{.025} * SE = 3.18 * 1.96 = 6$

$11 \pm 6 = 5 \text{ to } 17$

$5 < \Delta < 17$

Notice that the standard error is much smaller than in our unmatched pairs of equal sample size.

Confidence Interval:

Proportions

- Example:
  - Just before the 1996 presidential election, a Gallup poll of about 1500 voters showed 840 for Clinton and 660 for Dole.
  - Calculate the 95% confidence interval for the population proportion π of Clinton supporters.
    - n= 1500
    - Sample proportion P: $P = \frac{840}{1500} = .56$
      - That is, in our sample of 1500 individuals, 840 people responded that they preferred Clinton to Dole.

Confidence Interval:

Proportions

- Create a 95% confidence interval:
  - where π and P are the population and sample proportions, respectively, and n is the sample size.

$$\pi = P \pm \text{sampling allowance}$$

$$p = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$$

$$\pi = .56 \pm 1.96\sqrt{\frac{.56(1-.56)}{1500}}, \qquad \pi = .56 \pm .03.$$

  - That is, with 95% confidence, the proportion of voters for Clinton in the whole population was between 53% and 59%.

Variance of Binomial Dist.

- In general, the variance is the expected value of $(x-\mu)^2$
  - Take a binomial with $P(x=1) = \pi$, $P(x=0) = 1-\pi$
  - Mean $\mu = \pi * 1 + (1-\pi) * 0 = \pi$
  - Variance $= \pi * (1-\pi)^2 + (1-\pi) * (0-\pi)^2$

Prob x=1 | $(x-\mu)^2$ if x=1 | Prob x=0 | $(x-\mu)^2$ if x=0

Variance of Binomial Dist.

- In general, the variance is the expected value of $(x-\mu)^2$
  - Take a binomial with $P(x=1) = \pi$, $P(x=0) = 1-\pi$
  - Mean $\mu = \pi * 1 + (1-\pi) * 0 = \pi$
  - Variance $= \pi * (1-\pi)^2 + (1-\pi) * (0-\pi)^2$

$$= (1-\pi)*[\pi(1-\pi)] + \pi*[\pi(1-\pi)]$$

$$= [\pi(1-\pi)]$$