AP Statistics Notes – Unit Ten: Inference for the Proportion of a Population

Syllabus Objectives: 4.1 – The student will estimate population parameters and margins of errors for proportions and means.

In the last unit, we learned inferential procedures for means. In this unit, we will apply that logic to inference involving proportions. We will learn how to build confidence intervals and perform significance tests for one proportion as well as comparisons between two proportions.

- **Conditions for inference about a proportion**
  - We will be estimating the true population proportion with the sample proportion. The sample proportion is $\hat{p} = \frac{\#\text{ of successes}}{n}$.
  - Our data are from a **simple random sample** of size *n* from the population of interest. This condition is very important.
  - The population is at least 10 times as large as the sample.
  - For a test of $H_0 : p = p_0$, the sample size *n* is so large that both np_0 and $n(1-p_0)$ are 10 or more. For a confidence interval, *n* is so large that both the count of successes $n\hat{p}$ and the count of failures $n(1-\hat{p})$ are 10 or more.

- **Standard Error**
  - **Definition:** Provided that the conditions for inference are met above, the **standard error** of the statistic is the estimated standard deviation of the statistic.
  - **Formula:** $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

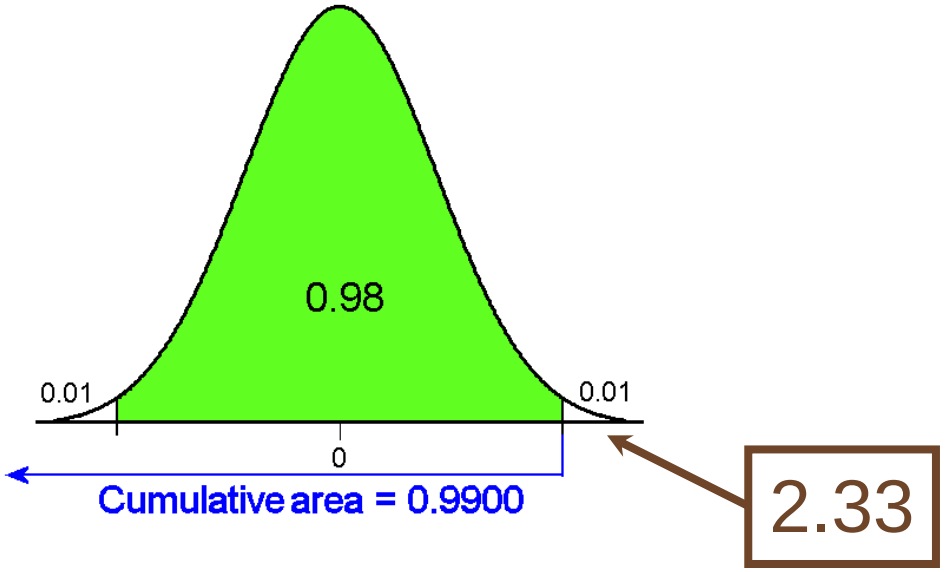
Syllabus Objectives: 4.4 – The student will calculate the large sample confidence interval for a proportion. 4.11 – The student will perform a large sample test for a proportion.

- **One-sample Proportion z Confidence Interval procedures**
  - Identify the population of interest and the parameter you want to draw conclusions about.
  - Choose the appropriate inference procedure. Verify the conditions for using the selected procedure. Make sure to substitute values in for *n* and $\hat{p}$ to check for normality.
  - If the conditions are met, carry out the inference procedure. CI = estimate $\pm$ margin of error.

    $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where z^* is the upper $\frac{(1-C)}{2}$ standard normal critical value.

  - Interpret your results in the context of the problem.

- **Finding a z Critical Value**
  - Finding a *z* critical value for a 98% confidence interval.



Looking up the cumulative area or 0.9900 in the body of the table we find $z = 2.33$.

  - Common Critical Values

Confidence level	z critical value
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.09
99.9%	3.29

- **Sample Size Formula**
  - The sample size required to estimate a population proportion, *p,* to within a margin of error, *m,* can be found using the formula: $n = \frac{(z^*)^2 p(1-p)}{m^2}$. The value of *p* may be estimated by prior information. If no prior information is available, use $p = 0.5$ in the formula to obtain a conservatively large value for *n*. Round up to the next integer and this would be the smallest sample required under the stated conditions.
  - **Example** – If a TV executive would like to find a 95% confidence interval estimate within 0.03 for the proportion of all households that watch a popular show regularly, how large a sample is needed if a prior estimate for *p* was 0.15?
    - **Solution:** We have the error as m = 0.03 and a prior estimate to substitute in. The critical value for a 95% interval is 1.96.

$$n = \frac{(z^*)^2 p(1-p)}{m^2} = \frac{(1.96)^2(0.15)(0.85)}{0.03^2} = \frac{0.489804}{0.0009} = 544.227 = 545$$

A sample of 545 or more would be needed for 95% confidence and a 3% margin or error.

  - **Example revisited** – Suppose this TV executive would like to make this same estimate, but this time does not have a prior estimate for the population proportion. We now have to use $p = 0.5$ in the sample size formula.
    - **Solution:**

$$n = \frac{(1.96)^2(0.50)(0.50)}{0.03^2} = \frac{0.9604}{0.0009} = 1067.11 = 1068$$

The required sample size is now 1068. Notice, a reasonable ball park estimate for *p* can lower the needed sample size.

- **Confidence Interval for a sample proportion example:**
  - **Example** – For a project, a student randomly sampled 182 other students at a large university to determine if the majority of students were in favor of a proposal to build a field house. He found that 75 were in favor of the proposal.
    - **Solution:** We will follow the four steps previously stated. First, we must state the population parameter and the population of interest: $p =$ the true proportion of students that favor the proposal. Secondly, we must verify that we can perform a 1-sample proportion *z* confidence interval. We must first assume the results are an SRS from the population of students. We must also assume that the population of students is greater than 10 times our sample size of 182 (Pop > 1820). We then find the sample proportion: $\hat{p} = \frac{75}{182} = 0.4121$ and check for normality. $n\hat{p} = 182(0.4121) = 75 \geq 10$ and $n(1-\hat{p}) = 182(0.5879) = 107 \geq 10$.
    - All of our assumptions hold, now we can move to step 3 and find the confidence interval.

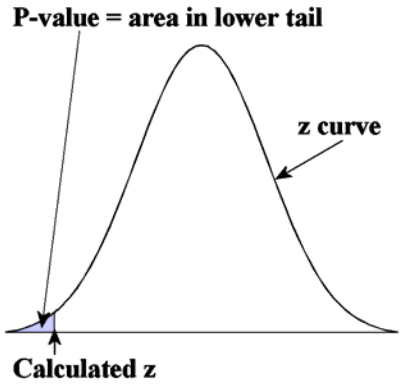
- $\hat{p} = 0.4121,\ z^* = 1.96,\ n = 182$
- $\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.4121 \pm (1.96)\sqrt{\frac{0.4121(0.5879)}{182}} = 0.4121 \pm 0.07151$
- The 95% confidence interval for *p* is: (0.341, 0.484)
- The last step is that we must interpret it correctly. Based on this sample, we are 95% confident that the true proportion of students that are in favor of a proposal to build a field house is between 0.341 and 0.484.

- **One-sample proportion *z* test procedures**
  - To test the null hypothesis, we use the following *z* statistic:

$$\mathbf{z = \frac{p - hypothesized\ value}{\sqrt{\frac{hypothesized\ value(1\text{-}hypothesized\ value)}{n}}}}$$

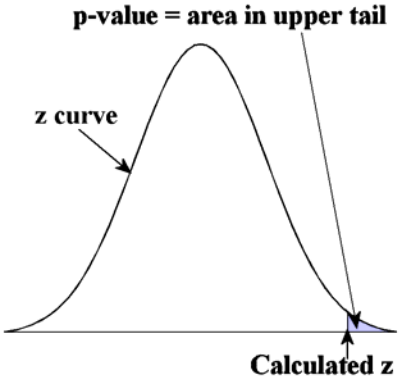
  - The approximate p-value for this test depends on the alternate hypothesis and is given for each of the possible alternate hypotheses.
  - **One-sided:**

H_0: *p* = hypothesized proportion
H_A: *p* < hypothesized proportion



$$\mathbf{P\text{-}value = P\left(z < \frac{p - hypothesized\ value}{\sqrt{\frac{hypothesized\ value(1\text{-}hypothesized\ value)}{n}}}\right)}$$

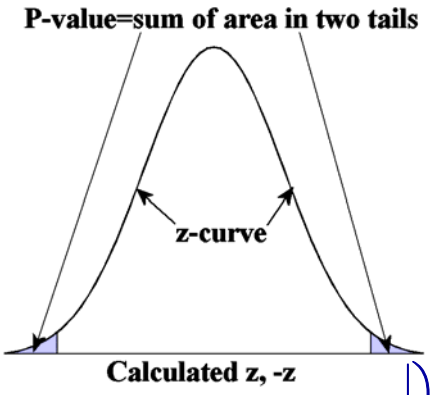
H_0: *p* = hypothesized proportion
H_A: *p* > hypothesized proportion



$$\mathbf{P\text{-}value = P\left(z > \frac{p - hypothesized\ value}{\sqrt{\frac{hypothesized\ value(1\text{-}hypothesized\ value)}{n}}}\right)}$$

- **Two-sided:**

H_0: p = hypothesized proportion
H_A: $p \neq$ hypothesized proportion



$$\text{P-value} = 2P\left(z > \left| \frac{\text{p} - \text{hypothesized value}}{\sqrt{\frac{\text{hypothesized value(1-hypothesized value)}}{\text{n}}}} \right| \right)$$

- **Tests of Significance for a sample proportion examples:**
  - **Example 1 –** An insurance company states that the proportion of its claims that are settled within 30 days is 0.9. A consumer group thinks that the company drags its feet and takes longer to settle claims. To check these hypotheses, a simple random sample of 200 of the company's claims was obtained and it was found that 160 of the claims were settled within 30 days.
    - **Solution:** Step 1: Identify the population of interest and the parameter you want to draw conclusions about. $p =$ the true proportion of the company's claims that are settled within 30 days.
    - Step 2: Write the hypotheses for the test. This will be a one-sided test.
      $H_0 : p = 0.9 \quad H_a : p < 0.9$
    - Step 3: Choose the appropriate inference procedure and verify the conditions for using the procedure. We are doing a one-sample proportion *z* test. It is stated we have an SRS of the population of claims. We must assume that the population of claims is greater than ten times the sample size of 200 (Pop > 2000) and we must check for normality:
      $np_0 = 200(0.9) = 180 \geq 10$ and $n(1-p_0) = 200(0.1) = 20 \geq 10$.
    - Step 4: Carry out the inference procedure. The sample proportion is
      $\hat{p} = \frac{160}{200} = 0.8$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.8 - 0.9}{\sqrt{\frac{(0.9)(0.1)}{200}}} = \frac{-0.1}{0.02121} = -4.71$$

$$P(z < -4.71) \approx 0$$

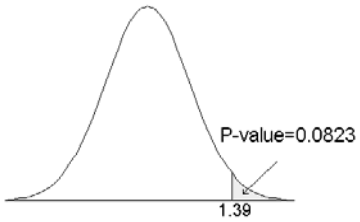
- Step 5: Make your decision. Our p-value is approximately zero. The probability of getting a result as strongly or more strongly in favor of the consumer group's claim if the company's claim was true is essentially zero. Since no significance level is given in the problem, we will use $\alpha = 0.05$. Since 0 < 0.05, our decision is to Reject H_0.
- Step 6: Interpret your results in the context of the problem. At a level of significance of 0.05, there is strong support for the claim that the proportion of the insurance company's claims that are settled within 30 days is less than 0.9. We have shown that the true proportion is statistically significantly less than 0.9.

o **Example 2 –** A county judge has agreed that he will give up his county judgeship unless there is evidence at the 0.10 level that more than 25% of his party is in opposition. An SRS of 800 party members included 217 who opposed him. Please advise this judge.

- **Solution:** Step 1: $p =$ the true proportion of his party that is in opposition.
- Step 2: $H_0 : p = 0.25$ and $H_a : p > 0.25$
- Step 3: We will use a one-sample proportion *z* test. We are told we have an SRS of party members. We must assume the population of party members is greater than ten times the sample size (Pop > 8000). We check $np_0 = 800(0.25) = 200 \geq 10$ and $n(1-p_0) = 800(0.75) = 600 \geq 10$.
- Step 4: The sample proportion is $\hat{p} = \frac{217}{800} = 0.27125$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.27125 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{800}}} = \frac{0.02125}{0.01531} = 1.39$$

$$P - Value = 0.0823$$



- Step 5: The problem states that $\alpha = 0.10$. Our p-value is smaller than the significance level (0.0823 < 0.10). This means that we will Reject H_0 at the 0.01 level of significance.
- Step 6: At the 0.01 level of significance, there is sufficient evidence to support the claim that the true percentage of the party members that oppose him is more than 25%. Under these circumstances, I would advise the judge not to run.

Syllabus Objectives: 4.5 – The student will calculate the large sample confidence interval for a difference between two proportions. 4.12 – The student will perform a large sample test for a difference between two proportions.

- **Comparing two population proportions**
  - We want to compare two populations or the responses to two treatments based on two independent samples.
  - New Notation: We use subscripts to define our two treatments or populations.

	Sample Size	Population Proportion of Successes	Sample Proportion of Successes
Population or treatment 1	n_1	p_1	$\hat{p}_1$
Population or treatment 2	n_2	p_2	$\hat{p}_2$

  - Sampling Distribution of $p_1 - p_2$

    If two random samples are selected independently of one another, the following properties hold:

    - $\mu_{p_1 - p_2} = p_1 - p_2$
    - $\sigma^2_{p_1 - p_2} = \sigma^2_{p_1} + \sigma^2_{p_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

      and $\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
    - If both n_1 and n_2 are large $[n_1 p_1 \geq 5,\ n_1(1-p_1) \geq 5,\ n_2 p_2 \geq 5,\ n_2(1-p_2) \geq 5]$, then p_1 and p_2 each have a sampling distribution that is approximately normal.
  - Assumptions:
    1. The samples are **independent**.
    2. The *n* sample differences can be viewed as a random sample for a population of differences.
    3. The populations are at least ten times as large as the samples.
    4. The number of sample differences is large (as noted above).

- **Two-sample Proportion z Confidence Interval procedures**
  - Identify both populations of interest and the parameters you want to draw conclusions about.

- Choose the appropriate inference procedure. Verify the conditions for using the selected procedure. Make sure to substitute values in for both sample sizes and sample proportions to check for normality.
- If the conditions are met, carry out the inference procedure. CI = estimate ± margin of error.

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$, where z^* is the upper $\frac{(1-C)}{2}$ standard normal critical value.

- Interpret your results in the context of the problem.

- **Confidence Interval for two sample proportion example:**
  - **Example** – A student assignment called for the students to survey both male and female students (Independently and randomly chosen) to see how they feel about the College's new drug and alcohol policy. A student went and randomly selected 200 male students and 100 female students and obtained the data summarized below. Use this data to obtain a 90% confidence interval estimate for the difference of the proportions of female and male students that approve of the policy.

	Number surveyed	Number that approve	Proportion that approve
Female	100	43	0.430
Male	200	61	0.305

  - **Solution:** We will follow the four steps previously stated. First, we must state the population parameters and the populations of interest: $p_1 =$ the true proportion of females that approve of the policy, $p_2 =$ the true proportion of males that approve of the policy. Secondly, we must verify that we can perform a 2-sample proportion *z* confidence interval. We are told that we have independent samples and that the females and males were randomly sampled. We must also assume that the population of females and males is greater than 10 times our sample sizes (Pop females > 1000, Pop males > 2000). We have the sample proportion: $\hat{p}_1 = \frac{43}{100} = 0.430$, $\hat{p}_2 = \frac{61}{200} = 0.305$ and check for normality. $n_1\hat{p}_1 = 100(0.430) = 43 \geq 5$, $n_1(1-\hat{p}_1) = 100(0.570) = 57 \geq 5$, $n_2\hat{p}_2 = 200(0.305) = 61 \geq 5$, and $n_2(1-\hat{p}_2) = 200(0.695) = 139 \geq 5$
  - All of our assumptions hold, now we can move to step 3 and find the confidence interval.
  - For a 90% confidence interval, the *z* value to use is 1.645. This value was obtained from the table.

- $(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$
- $(0.430 - 0.305) \pm \sqrt{\frac{0.430(1-0.430)}{100} + \frac{0.305(1-0.305)}{200}} = 0.125 \pm 0.097$
- The 90% confidence interval is: (0.028, 0.222)
- The last step is that we must interpret it correctly. Based on the observed samples, we believe that the proportion of females that approve of the policy exceeds the proportion of males that approve of the policy by somewhere between 0.028 and 0.222.

- **Two-sample Proportion *z* test procedures**
  - The **combined estimate of the common population proportion** is

$$p_c = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{\text{total number of success in 2 samples}}{\text{total sample size}}$$

  - To test the null hypothesis, we use the following *z* statistic:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}}$$

- **Tests of Significance for two sample proportion examples:**

  - **Example 1 –** A group of college students were asked what they thought the "issue of the day" was. Without a pause, the class almost to a person said "student retention". The class then went out and obtained a random sample and asked the question, "Do you plan on returning next year?" The responses along with the gender of the person responding are summarized in the following table. Test to see if the proportion of students planning on returning is the same for both genders at the 0.05 level of significance.

		Response		
		Yes	No	Maybe
Gender	Male	211	45	19
	Female	141	32	9

    - **Solution:** Step 1: Identify the populations of interest and the parameters you want to draw conclusions about. $p_1 =$ the true proportion of males that plan on returning next year, $p_2 =$ the true proportion of females that plan on returning next year.
    - Step 2: Write the hypotheses for the test. This will be a two-sided test. $H_0 : p_1 - p_2 = 0 \quad H_a : p_1 - p_2 \neq 0$

- Step 3: Choose the appropriate inference procedure and verify the conditions for using the procedure. We are doing a two-sample proportion *z* test. We assume that the two samples are independently chosen random samples. We must assume that the populations of both students is greater than ten times the sample sizes.(Pop males > 2750) (Pop females > 1820) and we must check for normality:

$$p_c = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{211+141}{275+182} = \frac{352}{457} = 0.7702$$

$n_1 p_c = 275(0.7702) = 211.8 \geq 5,\ n_1(1-p_c) = 275(0.2298) = 63.2 \geq 5$ and
$n_2 p_c = 182(0.7702) = 140.2 \geq 5,\ n_2(1-p_c) = 182(0.2298) = 41.8 \geq 5$.

- Step 4: Carry out the inference procedure.

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{275} + \frac{p_c(1-p_c)}{182}}}$$

$$= \frac{0.76727 - 0.77473}{\sqrt{\frac{0.77024(1-0.77024)}{275} + \frac{0.77024(1-0.77024)}{182}}}$$

$$= \frac{-0.0074525}{0.040198} = -0.19$$

P-value:
The P-value for this test is 2 times the area under the z curve to the left of the computed z = -0.19.
P-value = 2(0.4247) = 0.8494

- Step 5: Make your decision. Since our p-value is 0.849 which is much larger than our significance level of 0.05, we will fail to reject H_0.
- Step 6: Interpret your results in the context of the problem. At a level of significance of 0.05, there is no evidence that the return rate is different for males and females.

o **Example 2 –** A consumer agency spokesman stated that he thought that the proportion of households having a washing machine was higher for suburban households than for urban households. To test to see if that statement was correct at the 0.05 level of significance, a reporter randomly selected a number of

households in both suburban and urban environments and obtained the following data.

	Number surveyed	Number having washing machines	Proportion having washing machines
Suburban	300	243	0.810
Urban	250	181	0.724

- **Solution:** Step 1: $p_1 =$ the true proportion of suburban households having washing machines, $p_2 =$ the true proportion of urban households having washing machines.
- Step 2: $H_0 : p_1 - p_2 = 0$ and $H_a : p_1 - p_2 > 0$
- Step 3: We will use a one-sided, two-sample proportion *z* test. Assume that the two samples are independently chosen random samples. We must also assume the populations of both households are greater than ten times the sample sizes (pop suburban households > 3000, pop urban households > 2500). We now check for normality.

$$p_c = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{243 + 181}{300 + 250} = \frac{424}{550} = 0.7709$$

$n_1 p_c = 300(0.7709) = 231.3 \geq 5,\ n_1(1 - p_c) = 300(0.2291) = 68.7 \geq 5$ and

. $n_2 p_c = 250(0.7709) = 192.7 \geq 5,\ n_2(1 - p_c) = 250(0.2291) = 57.3 \geq 5$

- Step 4: Carry out the inference procedure.
-

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}}$$

$$z = \frac{0.810 - 0.724}{\sqrt{\frac{0.7709(0.2291)}{300} + \frac{0.7709(0.2291)}{250}}}$$

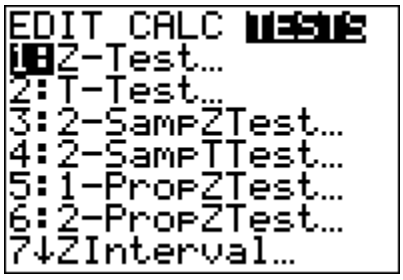
$$z = 2.390$$

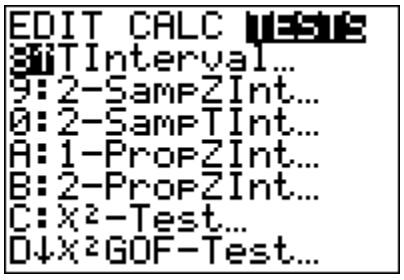
P-value:

The P-value for this test is the area under the z curve to the right of the computed z = 2.39. The P-value = 1 - 0.9916 = 0.0084

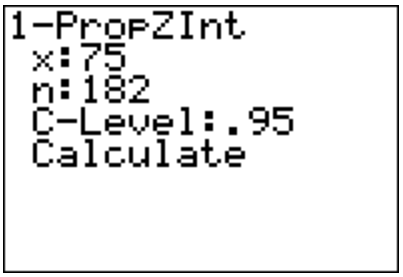
- Step 5: The problem states that $\alpha = 0.05$. Our p-value is smaller than the significance level (0.0084 < 0.05). This means that we will Reject H_0 at the 0.05 level of significance.
- Step 6: At the 0.05 level of significance, there is sufficient evidence that the proportion of suburban households that have washers is more than the proportion of urban households that have washers.

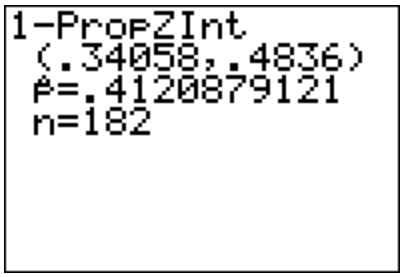
- **Using the TI-84 for inference problems.**
  - Tests and confidence intervals for one-sample and two-sample means and proportions can be done on the TI-84 calculator.
  - Note that the calculator will only help on the calculation step. All other steps for the confidence intervals and tests must still be done.
  - Use the following keystrokes to get to the test menu: 



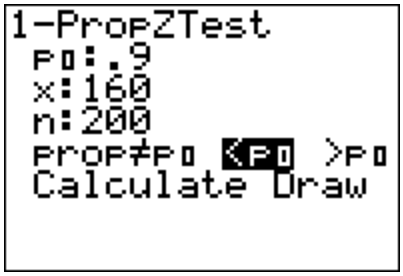


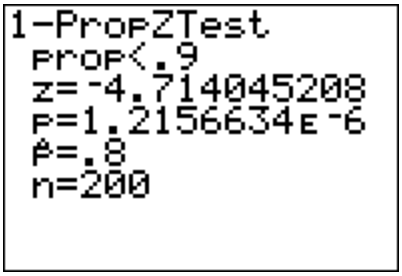
  - Select the inference procedure you wish to perform and input the necessary statistics, hypotheses and confidence levels.
  - **Example 1 – Revisit the 1-sample proportion confidence interval:**



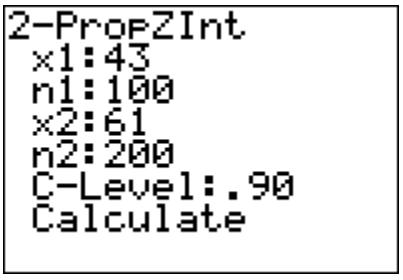


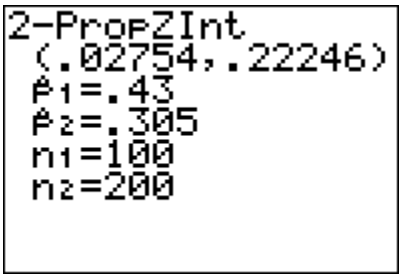
  - **Example 2 – Revisit the 1-sample proportion z test:**





- **Example 3 – Revisit the 2-sample proportion confidence interval:**





- **Example 4 – Revisit the 2-sample proportion z test:**