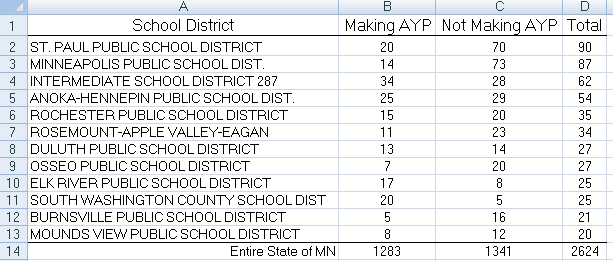
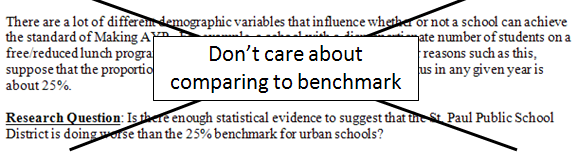
Example 5.1: Consider outcome the academic yearly progress (AYP) of several school districts in MN. Consider the following table that includes school districts within the State of MN with more than 20 schools.

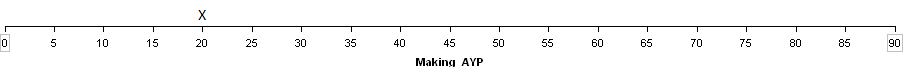


A confidence interval is necessary when there is no need or desire to compare an observed outcome to a pre-specified value.

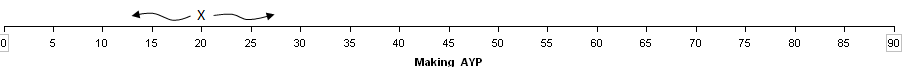


The purpose of a confidence interval is to understand the amount of inherent variation that exists in a statistic (without respect to some pre-specified value as is the case in hypothesis testing). For example, for the St. Paul School district, there were a total of 90 schools for which 20 of these schools meet AYP for the given year. A confidence interval will allow us to determine how much variation we’d expect over repeated sampling from a district with 90 schools.

Outcome for St. Paul School District:



A confidence interval will attempt to measure the inherent variation that exists in our statistic.

****

**Version #1: Simulation**

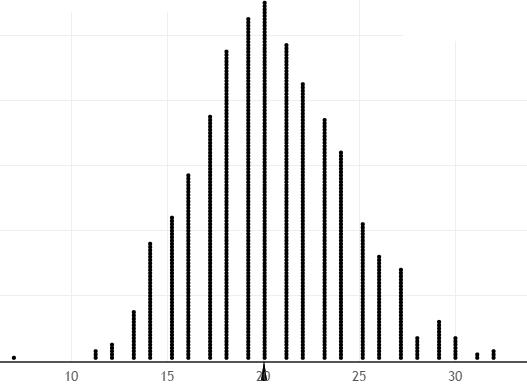
Simulation can be used to understand the variation that exists over repeated sampling under a given model. Set up the simulation using StatKey using the following parameters.

|  |  |
| --- | --- |
| Setting up the number line | Simulate under |

Questions:

1. Why is the sample size set to 90?
2. Where did the 0.2222 value come from? This value is called the **point estimate**.
3. Why was the value of 0.2222 used?
4. If a simulation is carried out, what value will my outcomes be centered around? How did you determine this value?

The following are the outcomes from 1000 simulations.



Questions:

1. This graph is centered around what value?
2. Using the graph above give a range of likely value for the number of schools making AYP under this model?

A more formal method of determining an appropriate range is advantageous and could be determined using the proportion of outcomes deemed “too low” or “too high” under this model.

Determine a lower endpoint so that the proportion of outcomes less than or equal to this value is close to 2.5%. Next, determine an upper endpoint so that the proportion of outcomes greater than or equal to this value is close to 2.5%.

|  |  |
| --- | --- |
| Lower Side | Upper Side |
|  |  |
|  | |

|  |  |
| --- | --- |
| Goal: Keep the cutoff values symmetric about the center… | |
| * 12, 8 below 20 | * 28, 8 above 20 |
|  | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1st try | | |  | 2nd try | | |
| Endpoint | Value | Probability in tail |  | Endpoint | Value | Probability in tail |
| Lower Tail | 13 |  |  | Lower Tail | 12 |  |
| Upper Tail | 29 |  |  | Upper Tail | 28 |  |
| Total | |  |  | Total | |  |

Question:

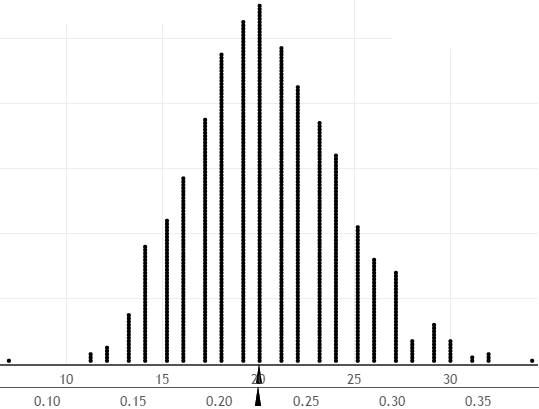
1. What is the purpose of finding this range of values? What information does this range of values provide?

There are some advantages to using percentages (or proportions) compared to counts. In the above simulation, suppose instead of tracking the number of schools making AYP, the proportion or percentage of schools making AYP was tracked instead.





The effect of tracking the proportion instead of the counts is a simple rescaling of the axis on my plot. The shape of the distribution will not change. Hence, when determining unusual observations (i.e. lower and upper endpoints), we can use either counts or proportions.



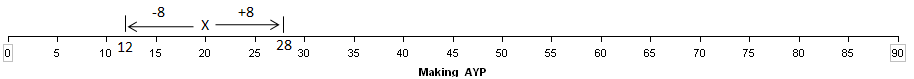
Fill-in the appropriate proportion for the following points of interest. Write each value in its corresponding position on the crossed-out count axis above.

* A value of 21 in the above plot would become \_\_\_\_\_\_\_\_\_ on the proportion or percentage scale.
* A value of 0 in the above plot would become \_\_\_\_\_\_\_\_\_ on the proportion scale.
* A value of 39 in the above plot would become \_\_\_\_\_\_\_\_\_ on the proportion scale.
* Our Lower Endpoint of 12 would become \_\_\_\_\_\_\_\_\_ on the proportion scale.
* Our Upper Endpoint of 28 would become \_\_\_\_\_\_\_\_\_ on the proportion scale.

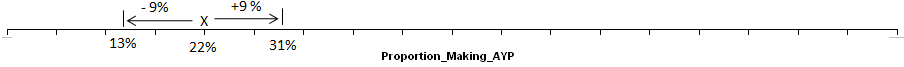
Question:

1. How would one interpret the meaning of the Lower and Upper Endpoint when the outcomes are measured as a proportion instead of a count? Explain.

95% Confidence Interval on Count Scale



95% Confidence Interval on Proportion or Percent Scale



**Version #2: Normal Theory Approach**

As we have mentioned before, statisticians do not use simulation each time we want to understand the inherent variation present in our point estimate. There are short-cuts. The most common short-cut involves the use of a normal curve or bell curve. This normal curve approach uses a proportion or percentage scale instead counts.

|  |  |
| --- | --- |
| Outcomes from simulation (on proportion scale) |  |
| Normal curve placed over the top of these outcomes |  |
| Normal curve without the outcomes shown |  |

The following steps are used to calculating a 95% confidence interval for a proportion using a normal curve.  
  
Step 1: Compute your **point estimate** which is the proportion observed in your data,

Step 2: Calculate the **standard error** for your observed proportion.



Step 3: Compute the **margin of error**. The margin of error is approximately 1.96 standard errors for a 95% confidence interval. The reason for using 1.96 is briefly discussed near the end of this section.



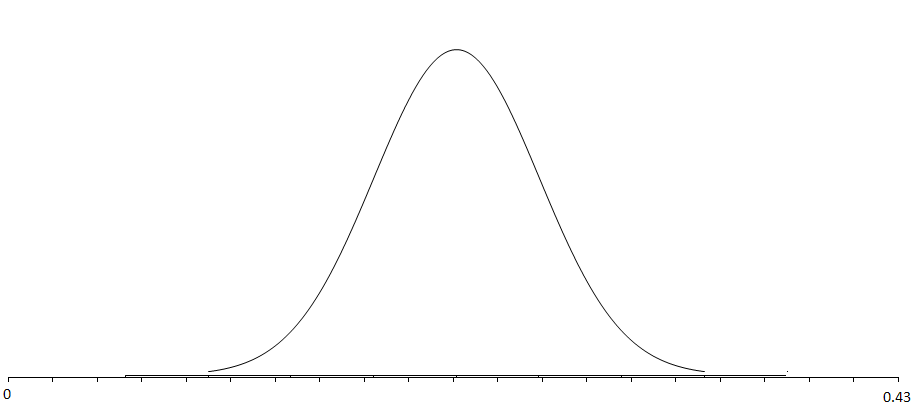
Step 4: Compute the lower and upper endpoint as follows:

Lower endpoint = =

Upper endpoint = =

Sketch the following values for our example on the normal curve below.

* Point Estimate
* Lower and Upper Endpoint
* Margin of Error



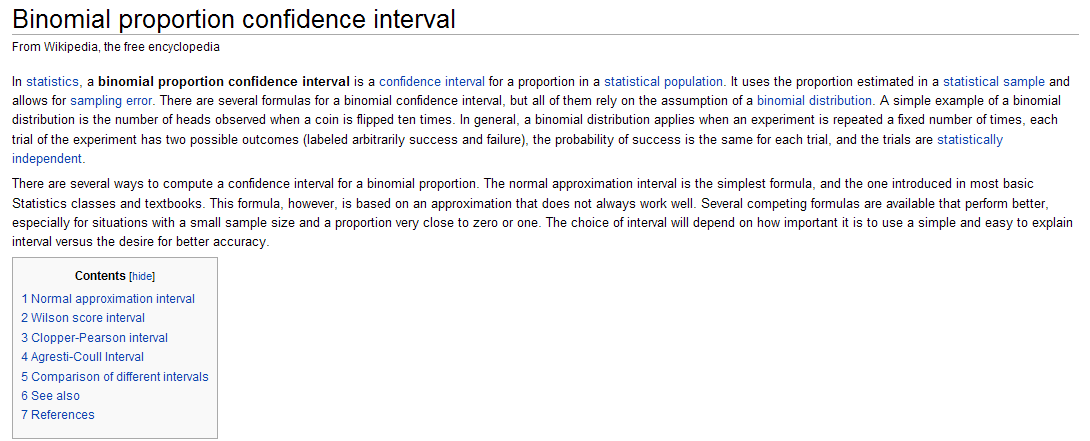
Questions:

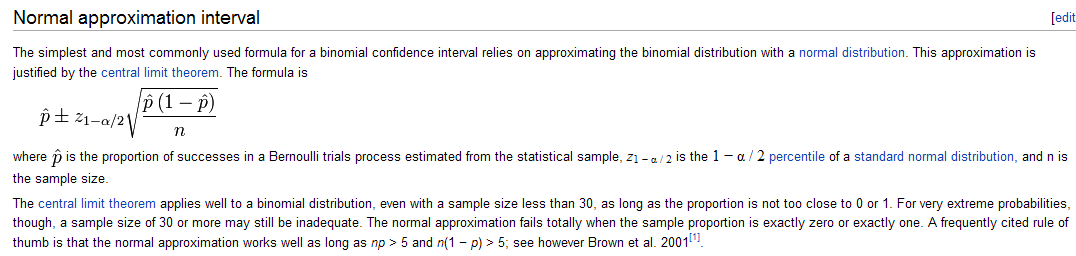
1. What is your 95% confidence interval based on the normal theory approach for this problem?
2. What is the meaning of this 95% confidence interval?
3. Are the lower and upper endpoints found using the normal theory approach close to the values found from the simulation? Compare and contrast the values in the following table.

|  |  |  |
| --- | --- | --- |
| Proportion or Percentage Scale | | |
|  | Simulation | Normal Theory |
| Lower Endpoint | 12 / 90 = 0.133 | 0.1363 |
| Upper Endpoint | 28 / 90 = 0.311 | 0.3081 |
|  | | |
| Count Scale | | |
|  | Simulation | Normal Theory |
| Lower Endpoint | 12 | 0.1363\* 90 = 12.3 |
| Upper Endpoint | 28 | 0.3081 \* 90 = 27.7 |

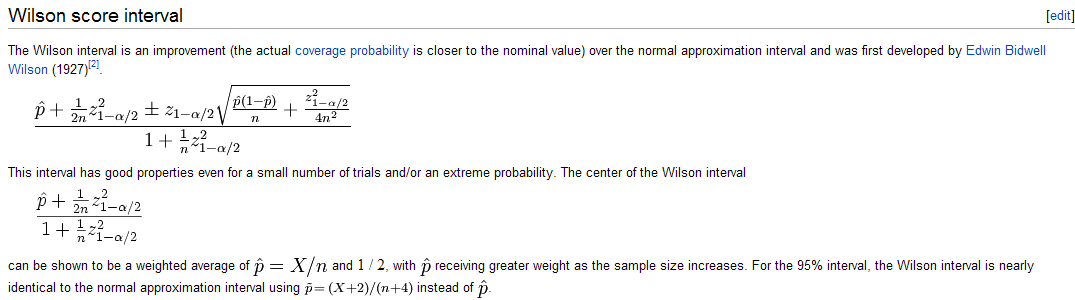
|  |
| --- |
| **Side note regarding the use of 1.96**  Question: What is 1.96 used in the formula for a two-sided 95% confidence interval?  Answer: Because for a normal distribution, using 1.96 \* standard error gives us an error rate of 2.5% in each tail, resulting in a total error rate of 5%.    Question: What if interest lies in only one end of the distribution (i.e. a one-sided confidence interval)?  Answer: The value of 1.645 should be used instead of 1.96. |

Comment: Several methods exist for constructing a confidence interval for a binomial proportion. The above is known as the normal approximation method or “Wald” interval. The Wald interval has limitations as is mentioned in the following Wikipedia entry for “Binomial proportion confidence interval”.



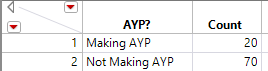


Wikipedia goes on to discuss the Wilson Score Interval. This is the interval done by JMP

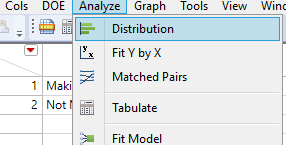


**Getting the 95% confidence interval in JMP**

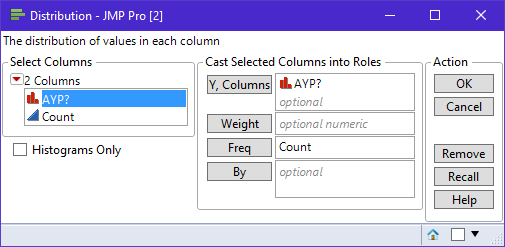
Enter the two categories and counts in JMP as is shown here.



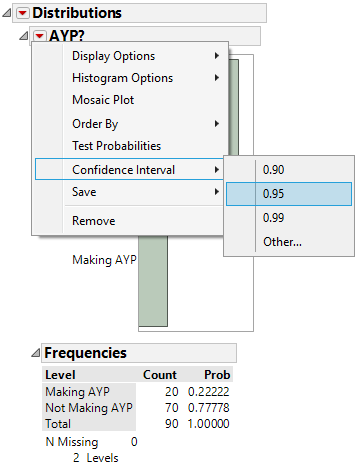
Select Analyze > Distribution



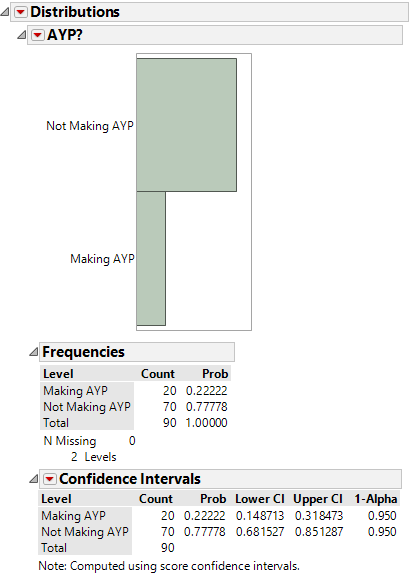
In the Distribution box, place the name of your categorical variable in the Y, Columns box and Count in the Freq box.



Finally, on the red drop down menu, select the 95% confidence interval.



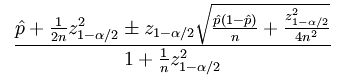
The resulting JMP output…



Questions:

1. What is the 95% Wilson confidence interval for this problem?
2. Interpret this 95% confidence interval in laymen’s terms.
3. How does the Wilson confidence interval differ from the simulation interval or the interval using normal theory methods?
4. Verify the calculations for the Wilson confidence interval computed in JMP. Use the following values from our example.

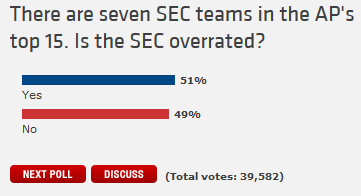
Wilson’s Formula:



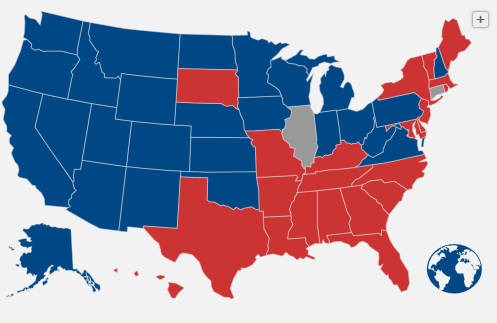
Example 5.2: ESPN regularly includes polls on its web site. These polls rarely mention the margin-of-error which to a statistician is unfortunate. The link below is the polling site for ESPN.

<http://espn.go.com/sportsnation/polls>

Consider the following question posed on ESPN’s site regarding the ranking of football teams from the SEC (Southeastern) Conference.



The following map is shows the most popular outcome for each state regarding this poll.



Consider the following results from Georgia, South Dakota, and Illinois.

|  |  |  |
| --- | --- | --- |
| Texas | South Dakota | Illinois |

Questions

1. What are the colors for Georgia, South Dakota, and Illinois on the graphic presented by ESPN?
2. What does it mean if a state has color blue? How about red?
3. What does it mean if a state is gray?
4. Notice that not very many of the states are gray. Why might this be the case? In particular, what guidelines do you think ESPN uses in identifying a gray state?
5. Consider the outcomes from South Dakota. Your friend makes the following statement, “I don’t like the fact that South Dakota is red because there is only a 4% difference between No and Yes. The outcomes may be the other way around on another sample in which case South Dakota would switch to blue. As a statistician, I agree with your friend. Why should we worry about “another sample” in a situation like this?
6. Calculate the appropriate margin-of-error for the poll in Georgia.

Step 1: Identify the statistic of interest

 =

Step 2: Calculate the **standard error** for 



Step 3: Calculate the **margin of error**.



Step 4: Compute the lower and upper endpoint using Approach #2: Normal Theory as follows:

Lower endpoint = =

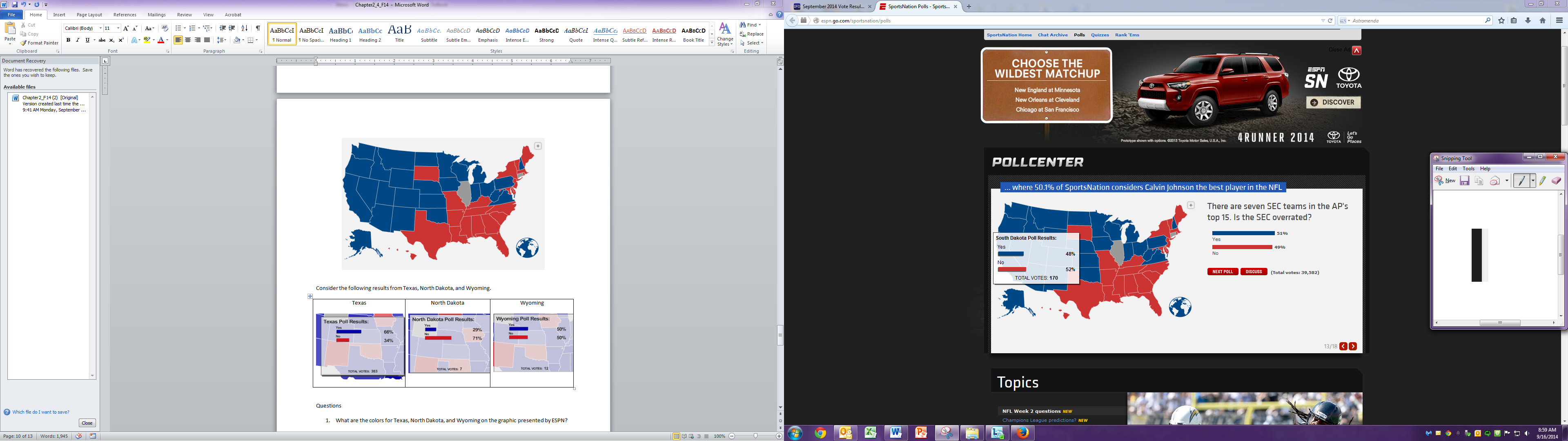
Upper endpoint = =

Sketch out the lower and upper endpoint on the number line below. Identify the margin-or-error on your sketch.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

­­­­­­­­­­­­­­­­­­­­­­­­­­

1. Use the margin-of-error for Georgia to defend the following statement: “We can be confident that over repeated samples, more people in Georgia will answer “No” to this question than “Yes”.” Explain.
2. Consider again the outcomes for South Dakota. Compute the margin-of-error to defend the following statement: “We cannot be certain that over repeated samples, more people in South Dakota will answer “No” to this question than “Yes”.” Explain.



1. Your friend makes the following statement: “South Dakota should be colored gray instead of red.” Why is this statement correct from a statistical perspective?

**Version #3: Binomial Exact Approach**

The binomial exact approach is another method of obtaining a 95% confidence interval. There are certain situations in which the normal theory approach does not work well. Statisticians have attempted to find guidelines for the use of the normal theory methods; however, all too often these guidelines get ignored or on the other extreme are taken too literally.

The most referenced guideline for the normal approximation to the binomial is having

Verify whether or not these guidelines were met for the above St. Paul School District example.

Example 2.2.2 Consider the following story in the Winona Daily News. This article discussed and compared the number of minority employees and the hiring of minorities at Winona County, Winona Area Public Schools, and Winona State University.



A summary of the information provided in this article follows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Employees | | Applicants Pools (2004-2010 for Winona County, 2010 for WSU) | |
| Employer | Total Number | Number of Minorities | Total  Number | Number of Minorities |
| Winona County | 332 | 2 | 3700 | 139 |
| Winona Area Public Schools | 642 | 11 | Data not provided | |
| Winona State University | 988 | 52 | 3940 | 580 |

In addition to this information, the article stated that Winona County has approximately 50,000 residents of which 4.5% are minorities.

Suppose interest lies in the variability present in the proportion of Winona County employees who are minority. Verify whether or not the guidelines for the normal approximation to the binomial are met for this example.

The lack-of-normality in the distribution can be seen in the simulated distributions. In the following setup there are two outcomes (Minority and Non-Minority) and the proportion used on the spinner for Minority is

We can see from the following graphs that the resulting distribution does not follow a normal curve very well.

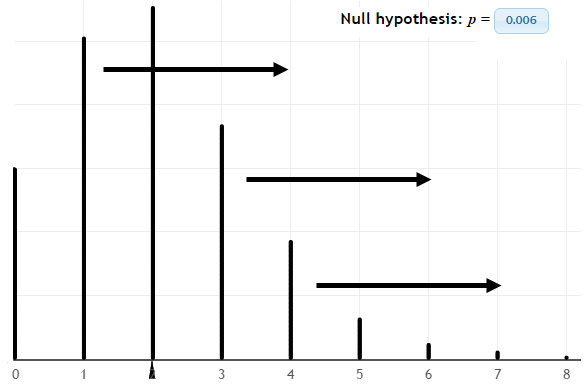
|  |  |
| --- | --- |
| Simulated Distribution | Simulated Distribution with Normal Curve |

Questions:

1. For what values (i.e. Counts) does the normal distribution fail?
2. Why is the left side of the normal distribution truncated in this example?

The idea behind the binomial exact approach to finding a confidence interval for a proportion is that we want to find all values for the proportion on our spinner that would result in 2 being a reasonable value. Certainly, in our pictures above, 2 is a reasonable value when the proportion on our spinner is set to 0.006.

For this example, we need to move the distribution to the right until our observed outcome is no longer a reasonable value.

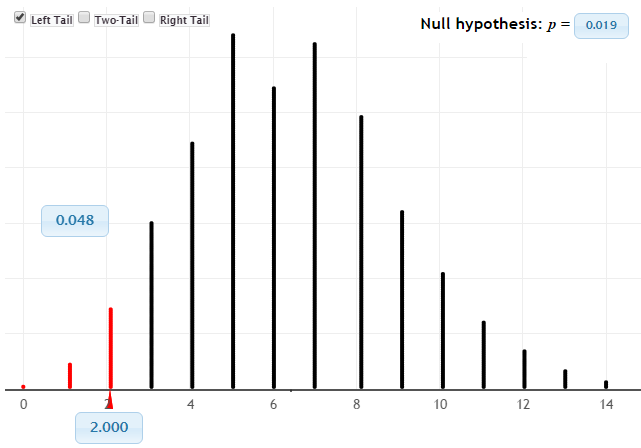


Questions:

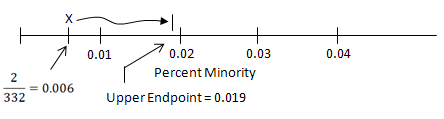
1. How can we shift this distribution to the right? Explain.

|  |  |
| --- | --- |
| Parameter values on Simulation for Proportion Minority | For what parameter values is 2 no longer a reasonable value? |
| 0.01 |  |
| 0.015 |  |
| 0.02 |  |

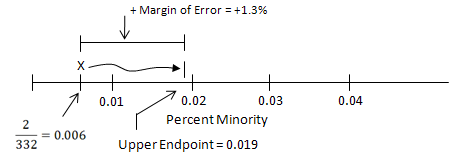
After trying several different values, a value of 0.019 or 1.9% seems to work well.



Thus, percent or proportion values that reach up to 1.9% result in 2 being a reasonable value.

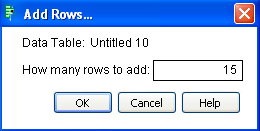


Thus, our 95% confidence interval using the binomial methods would reach up to 0.019 or 1.9%. Such a interval would result in a one-sided margin of error. The margin of error here would be computed as 1.9% - .6% = 1.3%.



Getting this done in JMP

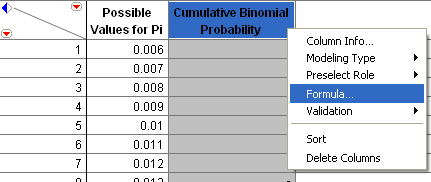
Open a new data table in JMP and add 15 rows to your table.



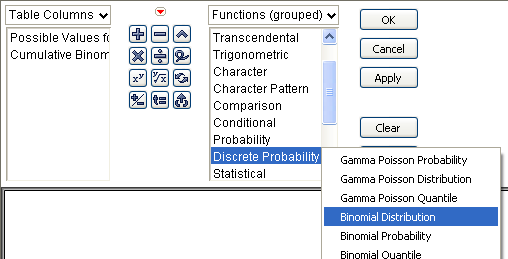
Label one column as the Possible Values for Pi. Use the sequence function to give a range of values from 0.006 up to 0.02. The step size should be set to 0.001.



Next, create a second variable that will contain the cumulative probabilities for the appropriate binomial distribution.



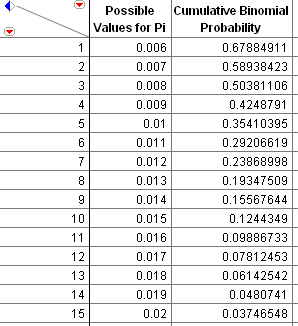
Select Binomial Distribution under Discrete Probability.



The value of Pi will vary here (in contrast to values of k). Use the Possible Values of Pi column for p, the total sample size is n=332 and we had k=2 minorities in our data.



The following values are returned from JMP.



Questions:

1. For what values of Pi is the cumulative probability less than 0.05? For these values of Pi, 2 would be considered an unusual or unlikely outcome.
2. For what values of Pi is the cumulative probability for 2 or less more than 0.05? These values of Pi are used to construct the 95% binomial confidence interval for a proportion.
3. Does the confidence interval here agree with what we discovered in the simulation? Explain.

**Understanding Margin of Error**

The margin of error formula for the normal approximation is



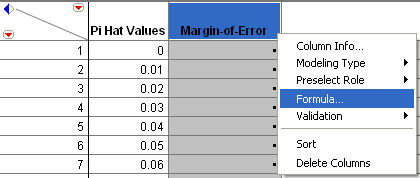
**Question**: For which value of pi-hat is this the margin of error maximized?

Method1: Draw a plot.

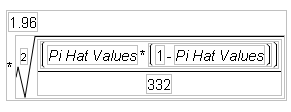
Step 1: Create a sequence of pi-hat value in JMP.



Step 2: Create a margin of error column an enter the formula in the formula editor as follows.



Margin of Error Formula



The resulting values with a plot.

|  |  |
| --- | --- |
|  |  |

Question:

1. For what value of pi-hat is the margin of error maximized? What is the maximum margin of error for a proportion when n = 332?

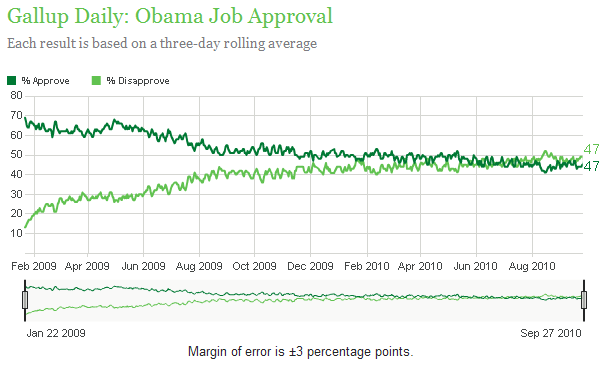
Method1: Calculus -- taking a derivative and setting equal to 0….

Setting the derivative equal to 0….

which yields…

Next, multiple through by yields

Consider next how the margin of error is influenced by the sample size. Consider the following graphic from Gallup.com. The reported proportion that approve Obama is about 47%. The reported margin of error for this poll was +/- 3%.



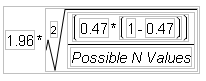
To estimate the sample size used in this poll, we need to create two column in JMP.



Step 1: The possible N values should have the following formula.



Step 2: The margin of error formulas is again typed into the formula editor for Margin of Error.



Why is 0.47 used in this formula? Explain.

|  |  |
| --- | --- |
| The effect of sample size on margin of error for Gallup poll with a proportion of 47% and a +/- 3% margin of error. | What happens to the margin of error when 50% is used for the proportion instead of 47%? |