CONFIDENCE INTERVAL FOR A PROPORTION

When carrying out hypothesis tests, we are testing claims about population parameters. Sometimes, however, our goal is to *estimate* a parameter of interest. Statisticians typically do this with a **confidence interval,** which is simply a range of likely values for the parameter of interest. The big difference between hypothesis testing and confidence intervals is as follows: the construction of a confidence interval does NOT require any hypotheses concerning our population parameter of interest.

Example 2.5: Star Tribune Poll Regarding Legalization of Medical Marijuana

In February of 2014, a Star Tribune Minnesota poll asked a random sample of adults in Minnesota, "Do you support or oppose legalizing marijuana for medical purposes in Minnesota?" Of the 800 adults surveyed, 408 (51%) answered that they would support this.

> # Minnesota Poll: Majority support legalization of medical marijuana
>
> Article by: JENNIFER BROOKS , Star Tribune | Updated: February 18, 2014 - 5:42 AM
>
> A bare majority of Minnesotans say the state should legalize medical marijuana.
>
> According to a new Star Tribune Minnesota poll, 51 percent of Minnesotans support legalization for medicinal uses, while 41 percent oppose a change in state drug laws. Twenty states, plus the District of Columbia, already allow doctors to prescribe marijuana for a range of medical conditions, from cancer to epilepsy and post-traumatic stress disorder.

Source: http://www.startribune.com/politics/statelocal/245910931.html

Questions:

Ultimately, the Star Tribune conducted this poll because they wanted to estimate the proportion of adults in Minnesota that felt (at the time of the survey) that medical marijuana should be legalized. Keep this in mind as you answer the following questions.

1. What is the population of interest?

2. What is the parameter of interest?

3. What is the sample?

4. What is the sample statistic of interest?

Based on the results of this poll, our best guess for the proportion of all Minnesotans that feel medical marijuana should be legalized is $\hat{\pi}$ = 408/800 = 51%. This is called a **point estimate**.

We know, however, that if a different sample of 800 Minnesotan adults had been polled, this point estimate would likely change. So, instead of using only this point estimate, we'll also use statistical methods to estimate the parameter of interest with a *range* of likely values (this range is called the **confidence interval**).

To get this range of likely values, we'll start with our point estimate (i.e., the sample statistic, $\hat{\pi}$ = 51%). This is our best guess so far. Then, we can use Tinkerplots to get an idea of how much we expect this statistic to change from one random sample to the next because of natural sampling variation. The distribution of sample proportions obtained by repeated sampling is shown below.
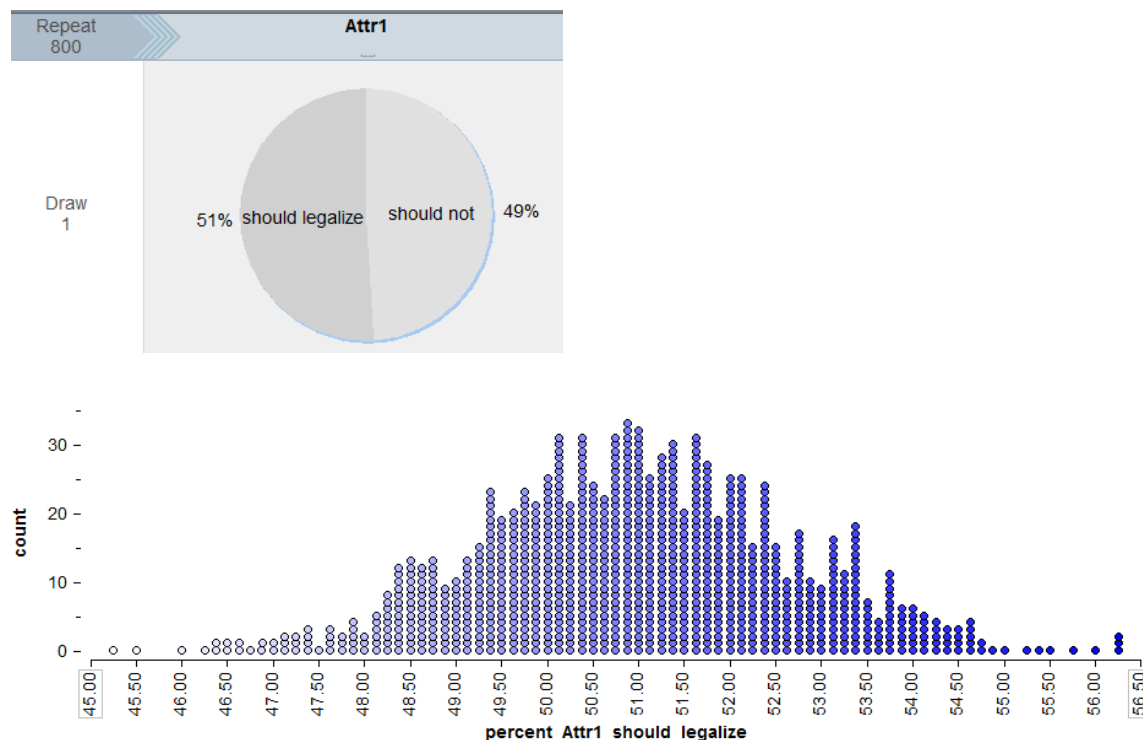
To get a 95% confidence interval for the parameter of interest, we find both the <u>lower endpoint</u> and the <u>upper endpoint</u> that separate the "middle" 95% of this distribution from the outer 5%.
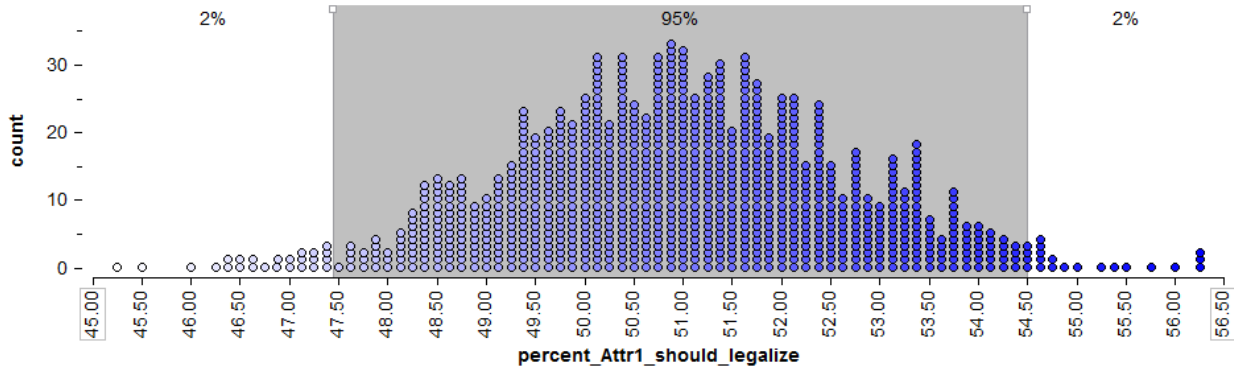
Questions:

1. Find the lower and upper endpoints for a 95% confidence interval for the parameter of interest, π = the proportion of all Minnesotan adults that felt medical marijuana should be legalized.

   Lower endpoint: _____

   Upper endpoint: _____

   Sometimes, this 95% confidence interval is also written as follows:

2. The sample statistic (also called the point estimate) was $\hat{\pi}$ = 51%. This describes only those in our *sample* (i.e., 51% of those surveyed supported the legalization of medical marijuana. Note that the confidence interval estimate, on the other hand, can be used to describe the *population*. Write an interpretation of this 95% confidence interval.

The General Form of a Confidence Interval

Note that in general, a confidence interval can be expressed in terms of both the *point estimate* and the associated *margin of error*, which is a measure of the accuracy of the sample statistic.

In general, a confidence interval can be computed as follows:

| General form of a confidence interval |
| --- |
| sample statistic ± margin of error |

For this example, recall that the sample statistic is $\hat{\pi}$ = 51%. The Star Tribune reported the margin of error as follows:

> The poll surveyed 800 Minnesota adults between Feb. 10 and 12 and has a margin of sampling error of plus or minus 3.5 percentage points.

Note that the confidence interval can be expressed as 51% ± 3.5% (i.e., 47.5% ≤ π ≤ 54.5%). This agrees with what we saw under repeated sampling on the previous page.

The Wald Interval

Our earlier discussion involving Tinkerplots helps us understand the intuition behind a confidence interval. Note that a statistician would not necessarily compute a margin of error (and thus a confidence interval) in this way. Instead, to calculate a confidence interval for a proportion, a statistician often makes use of the **normal distribution** (i.e., the bell-shaped curve). Note that the normal curve approximates the distribution of sample proportions obtained from Tinkerplots quite well:
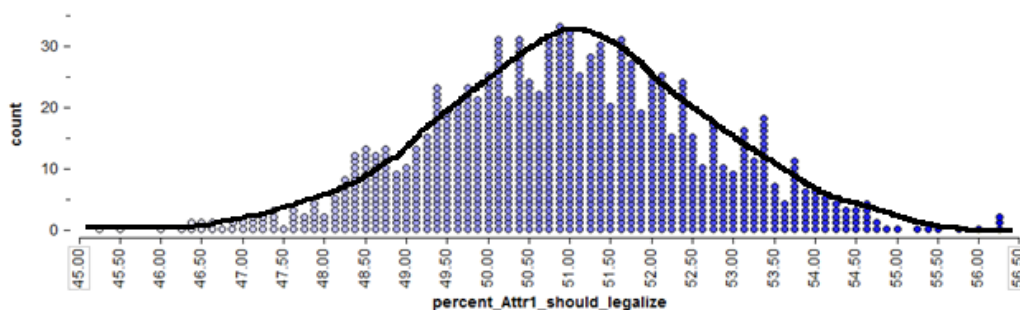
If we can reasonably assume that this will be the case (later, we'll discuss certain conditions under which the normal approximation may not be reasonable), then we don't even need the "dots" from Tinkerplots to find the middle 95%. We can use mathematical theory and a procedure known as the **Wald interval**, instead, to find the margin of error.
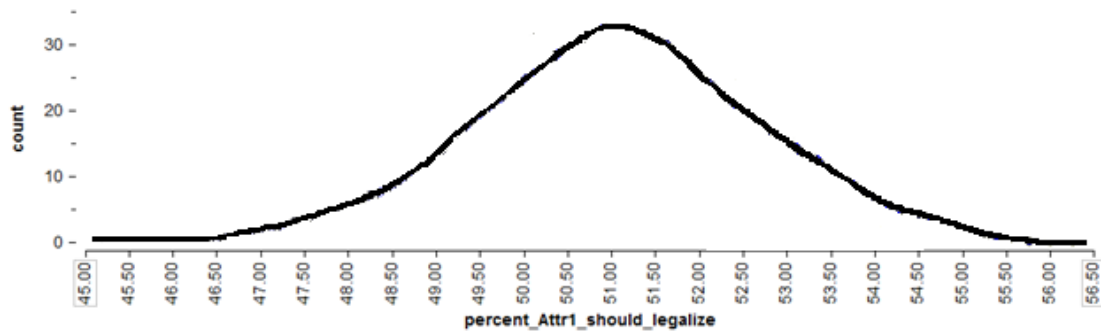
To find the endpoints for a 95% confidence interval, we need only find the values on the x-axis of the above graph that separate the middle 95% from the rest.

We can use Wald's method to do this with the following steps:

1. Start with the **point estimate,** $\hat{\pi}$:

2. Calculate the **standard error** associated with this point estimate:

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

3. Calculate the **margin of error**: This is defined as 1.96 standard errors for a 95% confidence interval (it's good enough to think of this as "about two" standard errors).

$$1.96 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} =$$

4. Find the **endpoints** of the confidence interval:

Lower endpoint = $\hat{\pi}$ - margin of error = _____

Upper endpoint = $\hat{\pi}$ + margin of error = _____

Finally, note that if you desired to construct a 90% or a 99% confidence interval, instead, the formula would change slightly. These formulas are shown in the table below.

Formulas for 90%, 95%, and 99% Confidence Intervals for a Proportion
For a 90% confidence interval, the margin of error is defined as 1.645 standard errors. The 90% confidence interval is then given by $\hat{\pi} \pm 1.645\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$.
For a 95% confidence interval, the margin of error is defined as 1.96 standard errors. The 95% confidence interval is then given by $\hat{\pi} \pm 1.96\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$.
For a 99% confidence interval, the margin of error is defined as 2.575 standard errors. The 99% confidence interval is then given by $\hat{\pi} \pm 2.575\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$.

Questions:

1. What happens to the margin of error when the confidence level decreases? For example, what will happen if we were to construct a 90% confidence level instead of 95%?

2. What will happen to the margin of error and the confidence interval if we increase the level of confidence?

3. What will happen to the margin of error if the sample size increases? Explain.

4. Recall the headline used by the Star Tribune and one of the first sentences of their article.

Minnesota Poll: Majority support legalization of medical marijuana

Article by: JENNIFER BROOKS , Star Tribune | Updated: February 18, 2014 - 5:42 AM

Carry out the binomial exact test to determine whether the poll provides evidence that this is really the case.

Null and Research Hypotheses	H_o: H_a:
p-value	
Conclusion	

5. Does the confidence interval agree with the decision at which you arrived using the hypothesis test? Explain.

General Comments Concerning the Confidence Interval for a Single Proportion:

1. A confidence interval allows us to estimate the population parameter of interest (note that the hypothesis test does NOT allow us to do this). Therefore, when available, a confidence interval should always accompany the hypothesis test.

2. Several methods exist for constructing a confidence interval for a binomial proportion. We have considered a method known as the "Wald" interval. **Note that this method does not typically perform well for very small sample sizes or when the point estimate is very close to either zero or one.**

    This is because the normal curve does <u>not</u> approximate the binomial distribution very well in these situations. For example, consider the binomial distribution with $\pi = .10$ and $n = 20$:
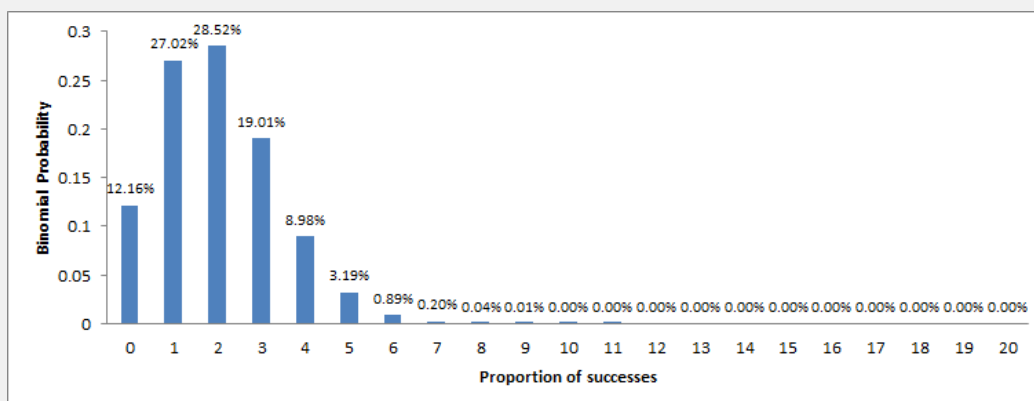
    If our point estimate had been $\hat{\pi} = .10$ and the sample size had been $n = 20$, then the Wald method would not have provided us with a very reliable interval estimate of the true population proportion. Other alternatives (such as adjusted Wald intervals, binomial exact intervals, or score intervals) will be more appropriate in these case.

    One rule of thumb is that as long as $n \times \hat{\pi} \geq 5$ and $n \times (1-\hat{\pi}) \geq 5$, the Wald interval should work reasonably well. If these conditions aren't met, you shouldn't use the Wald method.

Example 2.6: Helper vs. Hinderer, Revisited

Recall that in this study, 14 of the 16 ten-month-old infants chose the helper toy. Suppose that the researcher wants to estimate the parameter of interest, π = the proportion of all ten-month-olds that would choose the helper (you could also think of this as π = the probability that a randomly selected ten-month-old would choose the helper toy).

First, check the conditions to determine whether the Wald method is appropriate.

For illustrative purposes only, use Wald's method to obtain a 95% confidence interval for π.

1. Start with the **point estimate, $\hat{\pi}$**:

2. Calculate the **standard error** associated with this point estimate:

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

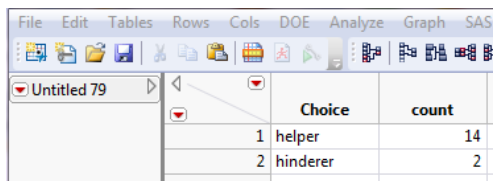3. Calculate the **margin of error**:

$$1.96 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} =$$

4. Find the **endpoints** of the confidence interval:
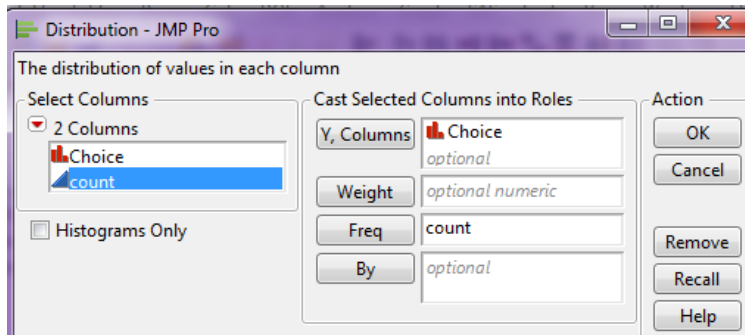
   Lower endpoint =

   Upper endpoint =

Note that we could also use JMP to find this confidence interval. JMP uses what is known as the **score method** instead of **Wald's method**, which would be more appropriate in this case.
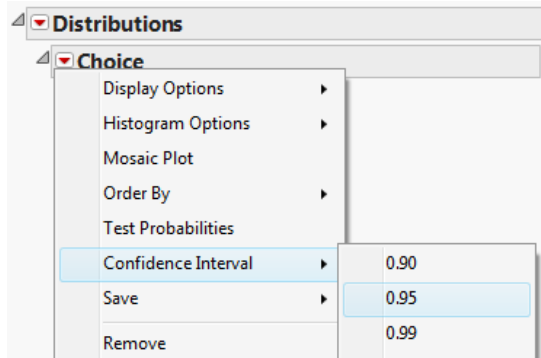
Select **Analyze > Distribution** and enter the following:

Then, click the red drop-down arrow next to the variable name (in this case "Choice") and select
Confidence Interval > .95.

JMP returns the following:

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
helper	14	0.87500	0.639772	0.965023	0.950
hinderer	2	0.12500	0.034977	0.360228	0.950
Total	16				

Note: Computed using score confidence intervals.

Interpret this confidence interval.

Comment: The hypothesis test for this problem provided evidence that the true population
parameter was greater than 50%. The confidence interval tells us this and more…
 HOW MUCH GREATER it is!