JMP Academic Case Study 008

Siblings

Log transformation, One Sample t Confidence Interval

Produced by

Dr. DeWayne Derryberry, Idaho State University
Department of Mathematics

jmp STATISTICAL DISCOVERY

Siblings
One Sample t Confidence Interval

Key Ideas

Logarithmic transformation, inverse transformation, mean versus median, power, robustness of t procedures, sampling distribution.

Background

The following data are a mixture of surveys from a small liberal arts school on the West Coast. Each semester students in introductory statistics courses complete a survey. One of the questions is the number of siblings in their immediate family. The results of the survey are similar each semester, but results have been combined in a haphazard manner so that there is no chance any particular class can be identified.

The Task

Estimate the average number of siblings for the general population.

The Data **siblings.jmp**

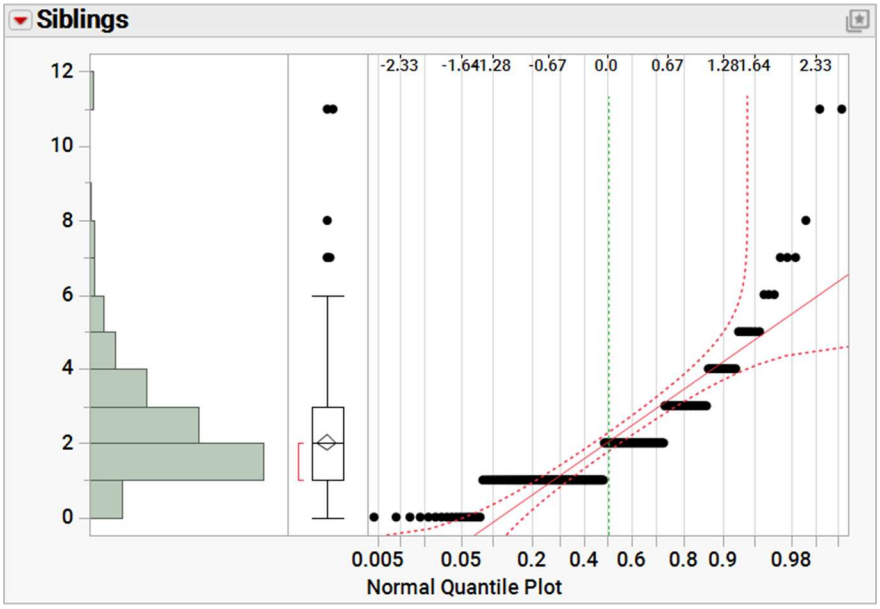
The data table contains the reported number of siblings for 220 students.

Siblings Number of siblings reported for each student

Analysis

An analysis of Siblings indicates that the variable is right skewed and only takes on discrete values.

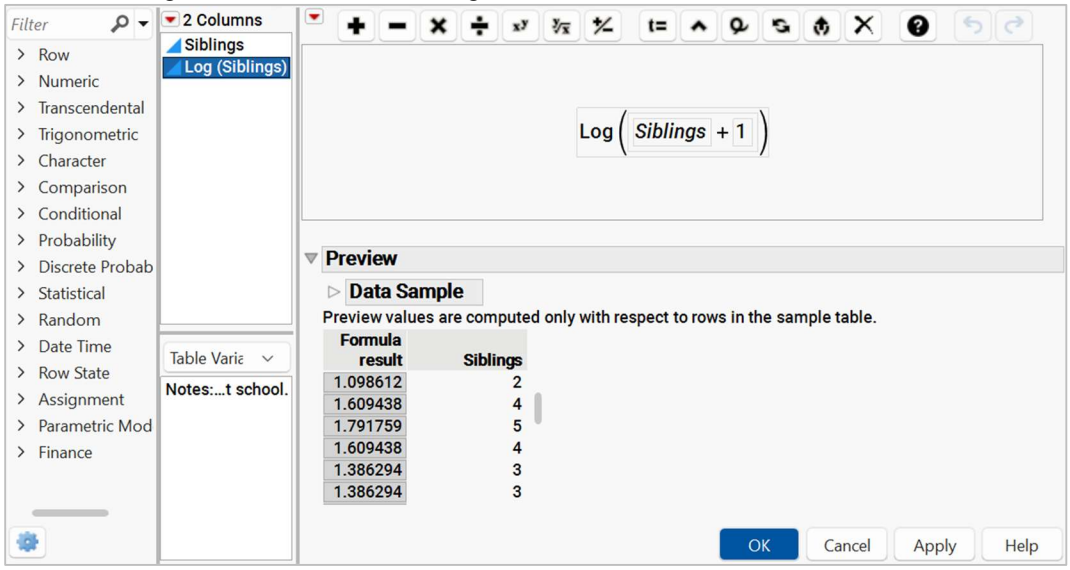
Exhibit 1 Distribution of Siblings



(Analyze > Distribution; select Siblings as Y, Columns and click OK. Select Normal Quantile Plot from the red triangle for Siblings. For a horizontal layout select Stack under the top red triangle.)

Since we're interested in drawing inferences about the number of siblings, we'll consider a transformation that normalizes the data. Because these data display a right skew, a logarithmic transformation makes sense. There are zeros in these data, but changing the data from "siblings" to "siblings + 1" is a simple remedy, because "siblings + 1" is just the number of children in a family (my siblings plus me). A new variable, Log(Siblings) is defined in a new column (Exhibit 2).

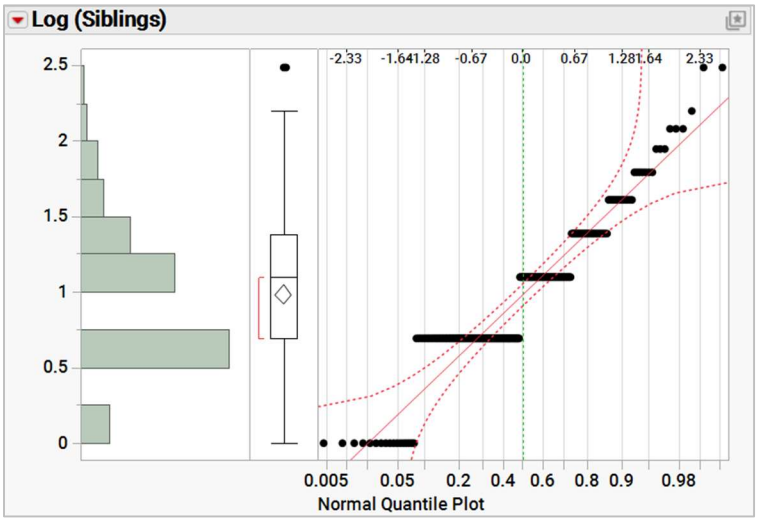
Exhibit 2 Log Transformation of Siblings



(Create a new column in the data table, and rename it Log(Siblings). Right click on the column header, and select Formula to open the Formula Editor. To create the formula: From the Functions(grouped) list select Transcendental, Log > Select Siblings from the columns list > Select the plus sign on the key pad > Type "1" > Click OK. (This could be done by right clicking column header and going to New Formula Column > Log > Log x+1.)

The new variable still displays unavoidable discreteness. No transformation can alter the fact that a substantial proportion of the data is at just three values (0, 1, and 2), but the skewed has been lessened (Exhibit 3).

Exhibit 3 Distribution of Log(Siblings)



(Analyze > Distribution; select Log(siblings) as Y, Columns and click OK. Select Normal Quantile Plot from the red triangle.)

Do we need to transform the data? Yes and no. It is not wrong to analyze the data on the untransformed scale. Because the sample size is large (n = 220), the central limit theorem assures us, that unless the data is pathologically skewed, the sample mean follows a normal distribution with the usual mean and variance (see sampling distributions). In other words, t procedures are robust to the normality assumption.

The notion of robustness is important and often not well understood. All statistical tests and confidence intervals are based on a number of assumptions. Every formula in statistics was derived in the purely mathematical sense. That is, beginning with some minimal set of assumptions, a variety of clever mathematical relationships were used to get the formulas found in books. Robustness is about taking an additional step – once we have the formula, do we really need the assumptions?! In other words, which assumptions, if any, could be dropped and have the formula still perform as expected.

Although the usual t formulas were derived based on an assumption of normal data, simulation studies have found that, unless the data are heavily skewed or there are extreme outliers, the t procedures produce reliable results:

- Confidence intervals capture the unknown population parameter at the advertised confidence level, and
- *p*-values are uniform[0,1] when the null hypothesis is true and favor smaller values when the alternative hypothesis is true.

However, some estimation procedures make better use of the data than others. Making better use of the data means:

- Producing confidence intervals that are narrower, yet still capture the unknown population parameter the correct proportion of the time.
- *p*-values that are still uniform[0,1] when the null hypothesis is true, but produce even smaller *p*-values when the alternative hypothesis is true.

In this case, the confidence interval using the transformed data is narrower (on the original scale) than the interval using the original data. Transforming right skewed data, to produce more bell-shaped data, almost always has this happy outcome.
We can directly read the confidence interval for the data on the original scale (Exhibit 4). We estimate the average number of siblings for the entire population to be between 1.79 and 2.24 persons.

Exhibit 4 Confidence Intervals for Siblings and Log(Siblings)

Distributions

Siblings		Log (Siblings)	
Summary Statistics		**Summary Statistics**	
Mean	2.0181818	Mean	0.9823774
Std Dev	1.6851856	Std Dev	0.4845724
Std Err Mean	0.1136152	Std Err Mean	0.0326699
Upper 95% Mean	2.2421009	Upper 95% Mean	1.046765
Lower 95% Mean	1.7942627	Lower 95% Mean	0.9179898
N	220	N	220
N Missing	0	N Missing	0

(Analyze > Distribution > Select Siblings and Log (Siblings) as Y columns. Only Summary Statistics is shown above)

However, the confidence interval for Log(Siblings) is not directly interpretable. Because we transformed the data, we must invert this transformation to get an interval on the original scale. Inversion just means applying the inverse steps in exactly the opposite order. Since we transformed the data by adding 1 to each value and taking logarithms, we invert by exponentiating and then subtracting 1:

$$e^{0.9179898} - 1 = 1.50425$$
$$e^{1.046765} - 1 = 1.84842$$

We estimate the average number of siblings for the entire population to be between 1.50 and 1.85 persons. The confidence intervals are different in two ways:

- The transformed interval is narrower (0.35 versus 0.45), a good thing.
- The transformed interval is lower. Why is this?

The logic of transformations

I used the rather ambiguous word "average" in the wording of the confidence intervals above, and the equivocation was intentional. The t procedure is used for means, and the confidence interval on the original scale is an interval for the population mean. However, transformations involve some subtle reasoning. In an exercise you will be asked to verify the following idea: If you transform the data, the mean acts in unpredictable ways, whereas the median acts in quite predictable ways.

Key Idea: The median of the transformed data = the transform of the median of the data
The mean of the transformed data ≠ the transform of the mean of the data

When we transform the data, we hope to make it more normal, then the mean and the median are about the same. Therefore, on this scale, a confidence interval for the mean (which is what a t procedure gives us) is also an approximate confidence interval for the median (because the data is approximately symmetric). When we transform the data back to the original scale, it is the median (not the mean), that can be recovered on the original scale.

So, 1.79 to 2.42 is a confidence interval for the mean and 1.50 to 1.85 (from the inverse transformation) is a confidence interval for the median. The interval for the median is lower because, for right skewed data, the median is lower than the mean. Both are only approximate, because unless data is normal, all t procedure intervals are approximate.

Which is better? They are both intervals for a measure of central tendency, and almost all textbooks mention that, for right skewed data, the median is a better summary of centering than the mean. All other factors being equal, we take the narrower interval.

Summary

Statistical Insights

The fundamental idea is that, while t procedures are robust to the normality assumption, transformations of the data often produce better results. We can anticipate that the results will be better when we find a transformation that produces graphs showing the data has been made more normal (or at least more symmetric).

In this context "better" means: i) narrower confidence intervals, and ii) smaller *p*-values when the alternative hypothesis is true.

But – avoid data dredging. Do not try several procedures and/or transformations and pick the one with the narrowest confidence interval. That would be the definition of data dredging. Data dredging can often be used to get any conclusion you want, and is one of those ways to "lie with statistics." Match procedures with their assumptions. We should use the transformed data with the t procedure if the data looks more normal when the data is transformed.

It should be mentioned that the log transformation, although the most common transformation, is not the only one. For right skewed data, sometimes the square root transformation is better. There are also some specialized transformations for specific situations discussed in more advanced books. Finally, for extremely left skewed data and data with extreme outliers, nonparametric methods often work well.

Implications

Be careful about the population to which we are making inference. As we will see in the exercises, we cannot conclude this is the average number of siblings for all college students. For a number of reasons, families on the East and West Coasts are smaller than families in the intermountain West. Once you have an interval based on the data, you need to think about what larger group, if any, this sample would generalize to. This often requires expertise in the life or social sciences and is separate from statistical expertise.

JMP® Features and Hints

This case study uses the formula editor to transform a variable using the log function. The Distribution Platform was used to visualize the distributions of the original and log-transformed data, and to compare confidence intervals.

Note that in JMP, variables can be dynamically transformed in any dialog window. To do this, right-click on the variable in the column selection panel, select Transform and then select the transformation of interest. This creates a temporary variable, which will be shown in italics. To save the transformed data to the data table (and create the formula for the transformation), right-click on the transformed variable and select Add to Data Table.

Exercises

1. Given the following five data points: 2, 4, 8, 16, 32
   Create a JMP data table, then:
   a. Find the mean of the data.
   b. Transform the data using a logarithmic or square root transformation.
   c. Find the mean of the transformed data.
   d. Show: The mean of the transformed data ≠ the transform of the mean of the data.
   e. Find the median of the data.
   f. Find the median of the transformed data.
   g. Show (at least in this case, but it is generally true): The median of the transformed data = the transform of the median of the data.

2. Perform a similar analysis for the SiblingsExercise2.jmp data, which is from a survey of students at an intermountain school. Consider two transformations "Log(Siblings + 1)" and "Square Root(Siblings)".

   a. Compare the plots of the data on the original scale and the two transformed scales. Rank them from most to least normal/symmetric.

   b. Complete the following table for confidence intervals on the original scale (you will need to perform inverse transformations in the cases where transformations were made).

Approach	**Lower limit of the interval**	**Upper limit of the interval**	**Width = upper – lower**
Original data			
Square root transformation			
Logarithmic transformation			

   How do the intervals compare? How does this relate to your answer in part 2a? (Write a sentence or two.)

   c. Why are the intervals for the transformed data at a lower set of values than the interval on the original scale?

   d. Based on the computations already done, is there evidence the intermountain school differs in overall size of family relative to the West coast school? (Hint: Do the confidence intervals overlap?)