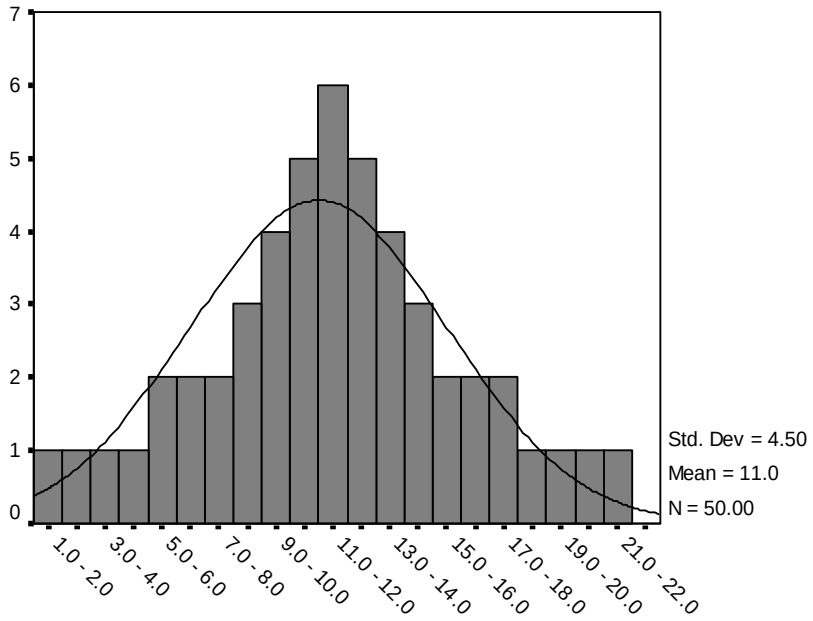
#15 Testing & Fixing for Normality

Purpose and Introduction
According to Cramer (1994) and others (see Sheet #3), one of the three assumptions of data used in 'parametric' tests (like t-tests, ANOVA, and regression) is that values are distributed normally. The purpose of this short paper is to outline a means of testing for univariate (single variable) normality and to suggest several techniques for resolving non-normal distributions.

Normality
Cramer (1994) has argued that of the three assumptions of data used in parametric testing: interval/ratio quality; equality of variance between groups; and normality, singular violations may not stop the researcher pursuing parametric testing. The point made by Cramer (see Sheet #3) is that simultaneous violations of normality and equality of variance assumptions is problematic and should lead researchers to try non-parametric testing, or transformations of the data.

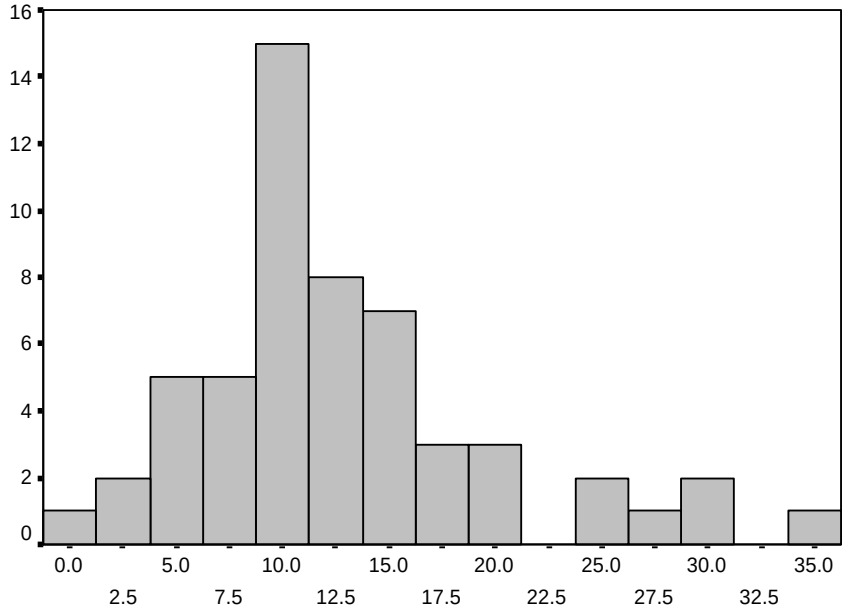
Normality refers to the 'shape' of the distribution of the data. Consider a histogram of values for one variable. By drawing a line across the 'tops' of the bars in the histogram, we are able to see the 'shape' of the data. When the 'shape' forms a 'bell' shape, we generally call this a normal curve. The figure below is approximately normally distributed. A perfect, normally-distributed 'bell-curve' is superimposed over the data.



Two Dimensions of Normality

Looking back at the perfect normal curve, there are two dimensions that we can test to examine for normality. The first is 'Skew' or 'Skewness', and the second is Kurtosis.

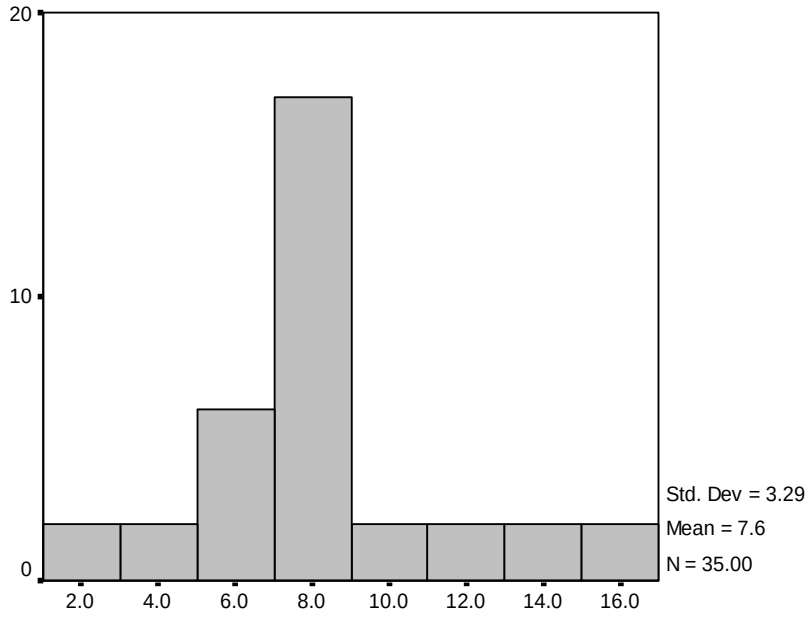
Skew
A variable that is positively skewed has large outliers to the right of the mean, that is, greater than the mean. In that case, a positively skewed distribution 'points' towards the right.



A negatively skewed variable 'points' towards the left, because it has outliers that are relatively lower than the mean. In cases of skewness, the mean is influenced by the position of outliers. The researcher may wish to examine the 5% trimmed mean (in Analyse, Explore command in SPSS) to see the influence of cutting the top 5% of scores and the bottom 5% of scores and recalculating the mean. Another option, possibly recommended by Tabachnick & Fidell, is to remove scores that are greater than + or – 3 standard deviations from the mean (First transform/save the variable in its standardised form – values are converted to standard deviations from the mean- through, Analyse Menu, Descriptives, then a small tick-box requesting variables be saved as standardised scores. Then second, the Data Menu, Select Cases, If condition is satisfied, using the new standardised scores as a condition < 3 or > - 3 as the filter). This is of course assuming the variable is approximately normally distributed- which is the purpose of this and following text. Potentially this latter identification and elimination of outliers could be conducted and the following section be re-attempted to gauge the affect of outlier elimination on normality.

Kurtosis
Whereas skewness examines the horizontal movement of a distribution from a perfect normal 'be;; shape, kurtosis examines the vertical displacement. A variable that is positively kurtic (has a positive kurtosis) is lepto-kurtic and is too 'pointed'. A variable that is negatively kurtic is platy-kurtic and is too

'flat'. The figure below shows a clear sign of positive kurtosis (lepto-kurtic) variable.



As all things statistical lead to questions about probability, so too are our estimates of skewness and kurtosis subject to error. SPSS can provide you with numerical estimates of skewness and kurtosis with their standard errors.

Assessing Normality

A perfectly normal distribution will have a skewness statistic of zero. Positive values of the skewness score describe positively skewed distribution (pointing to large positive scores) and negative skewness scores are negatively skewed. A perfectly normal distribution will also have a kurtosis statistic of zero. Values above zero (positive kurtosis score) will describe 'pointed' distributions, and values below zero (negative kurtosis scores) will describe 'flat' distributions. Like all estimates, we are unlikely to ever see the values of zero in either skewness or kurtosis statistics. The real question is whether the given estimates vary significantly from zero. For this question we need the standard error of skewness when looking at our skewness score, and similarly the standard error of kurtosis when examining our kurtosis statistic.

In SPSS, the Explore command provides skewness and kurtosis scores.

			Statistic	Std. Error
D	Mean		10.9796	.6491
	95% Confidence Interval for Mean	Lower Bound	9.6745	
		Upper Bound	12.2847	
	5% Trimmed Mean		10.9773	
	Median		11.0000	
	Variance		20.645	
	Std. Deviation		4.5437	
	Minimum		1.00	
	Maximum		21.00	
	Range		20.00	
	Interquartile Range		6.0000	
	Skewness		.014	.340
	Kurtosis		-.126	.668

The construction of a 95% confidence interval about a skewness score (or a kurtosis score) enables the evaluation of the variability of the estimate. The key value we are looking for is whether the value of 'zero' is within the 95% confidence interval. Strictly speaking, we are testing a null hypothesis that the skewness estimate is not significantly different from a value of zero- that is- our score is from a distribution whose population mean is zero. Thus a quick check on skewness and then kurtosis, using 1.96 or 2, for short, times the standard error of the estimate used evaluates the null hypothesis. Using the output on the previous page for a skewness stastistic of 0.014 with a standard error of 0.34 a 95% confidence interval is constructed. First the upper bound is 0.014 + (2 times the standard error of 0.34) = 0.694 and then the lower bound is 0.014 – (2 times 0.34) = -0.54 . Thus the 95% confidence interval for the skewness score ranges from –0.54 to +0.69 . The primary value of interest is zero. If zero is within our bounds (confidence intervals) then we can accept the null hypothesis that our statistic is not significantly different from a distribution of zero. As zero is within our bounds, we can indeed accept the null hypothesis and support the notion that our skewness statistic is not significantly different from a score of zero. If it was outside the bounds, we have cause to reject the null hypothesis, and conclude that our score is unlikely to come from a distribution whose average is zero.

The same process is undertaken for the kurtosis, using the kurtosis statistic and the standard error. A quick look reveals that zero will once again fall easily within a 95% confidence interval of the mean. This means that our kurtosis statistic is not significantly different from a value of zero and our distribution, here, is not distorted vertically from a normal distribution. Thus both skewness and kurtosis are not problematic for this variable. However, this was not surprising, given I constructed this variable for demonstration of the normal curve. You may find that many non-categorical variables will not be normally distributed.

SPSS actually provides a test, incorporating both skewness and kurtosis at once. This makes the manual testing of null hypotheses about skewness and kurtosis redundant- however it is useful to have a 'hands-on' ability with normal curve distortions, and I find the above process often provides the researcher with more insight and understanding than the overall test. On to the SPSS test for normality.

In SPSS, Analyse Menu, Explore Command, Plots Button, 'Normality Test with Plots' provides two tests for the normality of a variable.

The first is the Kolmogorov-Smirnov test for normality, sometimes termed the KS Lilliefors test for normality. The second is the Shapiro-Wilk's test for normality. The advice from SPSS is to use the latter test when sample sizes are small (n < 50). The null hypothesis, that there is no difference between your variable distribution and a normal distribution, is evaluated. The output, using the data from above, is presented below.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
D	.072	47	.200*	.984	47	.865

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

With less than 50 cases the Shapiro-Wilk's statistic for normality will be used of the KS Lilliefors test. The significance of 0.865 means we cannot reject the null hypothesis. What is the null hypothesis- that

our distribution is not significantly different from a normal distribution. This result was expected as our individual examination of the skew and kurtosis revealed normal distribution as well.

What to do if Not Normal

According to some researchers, *sometimes* violations of normality are not problematic for running parametric tests. However, when violation occurs in conjunction with violations of other assumptions (like interval/ratio level of measurement & equality of variance), the researcher becomes more concerned. The use of non-parametric statistical alternatives, or transformation of the data may be useful.

When transforming the data, the ease of communication of statistics, like means and standard deviations or errors, becomes more difficult. It is more difficult to write about the average logged number of employees in firms than it is to use the average number of employees (I did this in my PhD with the assistance of a long footnote that interpreted what logged values of 1, 2, 3, 4, 5 etc were equivalent to in terms of numbers of employees). You will find many texts make the case for NOT using transformations on data in order to retain a better interpretability of the numbers. But then again, many texts also take an instrumental approach (econometrics) that transformation may ease and solve for the violation of assumptions concerning data distribution.

There are different types of transformations that might be used.

Transforming the Data

The best resource for transforming data is a chapter from Tabachnick & Fidell (1996). Essentially the following table is summary of their diagnosis and solution for initial non-normal distributions of data.

– **To solve for Positive Skew**

Slightly Skewed = **Square root**

- problematic with negative numbers with positive numbers (solution- add a constant to make all values greater than 0)
- numbers between 0 and 1 increase while numbers above 1 decrease (this maybe a useful transformation and on other occasions maybe undesirable- solution add a constant to make all values greater than 1)

Skewed = **Logarithmic**

- Base 10 or exponent

Absolutely Skewed = **Inverse**

- The inverse is the value represented as a fraction- such that z = 1/z
- Large numbers become small, small numbers large.
- Data reflection* maybe useful because inverting will reverse the order of all values

– **To solve for Negative Skew**

Reflect the data first- then undertake the transformations above, then reflect* the data back.

- *(to reflect first, multiply by –1 and add a constant to make greater than positive 1- thus retaining the original order of the data).*