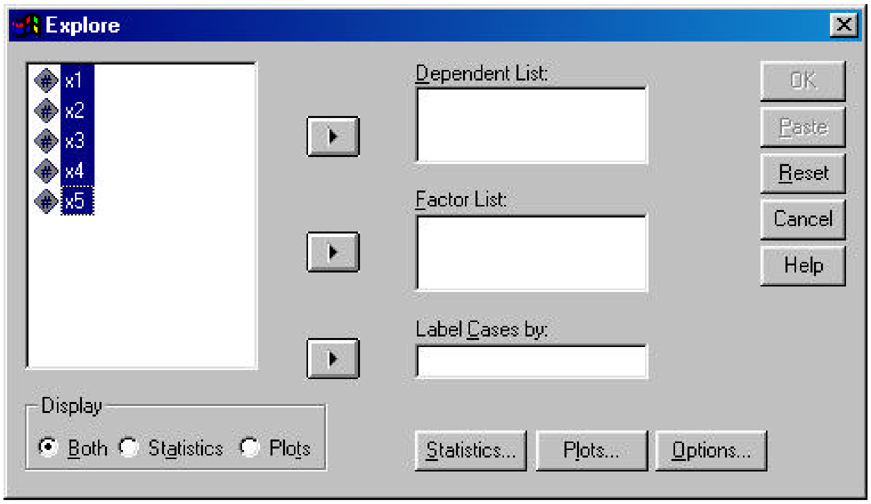
Univariate Data Cleaning

Taking a "Look" at your data

Useful analyses for taking a look at the univariate properties of your variables are "Explore" …

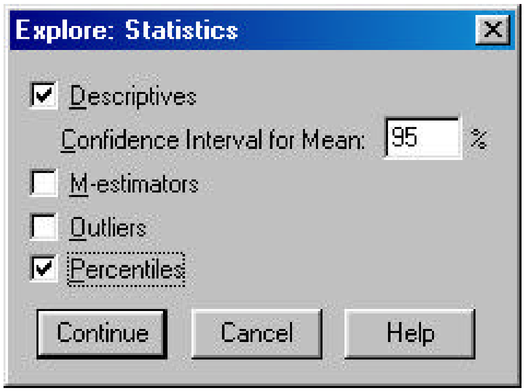
Analyze → Descriptive Statistics → Explore



x1
x2
x3
x4
x5

Display: ● Both ○ Statistics ○ Plots

- Move the variables you want to analyze into the Dependent List window
- Statistics -- allows you to select from various summary statistics
- Plots -- lets you select various graphical depicitons of the data
- Options allows you to determine how missing values are treated
- Display lets you pick what gets shown
- Here's the ones I like …



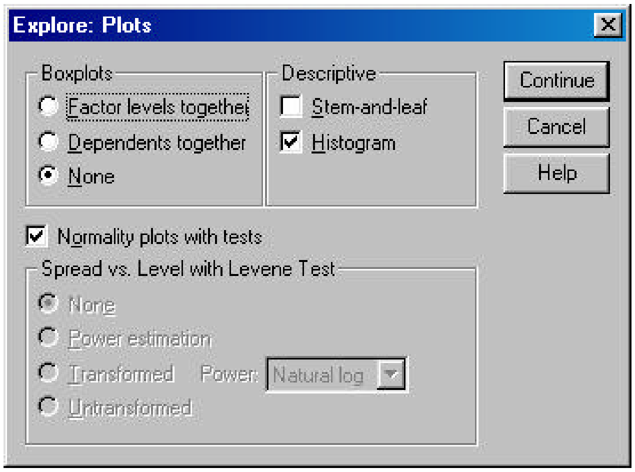
☑ Descriptives

Confidence Interval for Mean: 95 %

□ M-estimators

□ Outliers

☑ Percentiles



Boxplots: ○ Factor levels together ○ Dependents together ● None

Descriptive: □ Stem-and-leaf ☑ Histogram

☑ Normality plots with tests

Spread vs. Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

We'll use these to take a look at the variables in this data set.

The first part of the output tells you the sample size for each variable.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
X1	95	100.0%	0	.0%	95	100.0%
X2	95	100.0%	0	.0%	95	100.0%
X3	95	100.0%	0	.0%	95	100.0%

Descriptives

			Statistic	Std. Error
X1	Mean		49.2316	1.08666
	95% Confidence Interval for Mean	Lower Bound	47.0740	
		Upper Bound	51.3892	
	5% Trimmed Mean		49.1082	
	Median		50.0000	
	Variance		112.180	
	Std. Deviation		10.59150	
	Minimum		26.00	
	Maximum		77.00	
	Range		51.00	
	Interquartile Range		13.0000	
	Skewness		.074	.247
	Kurtosis		-.126	.490
X2	Mean		59.0947	2.01806
	95% Confidence Interval for Mean	Lower Bound	55.0878	
		Upper Bound	63.1016	
	5% Trimmed Mean		57.9678	
	Median		54.0000	
	Variance		386.895	
	Std. Deviation		19.66965	
	Minimum		26.00	
	Maximum		112.00	
	Range		86.00	
	Interquartile Range		25.0000	
	Skewness		.989	.247
	Kurtosis		.206	.490
X3	Mean		74.6737	2.58900
	95% Confidence Interval for Mean	Lower Bound	69.5332	
		Upper Bound	79.8142	
	5% Trimmed Mean		76.4415	
	Median		84.0000	
	Variance		636.775	
	Std. Deviation		25.23441	
	Minimum		3.00	
	Maximum		106.00	
	Range		103.00	
	Interquartile Range		34.0000	
	Skewness		-1.113	.247
	Kurtosis		.289	.490

The mean and std can be used to test whether or not the sample differs from a specific population mean t = (mean - pop mean) / Std. df = n - 1

The 95% CI is given, which can be used to test the same question. We are 95% sure the true population mean for X1 is somewhere between 69.5 and 79.8, so if we're looking to represent a population with a mean in that interval, this could be a good sample.

5% trimmed mean is found by tossing the top 5% of the scores and the bottom 5% of the scores and taking the mean of the middle 90% of the sample.

Notice the disparity of the mean, 5% trimmed mean and median. When they line up and differ by this much, you can expect a skewed distribution -- the mean will be toward the tail of the skewed distribution.

The mean and standard deviation will be more or less descriptive depending upon how "well-behaved" the distribution is. The greater the skewing the less well the distribution meets the assumptions of the mean and std computational formulas.

The interquartile range tells the boundaries of the middle 50% of the distribution.

The skewness is important to examine. "0" means a symmetrical distribution. The skewness value tells the direction of the tail of the asymmetrical distribution.

Kurtosis tells the relative amount of distribution that is in the middle vs. the tails of the distribution. + values means more of the distribution is in the middle (and so the middle is really high -- leptokurtic) - values means more of the distribution is in the tails (and so the middle is really flat -- platakurtic)

Standard errors are given for skewness and kurtosis. You can compute significance tests (e.g., t = skewness / std err skewness), but they haven't proved to be very useful.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	X1	31.6000	34.0000	42.0000	50.0000	55.0000	62.0000	67.6000
	X2	37.8000	39.0000	45.0000	54.0000	70.0000	94.0000	99.2000
	X3	23.0000	31.8000	59.0000	84.0000	93.0000	99.8000	103.0000
Tukey's Hinges	X1			42.0000	50.0000	55.0000		
	X2			45.0000	54.0000	69.5000		
	X3			62.0000	84.0000	92.5000		

The percentile information includes the 25th and 75th percentile Tukey's hinges, which are the basis for the most commonly used nonparametric approach to outlier analysis. These hinges have been shown to yield better outlier definitions than the standard percentiles.

Tests of Normality

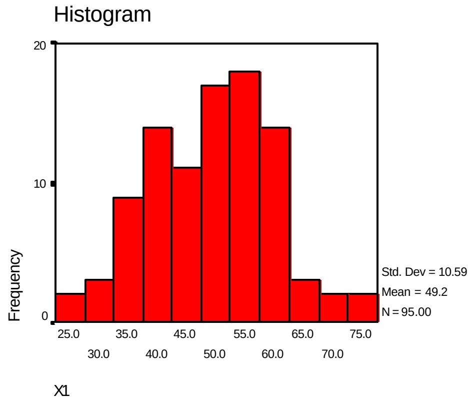
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X1	.199	95	.000	.867	95	.000
X2	.061	95	.200*	.991	95	.753
X3	.195	95	.000	.902	95	.000
X4	.260	95	.000	.789	95	.000
X5	.277	95	.000	.683	95	.000

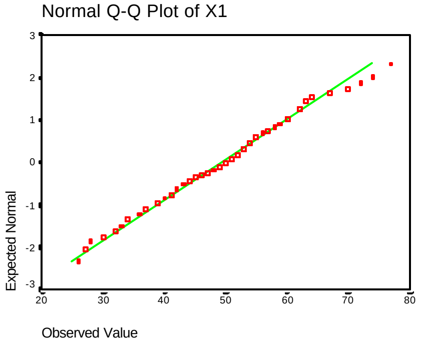
*. This is a lower bound of the true significance.

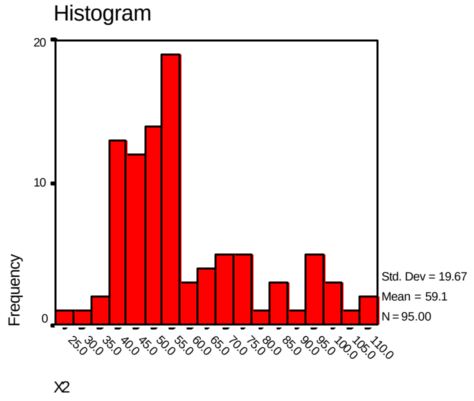
a. Lilliefors Significance Correction

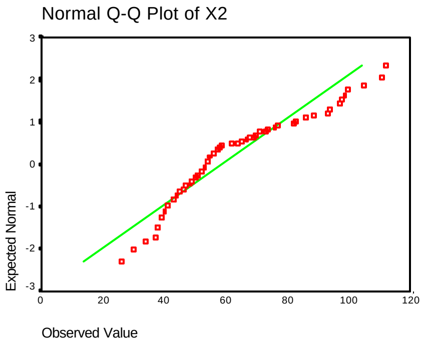
These are tests of whether the sample fits a normal distribution. They generally very sensitive -- meaning that the H0: (normal distribution) is often rejected.

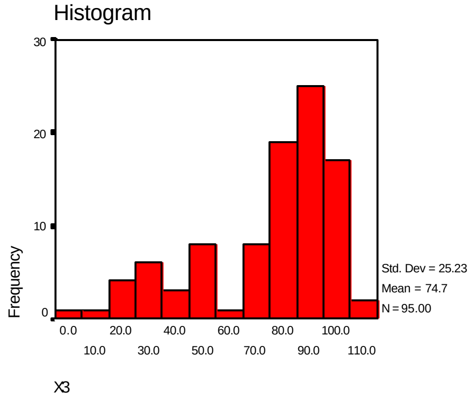
Other approaches to considering how "normal" the sample distribution are histograns and Q-Q plots. There are lots of other ways of looking at the data. SPSS also offers histograms with normal distribution overlays (in Freqencies or Charts), Boxplots, Stem-and-Leaf plots and others (e.g., de-trended Q-Q plots). With play you will develop preferences -- but rarely do different approaches reveal different things. Both of these, along with mean-median comparison and the skewness value tell the same storv…

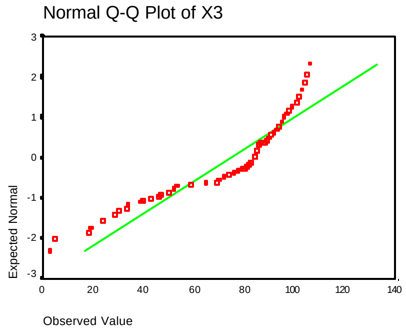












What does it mean to "clean" data?

Let's consider skewness. When a variable shows "appreciable" skewness we may want to un-skew it, for several reasons:

1. Skewing will distort the mean (draw it toward the tail) and inflate the std. Thus, descriptions of the sample and inferences to the target population are less accurate.
2. Skewing can distort bivariate statistics
   - For t-tests and ANOVA -- means comparisons
     - Means may not be "representative" measures of central tendency for skewed distributions, so comparisons of them lose meaningfulness. Depending upon the relative direction of skewing of the groups being compared, this can lead to an increased likelihood of Type I, II or III error.
     - Sum of squares, variance, std, sem and related error terms are all inflated, decreasing the power of significance tests and increasing Type II error (misses)
   - For correlations
     - Skewing, especially if variables are skewed in opposite directions, can lead to an underestimate of the population correlation and Type II errors of r or b (simple regression weight).
3. Multivariate analyses that are based in either of these analyses share related problems

Let's also consider outliers. When a variable has outliers we may want to deal with them, for several reasons:

1. Outliers will distort the mean (draw it toward the outlier) and inflate the std. Thus, descriptions of the sample and inferences to the target population are less accurate.
2. Outliers can distort bivariate statistics
   - For t-tests and ANOVA -- means comparisons
     - Means may not be "representative" measures of central tendency, so comparisons of them lose meaningfulness. Depending upon the relative values of the outliers of the groups being compared, this can lead to an increased likelihood of Type I, II or III error.
     - Sum of squares, variance, std, sem and related error terms are all inflated, decreasing the power of significance tests and increasing Type II error (misses)
   - For correlations
     - Outliers can change the r, and/or a values.
3. Multivariate analyses that are based in either of these analyses share related problems

Notice anything ???

Skewness and outliers have very similar detrimental effects upon univariate, bivariate and multivariate analyses -- producing results that don't describe that's going on in the population.

What we'll do about it

How to "un-skew" a distribution -- depends upon how skewed it is…

- The most common transformations and when to apply them -- be sure all X values are positive for all transformations
  - **Square root** **ÖX** skewness .8 - 1.5
  - **Base-10 Log** **log$_{10}$X** skewness 1.0 - 3.0
  - **Inverse** **1/X** skewness 2.0 - 4.0
  - The variability in "advice" about when to apply which is caused in part because the skewness isn't that great of an index of which transformation will work best -- the actual shape of the distribution for a given skewness can have a lot to do with what is the best transformation
- Please note, these transformations are for positive (+) skewing only
  - Negative (-) skewing is handled by first "transposing" the distribution ((max+1) - X) and then apply transformation

What to do about "outliers"

- Identify them…
  - For decades (and still sometimes today) common advice was to toss any score that had a corresponding Z-value of +/- 2-3, depending upon whom you believed
  - One difficulty with this is that outliers increase the std, and so "outliers can hide themselves".
  - In the late 70's, the use of nonparametric statistics for identifying outliers was championed. The reasoning was that nonparametric measures of variability (variations on the interquartile range) would be disrupted less by outliers than is the std.
- What to do with them -- opinions vary
  - "Trim" them -- delete them from the dataset
  - "Windsorize" them -- replace the "too extreme" value with the "most extreme acceptable value"
  - this has the advantage of not tossing any data -- large/small values are still large/small, but less likely to disrupt the mean and std estimates and statistics that depend upon them

Please note:

When deciding whether to transform and/or perform outlier analyses, it is useful to keep in mind that the skewness statistic, by itself, won't tell which of these is best-suited to improve your sample distribution. The reason for this is that both distributional skewing and asymmetrical outliers can produce high skewness values. So, it is important to take a look at the actual distribution before planning what to do.

Working with X1 - X5

Let's start with X1 -- Statistics and graphs are shown above. Not much to do with this one. Very low skewness value, Normal distribution tests are null, the histogram and the Q-Q plot look good. Before decide we can safely use the mean and std of this distribution we should check for outliers.

To do this we compute the upper and lower bound of "non-extreme" values and compare those with the maximum and minimum values in the distribution. We will use the 75% and 25% Tukey Hinges for these calculations.

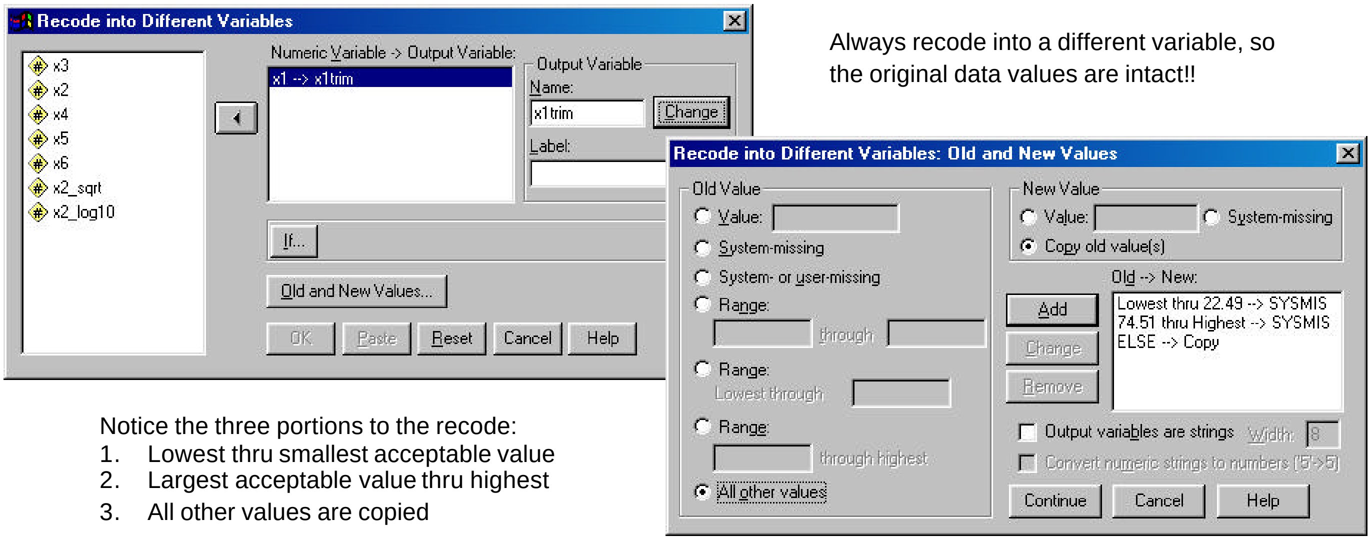
Upper bound = 75% + 1.5 * (75% - 25%) = 55 + 1.5 * (55 - 42) = 55 + 19.5 = 74.5

Lower bound = 25% - 1.5 * (75% - 25%) = 42 - 1.5 * (55 - 42) = 42 - 19.5 = 22.5

With a maximum of 77 and a minimum of 26, we can see that there are "too large" but not "too small" outliers. As mentioned, we can "trim" or "Windsorize" the outliers. Here's how to do each.

Trimming = turning "outliers" into "missing values".

Transform → Recode → Into Different Variables



Recode into Different Variables

Numeric Variable -> Output Variable: x1 --> x1trim

Output Variable Name: x1trim

Recode into Different Variables: Old and New Values

New Value: ● Copy old value(s)

Old --> New:
Lowest thru 22.49 --> SYSMIS
74.51 thru Highest --> SYSMIS
ELSE --> Copy

Old Value: ● All other values

Always recode into a different variable, so the original data values are intact!!

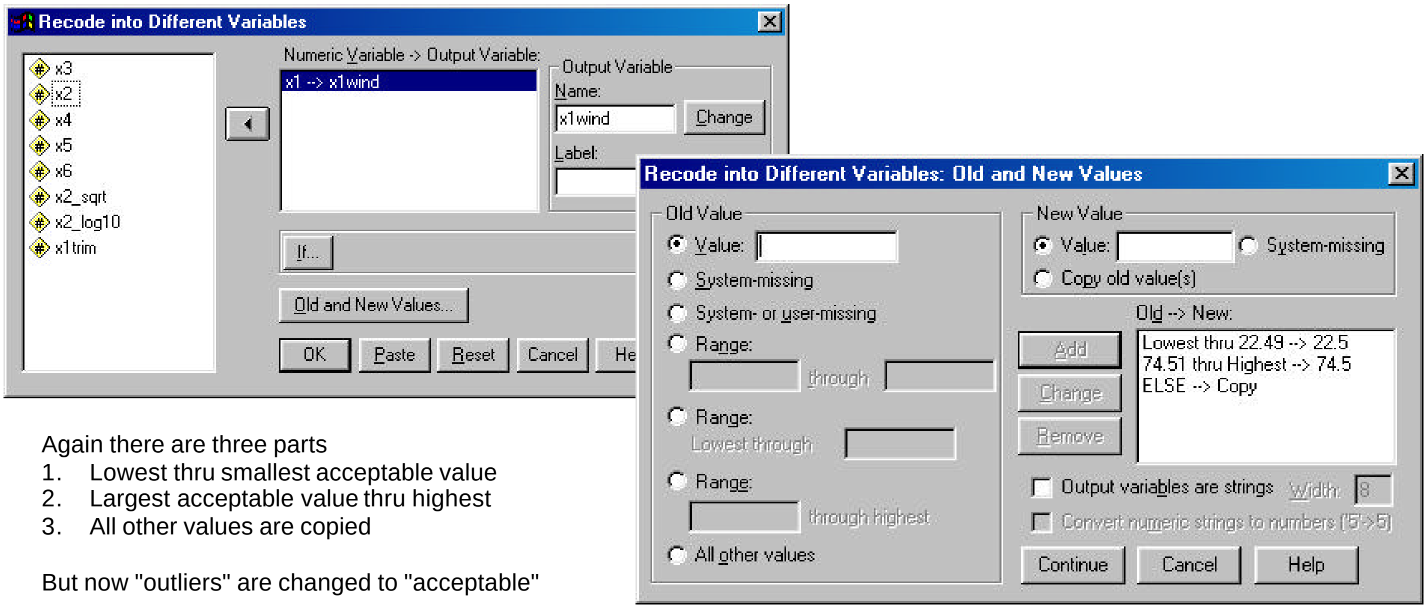
Notice the three portions to the recode:
1. Lowest thru smallest acceptable value
2. Largest acceptable value thru highest
3. All other values are copied

Here are the "final" stats on X1TRIM -- notice that N = 94

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X1TRIM	94	26.00	74.00	48.9362	10.24727	-.047	.249
Valid N (listwise)	94						

Windsorizing = turning "outliers" into "most extreme acceptable scores"



Recode into Different Variables

Numeric Variable -> Output Variable: x1 --> x1wind

Output Variable Name: x1wind

Recode into Different Variables: Old and New Values

Old --> New:

Lowest thru 22.49 --> 22.5
74.51 thru Highest --> 74.5
ELSE --> Copy

Again there are three parts
1. Lowest thru smallest acceptable value
2. Largest acceptable value thru highest
3. All other values are copied

But now "outliers" are changed to "acceptable" values rather than made "missing".

Here are the final stats on X1WIND -- notice that N = 95

Descriptive Statistics

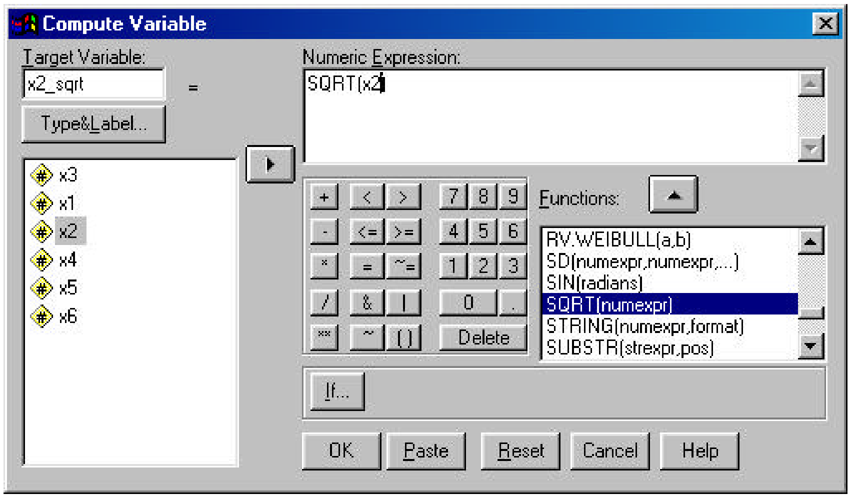
	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X1WIND	95	26.00	74.50	49.2053	10.52467	.034	.247
Valid N (listwise)	95						

As was found here. Trimming and Windsorizing lead to very similar results with the data are "well-behaved".

On to X2 -- Statistics and graphs are shown above. There is a definite positive skew to this variable. The shape seen in the histogram and Q-Q plot seem to correspond the skewness value -- no discontinuous sub-distributions, etc.

The question is, how skewed does a distribution need to be to warrant transformation? The usual skewness cutoffs vary from .7-.9, so this is "skewed enough" to transform.

Transformation → Compute



Compute Variable

Target Variable: x2_sqrt = Numeric Expression: SQRT(x2

The SQRT() and LG10() are the two nonlinear transformation functions provided by SPSS that you will use most often for "symmetrizing" you data.

Another Target Variable x2_log10

used the Numeric Expression LG10(x2)

The results of these transformations highlight a common dilemma.

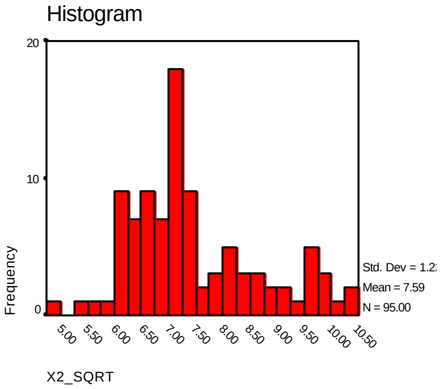
X2_sqrt has a skewness of .699

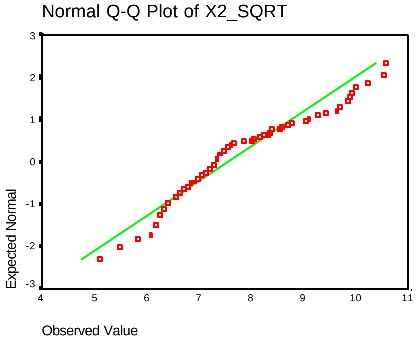
Which transformation to use??

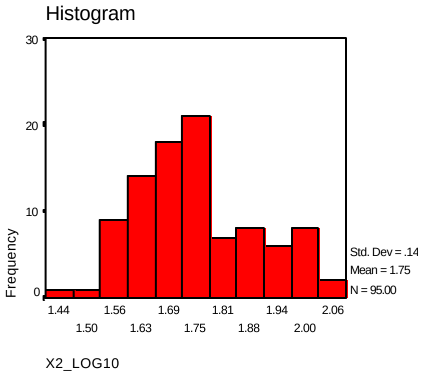
The usual advice is to use the less extreme transformation that gets skewness into the acceptable range.

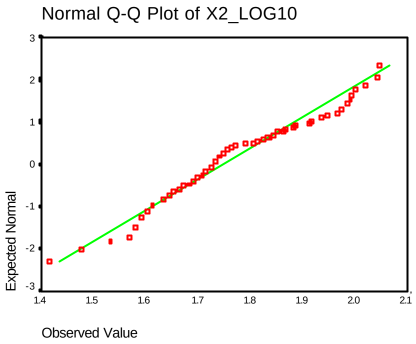
There's another reason to limit the use of log transforms. The sum or two log transformed variables is the same as the product of the original variables. So, the weighted sum or two log transformed variables is a weighted interaction between them, not usually what we intend.

X2_log10 has a skewness of .368









Checking X2_SQRT for outliers

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X2_SQRT	95	5.10	10.58	7.5904	1.22342	.699	.247
Valid N (listwise)	95						

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1	X2_SQRT	6.1481	6.2450	6.7082	7.3485	8.3666	9.6954	9.9599
Tukey's Hinges	X2_SQRT			6.7082	7.3485	8.3366		

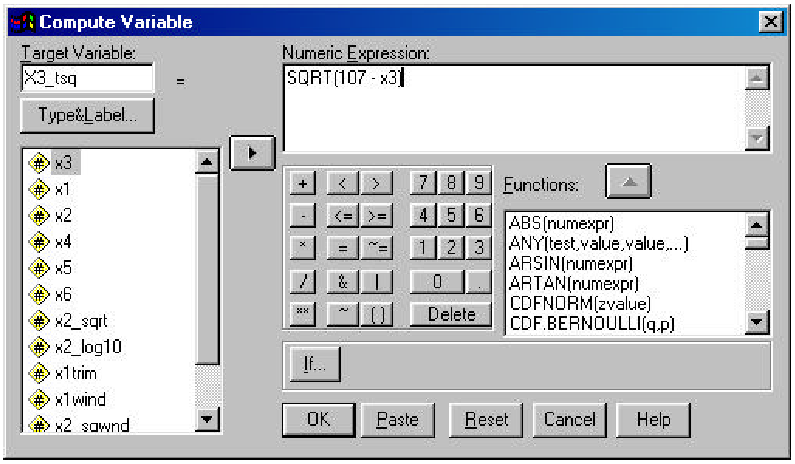
Upper bound = 75% + 1.5 * (75% - 25%) = 8.3366 + 1.5 * (8.3366 - 6.7082) = 8.3366 + 2.4426 = 10.7793
Lower bound = 25% - 1.5 * (75% - 25%) = 6.7082 - 1.5 * (8.3366 - 6.7082) = 6.7082 - 2.4426 = 4.2656

After transformation we have no outliers, since the min and max are "inside" of the outlier bounds.

Please note: Unlike X1, X2 is no longer in its original measured scale, but is the sqrt of that scale. Therefore, means and stds will be harder to think about.

Next is X3 -- Statistics and graphs are shown above. There is a definite negative skew to this variable. Again, the shape seen in the histogram and Q-Q plot seem to correspond the skewness value -- no discontinuous sub-distributions, etc.

Because the skewing is negative, we must transpose before transforming, so the compute window would include…



Target Variable: X3_tsq

Numeric Expression: SQRT(107 - x3)

Since 106 was the maximum for X3, subtracting from 107 guarantees that all the values are positive (can't take SQRT of 0 or

The skewness of the transformed variable is within acceptable bounds, so we wouldn't try more extreme transformations. Plus, the skew is now positive, like X2. The problems with skewness are lessened if all the variables are similarly skewed (say all +).

We would then continue with an outlier check.

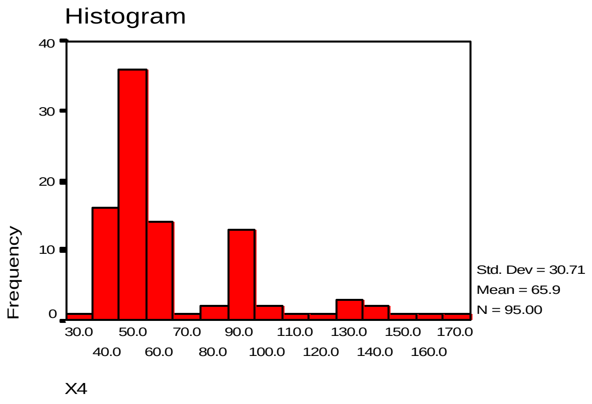
Descriptive Statistics

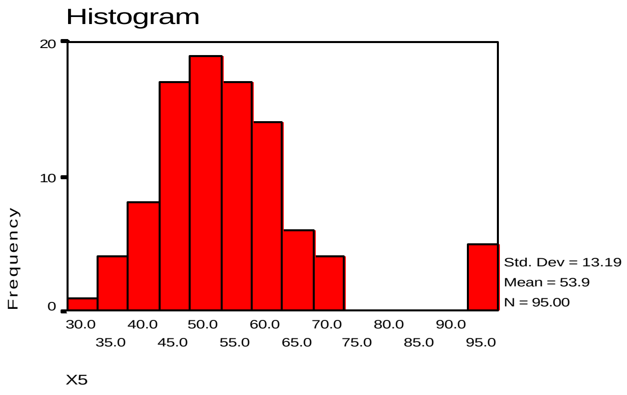
	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X3_TSQ	95	1.00	10.20	5.2701	2.14484	.464	.247
Valid N (listwise)	95						

X4 and X5 show another common "situation"

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X4	95	30.00	173.00	65.9053	30.71457	1.631	.247
X5	95	31.00	97.00	53.9263	13.19271	1.558	.247
Valid N (listwise)	95						





Both are positively skewed, and both probably have several "too large" outliers. But they have different distribution shapes.

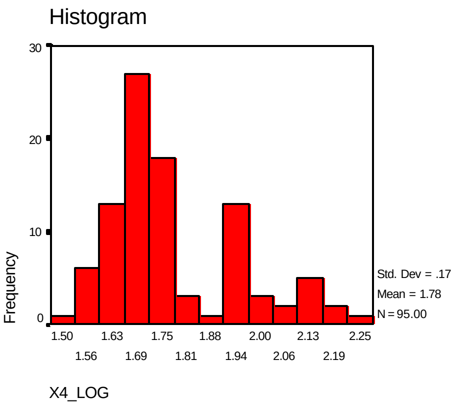
X4, being continuous, will probably clean up better by transforming, then re-checking for outliers. Whereas X5, should have its outliers cleaned up, and then be re-checked for skewness.

Let's look at the results from square root and log transforms of both (even though we're pretty sure that isn't the best thing to do with X5).

Descriptive Statistics

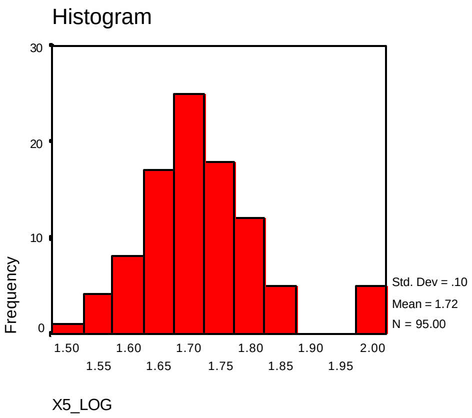
	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X4_SQRT	95	5.48	13.15	7.9408	1.69699	1.078	.247
X5_SQRT	95	5.57	9.85	7.2949	.84762	1.098	.247
X4_LOG	95	1.48	2.24	1.7824	.16964	.639	.247
X5_LOG	95	1.49	1.99	1.7206	.09692	.635	.247
Valid N (listwise)	95						

The sqrt transforms "aren't enough", but the log transforms bring the skewness into an acceptable range. But check the histograms…



X4_LOG looks better, but the body of X5_LOG looks slightly negatively skewed, but off set by obvious outliers.

Let's go back and start over with X5, cleaning up the outliers first and then re-assessing the skewness.



Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	X5	36.0000	41.6000	45.0000	52.0000	59.0000	67.4000	95.0000
Tukey's Hinges	X5			45.5000	52.0000	59.0000		

Upper bound = 75% + 1.5 * (75% - 25%) = 59 + 1.5 * (59 - 45) = 59 + 21 = 80
Lower bound = 25% - 1.5 * (75% - 25%) = 45 - 1.5 * (59 - 45) = 45 - 21 = 23

We have to decide whether to trim or Windsorize. If we Windsorize the "outrigger" of values will still be noncontiguous with the rest of the distribution. When this happens, it is very important to check up on these values and try to determine if they are collection, coding or entry errors or if they comprise an important sub-population that should be studied separately. Since these are simulated data, none of these evaluations are possible, and I'm going to trim the "too large" outliers.

I used Recode into Different variables to make X5trim using the Old→ New lowest thru 80 → copy ELSE → SYSMIS

An outlier check on X4_LOG revealed no outliers, so the "cleaned" versions of these variables have the following stats. You can see that trimming X5 was much more effective than was transforming it.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X4_LOG	95	1.48	2.24	1.7824	.16964	.639	.247
X5TRIM	90	31.00	72.00	51.5889	8.87731	.110	.254
Valid N (listwise)	90						