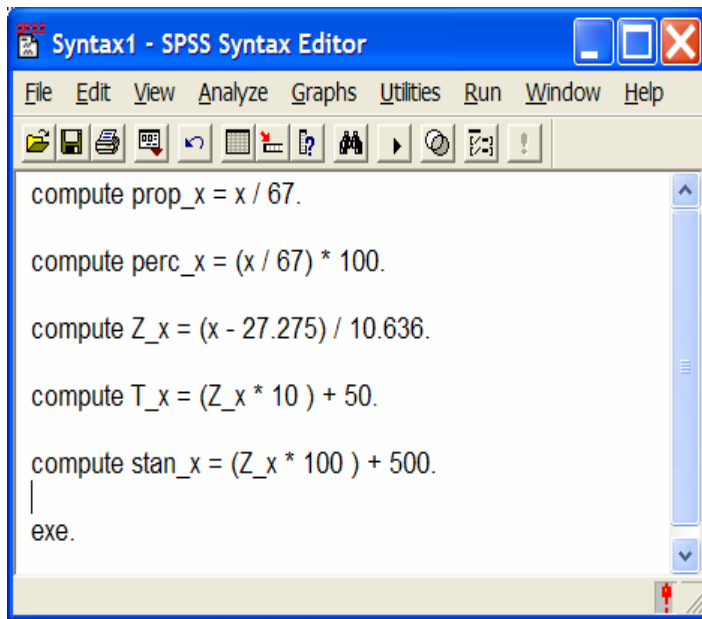


Transformations & Data Screening

Linear Transformations

Linear transformations are usually performed to change the mean and standard deviation of the data distribution to something that is easier to work with. Usually the transformations create new variables that are added to the right-hand side of the spread sheet in the Data Editor. Remember, linear transformations do not change the shape of the distribution, nor do they change the results of subsequent statistical tests.

SPSS provides two ways to perform transformations (more than that for Z-scores). The “old-fashioned” way is to write compute statements into a syntax window. Below are the statements to perform some common transformations.



File → New → Syntax

Some transformations require values that must be obtained from the data set.

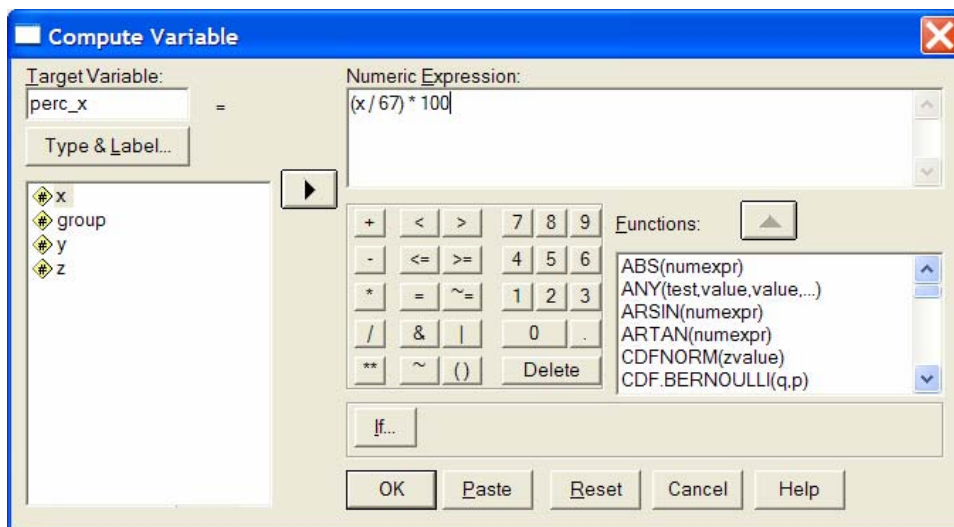
Computing the proportion and the percentage both require knowing the denominator value (in this case the largest value of that variable in the dataset)

Computing Z requires knowing the mean and standard deviation – either from the data set (as in this case) or population values

Computing T and Standard scores is easiest if you start with the Z-scores

Including the execute command causes the variables to be computed with the statements are run. Else, the phrase “Transformations Pending” appears in the bottom right of the Data Editor and the statements will not be executed until the next statistical analysis is requested.

SPSS also provides a point-&-click procedure using the Compute Variable window, that is available at **Transform → Compute** Here is what the percentage transformation would look like using this window.



The Numeric Expression will use the same formula as is used in the Syntax Window.

There many different functions available in the Compute window that can simplify more difficult transformations.

It is also possible to compute Z-scores when using the Descriptives procedure. At the bottom of the Descriptives window is a check box to “Save standardized values as variables. New variables will be made containing the Z-scores of all the variables included in the analysis. They will be named starting with a “z”, e.g., the Z-score for age will be stage.

Non-Linear Transformations

Non-linear transformations are usually performed to change the shape of the data distribution to something that is more symmetrical. Formulas for non-linear transformations can be applied to either a Syntax Window or the Compute window. Remember, non-linear transformations do change the shape of the distribution, and they do change the results of subsequent statistical tests.

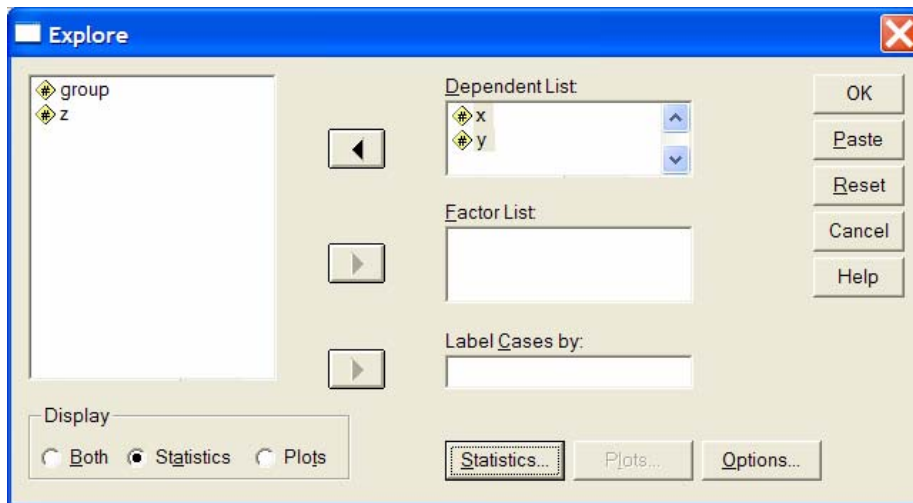
Two things to remember when performing non-linear transformations:

1. Log and inverse transforms require positive values and square-root transforms require non-negative values.
2. The transformations reduce positive skewing. To reduce negative skewing the data must first be reflected (each value subtracted from the maximum value + 1)

Data Screening

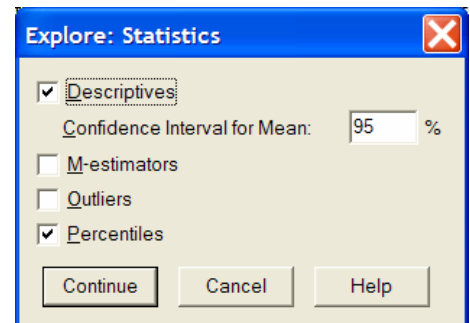
The Explore procedure makes data screening very simple. Just specify the variables and the statistics to compute.

Analyze → Descriptive Statistics → Explore



Move the variables into the Dependent List window, be sure "Statistics" is checked, and press the Statistics button.

Be sure "Percentiles" is checked in the Statistics window.



Descriptives

			Statistic	Std. Error
X	Mean		27.2750	1.68172
	95% Confidence Interval for Mean	Lower Bound	23.8734	
		Upper Bound	30.6766	
	5% Trimmed Mean		26.1389	
	Median		26.5000	
	Variance		113.128	
	Std. Deviation		10.63614	
	Minimum		8.00	
	Maximum		67.00	
	Range		59.00	
	Interquartile Range		7.5000	
	Skewness		2.318	.374
	Kurtosis		7.929	.733

Stats to pay attention to when data cleaning include:

1. Mean and std – the major purpose of data cleaning is to ensure that these are reasonable estimates of population values
2. minimum & maximum – helpful when deciding whether or not there are any outliers
3. Skewness – we usually want our variables to have skewness < +/- .80

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	X	10.4000	20.1000	22.2500	26.5000	29.7500	36.5000	63.6000
Tukey's Hinges	X			22.5000	26.5000	29.5000		

Tukey's Hinges are used as the basis for outlier analysis.

Example of Univariate Outlier Analysis

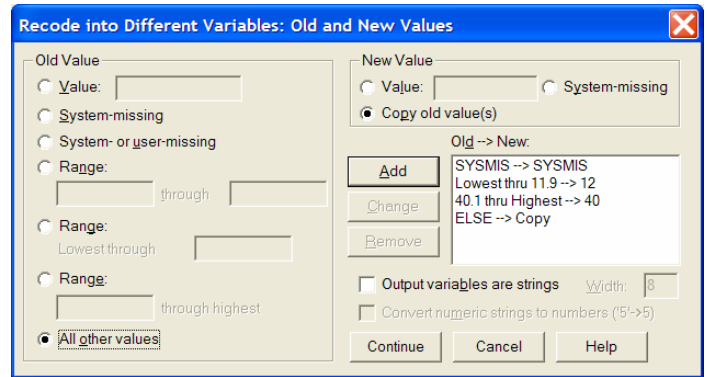
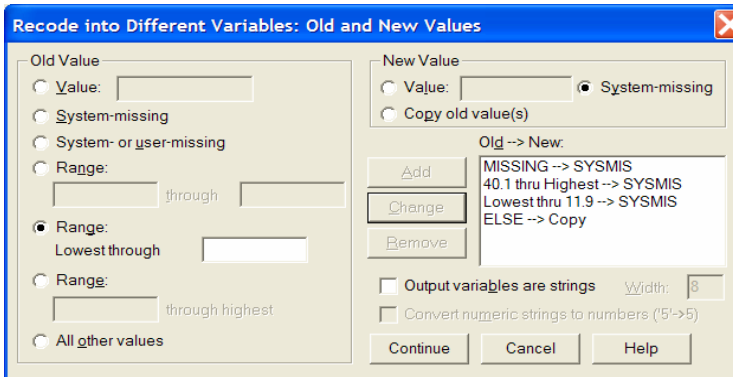
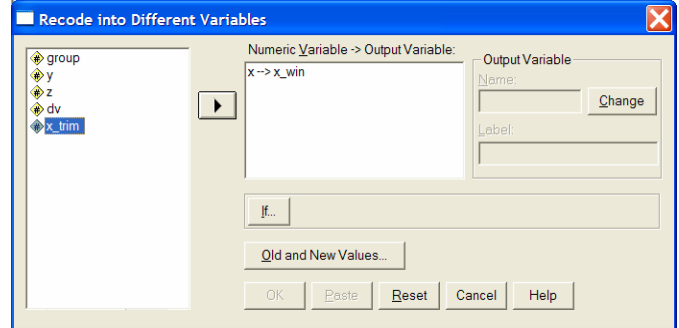
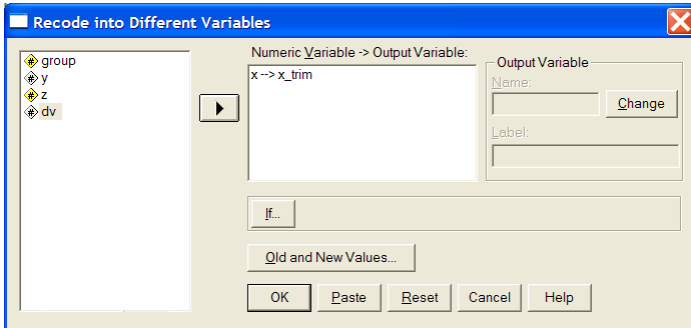
Working with the information about variable “x” from above... With Tukey’s Hinges of 22.5 and 29.5, the hinge spread is 7, so the bounds are:

$$\text{Lower Bound} = 22.5 - (1.5 * 7) = 22.5 - 10.5 = 12 \quad \text{any value smaller than 12 is “too small”}$$

$$\text{Upper Bound} = 29.5 + (1.5 * 7) = 29.5 + 10.5 = 40 \quad \text{any value larger than 40 is “too large”}$$

If we look at the minimum (8) we can see that we have at least one “too small” value in the data set. Similarly, if we look at the maximum (67) we can see that we have at least one “too large” value.

Transform → Recode → Recode into Different Variables



The following syntax will produce the same Winzorizing as above.

If (x lt 12) x = 12.
If (x gt 40) x = 40.

Here are the results of these transformations.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewnes
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
X	40	8.00	67.00	27.2750	10.63614	2.318
X_TRIM	36	18.00	37.00	26.1389	4.45391	.585
X_WIN	40	12.00	40.00	26.1250	6.15687	.167
Valid N (listwise)	36					

In this case the mean of the different versions of x are not very different, but the transformations led to much smaller (and probably more accurate) estimates of the std.

Example of Univariate Distribution “Symmetrizing”

Initial inspection of “y” and “z” variables shows that they both have substantial skewing.

Descriptive Statistics

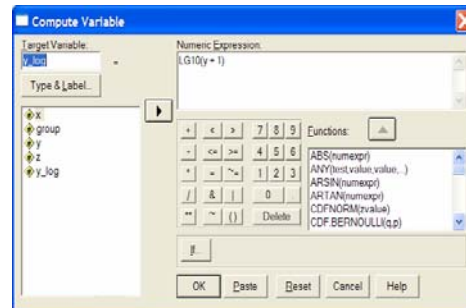
	N	Minimum	Maximum	Mean	Std.	Skewnes
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Y	40	.00	467.00	94.0500	87.87169	2.282
Z	40	8.00	19.00	15.1750	1.79583	-1.367

The skewness of “y” suggests a log or even an inverse transformation may be needed. Both of these transformations require positive values. Since the minimum y value is 0, we will need to modify this variable before we can perform the nonlinear transformations. All we need to do is add “1” to each y score before transforming it.

We could use the Syntax Window to obtain all 3 transformations – it is a good idea to try all 3, so we can use the least severe transformation that gives us an acceptable skewness. On the right below is the Compute Window to perform the log transformation. Available from **Transformation → Compute**

```

compute y_sqrt = sqrt(y).
compute y_log = lg10(y + 1).
compute y_inv = 1 / (y + 1).
exe.
    
```



Descriptive Statistics

	Mean	Std.	Skewnes
	Statistic	Statistic	Statistic
Y	94.0500	87.87169	2.282
Y_SQRT	8.6832	4.37384	.330
Y_LOG	1.7472	.59674	-1.802
Y_INV	.0908	.26252	-3.340

The square root transform seems to do the best (a reminder that the guidelines for when to use what transformation are just that – guidelines, not rules!).

If we hadn’t looked at the square root transform, we might have decided that the log-transformed skewness of -1.8 is somewhat better than the original skewness of 2.28 !!

It is best to check all three transforms each time.

When transforming “z” we have to remember that the transformations we know are designed to symmetrize positively skewed variables. Since z is negatively skewed, we’ll have to reflect its distribution before transforming it. To do that we need to know the maximum data value of z, which was 19. Again, we’ll look at the result of all three transformations.

```

compute z_sqrt = sqrt(20 - z).
compute z_log = lg10(20 - z).
compute z_inv = 1 / (20 - z).
exe.
    
```

Descriptive Statistics

	Mean	Std.	Skewnes
	Statistic	Statistic	Statistic
Z	15.1750	1.79583	-1.367
Z_SQRT	2.1583	.41346	-.191
Z_LOG	1.1681	.05060	.422
Z_INV	.2532	.18127	-3.714

The square root transform again seems to do the best, because a skew of -.19 is less than a skew of .42

Example of Data Cleaning for ANOVA

An initial ANOVA revealed...

DV	N	Mean	Std. Deviation
1.00	19	25.4211	7.39606
2.00	21	25.7619	5.69126
Total	40	25.6000	6.47203

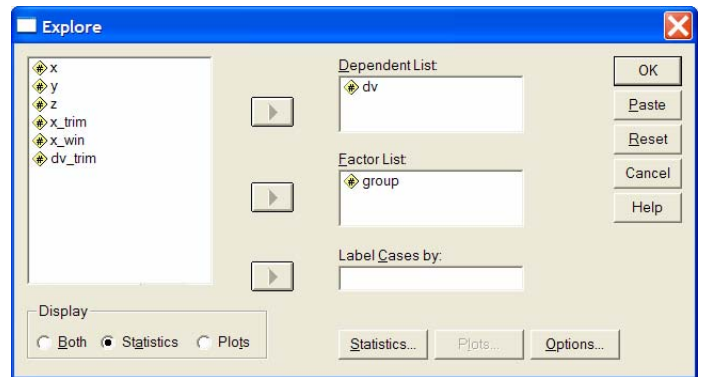
Small, non-significant mean difference.

Remember → you should do the data cleaning before the NHST analyses!

DV	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.159	1	1.159	.027	.870
Within Groups	1632.441	38	42.959		
Total	1633.600	39			

When performing a data cleaning associated with an ANOVA, the outliers must be identified and eliminated separately for each group.

SPSS Explore allows us to select a “grouping variable,” which will do the same thing as using a “select cases” → separate Explore analyses will be completed for each group.



The key output is the Tukey's Hinges values. Based on these, the outlier boundaries for each group was calculated.

		Percentiles			
	GROUP	25	50	75	
Tukey's Hinges	DV	1.00	22.5000	27.0000	29.0000
		2.00	22.0000	24.0000	27.0000

For Group = 1 the hinge spread = $29 - 22.5 = 6.5$

Lower Bound = $22.5 - (1.5 * 6.5) = 22.5 - 9.75 = 12.75$ any Group 1 value smaller than 12.75 is “too small”

Upper Bound = $29 + (1.5 * 6.5) = 29 + 9.75 = 38.75$ any Group 1 value larger than 38.75 is “too large”

For Group = 2 the hinge spread = $27 - 22 = 5$

Lower Bound = $22 - (1.5 * 5) = 22 - 7.5 = 14.5$ any Group 2 value smaller than 14.5 is “too small”

Upper bound = $27 + (1.5 * 5) = 27 + 7.5 = 34.5$ any Group 2 value larger than 34.5 is “too large”

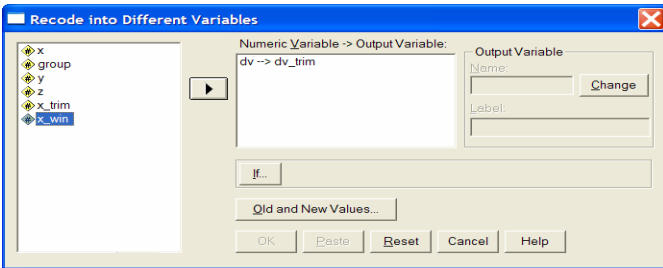
Remember, the boundaries are “the most extreme acceptable value” – values that large and small **are** acceptable values and should be included in the analysis. You have to be careful when applying these values. In this case, all the data values are integers, so we can use the exact boundary values and not eliminate “good” cases. However, when the data are decimals, you have to be sure to only trim or Winsorize values **more extreme** than the boundary values.

Remember the “logic” of each transformation!

- Trimming eliminates outliers as “too extreme” to probably have come from the target population
- Winsorizing keeps outliers as indicating there are population members with large & small values, but lessens the impact of them to “likely” large & small values.

Here’s how to apply the boundary values to trim the outliers.

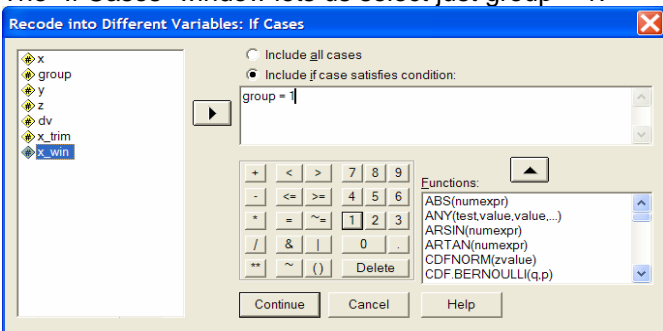
Set the name of the new variable



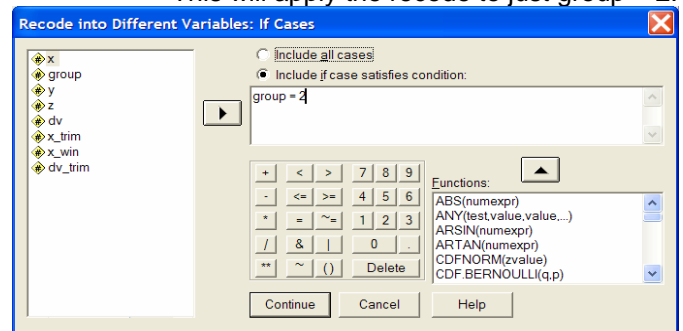
Please note:

- **These windows show you to truncate and Winsorize – they are just examples!!!**
- The windows below on the left were used to truncate outliers for group 1
- The windows directly below were used to Winsorize the outliers from group 2
- **You would never do this – always apply the same outlier procedure to all groups of an ANOVA**

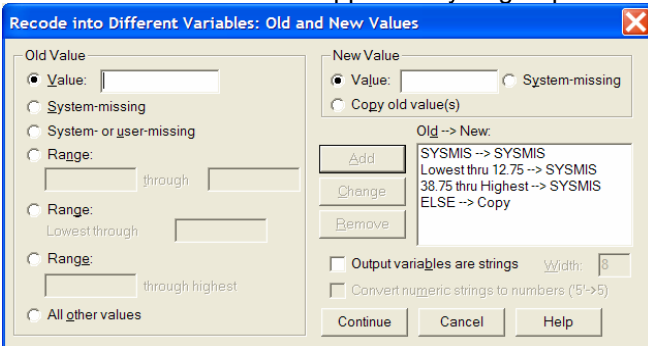
The “If Cases” window lets us select just group = 1.



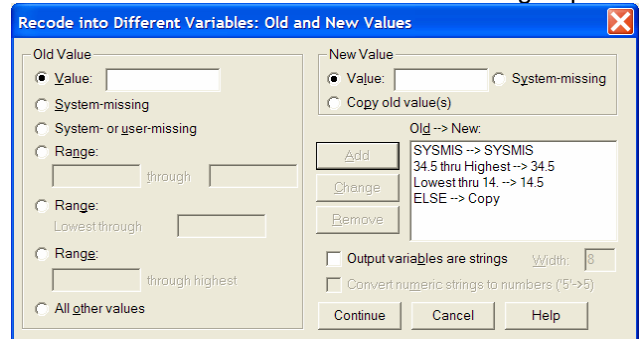
This will apply the recode to just group = 2.



This transformation will be applied only to group = 1



This recode uses the boundaries for group = 2.



Winsorizing the data would use the same information, but outlying values would be recoded into the nearest boundary.

Here are the results from the trimmed and Winsorized data.

Trimmed ANOVA results	Descriptives				ANOVA					
	DV_TRIM				DV_TRIM					
		N	Mean	Std. Deviation		Sum of Squares	df	Mean Square	F	Sig.
	1.00	17	27.3529	4.83401	Between Groups	79.919	1	79.919	4.297	.046
	2.00	19	24.3684	3.78903	Within Groups	632.303	34	18.597		
	Total	36	25.7778	4.51101	Total	712.222	35			

Winsorized ANOVA results	Descriptives				ANOVA					
	DV_WIN				DV_WIN					
		N	Mean	Std. Deviation		Sum of Squares	df	Mean Square	F	Sig.
Notice the difference between the results!	1.00	19	25.8158	6.47851	Between Groups	2.322	1	2.322	.074	.788
	2.00	21	25.3333	4.71257	Within Groups	1199.647	38	31.570		
	Total	40	25.5625	5.55155	Total	1201.969	39			

Example of Data Cleaning for Correlation

As a quick demonstration of the effects of outliers and skewed distributions, here are the correlations among the raw x, y & z and among the “cleaned” versions of these variables.

Correlations

		X	Y	Z
X	Pearson Correlation	1	.514	-.475
	Sig. (2-tailed)	.	.001	.002
	N	40	40	40
Y	Pearson Correlation	.514	1	-.904
	Sig. (2-tailed)	.001	.	.000
	N	40	40	40
Z	Pearson Correlation	-.475	-.904	1
	Sig. (2-tailed)	.002	.000	.
	N	40	40	40

Correlations

		X_TRIM	X_WIN	Y_SQRT	Z_SQRT
X_TRIM	Pearson Correlation	1	1.000	.822	.573
	Sig. (2-tailed)	.	.	.000	.000
	N	36	36	36	36
X_WIN	Pearson Correlation	1.000	1	.861	.588
	Sig. (2-tailed)	.	.	.000	.000
	N	36	38	38	38
Y_SQRT	Pearson Correlation	.822	.861	1	.779
	Sig. (2-tailed)	.000	.000	.	.000
	N	36	38	40	40
Z_SQRT	Pearson Correlation	.573	.588	.779	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	36	38	40	40

Clearly the two sets of intercorrelations would be interpreted differently!