Data Screening and Analysis

Prof. Ajai Pratap Singh

Dept. of Applied Psychology

VBS Purvanchal University Jaunpur

Uttar Pradesh

Email: ajaisingh27@gmail.com

Data Screening

- The purpose of data screening is to:
- (a) check if data have been entered correctly, such as out-of-range values.
- (b) check for missing values, and deciding how to deal with the missing values.
- (c) check for outliers, and deciding how to deal with outliers.
- (d) check for normality, and deciding how to deal with non-normality.

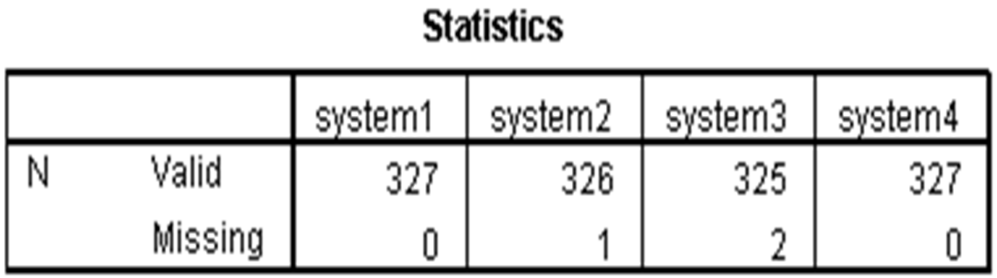
Incorrect Data

1. <u>**Finding incorrectly entered data**</u>
   - Your first step with “Data Screening” is using “Frequencies”

     1. Select **Analyze --> Descriptive Statistics --> Frequencies**

     2. Move all variables into the “Variable(s)” window.

     3. Click OK.
   - Output below is for only the four “system” variables in our dataset because copy/pasting the output for all variables in our dataset would take up too much space in this document.
   - The “**Statistics**” box tells you the number of missing values for each variable. We will use this information later when we are discussing missing values.



Statistics

		system1	system2	system3	system4
N	Valid	327	326	325	327
	Missing	0	1	2	0

Missing Data

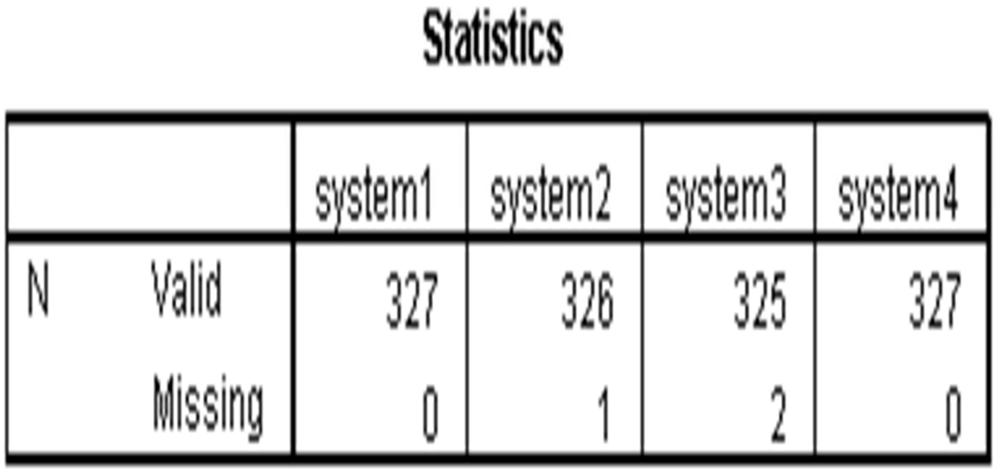
- Why is missing data a problem? Missing values means reduced sample size and loss of data. You conduct research to measure empirical reality so missing values thwart the purpose of research. Missing values may also indicate bias in the data. If the missing values are non-random, then the study is not accurately measuring the intended constructs. The results of your study may have been different if the missing data was not missing.
- How do I identify missing values?

  1. Select **Analyze --> Descriptive Statistics --> Frequencies**

  2. Move all variables into the "Variable(s)" window.

  3. Click OK.

- Output below is for only the four “system” variables in our dataset because copy/pasting the output for all variables in our dataset would take up too much space in this document.
- The “**Statistics**” box tells you the number of missing values for each variable.



Statistics

		system1	system2	system3	system4
N	Valid	327	326	325	327
	Missing	0	1	2	0

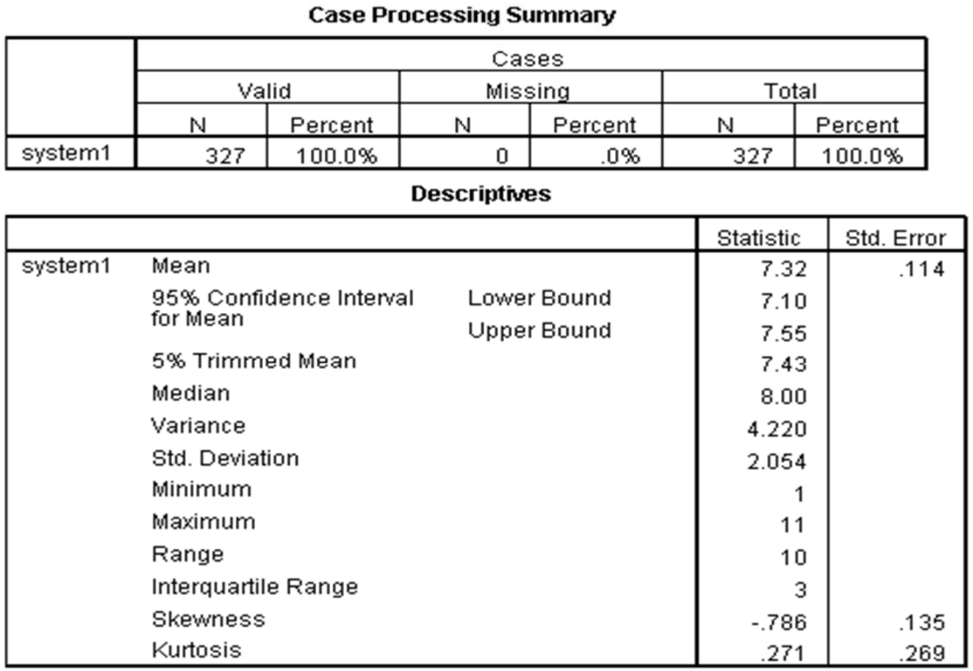
Outliers -Univariate

- **3**. **Outliers – Univariate**
- What are outliers? Outliers are extreme values as compared to the rest of the data. The determination of values as "outliers" is subjective. While there are a few benchmarks for determining whether a value is an "outlier", those benchmarks are arbitrarily chosen, similar to how "p≤.05" is also arbitrarily chosen.
- Should I check for outliers? Outliers can render your data non-normal. Since normality is one of the assumptions for many of the statistical tests you will conduct, finding and eliminating the influence of outliers may render your data normal, and thus render your data appropriate for analysis using those statistical tests.

- There are two categories of outliers – Univariate and multivariate outliers
  - Univariate outliers are extreme values on a single variable. For example, if you have 10 survey questions in your study, then you would conduct 10 separate univariate outlier analyses, one for each variable. Also, when you average the 10 questions together into a new composite variable, you can conduct univariate outlier analysis on the new variable. Another way you would conduct univariate analysis is by looking at individual variables within different groups.
  - For example, you would conduct univariate analysis on those same 10 survey questions within each gender (males and females), or within political groups (republican, democrat, other), etc. Or, if you are conducting an experiment with more than one condition, such as manipulating happiness and sadness in your study, then you would conduct univariate analysis on those same 10 survey questions within both groups.

- The second category of outliers is multivariate outliers. Multivariate outliers are extreme combinations of scores on two or more variables. For example, if you are looking at the relationship between height and weight, then there may be a joint value that is extreme compared to the rest of the data, such as someone with extremely low height but high weight, or high weight but low height, and so forth. You first look for univariate outliers, then proceed to look for multivariate outliers.
- Univariate outliers:
- 1.Select **Analyze --> Descriptive Statistics --> Explore**
- 2. Move all variables into the “Variable(s)” window.
- 3. Click “Statistics”, and click “Outliers”
- 4. Click “Plots”, and unclick “Stem-and-leaf”
- 5. Click OK.

- “**Descriptives**” box tells you descriptive statistics about the variable, including the value of Skewness and Kurtosis, with accompanying standard error for each. This information will be useful later when we talk about “normality”. The “5% Trimmed Mean” indicates the mean value after removing the top and bottom 5% of scores. By comparing this “5% Trimmed Mean” to the “mean”, you can identify if extreme scores (such as outliers that would be removed when trimming the top and bottom 5%) are having an influence on the variable.



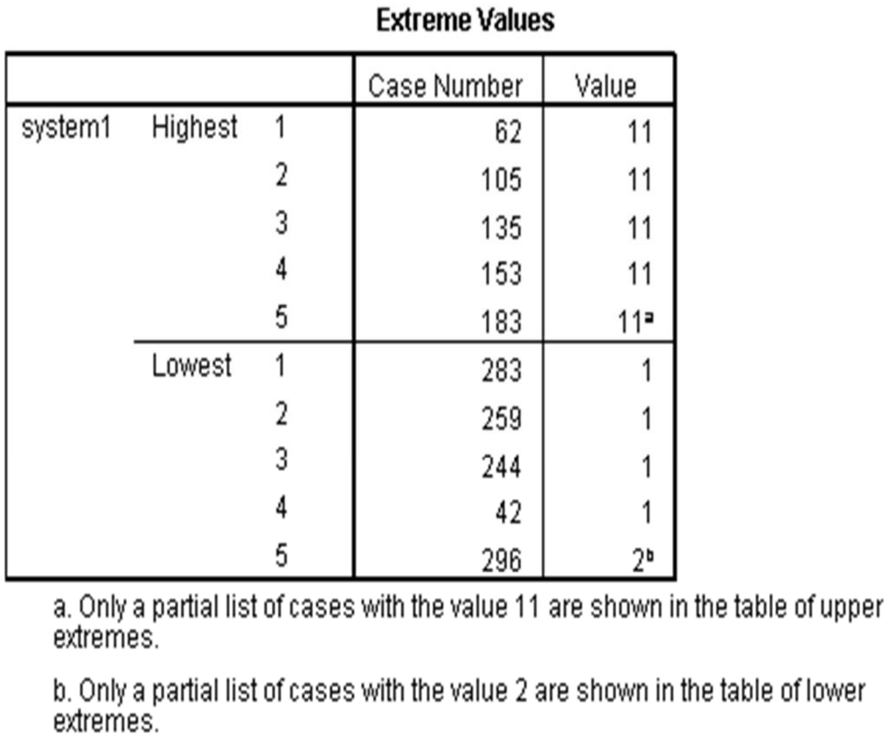
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
system1	327	100.0%	0	.0%	327	100.0%

Descriptives

			Statistic	Std. Error
system1	Mean		7.32	.114
	95% Confidence Interval for Mean	Lower Bound	7.10	
		Upper Bound	7.55	
	5% Trimmed Mean		7.43	
	Median		8.00	
	Variance		4.220	
	Std. Deviation		2.054	
	Minimum		1	
	Maximum		11	
	Range		10	
	Interquartile Range		3	
	Skewness		-.786	.135
	Kurtosis		.271	.269

- **“Extreme Values” and the Boxplot** relate to each other. The boxplot is a graphical display of the data that shows: (1) median, which is the middle black line, (2) middle 50% of scores, which is the shaded region, (3) top and bottom 25% of scores, which are the lines extending out of the shaded region, (4) the smallest and largest (non-outlier) scores, which are the horizontal lines at the top/bottom of the boxplot, and (5) outliers. The boxplot shows both “mild” outliers and “extreme” outliers.

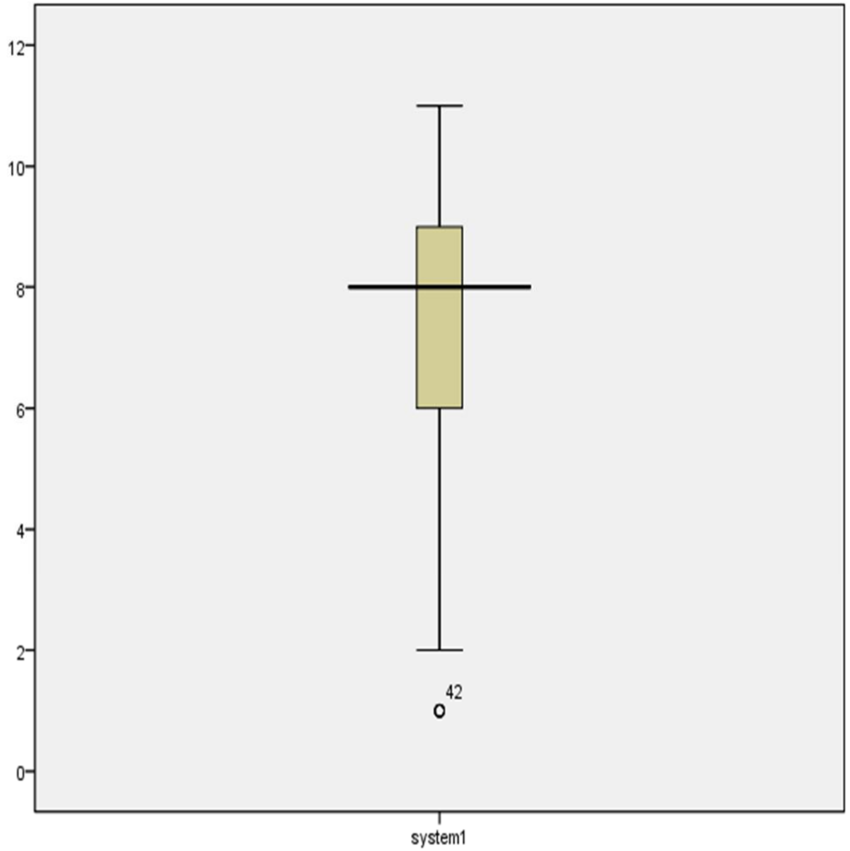


Extreme Values

			Case Number	Value
system1	Highest	1	62	11
		2	105	11
		3	135	11
		4	153	11
		5	183	11[a]
	Lowest	1	283	1
		2	259	1
		3	244	1
		4	42	1
		5	296	2[b]

a. Only a partial list of cases with the value 11 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 2 are shown in the table of lower extremes.



- . **Outliers - Multivariate**
- Multivariate outliers are traditionally analyzed when conducting correlation and regression analysis. The multivariate outlier analysis is somewhat complex, so I will discuss how to identify multivariate outliers when we get to correlation and regression.

- **If you want to reduce the influence of the outliers, you have four options**

1.Option 1 is to delete the value. If you have only a few outliers, you may simply delete those values, so they become blank or missing values.

2.Option 2 is to delete the variable. If you feel the question was poorly constructed, or if there are too many outliers in that variable, or if you do not need that variable, you can simply delete the variable. Also, if transforming the value or variable (e.g., Options #3 and #4) does not eliminate the problem, you may want to simply delete the variable.

- 3.Option 3 is to transform the value. You have a few options for transforming the value. You can change the value to the next highest/lowest (non-outlier) number. For example, if you have a 100 point scale, and you have two outliers (95 and 96), and the next highest (non-outlier) number is 89, then you could simply change the 95 and 96 to 89s. Alternatively, if the two outliers were 5 and 6, and the next lowest (non-outlier) number was 11, then the 5 and 6 would change to 11s.

4. Option 4 is to transform the variable. Instead of changing the individual outliers (as in Option #3), we are now talking about transforming the entire variable. Transformation creates normal distributions, as described in the next section below about “Normality”. Since outliers are one cause of non-normality, see the next section to learn how to transform variables, and thus reduce the influence of outliers.

- Third, after dealing with the outlier, you re-run the outlier analysis to determine if any new outliers emerge or if the data are outlier free. If new outliers emerge, and you want to reduce the influence of the outliers, you choose one the four options again. Then, re-run the outlier analysis to determine if any new outliers emerge or if the data are outlier free, and repeat again.

Normality

- **Normality**
- Below, I describe five steps for determining and dealing with normality. However, the bottom line is that almost no one checks their data for normality; instead they assume normality, and use the statistical tests that are based upon assumptions of normality that have more power (ability to find significant results in the data).
- First, what is normality? A normal distribution is a symmetric bell-shaped curve defined by two things: the mean (average) and variance (variability).

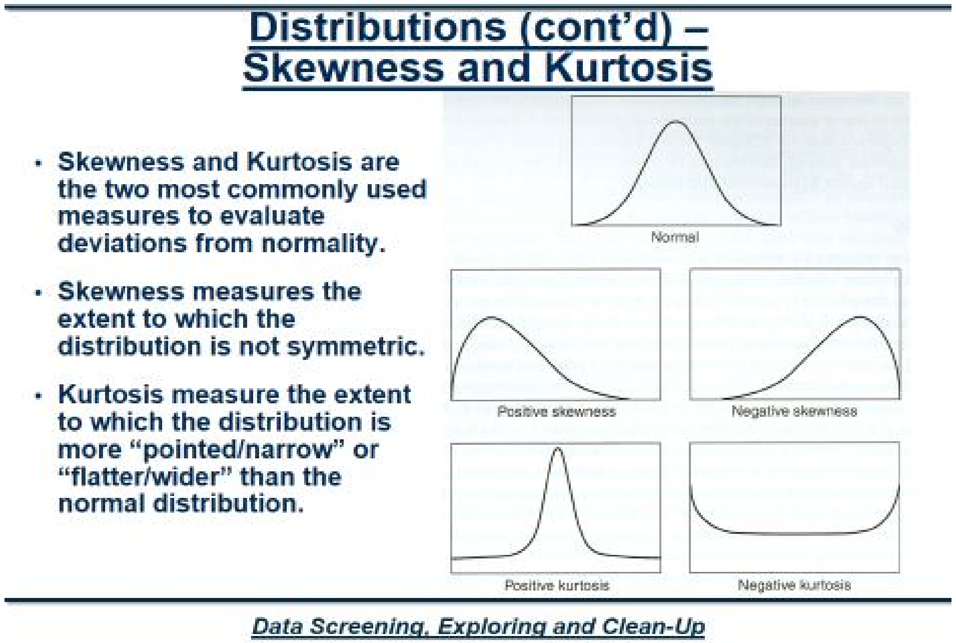
- Second, why is normality important? The central idea behind statistical inference is that as sample size increases, distributions will approximate normal. Most statistical tests rely upon the assumption that your data is “normal”. Tests that rely upon the assumption or normality are called **parametric tests**. If your data is not normal, then you would use statistical tests that do not rely upon the assumption of normality, call **non-parametric tests**. Non-parametric tests are less powerful than parametric tests, which means the non-parametric tests have less ability to detect real differences or variability in your data. In other words, you want to conduct parametric tests because you want to increase your chances of finding significant results.

- First, look at a **histogram** with the normal curve superimposed. A histogram provides useful graphical representation of the data. SPSS can also superimpose the theoretical “normal” distribution onto the histogram of your data so that you can compare your data to the normal curve.
- Second, look at the values of **Skewness** and **Kurtosis**. Skewness involves the symmetry of the distribution. Skewness that is normal involves a perfectly symmetric distribution. A positively skewed distribution has scores clustered to the left, with the tail extending to the right. A negatively skewed distribution has scores clustered to the right, with the tail extending to the left. Kurtosis involves the peakedness of the distribution. Kurtosis that is normal involves a distribution that is bell-shaped and not too peaked or flat. Positive kurtosis is indicated by a peak. Negative kurtosis is indicated by a flat distribution.

- Third, the descriptive statistics for Skewness and Kurtosis are not as informative as established tests for normality that take into account both Skewness and Kurtosis simultaneously. The **Kolmogorov-Smirnov test (K-S**) and **Shapiro-Wilk (S-W)** test are designed to test normality by comparing your data to a normal distribution with the same mean and standard deviation of your sample:
- 1.Select **Analyze** --**> Descriptive Statistics** --**> Explore.**
- 2. Move all variables into the “Variable(s)” window.
- 3. Click “Plots”, and unclick “Stem-and-leaf”, and click “Normality plots with tests”.
- 4. Click OK.

- Fourth, if your data are non-normal, what are your options to deal with non-normality? You have four basic options.
    - Option 1 is to leave your data non-normal, and conduct the parametric tests that rely upon the assumptions of normality. Just because your data are non-normal, does not instantly invalidate the parametric tests. Normality (versus non-normality) is a matter of degrees, not a strict cut-off point. Slight deviations from normality may render the parametric tests only slightly inaccurate. The issue is the degree to which the data are non-normal.
    - Option 2 is to leave your data non-normal, and conduct the non-parametric tests designed for non-normal data.
    - Option 3 is to conduct "robust" tests. There is a growing branch of statistics called "robust" tests that are just as powerful as parametric tests but account for non-normality of the data.
    - Option 4 is to transform the data. Transforming your data involving using mathematical formulas to modify the data into normality.

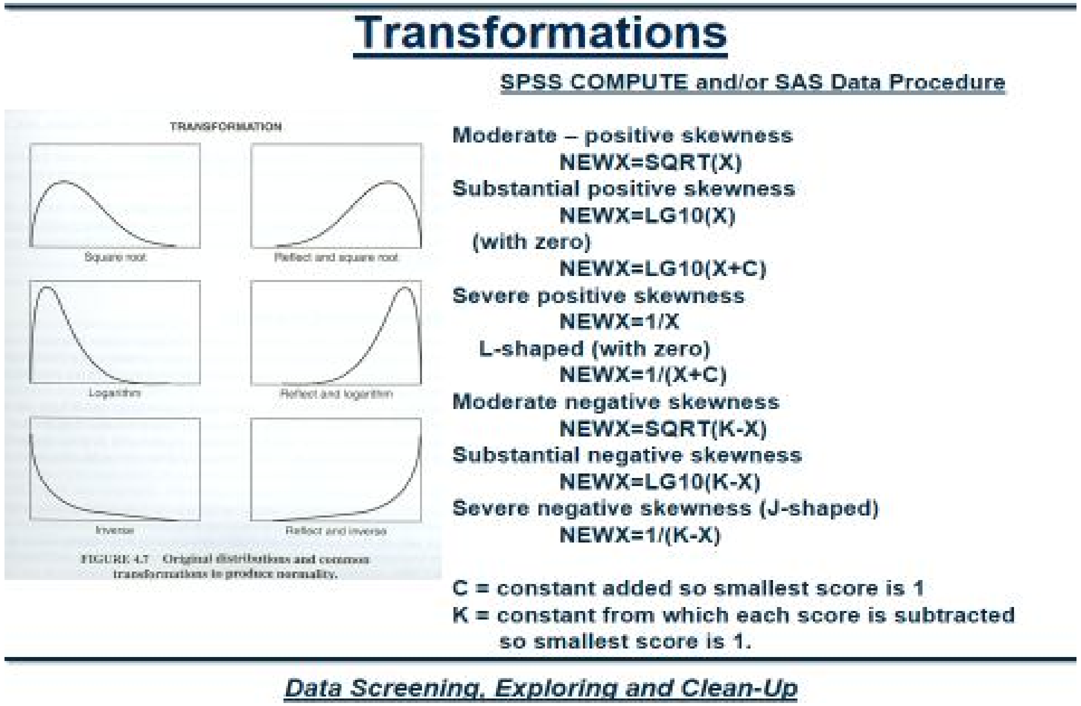
- Fifth, how do you transform your data into “normal” data? There are different types of transformations based upon the type of non-normality.



Distributions (cont’d) – Skewness and Kurtosis

- Skewness and Kurtosis are the two most commonly used measures to evaluate deviations from normality.
- Skewness measures the extent to which the distribution is not symmetric.
- Kurtosis measure the extent to which the distribution is more “pointed/narrow” or “flatter/wider” than the normal distribution.

Data Screening, Exploring and Clean-Up



Transformations

SPSS COMPUTE and/or SAS Data Procedure

Moderate – positive skewness
NEWX=SQRT(X)
Substantial positive skewness
NEWX=LG10(X)
(with zero)
NEWX=LG10(X+C)
Severe positive skewness
NEWX=1/X
L-shaped (with zero)
NEWX=1/(X+C)
Moderate negative skewness
NEWX=SQRT(K-X)
Substantial negative skewness
NEWX=LG10(K-X)
Severe negative skewness (J-shaped)
NEWX=1/(K-X)

C = constant added so smallest score is 1
K = constant from which each score is subtracted so smallest score is 1.

FIGURE 4.7 Original distributions and common transformations to produce normality.

Data Screening, Exploring and Clean-Up

Thank you