Data Screening using SPSS for beginner: Outliers, Missing Values and Normality

Donald Stephen
Institute of Borneo Studies, Universiti Malaysia Sarawak

Before we conduct the actual statistical tests, we need to screen our data for any irregularity. Usually, we check for:

(a) if data have been entered correctly, such as out-of-range values. It may be caused by human error in data entry (eg: entering "22" when it is supposed to be "2" for likert scale item)
(b) for other kind of outliers. Outliers are suspiciously larger or smaller observation (data) than the majority of the observations.
(c) for missing values. Is it because of you miss out entering some data or your participant did not provide a response for some questions.
(d) for checking assumptions before conducting tests (eg. Normality)

A. OUTLIERS

Outliers are observations that differ greatly from the majority of a set of data. Outliers can affect the normality of your data, although some researchers are against the idea of removing outliers simply because it does not fit the normality assumption.

If the analysis to be conducted does contain a grouping variable, such as t-test, ANOVA, among others, then data should be assessed for outliers separately within each group (gender, race). *refer to "Outliers within group" below

If the statistical analysis to be performed does not contain a grouping variable, such as linear regression, correlation, then the data set should be assessed for outliers as a whole (no need grouping variable).

There are different kind of outliers

I. Univariate outliers

Univariate outliers are extreme values on a single variable. This can be a good way to detect any wrong data entry (refer (a) above).

To find a incorrectly entered data,

1. Select **Analyze --> Descriptive Statistics --> Frequencies**
2. Move all variables into the "Variable(s)" window.
3. Click OK.

Using Exercise 1 data, I try to detect any mistakes in data entry for Q1.

Q1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	strongly disagree	2	10.0	10.0	10.0
	disagree	2	10.0	10.0	20.0
	neutral	6	30.0	30.0	50.0
	agree	5	25.0	25.0	75.0
	strongly agree	4	20.0	20.0	95.0
	11.00	1	5.0	5.0	100.0
	Total	20	100.0	100.0	

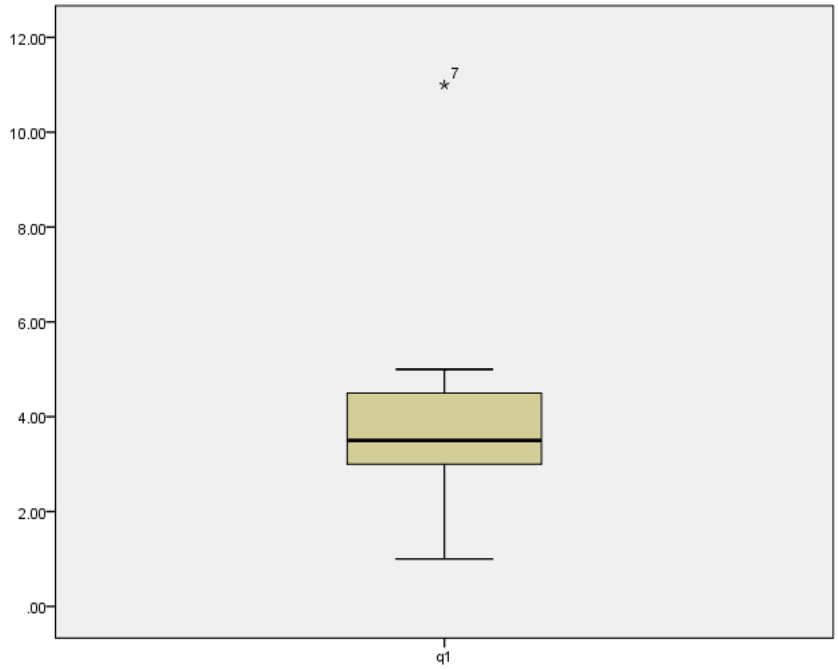
From this frequency table, there is suddenly a number "11", although Likert scale I used only ranges from 1-5. This may be caused by my mistake in entering data. It should be "1" not "11". So, I should change this score to "1". The question now, I know that I made a mistake, but how do I find the location of the data? To quickly find and replace:

1. **Edit --> Find**
2. Select Replace**,** fill in "11" in Find and "1" in Replace with
3. Click Replace

Using boxplot to find 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.

This will give you a boxplot.



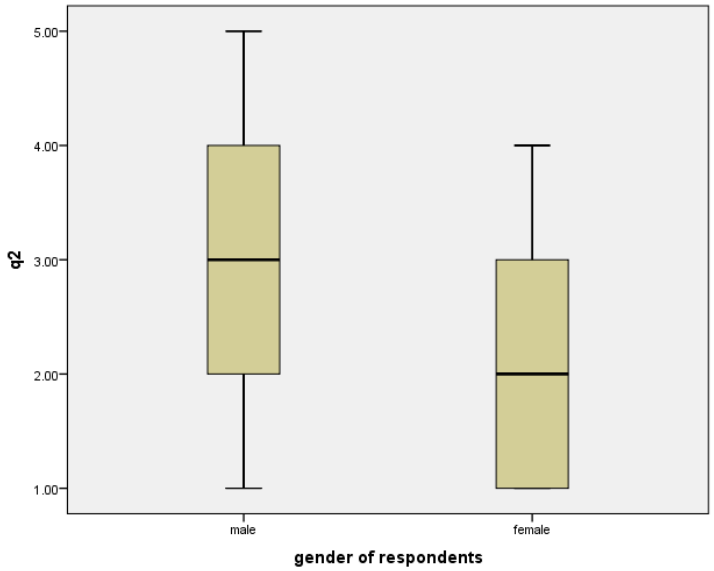
You can see there is a data point outside of the box that's shows extreme value. In this case, it is participant number 7. To pinpoint the exact location, you can double click on the boxplot, right click on the outlier, and then click "go to case". It will bring to straight to the outlier in your data view.

"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".

Outliers within group: If you need to conduct tests that involve categorical variable (sex, for example), you can detect outliers based on sex.

- Select **Analyze --> Descriptive Statistics --> Explore**
- Move all variables into the "Variable(s)" window.
- *Move "sex" into the "Factor List"*

- Click "Statistics", and click "Outliers"
- Click "Plots", and unclick "Stem-and-leaf"
- Click OK.



I try to detect outliers for Q2 and there seems to be no significant outliers in here.

II. Multivariate outliers

Multivariate outliers are traditionally analyzed when conducting correlation and regression analysis. Multivariate outliers are cases that have an unusual combination of values for a number of variables. For example, performing multivariate outliers for the set of independent variables in our data analysis. Multivariate outlier analysis is somewhat complex, most popularly computed using Mahalanobis D^2 (Multi-dimensional version of z-score). If there are only 2 variables, that is Bivariate outliers.

Generally, you first look for univariate outliers, then proceed to look for multivariate outliers. To identify multivariate outliers using Mahalanobis distance in SPSS, you will need to use Regression function:

- Go to **Analyze → Regression → Linear**
- Fill in your dependent variable (note that your dependent variable has no effect on mahalanobis distance)
- Pill in your independent variable (e.g: composite score of three constructs)
- Click **Save** and select **Mahalanobis** under option **Distances**
- Click **OK**

You will have a new variable in your data set named as MAH_1. You will need to compare this Mahalanobis distance to a chi-square distribution according to the same degree of freedom. The degree of freedom in this case equals to the number of predictors (independent variables).

- Go to **Transform → Compute Variables**
- Define a new name for this variable (e.g. PMAH)
- Find and select cumulative distribution function for chi-square **CDF.CHISQ**

- In **Numeric Expression** box, subtract this function with 1, and fill in your mahalanobis distance together with degree of freedom. You will get something like **1-CDF.CHISQ(MAH_1,3)** Note that 3 is the df.
- Click OK.
  You will have another variable in data set called PMAH.

In the PMAH variable, any value that is **less than** the cut-off point .001 are considered outliers. 0.001 is a common practice when identifying multivariate outliers. It would be a good idea to short your variable ascending/descending to help you detect.

There are several options on how to deal with outliers that is NOT caused by error in data entry. You can delete the value (so it becomes a missing value), delete the variables (if there way too many outliers), you can also transform the value (many ways to transform, and also a controversial way). You should re-run the outlier analysis after you deleted the outliers to determine if any new outliers emerge or if the data are outlier free.

B. MISSING VALUES

You will get missing values when you participants (purposely/accidentally) did not answer some questions. It may also occur through data entry mistakes. You can detect if there is missing value in your data using **Frequencies.**

1. Select **Analyze --> Descriptive Statistics --> Frequencies**
2. Move all variables into the "Variable(s)" window.
3. Click OK.

What to do when you have missing value? There are few options: The most popular option is to do nothing about the data because usually you don't have much of missing values. SPSS will treat your missing values differently depending on how you want SPSS to treat them: Listwise deletion (SPSS will simple omit your missing values in computation. This is the default option in SPSS), as well as pairwise deletion (SPSS will include all). You can also delete cases with missing values. Be careful because it can reduce your sample size and throw away all other valuable data (not recommended unless a participant really didn't answer most of the questions). Last option is to replace the missing value. There are several ways of how missing values can be replaced (of what to replace missing value with). For example, Mean Substitution, Multiple Imputation and Expectation Maximization algorithm.

Mean Substitution has a lot of critiques. In a way, it affect correlation between variables. Using Multiple Imputation method, you are restricted to a certain tests that allows imputation of data. SPSS will indicate it in special spiral icon to show if a certain analysis is available for imputated data. Expectation Maximization algorithm has becoming increasingly popular as a way to substitute missing values. It is available in **Analyze → Missing Value Analysis →** Select **EM**. Define a new name for this data set.

C. NORMALITY

Some statistical tests (parametric) requires you to have normal data. This is why you need to conduct normality test to ensure you actually have a normal distribution. A normal distribution is a symmetric bell-shaped curve defined by two things: the mean (average) and variance (variability). If your data is not normal, then you should consider using non-parametric tests, of which do not require you to have normality. There are also many ways to test normality of your data.

<u>Shapiro Wilk W/ Kolmogorov- Smirnov test</u>

Shapiro Wilk W test is considered by some authors to be the best test of normality. If you have small data, it is the best choice. It can handle up to 2000 data. For both tests, if you have more than 2000 data, use Kolmogorov-Smirnov test. If you reject Ho, it indicates a non-normal data.

Normality test hypotheses
H0: The observed distribution fits the normal distribution
H1: The observed distribution does not fit the normal distribution
(if we fail to reject Ho, we assume our data is normal)

What to do if your data is not normal? First option is to leave your data like that. Just because Shapiro-Wilk (S-W) said it is not normal it doesn't mean your data is unfit for parametric tests (tests that requires normal data). Your S-W test can give you significant result but your skewness and kurtosis look OK [between -2 and +2] (We will talk more about skewness and kurtosis later). Also use histogram to see how your data look like. Maybe you can proceed with parametric test. Secondly, you can also accept that your data is not normal and proceed with non-parametric tests. Bear in mind that non-parametric tests are generally less powerful (but more flexible) than its parametric counterpart. Alternately you can use Robust tests (not recommended in your case). Robust statistics seeks to provide methods that emulate popular statistical methods, but more flexible and not really affected by some popular assumptions for parametric tests (Outliers, normality). Next, you can check for outliers (if you have not) because outliers can influence the normality of your data. Remove outliers and you will likely to notice improvement on your normality. Lastly you can transform your data (also not recommended).

1. **Analyze --> Descriptive Statistics --> Explore**
2. Put the variable you want to test in **Dependent** box. I am still using lab one data so I use my horizontal mean of all the questions.
3. **Plots** --> **Normality plots with tests** (you can also select histogram)
4. Click **Continue** then **OK**

Descriptives

			Statistic	Std. Error
hor_mean	Mean		3.1700	.18254
	95% Confidence Interval for Mean	Lower Bound	2.7879	
		Upper Bound	3.5521	
	5% Trimmed Mean		3.1667	
	Median		3.3000	
	Variance		.666	
	Std. Deviation		.81635	
	Minimum		1.80	
	Maximum		4.60	
	Range		2.80	
	Interquartile Range		1.35	
	Skewness		-.190	.512
	Kurtosis		-1.066	.992

Sometimes, Skewness and Kurtosis is enough to see whether your data fit the assumption of normality. **Skewness** is a measure of symmetry. This is to know whether your data is skewed to the left or to the right of the center point. **Kurtosis** is a measure of whether the data are peaked or flat relative to a normal distribution (the height). The general rule of thumb is that, the values must be in between -2 to +2. Some researchers and scholars might suggest different range. It shows that this is not a rigid thing.

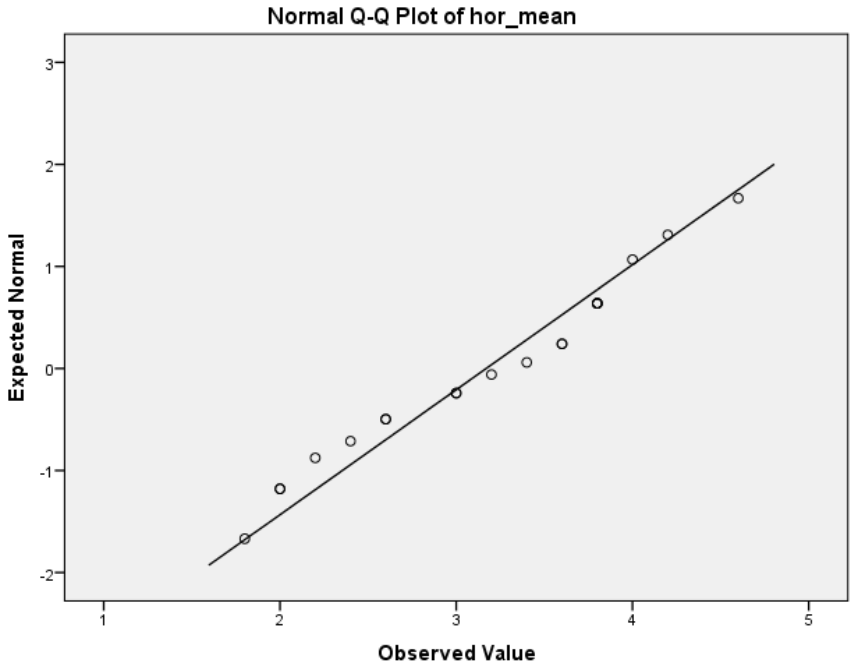
After all the normality assumption itself it not strict. So, my Skewness is -0.190 and my Kurtosis is -1.066. Both values fall under the range of -2 and +2.

What about the S-W test?

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
hor_mean	.151	20	.200*	.949	20	.357
*. This is a lower bound of the true significance.						
a. Lilliefors Significance Correction						

My data is really small (N=20) so I use Shapiro-Wilk. If the test is NOT significant (more than .05), then the data are normal. So, from here, my Sig. value is very large (>0.05), which is 0.357, indicating that I have a normal data.

Normal Q-Q Plot" provides a graphical way to determine the level of normality. The dots are your actual data. If the dots fall somewhere near the black line, then your data are normal.



My data points are close to the line. If you see some of your data points are located far away from the line, it means your data may not be normal.

End

Disclaimer: This note is for educational purpose only and author does not make profit from this. SPSS statistics is a property of IBM. Support the software by purchasing legal copy of it.