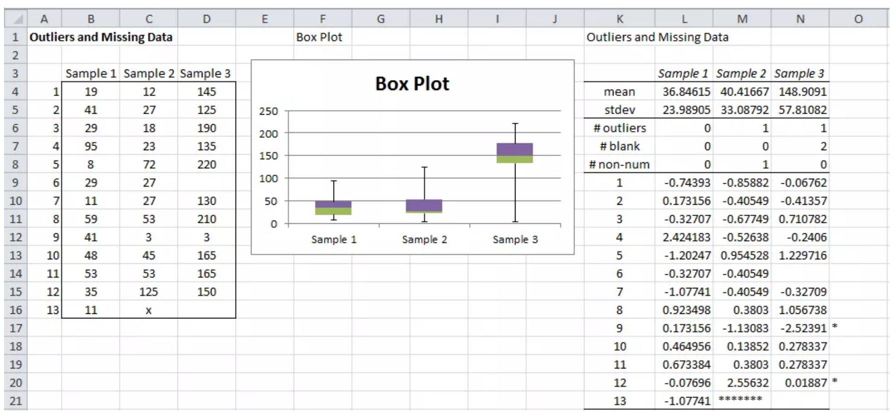
Identifying Outliers and Missing Data

The Real Statistics Resource Pack provides an option for identifying potential outliers in a sample. Assuming the sample is normally distributed (based on the Central Limit Theorem), we know that 1–NORMSDIST(2.5) = 0.621% of the data should have a z-score larger than 2.5 or less than -2.5. Here we use 2.5 as a somewhat arbitrary criteria for a potential outlier. E.g. for a sample of size 80, on average 80(.00621)(2) = .994, or about one, element will be viewed as a potential outlier.

Real Statistics Data Analysis Tool: One of the options of the **Descriptive Statistics and Normality** data analysis tool provided in the Real Statistics Resource Pack is the identification of potential outliers using a specified z-score (default 2.5).

Example 1: Identify potential outliers for the three data samples on the left side of Figure 1 (range B3:D16).



Outliers and Missing Data

	Sample 1	Sample 2	Sample 3
1	19	12	145
2	41	27	125
3	29	18	190
4	95	23	135
5	8	72	220
6	29	27	
7	11	27	130
8	59	53	210
9	41	3	3
10	48	45	165
11	53	53	165
12	35	125	150
13	11	x	

Outliers and Missing Data

	Sample 1	Sample 2	Sample 3	
mean	36.84615	40.41667	148.9091	
stdev	23.98905	33.08792	57.81082	
# outliers	0	1	1	
# blank	0	0	2	
# non-num	0	1	0	
1	-0.74393	-0.85882	-0.06762	
2	0.173156	-0.40549	-0.41357	
3	-0.32707	-0.67749	0.710782	
4	2.424183	-0.52638	-0.2406	
5	-1.20247	0.954528	1.229716	
6	-0.32707	-0.40549		
7	-1.07741	-0.40549	-0.32709	
8	0.923498	0.3803	1.056738	
9	0.173156	-1.13083	-2.52391	*
10	0.464956	0.13852	0.278337	
11	0.673384	0.3803	0.278337	
12	-0.07696	2.55632	0.01887	*
13	-1.07741	*******		

Figure 1 – Identifying potential outliers and missing data

Enter **Ctrl-m** and select the **Descriptive Statistics and Normality** data analysis tool. Fill in the dialog box that appears as shown in Figure 2. We leave the **Outlier Limit** field blank since we want to use the default value of 2.5.

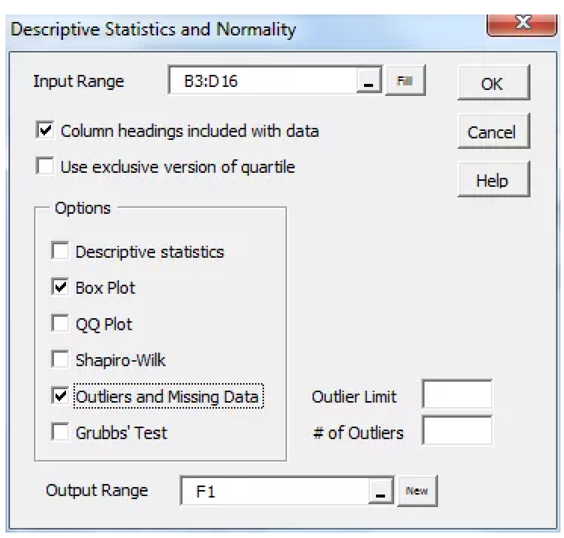


Figure 2 – Dialog box for Descriptive Statistics and Normality

The output as displayed in Figure 1 shows there are two potential outliers (indicated by the asterisks in column N): namely item 9 of Sample 3 and item 12 of Sample 2. Item 9 of Sample 3 has a z-score of -2.52391 (cell N17), which is less than -2.5. Item 12 of Sample 2 has a z-score of 2.62457 (cell M20), which is greater than 2.5. Item 4 of Sample 1 is just under the 2.5 threshold with a z-score of 2.424183.

Note that the z-scores are calculated in the usual way: e.g. the z-score for item 1 of Sample 1 (cell L9) is calculated by the formula =STANDARDIZE(B4, L4, L5).

We can also see evidence of potential outliers in the long tails in the Box plots for Sample 2 and 3.

Note that just because a data element is identified as a potential outlier doesn't mean that it is wrong or should be eliminated, but it does mean that that data element should be investigated to see if a typing mistake has been made or some other problem has occurred that will distort any analyses that are undertaken.

The output also totals up the number of blank cells (2 for Sample 3) and non-numeric cells (1 for Sample 2, indicated by a series of asterisks). These can represent potential missing data. See Dealing with Missing Data for how to deal with missing data.

Observation: Another popularly used method for identifying outliers is to denote any data element larger than Q3 + 1.5*IQR or smaller than Q1 – 1.5*IQR as a potential outlier, where Q1 and Q3 are the first and third quartiles (see Ranking) and IQR is the inter-quartile range (see Measures of Variability).

Real Statistics Function: The Real Statistics Resource Pack provides the following function, where if *type* = 0 then the test using the mean and standard deviation is employed while if *type* = 1 then the test using the IQR is employed.

STANDARD(*x*, R1, *type, exc*) takes the value

STANDARDIZE(x, $\bar{x}$, s)	if *type* = 0 (default)
(x–Q3)/IQR	if *type* = 1 and x > Q3
(x–Q1)/IQR	if *type* = 1 and x < Q1
0	otherwise (*type* = 1 and Q1 ≤ x ≤ Q3)

where $\bar{x}$ = AVERAGE(R1) and s = STDEV(R1).

If *exc* = TRUE then Q1 = QUARTILE.EXC(R1,1) and Q3 = QUARTILE.EXC(R1,3)
If *exc* = FALSE (default) then Q1 = QUARTILE.INC(R1,1) and Q3 = QUARTILE.INC(R1,3)

The STANDARD function plays a role similar to the STANDARDIZE function when *type* = 0 (except that the mean and standard deviation are calculated from R1). It plays the equivalent role using the median and IQR when *type* = 1.

For *type* = 0, if the value of the STANDARD function at x is larger than 2.5 or less than -2.5 we can consider x to be a potential outlier (although we can change 2.5 to 3.0 or some other value as we choose).

Similarly for *type* = 1, if the value of the STANDARD function at x is larger than 1.5 or less than -1.5 we can consider x to be a potential outlier (although we can change 1.5 to some other value as we choose).

Example 2: Identify potential outliers for the data set in range B3:D8 of Figure 3.

We insert the formula =STANDARD(B3,$B$3:$D$8) in cell F3, highlight the range F3:H8 and press **Ctrl-R** and **Ctrl-D** to fill in the range F3:H8 with the values shown in Figure 3. If we set a cutoff of ±2.5 for outliers, we see that the only value exceeding the cutoff is 2.816972 (cell H5), which means that the data element 99.5 (cell D5) is a potential outlier.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Identifying outliers					mean and stdev				q1, q3 and iqr		
2												
3		34.7	50.3	47.3		-0.5858	0.233386	0.075851		-0.28294	0.064795	0
4		23.7	44.2	39.4		-1.16343	-0.08694	-0.33899		-1.23326	0	0
5		45.8	39.9	99.5		-0.00292	-0.31274	2.816972		0	0	4.315335
6		37.5	41.7	52.8		-0.43877	-0.21822	0.364666		-0.04104	0	0.280778
7		42.8	36.1	71.1		-0.16045	-0.51228	1.325634		0	-0.16199	1.861771
8		10.6	41.4	66.6		-1.85134	-0.23397	1.08933		-2.36501	0	1.473002
9												
10						mean	45.85556			median	42.25	
11						stdev	19.0433			q1	37.975	
12										q3	49.55	
13										iqr	11.575	

Figure 3 – Identifying potential outliers using the STANDARD function

If we use the type 1 approach instead of the type 0 approach, we insert the formula =STANDARD(B3,$B$3:$D$8,1) in cell J3, highlight the range J3:L8 and press **Ctrl-R** and **Ctrl-D** to fill in the range J3:L8 with the values shown in Figure 3. If we set a cutoff of ±1.5 for outliers, we see that the only values exceeding the cutoff are 4.315335 (cell L5) and -2.36501 (cell J8), which means that the data elements 99.5 (cell D5) and 10.6 (cell B8) are potential outliers.

Note that values in the range B3:D8 between Q1 (37.975) and Q3 (49.55) take a zero value in range J3:L8.

Observation: Another approach to identifying outliers uses the Grubbs' test.

11 Responses to *Identifying Outliers and Missing Data*

Leo *says:*
September 29, 2016 at 10:48 am

Very useful info and detailed even for a beginner. Thank you for the Add-In, much appreciated for ist complexity.

Reply

Pandey *says:*
September 13, 2016 at 6:49 am

Is there a way to plot outliers outside the box plot? Currently the box plot includes outliers in the range.

Reply

Charles *says:*
September 13, 2016 at 11:09 am

Pandley,
So far I have not found a way to do this in Excel, except for Excel 2016 which has a built-in boxplot chart capability.
Charles

Reply

Pandey *says:*
September 14, 2016 at 6:36 am

I see. I think it might be possible in following way:

1. User selects outlier limit to identify outliers before doing 'descriptive statistics and normality'.

2. Once identified, outliers are separated from the original data. So we identify three data sets now:
A) Original dataset
B) Dataset containing outliers only
C) Dataset containing original data with outliers removed

3. Make a box plot with dataset C. Then plot dataset B as separate series in the same chart (as a scatter plot). This way we may end up with

I have performed this manually but it gets tedious with large number of outliers. I think this will further enhance this wonderful add-in.

Reply

Pandey *says:*
September 14, 2016 at 7:15 am

I think I did not mention clearly that in step 3, each row of outlier data has to be plotted as seperate scatter plot data series.

Reply

Charles *says:*
September 23, 2016 at 6:07 pm

Pandey,
What is the purpose of data set B?
Charles

Reply