# Removing Outliers

An outlier is a data value that is out of keeping with the other values. This could either be caused by a measurement or recording error, e.g. recording a long jump distance as 67.2 metres instead of 6.72 metres, or a genuine freak result, e.g. a long jump distance of 8.95 metres which stood as the World Record for 30 years.



Informally, outliers can be determined by eye. More formally, a statistical test can determine if a data value is an outlier. It is important to identify outliers, and if it is appropriate, remove them from the data, as they can affect any conclusions drawn.

**Example:** The following shows the height and weight of dogs in a kennel.

|  |  |
| --- | --- |
| Height | Weight |
| 53 | 2.5 |
| 55 | 2.7 |
| 56 | 3.9 |
| 58 | 2.8 |
| 59 | 3 |
| 62 | 2.2 |
| 65 | 3.3 |
| 66 | 3.9 |
| 67 | 8.1 |
| 70 | 4 |
| 75 | 4.3 |

(a) Type the information into excel.

(b) Create another table with a **Before** column and calculate the following;

Mean height, mean weight, standard deviation of height and weight, correlation and coefficient of determination.

(c) Create a scatter plot to show the information and remove any outliers.

(d) Repeat part b and create a column in the table labelled **After**.

|  |  |  |
| --- | --- | --- |
|  | Before | After |
| mean Height | 62.36363636 | 62.625 |
| mean Weight | 3.7 | 3.3125 |
| sd Height | 6.55428697 | 7.157819 |
| sd Weight | 1.5385944 | 0.631343 |
| correlation | 0.487702498 | 0.974696 |

(e) What numerical statistics do you see change and what does this tell us?

|  |
| --- |
| The correlation becomes much higher. |

**To drop or not to drop?**

1. If it is obvious that the outlier is due to incorrectly entered or measured data you should drop the outlier, e.g. if you are analysing the weight of a sample of women and the answer is 19 lbs, then you know this is physically impossible and may be a mistype.

2. If the outlier does not change the results but does affect some of the assumptions then you may drop the outlier, e.g. neither the presence of absence of this outlier would change the outcome.



3. More commonly, the outlier affects both results and assumptions. In this situation, it is *not* legitimate to drop the outlier.



4. If the outlier *creates* a significant association, you should drop the outlier and not report any significance. From the diagram below you can see that the linear relationship is only created due to the outlier.



**Example:** A group of patients are all suffering from the same condition in a hospital ward. Their temperature is in degrees Celsius are recorded as follows.

37.2 36.8 40.0 38.2 38.5 63.9 39.6 38.0 39.3 38.0

(a) Identify the outlier in these figures and explain why it should be discarded.

|  |
| --- |
| 63.9, much higher than the other numbers |

(b) Using Excel, Calculate the difference in the mean before and after removing the outlier.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TempBefore | TempAfter |  | Before | After |
| 37.2 | 37.2 |  | mean | 40.95 | 38.4 |
| 36.8 | 36.8 |  | sd | 7.709767 | 1.00995 |
| 40 | 40 |  |  |  |  |
| 38.2 | 38.2 |  |  |  |  |
| 38.5 | 38.5 |  |  |  |  |
| 63.9 | 39.6 |  |  |  |  |
| 39.6 | 38 |  |  |  |  |
| 38 | 39.3 |  |  |  |  |
| 39.3 | 38 |  |  |  |  |
| 38 |  |  |  |  |  |

 |

(c) Without calculating what would happen to the standard deviation and SIQR of the information above if you removed the outlier?

|  |
| --- |
| You would expect it to go down. |

With regards to linear regression, if you were to remove the outliers how might that be a benefit?

|  |
| --- |
| The line would become much better at being able to make predictions. |

**Exercise: Outliers**

1. The following data shows the amount of times a website is used every month.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Times used (Thousands) | 1.4 | 1.6 | 1.5 | 1.9 | 0.25 | 2.3 | 2.4 | 2.8 | 2.6 | 2.7 | 5.06 | 2.9 |

a) Type the following into an Excel Spreadsheet, create a scatterplot and calculate the following put your answers into the ‘Before’ column.

|  |  |  |
| --- | --- | --- |
|  | Before | After |
| Mean Times Used |  |  |
| Standard Deviation Times Used |  |  |
| Linear Expression |  |  |
| Correlation |  |  |
| R2 (Coefficient of determination) |  |  |

|  |  |  |
| --- | --- | --- |
|  | Before | After |
| mean | 2.284166667 | 2.21 |
| sd | 1.110071006 | 0.537494186 |
| correlation | 0.731663064 |  |
| linear regression | y = 0.2353x + 0.7548 | y = 0.1495x + 1.2831 |

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

2. The following data shows the amount of trees being cut down in a rain forest each year.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Amount of Tree(Ten Thousand) | 12 | 11.6 | 10.9 | 10.8 | 28 | 42 | 10.5 | 10.2 | 9.9 | 9.6 | 9.3 | 9.2 | 9.5 | 8.5 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

3. The following data shows the amount of rainfall each year in the month of November

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Average Rainfall cm | 0.04 | 4.8 | 5.2 | 5 | 5.5 | 5.8 | 1.7 | 6.7 | 6.5 | 6.9 | 11.8 | 6.8 | 7.1 | 7.5 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

4. The following shows a cars average speed vs its fuel consumption

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Average Speed | 28 | 29 | 33 | 36 | 37 | 48 | 49 | 50 | 50 | 52 | 55 | 56 | 57 | 61 | 70 | 75 | 81 |
| Fuel Consumption | 78 | 76 | 24 | 72 | 70 | 65 | 68 | 62 | 63 | 58 | 26 | 54 | 53 | 52 | 50 | 48 | 19 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.