



Raw Data Processing ^[1]

Admin Name ^[2]41.1K reads

Raw data processing refers to the refining of raw data that has been collected from the experiment.

Statistical data that is used to draw conclusions ^[3] and inferences should be accurate and consistent. This is important in order to ensure the validity ^[4] of all the inferences drawn on the basis of the data.

"Raw data is a term for data collected on source which has not been subjected to processing or any other manipulation." ([wikipedia.org](https://www.wikipedia.org)) ^[5]

Organizing the Data

Raw data is unprocessed/unorganized source data, such as the data from an eyetracker which records the coordinates and movement of the eye every millisecond. Output data ^[6] is the processed/summarized/categorized data such as the output of the mean position for a participant immediately after a stimulus was presented.

Raw data processing ^[7] is required in most surveys ^[8] and experiments ^[9]. At the individual level, data needs to be processed because there may be several reasons why the data is an aberration.

The raw data collected is often contains too much data to analyze it sensibly. This is especially so for research using computers as this may produce large amounts of data. The data needs to be organized or manipulated using deconstruction analysis techniques.

Removal of Outliers

While measuring the current flow through a resistor with the help of an ammeter, there may be one data point that is far away from the rest, an statistical outlier ^[10].

This may be due to a sudden surge in voltage in the source, and this data point is therefore a deviant.

Statistical raw data processing needs to be carried out in this case to eliminate this data point in order to ensure accuracy of the conclusions drawn based on the experiment.

Data Manipulation

In social experiments [11] involving surveys [8], there are a number of possibilities why a given data set might need to be edited or processed. For example, the researcher finds an error in a question which makes it invalid. Participants may also have checked the wrong answer or may have simply misunderstood or skipped a question.

It is also important to extract exactly the information that is needed from the overall experiment.

For example, census data provides a wealth of geographic and demographic data, but a researcher might need only certain segments of the data from certain locations.

Therefore raw data processing would be required in order to correctly extract the information required without errors.

Raw data processing can be a time consuming task and it is not always easy to catch anomalies. Therefore simple checks should be run that are quite effective in eliminating the abnormalities.

For example, a predefined range can be defined from most parameters that can be obtained in theory. Suppose a researcher is studying the amount of salts in a lake by averaging at different locations. At one particular location, it is possible that there is a sudden surge in salt levels, which is an anomaly and can happen if say, someone at a picnic dumped some salt there.

However, this anomaly can be caught by determining a predefined range for the value of salt content in the lake that are usually available in literature.

Such small tests often are very effective in raw data processing.

Source URL: <https://staging.explorable.com/en/raw-data-processing>

Links

[1] <https://staging.explorable.com/en/raw-data-processing>

[2] <https://staging.explorable.com/en>

[3] <https://staging.explorable.com/drawing-conclusions>

[4] <https://staging.explorable.com/types-of-validity>

[5] http://en.wikipedia.org/wiki/Raw_data

[6] <https://staging.explorable.com/data-output>

[7] <http://www.statcan.gc.ca/edu/power-pouvoir/ch3/editing-edition/5214781-eng.htm>

[8] <https://staging.explorable.com/survey-research-design>

[9] <https://staging.explorable.com/experimental-research>

[10] <https://staging.explorable.com/statistical-outliers>

[11] <https://staging.explorable.com/social-psychology-experiments>