Examine real world data

3 minutes

Data presented in educational material is often remarkably perfect, designed to show students how to find clear relationships between variables. 'Real world' data is a bit less simple.

Because of the complexity of 'real world' data, raw data has to be inspected for issues before being used.

As such, best practice is to inspect the raw data and process it before use, which reduces errors or issues, typically by removing erroneous data points or modifying the data into a more useful form.

Real world data issues

Real world data can contain many different issues that can affect the utility of the data, and our interpretation of the results.

It's important to realize that most real-world data are influenced by factors that weren't recorded at the time. For example, we might have a table of race-car track times alongside engine sizes, but various other factors that weren't written down—such as the weather—probably also played a role. If problematic, the influence of these factors can often be reduced by increasing the size of the dataset.

In other situations data points that are clearly outside of what is expected—also known as '*outliers*'—can sometimes be safely removed from analyses, though care must be taken to not remove data points that provide real insights.

Another common issue in real-world data is bias. Bias refers to a tendency to select certain types of values more frequently than others, in a way that misrepresents the underlying population, or 'real world'. Bias can sometimes be identified by exploring data while keeping in mind basic knowledge about where the data came from.

Remember, real-world data will always have issues, but this is often a surmountable problem. Remember to:

- Check for missing values and badly recorded data
- Consider removal of obvious outliers
- Consider what real-world factors might affect your analysis and consider if your dataset size is large enough to handle this

- Check for biased raw data and consider your options to fix this, if found

Next unit: Exercise - Examine real world data

Continue >