COMMON ERRORS IN DATA AND STATISTICS

There are an unlimited number of ways that data can be in error. The most common types, in both data and statistical analysis, will vary from application domain to application domain and from problem to problem. Furthermore, a feature of a data set may count as a mistake or error in certain contexts, but not in others.

EXAMPLES OF COMMON ERRORS ARISING IN DATA COLLECTION

- mistaken entry. e.g. entering 2 instead of 7;
- digit transposition. e.g. recording 21 as 12;
- misplaced decimal point. e.g. recording 0.067 as 0.67;
- mistaken units. e.g. recording a result as 3in instead of 3cm.

EXAMPLES OF COMMON ERRORS IN STATISTICAL ANALYSIS

- p-hacking or fishing expeditions: searching through results to find the largest or most significant, and reporting only that;
- selection distortion: using (possible unawares) only part of the data;
- failing to take account of both substantive and statistical significance;
- incorrect or missing axes on graphs;
- use of an inappropriate type of average;
- treating missing value codes as genuine data values;
- failure to model a missing data mechanism, and ignoring possible consequent bias;
- failure to recognise regression to the mean;
- constructing the hypothesis after seeing the data and testing it as if the data were new;
- failure to use adequate sample size;
- the ecological fallacy: assuming that what applies to groups also applies to individuals;
- confusing correlation with causation;
- failure to allow properly for baseline values;
- misunderstandings of statistical concepts, such as p-values;
- failure to recognise that the observations are not independent;
- evaluating a model on the data which was used to construct it;
- failing to allow for ceiling/floor (topcoding/bottomcoding) data, in which large (small) values are replaced by fixed maxima (minima);

EXAMPLES OF TYPES OF DATA FRAUD

- deliberately concealing or distorting axes on graphs;
- plots which do not make it clear if length, area, or three-dimensional volume represents the quantity of interest;
- concealing p-hacking or fishing expeditions;
- selecting data which support a hypothesis;
- modifying the data to make it conform more closely to the theory;
- copying data from elsewhere;
- fabricating data;

December 2017