A short note on Initial data analysis

Initial data analysis (IDA), also known as exploratory data analysis (EDA) is an aspect of statistics that has grown popularity in recent years. One of the reasons is due to the modern computer and dedicated software that can present data instantly in a wide range of graphical formats for viewing. The median and the IQR (interquartile range) are just two of the statistics that feature strongly in IDA.

What do we know about IDA?

IDA is basically geared to analyze data by plotting the data collected on a study sample or population over a period of time. This forms an initial step of all data analyses in checking their consistency and accuracy, exploring the study item and preparing the data for further analysis.

It is prudent that this is done before embarking on further and more complex statistical analyses of the data. Though IDA may look crude on the surface as its graphical display is meant for visual inspection of which some critics comment that it is too subjective, its outcome however provide reasonable statistical inferences for following up actions.

Examples of IDA techniques

Dot-plots to identify outliers

When you have got a series of data, you can easily put all the results in dot form over a straight line. Any suspect value(s) will become obvious as it (they) appear at high or low end or both ends of the measurement range. Figure 1 illustrates the spread of data in the form of dot-plots with a group of data in Table 1:

Table 1: 10 repeated analyses on a sample of animal feeds for its fat content in %w/w

Data	2.2	2.5	2.1	2.5	2.2	2.9	4.4	2.9	3.1	2.7

Figure 1: It is obvious from the dot-plot that the figure 4.5 is at the far extreme right of the measurement range

Another dot-plot example can be seen below for an interlaboratory cross-check exercise involving 4 laboratories testing on a similar sample for 4 replicates using the same analytical method. The reported results in terms of % recoveries are summarized as follows:

Lab #	1	2	3	4
Repeat 1	95	105	95	105
Repeat 2	98	97	98	112
Repeat 3	102	101	100	110
Repeat 4	99	104	97	107

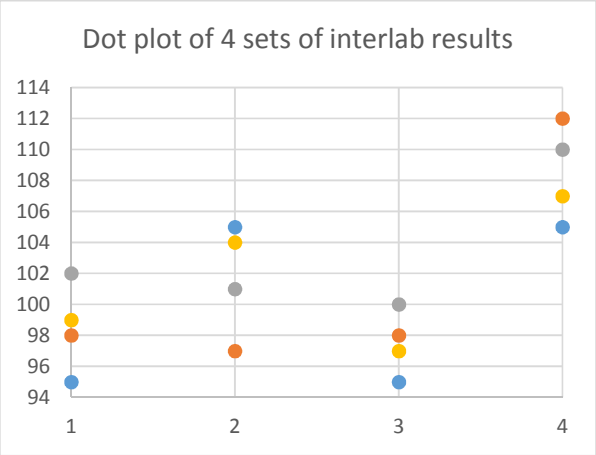

Figure 2: The dot-plot indicates that Lab#4 could have reported higher mean values than the others

The Boxplot

An informative graphical display that visually highlights the location and spread of a set of data and that is often highly suggesting of the need for inferential comparisons of means and standard deviations is a boxplot.

We designate the boxplots at the middle half of a data set by a rectangle whose lower and upper edges (or right and left edges, if drawn horizontally) , respectively, are the first quartile (Q1) and third quartile (Q3) of the data set. The median is indicated by a line segment drawn within the box parallel to the edges.

Hence, the box provides a useful visual impression that emphasizes the middle half of the data values and the line drawn at the median provides an explicit quantification of the center of the data. The mean value can also be indicated by a plotting symbol such as "+".

The result variability is portrayed by the lengths of vertical lines, often dashed, drawn from the edges of the boxes to upper and lower adjacent values calculated from the semi-interquartile range (SIQR) which is (Q3-Q1)/2.

Generally, the upper adjacent value is the largest observations in the data set that does not exceed Q3 + 3 x SIQR whereas lower adjacent value is the smallest data value that is no less than Q1 – 3 x SIQR. To be more critical, we have the liberty to choose a critical Student's t value (at α=0.025, degree of freedom, n-1) instead of the factor 3.

Any values larger or smaller, respectively than the upper and lower adjacent values are plotted individually because they are identified possible outliers, extreme data values whose magnitude are unusual relative to the bulk of the data under examination. A schematic boxplot example is in Figure 3.

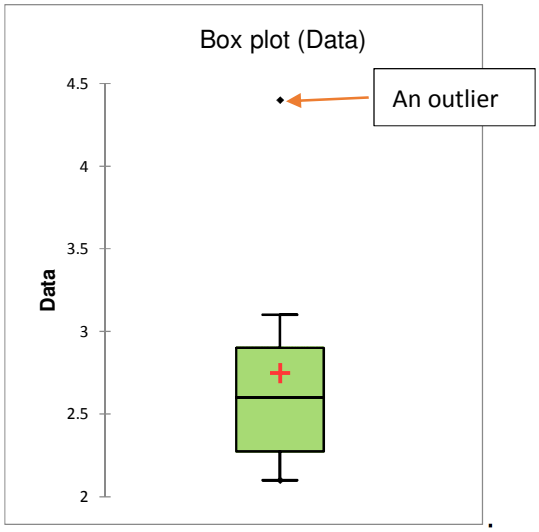

Figure 3: An example of schematic boxplot