1.2 Robust and/or believable?

Neither.

Robust is a term introduced by the statistician George Box (see *The Lady Tasting Tea*, D. Salsburg, Freeman / Owl Books, 2001) in the 1950s. It was relevant to devising tests on data that were 'contaminated', data that contained outliers, errors of measurement, or sample members simply drawn from a second and entirely different underlying distribution. Could testing processes be devised to reveal the true features of the original distributions under such circumstances?

One simple example of robustness is using the average as a measure of central location. What is the average salary? It is a very poor measure of what the average person gets because of the long tail to vast incomes, all the way up to Bill Gates. A far better measure of what most of us make is the median salary. The average generally is not a robust statistic and does not provide input to robust tests. The root-mean-square as a measure of standard deviation (see Chapter 2) is even worse; outliers get the squared treatment when added to the mix.

As a second example, described in Salsburg's book is the development in World War II by the US Navy of an optical range finder through which operators were require to place a triangle on the 3D image of the target. Several hundred operators were used in independent tests to determine the setting uncertainty. The results were a mess - because 20 per cent of humans cannot see stereoscopically, a fact not known at the time. 20 per cent of the measurements were random, contaminating the distribution. Robust statistical methods can now sort out the main distribution parameters from that of the contaminating distribution or distributions.

But when there is but one outlier as in this exercise, there is little chance of obtaining a 'robust' answer; a contaminating distribution cannot be identified by a single data point.

More interesting is the concept of *believability*. Yours or the readers? Context is crucial here. Was this a test to disprove a hypothesis? And was this a hypothesis that had stood the test of time, for which there was much other and independent supporting evidence? Or was it a straw-man hypothesis, perhaps invented by you? Was the hypothesis simply that objects of extreme luminosity such as the one quasar in question could NOT have the properties found for it in this sample? This is quite different.

The first question to ask is whether or not you yourself are convinced. Could there be an error of selection or an error of measurement? Would you be happy to see your conclusion based on this result in print? Now? Two or five years from now? Then ask whether the holder of the hypothesis that you are attacking / supporting will be convinced, or will at least pay heed to your result. Seems somewhat unlikely, does it not?

Then there is the word 'significance', a formal term in statistical inference. The formal calculation of significance is generally a scientific requirement in modern publications.

It is never usually done unless there is real significance apparent in the first place, or unless non-significance is the conclusion, disproving a hypothesis. A single data point, wildly separated from a cloud of points in a two-dimensional diagram, will yield a very high significance for a correlation if formal correlation tests are applied. It will be argued in Chapter 4 that any tempting inference from this should be treated with extreme caution.

The best way to look at a 'conclusion' when influenced by a single object - or a small proportion of the data-set - is to regard it perhaps as an indication that there is a hypothesis worth testing. You have reached stage one in the experimental process - *now is the time to design an experiment*. The original sample with its extreme object provides a clear pointer as to how to design the sample; or at least how to find a sample giving a reasonable chance of achieving robustness and believability.