

Re: Sampling Threshold for Distributions

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2007-04/msg00041.html>

- *From:* "Ray Koopman" <koopman@xxxxxx>
 - *Date:* 2 Apr 2007 00:44:32 -0700
-

Hosley wrote:

I am charting distributions of one variable across another, where the dependent variable has been averaged at discrete independent values. There are fewer samples at larger numbers, and at certain higher values there may only be one or two samples that went into the averaged values. These extreme values are not very representative of their population since they have low sample size, so I would like to pre-determine a threshold of how many (or what percentage) of samples are sufficient to be included in the analysis. Is there any commonly used rule in stats used to decide when to draw a line b/t values that are of high enough sampling power to be included in a distribution chart and those that are not?

Here is an example, just in case the above was painful and confusing (don't make yourself read this if you already understood): Say for a 100 tree branches, I have measured their length and their weight, both rounded to the nearest whole number (inches and lbs., respectively). Now I want to create a chart that shows how branch length varies with tree weight. I average the length of all of the branches of the same weight, and plot this on a chart where branch weight is my x axis, and the averaged branch length is my y axis. I would assume that such a chart would have a positive slope, but there may be certain weights, particularly those of smaller and higher values, that only 1 or 2 branches fell into. Thus the average length values at these given weights will only have 1 or 2 samples. If there is a lot of natural variability in my population then there is a decent chance that these low sample values will not be representative of the true population. Moreover, if I include them in my distribution chart w/o including sample size at each point (which may be unfeasible), then the viewer cannot tell which values are more representative than others. Therefore, I need to determine at what point I am justified in cutting off values from my population. Note that this not some scalar statistical analysis, but instead meant to provide a visual distribution of how the two variables (length and weight) correlate with respect to each other, and thus it is the overall characteristic that is important here.

Sorry for the long post. Thanks!

Re: Sampling Threshold for Distributions

Why not give an ordinary x - y scatterplot, with y -jittering as needed, with a superimposed regression line and its confidence region?

.