From Vincent Lemaire, 2013/01/11

Thank you for the paper about histograms.

I liked your discussion about minimizing the "loss of information" in histograms by the choice of c_1 (the center of the first bin).

However, in your discussion, you assume a priori knowledge of l the bin width, or K, , the number of bins.

But I think a "good" value of l is crucial to achieve a good representation of the random variable X using the histogram tool. I agree that first, one has to explain what one means by "good representation"...

Did you think about if there would be a way to compute some sort of optimal bin width that would best represent the random variable X with the histogram?

I think it's an interesting question, and it does not seem trivial...

I found some information about this question here:
http://en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width
and here:
http://toyoizumilab.brain.riken.jp/hideaki/res/histogram.html

SCM's answer :

In our opinion, the width of the bin, denoted by l (in fact, this is the half of the width), is linked only with the precision we expect, and not with the size of the sample. When we put some numbers in the same bin, we consider them as identical, we do not want to differentiate between them anymore. So, there is a loss of information, when performing this operation. The width of the bin indicates the value of the loss we accept. We consider it as independent from the size of the sample; in other words, we accept the fact that, if the sample is small, many bins will be empty. This is what happens for instance with extreme phenomena (temperatures), where the data are not numerous.

Of course, anyone may decide "since we have very few data, we will require only a very low precision", and therefore increase the size of the bins. But, once again, these two preoccupations : precision and amount of data, are conceptually distinct.