

Re: estimate the cdf 95% with a confidence interval of a 95%

Source: <http://sci.tech-archive.net/Archive/sci.math/2006-09/msg05404.html>

- *From:* mayost@xxxxxxxx (Daniel Mayost)
 - *Date:* 26 Sep 2006 18:09:50 -0400
-

What BinomF (MN, F(x), 0.95 MN) represents is the probability that of your MN samples, 95% of them *or less* will be less than x i.e. BinomF is the distribution function for a binomial distribution, not the density function. Thus, 1 - BinomF represents the probability that 95% *or more* of the sample will be less than x. But saying that 95% or more of the sample is less than x is equivalent to saying that the 95th percentile of the MN observations is less than x.

—
Daniel Mayost

In article <1159304444.093308.194460@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>, Dani Camps <danicamps81@xxxxxxxx> wrote:

Hi Daniel,

First of all thank you for reply.
I spend some thinking on your argument but I do not completely understand, I probably have to refresh my old maths :)

I think I understand what you mean by:

BinomF (MN, F(x), 0.95 MN)

This represents the probability that the 95%, and only the 95%, of my samples have a delay below x, so this would be the probability that $X=x$ (Prob{X=x}).

Then you say that:

$$\text{Prob}\{X < x\} = 1 - \text{Prob}\{X = x\}$$

I do not understand this relation, should not it be:

$$\text{Prob}\{X < x\} = \sum_{k=0..x} \text{Prob}\{X=k\} ?$$

Re: estimate the cdf 95% with a confidence interval of a 95%

I think I need first to understand this to be able to continue with the rest.

Thanks!

dani

Daniel Mayost wrote:

Here is an intuitive way to come up with a confidence interval around the empirical percentile score; I haven't tested this rigorously. I'm assuming here that each observation from each server is independent of all the other observations, and that all the observations are identically distributed. I'm also assuming that you don't know the distribution of the observations, because otherwise you could just fit the parameters of the distribution from your data and then compute the percentile from the fitted distribution.

If F is the unknown distribution function of the observed delays, and X is a random variable representing the observed 95th percentile (you should pool all your MN observations together to get this statistic), then the distribution function of X is given by:

$$\text{Prob}(X < x) = 1 - \text{BinomF}(MN, F(x), 0.95 MN)$$

where $\text{BinomF}(MN, F(x), 0.95 MN)$ is the distribution function for a binomial distribution, representing the probability that for MN trials with probability of success $F(x)$, 95% or less of them will be successful. For large M and N you can approximate this with a beta distribution.

Anyhow, if d_0 is the true percentile value, then to get a 95% confidence interval around the observed percentile we need to find a c such that:

$$\text{Prob}(d_0 - c < X < d_0 + c) = 95\%$$

This condition is equivalent to:

$$\text{BinomF}(MN, F(d_0 - c), 0.95 MN) - \text{BinomF}(MN, F(d_0 + c), 0.95 MN) = 95\%$$

Since this equation is not linear, you will have to find c by trial and error. For d_0 you can use the observed percentile, while to evaluate F you can use the empirical distribution from the observed data. The more observations you have around the observed percentile, the tighter your confidence interval will be.

Again, I haven't proved that this method will converge to the true answer as MN goes to infinity, but it does look like an obvious starting point.

Re: estimate the cdf 95% with a confidence interval of a 95%

—
Daniel Mayost

In article <1159217795.167906.267060@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>, Dani Camps <danicamps81@xxxxxxxx> wrote:

Dear all,

I do networking simulations and I have the following problem. Imagine that I set up a certain scenario, for instance N stations in my wireless network, and I want to obtain an estimation of which is the delay that my stations are suffering when they sent a packet to a certain node that acts as a gateway.

What I do is that every station writes in a file the delay that it took for it to send every packet. Then at the end of a simulation I have N files, each file containing the delays of all the packets that every station has sent. If I run the simulation M times, with M different random seeds, then I have $N * M$ files each one containing the delays that one station suffered for each one of its packets.

Now assume that the delay that one station can suffer when sending a packet in my network is a random process, what I want to do is to compute an estimate of the 95% cdf of such the delay, i.e the delay d_0 such that a packet sent by the station will suffer a delay lower than d_0 with probability 0.95.

What I do to compute this delay is to put all the samples from all the $N * M$ files together, as a samples of the same random process or realizations of the random variable delay, and then I compute an empirical cdf out of all these samples, and from that empirical cdf I

Re: estimate the cdf 95% with a confidence interval of a 95%

compute the 95% cdf. To compute the cdf I use the 'ecdf' function in matlab. This in my understanding is the best estimation of the 95% cdf that I can obtain given all my samples.

Now my problem is that I want to give this 95% cdf delay with a certain level of confidence. The point is that using the method described above I can not give any confidence because I need samples to compute an interval, and according to what I described previously I only have one sample. The other alternative is to compute an estimation of the 95% cdf for each one of the simulation runs, having thus M samples of the 95% cdf, and then I can compute the average of my M samples, having the "expected" 95% cdf, and compute the confidence interval around the expected 95% cdf using a gaussian approximation, or the T-student method.

Which is according to you the best procedure to follow? Would there be another way to give the estimation of the 95% cdf with a certain level of confidence ?

Best Regards

Dani