

Re: Confidence interval on mean for a set of numbers

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2004-09/0318.html>

From: George Kahrmanis (*anakreon_at_hol.gr*)

Date: 09/18/04

Date: 18 Sep 2004 04:03:05 -0700

Peter Michaux (*petermichaux@yahoo.com*) wrote on 2004-09-17:

*>I have a list of numbers all of which are nonnegative. I have no idea
>from what distribution this list of numbers comes. I also have no time
>to check which distribution because I have many of these lists.
>
>I can calculate and plot the mean. I also want to have a confidence
>interval on the mean.*

and on 2004-09-17:

*>[...] I realize that I don't really want just a
>confidence interval on the mean. [...] What I also want to do is
>characterize the width of the underlying distribution.*

I agree with "andre" (*andrevh@sci.kun.nl*), that a quick (and not too-dirty) way is to assume normality in the distribution of the mean (provided that the size of the sample is large enough). Here I post my suggestion on how we can "do it right". Like Peter Michaux prefers, we avoid estimating the PDF of each outcome; we work instead with the corresponding predictive CDF ("cumulative distribution function").

E.g., say that we have recorded one thousand outcomes, lumping all old samples together. Let us rank these outcomes. Then the probability of the next outcome being, say, between the outcomes ranked #393 and #394 is $1/1001$, provided that we admit no prior information of any kind regarding the true underlying distribution (in other words, we apply only "low structure assumptions"). This "posterior distribution of percentiles" is still debatable wrt the foundation. (Imho, it *is* possible to supply a classical proof (i.e. based on properties of conventional, orthodox probability) but no such proof is in print yet, afaik.)

For references, if you need them, see Section 3a of my news:
<3ce8f26b.0409070825.7c799b23@posting.google.com>
"prediction versus parameter estimation (was: literature: ...)"

7 Sep 2004 09:25:57 -0700).

If we accept (OK, just provisionally) the above posterior distribution of percentiles, we obtain an inexact CDF for the next outcome. E.g., the CDF at $x=393$ is $393/1001$, the CDF at $x=394$ is $394/1001$, and the CDF at any intermediate point is undefined as an exact value but defined as an "interval value": $[393/1001, 394/1001]$.

Now it is elementary to derive a CDF for the mean of the next sample of size n , given the CDF of single iid outcomes. The upshot is, in this problem we obtain an interval-valued CDF for the mean.

(If we wanted the CDF of the median rather than the CDF of the mean, it would be an easier calculation.)

Moreover, if we have recorded a large number of outcomes, this "second-order uncertainty" will be negligible.

Not only this method is trivial to implement, but also it is *the* correct way, imo. (Of course I will welcome any comment or disagreement.)

Good Day, ~ George Kahrmanis