Two Comments on LEE with Two Data Sets

Bob Cousins

Univ. of California, Los Angeles

LL Statistics Workshop

13 February 2013

Mythology of Two Data Sets

Very **common myth re bump-hunting:**

Look for bump in one data set (or one experiment), and identify candidate bump and its location in spectrum. Then in the next data set (or other experiment), look in same location. ***Then no LEE correction needed.***

This is "bad" on several levels, and the Gross-Vittels paper gives another reason why.

This is presumably known to experts, but I think it is useful to advertize, given the persistence of the myth.

1st Comment (pre-Gross-Vitells):

If you have data-taking periods under identical conditions, with identical cuts, it is a property of the Poisson probability that:
The Z-value for the whole run is independent of the distribution of events within the run.

The distribution of signal events as a function of integrated lumi is a check on the model (systematics): goodness of fit to hypothesis that data-taking conditions were identical, etc.

Case 1: 2σ effect in first half of run, 4σ in second half.
Case 2: 4σ effect in first half of run, 2σ in second half.

Inference should be the same in both cases.
This is obvious to this group, but for some reason, the myth seems never to die.

2nd Comment:

Consider LEE applied to a single plot.
As I first learned from Gross-Vitells paper v1,
the LEE factor depends on the significance Z (!)

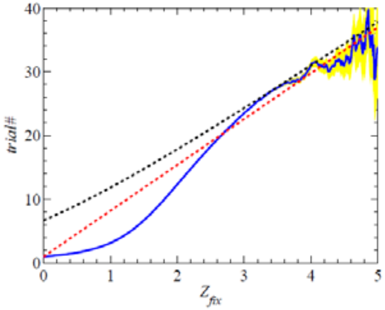

Fig. 3. The trial factor estimated from toy Monte Carlo simulations (solid line), with the upper bound of eq.(3) (dotted black line) and the asymptotic approximation of eq.(12) (dotted red line). The yellow band represents the statistical uncertainty due to the limited sample size.

Trial factors for the look elsewhere effect in high energy physics

Eilam Gross and Ofer Vitells

Weizmann Institute of Science, Rehovot 76100, Israel

1) **Since one does not know the "final" Z before obtaining all the data, it is impossible to correct correctly (or remove) the LEE after obtaining half the data.**
2) **A key part of the GV result is the "local" LEE from finding Max-Liklihood location *within* the bump. This is not accounted for in definition of "same" used in myth.**

Please Help Kill the Myth

In my opinion, it is such a bad “approximation” so as not to be useful even for “hallway statistics”.

But... There *is* truth to the idea that in the long run, Z should get larger and larger as sample size grows. See tomorrow’s talk.