DRAFT 0.1

Glen Cowan
22 December 2017

Thoughts on Limits

This note is part of an ongoing conversation about limits. The thoughts here overlap with Refs. [1, 2] and the talks by myself and Bob Cousins from the Friday session of the 2011 Terascale School at DESY [3].

Consider a parameter μ that is proportional to the rate of a signal process. A test of μ is carried out by specifying a critical region w_μ in data space with a summed probability, under assumption of μ, is less than or equal to a given size α (e.g., $\alpha = 0.05$). If the data are observed in w_μ, the value of μ is rejected. This procedure can be carried out for all μ and the set of μ values not rejected constitutes a confidence interval.

Equivalently, one can define a p-value p_μ for all μ, and reject μ if $p_\mu \leq \alpha$. In this way, the confidence interval consists of values with $p_\mu > \alpha$, the rejected values have $p_\mu \leq \alpha$, and the boundary is determined from the condition $p_\mu = \alpha$.

One normally chooses the critical region to maximize the power of the test with respect to some alternative deemed to be most relevant. Equivalently, the alternative determines what region of data space is deemed to constitute greater incompatibility for purposes of computing the p-value (saying an outcome less compatible with μ means it is more compatible with the alternative). In the usual Neyman-Pearson framework of statistical tests, the relevance of the alternatives is not (or at least does not have to be) quantified. It is impossible to construct a test that has maximum power with respect to all alternatives and thus the analyst must decide how to optimise the choice of critical region. This subjective element of a frequentist test can be contrasted with the Bayesian requirement that one specify prior probabilities for all hypotheses.

For purposes of discovering the signal process one tests the background-only hypothesis ($\mu = 0$). Suppose the relevant alternative is a positive signal rate, i.e., an excess of events over the expected background. Therefore in a counting experiment one would define the p-value of $\mu = 0$ to be

$$p_0 = P(n \geq n_{\rm obs} | \mu = 0) \ . \tag{1}$$

Note that in principle the alternative to the null hypothesis could imply an increase or decrease in the expected number of events (e.g., neutrino oscillations). In such a case, both $\mu > 0$ and $\mu < 0$ are relevant alternatives and one could then choose the critical region to have power with respect to both possibilities.

Suppose that the existence of the signal process has not yet be established. One may then be interested in knowing whether a large value of μ is disfavoured by the data on the grounds that its predicted rate is too high relative to what is observed. In such a situation, the relevant alternative to a hypothesized value μ is $\mu = 0$ (or in any case, a lower value of μ). Therefore the critical region is chosen to be characteristic of low μ, which in a Poisson counting experiment corresponds to low numbers of events. In this case the p-value for μ would be

$$p_\mu = P(n \leq n_{\rm obs} | \mu) \ . \tag{2}$$

The upper limit is thus useful for deciding between a hypothesized μ and the alternative of $\mu = 0$. It answers the question of how large the considered parameter value can be before it becomes incompatible with the data, where the notion of "compatibility" is defined with respect to the alternative deemed here to be most relevant, namely, $\mu = 0$.

Suppose the existence of the process is already established, i.e., the value $\mu = 0$ has been strongly disfavoured by the data and thus is no longer a relevant alternative. Then in testing a certain value of μ, one would like to have power with respect to parameter values both higher and lower. One may the choose to define the p-value using the statistic $t_\mu = -2 \ln \lambda(\mu)$, where $\lambda(\mu)$ is the likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})} \ , \tag{3}$$

and $\hat{\mu}$ is the maximum-likelihood estimator. The p-value of a hypothesized μ is thus

$$p_\mu = P(t_\mu \geq t_{\mu,{\rm obs}} | \mu) \ . \tag{4}$$

Here the critical region where $p_\mu \leq \alpha$ will contain data outcomes characteristic both of low and high μ, which is to say one has a two-sided test. The part of the critical region characteristic of the $\mu = 0$ alternative is therefore smaller than it was in the one-sided case. Thus the power of the test of μ with respect to $\mu = 0$ (i.e., low μ) is not as high as the corresponding power from a one-sided test used to establish an upper limit.

That is, for purposes of establishing an upper limit, a value of μ is deemed to be disfavoured if it predicts a rate that is too high relative to what is observed. But in the two-sided test a value of μ maybe rejected if the observed rate is too high or too low, and it thus answers a different question. Whether the one- or two-sided alternative is deemed more important is a matter of some debate; one could argue that in the search phase the focus is on knowing which parameter values are excluded because their predicted rates are too high. That is, in the absence of a discovery one primarily wants to know how large the undetected signal still might be. Or at least one would like to have a single number (the upper limit), that specifically addresses that question.

A well-known problem with frequentist upper limits is that one may exclude parameter values to which one has very little sensitivity. The problem is avoided in the Bayesian and CLs methods, and is reduced but not eliminated by unified (Feldman-Cousins) limits [4].

Another problem pointed out by Feldman and Cousins is flip-flopping. One can show that the coverage probability of an interval is unequal to the nominal value if one decides, using the observed data, whether to carry out a one- or two-sided test. Doing so would in fact violate the procedure outlined above, in which the critical region is fixed before seeing the data.

In order to avoid the coverage problem associated with flip-flopping, it is only necessary to specify the critical region before carrying out the test. One can do this, in effect, by stating what tests will be carried out prior to looking at the data. In practice one would report an upper limit as long as one is in the search phase for the signal process. The paper that establishes discovery of the signal would thus also quote an upper limit (as was done, e.g., for the discovery of the Higgs boson).

References

[1] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1554.

[2] Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells, *Power-Constrained Limits*, arXiv:1105.3166 [physics.data-an] arXiv:1105.3166 [physics.data-an] (2011).

[3] School on data combination and limit setting, DESY, 4-7 October 2011, `https://indico.desy.de/indico/event/4489/`.

[4] Robert D. Cousins and Gary J. Feldman, Phys. Rev. D 57, 3873 (1998).