

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2006-09/msg00184.html>

- *From:* "Kevin E. Thorpe" <kevin.thorpe@xxxxxxxxxxx>
 - *Date:* 4 Sep 2006 12:35:45 -0700
-

Reef Fish wrote:

Kevin E. Thorpe wrote:

<SNIP>

First, I suspect that modern statistical education is no longer purely NP or purely Fisherian. I realize my education was a mixture with no reference (that I understood at the time anyway) to the fact that there were two philosophies at odds here. On the other hand, I suspect that Reef Fish's education was in a time when there were very firmly established boundaries in statistical inference. I have only recently started to think through these things a bit more deeply and am beginning to see a richer landscape. What follows is my simplified overview of the two methods.

In the Fisher approach, called "pure significance tests" in Cox and Hinkley, you have the following set-up. You define an hypothesis H (eg. $\mu = 0$ or $\pi = 0.5$). You define an "appropriate" test statistic D .

You gather data and compute D_{obs} . You compute a p-value or significance level as $P(D_{obs} \geq D|H)$. I will mention that how you determine what is an appropriate test statistic was one of the points of contention between Fisher and Neyman.

Note my correction of the above probability statement in another post -- $P(D \geq D_{obs}|H)$ is what I should have written.

In the classical NP approach you have the familiar H_0 and H_1 , a critical value c is chosen such that the test has level α . The test statistic T is calculated from the observed data. Then H_0 is rejected if $T > c$ (or $< -c$ or $|T| > c$ depending on H_1). Also, the significance level is again calculated with strict inequality.

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

For the continuous case, this difference doesn't matter since $P(T=c|H_0) = 0$. For the discrete case it does. In one of the books on my shelf at home it says (slightly paraphrased) that in the discrete case $P(T=c|H_0)$ may be

0 and to make the test have level alpha, we may need

to use randomization (ie. a randomization test).

This may be precisely the loophole for some N-P situations such as the $c = 0$ case in Bruce Weaver's example.

It is probably also one of the things that prompted some to consider mid-p.

I know of no one who ever said that the N-P approach to statistical inference is perfect, or even near perfect. Instead, there are cases of glaring holes, such as the Hogg and Craig example of a 96% confidence interval (based on a random interval), that may turn out to be a 100% confidence interval AFTER the data is observed!

Now, one thing I haven't mentioned yet is that the test indicated by the NP Lemma is the likelihood ratio test. If one uses that and the theorem that tells us the LRT is approximately chi-square, the problem again goes away for the discrete case since the reference distribution for the test is continuous.

But here's the nit on the Bruce Weaver example relative to your mention of LRT. The LRT is the same as the LRT used for Bayesian inference of the parameter p , which is indeed continuous because of the Bayesian approach which is distinctly different from both the N-P and Fisherian approach. The LRT for the binomial p , combined with any continuous prior for p , or the versatile conjugate prior of the Beta family for p will yield a posterior distribution which is continuous in p for the Bayesian inference, whether in terms of credible interval (as opposed to C.I.) for p , or in setting up the likelihood ratio of the MODELS of p , based on the posterior distribution.

NONE of that artificiality in the Neyman-Pearson nor the Fisherian approach. You simply CANNOT patch up an imperfect system with inherent DEFECTS (in both the N-P and Fisherian) to make it a

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

perfectly workable consistent system.

THAT's the crux of the matter.

But given the H-P approach in the use of p-values for the DISCRETE case of Bruce Weaver, fixing n , so that the only POSSIBLE observable Test Statistic values are $0, \dots, 13$, which corresponds to the only POSSIBLE p under consideration given n to be $0/13, 1/13, \dots, 12/13, 13/13$.

No other value of p can be OBSERVED. The LRT for that problem is anything but a continuous distribution in p . I don't know what the LRT is for the two discrete alternatives that differ by $c=2$, which is no longer a point of measure zero. That's why the LRT scenario breaks down, and one resorts to the perfectly unambiguous use of the p -value, of "more extreme" <unambiguously NOT "2"> for the N-P definition and usage, and leave the $X = 0$ and $X = 13$ cases for patch up by randomization.

I guess I was thinking that although the LRT would be discrete, the reference distribution (chi-square) is not and so the "equality" problem might be less.

That seems to be the only sensible approach to THAT problem.

The old saw, "don't fix it if it ain't broke" clearly apply to the EXCLUSION of $X = 2$ from the p -value definition unambiguously defined under the N-P approach, instead of fixing it (when it ain't broke) as some textbook writers tried to do. The Agresti approach is completely OFF THE WALL -- it has NO justifiable support of his $p+$ and $p-$ and other witchcraft brew cooked by him, unjustified by all THREE approaches -- Bayesian, N-P, or Fisher.

It would appear that most statisticians I have encountered would compute a p -value, particularly in the discrete case, in the manner of Fisher even when using NP hypothesis testing. This will naturally be a conservative test in terms of α whereas excluding the observed value will increase the type I error probability since, in general you don't have an exactly level α test for the discrete case. In my opinion, the "best" practice may well depend on the consequences of a type I error. In my work with primarily clinical trials, you REALLY do not want to conclude one treatment is better when it is not, so I would err on the conservative side.

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

Re: Two nit-picks re definition of p-value (Was: goodness of fit ?)

Kevin E. Thorpe
Assistant Professor, Department of Public Health Sciences
Faculty of Medicine, University of Toronto

<SNIP REST>

.