EVERYTHING WRONG WITH P-VALUES UNDER ONE ROOF

WILLIAM M. BRIGGS

Abstract. Use of p-values should be abandoned forthwith, and replaced with probability methods based on observables.

1. They are based on a fallacious argument

Repeated in introductory texts, and began by Fisher himself, are words very like these (these were adapted from Fisher, R. 1970. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, fourteenth edition):

> Belief in a null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: Either the null hypothesis is false, or the p-value has attained by chance an exceptionally low value.

Fisher's choice of words was poor. This is evidently not a logical disjunction, but can be made into one with slight surgery:

> Either the null hypothesis is false and we see a small p-value, or the null hypothesis is true and we see a small p-value.

Stated another way, "Either the null hypothesis is true or it is false, and we see a small p-value." Of course, the first clause of this proposition, "Either the null hypothesis is true or it is false", is a tautology, a necessary truth, which transforms the proposition to "TRUE and we see a small p-value." Or, in the end, Fisher's dictum boils down to:

> We see a small p-value.

In other words, a small p-value has no bearing on any hypothesis (unrelated to the p-value itself, of course). Making a decision *because* the p-value takes any particular value is thus always fallacious. The decision may be *serendipitously* correct, as indeed any decision based

Date: October 6, 2013.

on any criterion *might* be, and as it often likely correct because experimenters are good at controlling their experiments, but it is still reached by a fallacy.

2. People believe them

Whenever the p-value is less than the magic number, people believe or "act like" the alternate hypothesis is true, or very likely true. (The alternate hypothesis is the contradiction of the null hypothesis.) We have just seen this is fallacious. Compounding the error, the smaller the p-value is, the more likely people believe the alternate hypothesis true.

This is also despite the strict injunction in frequentist theory that *no* probability may be assigned to the truth of the alternate hypothesis. (Since the null is the contradiction of the alternate, putting a probability on the truth of the alternate also puts a probability on the truth of the null, which is also thus forbidden.) Repeat: the p-value is silent as the tomb on the probability the alternate hypothesis is true. Yet nobody remembers this, and all violate the injunction in practice.

3. People don't believe them

Whenever the p-value is less than the magic number, people are supposed to "reject" the null hypothesis forevermore. They do not. They argue for further testing, additional evidence; they say the result from just one sample is only a guide; etc., etc. This behavior tacitly puts a (non-numerical) probability on the alternate hypothesis, which is forbidden.

It is not the non-numerical bit that makes it forbidden, but the act of assigning any probability, numerical or not. The rejection is said to have a probability being in error, but this is *only* for samples in general in "the long run", and *never* for the sample at hand. If it were for the sample at hand, the p-value would be putting a probability on the truth of the alternate hypothesis, which is forbidden.

4. They are not unique: 1

Test statistics, which are formed in the first step of the p-value hunt, are arbitrary, subject to whim, experience, culture. There is no unique or correct test statistic for any given set of data and model. Each test statistic will give a different p-value, none of which are preferred (except by pointing to evidence outside the experiment). Therefore, each of the p-values are "correct." This is perfectly in line with the p-value having *nothing* to say about the alternate hypothesis, but it encourages bad and sloppy behavior on the part of p-value purveyors as they seek to find that which is smallest.

5. They are not unique: 2

The probability model representing the data at hand is usually *ad hoc*; other models are possible. Each model gives different p-values for the same (or rather equivalent) null hypothesis. Just as with test statistics, each of these p-values are "correct," etc.

6. They can always be found

Increasing the sample size drives p-values lower. This is so well known in medicine that people quote the difference between "clinical" versus "statistical" significance. Strangely, this line is always applied to the other fellow's results, never one's own.

7. They encourage magical thinking

Few remember its definition, which is this: Given the model used and the test statistic dependent on that model and given the data seen and assuming the null hypothesis (tied to a parameter) is *true*, the p-value is the probability of seeing a test statistic larger (in absolute value) than the one actually seen *if* the experiment which generated the data were run an indefinite number of future times and where the milieu of the experiment is precisely the same except where it is "randomly" different. The p-value says *nothing* about the experiment at hand, by design.

Since that is a mouthful, all that is recalled is that if the p-value is less than the magic number, there is success, else failure. P-values work as charms do. "Seek and ye shall find a small p-value" is the aphorism on the lips of every researcher who delves into his data for the umpteenth time looking for that which will excite him. Since wee p-values are so easy to generate, his search will almost always be rewarded.

8. They focus attention on the unobservable

Parameters–the creatures which live inside probability models but which cannot be seen, touched, or tasted—are the bane of statistics. Inordinate attention is given them. People wrongly assume that the null hypotheses ascribed to parameters map precisely to hypotheses about observables. P-values are used to "fail to reject" hypotheses which nobody believes true; i.e. the parameter in a regression is precisely, exactly, to infinite decimal places zero. Confidence in real-world observables must always be *necessary lower* than in confidence in parameters. Null hypotheses are never “accepted”, incidentally, because that would violate Fisher’s (and Popper’s) falsificationist philosophy.

9. They base decisions on what did not occur

They calculate the probability of what did not happen on the assumption that what didn’t happen should be rare. As Jefferys famously said: “What the use of P[-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.”

10. Fans of p-values are strongly tempted to this fallacy

If a man shows that a certain position you cherish is absurd or fallacious, you multiply your error by saying, “Sez you! The position *you* hold has errors, too. That’s why I’m going to still use p-values. Ha!” Regardless whether the position the man holds is keen or dull, you have not saved yourself from ignominy. Whether you adopt logical probability or Bayesianism or something else, you must still abandon p-values.

11. Confidence intervals

No, confidence intervals are not better. That for another day.

wmbriggs.com, matt@wmbriggs.com, 917-392-0691.