Thailand Statistician
July 2016; 14(2): i-iv
http://statassoc.or.th
Editorial

Why P-values are Banned?

We elaborate and offer comments on the recent ban on the use of p-values in hypothesis testing.

1. The concept of p-values (invented by R. Fisher, 1925) is familiar to those who use statistical methodology to conduct empirical research via testing of hypotheses, especially in social sciences such as psychology. Specifically, p-values are used as empirical evidence to reject or accept hypotheses under consideration. In other words, scientific conclusions are based only upon the values of the p-values.

Despite severe criticisms of the possible misuse of p-values in recent past, this "golden standard" continues to be used. Finally, in 2015, the editors (Trafimow, D. and Marks, M., 2015) of the Journal *Basic and Applied Social Psychology* (BASP) issued an "official" ban on using p-values in null hypothesis significant testing (NHST) because the testing procedure based on them is invalid. And, recently, the American Statistical Association reacted to it by releasing a "statement" acknowledging the flaw of statistical reasoning using p-values (ASA News, March 7, 2016).

Although the ban is intended for BASP, it could spread in view of the ASA's statement to the whole statistical community, and in view of the real problem with using p-values for decision-making as we are going to elaborate shortly. As such, the consequences of this ban will obviously affect statistical teaching and research, and hence this "crisis" should be emphasized to all.

2. Just like his Maximum Likelihood Estimation (MLE) method, Fisher proposed the use of p-values as a quantitative measure of evidence obtained from the data concerning a null hypothesis. The MLE is a principle, not a rule. It only suggests a plausible way to obtain estimators. There is another step to achieve the estimation goal, namely establishing desirable properties of a "good" estimator, such as consistency, efficiency, limiting distribution. And this is attainable for regular models. It seems that the misuses or misconceptions of p-values came from the misunderstanding that there is only one step to perform in testing.

To be specific, the p-value of a test statistic T_n is, say, $P(T_n \geq t_n|H_o)$, the probability under the null (i.e., if H_o were true) that the statistic T_n will exceed the (already) observed value t_n. Clearly, this is an useful information coming from the data, but the question is "can we use only this information to decide whether H_o is true or not?". Well, it depends not only on how we interpret the meaning of a p-value, but also on how we **reason** (i.e., using which approximate **logic**) with it.

For example, what is the meaning of a p-value of 2%? It could be the "likelihood" that the null is true? But, definitely, *it does not mean that there is only a* 2% *chance that the null is true!* Why? Well, we cannot have both: in calculating the p-value, we assume the null is true, thus this p-value cannot, at the same time, be the probability that the null is true!

Remark. A probability value for the null to be true, given the data, seems enough for rejecting or accepting the null. With the cost of using (subjective) priors, a Bayesian approach could provide such a probability value.

But why p-values have been used to carry out tests so far? Well, as Cohen (1994) correctly pointed out, the answer is that we have used an invalid logic for reasoning with p-values.

In binary logic, the way to deny something is the modus tollens rule: $A \Longrightarrow$(implies) B, and B^c (not B), then A^c , as in

If H_o is true, then this data cannot occur
This data has occurred
Therefore, H_o is false

But if we use an **approximate reasoning logic**, say, a probability logic, then the above modus tollens rule is no longer valid:

If H_o is true, then this data is highly unlikely
This data has occurred
Therefore, H_o is highly unlikely

Why? Without going into probability logic, just consider this:

If she is a person, then probably she is not rich
She is rich
Therefore, she is probably not a person

With a wrong interpretation of p-values, we derived the following decision rules

(i) Reject the null if the p-value is small,

(ii) Otherwise, do not reject the null.

To apply these rules, statisticians need to know "how small is small?". Without using linguistics or fuzzy set theory, we just need a threshold. Fisher suggested 0.05 as the cutoff point. It is not really this suggested threshold of 0.05 which causes the problem, because we can simply report our p-values and leave the thresholds to decision-makers to assign subjectively using their own perception.

As reported by Goodman (2008), the misconception of p-values leads to **invalid rules** such as

a) If the p-value is 0.05, then the null has 5% chance of being true,

b) If the p-value is greater than 5%, then there is no difference between the treatments,

c) A statistically significance is clinically important: no! the p-value carries no information about the effect size,

d) A scientific conclusion should be based on whether or not the p-value is significant (i.e., less than 0.05).

As a final note, do not forget that p-values depend on sample sizes!

3. The BASP's ban on the use of p-values: "From now on, BASP is banning the NHSTP (Null Hypothesis Significance Testing Procedure)" seems justified, as it is clear that it is a ban on its use in NHSTP.

The ASA's statement listed six principles regarding the use of P-values in testing.

(1) P-value can indicate how incompatible the data are with a specified statistical model,

(2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone

(3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold

(4) Proper inference requires full reporting and transparency

(5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result

(6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

and adding that "In light of misuses and misconceptions concerning p-values, statisticians should supplement or even replace p-values with other approaches".

However, few things need to be spelled out from the "Editorial BASP's ban":

(i) There are no suggestions to researchers in Psychology (or else) what to do when their bread-and- butter tool is prohibited. Is Bayesian statistics the way out?

(ii) Is banning a statistical practice the best way to prevent bad research results? Noting that "misuse" of p-values (in testing) does not mean necessarily that the concept of p-values itself is wrong,

(iii) In the answer to Question 2 in BASP's Editorial, the phrase "Therefore, **confidence intervals** also are banned from BASP" seems strange, and not justified! This should be ignored. In fact, the "opposite" should be considered: it has been advocated that an **alternative** to p-values is confidence intervals got testing. Moreover, confidence intervals can estimate the effect size and provide information on precision (through their widths).

(iv) It was its misuse in testing that the concept of p-value is regarded as undesirable in the repertoire of statistical tools. By itself, the concept provides some useful information which could be used, say, in conjunction with some further steps for valid inference in an appropriate decision framework.

(v) For a compromise between frequentists and Bayesians, and with confidence intervals as a possible alternative to p-values in testing, an newly statistical framework, known as *Inferential Models* (Martin and Liu, 2016) providing a prior-free posterior analysis, seems attractive for research, in which, possibly p-values have some role to play.

In conclusion, clearly, the above "news" will affect (or "reform") statistics at least at some levels of teaching and applied research works.

How do we tell our students? Should we skip sections in text books having p-values and how to use them?

A positive side is this. This is a good example to remind our students and applied workers that statistical methodologies and tools are really useful and valid only when using them **with critical thinking**.

References

American Statistical Association releases statement on statistical significance and p-value. ASA News. March 7, 2016.

Cohen J. The earth is round. Am. Pshychol. 1994; 49: 997-1003.

Fisher R. A. Statistical Methods for Research Workers. Oliver and Boyd (Google); 1925.

Goodman S. A dirty dozen: Twelve p-value misconceptions. Semin. Hematol. 2008; 45: 135-140.

Martin R., Liu C. Inferential. Models: Reasoning with Uncertainty. Chapman and Hall: CRC Press; 2016.

Nuzzo R. Statistical errors. Nature. 2014; 506: 150-152.

Trafimow D., Marks M. Editorial. Basic Appl. Soc. Psych. 2015; 37: 1-2.

Wasserstein R. L., Lazar N. A., The ASA's statement on p-value: Context, process, and purpose. Am. Stat. 2016.

Hung T. Nguyen
Associate Editor