HOSTED BY

ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Journal of Sport and Health Science 5 (2016) 77–79

www.jshs.org.cn

Opinion

$p < 0.05, < 0.01, < 0.001, < 0.0001, < 0.00001, < 0.000001,$ or < 0.0000001 . . .

Weimo Zhu

Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 6 January 2016; accepted 8 January 2016
Available online 21 January 2016

These days when I look at scientific research papers or review manuscripts, there seems to be almost a competition to have a smaller p value as a means to present more significant findings. For example, a quick Internet search using "$p < 0.0000001$" turned up many papers even reporting their p values at this level. Can and should a smaller p value play such a role? In my opinion, it cannot. The current statistical software making possible p value-centered statistical reporting, I believe, is leading scientific inquiry into a quagmire and dead end.

To fully understand why the p value-centered inquiry is the wrong approach, let's firstly understand what p value and hypothesis testing (HT) are and examine how statistical hypothesis testing (SHT) was run prior to the computer era. While p value and HT are both now used under the umbrella of SHT, they had different roots. The p value and its application in scientific inquiry is credited to the English statistician Sir Ronald Aylmer Fisher[1] in 1925. In Fisher's inquiry system, a test statistic is converted to a probability, namely the p value, using the probability distribution of the test statistic under the null hypothesis and the p value was used solely as an aid, after data collection, to assess if the observed statistic is a simply random event or indeed belongs to a unique phenomenon fitting the researchers' scientific hypothesis.[2] Furthermore, 0.05 or 0.01 are not the only p value cutoff scores for the decision. Thus, Fisher's p value inquiry system belongs to *a posteriori* decision system, which also features, "flexibility, better suited for *ad-hoc* research projects, sample-based inferential, no power analysis and no alternative hypothesis" (p. 4).[3]

HT, on the other hand, was credited to the Polish mathematician Jerzy Neyman and American statistician Egon Pearson[4] in 1933, who sought to improve Fisher's method by proposing a system to apply repetition of experiments. Neyman and Pearson believed that a null hypothesis should not be considered unless one possible alternative was conceivable. In contrast to Fisher's system, Type I error or the error the researchers want to minimize, the corresponding critical region and value of a test must be set up first in the Neyman–Pearson's system, which, therefore, belongs to *a priori* decision system. In addition, the Neyman–Pearson's system is "more powerful, better suited for repeated sampling projects, deductive, less flexible than Fisher's system and defaults easily to the Fisher's system" (p. 8).[3]

The current commonly used SHT is mainly derived from the Neyman–Pearson's system. With the p value conveniently provided by modern statistical software, researchers have started to mix the two systems together and with the result that SHT has started to become a means to foster pseudoscience.[3]

A quick review of the SHT practice prior to the computer era may help better explain the above points. A typical SHT can be considered as a decision system by including the following steps:

1. Determine the null hypothesis (H_0) and alternative hypothesis (H_1):
   H_0 (The hypothesis is tentatively held to be true),
   H_1 (Often = research hypothesis);
2. Set Type I error (α or critical value), which represents the error rate when an H_0 was rejected when it is true (should not be rejected); in practice, α is often replaced by a p value, which forms a specific boundary between rejecting or not rejecting the null hypothesis (note: in contrast to Type I error, a Type II error means an H_0 was not rejected when it is false);
3. Select a statistical test and set the decision rule, which is the statement that designates the statistical conditions necessary for rejecting the null hypothesis, based on a Type I error, one or two tailed tests and sample size;
4. Compute statistic using the selected test;
5. Make a decision based on the decision rule set in Step 3.

By going through these steps, you should be able to quickly realize two things: first, SHT is similar to the US criminal court trial system, in which "innocent until proven guilty" is the guiding principle:

H_0: The accused is innocent.
H_1: The accused is guilty.

Peer review under responsibility of Shanghai University of Sport.

E-mail address: weimozhu@illinois.edu

http://dx.doi.org/10.1016/j.jshs.2016.01.019

If H_0 is rejected when it is true (i.e., Type I error happened), an innocent person may be convicted for a crime they did not commit. Therefore, Type I error in practice is often strictly controlled since the consequences of having a Type I error could be much more serious than a Type II error (failed to convict a criminal). Secondly, before the use of computer software, Type I error or p value had to be determined prior to computing statistics and there were usually only two choices, $p = 0.05$, which is commonly used in kinesiology research, or $p = 0.01$, which is commonly used in pharmaceutical research. So, SHT belongs to *a priori* decision system, i.e., a probability-based evaluation standard or the confidence has to be set up before computing a statistic and making a decision.

An example may be helpful to illustrate the above steps. Say a researcher observed a difference between males and females in body composition and wants to test her research hypothesis that females have a higher percentage of body fat. To do so, she recruited 10 adults (5 females and 5 males) and measured their fat percentage using the underwater weighing method (Table 1).

Table 1
An example of sex difference on percentage of body fat.

ID	Sex	Fat%
1	Female	17.55
2	Female	35.77
3	Female	29.55
4	Female	16.84
5	Female	20.08
6	Male	20.97
7	Male	25.59
8	Male	3.71
9	Male	5.17
10	Male	24.27

Following the SHT steps, she tested her research hypothesis:

1. Determine H_0 and H_1:
   H_0: Female fat% population mean = Male fat% population mean.
   H_1: Female fat% population mean ≠ Male fat% population mean.
2. Set Type I error
   $\alpha = 0.05$.
3. Select a test and set the decision rule
   Since there are two groups, she selected the independent t test; given $\alpha = 0.05$, two-tailed test, and $df = 5$ (n of male group) + 5 (n of female group) − 2 = 8, the critical value according to t value table is 2.306; such, the decision is set as below:
   If $-2.306 < t$ statistic observed < 2.306, do not reject H_0;
   If t statistic observed ≤ -2.306 or if t statistic observed ≥ 2.306, reject H_0.
4. Compute t statistic
   Female fat%: M = 23.958, SD = 8.330
   Male fat%: M = 15.942, SD = 10.646

$$t = (M_{\text{female}} - M_{\text{male}}) / \surd\, (s^2{}_{female} / n_{female} + s^2{}_{male} / n_{male})$$
$$= 8.02 / \surd\, 36.55 = 1.33$$

5. Make a decision
   H_0 was NOT rejected since the observed t statistic is larger than −2.306 and smaller than 2.306.

With convenient and powerful statistical software now available, an extra piece of information is generated when the statistic is computed, i.e., the exact p value along with a specific statistic condition of the sample size and the direction of the test. For the example, for the above research data, if we run the t test using a statistical software, we also get a specific p value corresponding to the t statistic of 1.33, which is $p = 0.221$. Since it was larger than $p = 0.05$, one may normally conclude that since H_0 was not rejected, there is no significant difference between males and females in fat percentage. As a result of this additional information, you can see that researchers start to report these specific p values in their research reports and omit other related important information (e.g., the statistics themselves, *df*, *etc.*), especially if they have one less than 0.05 or 0.01, which has resulted in the "p value competition".

What is the issue with this approach if the p value itself could reach a similar conclusion without other information (e.g., the statistics themselves, *df*, *etc.*)? Unfortunately, there are two problems related to this p value only practice. Firstly, it changed the *priori* nature of the SHT decision deriving, i.e., a Type I error should be selected before one can make a decision. As mentioned above, only two p values, 0.05, which corresponds to a 95% confidence for the decision made or 0.01, which corresponds a 99% confidence, were used before the advent of the computer software in setting a Type I error. Secondly, and a more serious problem, the p value could be impacted by the sample size employed, making it an inconsistent standard in decision-making.

Let's go back to our example to illustrate why p value is not a consistent standard. By looking at the fat percentage means of males and females, you may quickly realize that the difference between the two means is rather large. How was the p value larger than 0.05 when there seems to be an obvious difference between the two means? To get a less-than 0.05 p value or to reject the null hypothesis is, in fact, not difficult as long as we have a large enough statistical power, which is the probability of rejecting the null hypothesis when it is false (i.e., detecting a real difference). There are four factors that may impact the statistical power: (a) α level, (b) one-tailed or two-tailed test, (c) effect size (ES), and (d) sample size. Since the α level (0.05 or 0.01) or the direction of the test (we use a two-tailed test the majority of the time) are often fixed, two things that can affect the statistical power in practice are ES or sample size. For the ES of our example, we computed Cohen's d index:[5]

$$ES\ (Cohen\text{'}s\ d) = \frac{\text{Mean of female} - \text{Mean of male}}{\text{SD}_{\text{pooled}}}$$

$$= \frac{23.958 - 15.942}{\surd\, ((8.330)^2 + (10.646)^2) / 2} = 0.839$$

According to the Cohen's ES standard (≥0.8 = large; <0.8 to > 0.2 = medium; ≤0.2 = small), ES of between male and female mean difference in our example indeed belongs to "large". Thus, the reason the H_0 was not rejected is likely due to the small sample size ($n = 5$ for each group) employed. To verify this finding, we compute the sample size needed to get enough power by entering ES of 0.839, the desired statistical power of 0.8 and α level of 0.05 into an online sample size calculator for *t* test (http://www.danielsoper.com/statcalc3/calc.aspx?id=47). For a two-tailed hypothesis, the recommended sample size per group is 24. For the purpose of illustration, rather than to collect another 19 data points for each group, we simply copied and pasted the existing data three times, which made the sample size of each group 20 and recalculated means, SDs and *t* test. Here are the results:

Female fat%: M = 23.958, SD = 7.644
Male fat%: M = 15.942, SD = 9.770

$$t = (M_{\text{female}} - M_{\text{male}}) \big/ \sqrt{}\,(s^2{}_{female} \big/ n_{female} + s^2{}_{male} \big/ n_{female}) = 8.02 \big/ \sqrt{}\,7.69 = 2.89$$

p value = 0.006.

As expected, the means remained the same, SDs became slightly smaller, *t* statistic became larger, and the most important change, of course, is that *p* value is now less than 0.05 so that the earlier "no difference" conclusion suddenly changed to a "significant" difference. It should be pointed out the *p* value problem is not only in the situation where a true difference could not be detected when a small sample was employed, but also a little, meaningless difference or no/low correlation could become "significant" when a large sample was employed.[6] It is this inconsistency that makes the *p* value useless in decision-making.

The above procedures also demonstrated that as long as ES is determined, needed sample size to get a less than 0.05 *p* value can be easily estimated. Since an absolute evaluation system has been developed for ES (e.g., the small-medium-large rating for Cohen's *d*), there is no need to use an extra inconsistent decision-making system. Criticism of the *p* value and the SHT is not new; in fact, it has a rich history of more than 80 years.[6–9] The problem of the abuse of the *p* value, which is often incorrectly used as a symbol of a significant finding, is clearly getting worse due mainly to the exact *p* values provided by modern statistical software. It is my strong opinion that this reporting practice be stopped. In addition to using ES[5] as an alternative, other recommendations of alternative approaches include exploratory data analysis,[10] confidence interval,[11] meta-analysis,[12,13] and Bayesian applications,[14] *etc.*

Considering *p* value is currently required by the most journals in the submission process and expected by peer-reviewers, a more practical recommendation to report statistics and *p* value is as follows:

1. In the method section, clearly state the α level (the *a priori* criterion for the probability of falsely rejecting your null hypothesis, which is typically 0.05 or 0.01) used as a statistical significance criterion for your tests. Example: "We used an α level of 0.05 for all statistical tests";
2. For correlations, use the absolute criterion:[6]
   0–0.19: no correlation,
   0.2–0.39: low correlation,
   0.40–0.59: moderate correlation,
   0.60–0.79: moderately high,
   ≥ 0.80: high correlation, or report the correlation determinations, i.e., squared correlation coefficients;
3. For regression or similar statistics, report the proportion of variance explained (e.g., R^2);
4. For all other inferential statistics, report statistics themselves, corresponding *df* and ES;
5. There are two ways to report *p* values: (a) report *p* value based on the α level determined, e.g., "$p > 0.05$ or $p < 0.05$" or "$p > 0.01$ or $p < 0.01$" and (b) report the exact *p* value (the *posteriori* probability reported by the statistical software). If the exact *p* value is less than 0.001, it is conventional to state merely $p < 0.001$;
6. Use "statistically significant or statistically not significant", rather than "significant or not significant" when reporting a *p* value based finding.

In summary, due to the conveniently available exact *p* values provided by modern statistical data analysis software, there is a wave of *p* value abuse in scientific inquiry by considering a $p < 0.05$ or 0.01 result as automatically being significant findings and that a smaller *p* value represents a more significant impact. After explaining the roots of the problem and why *p* value should not be used in this way, some practical recommendations on appropriately reporting statistical findings, including *p* value, are provided.

Competing interests

The author declares no competing financial interests.

References

1. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd; 1925.
2. Biau DJ, Jolles BM, Porcher R. *p* value and the theory of hypothesis testing. *Clin Orthop Relat Res* 2010;**468**:885–92.
3. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data test. *Front Psychol* 2015;**6**:223. doi:10.3389/fpsyg.2015.00223
4. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A* 1933;**231**:289–337.
5. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. New York, NY: Psychology Press; 1988.
6. Zhu W. Sadly, the earth is still round ($p < 0.005$). *J Sport Health Sci* 2012;**1**:9–11.
7. Carver RP. The case against statistical significance testing. *Harv Educ Rev* 1978;**48**:378–99.
8. Carver RP. The case against statistical significance testing, revisited. *J Exp Educ* 1993;**61**:287–92.
9. Savage IR. Nonparametric statistics. *J Am Stat Assoc* 1957;**52**:331–44.
10. Tukey JW. *Exploratory data analysis*. Reading, MA: Addison-Wesley; 1977.
11. Neyman J. On the problem of confidence intervals. *Ann Math Stat* 1935;**6**:111–6.
12. Glass GV, McGaw B, Smith ML. *Meta-analysis in social research*. Thousand Oaks, CA: Sage Publications; 1981.
13. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. New York, NY: Academic Press; 1985.
14. Wagenmakers EJ. A practical solution to the pervasive problems of *p* values. *Psychon Bull Rev* 2007;**14**:779–804.