Statistics In Brief

When to Use and When Not to Use a Threshold P Value

Bruno Falissard MD, PhD

Published online: 20 September 2011
© The Association of Bone and Joint Surgeons® 2011

Background

Hypothesis tests are statistical tools designed to help investigators control the risk that is taken when making inferences or basing decisions on probabilistic data. Although these tools are easy to implement in practice, they are conceptually more complex than generally is assumed [3, 5]. Statistical tests of an hypothesis can be presented according to two different theoretical frameworks: the Neyman-Pearson approach, which uses a threshold probability value, and the Fisher approach, which does not.

Question

When does a researcher use the Neyman-Pearson approach or when should he or she use the Fisher approach?

The author certifies that he, or a member of his immediate family, has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.

B. Falissard (✉)
INSERM U669, Maison de Solenn, Univ. Paris-Sud and Univ. Paris-Descartes, 97 Bd du Port Royal, 75679 Paris Cedex, France
e-mail: falissard_b@wanadoo.fr

B. Falissard
AP-HP, Hôpital Paul Brousse, Villejuif, France

Discussion

The Neyman-Pearson approach is presented in many textbooks and statistical courses. Here, two hypotheses are considered and the objective is to find which is the more compatible with data. Consider for instance the classic situation of a randomized controlled trial in which two proportions p_1 and p_2 are to be compared; there is the so-called null hypothesis $H_0: p_1 = p_2$ (H_0 corresponds to no difference between groups) and, the alternative hypothesis $H_A: p_1 \neq p_2$ (H_A , showing a difference between groups, is in general the hypothesis that the investigator wants to confirm). In this situation there are two ways the investigator can make the wrong decision: accepting the alternate hypothesis when the null hypothesis is true (the probability of this error is called alpha and the error is a type I error) and of accepting the null hypothesis when the alternative hypothesis is true (the probability of this error is called beta and the error is a type II error). Neyman and Pearson developed a theorem which leads to a series of decision algorithms for choosing H_0 or H_A that minimizes beta when alpha is fixed (alpha usually is set at 5%) [6]. These decision algorithms have well-known names like the chi square test and Student's t-test. As an illustration, if, in a randomized controlled trial, 1000 subjects are compared with 1000 subjects and if one observes that $p_{1\text{obs}} = 13\%$ and $p_{2\text{obs}} = 21\%$ of patients are cured in each group, for $\alpha = 5\%$ the decision algorithm will decide that the alternate hypothesis should be accepted (ie, there is a difference).

The Fisher approach, which is more of a probabilistic nature, uses the well-known p values. Imagine that the same two proportions, p_1 and p_2 , are to be compared and that one observes again (in 2000 subjects) $p_{1\text{obs}} = 13\%$ and $p_{2\text{obs}} = 21\%$. The idea is to determine to what extent

these two observed values are compatible with the hypothesis: $p_1 = p_2$. This can be done by estimating with statistical software the probability p (the p value) that a difference between $p_{1\text{obs}}$ and $p_{2\text{obs}}$ observed by chance is at least as large as the difference $21\% - 13\% = 8\%$ actually observed in the experiment. In practice if p is close to 0 it is unlikely that $p_1 = p_2$. Many statisticians and investigators suggest that if $p < 5\%$, the difference between p_1 and p_2 can be considered statistically significant, so that $p_1 \neq p_2$. Some even suggest that if $p < 10\%$ there is a trend that $p_1 \neq p_2$, or that if $p < 1\%$ or even less, the difference between p_1 and p_2 is highly significant [4].

The Neyman-Pearson approach gives a clear-cut answer to the question, should the null hypothesis or the alternate hypothesis be accepted considering the two risks alpha and beta? That is, this approach establishes more or less arbitrary thresholds of alpha and beta for making a decision, and the exact value of p is not otherwise important (it is sufficient to know that p is inferior to alpha or not). In contrast, the Fisher approach proposes a number, the p value, which can be interpreted as the level of plausibility of the hypothesis $p_1 = p_2$. In this case, the exact p value is important to judge the level of plausibility. There is no notion of power calculation with the Fisher perspective, and this likely is why statisticians working with this approach often consider that when it is not possible to conclude that $p_1 \neq p_2$, it is not possible nevertheless to conclude that $p_1 = p_2$ [2], while it is possible to accept H_A or H_0 with the Neyman-Pearson perspective.

From a practical point of view, the Neyman-Pearson approach is preferred in a confirmatory setting, where a main hypothesis of interest is clearly defined a priori. A clear-cut ‘yes or no’ answer to the question, can we claim that p_1 is different from p_2 ?, is the answer many investigators seek. Such an answer would be sought in a randomized, controlled trial. However, the probabilistic use of a p value is useful in an exploratory setting such as in an epidemiologic study in which a series of risk factors are studied in relation to a given disorder.

Myths and Misconceptions

Statistical tests are not as elementary as is generally assumed.

There is no unique way to consider statistical testing and this brief article describes only two approaches (a Bayesian approach [1] is another).

The debate between the two approaches is not merely theoretical: according to the Neyman-Pearson approach, only the acceptance or rejection of the null hypothesis is relevant, the magnitude of the p value is not (it is sufficient to know whether p is inferior to alpha or not). This is the opposite of the Fisher approach.

The Neyman-Person approach can appear more rigid. However, it allows power calculations so that the risk taken when the null hypothesis is accepted is known.

Conclusions

The dimensional use of p values should be limited to exploratory contexts, including most epidemiologic studies. In confirmatory settings, like randomized controlled trials, the classic threshold of 5% may be used and there is no interpretation of the magnitude of p .

References

1. Bolstad WM. *Introduction to Bayesian Statistics*. ed 2. Hoboken, NJ: John Wiley; 2007.
2. Georgetown University Department of Psychology. Available at: <http://psychology.georgetown.edu/resources/researchmethods/statistics/8315.html>. Accessed August 24, 2011.
3. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol*. 1993;137:485-496.
4. Greenhalgh T. How to read a paper: statistics for the non-statistician. II: “Significant” relations and their pitfalls. *BMJ*. 1997;315:422-425.
5. Lehman EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc*. 1993;88:1242-1249.
6. Wikipedia contributors. Neyman–Pearson lemma. Wikipedia, The Free Encyclopedia; Available at: http://en.wikipedia.org/w/index.php?title=Neyman%E2%80%93Pearson_lemma&oldid=424348004. Accessed August 24, 2011.