Ojaas Hampiholi – 01/23/2020

Most of the decisions that are taken in the world for any subject fall into the Statistical Inference domain. Statistical Inference is science of performing experiment where data is collected keeping in mind a certain hypothesis that is to be tested and the null hypothesis and alternate hypothesis are decided prior to plugging the data into the statistical tests to find the significance levels (also known commonly as P-value). P values measure the strength of evidence against the null hypothesis. Stronger evidence against the null hypothesis are provided by lesser P values; however, it is interesting to note that the division of result as 'significant' and 'non-significant' was not the intentions of the founders of statistical inference. Fisher (Sir Ronald Aylmer Fisher) saw the P value as an index measuring the strength of evidence against the null hypothesis. He advocated 5% significance as a standard level for concluding that there is evidence against the hypothesis tested, though not as an absolute rule. Fisher proposed that "If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05 …." [1]. Another major problem that arises with the hypothesis testing methods is that the accuracy of such tests cannot be confirmed, this is since most of the tests are based on probability. According to the words of Neyman and Pearson "no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint." [2]. P values are largely misunderstood and misinterpreted. The most common misinterpretation that P value shows the probability of null hypothesis being true leads to the issue of capping the P value at a 5% cutoff.

Let us take the example of a clinical hypothesis study that must be done, where the results of the study affect the type of treatment given to patients suffering from the same problems in the future. Let us take the clinical case study of Vasodilator Trial. There are two types of vasodilator treatments for patients who suffer heart failures. The number of survivor patients treated with Enalapril were compared to the number of survivors among the patients who were given a combination of Hydralazine and Nitrates. According to G Guyatt, R Jaeschke, N Heddle, D Cook, H Shannon and S Walter "The null hypothesis in the vasodilator trial could be stated as - The true difference in the proportion surviving between patients treated with enalapril and those treated with hydralazine and nitrates is zero……. As the results diverge farther and farther from the finding of no difference, the null hypothesis that there is no difference between treatments becomes less and less credible." [3]. The research and subsequent testing here are done to observe difference between the values obtained for the treatment group who receive enalapril as compared group who receive hydralazine and nitrates. This trial illustrates hypothesis testing when there is a dichotomous (Yes-No) outcome, in this case, life or death. During the follow-up period, which ranged from 6 months to 5.7 years, 132 (33%) of the 403 patients assigned to the enalapril group died, as did 153 (38%) of the 401 assigned to the hydralazine and nitrates group. The statistics tests that are applied to the proportions of the observed data show that there is no significant difference between the mortality rates associated with the two groups, while the actual data shows up to 11% of significance level between the mortality rates for both the groups. However, when we use the hypothesis-testing framework and the conventional cut-off point of 0.05, we conclude that we cannot reject the null hypothesis, which is inconsistent with the observed data. If the sample size is increased or if the data samples are captured again with more accuracy, we can confirm the significance levels again. When a trial fails to reject the null hypothesis ($p > 0.05$) the investigators may have missed a true treatment effect, and we should consider whether the power of the trial was adequate. It is interesting to note that in such "negative" studies, the stronger the trend in favor of the experimental treatment, the more likely the trial missed a true treatment effect. The paper also states that there is one other problem with dichotomic tests, “Some studies are designed to determine not whether a new treatment is better than the current one but whether a treatment that is less expensive, easier to administer or less toxic yields the same treatment effect as standard therapy. In such studies (often called "equivalence studies") recruitment of an adequate sample to ensure that small but important treatment effects will not be missed is even more important. If the sample size in an equivalence study is inadequate, the investigator risks concluding that the treatments are equivalent when, in fact, patients given standard therapy derive important benefits in comparison with those given the easier, cheaper or less toxic alternative.” [3]

It is very important to note that correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise *P* values. Any opinion offered about the probability, likelihood, certainty, or similar property for a hypothesis **cannot** be derived from statistical methods alone. Significance tests and confidence intervals do not by themselves provide a logically sound basis for concluding an effect is present or absent with certainty or a given probability. This point should be borne in mind whenever one sees a conclusion framed as a statement of probability, likelihood, or certainty about a hypothesis. [4] Information about the hypothesis beyond that contained in the analyzed data and in conventional statistical models (which give only data probabilities) must be used to reach such a conclusion; Reproducing the test multiple times to confirm whether the test comes up with the same results is also one of the suggested ways to make sure that the inferences are proper. All statistical methods (whether frequentist or Bayesian, or for testing or estimation, or for inference or decision) make extensive assumptions about the sequence of events that led to the results presented—not only in the data generation, but in the analysis choices. Hence, it is extremely important that the attached report with the test mentions all the steps from objective of study, type of study, data collection methods, the subsequent tests and their parameters to the results of the tests drawn alongside the reasoning for the tests. All the data collected, and the analysis steps should be made available for confirmation of the test as and when required. All the above steps are equally important alongside the analysis of p-value to determine the significance.

References:

1. Fisher. *Statistical methods for research workers*. London: Oliver and Boyd, 1950:80
2. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans Roy Soc A 1993; 231:289-337*
3. G Guyatt, R Jaeschke, N Heddle, D Cook, H Shannon, and S Walter. *Basic statistics for clinicians: 1. Hypothesis testing.*
4. Greenland, S., Senn, S.J., Rothman, K.J. *et al.* Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* **31,** 337–350 (2016).