Week 6 Reading Responses

Maria Feiler

03/02/2022

Bender, R. and Lange, S. 2001. Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology 54*: 343-349.

Key Points

- Multiple test procedures are underutilized in biomedical and epidemiological research.
- Adjustments for multiple testing are necessary for confirmatory studies.
- Use of multiple tests for one final conclusion or decision is enough to have to consider this as a factor of your research.

Summary: There is a choice to be made by the researcher: control the comparison-wise error rate (CER) or the experiment-wise error rate (EER). If the EER needs to be controlled for (or you're conducting a confirmatory experiment), then a statistical adjustment for multiple tests is required. The procedure of Bonferroni adjustments and resampling-based procedures are discussed. In addition, five circumstances that require multiple test adjustments are discussed: when studying more than two groups, when there is more than one endpoint of an experiment, when there are repeated measurements, when subgroup analyses are employed, and when interim analyses are performed. Finally, the alternative of the Bayes method of statistics is briefly discussed. This is considered a better procedure because control of the type I error rate is not necessary to make valid inferences.

Berger, J.O. and Berry, D.A. 1988. Statistical analysis and the illusion of objectivity. *American Scientist 76*: 159-165.

Key Points

- The supposed credibility that is given to statistical procedures is still affected by human subjectivity
- The lab coat/statistical procedure does not erase the human underneath it.

Summary: This article presents the argument for the Bayesian school of statistics as the solution for subjectivity in scientific studies. Though most statistical applications will have similar results, differences in interpretation can cause major differences in reported results. Understandard statistical procedures, their example of 17 treatment/control trials (is it 17 trials? Or stop testing when you get 4 of either side?) exemplifies how the difference in what is "as or more extreme" in an experiment can influence your interpretations of the results. Rather than the standard, the Bayesian approach would instead calculate the probability that the hypothesis is true in light of the data. In the example provided, if the experiment was designed to stop if conclusive evidence is found, else conduct 44 trials, the Bayesian final probabilities will not be affected.

Cohen, J. 1994. The earth is round ($p < 0.05$). *American Psychologist 49*(12): 997-1003.

Key Points

- Just using a p-value to "prove the null hypothesis false" is not sufficient
- Cohen suggests using graphic methods, estimating effect sizes, creating confidence intervals, and making more informed decisions about the statistical procedures used is necessary to fix scientific research
- $P(D|H_0) \neq P(H_0|D)$

Summary: Hull hypothesis significance testing is interpreted as "Given these data, what is the probability that H_0 is true?" when really is tells us "Given that H_0 is true, what is the probability of our (or more extreme) data?" Their American member of Congress example shows how the introduction of improbability makes a poor premise sound more acceptable. The rejection of a hil hypothesis (there is no difference between A and B) only tells us that there is a difference. Not a direction or any sort of causal relationship. Finally, stop hiding your confidence intervals and start making better measurements.

Gawande, A. 1999. The cancer-cluster myth. *The New Yorker* Feb. 8, 1999: 34-37.

Discussed in class, see class notes.

Johnson, D.H. 1999. The insignificance of statistical significance testing. *The Journal of Wildlife Management 63*(3): 763-772.

Key Points

- Usually statistical significance testing is not helpful to scientific pursuits; instead it confuses interpretations of the data
- $p = P(observed\ or\ more\ extreme\ data|H_0)$
- p is arbitrary
- Null hypotheses are straw men

Summary: There are common three interpretations of hypothesis testing: (1) p is the probability that the results obtained were due to chance, (2) 1-p is the reliability of the result, or the probability of getting the same resultif the experiment were repeated, and (3) p is the probability that the null hypothesis is true. In reality, P is the probability of your OR MORE EXTREME data given the null hypothesis is true. the null hypothesis set up is designed to be shot down, because of the sheer unlikelihood of the average difference between two sets of data being 0, or them being exactly the same. The reaction of the investigator and the interpretation of the sample size are also important.

Practical importance of observed difference	Not statistically significant	Statistically significant
Not important	Happy	Annoyed
Important	Very sad	Elated

Practical importance of observed difference	Not statistically significant	Statistically significant
Not important	n okay	n too big
Important	n too small	n okay

Defining the type of significance is also important. Something may be statistically significant but not biologically significant, for example. Potential good alternatives to hypothesis testing include estimates and confidence intervals, decision theory, model selection, and Bayesian approaches.