Statistical Significance

J. Dana Stoll

University of Liverpool

November 12, 2016

Statistical Significance

When testing hypotheses in real-world situations, absolute truth can never be established. Complete induction is only available in mathematics as the domains are completely known. For real-world models phenomena may be observed in the future that are not known yet, that can falsify the theory. Theories are abstractions, leaving some terms as explanatory principles with uncertainties that are not further challenged. The question arises, how likely may phenomena be observed, that cannot be explained by the model, and the model needs refinement (Bateson, 1972, ch. 1, "Metalogue: What is an instinct?"). The probability to investigate is, given that the hypothesis is true, how likely is it that this particular set of data will be observed.

At p = .05 as confidence, there is a likelihood of one in 20 that a subsequently drawn sample will contradict the hypothesis and 19 samples will match the hypothesis. This confidence facilitates that subsequent tests will be able to recreate the findings of the study. The number goes back to Ronald Fisher, who used it as an evidence indicator to accept a hypothesis (not rejection rate). It subsequently found its way into null hypothesis significance testing (NHST) that tries to find out the probability with which an alternative hypothesis (H1) is true on based on the likelihood of rejecting a null hypothesis (H0).

NHST creates a probabilistic modus tollens that does only hold true if the base rate of the phenomenon in the population is high (reversibility condition). To be combined by Bayes's rule, base rates need to be known *a priori*. For low base rates probabilistic inversions are invalid and create significant numbers of false positives (cf. Kahnemann, 2011, ch. 16, "Causes trump statistics"). In physics, for example, some events are rarely observed, and significance testing is done to a confidence of 3σ (1 in 370) or 5σ (1 in 1.7 million). PBS (2016) shows an example of misinterpretation. Recent research suggested that the supernova data, on which the theory of an expanding universe was founded, may also allow for a non-expanding theory. By mapping out the confidence spaces of 1, 2 and 3 sigma for lambda dark matter/dark energy it becomes clear that they (possibly) contain many theories, some are just not very likely and may be ruled out by prior knowledge on one of the parameters (PBS, 2016, 7:00).

In medicine, a study with 2000 patients may not be representative for the population (sampling error), particularly if the prevalence in the general population is low. The explanation then may or may not be the reason for the observed data, and it is unclear whether corresponding treatment will match the condition. Lindmark et al. (2016) emphasize that clinical relevance must be paired with statistical significance, because "large hospitals can have statistically significant results even for clinically irrelevant deviations while important deviations in small hospitals can remain undiscovered" (p. 1). Moyé (2006) points out that statistical significance must beat advocated, "well-developed and motivated theories" (p. 21). The p-values for effects on total mortality or myocard infarctions in an experiment may be significantly different, while the percent reduction in events for both remains the same. To honor “first do no harm”, p-values in studies may not only reflect the benefit case (ibid., p. 139, 177).

References

Bateson, G. (1972). *Steps to an ecology of mind*. Northvale, NJ: Jason Aronson.

Cohen, J. (1994). The earth is round (p < .05). *American Psychological Association*, 49(12), 997–1003. Retrieved from http://www.stats.org.uk/statistical-inference/Cohen1994.pdf

Field, A. (2013). *Discovering Statistics using IBM* (4th ed.). London, UK: SAGE.

Kahnemann. D. (2011). Thinking fast and slow. London, UK: Allen Lane.

Lindmark, A., van Rompaye, B., Goetghebeur, E., Glader, E., & Eriksson, M. (2016). The importance of integrating clinical relevance and statistical significance in the assessment of quality of care – Illustrated using the swedish stroke register. *Plos ONE*, 11(4), 1-13. doi:10.1371/journal.pone.0153082

Moyé, L. A. (2006). Stastistical reasoning in medicine: The intuitive p-value primer (2nd ed.) . Houston, TX: Springer.

PBS Space Time [PBS]. (November 9, 2016). Did dark energy just disappear? Retrieved November 12, 2016 from https://www.youtube.com/watch?v=7UNLgPIiWAg