0.01 and 0.05

R.R. Hoffman
September 2020

rhoffman@ihmc.us

The Concept of Statistical Significance

This concept is characterized by disagreements among statisticians and scientists, with some arguing that tests of statistical significance actually harm scientific progress because of faulty interpretations (e.g., Armstrong, 2007; Glaser, 1999; Kline, 2004 Ch. 3). Hence, many suggest shifting from significance testing to evaluation in terms of confidence intervals or prediction intervals. Another criticism is that the concept of statistical significance is often subject to misinterpretation by non-statisticians (see Gigerenzer, 2004; Wasserstein and Lazar, 2016).

In the tradition established by Ronald Fischer and Karl Pearson, among others, statistical significance is understood as a difference between means expressed as some proportion of the standard deviation, a difference which is transformed to values of a probability integral (that is, a continuous dimension of numbers falling between zero and 1.00). This is interpreted as the probability that the difference is due to chance. If the probability is high, the sample means are believed to be indistinguishable from the estimated mean of the hypothetical population from which the samples might have been drawn. If the probability is low, the sample means are believed to represent some effect, that is, the samples are not representative of some single hypothetical population from which the samples might have been drawn.

For one extreme or ideal case, the difference is infinitesimal with regard to the probable error, in which case the likelihood of the difference is essentially zero and one can conclude with confidence that the difference is not random. The sample means are believed to represent some effect, that is, the samples are not representative of a single hypothetical population from which the samples might have been drawn. In the other ideal case, the difference is infinite with regard to the probable error, in which case the likelihood of the difference is essentially 1.00 and one can conclude with confidence that the difference is random.

But...

Why 5% and 1%?

Claims that one or another psychological paradigm is flabby or heretical seem to roar themselves upon the stage on an occasional basis. Thus, today we see the imposition of requirements to pre-register hypotheses and conduct power analyses, and other procedures that sometimes actually seem to stifle scientific activity. One way that this yearning for scientific holiness has manifested itself is in the reliance on statistical significance.

The general tenet is that in order for the results of an experiment to be scientifically acceptable and the results determined to provide genuine evidence of a cause-effect relation between independent and dependent variables, there must be some form of parametric statistical test with a primary statistic having an associated probability of 5% or less, showing that the result reflects something other than random variation within the population that has been sampled. We love it if (and only if) the t-test or F-test yields $p < 0.01$ or $p < 0.05$. Probability less than 0.01 is better, of course. One often sees experimental reports in which the authors bemoaned the failure to attain such specific probability levels, and then point out something like "While the value of $p = 0.07$ did not attain statistical significance, the difference was in the expected direction" or some such waffling that admits, merely, "Gee whiz, we'd really like to believe we found the result we were looking for."

The question is not just why so specific, but why so why so strict?

The answer to this question, and its pertinence to the science of psychology, begins in 1880. In that year, and five years prior to his publication of *Über das Gedächtnis* ("On Memory"), Herrmann Ebbinghaus had submitted to Berlin University a document satisfying the requirements of an *Habilitaiton,* qualifying him to be a university lecturer (*Privatdozent*). This *Habilitationschrift* was essentially a draft of *Über das Gedächtnis.* (For a detailed run-down on the experiments Ebbinghaus conducted, see Hoffman, et al., 1987.) But the *Habilitationschrift*, and the draft manuscript of "On Memory" *(Uhrmanustript),* elaborated material that was only footnoted in the eventual book (see Footnote 1 in Chapter IV). In a clear stroke of genius, he presented an empirical case for the claim that psychology could be a science. He compared the average deviation (then called "probable errors") (see Chapter II, Sections 5, 10) that he obtained for his list learning experiments with probable errors that had been reported in the physical and biological sciences—Hermann von Helmholtz's measurements of the speed of neural conduction, and James Prescott Joule's measurements of the mechanical equivalent of heat. In fact, Ebbinghaus' probable errors of about seven percent were an order of magnitude smaller than the physical measurements and very close to those for the biological measurements.

It is no surprise that Ebbinghaus was praised for his rigorous program of experiments on memory. Indeed, the praise that was laid upon him after publication of *Über das Gedächtnis* includes recognition that psychology had arrived, as a genuine science independent from its philosophical origins (see for example, James, 1885). Thanks largely to this thing called the probable error.

The next development, also particularly linked to psychology, was the determination of a metric for differences on this scale of the probable error. How much of a difference makes a difference?

Through the late 19th and early 20th centuries there was actually considerable discussion and debate of the issue of what would be good or best criterion levels (2%, 3.25%, 7%, and even 20%) for determining scientific significance, and whether one should refer to results that are "almost" significant or "highly significant," or "very improbable" (see Cowles and Davis, 1982; Stigler, 2008). There were even more cumbersome attributions, such as "not very improbable that the observed frequencies are compatible with a random sampling" (Pearson, 1900, p. 171). Dublin brewer William Gosset, writing using the pen name Student (1908), published his method

called the t-test, saying "three times the probable error in the normal curve, for most purposes, would be considered significant" (p. 18). Debate continued through the 20th century (see Yule and Kendall, 1950), but the die had been cast by Pearson. Apparently, he did not feel that the 0.10 level was strict enough, and did feel that 0.01 was convincing. And 0.05 falls between these values (see Cowles and Davis, 1982) (See Note 1, below.)

In 1935, Ronald Fisher introduced methods of analysis of variance, a significant advance in statistics. Fisher also expressed differences as a proportion of the standard deviation, rather than the probable error.

> Odds of about 20 to one, then, seem to have been found a useful social compromise with the need to allow some uncertainty, a compromise between (say) 0.20 and 0.0001. That is, 5% is arbitrary (as Fisher knew well), but fulfills a general social purpose. People can accept 5% and achieve it in reasonable size samples, as well as have reasonable power to detect effect-sizes that are of interest... One may consider the formatting of [Fisher's] tables as a brilliant stroke of simplification that opened the arcane domain of statistical calculation to a world of experimenters and researcher workers who would begin to bring statistical measure to their data analyses (Stigler, p. 12).

Cowles and Davis (1982) speculated that the 0.05 level was *not* arbitrary.

> ... the conventional rejection level of three times the probable error is equivalent to two times the standard deviation (in modern terminology a z-score of 2), which expressed as a percentage is about 4.56%... one may hazard a guess that Fisher simply rounded off this value to 5% for ease of explanation... 5% could be more easily digested by the uninitiated than the report that the result represented a z-score of approximately 2 (p.557).

The argument was that a probability of one-in-twenty would be in most scientist's comfort zones for rejecting the null hypothesis. The entrenchment of this metric, spoon fed to generations of psychology majors, has left the field in a somewhat closed-minded state about this notion of "significance." This is the way we do it. Period.

But...

It has been understood for decades in the field of statistics that statistical significance, in the sense of null hypothesis significance testing, can be readily achieved by merely increasing the sample size:

> If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large-for instance, on the

> order of 200,000-the chi-square p will be small beyond any usual limit of significance (Berkson, 1938, p. 526).

And what we see today in many research programs is the mindless requirement for power analysis in order to determine in advance how many "subjects" are needed to give one sufficient confidence that any difference between group means will achieve statistical significance. (The n is usually frighteningly large, and this feeds into the senseless drive to make science easy by having Mechanical Turk do all the work).

Keeping in mind that t-test and ANOVA results reference differences between means, one can acquire data on control and experimental groups where there is a slight but statistically significant difference between means, yet the frequency distributions of the two groups show an amazing degree of overlap. Figure 1 below is a case in point —based on real data (not collected by me, I should point out). The x-axis is a performance measure (proportion correct on the primary task), blocked over sub-ranges. The difference is statistically significant: One can see that the average for the Control condition (black histobars) is pulled towards the left by the high frequency at proportion correct = 0.33. But one can also see that a considerable number of participants in the Control condition performed *better* than a considerable number of participants in the Experimental condition.

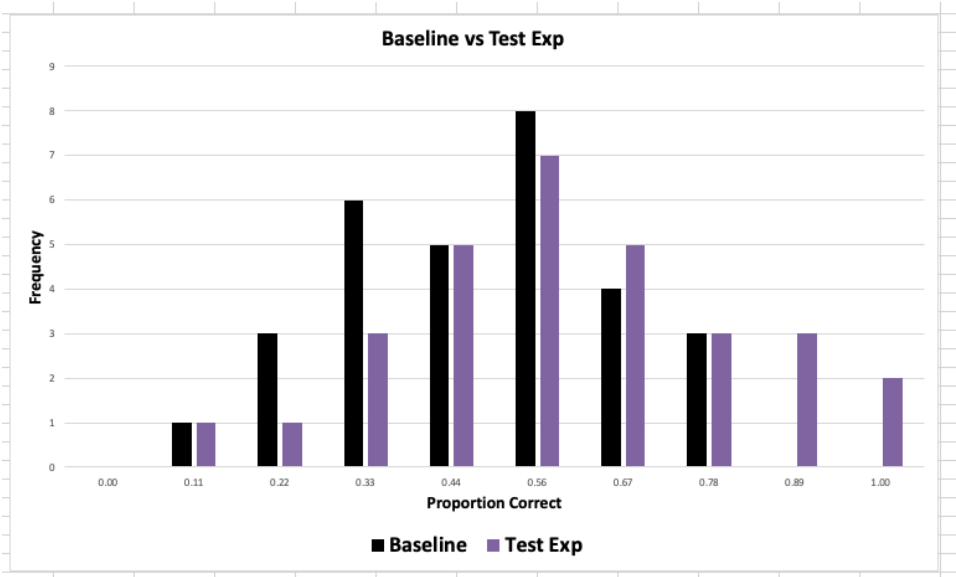

Figure 1. Frequency distributions that seem approximately "normal." Black histobars = Control Condition; Purple histobars = Experimental Condition.

Figure 2 below also presents actual frequency distributions, also with a statistically significant difference. But for "real" data such as we see here, it is unclear that the mathematical average is the best measure of central tendency, as illustrated by the distribution for Experimental condition. Indeed, in this example the very concept of a single best measure of central tendency gets called into question. And if the average is not a best measure of central tendency, is the standard error is a good measure of variability?

I have been collecting data sets (some with very large n, I should point out), and have done something that researchers (apparently) never do: Create and examine their actual frequency distributions. The inescapable conclusion is this: *The normal curve is not normal*. The only thing that comes to the researchers' rescue is the Central Limit Theorem, but the paradigm of parametric null hypothesis nevertheless runs head-long into its own assumptions. To paraphrase T.H. Huxley, we see here 'the slaying of a beautiful paradigm by ugly facts.'

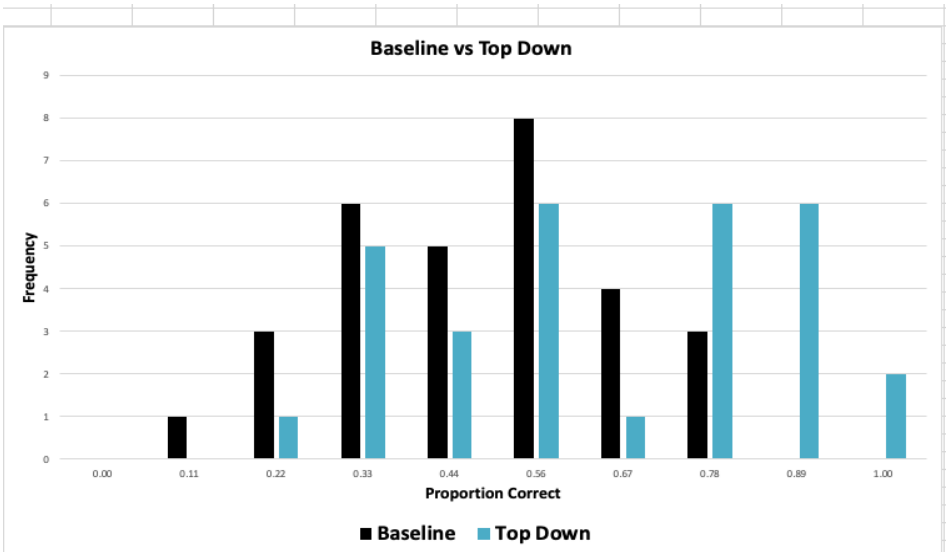

Figure 2. Frequency distributions very often do not seem even approximately "normal." Black histobars = Control condition; Blue histobars = Experimental Condition.

Conclusion

Statistics as a field is wide-open for innovation. The world we see, and study, does not orbit in some sort of gravitational dependence on parametric statistical tests of the null hypothesis. Statistical analysis is an exploratory activity. It is not, as some might wish, a means of calculating proper scientific judgments and decisions. F-ratios and t-tests are not a means of abrogating responsibility. They are exploratory tools.

In 1919, prominent experimental psychologist Edwin G. Boring raised some concerns about the then-new statistical procedures with respect to their experimental contexts. He concluded his discussion with a distinction between statistical and scientific significance: "Statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere" (p. 338). But Boring's suggestion itself does lead somewhere—to the concept of *practical* significance, which is hinted at in Figures 1 and 2. Can there be a means of calculating practical significance? This will be a topic for a forthcoming Concept Blog. As a tease, the answer is "yes."

A Final Thought

It is ironic that parametric statistical analysis and the achievement of statistical significance are regarded as a main aspect of "objective" science, distinguishing science from subjectivity. The irony is that the significance metric is relied upon by countless thousands of people, when the 0.05 and 0.01 levels were decided by a debate among just a few people, a debate centered on their individual judgments. All measures are determined by a consensus group, be it large or small, in an aggregation of the "subjective" judgments of individuals. All measures have objective and subjective characteristics (see Annett, 2002; Hoffman, 2019; Muckler, 1992)

References

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics, 45,* 966-987.

Armstrong, J.S. (2007). Significance tests harm progress in forecasting. [http://repository.upenn.edu/ marketing papers/99]

Berkson, J. (1938). some difficulties of interpretation encountered in the application of the Chi-Square test. *Journal of the American Statistical Association, 33*, pp.526-536

Boring, E. G. (1919). Mathematical versus scientific significance. *Psychological Bulletin, 16,* 225-338.

Cowles, M., and Davis, C. (1982). On the origins of the 0.05 level of statistical significance. *American Psychologist, 37,* 553-558

Ebbinghaus, H. (1885). *Über das Gedächtnis.* ("On Memory") Leipzig: Duncker and Humblot. Translated by H. Ruger and C. Busenius (1913.) New York: Columbia University Teacher's College.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics,* 33, 587-606.

Glaser, D.N. (1999). the controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care, 8,* 291-296.

Hoffman, R.R. (2019, November). Concept Blog Episode No. 2: "Subjective vs. Objective: Here We Go Again" [https://cmapsinternal.ihmc.us/viewer/cmap/1V5NF4884-271MKCX-2N98]

Hoffman, R.R., Bringmann, W., Bamberg, M. and Klein, R. (1987). Some historical observations on Ebbinghaus. In D.S. Gorfein and R.R. Hoffman (Eds.) *Memory and Learning: The Ebbinghaus Centennial Conference* (pp, 77-88). Hillsdale, NJ: Lawrence Erlbaum Associates.

James, W. (1998). Experiments on memory. *Science, 6,* 198-199.

Kline, R.B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Muckler, F.A. (1992). Selective performance measures: "Objective" versus "subjective" measurement. *Human Factors, 34,* 441-455.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlational system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*, 157-175.

Stigler, S. (2008). Fisher and the 5% level. *Chance, 21,* 12.

Wasserstein, R.L., and Lazar, N.A. (2016). The ASA statement on p-values: Context, process and purpose. *The American Statistician, 70,* 129-133.

Yule, G.U., & Kendall, M.G. (1950). *An introduction to the theory of statistics*. London: Griffin.