

11-1-2015

## In (Partial) Defense of .05

Thomas R. Knapp

University of Rochester and The Ohio State University, tknapp5@juno.com

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

### Recommended Citation

Knapp, Thomas R. (2015) "In (Partial) Defense of .05," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 2 , Article 6.

DOI: 10.22237/jmasm/1446350700

Available at: <http://digitalcommons.wayne.edu/jmasm/vol14/iss2/6>

This Invited Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

## ***Invited Article*** **In (Partial) Defense of .05**

**Thomas R. Knapp**  
University of Rochester  
Rochester, NY

---

Researchers are frequently chided for choosing the .05 alpha level as the determiner of statistical significance (or non-significance). A partial justification is provided.

*Keywords:* .05 level, statistical significance, R. A. Fisher

---

### **Introduction**

For the last 50 or 60 years it has been fashionable to deride the insistence on using an alpha level of .05 for testing the statistical significance of a sample finding. It is commonplace to read critical comments such as “The current obsession with .05” (Skipper, Guenther, & Nass, 1967, p. 16; see also Labovitz, 1968) and “God loves the .06 nearly as much as the .05” (Rosnow & Rosenthal, 1989, p. 1277). In the spirit of Robinson, Funk, Halbur, and O’Ryan (2003) I would like to provide an explanation for ‘why .05?’ and an argument in favor of its prevailing use. Near the end of the paper I will give a similar argument for 95% confidence (.05’s interval estimation counterpart), and I will conclude with a few cautionary statements regarding total devotion to .05 and/or 95%.

### **A bit of history**

Although there is some evidence for earlier recommendations of .05 as a defensible level of statistical significance, most people claim that it was first suggested by Fisher (1926):

[T]he evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the

---

*Dr. Knapp is Professor Emeritus of Education and Nursing at the University of Rochester and The Ohio State University. Email him at [tknapp5@juno.com](mailto:tknapp5@juno.com).*

level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 per cent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) or one in a hundred (the 1 per cent point). Personally, the writer prefers to set the low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. (p. 504)

There are several things to note about what Fisher said:

1. He used the interesting phrase “the verge of significance”. As far as I have been able to determine, none of his critics have commented about that choice of words.
2. He did not insist on .05, as the second part of the quote indicated. Many of his critics unfairly charged him with being unwavering regarding .05.
3. Surprisingly, he confused probability with odds (and high with low). The alpha level of .05 has to do with a probability of one in twenty; the corresponding odds are one to nineteen (in favor) or nineteen to one (against).

Fisher didn't write about .05 being the probability of making a Type I error. That concept (along with the probability of making a Type II error) was yet to come in the Neyman-Pearson approach to hypothesis testing. Also yet to come were several acrimonious arguments between Fisher and W. S. Gosset (who had previously developed the t-test), between Fisher and Karl Pearson, and between Fisher and both Jerzy Neyman and Egon Sharpe Pearson (Karl's son), as documented by Fienberg and Tanur (1966), Cowles and Davis (1982), Inman (1994), Wainer and Robinson (2003), and others.

In the intervening years between 1926 and the present there were several criticisms of .05, e.g., Cohen (1994), along with some defenders, e.g., Robinson, et al. (2003). Cohen (1994) was particularly puzzling (see the collection of comments regarding it in the December, 1995 issue of *American Psychologist*). The title is difficult to understand. Was he trying to be clever in considering “The earth is round” as a null hypothesis that should be rejected at the .05 level,

because it is actually slightly elliptical rather than perfectly round? He also made an error where he claimed many people believe a  $p$ -value is the probability that the null hypothesis is false. No; some people mistakenly believe that a  $p$ -value is the probability that the null hypothesis is true; no one believes  $p$  is the probability of a false null.

After discussing some of the historical origins of the use of an alpha level of .05, Robinson, et al. (2003) provided the results of empirical studies in which students were asked how many heads in each of the first  $n$  flips of a coin would lead them to claim that the coin was not “fair”. The modal response in most of those studies was five. The probability of heads on the first five tosses of a fair coin is .03125, which is close to the traditional .05 (see Figure 1 below).

### **A rationale for .05**

Although Fisher didn't use the following argument, some of the students in the Robinson, et al. (2003) studies apparently did, implicitly if not explicitly. (Comparable arguments have been made by Tintle, et al., 2014 and at the EMBstats website, <http://www.embstats.com>. See Figure 1 below for the latter.) Suppose you were asked your opinion about the fairness of a coin. You want to make a decision if its probability of landing as heads is equal to .5. How many heads would have to be obtained in the first five tosses for you to call a halt and conclude it's not a fair coin? The probability of one head in one toss of a fair coin is .5. (You wouldn't call a halt.) The probability of two heads in two tosses is  $.5 \times .5 = .25$ , and the probability of three heads in three tosses is  $.5 \times .5 \times .5 = .125$ . (Still no clear decision to halt.) The probability of four heads in four tosses is  $.5 \times .5 \times .5 \times .5 = .0625$ . (Perhaps the decision to halt is near, and note .0625 is close to .05.) If you want to wait for the result of one more toss, the probability of five heads in five tosses is  $.5 \times .5 \times .5 \times .5 \times .5 = .03125$ . At this point you are likely to claim that the coin is not fair. (The difference between .0625 and the .03125 is .046875, which is very close to .05.) However, you know you might be wrong.

Figure 1 details the argument presented at the EMBstats website. Note the interpretations of “Unusual” (for 4 heads in 4 tosses), “Surprising” (for 5 heads in 5 tosses), “Strange” (for 6 heads in 6 tosses), and “I don't believe it!” (for 7 heads in 7 tosses). Fisher's .05 would come between “Unusual” and “Surprising”. He avoided the matter of proof and exhibited a commendable tolerance for uncertainty. Similarly, statisticians are so comfortable with uncertainty that they occasionally advocate the use of the randomized response technique for

estimating a proportion where only some of the respondents to a survey actually answer the question of interest (Campbell & Joiner, 1973).

<b>Testing:</b> Is my coin fair?			
<b>Formally:</b> We want to make some inference about P(head)			
<b>Try it:</b> Toss coin several times (say 7 times). Assume that it is fair (P(head) = 0.5), and see if this assumption is compatible with the observations.			
# tosses	# heads	Comment	Probability
1	1	OK	0.50
2	2	OK	0.25
3	3	OK	0.12
4	4	Unusual	0.06
5	5	Surprising	0.03
6	6	Strange	0.02
7	7	I don't believe it!	0.01

Figure 1. EMBstats dialogue on tossing a coin (<http://www.embstats.com>).

## 95% confidence intervals

In the last 25 or 30 years there has been a pronounced shift from an emphasis on significance testing to a preference for confidence intervals. Some methodologists suggest reporting both; some journal editors require it. (Reporting both is not a good idea. See the third statistics commandment in Knapp & Brown, 2014). But the continuing choice of 95% for confidence (the interval estimation counterpart to .05 for hypothesis testing) has not been subject to the same sort of scrutiny that has been directed at .05. Why is that?

Perhaps consumers are more convinced by a 95% confidence argument than by the .05 significance argument. Consider the coin-tossing problem above, but change it to a desire for estimating the degree of bias associated with the coin rather than testing its fairness. If the coin-tosser got five heads in five tosses and was interested in estimating the population proportion of heads for that coin, he could get a confidence interval by using Pezzullo's online computing routine (<http://www.statpages.org>) based on Clopper and Pearson's (1934) formulas, tables, and graphs.

## IN (PARTIAL) DEFENSE OF .05

For example, at <http://www.statpages.org> for Exact Binomial Confidence Intervals input 5 heads (the numerator) in 5 tosses (the denominator, chose 95% confidence (the default). The results returned are 1.0000 as the statistic and .4782 to 1.0000 as the confidence interval. A choice of 99% confidence (corresponding to .01 significance) or 99.9% confidence (corresponding to .001 significance) serves only to reduce the lower limit (.3466 for 99% and .2187 for 99.9%) and therefore provides more confidence. Could it be that some people regard 99% confidence intervals and 99.9% confidence intervals to be too wide and are willing to stick with 95% for its greater precision despite its lesser confidence?

### Asterisks

Consider the still-common practice of labeling with a single asterisk a finding for which  $p < .05$ , two asterisks for  $p < .01$ , and three asterisks for  $p < .001$  (or what [Leahey, 2005](#) refers to as the three-star system”, Abstract). That is not sound practice (see [Slakter, Wu, & Suzuki-Slakter, 1991](#)), because if an alpha of .05 has been used in a power analysis to select an appropriate sample size, then all that is necessary to determine is whether  $p$  is less than or greater than .05. (Similarly, for alphas of .01 and .001.) Some journal editors require the reporting of the actual  $p$ , and that is the preferred practice according to the American Psychological Association manual ([APA, 2010](#)), which is not perfect, but is more sound than using asterisks.

To be consistent, why aren't asterisks or similar symbols used in the tables where authors report 95%, 99%, or 99.9% confidence intervals? If this statistic is significant at the .05 level and that statistic is significant at the .01 level, doesn't it make sense to put a 95% confidence interval around the first statistic and a 99% confidence interval around the second statistic?

All of the references so far have been to journal articles. There are three books on this topic that are recommended: [Fisher \(1925\)](#), [Salsburg \(2001\)](#), and [Moye \(2006\)](#). These three authors addressed the choice of .05 for statistical significance. [Fisher \(1925\)](#) contained some of the same views later expressed in [Fisher \(1926\)](#). [Salzburg](#) related [Fisher's](#) classic experiment regarding a lady's ability to determine whether milk has been added to tea or tea added to milk. [Moye](#) provided a thorough discussion of the advantages and disadvantages of  $p$ -values (mostly disadvantages). Both [Salzburg](#) and [Moye](#) gave fascinating accounts of [Fisher's](#) battles with [Neyman](#) and [Pearson](#) (and with [Gosset](#)). [Moye](#) noted that [Fisher](#) was not wedded to .05, as stated above.

## Some cautions

Continuing to emphasize .05 as the cut-off between statistical significance and non-significance is not all that bad. The same holds for continuing to emphasize 95% for confidence intervals. But there are exceptions.

1. If there might be very serious consequences should a Type I error be made, a more stringent alpha is necessary. For example, suppose a randomized clinical trial (RCT) were to be carried out comparing the effectiveness of a new and very expensive drug with an existing much less expensive drug. Suppose further that a decision might be made to reject the null hypothesis of no effect because of a statistically significant effect in favor of the new drug, but in reality it is no better. That could lead to the adoption of a drug that is not only no better than the existing drug but could result in an unnecessary cost of thousands or millions of dollars. In that case an argument could be made to use .01 or .001 or an even smaller significance level.
2. If the committing of a Type II error would have much greater consequences than a Type I error, the argument is reversed; i.e., change alpha to a more liberal level, such as .20. An example of this would be a medical diagnosis of no disease if a patient is in fact ill. Generally, it would be worse to not treat a patient who has a disease than to treat a patient when the disease is not present.
3. If the estimate of a population parameter must be both precise and defensible, a confidence coefficient of 99.9% might be chosen, as well as a huge sample size. For example, if an estimate of the proportion of people who are below the poverty line is to be made, we might want to do that in order to have both politically defensible and morally desirable evidence for so doing.

## References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Campbell, C., & Joiner, B. L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27(5), 229-231. doi:10.1080/00031305.1973.10479043
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413. doi:10.2307/2331986
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *The American Psychologist*, 49(12), 997-1003. doi:10.1037/0003-066X.49.12.997
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *The American Psychologist*, 37(5), 553-558. doi:10.1037/0003-066X.37.5.553
- Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review / Revue Internationale de Statistique*, 64(3), 237-253. doi:10.2307/1403784
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503-513.
- Fowler, R. D. (Ed.) (1995). Bridging science and practice [Full issue]. *American Psychologist*, 50(12).
- Inman, H. F. (1994). Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from Nature. *The American Statistician*, 48(1), 2-11. doi:10.1080/00031305.1994.10476010
- Knapp, T. R., & Brown, J. K. (2014). Ten statistics commandments that almost never should be broken. *Research in Nursing & Health*, 37(4), 347-351. doi:10.1002/nur.21605
- Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of .05. *The American Sociologist*, 3(3), 220-222. Retrieved from <http://www.jstor.org/stable/27701367>

THOMAS R. KNAPP

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance standards in sociology. *Social Forces*, 84(1), 1-24.  
doi:10.1353/sof.2005.0108

Moye, L. A. (2006). *Statistical reasoning in medicine: The intuitive p-value primer* (2nd. ed.). New York: Springer.

Robinson, D. H., Funk, D. C., Halbur, D., & O'Ryan, L. (2003). The .05 alpha level in educational research: Traditional, arbitrary, sacred, magical, or simply psychological? *Research in the Schools*, 10(2), 79-86.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *The American Psychologist*, 44(10), 1276-1284. doi:10.1037/0003-066X.44.10.1276

Salsburg, D. (2001). *The lady tasting tea*. New York: Freeman.

Skipper, J. K., Jr., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, 2(1), 16-18. Retrieved from <http://www.jstor.org/stable/27701229>

Slakter, M. J., Wu, Y.-W. B., & Suzuki-Slakter, N. S. (1991). \*, \*\*, and \*\*\*: Statistical nonsense at the .00000 level. *Nursing Research*, 40(4), 248-249.

Tintle, N. L., Chance, B., Cobb, G., Rossman, A. Roy, S., Swanson, T., & VanderStoep, J. (2014). *Introduction to statistical inference*. (Preliminary edition). Hoboken, NJ: Wiley.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, 32(7), 22-30.  
doi:10.3102/0013189X032007022