

Announcement**Why P=0.05?**

The standard level of significance used to justify a claim of a statistically significant effect is 0.05. For better or worse, the term *statistically significant* has become synonymous with $P \leq 0.05$.

There are many theories and stories to account for the use of $P=0.05$ to denote statistical significance. All of them trace the practice back to the influence of R.A. Fisher. In 1914, Karl Pearson published his *Tables for Statisticians & Biometricians*. For each distribution, Pearson gave the value of P for a series of values of the random variable. When Fisher published *Statistical Methods for Research Workers* (SMRW) in 1925, he included tables that gave the value of the random variable for specially selected values of P. SMRW was a major influence through the 1950s. The same approach was taken for Fisher's *Statistical Tables for Biological, Agricultural, and Medical Research*, published in 1938 with Frank Yates. Even today, Fisher's tables are widely reproduced in standard statistical texts.

Fisher's tables were compact. Where Pearson described a distribution in detail, Fisher summarized it in a single line in one of his tables making them more suitable for inclusion in standard reference works*. However, Fisher's tables would change the way the information could be used. While Pearson's tables provide probabilities for a wide range of values of a statistic, Fisher's tables only bracket the probabilities between coarse bounds.

The impact of Fisher's tables was profound. Through the 1960s, it was standard practice in many fields to report summaries with one star attached to indicate $P \leq 0.05$ and two stars to indicate $P \leq 0.01$. Occasionally, three stars were used to indicate $P \leq 0.001$.

Still, why should the value 0.05 be adopted as the universally accepted value for statistical significance? Why has this approach to hypothesis testing not been supplanted in the intervening three-quarters of a century?

It was Fisher who suggested giving 0.05 its special status. Page 44 of the 13th edition of SMRW, describing the standard normal distribution, states

The value for which $P=0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

Similar remarks can be found in Fisher (1926, 504).

... it is convenient to draw the line at about the 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."...

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 a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

However, Fisher's writings might be described as inconsistent. On page 80 of SMRW, he offers a more flexible approach

In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2 , but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: *Either* the hypothesis is untrue, *or* the value of χ^2 has attained by chance an exceptionally high value. The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.

These apparent inconsistencies persist when Fisher dealt with specific examples. On page 137 of SMRW, Fisher suggests that values of P slightly less than 0.05 are not conclusive.

[T]he results of *t* shows that P is between .02 and .05.

The result must be judged significant, though barely so; in view of the data we cannot ignore the possibility that on this field, and in conjunction with the other manures used, nitrate of soda has conserved the fertility better than sulphate of ammonia; the data do not, however, demonstrate this point beyond the possibility of doubt.

On pages 139-140 of SMRW, Fisher dismisses a value greater than 0.05 but less than 0.10.

[W]e find...*t*=1.844 [with 13 df, P = 0.088]. The difference between the regression coefficients, though relatively large, cannot be regarded as significant. There is not sufficient evidence to assert that culture B was growing more rapidly than culture A.

while in Fisher [19xx, p 516] he is willing pay attention to a value not much different.

...P=.089. Thus a larger value of χ^2 would be obtained by chance only 8.9 times in a hundred, from a series of values in random order. There is thus some reason to suspect that the distribution of rainfall in successive years is not wholly fortuitous, but that some slowly changing cause is liable to affect in the same direction the rainfall of a number of consecutive years.

Yet *in the same paper* another such value is dismissed!

[paper 37, p 535] ...P=.093 from Elderton's Table, showing that although there are signs of association among the rainfall distribution values, such association, if it exists, is not strong enough to show up significantly in a series of about 60 values.

Part of the reason for the apparent inconsistency is the way Fisher viewed P values. When Neyman and Pearson proposed using P values as absolute cutoffs in their style of fixed-level testing, Fisher disagreed strenuously. Fisher viewed P values more as measures of the evidence against a hypotheses, as reflected in the quotation from page 80 of SMRW above and this one from Fisher (1956, p 41-42)

The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests.

A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.

Still, we continue to use P values nearly as absolute cutoffs but with an eye on rethinking our position for values close to 0.05^{**}. Why have we continued doing things this way? A procedure such as this has an important function as a gatekeeper and filter--it lets signals pass while keeping the noise down. The 0.05 level guarantees the literature will be spared 95% of potential reports of effects where there are none.

For such procedures to be effective, it is essential there be a tacit agreement among researchers to use them in the same way. Otherwise, individuals would modify the procedure to suit their own purposes until the procedure became valueless. As Bross (1971) remarks,

Anyone familiar with certain areas of the scientific literature will be well aware of the need for curtailing language-games. Thus if there were no 5% level firmly established, then some persons would stretch the level to 6% or 7% to prove their point. Soon others would be stretching to 10% and 15% and the jargon would become meaningless. Whereas nowadays a phrase such as *statistically significant difference* provides some assurance that the results are not merely a manifestation of sampling variation, the phrase would mean very little if everyone played language-games. To be sure, there are always a few folks who fiddle with significance levels--who will switch from two-tailed to one-tailed tests or from one significance test to another in an effort to get positive results. However such gamesmanship is severely frowned upon and is rarely practiced by persons who are *native speakers* of fact-limited scientific languages--it is the mark of an amateur.

Bross points out that the continued use of P=0.05 as a convention tells us a good deal about its practical value.

The continuing usage of the 5% level is indicative of another important practical point: it is a feasible level at which to do research work. In other words, if the 5% level is used, then in most experimental situations it is feasible (though not necessarily easy) to set up a study which will have a fair chance of picking up those effects which are large enough to be of scientific interest. If past experience in actual applications had not shown this feasibility, the convention would not have been useful to scientists and it would not have stayed in their languages. For suppose that the 0.1% level had been proposed. This level is rarely attainable in biomedical experimentation. If it were made a prerequisite for reporting positive results, there would be very little to report. Hence from the standpoint of communication the level would have been of little value and the evolutionary process would have eliminated it.

The fact that many aspects of statistical practice in this regard *have* changed gives Bross's argument additional weight. Once (mainframe) computers became available and it was possible to calculate precise P values on demand, standard practice quickly shifted to reporting the P values themselves

rather than merely whether or not they were less than 0.05. The value of 0.02 suggested by Fisher as a *strong* indication that the hypothesis fails to account for the whole of the facts has been replaced by 0.01. However, science has seen fit to continue letting 0.05 retain its special status denoting statistical significance.

*Fisher may have had additional reasons for developing a new way to table commonly used distribution functions. Jack Good, on page 513 of the discussion section of Bross (1971), says, "Kendall mentioned that Fisher produced the tables of significance levels to save space and to avoid copyright problems with Karl Pearson, whom he disliked."

**It is worth noting that when researchers worry about P values close to 0.05, they worry about values slightly greater than 0.05 and why they deserve attention nonetheless. I cannot recall published research downplaying P values less than 0.05. Fisher's comment cited above from page 137 of SMRW is a rare exception.

References

- Bross IDJ (1971), "Critical Levels, Statistical Language and Scientific Inference," in Godambe VP and Sprott (eds) *Foundations of Statistical Inference*. Toronto: Holt, Rinehart & Winston of Canada, Ltd.
- Fisher RA (1956), *Statistical Methods and Scientific Inference* New York: Hafner
- Fisher RA (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.
- Fisher RA (19xx), "On the Influence of Rainfall on the Yield of Wheat at Rothamstead,"

Gerard E. Dallal

Last modified: 05/23/2012 03:52:42.

Statistical Rules of Thumb

Second Edition

Gerald van Belle

University of Washington
Department of Biostatistics
and Department of Environmental and Occupational Health Sciences
Seattle, WA

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith,
Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.



WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

data, the uncertainty due to this estimation, which is incorporated in the t -statistic, decreases also as the sample size increases.

Parker and Berman (2003) provide another way of looking at this issue by considering the amount of information provided by a confidence interval based on the t -distribution. They reach similar conclusions.

1.15 ESTIMATE \pm TWO STANDARD ERRORS IS REMARKABLY ROBUST

Introduction

Interval estimates such as confidence intervals require specification of the distribution from which the observations were drawn. Often this is not known or obtainable.

Rule of Thumb

For sample sizes ≥ 20 a point estimate \pm two standard errors has approximately 95% coverage for a wide variety of distributions.

Illustration

Figure 1.2 illustrates the somewhat curious phenomenon that the non-coverage probability is approximately 5% for the interval defined by the estimate of the mean \pm two standard errors for a very diverse set of distributions. The distributions are means of exponentials of sample size 1, 2, 5, and 20. This ranges from a very long-tailed distribution to one becoming reasonably symmetrical for means of 20 observations. Around the standardized value of 2 the coverage is about 95%, the non-coverage is about 5%.

Note that the ordinate scale in the figure is logarithmic; this tends to emphasize the divergence of the coverage probabilities from the crossings at 2.

Basis of the Rule

The basis of the rule is already found in Pearson and Tukey (1965) as pointed out by Andrews (2007). He shows that the 95% coverage probability is remarkably robust, that is, under a wide variety of distributions the 95% level intervals for these distributions tend to be similar.

Discussion and Extensions

Andrews (2007) demonstrates the same pattern using the chi-square distribution. The requirement that the number of observations be ≥ 20 is needed to provide a reasonably

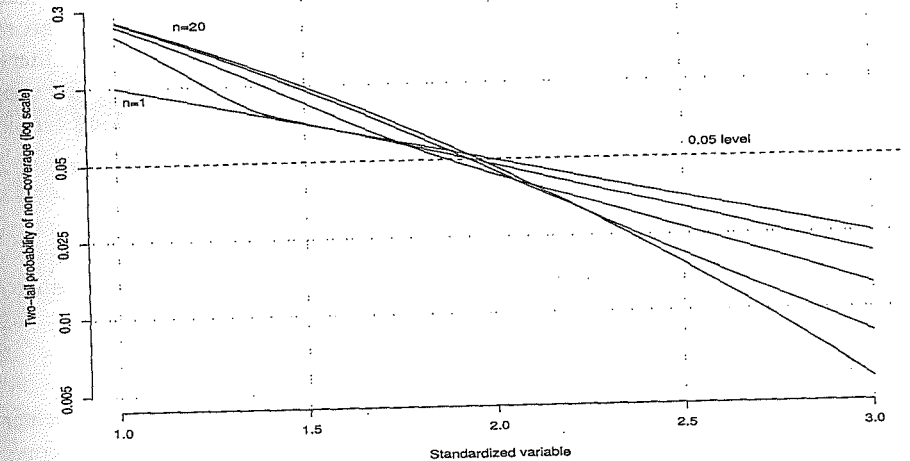


Fig. 1.2 Tail probability of non-coverage of estimate ± 2 standard errors. Curves are for exponential distribution with sample sizes 1, 2, 5, and 20.

stable estimate of the variance; an added benefit is that the distribution of means of 20 observations is reasonably symmetric.

The interesting feature of this rule is that it only applies to a 95% confidence interval; it does not work for example, for an 80% confidence interval—as indicated by the graph.

1.16 KNOW THE UNIT OF THE VARIABLE

Introduction

Variables come with units. Sometimes these units are considered “natural.” For example, annual income is usually measured in dollars, or euros or pounds. It is not “natural” to consider the logarithm of this income. However, statisticians frequently transform variables in order to achieve homogeneity of variance, normality, or some other statistical property.

Rule of Thumb

Always know what the unit of a variable is.