On the Origins of the .05 Level of Statistical Significance

MICHAEL COWLES *York University, Canada*
CAROLINE DAVIS *York University, Canada*

ABSTRACT: *Examination of the literature in statistics and probability that predates Fisher's Statistical Methods for Research Workers indicates that although Fisher is responsible for the first formal statement of the .05 criterion for statistical significance, the concept goes back much further. The move toward conventional levels for the rejection of the hypothesis of chance dates from the turn of the century. Early statements about statistical significance were given in terms of the* probable error. *These earlier conventions were adopted and restated by Fisher.*

It is generally understood that the conventional use of the 5% level as the maximum acceptable probability for determining statistical significance was established, somewhat arbitrarily, by Sir Ronald Fisher when he developed his procedures for the analysis of variance.

Fisher's (1925) statement in his book, *Statistical Methods for Research Workers*, seems to be the first specific mention of the $p = .05$ level as determining statistical significance.

> It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (p. 47)

Cochran (1976), commenting on a slightly later, but essentially similar, statement by Fisher (1926), says that, "Students sometimes ask, 'how did the 5 per cent significance level or Type I error come to be used as a standard?' . . . I am not sure but this is the first comment known to me on the choice of 5 per cent" (p. 15).

In the 1926 article Fisher acknowledges that other levels may be used:

> 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. (p. 504)

Cochran feels that Fisher was fairly casual about the choice, "as the words *convenient* and *prefers* have indicated" (p. 16). However, the statement quoted above leaves no doubt about Fisher's acceptance of the level as the critical cutoff point, once he had decided upon it.

Other writers, well-versed in the history and development of probability, have also fostered the attitude that the level is an arbitrary one. Yule and Kendall (1950), in the 14th edition of a book first published by Yule in 1911, state,

> In the examples we have given . . . our judgment whether P was small enough to justify us in suspecting a significant difference . . . has been more or less intuitive. Most people would agree . . . that a probability of .0001 is so small that the evidence is very much in favour. . . . Suppose we had obtained $P = 0.1$. . . . Where, if anywhere, can we draw the line? The odds against the observed event which influence a decision one way or the other depend to some extent on the caution of the investigator. Some people (not necessarily statisticians) would regard odds of ten to one as sufficient. Others would be more conservative and reserve judgment until the odds were much greater. It is a matter of personal taste. (pp. 471–472)

Cramer (1955), in a completely rewritten version of a Swedish text first published in 1926, tells his readers,

> a value of t . . . will be denoted as *almost significant* if t exceeds the 5% value, but falls short of the 1% . . . called *significant* if t lies between the 1% and the 0.1% values and *highly significant* if t exceeds the 0.1% value. This is, of course, a purely conventional terminology. (p. 202)

The issue to be considered is whether the choice of the 5% value was as arbitrary and casual as is so often implied. An examination of the history of probability and statistical theory, however, indicates that the choice was far from arbitrary and was influenced by previous scientific conventions

Request for reprints should be sent to Michael Cowles, Department of Psychology, York University, Downsview, Ontario, Canada M3J 1P3.

0003-066X/82/3705-0553$00.75

that themselves were based on the notion of "chance" and the unlikelihood of an event occurring.

Origins

As David (1962) has so articulately and elegantly described, the first glimmerings of an appreciation of long-run relative frequencies, randomness, and the unlikelihood of rare events being merely fortuitous go back at least to the Greek mathematicians and the Roman philosophers. Later, however, the spread of Christianity and the collapse of the Roman Empire made the Church the sole haven for scholars. This religious philosophy that accepted a universe in which every event, no matter how trivial, as being caused by an omnipotent God left no place for the investigation of random events. This is very likely the reason why the seeds of mathematical probability theory were not sown until late in 17th-century France. The opportunities had always been there: Because both the archaeological and the written records show that gambling has been an ever-popular pastime, informal and relatively unsystematic "systems" for calculating "odds" were undoubtedly developed.

The questions posed by Antoine Gombauld, the Chevalier de Méré, related to certain gaming problems, sparked off the correspondence between Blaise Pascal and Pierre Fermat in 1654. Here are the beginnings of combinatorial algebra and mathematical probability theory (again see David, 1962).

In a slightly later (1662) development, John Graunt, a London haberdasher, constructed tables from the Bills of Mortality, parish accounts regularly recorded from early in the 17th century and, most importantly, used these tables for a series of statistical, actuarial inferences.

Graunt was, for example, able to reassure readers of his quite remarkable, unassuming, and refreshing work that,

> This *casualty* [Lunacy] being so uncertain, I shall not force myself to make any inference from the numbers, and proportions we finde in our Bills concerning it: onely I dare ensure any man at this present, well in his Wits, for one in the thousand, that he shall not die a *Lunatick* in *Bedlam*, within these seven years, because I finde not above one in about one thousand five hundred have done so. (Graunt, 1662/1956, p. 1430)

Here is a statement based on numerical data and couched in terms not so very far removed from those in reports in the modern literature.

In 1657, Huygens (1657/1970) published a tract, *On Reasoning in Games of Dice*, that was based upon the exchanges between Pascal and Fermat, and in 1713 Jacques Bernoulli's (1713/1970) book, *The Art of Conjecture*, developed a theory of games of chance. De Moivre's (1756/1967) *The Doctrine of Chances* was the most important of the gambling manuals; it appeared in three editions in 1718, 1738, and 1756. In the two later editions De Moivre presents a method, which he had first published in 1733, of approximating the sum of a very large number of binomial terms. It is safe to say that no other theoretical mathematical abstraction has had such an important influence on psychology and the social sciences as that method, for it generates the bell-shaped curve now commonly known by the name Karl Pearson gave it: the normal distribution.

The law of frequency of errors is often attributed to Laplace (1749–1827) and Gauss (1777–1855). Both men developed the use of the distribution outside of gaming and in particular demonstrated its utility in evaluating the variable results of measurements and observations in astronomy and in geodetic surveying. With the introduction of this distribution into the field of the biological and social sciences, we may start to trace the path that leads to the $p = .05$ level.

The Normal Distribution

The credit for the extension of the use of calculations used to assess observational error or gaming expectancies into the organization of human characteristics goes to Lambert Adolphe Quetelet (1796–1874), a Belgian astronomer.

Quetelet (1849) found, for example, that the frequency distribution of the chest girths of 5,738 Scottish soldiers closely approximated the normal curve. Moreover, he used the curve to infer what he took to be a non-chance occurrence. In examining the distribution of the heights of 100,000 French army conscripts, he observed a discrepancy between the calculated and reported frequencies of men falling at the minimum height for military service. "Is it not a fair presumption that the . . . men who constitute the difference of these numbers have been fraudulently rejected?" (p. 97).

Sir Francis Galton (1822–1911) eagerly adopted the curve in the organization of the anthropometric data that he collected and introduced the concept of percentiles.

All persons conversant with statistics are aware that this supposition brings Variability within the grasp of the laws of Chance, with the result that the relative frequency of Deviations of different amounts admits of being calculated, when these amounts are measured in terms of any self-contained unit of variability, such as our Q. (Galton, 1889, pp. 54–55)

Q is the symbol for the semi-interquartile range, defined as one half of the difference between the score at the 75th percentile (the third quartile) and the 25th percentile (the first quartile). This means that in a distribution of scores, one half of the deviations fall within ±Q of the mean, which in the normal distribution falls at the 50th percentile (the second quartile). This measure of variability is equivalent to the *probable error*.

Probable Error

The unit of measure of the abscissa of the normal distribution has had many forms. Today the *standard deviation* is the unit of choice, but for many years the probable error (*PE*) was in common use, and it is still used occasionally in the physical sciences. Fundamentally, probable error defines the deviation from a central measure between whose positive and negative values one half of the cases may be expected to fall by chance.

The term appeared in the early 19th century among German mathematical astronomers. Although De Moivre refers to the concept on which *PE* is based, Bessel used the term (*der wahrscheinliche Fehler*) for the first time in 1818. It was subsequently adopted by Gauss, who developed several methods of computing it (Walker, 1929). It was first used with the normal distribution in instances where it was necessary to determine the best possible value of the true position of a point from a series of measurements or observations all of which involved an element of error.

It remained for Karl Pearson (1894) to coin the term *standard deviation*, but the calculation of an equivalent value had existed since De Moivre. Simple calculation shows that the *PE* is equivalent to 0.674560, or roughly 2/3 of a standard deviation.

It was apparently normal practice for Quetelet and Galton to express values in a normal distribution as a function of *PE*, and it seems reasonable to assume that their preference was the overriding influence in its being used in subsequent statistical practice. It should be noted in passing that Galton (1889) objected to the name probable error, calling it a "cumbrous, slipshod, and misleading phrase."

The probable error is, quite clearly, not the most probable of all errors, and the use of the term *error* in describing the variation of human characteristics perhaps carries the analogy with measurement error distribution a shade too far.

Statistical Tests

In 1893 Pearson began his investigations into the general problem of fitting observed distributions to theoretical curves. The work led eventually to the formulation of the χ^2 test of "goodness of fit" in 1900, one of the most important developments in the history of statistics.

Weldon, the co-founder with Pearson of the biometric school (both men, of course, being much influenced by Galton), approached the problem of discrepancies between theory and observation in a much more empirical way, tossing coins and dice and comparing the outcomes with the binomial model.

In a letter written to Galton in 1894, Weldon asks for a comment on the results of some 7,000 throws of 12 dice collected for him by a clerk at University College, London.

A day or two ago Pearson wanted some records of the kind in a hurry, in order to illustrate a lecture, and I gave him the record of the clerk's 7,000 tosses . . . on examination he rejects them because he thinks the deviation from the theoretically most probable result is so great as to make the record intrinsically incredible. (E. S. Pearson, 1965/1970, p. 331)

This incident set off a good deal of correspondence and discussion among the biometricians. These interchanges contain various references to odds and probabilities beyond which one would be ready to assert that the outcome was unlikely to be chance. Certainly it seems to have been agreed that what we now call the alpha level should have a relatively low value.

But only with the publication of the χ^2 test, the first test that enabled us to determine the probability of occurrence of discrepancies between expected and measured frequencies in a distribution, are indications of specific criteria to be found. Here we see the beginnings of standard rejection levels (i.e., points at which the probability of occurrence is so small as to make it difficult, perhaps impossible, for one to regard the observed distribution as a random variation on the theoretical distribution).

Pearson did not choose one particular value as

the point of rejection. However, from an examination of the various examples of χ^2 calculations presented, with their corresponding probability values, one can see the range within which what might be described as a mixture of intuitive and statistical rejection occurred. The following remarks are from Pearson's paper: $p = .5586$ ("thus we may consider the fit remarkably good" [p. 170]); $p = .28$ ("fairly represented" [p. 174]); $p = .1$ ("not very improbable that the observed frequencies are compatible with a random sampling" [p. 171]); $p = .01$ ("this very improbable result" [p. 172]).

From Pearson's comments, it appears that he began to have some doubts about the goodness of fit at the .1 level ("not very improbable" implies that the results were perhaps *a little* improbable); however, he was obviously convinced of the unlikelihood of the fit at the .01 level. The midpoint between the two is, of course, the .05 level.

William Gosset (who wrote under the pen name of "Student") began his employment with the Guinness Brewery in Dublin in 1899. Scientific methods were just starting to be applied to the brewing industry. Among Gosset's tasks was the supervision of what were essentially quality control experiments. The necessity of using small samples meant that his results were, at best, only approximations to the probability values derived from the normal curve. Therefore the circumstances of his work led Gosset to formulate the small-sample distribution that is called the *t* distribution.

With respect to the determination of a level of significance, Student's (1908) article, in which he published his derivation of the *t* test, stated that "three times the probable error in the normal curve, for most purposes, would be considered significant" (p. 13).

A few years later, another important article was published under the joint authorship of an agronomist and an astronomer (Wood & Stratton, 1910). This paper was essentially to provide direction in the use of probability in interpreting experimental results. These authors endorse the use of *PE* as a measure: "The astronomer . . . has devised a method of estimating the accuracy of his averages . . . the agriculturist cannot do better than follow his example" (p. 425). They recommend "taking 30 to 1 as the lowest odds which can be accepted as giving practical certainty that a difference is significant" (p. 433). Such odds applied to the normal probability curve correspond to a difference from the mean of 3.2 *PE* (for practical purposes this was probably rounded to 3 *PE*).

What specifically determined the adoption of this convention is largely a matter of speculation. Perhaps it was a combination of the preferred use of the *PE* as a measure by early statisticians like Galton and the influence of Pearson and his statements about the unlikelihood of particular results. In any case, it is clear that as early as 1908 $X \pm 3PE$ was accepted as a useful rule of thumb for rejecting differences occurring as the result of chance fluctuations.

Certainly by the time Fisher published his first book on statistical methods 17 years later, $3PE$ was a frequently used convention for determining statistical significance in a variety of sciences that employed statistical tests as experimental tools. For example, an article in the 1925 volume of the *British Journal of Psychology* reports that the chance occurrence of all calculated correlations is "greater than 3 times the *PE*" (Flugel, 1925).

McGaughy (1924) uses the term *critical ratio* for the expression $X/3PE$, where X represents a difference. This, he says, is "the accepted standard for the undoubted significance of an obtained difference between averages" and cites Jones (1921).

Having examined the events preceding Fisher's 1925 publication and remembering the context of his discussion, consideration of his first reference to $p = .05$ quite clearly indicates nothing startling or new, or for that matter arbitrary, about what he was suggesting.

A fact that would have been no surprise to most of those reading his book (and which, indeed, Fisher pointed out) is that "a deviation of three times the probable error is effectively equivalent to one of twice the standard error" (Fisher, 1925, pp. 47–48).

Fisher then cannot be credited with establishing the value of the significance level. What he can perhaps be credited with is the beginning of a trend to express a value in a distribution in terms of its own standard deviation instead of its probable error. Fisher was apparently convinced of the advantages of using standard deviation (*SD*), as evidenced by his remark that "The common use of the probable error is its only recommendation" (p. 48).

Fisher provided calculations for a "probability integral table," from which for any value (described as a function of its *SD*), one could find what proportion of the total population had a larger deviation. Therefore, when conducting any critical test, use of this table necessitated expressing the deviation of a value in terms of its *SD*.

Although, strictly speaking, the conventional rejection level of 3*PE* is equivalent to two times the *SD* (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. Furthermore, it seems reasonable to assume that as the use of statistical analysis was extended to the social sciences, the tendency to report experimental results in terms of their associated probability values rather than transforming them to *z* score values provided a broader base for general understanding by those not thoroughly grounded in statistical theory. In other words, the statement that the probability of obtaining a particular result by chance was less than 5% could be more easily digested by the uninitiated than the report that the result represented a *z* score of approximately 2.

Subjective Probability

How the 5% significance level came to be adopted as a standard has been considered. However, *why* this level seemed appropriate to early statisticians, or why it has continued to prevail in statistical analysis for so long, must be approached not so much from a historical point of view, but from a consideration of the concept of *probability*.

Definitions of the term are most frequently based on expositions of the formal mathematical theory of probability. This may reflect the need to bridge the reality of events in everyday life and the philosophy of logic. Probability in this sense is an objective exercise that uses numerical calculations based on the mathematical theories of arrangements and frequency for the purpose of estimation and prediction.

What often eludes precise definition is the idea that, fundamentally, probability refers to the personal cognition of individuals whereby their knowledge of past experience aids in the formation of a system of expectations with which they face future events. This has been called *subjective probability* to distinguish this notion from its more formal mathematical counterpart.

Alberoni (1962a, 1962b) has conceptualized the intellectual processes that underlie the operation of subjective probability. When individuals cannot find a cause or a regular pattern to explain some differences or variation in the real world, they arrive at the idea of *chance*. This, in turn, forms their expectations for future events. If, however, at some point the events begin to contradict the expectations they have formed, they introduce *cause* and abandon the idea of chance. The point at which this rejection occurs depends largely on the degree of discrepancy and how it is interpreted by each individual. Alberoni refers to this point as the "threshold of dismissal of the idea of chance."

The fundamental questions that remain are straightforward and simple: Do people, scientists and nonscientists, generally feel that an event which occurs 5% of the time or less is a rare event? Are they prepared to ascribe a cause other than mere chance to such infrequent events?

If the answer to both these questions is "Yes," or even "Generally speaking, yes," then the adoption of the level as a criterion for judging outcomes is justifiable.

There is no doubt that the "threshold of dismissal of the idea of chance" depends on a complex set of factors specific to each individual, and therefore varies among individuals.[1] As a formal statement, however, the level has a longer history than is generally appreciated.

---

[1] We have some evidence, based on both formal and informal data, that people, on average, do indeed approach this threshold when the odds reach about 1 in 10 and are pretty well convinced when the odds are 1 in 100. The midpoint of the two values is close to .05, or odds of 1 in 20. One is reminded that these subjective probability norms are congruent with the ideas expressed in Pearson's 1900 publication.

REFERENCES

Alberoni, F. Contribution to the study of subjective probability. Part I. *Journal of General Psychology*, 1962, *66*, 241–264. (a)

Alberoni, F. Contribution to the study of subjective probability: Prediction. Part II. *Journal of General Psychology*, 1962, *66*, 265–285. (b)

Bernoulli, J. *The art of conjecture* (F. Maseres, Ed. & trans.). New York: Redex Microprint, 1970. (Originally published, 1795.)

Bessel, F. W. *Ueber den Ort des Polarsterns*. Berlin: Berliner Astronomische Jahrbuch für 1818, 1818.

Cochran, W. G. Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), *On the history of statistics and probability*. New York: Dekker, 1976.

Cramer, H. *The elements of probability theory*. New York: Wiley, 1955.

David, F. N. *Games, gods and gambling*. New York: Hafner, 1962.

De Moivre, A. *The doctrine of chances* (3rd ed.). New York: Chelsea, 1967. (Originally published, 1756.)

Fisher, R. A. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 1925.

Fisher, R. A. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 1926, 33, 503–513.

Flugel, J. C. A quantitative study of feeling and emotion in everyday life. *British Journal of Psychology*, 1925, *15*, 318–355.

Galton, F. *Natural inheritance*. London: Macmillan, 1889.

Graunt, J. Natural and political observations made upon the bills of mortality, 1662. In J. R. Newman (Ed.), *The world of mathematics*. New York: Simon & Schuster, 1956. (Originally published, 1662.)

Huygens, C. On reasoning in games. In J. Bernoulli (F. Maseres, Ed. & trans.), *The art of conjecture*. New York: Redex Microprint, 1970. (Originally published, 1657.)

Jones, D. C. *A first course in statistics*. London: Bell, 1921.

McGaughy, J. R. *The fiscal administration of city school systems*. New York: Macmillan, 1924.

Pearson, E. S. Some incidents in the early history of biometry and statistics, 1890–94. In E. S. Pearson & M. G. Kendall (Eds.), *Studies in the history of statistics and probability*. London: Griffin, 1970. (Originally published, 1965.)

Pearson, K. Contributions to the mathematical theory of evolution: I. On the dissection of asymmetrical frequency curves. *Philosophical Transactions*, 1894, Part I, pp. 71–110.

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

Quetelet, L. A. *Letters on the theory of probabilities* (O. G. Downes, Trans.). London: Layton, 1849.

Student [W. S. Gosset]. The probable error of a mean. *Biometrika*, 1908, *6*, 1–25.

Walker, H. M. *Studies in the history of statistical method*. Baltimore, Md.: Williams & Wilkins, 1929.

Wood, T. B., & Stratton, F. J. M. The interpretation of experimental results. *Journal of Agricultural Science*, 1910, 3, 417–440.

Yule, G. U., & Kendall, M. G. *An introduction to the theory of statistics* (14th ed.). London: Griffin, 1950.