
SCIENCE RESEARCH
NEWS MAGAZINE OF SCIENCE • FEBRUARY 1968
McGRAW-HILL
William Feller:
too much
faith in
statistics

We heard a lecture you gave recently at Rockefeller University in which you were quite unhappy about what you feel is an unhealthy influence of statistics on the life sciences.

I do not criticize statistical theory as such, or the proper uses of statistics. In fact, statistical theory has developed ingenious methods, which are extremely efficient in various fields. The trouble is that these methods are often used thoughtlessly and routinely by researchers for purposes for which they were not intended and the results are sometimes ridiculous.

In biological experimental work, for instance, a major abuse of statistics has been overemphasis of the role of statistics in evaluating results. The aim of basic research is not to produce statistically valid results but to study new phenomena. An evaluation of experimental findings depends on many factors, such as compatibility with other results, predictions to which it leads and so on—such evidence can rarely be evaluated statistically. This point of view was emphasized by the great pioneer of modern statistics, R. A. Fisher himself.

The purpose of statistics in laboratories should be to save labor, time, and expense by efficient experimental designs. But all too frequently statisticians impose all kinds of nonsensical conditions on the poor biologist or psychologist—conditions which, although they produce unequivocal statistical results, actually hinder him in his research.

Originally the statistical theories for efficient experimental designs, measurements and so on were developed for applications in industry, agriculture and the applied physical sciences, and in these situations they are exceedingly useful. They tell you the most efficient way to go about gathering information to reach a statistical decision—for example whether to use drug A or drug B, or whether to use this kind of sugar beet or that. In these situations, statisticians have learned that you should not use prior information, that you should decide in advance what you are going to do in terms of experimental procedure, and so on.

But in basic research we are concerned with experiments of a totally different kind in which the object is to discover new basic facts that may lead to new insights. Such experiments have little in common with standard routines and must be considered on their own merits. Unluckily, the pattern of the traditional statistical techniques has now penetrated the statistical way of thinking, and statisticians get trained in these methods without understanding to what situations they really apply.

So, when statisticians are working with life scientists in basic research areas, too many assume that the purpose of this kind of experimentation is to produce statistically valid tests for something, while in reality the purpose is simply to discover new things.

To illustrate. A biologist friend of mine was planning a series of difficult and laborious observations which would extend over a long time and many generations of flies. He was advised, in order to get "significant" results, that he should not even look at the intervening generations. He was told to adopt a rigid scheme, fixed in advance, not to be altered under any circumstances.

This scheme would have discarded much relevant material that was likely to crop up in the course of the experiment, not to speak of possible unexpected side results or new developments. In other words, the scheme would have forced him to throw away valuable information—an enormous price to pay for the fancied advantage that his final conclusions might be sustained by some mystical statistical court of appeals.

No statistics should stand in the way of an ex-

A MATTER OF OPINION

Are life scientists overawed by statistics?

William Feller

perimenter keeping his eyes open, his mind flexible, and on the lookout for surprises.

How do you explain that biologists become overawed by statistics in this way?

This attitude is not restricted to biologists—most of us react similarly in ordinary life. Take the procedure used by a consumer research organization to test "consumer preferences for can openers." Surmounting great organizational difficulties, they made comparisons using 4,312 cans. Alas, these cans were identical—round and of a size chosen exclusively for the convenience in *testing*. Now the practical housewife has to deal with cans of various shapes and qualities with high and low edges etc., and her preference will depend on the ease of cleaning and other factors.

Thus, the impressive experiment and its refined statistical analysis have little relevance to the problem at hand, and a few trials by an experienced housewife would have been a safer guide. Yet most of us instinctively reject such "subjective" methods and are awed by the "scientific" approach.

How extensive would you say such an attitude is among the sciences?

That is hard to say. You can get an indication, however, from the rather widespread but preposterous discussion among statisticians about whether or not Gregor Mendel used proper methods—indeed, whether he was honest in collecting data supporting his theory of genes.

The evidence, to put it mildly, sufficed to justify further experimentation, and everybody knows to what results it led. Of course, Mendel lived before the advent of modern statistical methods and he did not support his evidence by an acceptable statistical test. I also admit that Mendel used his judgment as a scientist to omit certain observations which, in modern jargon, he would have attributed to "assignable causes" or disturbances. In other words, he seems to have dropped "outliers."

This is a reasonable procedure and extensively used. We now have good statistical theories about outliers, applicable under certain circumstances. The great astronomer Eddington was disturbed by outliers in his measurements and devised a special theory justifying the omission of some of them. Mendel could have constructed similar explanations but luckily spared us. Anyway, the evidence that he produced was so complicated and manifold that it would not have been accessible to a statistical test. The modern criticism that he did not use proper statistical methods is absurd, and the accusation that he was not honest is a sad commentary on the mentality of those who indulge in such discussions.

Are you saying statistical tests in experimental biology are useless in general?

No, not when they are properly used. A good example is the work of the geneticist Sewall Wright at the University of Wisconsin, who found it worth his while to develop new statistical methods to extract all conceivable information from a comparison of observations with the assumed theoretical models.

But, too often, statistical tests *are* misused. A common abuse is to use a statistical test to try to "prove" a hypothesis. People are always trying to do this, for instance, with the chi-square test, the goodness-of-fit test that is often regarded as a panacea for everything. A *bad* chi-square fit may be a sufficient reason for rejecting a hypothesis but a good fit *taken by itself* is absolutely inconclusive: it may provide additional evidence for a hypothesis that is already plausible for other reasons, but the final judgment must depend primarily on scientific intuition.

This is so because usually many hypotheses (or

William Feller, Higgins professor of mathematics at Princeton, is in a fighting mood over the abuse of statistics in experimental work. Author of a widely used text on the theory of probability and a leading authority on statistics, he has worked closely with biologists as a statistical consultant

mathematical models) are compatible with one given set of observations, and a good chi-square fit will simultaneously "confirm" all such hypotheses. Nevertheless, two hypotheses that are compatible in one situation may lead to entirely different predictions and results in other situations.

Let me illustrate. Some time ago a great fuss was made about the so-called logistic law of growth. All sorts of biological populations, the incomes of towns and countries, the lengths of railroads, the heights of sunflowers, the weight of rats, etc., were supposed to follow this same law and new chi-square fits seemed daily to confirm the universality of this "law."

When Hitler invaded Czechoslovakia, I felt too depressed for more serious work and passed my time by testing this apparent miracle in another way. I constructed two alternative models—a "law of random growth" and a "law of arc tangent growth"—and fitted them to the same material that was supposed to prove the logistic law. Lo and behold, the fit was always as good and often better.

In other words, all the chi-square tests proved not only the logistic law but also the contradictory hypotheses. Since most practical observations refer only to the initial stages of the growth curve, the different hypotheses lead to entirely divergent predictions for the ultimate growth plateau.

Even worse abuses of statistics occur in the perpetual sampling experiments conducted in medicine, psychology, sociology, education, etc. Indeed, it seems that institutes in these areas send their students out to correlate everything in sight—how the liver is related to the brain, how your maiden aunt's weight at birth is correlated to your I.Q., or what not.

Even if no correlations should exist in any of these cases, on the 5-percent level of significance you expect that 5 percent of these experiments will report "significant" correlations on the basis of chance alone. The others are mercifully forgotten, but the scandal is that the "significant" results are published as though they had meaning. This method of gathering "scientific" insights is ridiculous.

Unfortunately, this sin can be compounded. A shuddering example of this is the paper on schizophrenia that you sent me for inspection. In it three psychiatrists compared 77 items of historical data for a population of 28 nonschizophrenic patients and a population of 29 schizophrenic patients in an attempt to identify significant factors in the etiology of schizophrenics. On the basis of a chi-square test, only two items in the list of 77 items showed differences between the two populations greater than would be expected from chance at the 5-percent level of confidence. At that level you would expect more than two items to show differences on the basis of chance alone. Nevertheless, the authors conclude that the differences for these two items are "significant" and that it would be profitable to pay special attention to them in future studies of the causes of schizophrenia. This is sheer nonsense.

Thus, in its concept and design we see that their investigation involves a play with meaningless numbers. Worse than that, even if we forget this theoretical point, we remain confronted by a horrible abuse of statistics and common sense on a different level.

What even the layman can see is that the so-called significance of the two selected items depends on the hazards of tabulation rather than on realities. For example, the first "significant" item, absence of the father, is tabulated as follows:

	Father away: ages 11-15	Father away: other ages	Father never away:	Unknown:	Total:
Schizophrenic	10	5	10	4	29
Nonschizophrenic	2	10	14	2	28

The chi-square test is a measure for the disparity between the two rows, and it is intuitively clear that the main contribution is due to the huge ratio 10/2=5 of the entries in the first column. The classification of a patient as schizophrenic was so uncertain that it varied during the treatment, and the entries in the table represent a compromise between the diagnoses of two or three observers. But if one single schizophrenic is reclassified, the first column changes from 10:2 to 9:3 and the magical "significance" discovered by the author disappears.

Another favorite method of producing "significant differences" is to test only selected portions of the data. The procedure is extensively used in many fields, and is illustrated by an example taken from an early volume of the *Journal for Parapsychology*. To test his extrasensory abilities, a "prominent experimental physicist" made a series of 3,500 card guesses in which the probability of a hit was ⅕. The expected number of guesses was therefore 700, whereas the experimenter obtained 711 hits. The difference of 11 was not judged significant. However, the series was broken up into three subseries according to the time of day and the feeling of the author. He presented the outcome as follows:

	Number of guesses	Number of correct guesses exceeding chance
Morning, well:	2,100	56
Evening, tired:	1,100	—29
Morning, ill:	300	—16

The subseries "morning well" has a significant critical ratio of 2.99, and because of this the experimenter would have us believe that he is endowed with extrasensory perception despite his poor overall performance. In this sense, clairvoyance is, of course, absolutely universal. The trouble is that breakfast or lunch, fitness or illness, coffee or beer, or combinations of these may be required to produce the desired effect—and the nature of the favorable conditions is subject to instant change.

When we contemplate the fantastic successes of the various experimental sciences and the ingenuity and imagination that go into them, then it is saddening that also this black magic passes for art.