REDEFINING THE CONFIDENCE INTERVAL

JOHN C. BAILAR III

University of Chicago
Chicago, IL, USA

Abstract. The confidence interval, its close cousin the p-value, and other statistical summaries of data are widely used in science, including the occupational and environmental health sciences, but they are almost as widely misused. My purpose here is to review some basic matters that should be covered in every introductory statistics course, but are generally omitted. More specifically, some of the many assumptions that are needed in the usual forms of analysis are reviewed.

Key words: Statistics, Probability, Confidence intervals, Data analysis

Standard statistical methods, including the arbitrary selection of 5% (or 95%) in most applications of p-values and confidence bounds, are rooted in laboratory and field studies undertaken to determine how and to what extent things happen. In that context they have been highly productive, and their continued use is sanctioned by decades of scientific progress. They were not developed as tools for making decisions outside the realm of scientific study, but they are often and inappropriately used as a basis for public policy. Methods that are useful for drawing scientific conclusions, where individual findings are embedded in a matrix of other established observations, and new data can be generated if important questions remain, are simply not appropriate as a basis for public decisions.

One may ask why these statistical methods are so widely misused and misinterpreted. I believe it is because they are deceptively simple in concept, easy to apply, and give an air of invincible scientific authority to findings. In short, uncertainty of a nature and degree that is generally accepted in scientific investigation is not appropriate in public policy. These matters have profound implications for the Precautionary Principle: We never know as much as we think we do about either the costs or the benefits of some innovation. This requires a heightened level of sensitivity to the possibility of error, and hence strengthens arguments for caution in adopting innovations.

While the statistician should be deeply involved in planning and conducting many kinds of investigation, formal analysis begins with a collection of observations that we believe are relevant to some phenomenon under study. The proper statistical interpretation of the data depends on using data from a well-characterized and well-understood sampling plan, having the correct mathematical model for the population phenomenon of interest (e.g., that the population data follow a linear model), and of course correct selection and application of the statistical methods and the correct interpretation of their results. Failures in each of these matters are common, and at least a few small failures are always unavoidable, but each failure adds to the variance and/or bias in the findings. Mathematical considerations show that the change in uncertainty is always upward, never downward. The net result is that point estimates may be in serious error, calculated p-values are always too small, and confidence bounds are always too narrow. Sometimes the errors are many orders of magnitude.
I am not objecting to the use of these statistical tools – after all, I am a statistician – but their proper use is to provide a lower limit on the uncertainty in some estimate. In well-designed, well-conducted experiments the additional error may be small. In observational studies it is commonly, and generally unavoidably, large. Thus the person conducting an observational study usually has a major task in understanding all of the potentially important assumptions in the analysis, and in estimating the approximate size and direction of the errors that may be introduced.

The various kinds of observations are commonly divided into one or more independent variables (also called input variables) and a single dependent (output) variable. Most statistical analyses, including t-tests, ANOVA, regressions, and many other things, use what are called least-squares methods, and I will focus on these. Now I will use some mathematical notation, but only for a short time. In least-squares analysis, a model is developed, often just implicitly (as required by some statistical procedure such as a t-test), and generally of the form

\[ Y_i = x_{i1}b_1 + x_{i2}b_2 + \ldots + x_{ir}b_r, \]

where:

- \( N \) is the sample size, the units of study \( Y \) (such as a human, or a mouse, or a piece of equipment) are numbered \( i = 1, 2, \ldots, N \), the mean or “expected value” of the \( i \)-th dependent variable is designated \( Y_i \) and for each \( Y_i \) there are \( r \) independent variables, \( x_{i1}, \ldots, x_{ir} \). The \( b \)'s are unknown constants (if a constant term is required in the analysis, one of the \( x \)'s is set to unity for every \( Y_i \)). More specifically, \( Y_i \) is the expected value of the \( i \)-th outcome variable when its \( x \)'s all have the stated values. This introduces an assumption:

1. **The independent variables are all measured without error.** They need not have a pre-specified “target” value, but their actual values must be used in the computations. Deviations may be nil or virtually nil for some variables, such as gender, or dose of a toxic agent administered, but there is generally some error in many other variables, such as laboratory values of independent variables, or an estimate of the dose of some toxic agent inhaled on the factory floor. The analysis then proceeds by the development of some statistical model that summarizes the data in a probability form, generally either a p-value or upper and lower confidence bounds. A model is always used, even for non-parametric methods, whether or not it appears explicitly in the analysis. We also assume that:

2. The dependent (output) variable is in fact the matter we intend to study in this analysis. This can be approximated be narrowing the target of the analysis sufficiently, but that may not serve our needs well, so this assumption is often violated. For example, unless we are working on new methods of rodent control, we do not usually care much about the effects of some compound on mice; we care about human exposures and outcomes. However, to estimate human risks we may have to study animal responses to some exposure, in which the data are for high rather than low doses, or use data for lifetime exposure when we are interested in intermittent, irregular exposure, or use inbred strains rather than the vast variety of genetic types that characterize wild populations of mice (and humans). These problems are of course widely recognized, and much effort has gone into finding solutions, but the needed assumptions inevitably add some uncertainty to the analysis – often a lot of uncertainty. That uncertainty is not reflected in the ordinary use of p-values or confidence limits. To calculate confidence bounds from the results of a high-dose lifetime animal experiment and convert those bounds directly to the estimation of human risk at low-dose intermittent exposures, requires assumptions regarding the relations between what we have observed and what we want to know. These assumptions are so sweeping that they have been called “heroic”.

The heroism may be somewhat reduced by further assumptions about the dose-response relationship in both animals and humans (both commonly taken as linear, or at least not supra-linear; they are also assumed to have the same mathematical form on some scale, though this assumption is rarely considered). Further assumptions are needed regarding the proper measure of effective dose (generally that effective doses are directly proportional to exposures, or to releases, or to years worked in some plant, etc., and perhaps that effective dose is a power function of weight), and about the effects of variations in dose rate over time.
(generally ignored, which is equivalent to assuming that only total dose over some time period matters). However, these further assumptions do not eliminate the uncertainty required by 2. above, and may in fact introduce even more error. An example is the common assumption that dose-response relations are almost never supra-linear. Our late friend and colleague Cesare Maltoni [1] showed this supra-linearity very clearly for liver angiosarcomas induced by increasing doses of vinyl chloride and more recent work has shown that supra-linear dose-response relationships are quite common, at least at high doses [2]. There is nothing conservative about linear extrapolations from high to low doses. More generally, replacing one or a few big assumptions by a larger number of smaller assumptions is not necessarily progress.

Of course, we cannot observe the expected values defined in the equation above; instead, we have observations that are affected by inherent variation as well as errors of many kinds. Thus in practice the equation must be modified to:

\[ y_i = x_1b_1 + x_2b_2 + \ldots + x_rb_r + \epsilon_i, \]

where:

- \( y_i \) (rather than \( Y_i \)) indicates the actual observation, and the last term \( \epsilon_i \) is the difference between the observation and its expected value, or expectation. In practice, \( \epsilon_i \) is the sum of all the (generally unwanted) influences on the data other than the \( x \)'s, and laboratory experiments are usually designed to keep \( \epsilon \) as small as possible.

Now, what about those \( b \)'s? They are constants specific to the problem, but are not generally observable. They must be estimated, most often from the same data that are under study to learn something about the \( Y \)'s, and since the data have some errors and variations, the estimates of the \( b \)'s will also be imprecise. Thus we need a further modification of the formula above to show that the values of the \( b \)'s are estimated from the data:

\[ y_i = x_1b_1 + x_2b_2 + \ldots + x_rb_r + \epsilon_i, \]

where:

- the primes indicate estimates. Note that the error is now estimated, as well as the \( b \)'s.

It is to get at this last term \( \epsilon_i \) that I have lapsed into this mathematical notation. This is the working equation the investigator must deal with in standard least-squares analysis, whether it ever appears explicitly or merely lurks in the background of a statistical computing formula or a computer program. Standard procedures account for the randomness in the data, but some kinds of error escape that correction. In particular,

3. The value of each \( b \) must be properly formulated and appropriate to the problem. For example, a set of data about some toxic effect in healthy young males may produce values of some or all \( b \)'s that are not appropriate for sick old people of both sexes.

A very common source of problems here is missing data, either for a whole unit of study (such as a worker in a factory) or for some of the independent variables for that unit (such as a job title, or age). An example of such a correlation might be in a study of toxic responses to some workplace exposure, where sensitive persons quickly move on to other jobs; toxicity in the remaining workers will underestimate toxicity in the general population. The absence of data, especially when absence is correlated with some \( b \) or with \( Y_i \) or \( y_i \), can be very serious. Missing and incomplete data can upset the most stable of applecarts.

The standard statistical least-squares analysis rests on four mathematical assumptions about the \( \epsilon \). These do not apply to non-parametric statistical methods, but non-parametric methods have their own assumptions and limitations, which I cannot discuss here, as well as much reduced statistical power. The trade-off here is that when a parametric model is at least approximately correct, the model can add much strength to the analysis; but when the violations of assumptions are larger, the parametric model can lead to serious, even disastrous, error. Violations of the assumptions in non-parametric analysis are likely to have smaller effects, but that benefit comes at a high cost in statistical power. I will discuss the four mathematical assumptions in least-squares analysis one by one. First is:

4. Every \( \epsilon_i \) has a mean (or expected) value of zero. That is, the values of \( \epsilon \) over a lot of repetitions of the observation with these values of the \( x \)'s would tend to average out
to zero. What this means in practice is that the model must be correct with respect to the relationships among the variables in the analysis. \( Y_i \) must in fact be the linear combination of \( x \)'s specified in the formula above. But the relationship may not be linear. There are two good ways to get around this problem – to use one or another form of curvilinear regression, or to transform the data (as in the frequent logarithmic transformation of dependent variables) so that they become closer to linear. Transformation is easier, as well as more familiar, and often appropriate, but has the additional effect of transforming the e's. That may be good or bad; in many situations where the data are right-skewed and call for a logarithmic transformation, the variance of the observations (that is, the variance of the e's) increases with the value of \( Y \). In this case, the transformation may also improve the statistical properties of the e's (see just below).

5. The variance of the e's is the same for each of the values of \( Y \). There are statistical means for dealing with known variation in the variances, but they rely on still further assumptions. This assumption is not vacuous, and violations sometimes lead to serious consequences. For example, if a laboratory process goes out of control (perhaps unknown to the investigator) the variance may increase substantially for many but not all \( y \)'s. Or, there may be outliers in the data – that is, observations with different means or different variances that really do not belong there, such as data from a contaminated sample, or from a subject in a drug study who is careless about taking his/her doses of the drug. Problems can be especially serious when variance depends on the values of one or more \( x \), or on \( Y_1 \).

6. The values of the e's are statistically independent. Violation of this assumption can also have serious consequences. An example is a drug study in which the serum creatinine of each subject is measured many times, and all of the values are pooled for study without adjusting the procedure to account for the fact that within-subject variances will be smaller than the variance across the whole pool of data. Unexpectedly high p-values (e.g., 99.9%) or unexpectedly narrow confidence bounds are often a symptom of failure to consider and account for lack of independence among the e's.

7. The statistical distribution of the e's is normal (or Gaussian – a bell-shaped curve of a quite specific mathematical form, not just any more or less bell-shaped curve). Most major computer packages of statistical procedures offer a means to examine this assumption, but investigators often skip that step, perhaps because they do not realize its importance. Highly skewed distributions of e's, which are especially common with highly skewed \( Y \)'s, can seriously undercut the accuracy of calculated p-values and confidence limits.

Assumptions 4. through 7. are the basis of statistical theory about p-values and confidence limits, but fortunately most of the standard procedures are not affected much by modest deviations from perfection, which usually have little effect on the calculated p-values or confidence limits (that is, the procedures are “robust”). More serious deviations are a matter of concern, and the concept of robustness does not apply in the same way to any of the other assumptions discussed here.

Now I am done with the mathematics, but there are still some other important assumptions to consider. One is:

8. The chosen probability level (commonly 5% for p-values and 95% for confidence bounds) is appropriate for the problem. First, a 5% chance of an error leading to a major catastrophe is likely to be quite unacceptable. These figures of 5% and 95% are entirely arbitrary. They arose early in the development of statistical theory, and their justification has always been that they have worked well in practice to sort out real effects from “noise”. However, that favorable experience has been largely in the analysis of designed laboratory experiments, where the consequences of making a wrong decision may be small and where independent replication of an experiment (or of some of the observations) is feasible if the results call for further study. No such justification can be advanced for using these probability levels in most other uses of statistical methods, including studies of occupational and environmental health, such as large observational studies of populations, or clinical studies of some rare exposure or rare disease where the first study used all of the available
patients. It may not even be appropriate for a long-term carcinogen bioassay, where tight protocols can control many sources of random variation, but the resources and time required for replication make independent validation in a new study unlikely even if the results seem to require further investigation.

In choosing probability levels, investigators often face a cruel choice. They must balance false positive results (from lenient standards that find a lot of real effects, but also appear to find a lot of effects that are not really there) against false negative results (of failing to find effects that are in fact there). There is no way to escape this choice – if everything else is done correctly, pushing down the false positive rate inevitably raises the false negative rate, and vice versa. My assessment, after a career in applied statistics, is that false positives are more common and more severe than false negatives, so that we should often use smaller values of $p$ in interpreting $p$-values (such as $p < 0.01$, or sometimes even smaller) and higher $p$’s in confidence limits (such as 99% confidence limits, or even wider), but there are exceptions, especially when no additional data are likely to be forthcoming.

Confidence limits and $p$-values are mathematically linked: a $p$-value is “statistically significant” if and only if the matching confidence limits exclude the value specified by the null hypothesis. However, they are based on different views of the problem. A $p$-value uses the temporary assumption that the null hypothesis is true, and measures the probability that (if everything else is as specified) one would obtain the result actually observed or a more “extreme” result. What is more extreme can sometimes be debated, but is usually evident from the problem.

It is critical to understand that the $p$-value does not measure the probability that the null hypothesis is true – it measures the likelihood of the observed outcome (or something more extreme) if the null hypothesis is true, and its correct interpretation involves many things, including our prior assessment that the null hypothesis is in fact true. The probability that the observed data indicate a true null hypotheses and the probability that a true null hypotheses would produce the observed data are entirely different matters. In an analogy, the probability that my coin will turn up heads when the coin is two-headed is not the same as the probability that the coin is two-headed if it turns up heads.

Rare events do happen, with precisely the predicted frequency (such as 5% for the usual $p$-value) and the analyst must assess the likelihood that a statistically significant result is a “real” finding vs. one of those uncommon events. Confidence limits use no assumption about the truth of the null hypothesis, but (again, if everything else is as specified) are calculated so as to have the specified probability of including the unknown parameter. Since the confidence bounds are calculated from the data, they are random points and define a random interval. A repetition of the study would produce a different random interval, defined by different confidence limits. Bayes’ Theorem, which does deal with the probability that the null hypothesis is true, can be a great help here, but that topic is outside the scope of this paper.

Some comment on one-tail $p$-values is in order here. These are variations on the usual statistical procedures that effectively double the chance of finding a “significant” result. Nearly every instance of the use of one-tail $p$-values that I have seen has violated two cardinal rules of interpretation. First, the intention to use one-tail tests must be spelled out in writing, including the specific procedures to be used, prior to the collection of the data. Second, we must be told, also ahead of time, what the investigator would do if the findings come out “too good to be true”; that is, if they point in the direction opposite to what was expected or hoped for. An example is a $p$-value of 99%. If the results would be accepted as simply strong confirmation of the null hypothesis, there may be no problem. If the unexpected findings would be taken as indicating some important and unexpected phenomenon, perhaps published as such (and perhaps even with a one-tail $p$-value in the opposite direction!) the investigator is using statistics to lie to his readers. The difficulties with one-tail procedures are so great that I recommend that they never be used.

More broadly, there are places for the ordinary $p$-value (such as when a model is tested by study of the distribution of $p$-values over a large collection of data sets) [2],
but they are uncommon. Confidence limits include all of the information in the corresponding p-values and a lot more, and are almost always to be preferred. I refer here to confidence limits that are calculated correctly given the underlying statistical model. Such absurdities as a negative estimate of a dose, or a confidence range that includes values greater than 1.0 for a correlation coefficient, are a result of a poor statistician, not poor statistical theory.

9. **The analyst has honestly and correctly presented everything relevant to the analysis.** Problems here are, I fear, very common. There must be no covering up of problems in the data or the analysis. The investigator must state, in writing, and before the data have been seen (which generally means before they have been collected) what primary hypothesis or small number of hypotheses will be examined, and precisely how they will be examined. There must be no *post hoc* hypotheses presented as if they were what the investigator was planning to study, no fiddling of p-values to obtain (or less commonly, to exclude) a “statistically significant” result, no scanning of pages of computer output to see which statistical procedure will be quoted because it gives the most “favorable” result, etc.

Science is not a game, with points for the most significant p-values. In short, the investigator must tell the reader everything the reader should know to make an independent judgement about the reliability of the data and the analysis, and the author must explain his reasoning in coming to the stated conclusions. Of course, one cannot present every last detail, or a paper would grow to the size of a book and become unreadable, so some judgement is needed in what to present. Perhaps the best guide here is what the author himself, in his most critical mode, would want to know if someone else had done the work.

I am not arguing against exploratory and imaginative analysis of data to discover what may be there beyond the point of initial interest. That is in fact how most of the “big” discoveries in science are made. However, exploratory analysis must be clearly labeled, because the absence of a well-defined prior hypothesis violates the rules of ordinary statistical procedures, and neither p-values nor confidence bounds have their usual meaning. Interesting findings from exploratory analysis must usually be validated in other ways, generally by the generation of new data designed to test a specific hypothesis about the unforeseen result.

10. **There have been no blunders in the whole of the analysis and interpretation.** Examples of such a blunder are the mislabeling of a variable, or the use of the wrong units in the analysis. Errors in data input were responsible for a large part of the confusion in the outcome of the Florida vote in the US 2000 presidential election [3]. I have commonly seen such problems in technical reports and draft papers, though they seem to be rare in published papers. Perhaps editors and reviewers tend to find them before publication. But wouldn’t it be better to avoid the embarrassment of retracting and redoing some work by making sure that it is done right the first time? There are other, usually lesser, kinds of general assumptions in statistical analysis, and of course additional assumptions may be needed that are specific to any particular analysis. The import of what I have covered here should, however, be clearly understood. The variability that is expressed in ordinary p-values or confidence bounds is only a part of the total uncertainty in any analysis – often a very small part. The investigator must always make a serious effort to understand how far the analysis may be in error because of the violation of the (generally hidden) assumptions. Statistics is not a matter of putting a lot of data into a computer, running some computer program, and taking the printout at face value.

It is time to return to the title of my talk here: Redefining the Confidence Interval. Confidence intervals and p-values provide absolute minimum estimates, and they are estimates, of the uncertainty surrounding statistical results, and we should always consider how much greater the real uncertainty may be. There are ways to do this in a more or less formal manner [4], but they are not widely known, they are expensive in investigator time, and they still will not capture the whole of the uncertainty.

What does this mean for the Precautionary Principle? It means a great deal. Statistical results must always be regarded as providing a minimum estimate of error and variability, and an extra margin of error – often large – must
be added to account for uncertainty in the results that is not captured by the statistical analysis.

In summary, we never know as much as we think we do. Standard statistical procedures measure one kind of uncertainty in the data, and one kind only – that arising from randomness in the data under a specified model. They do not account for many other kinds of uncertainty and error, including error in the model itself. Statistics is the art of interpreting quantitative data; it is not a science and most certainly not a mathematical science, though it uses mathematical tools (as do other arts and sciences). And the good applied statistician must be an artist, closely familiar with his tools, knowing which tool to use when, and understanding what, if anything, might be done to reduce the total uncertainty in the analysis at hand.

REFERENCES


