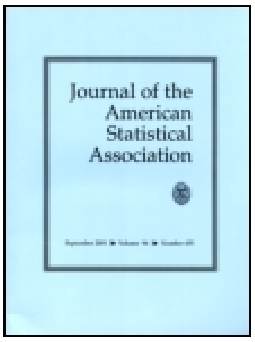

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Errors of the Third Kind in Statistical Consulting

A. W. Kimball

To cite this article: A. W. Kimball (1957) Errors of the Third Kind in Statistical Consulting, Journal of the American Statistical Association, 52:278, 133-142

To link to this article: https://doi.org/10.1080/01621459.1957.10501374

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 278 JUNE 1957 *Volume 52*

ERRORS OF THE THIRD KIND IN STATISTICAL CONSULTING

A. W. KIMBALL
Oak Ridge National Laboratory

Because graduate students in statistics are given little, if any, preparation for actual consulting, they are prone, particularly in their early years, to commit errors of the third kind, many of which could be avoided if the students were properly trained. Errors of the third kind are defined and are illustrated with actual examples from consulting experience. The cases used represent types of error which result from different situations that arise frequently in practice. Some discussion is included of possible remedies for this problem that are suggested by the experience of educators in other fields.

INTRODUCTION

At a relatively early age in their graduate academic life, students of statistics become familiar with certain risks associated with what they come to know as the first and second kinds of error in the theory of testing hypotheses. They soon learn that in many widely used statistical tests the first kind of error is easy to control but that often the risk of the second kind of error is difficult to compute and more often neglected entirely in practice. The importance of these errors is constantly brought to their attention through emphasis in their course work on such things as uniformly most powerful tests and sequential procedures which control the risks of both kinds of error. More recently the theory of decision making, the natural sequel to hypothesis testing, has elevated the notion of risk to an even higher place in the hierarchy of ideas passed on from professor to student.

As a result of these teachings many of today's statistics graduates come away from the warm comfort of university complacence into the coldly realistic outside world imbued with the idea (and probably rightly so) that the statistician's only real function in this world is to compute risks of error for other people who have to make decisions. To be sure, there is a vast amount of planning (design of experiments, model building) and intermediate adjustment (missing data, extreme observations) necessary before the statistician can estimate these risks, but essentially this is his main task, and the student finds it out usually before the end of the first semester.

Consider then the embryo statistician who has been released from the university's uterus with a shiny new degree and who proceeds on his mission as a risk computer fully equipped with the tools of his trade and the mental wherewithal to apply them. Let us assume that during the first few years of his initiation as a consulting statistician he is lucky, from a mathematical statistics point of view, and computes correctly the risks of error for all problems he tackles. The chances are, speaking nonmathematically, that during this time he will commit the third kind of error more often than he or anyone else realizes. What is even more tragic is that, although as a student he was constantly reminded of the importance of the first two kinds of error and duly vowed always to keep sight of them, he was probably never made aware of the *existence* of a third kind of error, let alone told what to do about it.

The purpose of this paper is to draw attention to the third kind of error by quoting actual examples in which the error was made and later rectified. The hope is that the paper will serve simultaneously as a warning and as a moderator for newly trained consultants who tend to descend on research workers with the sometimes frightening enthusiasm and confidence of a freshman at his first football practice, and that perhaps it will help stimulate responsible educators to move more rapidly in filling this wide gap in graduate statistics training. Most conscientious teachers of statistics recognize this need and are searching for effective methods of correcting the situation, but very little real progress has been made.

In this connection there is an interesting analogy between graduate statistical training and medical training. The physician of today, after he completes internship and residency, is well trained to practice medicine but not so well trained to do research. This fact is recognized by many schools in which the M.D. who wants to do research in physiology is advised to get a Ph.D. in this field after he completes medical school. The emphasis in medical school is on practice since most medical graduates never see the inside of a research laboratory. The graduate statistician, on the other hand, is for the most part well trained to do research in statistics but more often than not inadequately trained to "go into practice," that is, to do statistical consulting. A safe guess is that over half of the graduates in statistics each year are lured into industry or government where their principal work is consulting, and those who do go to universities frequently find their nonteaching time fully occupied with consulting both on and off campus. It is of utmost importance, therefore, that the third kind of error in statistical consulting be emphasized and brought out into the open. Otherwise nothing may ever be done about it.

THE ERROR OF THE THIRD KIND

A simple and almost ludicrous definition of the error of the third kind is *the error committed by giving the right answer to the wrong problem.* In defining it this way we are allowing the statistician the benefit of the doubt by rejecting the possibility that he would give the wrong answer to the wrong question. We are also protecting ourselves against the occurrence of a false positive, that is, the situation in which the wrong answer to the wrong problem turns out to be the right answer to the right problem. At this point the reader who finished the introduction without succumbing to the temptation to look ahead for a definition may well feel like the reader of a murder mystery who on the last page discovers that the victim committed suicide. Why, he may ask, should we concern ourselves with any consulting statistician who could be stupid enough to commit such an error? Admittedly, there may be many mature statisticians who prefer to take this attitude rather than face the consequences of accepting its alternative. If this is so, the situation is indeed a grave one.

There is no way of knowing how many of us, particularly in our early years as consultants, were guilty of errors of the third kind, but it is almost certain that few have escaped an occasional mistake of this nature. The reason is simple enough. Many of us, in good faith, have helped research workers make t-tests, or compute analyses of variance, or design experiments thinking we were giving the right answer to the right problem; and usually we do give the right answer to the question that is asked. Unfortunately it often happens that the question asked has little bearing on the real problem, and we are led into committing the third kind of error.

A stranger to the intimacies of statistical consulting might well doubt that such ridiculous events could ever occur, but the experienced statistician knows that they do occur and will probably never be completely eliminated. Basically, errors of the third kind are caused by inadequate communication between the consultant and the research worker. In some instances, the research worker is at fault for failing to discuss his problem in complete perspective. He may feel that the statistician is weak in the subject matter field and that any attempt at a complete explanation would be a waste of time; or he may not have his ideas completely crystallized and may not want to be "confused" by a mathematician; or he may know a little statistics and feel that he can state the question adequately himself; or he may simply not want to take up too much of the consultant's time. At the same time the statistician is at fault for not becoming sufficiently familiar with the problem to enable him to advise intelligently. With proper preparation, sufficient patience, and persistent questioning of the experimenter, the consultant should be able to avoid most errors of the third kind, but not until he recognizes that they exist. In the next section an attempt is made to show that such errors can happen and under circumstances that ordinarily would not be regarded as unusual or bizarre.

EXAMPLES OF ERRORS OF THE THIRD KIND

The material for these examples is drawn for the most part from the author's own experience, with the natural result that most of the problems come from the field of biology. The main theme of the paper, however, is not biological and except for weakness in the subject matter field, either on the part of the author or the reader, the message should be clear. It should not be inferred that the errors illustrated are necessarily those of the author, although he would not deny this possibility.

Example I. An engineer was engaged in particle size determinations in connection with corrosion studies. He wanted to estimate the particle size distribution, which he was willing to assume normal, but his method prevented him from observing particle sizes below a certain diameter. He knew very little about statistics but he had heard that there were ways of estimating distributions when samples are restricted. There was no statistician in his own group to whom he could turn for help, but there was one nearby who, although very busy, might give him a reference.

So he visited the statistician and presented him with the following sample of particle sizes: 25.6, 7.1, 5.1, 4.2, 3.7, 3.0, 2.6, 2.0, 1.8, 1.6, 1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7—and pointed out that his method would not allow him to determine particle sizes less than 0.7. Assuming the distribution normal, he wanted to know how he could estimate its mean and variance. The statistician was indeed quite busy and not inclined to spend much time on a problem he knew very little about and which did not originate in his group. On the other hand he did not want to cause any ill feelings by refusing to give any help at all. An easy way out was simply to hand the engineer one of his many reprints on truncated normal distributions (after all the engineer had asked for a reference), and this he did. Both participants in this short conference went away happy, the engineer because he thought he had an answer to his problem and the statistician because he disposed of an uninteresting problem in short order. But, as any reader who carefully inspected the "sample" of particle sizes already knows, an error of the third kind was committed. It might easily have gone unnoticed indefinitely, as do many others, but fortunately this error was caught.

The engineer returned to his desk armed confidently with the newly acquired reprint and began to apply the method with the help of his 1935 model calculator. He had not gotten very far along before he found that one of the statistics he computed was far outside the range of a key table given in the reprint to facilitate solution of the equations. After checking for and finding no arithmetical inaccuracies, he reluctantly returned to the statistician who inwardly was not too happy to see the engineer back. This conference lasted longer than the first, and with great chagrin the statistician finally realized what a stupid blunder had been made.

Among the methods used in particle size determination is one known as the sedimentation method. Briefly, it consists of the preparation of a liquid suspension of the material to be analyzed and the measurement of the decrease in concentration of particles at or above a particular level in the suspension as sedimentation proceeds. Under suitable conditions, Stoke's law can be used to compute the percentage of particles in the suspension having diameters greater than d, say, where the value of d is determined by the time elapsed after sedimentation starts. Thus the random variables are the percentages, and d is a fixed or independent variate. It was this technique that the engineer had used. The appropriate method of estimation is, of course, probit analysis or one of its counterparts, and the "truncation" is not a problem except insofar as it increases the errors of estimate.

If the statistician had been familiar with particle size methods, or even if he had carefully scrutinized the "sample" that was presented to him, the error could never have occurred. It might be argued that both parties to this near-fiasco were the victims of circumstance and not really responsible, but if we are honest we must admit that the statistician has a duty to be more careful in avoiding this kind of error than perhaps any other. If he commits an error of the third kind, he is no less at fault than the physician who inadvertently administers arsenic instead of aspirin.

Example II. A geneticist working in the field of radiation biology became interested in the relative biological effects of different kinds of radiation. In one experiment he hoped to compare the effects of gamma radiation and neutron radiation by exposing two groups of organisms separately to graded doses of each kind of radiation and then determining the frequency of mutations at each dose. In previous experiments it had been found that mutation frequencies increase linearly with dose, so he planned to evaluate the relative biological effect by a comparison of the two slopes for the two kinds of radiation.

After the experiment was completed, he visited a newly hooded statistician and asked him to estimate the two slopes and make a statistical test of the difference between them. He explained that the gamma source used in the experiment was radioactive cobalt which provided an essentially pure source of gamma rays, but that the neutron experiment was carried out in a cyclotron and he had "corrected" the neutron doses for a known gamma ray contamination of about 7 per cent. The young statistician, who had little or no experience with radiation experiments and who at the moment was not particularly interested in learning about radiation, proceeded promptly and, as it turned out, rashly with his analysis. From the biologist he had obtained the following data:

Gamma experiment $(i=1, \cdots, n)$
y_i = proportion of mutations
x_i = dose of gamma radiation

Neutron experiment $(j=1, \cdots, m)$
u_j = proportion of mutations
v_j = "corrected" dose of neutron radiation.

Originally there were several replications at each dose point and the statistician had carefully tested for homogeneity. Finding no significant departure from binomiality, he pooled the replications and proceeded with a weighted linear regression for each experiment. He ended up with the two equations

$$\hat{y} = a + b_\gamma x$$
$$\hat{u} = a + b_n v,$$

for the gamma and neutron experiments, respectively. Finally he made the requested test of significance and chalked up (he thought) another successfully completed problem.

The third kind of error made by this statistician was most certainly avoidable. He had only to question the geneticist about the nature of the "correction" of the neutron dose, and without having to learn much at all about radiation dosimetry, he would have discovered his error. The consulting statistician, particularly in the physical science and engineering fields, soon learns to question any "corrections" applied by the experimenter before the data are presented for analysis. In the problem at hand it turned out that the geneticist had simply reduced the original neutron dose by 7 per cent intending thereby to evaluate the effect of neutrons uncontaminated by gamma rays. Overlooked was the fact that the corresponding biological effect still included the gamma component. When the error was uncovered, a somewhat different approach was taken. The two experiments were analyzed simultaneously by minimizing

$$\sum_{i=1}^{n} \lambda_i(y_i - \hat{y}_i)^2 + \sum_{j=1}^{m} \nu_j(u_j - \hat{u}_j)^2,$$

where

$$\hat{y}_i = a' + b_\gamma' x_i$$
$$\hat{u}_j = a' + b_\gamma'(0.07w_j) + b_n'(0.93w_j),$$

where the uncorrected neutron doses (w_j) were determined from the relation, $v_j = 0.93\ w_j$, and where λ_i and ν_j are the appropriate weights. Needless to say, the second approach yielded estimates and standard errors somewhat different from those of the first approach, and the new significance test had to allow for the covariance between b_γ' and b_n'.

Once again in this example the blame must rest primarily with the statistician. Perhaps in his eagerness to apply his newly acquired skills to a problem which he thought fell into a pattern he had seen in graduate school, he temporarily lost his common sense. Whatever the explanation it is hard to draw any conclusion other than one which reflects the fact that he was just not ready to do statistical consulting on his own.

Example III. This example illustrates in a sort of general way a situation which must occur many times in the life of every consulting statistician. It might be called "Consulting by remote control," or "Communication without representation." Frequently the situation arises in a manner similar to the one in this example.

A research worker who, mostly through experience, had become fairly adept with many text-book statistical methods, encountered a problem which was new to him and which he could not find in his elementary text-book. He had computed two product-moment correlation coefficients and wanted to test the hypothesis that the population correlations were equal. He was reasonably sure that the t-test would not be appropriate, but he was also sure that some method must exist. The research organization to which he belonged did not employ a statistician, but he had a statistician friend in the same city who he felt would certainly have the answer. For such a minor problem the trip across town was hardly worthwhile, but thanks to Alexander Graham Bell, he knew he could solve his problem without leaving his desk. The phone call was made and the statistician, not wanting to be impolite or difficult by suggesting a meeting in person, and being allergic to long telephone conversations, quickly told his friend about the z-transformation and where to find an example of its use.

Sometime later both men happened to attend the same local seminar, and upon seeing his friend, the research worker rushed over to thank him for the useful advice about the z-transformation. During the course of the conversation, the statistician discovered to his horror that the experimenter had taken N simultaneous observations on three mutually correlated variables, x, y and

z, and the two correlation coefficients which had been the subject of the aforementioned telephone conversation turned out to be the correlations between x and z and between y and z. With much embarrassment he realized that he had recommended a t-test between two z-transformed correlation coefficients which were not independent. Summing up all his courage he confessed his mistake and referred the experimenter to the paper by Hotelling [1] in which it is shown that under the null hypothesis, $\rho_{xz} = \rho_{yz}$,

$$t = \frac{\sqrt{N-3}(r_{xz} - r_{yz})\sqrt{1 + r_{xy}}}{\sqrt{2D}}$$

is distributed approximately as "Student's" t with $N-3$ degrees of freedom, where

$$D = \begin{vmatrix} 1 & r_{xz} & r_{xy} \\ r_{xz} & 1 & r_{yz} \\ r_{xy} & r_{yz} & 1 \end{vmatrix}.$$

The experimenter tried to accept the blame for this mistake contending that he should have taken the time to explain the actual problem more completely. Actually in this error of the third kind it would appear that both parties were at fault and for essentially the same reason—neither wanted to take the time to find out what the other was really doing.

Example IV. It seems desirable to include, as one of the examples of errors of the third kind, an error of omission. Essentially these errors occur when the statistician fails to do the best job possible simply because he has not taken enough time to question the research worker thoroughly about his experiment. In these cases, the answer given is often the right answer to the right problem but not always the *best* right answer. The following example illustrates an error of this kind.

A geneticist was engaged in a series of recombination experiments with bacteriophage T4. He was interested in testing for independence of the occurrence of two markers, r and tu. Under the hypothesis of independence, in an experiment in which plaques are counted for all four types of progeny, the observed and expected plaque counts can be represented as shown in the following table:

PLAQUE COUNT FREQUENCIES

Frequency	Type of Progeny				Total
	Parental	r^+	tu^+	r^+tu^+	
Observed	a_1	a_2	a_3	a_4	M
Expected	Mq_1q_2	Mp_1q_2	Mq_1p_2	Mp_1p_2	M

where p_1 and p_2 are the probabilities of events leading to recombinants r^+ and tu^+, respectively, and $q_1 = 1 - p_1$, $q_2 = 1 - p_2$. Typical experiments of this type yield about 90 per cent of parental type progeny and 10 per cent recombinants.

The geneticist who was doing these experiments had had some experience using chi-square in testing for independence with genetic frequency data, but since there were two parameters to be estimated in this case, he was not quite sure how to proceed. So he visited a young biometrician and presented him with data of the type shown in the above table. After explaining the experiment, he mentioned casually that he had much more data from another replication of this experiment but that it would probably be of little use since not all of the four classes of progeny were counted.

Perhaps it was too early in the morning, or perhaps the biometrician had his mind on something else. In any event he ignored the experimenter's casual remark about the other replication, proceeded to obtain maximum likelihood estimates of the parameters p_1 and p_2 from the complete experiment and correctly computed a chi-square wlth one degree of freedom which provided the required test for independence.

The results of the test were somewhat inconclusive, at least in the mind of the experimenter, and he began to reflect on why he had done the second replication in the first place. The greatest labor in experiments of this type is the counting of plaques, and since about 90 per cent of them represent parental type progeny, most of the work is done in counting plaques which provide little information about independence. It seemed reasonable to him, therefore, to do an experiment in which only the recombinants were counted. This was the second replication which he had mentioned to the statistician and it was about twice the size of the first.

With these points in mind he returned to the statistician and asked specifically if there wasn't some way in which the information from the second replication could be combined with the first so as to provide a more sensitive test for independence. As a result of this gentle prodding by the experimenter, who was obviously thinking more clearly than our young biometrician, an approach was found which would make use of all the data. The result of the second experiment was representable as:

PLAQUE COUNT FREQUENCIES

Frequency	Type of Progeny				Total
	Parental	$r+$	$tu+$	$r+tu+$	
Observed	—	a_5	a_6	a_7	N
Expected	—	$\frac{Np_1q_2}{(1-q_1q_2)}$	$\frac{Nq_1p_2}{(1-q_1q_2)}$	$\frac{Np_1p_2}{(1-q_1q_2)}$	N

Under the hypothesis of independence the joint probability of both samples is

$$\frac{M!}{a_1!a_2!a_3!a_4!}(q_1q_2)^{a_1}(p_1q_2)^{a_2}(q_1p_2)^{a_3}(p_1p_2)^{a_4}$$

$$\times\frac{N!}{a_5!a_6!a_7!}(1-q_1q_2)^{-N}(p_1q_2)^{a_5}(q_1p_2)^{a_6}(p_1p_2)^{a_7}.$$

The maximum likelihood equations for p_1 and p_2 can be reduced to a quadratic equation in p_2 with only one admissible root, and an equation in p_1 which is linear in p_2. A chi-square with three degrees of freedom is then easily computed. In this particular experiment the added strength of the second replication was sufficient to convince the geneticist that he had no reason to suspect lack of independence, whereas the significance level of chi-square based on the first replication alone had left him in doubt.

Perhaps there are only a few young statisticians who would commit an error of this kind, but the temptation must be great in many practical situations for the new consultant to discard extra observations which make the pattern of an experiment look different from what he has been accustomed to seeing in class examples. We so often hear it said that many research workers never come to the statistician until after the experiment is completed, and that frequently much of the data is worthless for statistical analysis. Certainly this does happen more often than it should, but in many apparently hopeless cases it also happens, as in the foregoing example, that a little extra effort on the part of the consultant will yield a workable, relatively simple method of analysis. A feel for these situations comes only with experience, but the graduate student should be given a chance to get some of this experience before he starts out completely on his own.

A POSSIBLE SOLUTION TO THE PROBLEM

Many readers may object to the examples which were chosen to illustrate errors of the third kind as being unrealistic and unlikely to happen in actual practice. To a large extent they are right because all of the errors discussed were eventually corrected and hence no longer qualify as errors. But it should be obvious that the only errors of the third kind which become known are those which are corrected, and for every one which is corrected there must be many which we will never know about. If we are ready to admit that these errors are committed and perhaps in large numbers, then we should also be ready to do something about it.

The obvious place to start is in graduate schools where degrees in statistics are awarded to students who expect to do statistical consulting. For some time to come these institutions will provide the largest part of the supply of consulting statisticians. If the consulting statistician were required by law to obtain a license before he could go into practice, we could take our cue from the medical profession. Every statistics graduate who expects to consult would be required to intern for, say, one year, and at the end of this time would be required to take an examination to obtain his license. This arrangement might or might not prove satisfactory but most people would admit that it is not practicable, at least not in the foreseeable future.

Let us turn then to the teaching profession. In many states licenses to teach are either not required or can be obtained merely by payment of a fee, and the teachers colleges, in addition to providing a comprehensive curriculum of course work must somehow prepare students for actual teaching. They accomplish this by the long established requirement of practice teaching. Every conscientious teachers college includes as part of its curriculum a period in which the student leaves the campus and under the direction of an experienced teacher learns to teach by teaching. In some schools practice teaching begins at the junior level, and college administrators have found that there is absolutely no substitute for it. Why then should not the statistics student be required to learn to consult by consulting?

Some statistics departments have attempted to achieve this goal by having the student "sit in" on consultations held by members of the staff. This undoubtedly helps to some extent, but frequently the student participates very little in the discussion and some staff members complain that their clients are reluctant to talk in the presence of graduate students. Whereas attendance at staff consultations may serve to introduce the student to the complexities of consulting, he can never learn to cope with them until he tries it on his own. To achieve this opportunity it is imperative that he leave the campus and "intern" in the field.

Exactly how this can best be accomplished is anybody's guess. As a start it would seem that graduate schools should attempt to obtain affiliations with consulting groups in government and industry, much as medical schools are affiliated with hospitals, or teachers colleges with practice schools. Universities contribute heavily to government and industry through the medium of the research contract. Both parties benefit, of course, even under the present system, but certainly both would benefit more in the long run if programs of student participation could be arranged. There must be many instances in which essentially this sort of arrangement has been made and proved successful, but only for an isolated student here and there. To be really effective such a program would have to be made an integral part of the graduate curriculum and listed in the catalog as one of the requirements for a degree.

Those of us in the profession of statistical consulting who take honest pride in our work face a real challenge. Two avenues are open to us. One is to ignore the presence of this situation and to continue along our narrow paths of individual self-satisfaction, oblivious of the effect it might have on the future of our profession. If this course is followed, when the production rate of new statisticians begins to catch up with the demand, we will face loss of prestige and public confidence, and possibly even virtual extinction. The other avenue is to recognize the problem, to appreciate that it is constantly increasing in intensity and to push hard for positive action as soon as practicable. We should have begun yesterday; today we are only thinking about it; tomorrow we must act.

REFERENCE

[1] Hotelling, Harold, "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," *Annals of Mathematical Statistics*, 11 (1940), 271–83.