Estimation of Upper Limits from Experimental Data

Virgil L. Highland

July 1986, Revised February 1987

Temple University

Philadelphia, PA 19122

Report COO-3539-38

Addendum

Since the initial version of this report was written, the Particle Data Group has given some specific advice about the confidence limit problem discussed here (Physics Letters 170B, 55(1986)). They indeed endorse Method 4, the Bayesian method. They do, however, add a caveat that the method can give absurd results if applied thoughtlessly, as has been emphasized in this paper. At any rate there is now an authoritative (though still very brief) reference to which one can refer when wishing to explain his derivation of a confidence limit. If the limitations are respected, this reference would bring some uniformity to the field and this report would become superfluous.

I have also spent some time studying the mathematical foundations of the likelihood method (using Kendall and Stuart) as applied to limits on branching ratios. I have convinced myself that the worries expressed above about possible failure of the likelihood theorems in these cases are unfounded.

Upper Limits

Virgil L. Highland
Temple University

A. Introduction

An unsystematic survey of the physics literature concerned with the measurement of upper limits to small quantities shows a variety of statistical methods in use. An adaptation of methods to specific experimental problems explains some of this variety, but it is clear that the lack of an authoritative reference which deals with the real situations facing experimenters is responsible for much of it. Despite this lack, quite a few papers give a very scant explanation of their methods, apparently in the belief that there is general agreement on methods and that their particular procedure is obvious. This is frequently not the case. The arithmetic involved is usually simple, but it can be very difficult to reproduce the limit which is given.

Most books on statistics discuss only central confidence limits, and do not deal in any detail with the question of upper limits. Two of the most advanced books on the subject, Kendall and Stuart[1] and Eadie et al[2], do go into considerable detail on the basic principles involved. They show in the general case how to determine a non-central confidence interval, the limiting case of which is an upper (or lower) limit. They do not, however, discuss the practical problems that arise. The Particle Data Group does not appear to have published anything on the subject (except the material dealing with central intervals and simple Poisson limits in the bluebook) despite several references by Daum et al[4,5] to a "prescription" of the PDG. All this suggests that an elementary discussion of the problems involved with upper limits would be useful.

One source of the diversity in methods is the division between "Bayesian" and "classical" statistics. It happens that the calculation of upper limits is one of the few areas where these two approaches give different results as far as normal use by physicists is concerned. Since the controversy between Bayesian and classical statisticians has been going on for a long time, it is unlikely to be resolved here. Many scientists have been advocates of the Bayesian viewpoint, arguing that it is most like the actual scientific view, but two of the most respected statistical references[1,2] used by physicists are clearly classical. To recapitulate the problems with the Bayesian approach: (1) one has to develop the concept of a probability distribution for a constant of nature; (2) one needs to know the prior probability distribution for this constant; (3) total lack of knowledge is conventionally represented by a uniform distribution, but there seems to be no convincing way of specifying what function of the constant is to be uniform (e.g. m_ν^2 or m_ν). Justifications of (1) and (2) point to the effect of prior experimental knowledge on "degrees of belief" in the value of a quantity. No one seems to have attempted to incorporate this previous knowledge into a real data analysis formally except to include logical constraints (e.g. a square cannot be negative) in an UL analysis. (It would be interesting to see a Bayesian conclusion drawn about the flux of magnetic monopoles, based on the two candidate events so far seen. This is precisely the sort of sparse data situation where Bayesian theorists suggest that their method is most powerful.) Kendall[1] and Eadie[2] both

consider the Bayesian viewpoint, but they come down on the side of classical statistics: Only experimental quantities have probability distributions. The 90% confidence level upper limit on a physical constant deduced from an experimental value is that value of the constant for which an experimental value as small as the observed one would be expected in less than 10% of identical experiments. In other words, if the true value were any bigger than this upper limit, the chances of a fluctuation to the small value actually observed would be at most 10%.

B. The Normal Distribution

To illustrate the problem in its simplest form, consider an experiment in which an intrinsically positive physical quantity $X = A-B$ is calculated from measured quantities $A \pm \sigma_A$ and $B \pm \sigma_B$. Assuming for simplicity that A and B are normally distributed, X is also normal with known $\sigma^2 = \sigma_A^2 + \sigma_B^2$.

Consider the situation where A and B are nearly equal. An example would be $m_\nu^2 = E_\nu^2 - P_\nu^2$, which is essentially the experiment of Daum et al[3-6]. If the true value X_o is in fact zero, an unbiased estimator of X_o, e.g. $X = A-B$, will be negative in half of all experiments. As Kendall and Stuart[1] remark in several places (see section 27.34), this is a case in which a biased estimator, e.g. Max(X,0), is clearly preferable to an absurd one.

Since it is the custom in these cases to quote not the estimate and a central confidence interval but rather the 90% CL upper limit, the problem is how to calculate this limit. (One might argue that this custom is one way that a priori assumptions are incorporated into the data analysis.) Several possible algorithms for this calculation are discussed below. The computed upper limit as a function of the measured value is given in Fig.(1) for each algorithm. The expected fraction of cases in which the the upper limit will be wrong (i.e. smaller than the true value) is given in Fig.(2) as a function of the true value.

1. Pure Classical Method

Since the distribution of X is normal $N(X_o,\sigma)$, it is straightforward to calculate the value X_u such that if the true value X_o were X_u the probability of obtaining a value as small as the observed X would be $\leq$ 10%. The result is $X_u = X + 1.28\sigma$. When $X < -1.28\sigma$, X_u is negative, but this will happen less than 10% of the time. This X_u is therefore consistent with the classical definition of a 90% CL upper limit.

A method that gives a negative upper limit requires some explanation. As a purely mathematical problem, a fluctuation to a very negative value of X makes one certain that the value of X_o is not very large. For example, $X = -3.0\sigma$ would imply not that $X_o < -1.7\sigma$ with 90% confidence, but that $X_o < 0.1\sigma$ with 99.9% confidence. Physically this is ridiculous. Such a result would make one pretty certain that there was something wrong with the experiment, rather than more confident in a small upper limit.

2. Truncated Classical Method[3]

If negative values of X_u were the only problem with Method 1, it could be cured by truncation: $X_u = \text{Max}[X + 1.28\sigma, 0]$. This limit is violated less than 10% of the time for any X_o, and the limit is never negative. But another objection to Method 1 is not removed by truncation: the more

negative the fluctuation (or possibly the systematic error) in an experiment, the smaller the upper limit. In particular, a very suspect experiment leads to upper limit 0. This is clearly unsatisfactory.

3. Shifted Classical Method[8-11]

In order to avoid the problem with Method 2, consider the algorithm $X_u = \text{Max}(X,0) + 1.28\sigma$. The rationale for the specific form chosen is as follows: If $X > 0$, X_u is identical to the plain classical method which gives a physically and statistically reasonable upper limit which is wrong exactly 10% of the time. For $X < 0$, the mathematical calculation of Method 1 implies (discounting the possibility of systematic error) that X_o is not very large and that the observed value is a statistical fluctuation. In that case a typical similar experiment would obtain $X \approx 0$ and an upper limit of about 1.28σ. If the experimenter does not wish to have a fluctuation lower his limit, an upper limit typical of the intrinsic sensitivity should be used. (J. Franklin has emphasized this point.) This algorithm is consistent with that constraint.

4. Bayesian Method[4-7,12]

A very usual method (what Daum et al[4,5] refer to as "the Particle Data Group prescription") is to use

$$X_u = X + \sigma\Phi^{-1}[1-0.1\Phi(X/\sigma)],$$

where Φ is the cumulative distribution of the Gaussian function. This is the "10% of the positive area" method. It can be derived from the Bayesian discussion given by Eadie[2] (p. 213). Its derivation is based on using a step function at zero as the prior probability distribution in order to convey the knowledge that $X_o \geq 0$. The behavior of this X_u as a function of X is shown in Fig. 1. At large X it is the same as method 3. (The smooth transition to the classical result is perhaps the best justification of the prior probability.) For $-.6 < X < 1$, it is somewhat more conservative than method 3. In particular, for $X = 0$ it gives an upper limit that is violated only 5% of the time if $X_o = 0$. For $X < -.6$ it becomes progressively less conservative than method 3, but since it always lies above curve 2, it is wrong less than 10% of the time. It does share the problem of methods 1 and 2 that more negative X gives progressively smaller values of X_u[6,12].

5. Shifted Bayesian Method[5]

A very conservative method would be to apply the 90%-10% formula after shifting the mean of the curve upward to 0 if it were negative. As in method 3, the rationale would be one of intrinsic experimental sensitivity. This approach does not make a great deal of sense, since the Bayesian philosophy is inconsistently applied.

6. Loss of Confidence Algorithm

K. McFarlane has suggested[13] that as X becomes more negative the upper limit should increase rather than decrease or stay constant. The idea is that if X is negative by more than 1 or 2 standard deviations, then the most likely explanation is that there are uncontrolled systematic effects in the experiment. The magnitude of the errors are probably at least as large as the negative value, which is therefore a measure of the intrinsic sensitivity of the experiment. An example is given in Fig.(1), corresponding to:

$$X_u = \text{MAX}(X,0) + 1.28\sigma, \quad X > -1.28\sigma$$
$$= |X| \quad , \quad X < -1.28\sigma$$

C. The Poisson Distribution

1. Single Poisson Measurement

The one thing upon which everyone agrees is the case of a simple Poisson variate. If the classical definition is applied, the result is that an observation of zero events gives an UL of 2.3, one event gives 3.9, etc. A Bayesian treatment gives identical results, provided that the prior distribution of the parameter is taken to be uniform. Since negative values do not enter into consideration, it seems likely that most people think of this case classically.

2. Difference of Two Poisson Variates

Consider a foreground measurement of n_f Poisson distributed events and a background measurement of n_b such events. The estimator of the signal is $s = n_f - n_b$ with variance $n_f + n_b$. If the samples are moderately large the estimator will be approximately normally distributed and the Gaussian methods discussed above can be applied. It is an interesting, though apparently little known, fact that the distribution of s is known exactly.[14] It is not difficult to show that if the true parameters are μ_b and μ_s for background and signal respectively, then the distribution of the estimator s is given by

$$P(s;\mu_s,\mu_b) = \exp[-(2\mu_b+\mu_s)][(\mu_b+\mu_s)/\mu_b]^{s/2} I_s[2((\mu_b+\mu_s)\mu_b)^{1/2}]$$

where I_s is the modified Bessel function of order s ($s = n_f - n_b$). It can be shown that this expression has the expected mean and variance, and that it rapidly approaches normality as the numbers increase.[15] Given the estimate $\mu_b = n_b$ from the background experiment, the probability of obtaining a measured s for any true value μ_s can be calculated. These values can then be graphed and used to obtain upper limits following the mechanics described by Eadie et al (section 9.2). Note that in this case the variance of the probability distribution depends on the parameter μ_s so that the confidence lines are not straight lines but curves concave downward as in Eadie's Fig. 9.6. The effect is to make the upper limit larger than it would have been if the variance were constant at the value associated with the point estimate.[9] This effect is already incorporated in the treatment of the simple Poisson case, but should be remembered in making Gaussian approximations to more complicated Poisson situations. For example, the case $s = n_f - \alpha n_b$ apparently has no analytic solution and so must be handled by a normal approximation[9,14] or by a Monte Carlo generation of the distribution.

D. Likelihood and Other Methods[7]

Although it is not so easy to describe the UL methods used in more elaborate analyses, such as maximum likelihood, the general ideas and choices involved are not so different. In particular, the significance of statistical fluctuations is the same. In many cases the likelihood function is assumed to be normal and then the alternatives discussed above should be applicable, at least in principle.

It should be noted that normality is an asymptotic property of the likelihood function, as is the maximal efficiency of the maximum likelihood estimator. These properties are guaranteed only as the number of events becomes very large. It seems to me to be open to question whether these conditions are met in the typical UL situation. There the number of data may be large, but it is usually all background, while

there are very few or zero signal events. The likelihood function frequently seems to attain its maximum (in the zero derivative sense) at a negative value, though the full curve is rarely shown in that case. The standard theorems seem dubious in this situation. It would seem that the best plan in all cases is to give enough information so that the reader can evaluate the situation himself. In the likelihood or minimum χ^2 methods, that would call for giving the curve all the way to its maximum (or minimum), even if the independent variable is unphysical in that region.

E. Conclusion

The key idea discussed above is that one should avoid having statistical fluctuations or, worse yet, systematic errors produce a lower value for an upper limit than would be obtained by the experiment on the average and in the absence of systematic errors. The shifted classical algorithm (method 3) seems to me to be the most reasonable alternative that satisfies this criterion while staying close to a reasonable understanding of a statistical confidence limit.

The only reasonable alternative would be the Bayesian method (4), which might be chosen on one of several grounds. If it is chosen on philosophical grounds, there is not much to say, but one could reasonably expect a discussion of the reasons for the prior probability distribution in each case. Method (4) also might be chosen in the belief that it is mandated by some authority or that it is at least customary. Although the former belief does not seem to be correct, there is some truth to the latter, but not so much as to exclude other considerations. Finally one might simply prefer the perfectly smooth behavior of the Bayesian formula. As shown in Fig (2), it is not very likely that it will give an excessively small limit provided negative measurements are due only to statistical fluctuations. There is the distinct danger that systematic errors (unfortunately not so rare) will cause a very low limit to be computed. When large negative values occur, it is clearly necessary to consider the implications for the reliability of the experiment, as discussed in method (6).

In any case, since there is a lack of consensus on the statistical principles involved, it is particularly necessary to give a clear statement of the procedure used to calculate an upper limit if a reader is to be able to follow the conclusions.

Figures

1. Curves giving the 90% confidence upper limit computed from a measured value according to the various algorithms discussed in the text for the case of a normal distribution with fixed variance. The independent variable is the measured value divided by the standard deviation.

2. Curves giving the probability that an upper limit will give a value smaller than the true value of the parameter, for several of the algorithms discussed in the text. The independent variable is the true value of the parameter divided by the standard deviation.

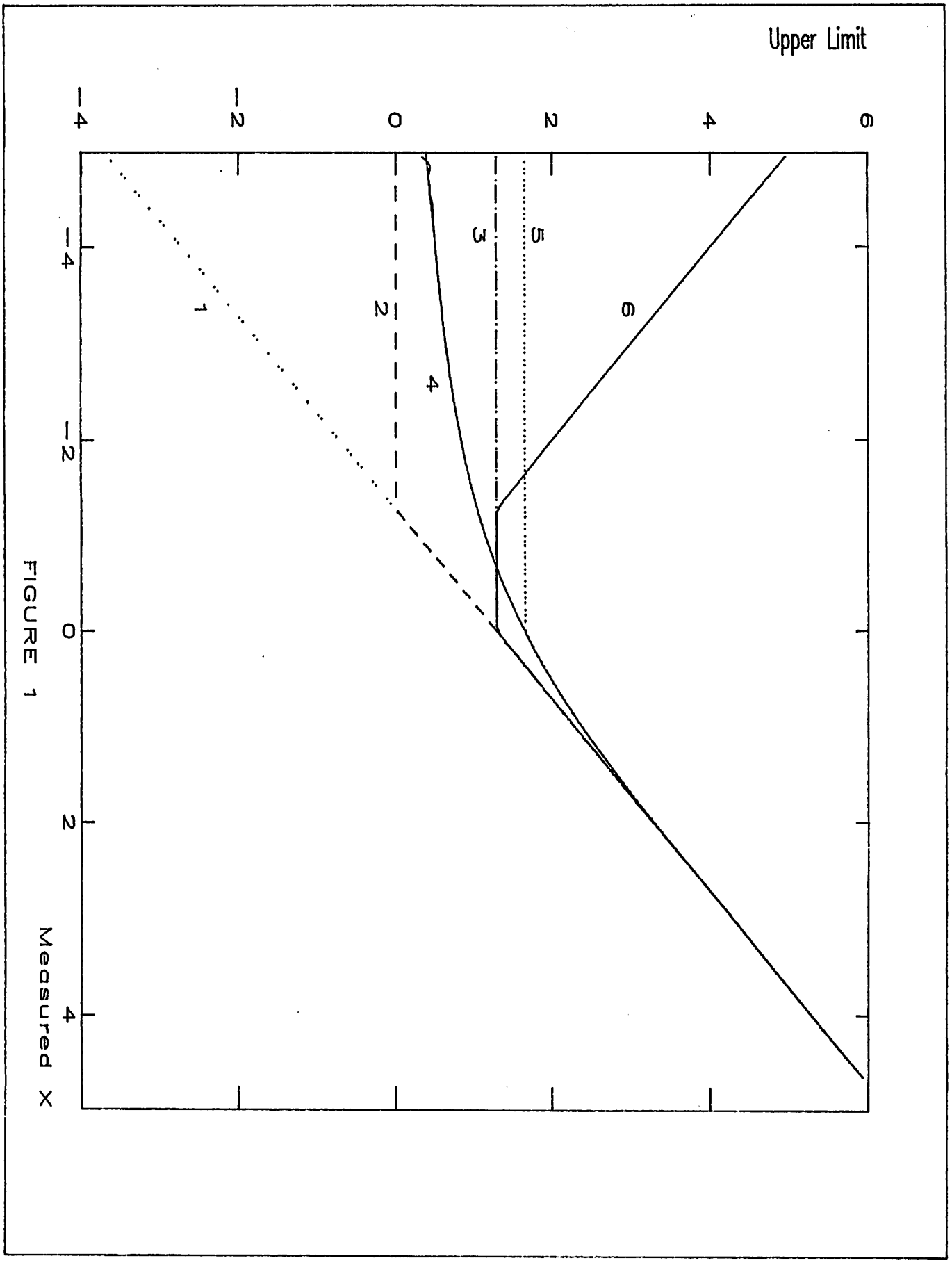

FIGURE 1

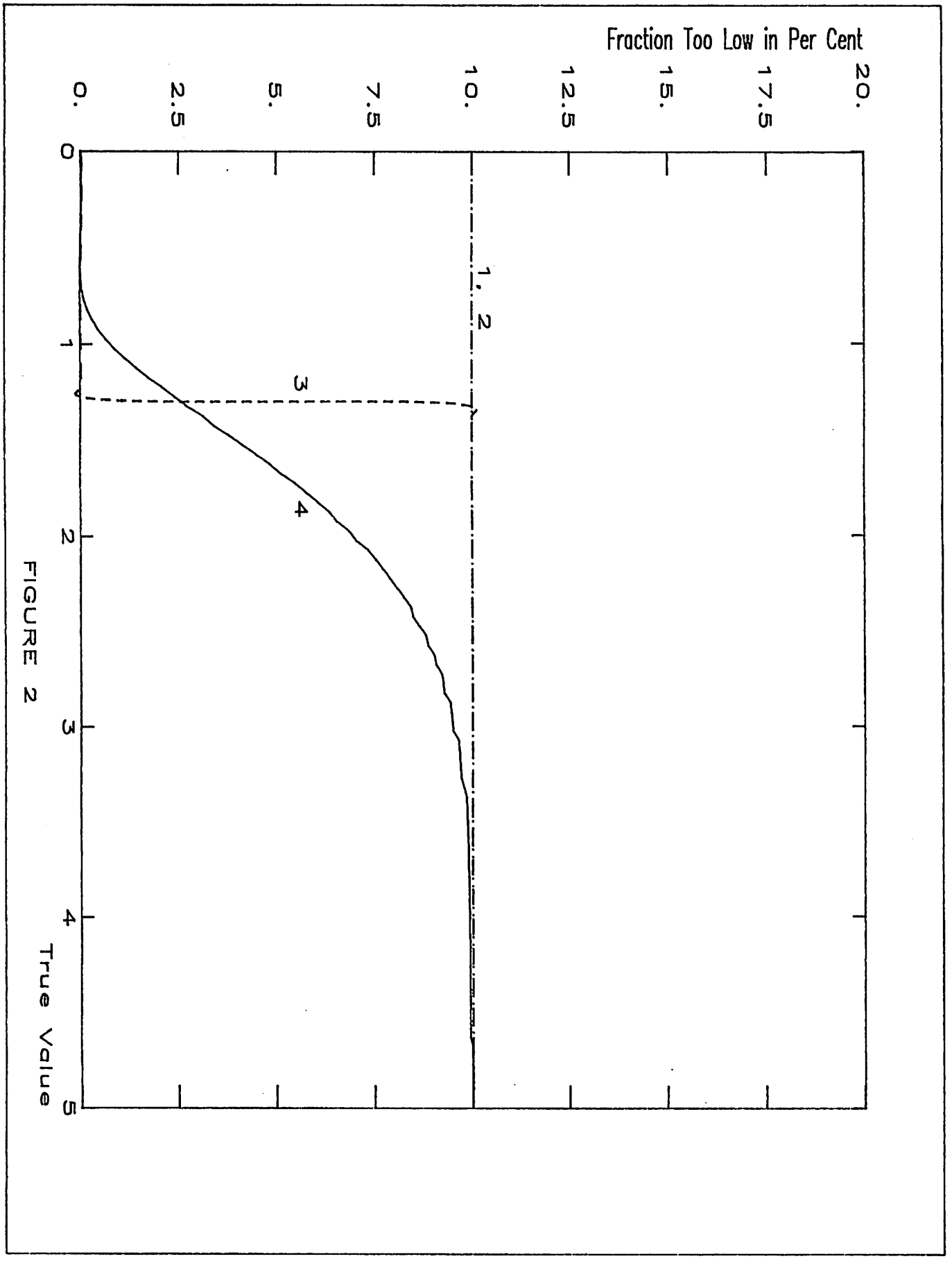
Fraction Too Low in Per Cent
0.
2.5
5.
7.5
10.
12.5
15.
17.5
20.
0
1
2
3
4
5
1.2
3
4
True Value

FIGURE 2

References

1. M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol II (Hafner, New York, 1973).

2. W. T. Eadie, D. Drijard, F. James, M. Roos, and B. Sadoulet, Statistical Methods in Experimental Physics (American Elsevier, New York, 1971).

3. M. Daum, L. Dubal, G. H. Eaton, R. Frosch, H. Hirschmann, J. McCulloch, R. C. Minehart, E. Steiner, C. Amsler, and R. Haussmann, Phys. Lett. 60B, 380 (1976). m_ν^2, Method B1 or B2, but using an 84% CL.

4. M. Daum, G. H. Eaton, R. Frosch, H. Hirschmann, J. McCulloch, R. C. Minehart, and E. Steiner, Phys. Rev. D20,2692 (1979). m_ν^2, Method B4.

5. R. Abela, M. Daum, G. H. Eaton, R. Frosch, B. Jost, P. Kettle, and E. Steiner, Phys. Lett. 146B, 431(1984). m_ν^2, Method B4. In a 1984 SIN report of their work, the group also quotes a result using Method B5.

6. H. Anderhub, J. Boecklin, H. Hofer, F. Kottmann, P. LeCoultre, et al, Phys. Lett. 114B,76 (1982). m_ν^2, Method B4. Making the same measurement as Ref 1-3, this group attained a 15% lower UL despite a 50% worse precision, thanks to a negative fluctuation of -0.7σ.

7. W. Kinnison, H. Anderson, H. Matis, S. Wright, R. Carrington, R. Eichler, R. Hofstadter, E. Hughes, T. McPharlin, et al, Phys. Rev. D25, 2846, (1982). $\mu \rightarrow e\gamma$, Method B4, in the context of a maximum likelihood analysis.

8. C. Baltay, P. Franzini, J. Kim, R. Newman, N. Yeh, and L. Kirsch, Phys. Rev. Lett. 19, 1495(1967). $\eta \rightarrow \pi\gamma\gamma$, Method B3.

9. V. L. Highland, L. B. Auerbach, N. Haik, W. K. McFarlane, R. J. Macek, J. C. Pratt, J. Sarracino, and R. D. Werbeck, Phys. Rev. Lett. 44, 628(1980). $\pi^o \rightarrow 3\gamma$, Method B3.

10. J. Duclos, D. Freytag, K. Schlupmann, V. Soergel, J. Heintze, and H. Rieseberg, Phys. Lett.19, 253(1965). $\pi^o \rightarrow 3\gamma$, Method B3.

11. J. Cole, J. Lee-Franzini, R. J. Loveless, C. Baltay, et al, Phys. Rev. D4, 631(1971). $\Sigma^+ \rightarrow ne^+\nu$, Method B3. A classical calculation with a fixed variance seems to be used for the case of a Poisson subtraction.

12. S. Berko and A. L. Mills, Phys. Rev. Lett. 18, 425(1967). Positronium $\rightarrow 3\gamma$, Method B4. The experiment gave two semi-independent results: $(-11.3 \pm 6.3)\times10^{-6}$ and $(10.8 \pm 11.9)\times10^{-6}$. The final limit was based only on the first result, for which they computed a 68% CL upper limit of 2.8×10^{-6}, i.e. 0.45 standard deviations. A similar calculation of a 90% limit would have given 5.4×10^{-6} (0.87 standard deviations). The probability of measuring a result 1.82 standard deviations negative is $\leq$3.4%

13. W. K. McFarlane, private communication.

14. F. Haight, Handbook of the Poisson Distribution, (Wiley, New York, 1967).

15. The skewness is $\mu_s/(2\mu_b+\mu_s)^{3/2}$ and the kurtosis is $(2\mu_b+\mu_s)^{-1}$.