Degrees of Freedom: A Correction to Chi Square For Physical Hypotheses

by John Michael Williams

P. O. Box 2697
Redwood City, CA 94064
jwill@BasicISP.net

Abstract

In common practice, degrees of freedom (*df*) may be corrected for the number of theoretical free parameters as though parameters were the same as data categories. However, a free physical parameter generally is not equivalent to a data category in terms of goodness of the fit.

Here we use synthetic, nonrandom data to show the effect of choice of categorization and *df* on goodness of fit. We then explain the origin of the *df* problem and show how to avoid it in a three-step process:

- First, the theoretical curve is fit to the data to remove its variance, leaving what, under the null hypothesis, should be structureless residuals.
- Second, the residuals are fit by a set of orthogonal polynomials up to the degree, should it occur, at which significant variance was removed.
- Third, the number of nonsignificant polynomial terms in the original + orthogonal set become the *df* in a standard chi square test.

This process reduces a general *df* problem to one of polynomial *df* and allows goodness of a fit to be determined by data categorization and significance level alone. An example is given of an evaluation of physical data on neutrino oscillation.

Table of Contents

I. Introduction

Parametric Statistics

Hypotheses are theoretical assumptions. Scientific hypotheses may be tested against empirical data. It is fundamental to all science, that no hypothesis ever can be proven by experimental data; an hypothesis only may be rendered more or less likely, or disproven.

The most famous modern use of formal statistics to test scientific hypotheses dates back to R. A. Fisher [1], who, in the early twentieth century, published a method for analysis of the variance in experimental data. Fisher developed sensitive ways of detecting differences among data as a function of category of treatment. The treatments in his favorite examples were on agricultural variables such as crop fertilization, light, watering, and so forth. Fisher's methods were broad in scope--the nonparametric *Fisher Exact Test* is named for him--but his analysis of variance was based on the apparently narrow assumption that the data were sampled independently and would be normally distributed with constant variance.

Defining a random variable $\boldsymbol{X}$ with normally distributed probability density $N\left(X;\mu,\sigma^2\right)$ by

$$P(\mathbf{X}=x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

it easily may be seen that the two parameters of such a distribution are the mean μ and the variance σ^2. The mean (expected value) of data may be taken to define the effect of an experimental treatment. A sample taken of data depending on a mixture of different effects will have a bigger variance than a sample depending on one; so, the variance may be analyzed to show differences in the underlying means.

If we use mean values to represent categories of data, then hypothesis testing by analysis of variance may be used to test differences among the categories. Fisher's approach to analysis of variance came to be known as *parametric statistics*.

Fisher's analysis of variance rationale has been supplemented by Bayesian methods [2, 4]. Problems not fitting Fisher's paradigm have extended the scope of hypothesis-testing statistics into the nonparametric realm [3, Ch. 7; 6, Sect. 9.2 & Ch. 12]. Nonparametric statistics attempt to resolve questions for data in very small samples, for mismatched variances, and for experimental results sometimes more qualitative than numerical.

Statistical Tests in Physics

A problem with hypothesis testing in physics is that much of the subject matter is multiply parameterized. In physics, a good hypothesis rarely states anything about categories; instead, it provides an explanation of a set of one or more phenomena which may vary continuously. The data typically involve broad ranges of energy or momentum, or averages of many carefully winnowed events. Questions arise of, Is the spin 1/2? Does the field vary as $1/r$? The connection of any physical hypothesis with data typically is through multidimensional continua of complexity.

Rather than analyzing experimental variations to discover differences in means on narrow domains of one or more continuua, the preferred way to demonstrate support for a physical hypothesis is to show that it may be used unequivocally to control some physical process in such a way as no other known hypothesis might explain.

For example, a demonstration of the Hall effect would involve an *unequivocal* excess of charge along one side of a flat conducting strip. To achieve this confirmation of the predicted effect, superior instrumentation or measurement of previously ignored quantities would be adopted. The gradient of the charge, the effect's time-course of development, and other factors would be examined for consistency with the hypothesis. Rarely would physical measurement be on such a narrow range of values that the variance would be constant, independent of the mean. Also, if not all evidence supported the hypothesis, alternatives would be explored instead. Merely showing that one side of a conducting strip seemed more likely to have a charge excess than the other, a test between two categories on one variable, would be an unsatisfying result.

Doing Physics by Category

Let's consider the special case of curve-fitting hypotheses. The problem is to distinguish among various hypotheses of the shape of the curves. We might start by arranging the data along one or more continua. However, on a continuum, two data (probably) never coincide; this means we have slim bases for estimating the mean, variance, or any other parameter of the assumed underlying random variable.

So, to use parametric statistics, we define category boundaries (bins). By making the bins wide enough, the number of data available in each will be large enough to provide a good estimate of the bin mean. We may assume the variance in a bin to be constant, but only if the bins are narrow enough. Sometimes riskily, we may assume the variance over all bins to be the same, constant value.

A simple example will illustrate this approach. Imagine that two different physical theories are to be tested against a certain data set. Perhaps, the theories might relate to field intensity I as a function of distance x.

Hypothesis I: The data follow a quadratic function. This would strengthen belief in Theory I.

Hypothesis II: The data follow a cubic function. This would support Theory II.

Here are the two hypothetical predictions on a data set which has been generated purely artificially, with no randomization:

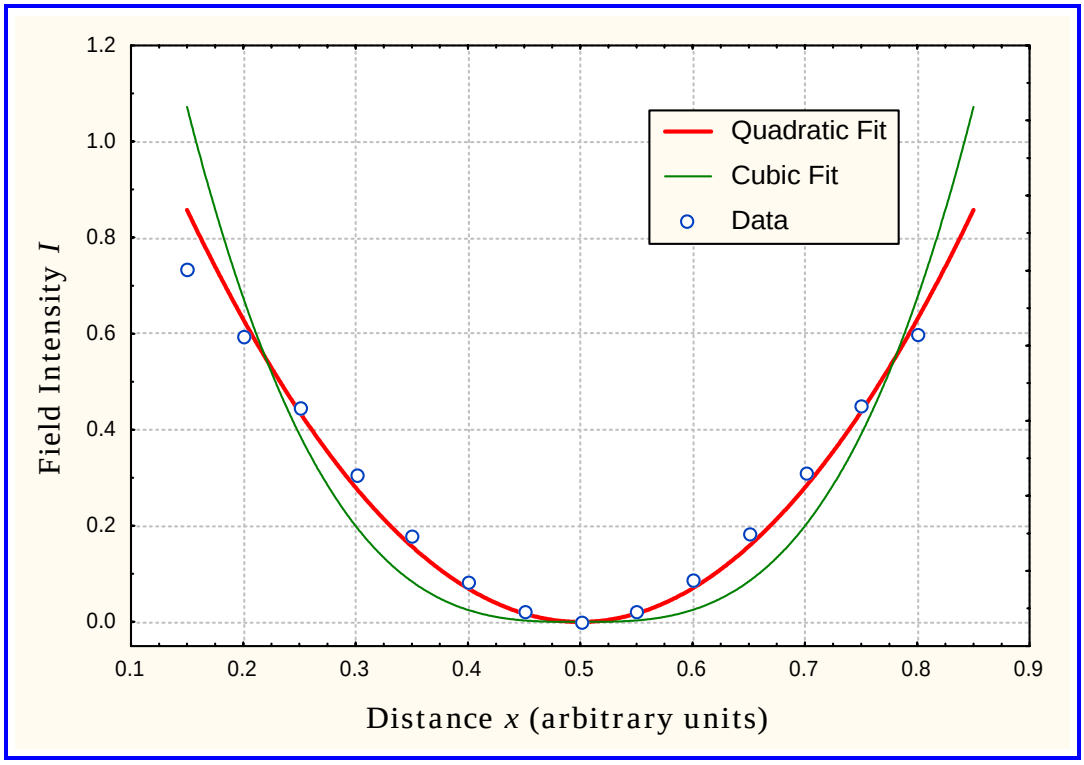

Figure 1. Simulated data fit by eye. Deviations from the data seem greater for the cubic than for the quadratic. The curves represent $I = 7\left(x - 1/2\right)^2$ for the quadratic, and $I = 25\left|\left(x - 1/2\right)^3\right|$ for the cubic.

Clearly, visual inspection tells us that the quadratic fit is better. So, we immediately should drop Theory II in favor of Theory I. Or, should we? How can we test how well Theory I has been confirmed? Is it possible Theory II might be more elegant to derive, simpler, or otherwise more attractive than Theory I, even though this data set was not fit so well?

We shall discuss the special case in which we wish to test Theory I against the data, leaving out Theory II entirely. This kind of statistical test, which measures the fit rather than rejecting a competing fit, is called a *goodness of fit* test.

The most popular goodness of fit test is by ***chi square***, a statistic first derived by Helmert [2, Sect. 3.4.3] and later independently derived and popularized by Karl Pearson; it is named by Pearson after the Greek letter *chi* (χ). Chi square simply is the distribution of the sum of squares of a set of standardized, normally distributed data.

Categories in NonRandom Chi Square

Before presenting formalities about chi square, we first develop its application to the curve-fit example in Figure 1 above. We ignore the cubic curve for a while.

One Category

Suppose we computed the mean values of the data and of the quadratic curve on the interval of interest, $0.1 \le x \le 0.9$. We would get, using all 14 available data points shown in Figure 1, a statistically invalid analysis but an instructive example as shown in Table 1.

Table 1. Comparison of data vs quadratic means on one category (1 degree of freedom)

Distance *x*	**Data of *I***	**Quadratic**
0.151	0.73	
0.201	0.59	
0.251	0.44	
0.301	0.30	
0.351	0.18	
0.401	0.08	
0.451	0.02	
0.501	0.00	
0.551	0.02	
0.601	0.08	
0.651	0.18	
0.701	0.31	
0.751	0.45	
0.801	0.60	
mean =	0.2855	μ_I = 0.3733
sample sd of *I* =	0.2349	σ_I = 0.3339
sample sd of mean	0.0628	$\sigma_{\bar{I}}$ = 0.0892

The mean of the quadratic was computed as follows, using the expected value operator, E,

$$\mu_I = E(I) = \frac{\int_{0.1}^{0.9} dx \cdot I(x)}{\int_{0.1}^{0.9} dx} = \frac{\int_{0.1}^{0.9} dx \cdot 7\left(x - 1/2\right)^2}{x\big|_{0.1}^{0.9}} \cong 0.3733. \tag{2}$$

The operator $E(\cdot)$ means the same as the familiar $\langle\cdot\rangle$; E is used here for symmetry with the variance operator, $V(\cdot)$.

The standard deviation of the quadratic was computed as follows,

$$\sigma_I = \sqrt{V(I)} = \sqrt{E\left(I - E(I)\right)^2} = \sqrt{E\left(I^2\right) - \left[E(I)\right]^2} = sqrt\left(\frac{\int_{0.1}^{0.9} dx \cdot I(x)^2}{\int_{0.1}^{0.9} dx} - \left[\frac{\int_{0.1}^{0.9} dx \cdot I(x)}{\int_{0.1}^{0.9} dx} \right]^2 \right)$$

$$\sigma_I = sqrt\left(\frac{\int_{0.1}^{0.9} dx \cdot \left(7\left(x - 1/2\right)^2\right)^2}{\int_{0.1}^{0.9} dx} - [0.373]^2 \right) \cong 0.3339. \tag{3}$$

The standard deviations of the means were computed by dividing the standard deviations of the variables (data or quadratic) by the square root of the sample size, $\sqrt{N} = \sqrt{14} \cong 3.742$.

Now, instead of just noticing in Table 1 that the means are about 0.09 apart, and that the standard deviations of the means are about the same size, suggesting no significant difference between the theory and the data, let's look at the chi square statistic. A *unit-normal* distribution is defined as one with mean 0 and variance 1, we describe such a distribution as $N(0,1)$, a notation which may be related to equation (1) in an obvious way.

We follow Brownlee [3, Sect. 1.27], and others in defining chi square as a sum of squared unit-normal deviates: Given a number N_C of data categories,

$$\chi^2_{Data} = \sum_{i=1}^{N_C} N_i^2(0,1), \tag{4}$$

with computed or tabulated cumulative probability P at significance level x_p, $P\left[\chi^2(df) \geq x_p\right] \leq p$, represented as $\chi_p^2(df)$.

In common usage, which we shall question later, the degrees of freedom variable *df* is the number of statistically independent categories of the data, and the unit-normal deviates are differences between theoretical (or otherwise anticipated) values and data values, one for each category. Because the differences are defined standardized, their means all will be expected to be 0 under the null hypothesis that the fit is good.

Usually, the variance in each *df* category would be the variance under the null hypothesis; in our introductory examples, this is the variance under the assumption that the data were distributed randomly in each category as determined by the curve being fit. In this introduction, we shall be looking at the curve fit to data generated under a theory and examining the adequacy of the representation of that theory by the fitted curve. So, we are looking at two different theories, in a sense, one a simplification of the other. In our introductory examples, there is no random component other than rounding error of the numerical values.

In common usage, we would estimate the variance in each category using the data, thus losing a degree of freedom [4, Sect. 10.11]. In these introductory examples, we assume the apparatus returns good data, and that we are measuring the lawful realization of a physical principle. We assume we know the means and the change in them within a category; so, we need not use data to estimate anything; we merely are measuring the fit of a curve. So, in the first few examples following, we shall use the variance of the quadratic curve to standardize in each category, rather than the variance of the equally theoretically correct data. The actual difference in the variances is not large anyway, as may be seen in the tabulations below.

So, to evaluate chi square, we may compute the following sum:

$$\chi_{Data}^2 = \sum_{i=1}^{N_C} \left(\frac{X_i - E(X)_i}{\sqrt{V(X)_i}} \right)^2 = \sum_{i=1}^{N_C} \frac{\left[X - E(X)\right]_i^2}{V(X)_i}, \tag{5}$$

in which, in each category, the theoretically expected value $E(X)$ is subtracted from the value of the observed datum X. In general, each *i*-th category represents a mean of some observations, from which observed mean will be subtracted the *i*-th expected mean. Under the null hypothesis that the fit is good, the difference in the numerator of the sum in (5) will, of course, be zero. Dividing each difference by the expected standard deviation of X standardizes the variance of each of the *i* elements in the sum.

We notice here, immediately, that our chi square ignores the theoretically expected shape of the curve: It just sums squared differences. This means that chi

square is not truly a parametric statistic. Each category is standardized individually by its own expected variance; therefore, chi square does not require a constant variance over all categories tested. This is a great advantage in physics. The shape of the theoretical curve enters only locally, at the granularity of the individual categories, when a difference ($X_i - E_i$) is taken and a variance estimated perhaps on the assumption that the variance will be constant.

Anyway, let's decide to reject the fit at the p = .001 level. Keep in mind that we do not have a valid random sample and that this is meant as a thought-provoking introduction to a problem of statistical inference in physics. The tables of chi square tell us that $\chi^2_{.001}(1) \cong 11$. We trivially compute chi square from the Table 1 means this way:

$$\chi^2_{Data} = \frac{\left(X_1 - E_1\right)^2}{V_1} = \frac{(.2855–.3733)^2}{(.0892)^2} \cong 1.0; \qquad (6)$$

and, confirming our visual impression, the result is not significant, so the fit may be considered good to the extent the overall averages of the data and quadratic hypothesis in Table 1 are not shown different. And we are confident the fit would have been shown good with likelihood of error p = .001, had the statistical requirements been fulfilled.

We should mention here that the difference in means in Table 1 also might have been tested by other statistics, such as Student's *t*. Above, we have accepted the mean and variance of the quadratic function as theoretically errorless, and have accepted the quadratic's computed $N(.3733,.0892)$ distribution as that of the data, under the null hypothesis. In contexts other than the present introduction, this would be quite wrong statistically, because, visibly, the mean and the variance change systematically over the domain of interest (standard deviation is a measure of slope)--and we have a quadratic curve, not a random normal variate.

As a check, ignoring these problems, at p = .001 (two-tailed), a normally distributed data mean would have to lie about 3.3 standard deviations of the mean away from the quadratic mean to reject the goodness of the fit. This would be about 0.30 distance units, clearly about 3 x greater than the 0.09 difference in Table 1.

Using the normal significance value in a second check, if we substitute 0.30 in the chi square numerator of (6) above, we get $0.30^2/0.09^2 \cong 11.1$, closely matching the chi square significance level of 11. This is a tautology, confirming our arithmetic, because by definition $\chi^2_{.001}(1) \equiv \left[N_{.001}(0,1)\right]^2$.

Two Categories

Now, let us double the number of categories and the sample size, too. We increase the density of the data points but keep the same quadratic fitting function:

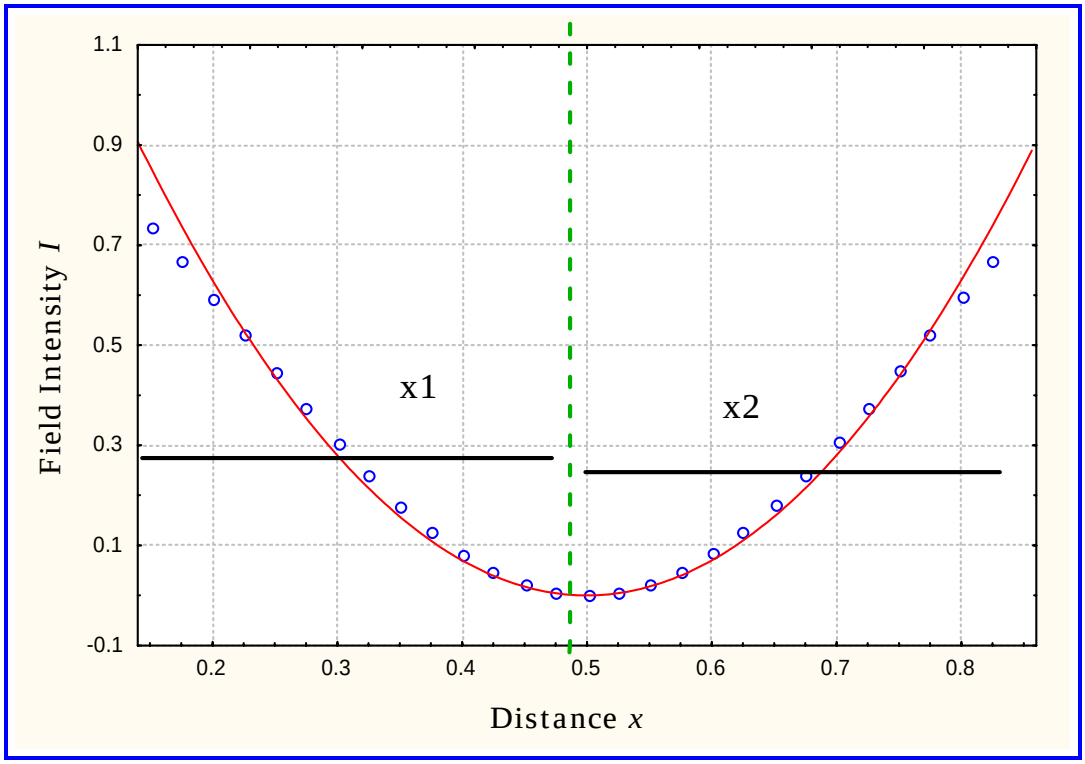

Figure 2. Two-category fit by eye. More of the same artificial data as in Fig. 1, for the same quadratic fit $I = a(x-b)^2 = 7\left(x-1/2\right)^2$.

This time, we partition the domain of analysis *x* into two regions, or bins, called *x1* and *x2*, as shown in Figure 2. The horizontal lines show the different means of *I* in the two bins. We tabulate the data shown, and recalculate the parameters of the quadratic separately for the two halves, as was done above. We change the problem slightly here, by integrating this time on the smaller, more precise domain of $x \in (.15,.85)$ instead of $x \in (.10,.90)$, making the domain of evaluation of the quadratic match the domain of the sampled data more closely. Each half now has 14 data points. The result is in Table 2.

Looking at Figure 2, with two categories, the obviously-changing data at least are more or less monotonic in each bin. As confirmed in Table 2, the two bins, dividing the data exactly in half at a point of symmetry, make the two quadratic categories almost identical. The quadratic statistics all are lower in Table 2 than in Table 1 because of the small reduction in the domain of integration, which eliminated the largest values of dI/dx. The sample standard deviations and standard deviations of the sample means in Table 2 agree reasonably closely with those in Table 1, because we have essentially the same functional change in each "data" bin in both cases, making slopes all about the same. However, the quadratic curve-fit standard deviations reflect the narrower domain of integration. If we calculated a standard

deviation of the grand mean in Table 2, using both bins, we would find it $\sqrt{2}$ smaller than in Table 1, because of the doubling of the data (sample size).

Table 2. Comparison of data vs quadratic means on two categories (2 degrees of freedom)

	***x1*: Low Half**			***x2*: High Half**		
	x	**Data**	**Quad**	***x***	**Data**	**Quad**
1	0.151	0.7338		0.501	0.0000	
2	0.175	0.6677		0.525	0.0053	
3	0.201	0.5926		0.551	0.0220	
4	0.225	0.5216		0.575	0.0471	
5	0.251	0.4445		0.601	0.0844	
6	0.275	0.3747		0.625	0.1273	
7	0.301	0.3023		0.651	0.1822	
8	0.325	0.2396		0.675	0.2396	
9	0.351	0.1777		0.701	0.3077	
10	0.375	0.1273		0.725	0.3747	
11	0.401	0.0811		0.751	0.4504	
12	0.425	0.0471		0.775	0.5216	
13	0.451	0.0203		0.801	0.5985	
14	0.475	0.0053		0.825	0.6677	
mean (n=14)		0.3097	μ_I = 0.2858	mean	0.2592	μ_I = 0.2867
sample sd of I		0.2494	σ_I = 0.2557	sd I	0.2304	σ_I = 0.2556
sample sd of mean		0.0667	$\sigma_{\bar{I}}$ = 0.0683	sd m	0.0616	$\sigma_{\bar{I}}$ = 0.0683

As before, imagining the two categories to be statistically independent, we may look for a rejection of goodness of fit in Figure 2 at the p = .001 level at $\chi^2_{.001}(2) \cong 13.8$. We may compute the value of chi square from Table 2 as follows:

$$\chi^2_{Data} = \frac{\left(X_1 - E_1\right)^2}{V_1} + \frac{\left(X_2 - E_2\right)^2}{V_2} = \frac{(.3097 - .2858)^2}{.(0683)^2} + \frac{(.2592 - .2867)^2}{(.0683)^2} \cong 0.28; \qquad (7)$$

Again, we get nowhere near a value allowing us to reject goodness of fit; in fact, we are farther away, because the slight adjustment of the domain away from the largest values of I (above) has removed the worst deviations from a perfect fit.

Two Categories with More Data

Now, suppose we complete our experimental work and have a set of some 1,000 data at equal 0.001 intervals on the domain $x \in (0,1)$. As in Figure 2 and (7) above,

we have decided to keep just $x \in (.15,.85)$. This means we will keep just 700 useful data.

Without tabulating data, the result of a chi square test on two categories with 350 data each is shown in Table 2a:

Table 2a. Comparison of numerous data vs quadratic means on two categories (2 degrees of freedom)

	***x1*: Low Half**			***x2*: High Half**	
	Data	**Quad**		**Data**	**Quad**
mean (n=350)	0.2831	μ_I = 0.2858	mean	0.28520	μ_I = 0.2867
sample sd of I	0.2321	σ_I = 0.2557	sd I	0.23290	σ_I = 0.2556
sample sd of mean	0.01241	$\sigma_{\bar{I}}$ = 0.01367	sd m	0.01245	$\sigma_{\bar{I}}$ = 0.01367

Of course, the quadratic means and variances have not changed; however, we now must standardize the category means by the much larger sample size. The result is,

$$\chi^2_{Data} = \frac{\left(X_1 - E_1\right)^2}{V_1} + \frac{\left(X_2 - E_2\right)^2}{V_2} = \frac{(.2831-.2858)^2}{.(01367)^2} + \frac{(.2852-.2867)^2}{(.01367)^2} \cong 0.05; \qquad (7a)$$

and this is even farther from the significance level of 13.8. We thus might conclude from chi square on two categories that the fit of the quadratic was good.

Ten Categories

As above, we have 700 usable, equally-spaced data on $x \in (.15,.85)$. Let us test the fit of the quadratic function again, this time dividing our x domain into 10 categories. Using "box plot" representation, the result, again a fit by eye, is in Figure 3. Each of the 10 bins has been processed as we did in the preceding analyses. Again, we have such precision of measurement of the data in this artificial example, that there is no legitimate randomness.

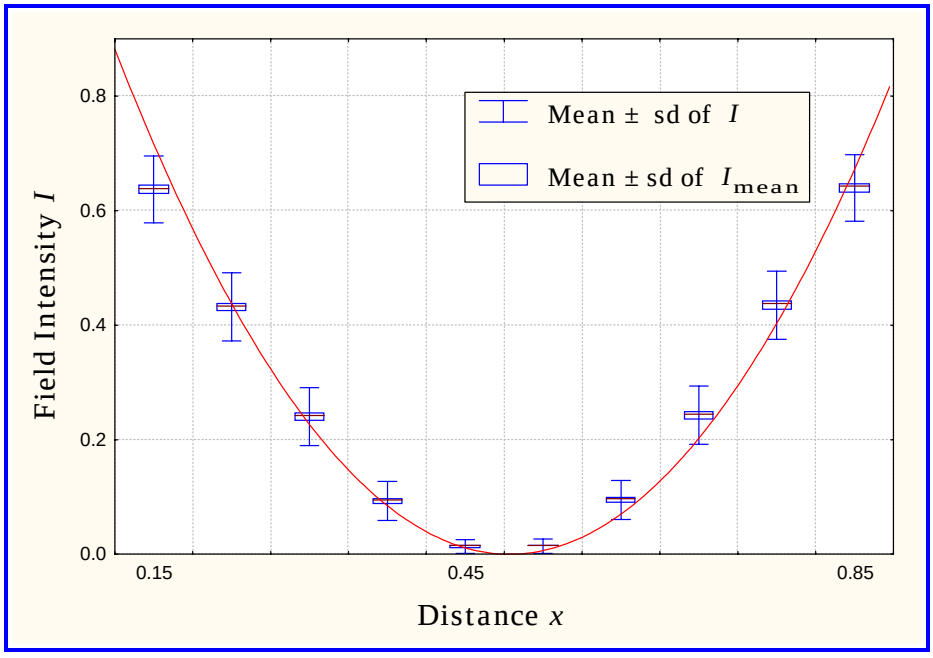

Figure 3. Ten-category fit by eye. Summarized set of 700 of the same artificial data as in Figs. 1 and 2, for approximately the same quadratic fit, $I = 7\left(x - 1/2\right)^2$.

Notice in Figure 3 that the variance in each bin clearly increases with the local slope of the curve. Each mean represents 70 data; the equal-spacing of the data project on the local slopes to determine the standard deviations shown.

Also notice that the large number of 70 data in each bin, treated as though sampled randomly and independently, yield such a small standard deviation of the bin mean, that for only one of the data, the second from the left, does the quadratic actually fall within a standard deviation of the mean.

To prepare our chi square test of goodness of fit of the quadratic, we summarize in Table 3 the statistics plotted in Figure 3:

Table 3. Comparison of data vs quadratic means on ten categories (10 degrees of freedom)

Bin # (70 data)	Bin mean	Bin sd	sd of Bin mean	Quad mean	Quad sd	sd of Quad mean	χ^2 element
1	0.63680	0.05840	0.006976	0.695150	0.087730	0.010486	30.97
2	0.43190	0.05945	0.007106	0.421240	0.068241	0.008156	1.71
3	0.24020	0.05077	0.006069	0.215930	0.048726	0.005824	17.37
4	0.09301	0.03404	0.004068	0.079220	0.029227	0.003493	15.58
5	0.01348	0.01220	0.001459	0.011110	0.009933	0.001187	3.99
6	0.01407	0.01253	0.001498	0.011599	0.010206	0.001220	4.10
7	0.09468	0.03432	0.004102	0.080689	0.029524	0.003529	15.72
8	0.24270	0.05096	0.006091	0.218379	0.049003	0.005857	17.24
9	0.43480	0.05901	0.007113	0.424669	0.068529	0.008191	1.53
10	0.63970	0.05828	0.006966	0.699560	0.088010	0.010519	32.38
						Value of $\chi^2_{Data}(10) \cong$	141

But, $\chi^2_{.001}(10)$ is about 29.5; so, now, with a larger sample and *df* = 10, we easily reject the null hypothesis and conclude that the fit by the quadratic is not good to explain the data.

Fifty Categories

Before leaving this introduction, we look at our nonrandom data once more to see what would happen if we tried a chi square goodness of fit test as above, but with the data subdivided into more numerous categories, say, 50 of them.

A priori, if we had some randomness, we might hope that 50 categories would make for fewer data per bin and thus larger variances of the mean in each bin; so, with higher *df* at the significance point, the chance for a good fit might be improved over what it was with 10.

The result is plotted in Figure 4 and dramatically shows the effect of systematic versus random variation. With 50 categories and 700 data, the real shape of the data curve becomes evident; and, quite contrary to our conclusion based on Figure 1 or 2 above, and maybe to our first impression of Figure 3, we see that the quadratic isn't even close to a good fit to the data. Perhaps a slightly better quadratic might have been chosen by eye, but the shape of the data curve obviously prevents chi square ever from accepting the quadratic on 50 degrees of freedom.

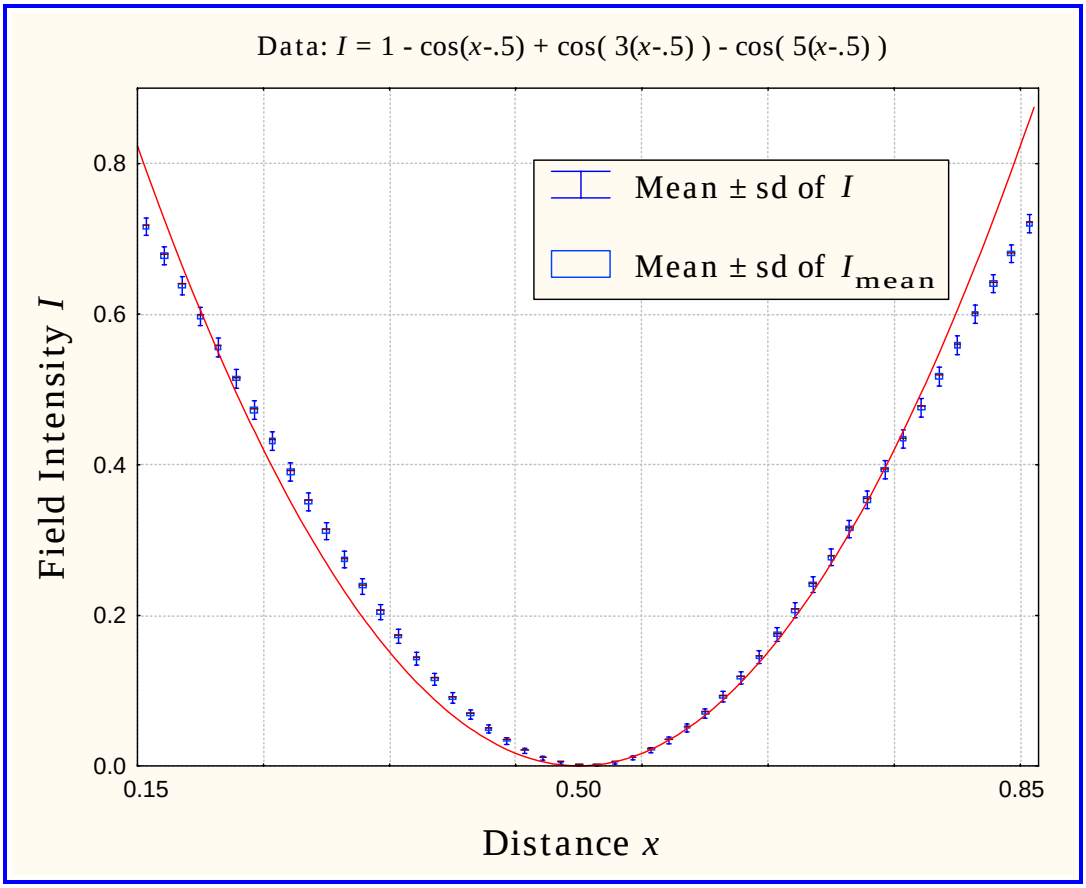

Figure 4. Fifty-category fit by eye. Summarized set of 700 artificial data for the quadratic fit, $I = 7\left(x - 1/2\right)^2$. The function used to generate the data throughout the Introduction is given in the graph title.

Tabulation of the Figure 4 statistics has been omitted; however, the significance point is $\chi^2_{.001}(50) \cong 87$; the obtained value is computed at $\chi^2_{Data} = 304$; so, again, the chi square test reassures us that we should reject the fit as not good.

II. Meaning of Chi-Square

At this point, we leave the introductory examples for some formalism about the statistics. Generalization of these formalisms to the multidimensional case [4, 5] is nontrivial but reasonably straightforward.

Formalisms

Relation to Gamma Distribution

The factorial of an integer n, defined as $\text{Fact}(n) \equiv n! = \prod_{i=1}^{n} n_i = n \cdot (n-1) \cdot \ldots \cdot 1$, appears in combinatorial analysis everywhere. The factorial may be expressed as a special case of the continuous *gamma function*. The gamma function is discussed in the context of Bayesian inference in [6, Sect. 7.3 ff.]; the derivations below follow those of Feller [7 vol II, Ch. 2]. It isn't much of an overstatement to say that the gamma function is as fundamental to statistical inference as the exponential function is to the solutions of differential equations.

The gamma function is defined as,

$$\text{gamma}(x) \equiv \Gamma(x) = \int_0^\infty d\tau \cdot \tau^{x-1} e^{-\tau} . \tag{8}$$

Integrating (8) formally by parts,

$$\Gamma(x) = \frac{1}{x} \int_0^\infty d\tau \cdot \tau^{x} e^{-\tau} = \frac{1}{x} \Gamma(x+1) \text{; therefore,}$$

$$\Gamma(x+1) = x\Gamma(x) \text{; and so,} \tag{9}$$

$$\Gamma(x) = (x-1) \cdot (x-2) \Gamma(x-2) , \tag{10}$$

which clearly shows the recursive relation making $\Gamma(x) = (x-1)!$, for the special case of x an integer.

Using the gamma function Γ in (8), we may define the *gamma probability density* of random variable X on $(0,\infty)$ by,

$$P(X = x) = \text{gamma}(\alpha, \nu) = \frac{1}{\Gamma(\nu)} \alpha^{\nu} x^{\nu-1} e^{-\alpha x} . \tag{11}$$

Looking at (11), we notice that the familiar exponential density then may be defined as gamma$(\alpha,1)$; and so, by analogy, in (11) we will have the mean $\mathrm{E}(X)=\nu/\alpha$ and the variance $\mathrm{V}(X)=\nu/\alpha^2$.

In a small digression, recall the combinatorial expression for the number of ways (unordered) of taking a sample of *j* objects from a discrete population of *m* of them:

$$\binom{m}{j}=\frac{m!}{j!(m-j)!}=\frac{\big(j+(m-j)\big)!}{j!(m-j)!}=\frac{(j+n)!}{j!n!}=\frac{\Gamma(j+n+1)}{\Gamma(j+1)\Gamma(n+1)}. \tag{12}$$

The discrete case may be generalized to the *beta integral*, in terms of the gamma function, starting this way:

$$\mathrm{B}(\nu,\vartheta)=\left[\frac{\Gamma(\vartheta+\nu)}{\Gamma(\vartheta)\Gamma(\nu)}\right]^{-1}. \tag{13}$$

Other uses of the gamma function are discussed in [4, Ch. 4] and [7 vol II, Ch. 2].

Returning to the subject, we may redefine the *unit normal density* as in (1) above, in the following way: From (8) and (11), setting α to $1/2$ and ν to $1/2$ yields,

$$gamma\left(\frac{1}{2},\frac{1}{2}\right)=\frac{1}{\Gamma\left(\frac{1}{2}\right)}(1/2)^{\frac{1}{2}}x^{-\frac{1}{2}}e^{-\frac{1}{2}x}=\frac{1}{\int_0^\infty d\tau\cdot\tau^{-\frac{1}{2}}e^{-\tau}}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{x}}e^{-\frac{x}{2}}. \tag{14}$$

The definite integral $\Gamma(1/2)$ evaluates to $\sqrt{\pi}$; so,

$$gamma\left(\frac{1}{2},\frac{1}{2}\right)=\frac{1}{\sqrt{x}}\left[\frac{1}{\sqrt{2\pi\cdot1^2}}e^{-\frac{1}{2}\frac{\left(\sqrt{x}-0\right)^2}{1^2}}\right]. \tag{15}$$

In the brackets in (15), we have $\sqrt{x}$ in a unit normal distribution. Because $\frac{1}{\sqrt{x}}N\left(\sqrt{X};0,1\right)$ is the density of the square of the unit normal random variable *X* [7 vol II, p.47], we may say that

$$gamma\left(\frac{1}{2},\frac{1}{2}\right)\equiv\left[N(0,1)\right]^2. \tag{16}$$

So, (16) means that gamma$(1/2,1/2)$ represents the probability density of the square of a random variable *X* distributed as $N(0,1)$. It is easy to see the generalization of

(16) to a sum of *n* variates, each distributed as $N(0,1)$: Therefore, recalling (4) above, we may write

$$gamma\left(\frac{1}{2},\frac{n}{2}\right) \equiv \chi^2(n)\text{; and,} \tag{17}$$

this is just the definition of the distribution of chi square.

Thus, we have used the gamma density to show that a sum of *n* squared unit normal variates will be distributed as $\chi^2(n)$. By analogy to (11) above, the mean of $\chi^2(n)$ will be *n*, and the variance will be 2*n*.

Moment Generating Function

Not many readers would see the relation of (15) to (16) above. An easier way to show the relation between the normal and chi square probability densities is by comparison of their moment generating functions (mgf*s*).

The *moment generating function* of a random variable is a series expansion of its density which, term by term, yields as coefficients the *n*-th moments of the random variable about the origin. The first moment is the *mean*, the second the *variance*, and so forth. What is important for us here, is that two random variables, under very general conditions, have the same density if and only if all their moments are identical [6, 6.7.8; *cf*. 6.3.3]. And, if they both have the same mgf, their moments all will be identical.

The mgf of a random variable *X* is defined [6, Sect. 7.1] for continuous distributions on $(-\infty,\infty)$ as

$$\text{mgf}(t;X) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} dx \cdot e^{tx}\,\text{f}(X=x)\text{,} \tag{18}$$

in which $\text{f}(X=x)$ is the probability density function of *X*. The moments will be obtained by a power series expansion on *t*; the interval of series convergence does not concern us here.

The rest is very simple: From (14),

$$f_{\chi^2(1)}(X=x) \equiv \text{gamma}\left(\frac{1}{2},\frac{1}{2}\right) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{x}}e^{-\frac{x}{2}}\text{; so, from (18),}$$

$$\text{mgf}\left(t;\chi^2(1)\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dx \cdot e^{x\left(t-1/2\right)}x^{-\frac{1}{2}}\text{.} \tag{19}$$

Changing variables in (19) with $y = x^{1/2}$ and fixing the lower limit of integration immediately yields, for chi square,

$$\mathrm{mgf}\left(t; \chi^2(1)\right) = \frac{1}{\sqrt{2\pi}} \int_0^\infty dy \cdot e^{-\frac{(1-2t)}{2} y^2} \quad . \tag{20}$$

On the other hand, the unit normal density (1) may be written,

$$\mathrm{f}_{N(0,1)}(\mathbf{X} = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad . \tag{21}$$

So, the mgf for X^2 will be,

$$\mathrm{mgf}\left(t; N^2(0,1)\right) = \frac{1}{\sqrt{2\pi}} \int_0^\infty dx \cdot e^{tx^2 - x^2/2} = \frac{1}{\sqrt{2\pi}} \int_0^\infty dx \cdot e^{-\frac{(1-2t)}{2} x^2} \quad , \tag{22}$$

which clearly is the same as for the chi square in (20).

While on mgfs, it should be mentioned that the mgf also is perhaps the easiest way to derive the variance of the Poisson distribution [7 vol I, Sect.6.5 - 6.7] . The Poisson distribution, which gives the (discrete) probability of *k* events each with small probability λ, has probability function [6, 6.3.7],

$$P(\mathbf{X} = k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{23}$$

and may be shown to have mean = variance = λ. For *n* repeated Poisson trials, the mean will be $nE(X) = n\lambda$, and the variance will be $nV(X) = n\lambda$. For *n* repeated trials, the standard deviation of the Poisson mean will be $\sqrt{nV(X)}/\sqrt{n} = \sqrt{n\lambda}/\sqrt{n} = \sqrt{\lambda}$; so, by the central limit theorem, we would expect a Poisson distribution sum of *n* repeated trials with mean λ to approach a normal distribution with mean $n\lambda$ and standard deviation of the mean $\sqrt{n\lambda}$ (= standard error).

Sufficiency

A sample of a random variable may be seen as a measure of the information known about the underlying physical process. The sample size and the variance of the measurement determine the information in the sample: Variance adds unusable information (possible interpretations of the physical process) and so reduces the precision of the sample in representing the physics. Sample size decreases the variance of the mean of the sample and so increases the precision of the sample in representing the mean of the physics. Because physical processes are of little value unless repeatable, and because the mean may be used as an expectation of future

repetitions, size of the sample and variance of the measure work inversely in determining the information in a random sample.

Given a variety of statistics describing a random sample, a *sufficient* statistic [6, Sect. 10.4] is one which describes the data with no loss of information, so that information beyond that of the sufficient statistic does not improve prediction of future samples. Given the same physical process, a set of measurements of small variance generally will be *sufficient* to predict a set with greater variance: The latter set may be viewed as the former set plus some extra random variation.

Generally, sufficiency requires choice of a statistic good enough to represent the underlying physical process; otherwise, error in choice of statistic will add systematic variance which will reduce predictability. An example of this was in the introductory example above, in which the wrong choice of curve (quadratic) added systematic error--clearly systematic, because there was no random error.

In the present context, we may see chi square as a way to represent the goodness of choice of curve, such goodness being an unavoidable prerequisite to examination of the random error in the data. This goodness must be relative to the categories of the data, not to competing curves. But, in any case, we cannot look at sufficiency without goodness of fit.

So, how can we know that chi square might be useful generally as a test, when, perhaps, the chi-square assumption of a sum of squares of unit normal deviates might itself be bad?

Central Limit Theorem

The goodness of chi square as a test statistic for goodness of fit of reasonably large samples generally need not be questioned. The reason is the *central limit theorem* [7 vol II, Sect. 8.4 & Ch. 14], which holds under almost all conditions in which stable apparatus produces consistent data:

> *The sum X of any set of independent random variates, regardless of how distributed, will approach a normal distribution* $N\left(\mu_X, \sigma_X^2\right)$ *as the sample size approaches infinity.*

Of course, that being true, the sample mean will represent $E(X) \equiv \mu_X$ and the sample variance will represent $V(X) \equiv \sigma_X^2$.

As applied to the chi square question, if the data in each category fulfil the requirements of the central limit theorem, the mean in that category always may be standardized for use as a valid element of a chi square sum.

Furthermore, if errors of measurement caused by instrumentation are themselves subject to the central limit theorem, and therefore are normally distributed, it can be proven [7 vol II, Sect. 2.2] that the convolution of the instrumentation error with

the physical process also will be subject to the central limit theorem, assuming a reasonably consistent, repeatably measurable process.

Categories in Random-Sample Chi Square

Number of Categories

Above, it was stated that a sufficient statistic will have less variance than one which is not, for a good fit and the same data. From (17) and comments immediately following, as the *df* of chi square increases, its variance increases, too. However, to increase *df*, we change the experiment by changing the way the categories for averaging the data are defined. Chi square represents the categories, not the data.

In terms of the data, a set of means on many categories (in the limit, one category per datum) always may be combined to make fewer categories. In this sense, a chi square of greater *df* always will be sufficient relative to one of lesser *df*, for a given set of data. Looked at in terms of individual categories, a wider category means greater likelihood of a change in the physics, making for a bigger variance, given the same set of data being categorized. So, in chi square testing, more numerous, narrower categories (always fulfilling normality) will be sufficient relative to fewer, wider ones. The more the *df*, the closer to sufficiency.

Number of Parameters

For a given data set, it seems obvious that adding free theoretical parameters will make the fit of the theory better. In the limit, one always may fit N data categories by an $(N-1)$-th degree polynomial, resulting in zero variance in each category and therefore a perfectly good chi square fit. However, this would be very uninteresting physics--to claim, in the limit, that everything was a certain number of terms of its own, unique polynomial.

Clearly, a physical goodness of fit somehow should involve the complexity of the hypothesis, as well as the categorization of the data. The issue here is, how should we view the degrees of freedom in a goodness of fit test when we are free to vary both the data categories and the number of theoretical parameters? Is every theory with numerous parameters and wide-sweeping flexibility better than one which leaves some variance unaccounted?

Meaning of df

Bock [5, Sect. 2.6.3] defines degrees of freedom in terms of the dimension of a subspace on which the data are projected.

Brownlee [3, Sect. 8.1] defines degrees of freedom as "the number of variables minus the number of independent linear relations or constraints between them [and equal to the degrees of freedom of their sum of squares]."

In describing the chi square approximation to the multinomial distribution, Brownlee counts the degrees of freedom as the number of degrees of freedom for the chi square minus the number of parameters estimated in order to perform a curve fit. In an example [3, Sect. 5.2], estimation in the sample makes the *df* of the chi square equal to the number of categories minus 1. He assumes a Poisson distribution for which he subtracts another 1 from *df*, because the Poisson density has one free parameter. Thus, Brownlee's final *df* is equal to the number of categories minus 2.

If Brownlee's approach were applied to our problem of counting *df*, we would use the formula,

$$df = \left(N_C - 1\right) - N_P, \tag{24}$$

in which N_C is number of categories and N_P is number of free parameters. At least one popular statistical software package uses formula (24) for chi square goodness of fit.

For example, consider the artificial data in Figure 4b, and the effect of sequentially fitting them with a line of slope *b* and then with a constant function *a*, each time subtracting away the fit from the data:

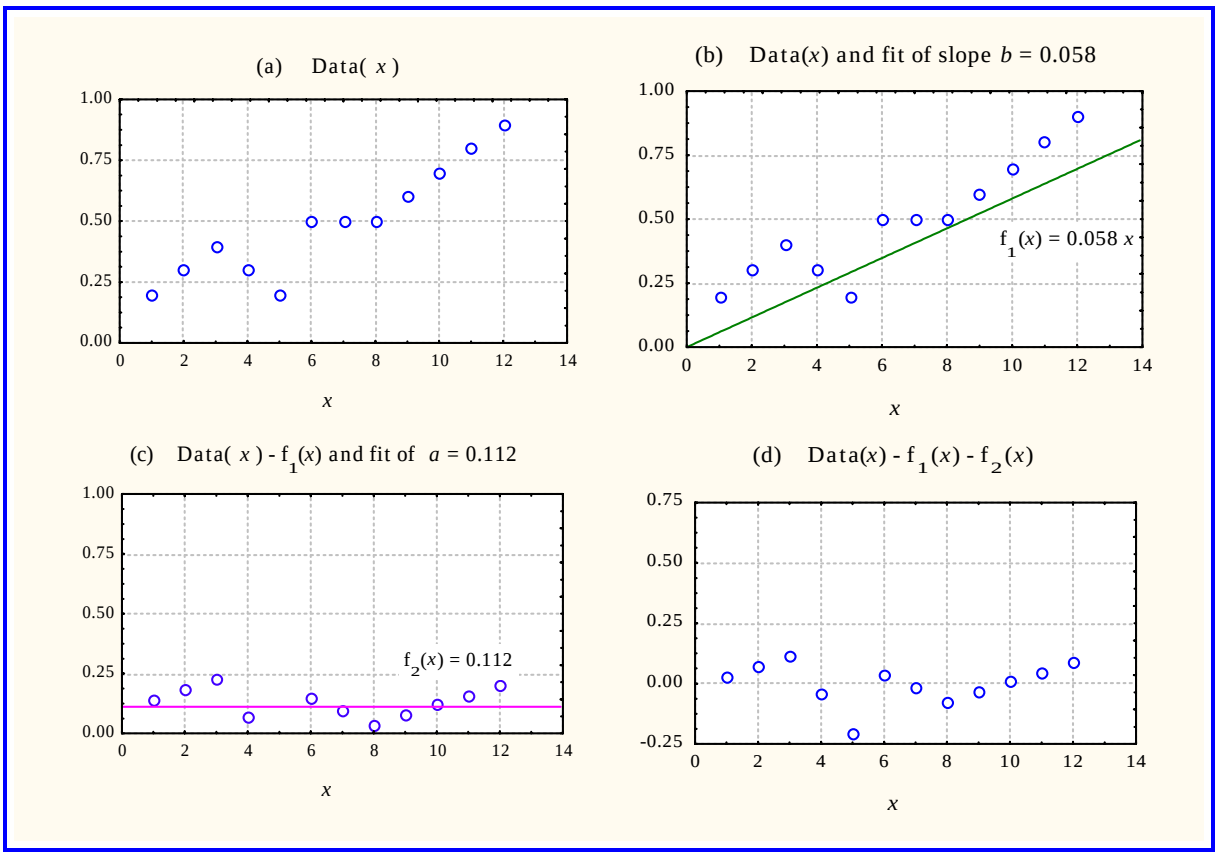

Figure 4b. Effect of linear parameters on residuals. The data were fit by f(*x*) = *a* + *bx* = 0.112 + 0.058*x* before the successive subtractions.

Using these two parameters *a* and *b* in a curve fit in some sense does reduce the orthogonality of the remainder (residuals). The smaller vertical range means the variance was reduced, too. With 12 data and 2 parameters to fit, it would seem reasonable in this example to use the rule above to test goodness using $\chi^2_p(9)$, with *df* computed as 9 = (12 - 1) - 2. For example, in Figure 4b, if theory forced *a* = 0.10, then only *b* would be fit, allowing a test on $\chi^2_p(10)$, and possibly making acceptance of the fit more likely for the same data. Clearly, in such an example, the acceptance would be more likely for a theoretical value of *a* equal to its value best fit to the data. So, a "stronger" theory which predicted parameters accurately without data would get preference in goodness. The *Galahad Effect*, one might say, in which a curve gains strength because its theory is pure.

The problem with the approach of simply subtracting the number of free parameters is that it treats everything as linear and orthogonal. It gives no more importance to a parameter of the curve than to just one of the possibly numerous categories. This approach seems unquestionably valid when the curve to be fit is linear. In other cases, although the categories might be linearly independent and

contribute equally to the random variance removed by the fit, constraints on the parameters in general will introduce nonlinear constraints on the category means.

Another, entirely different problem, is the efficiency of the estimator used for a parameter to be fit. Surely, we should subtract a *df* when testing a polynomial hypothesis, if we are fitting an intercept parameter using the data mean. But, what if we are using the less efficient, nonparametric statistic, the median, instead of the mean? The median only accounts for the rank order of the data and is invariant under any strictly monotonic transform of the data. Should we subtract 1/2 *df* if we use the median of the data to estimate the hypothetical intercept?

For a theorist interested in an objective test of a hypothesis with a small number *r* of free parameters, testing by (24) above on a large number of categories, say 10*r* or more, essentially ignores the number of parameters. Yet, under the null hypothesis that the data are well fit by the curve, the data in some sense should be expected to be interdependent in a way similar to the way they respond to small changes in the parameters.

Each parameter is *critical*, to an hypothesis with two parameters to fit. With 100 categories of data, though, dropping ten categories might not make much difference in the conclusion.

Need for a Correction to df

The practicality of the problem may be seen by comparing two hypothetical curves as shown here, one a sine curve with 3 free parameters (vertical displacement, phase, and frequency) and the other a 2-term polynomial (constant + coefficient of square).

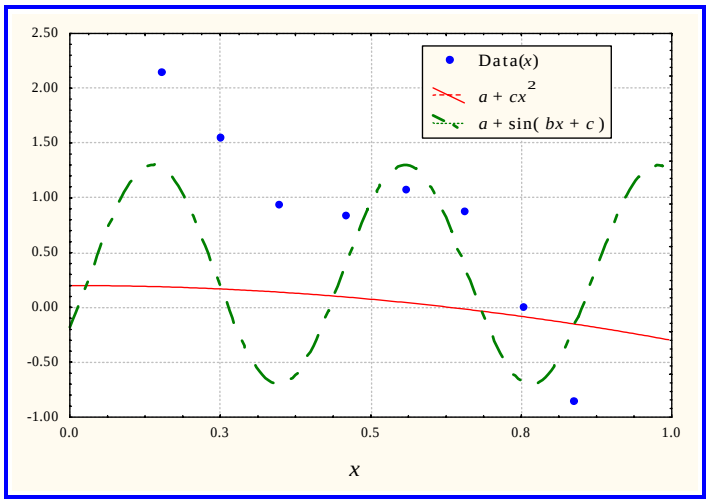

Figure 4a. Two ill-fit curves.

Suppose we were allowed to adjust the parameters shown. If the data ranged only over $a \pm 1$, the sine fit might be better than that of the polynomial; with the data as shown, a good sine fit might be hopeless. By the rule above, we would use a more stringent test (more likely to reject goodness) for the sine curve than for the polynomial.

We note that adding a free parameter to the sine (say, amplitude) or a couple of more terms to the polynomial, might make the fit good in both cases. Alas! In the context of a theory, many parameters will be constrained by the physics and

unavailable for arbitrary fitting. The problem of the present paper is that we cannot know in advance what the relation will be, for those that remain free to fit.

Clearly, ignoring the number of parameters in a goodness of fit test will make the test value of chi square such that $df_{test} = df_{categories}$, so that, for the same curve fit and data, acceptance of the null hypothesis (that the fit is good) will be favored. However, as above, merely subtracting the number of parameters from ($N_{categories} - 1$) would seem likely to undercorrect *df* except in the special case of linear regression.

We seek a way of correcting the *df* in a way equally fair, in some sense, to all curves to be fit.

III. The Parameter Correction

Definition of Correction

Development and Rationale

Before starting, we emphasize that we do not address the question of categorization of the data: Presumably, an optimum categorization already has been achieved, or perhaps the problem may not be flexible in these terms. We accept the categories of the data set as a given.

We also accept as fixed *a priori*, the level of significance *p* to be required for a rejection of goodness of fit.

A clue to the solution suggested here may be seen in the example in Figure 4b above: In that example, the data residuals were just as random as the original data unless fit by a straight line. If the residuals were fit by a straight line, the slope would be 0 and the intercept would be 0, because these orthogonal components were the ones subtracted down to 0. But, they were the only ones subtracted away, and no other structure in the data was affected. If we wished to test a linear fit to the residuals by chi square, we would subtract away $df = 2$; however, we assume here that to test a trigonometric fit, we would subtract away nothing from the *df*. Even though the variance of the residuals in Figure 4b was less than that of the original data, the residuals remain just as dimensionful as the data, except as projected on the intercept or the slope of a linear basis set.

Given categories *C*, N_C in number, we consider a set of arbitrary curves *K* to fit, each curve itself with an arbitrary and probably not unique set of parameters N_P in number. All such curves are assumed single-valued on the domain of the categories of data. The question is how to count the *df* to be used in a chi square test. This question is equivalent to the one of counting how many *df* should be subtracted from the number of data categories. We know that for realistic problems we must have $N_P < df < N_C$.

The answer we propose here is to account for the diversity of possible parameters by ignoring their number and considering only their effect. We therefore propose to *count* the degrees of freedom in the data by measuring the orthogonality of the data after the curve has been fit.

Here is how we prepare this quantification: For data on random variable *X* with domain $\{x\}$, described categorized by some function, $Data\{x\}$, we can compute the total variance in the data categories, $V_{cat}(Data)$. We choose the curve *K* to be fit, fit it, subtract it, and then examine the residual variance, $V_{res}(Data, K)$.

Under the null hypothesis that the fit was good, the residuals should represent solely random variation; they of course may not be assumed constrained by *K*, *K* having been removed. Under the null hypothesis, we assume here that the orthogonality of these residuals may be seen as being equal to the degrees of freedom of the curve fit to the original data, except when considered in regard to a fit by *K* itself.

Counting Process

So, having removed the effect of the curve *K*, we propose to count degrees of freedom by removing further variance by fitting the residuals with the sum of a set $L \equiv \{L(x)\}$ of linearly orthogonal functions, leaving us with a new residual variance, $V_{res}(Data, K, L)$. Statistical significance of the sum will be tested after adding each term of L to it; any term removing statistically significant variance will be omitted from L.

The size of the set L then will define the *df* of the chi square to be used in testing goodness of fit. We see no reason not to use an orthogonal set of polynomial functions; because we will enumerate terms in L beginning with L_0, we call the size, $n+1 = df$, the *nomial* of the fit:

$$df_{corrected} \equiv \text{nomial}(Data, K, L) = n(L) + 1,\ n \text{ s. t. } Res(Data, K, L_0) = \sum_{j=0}^{n} L_j(x), \qquad (25)$$

in which the *j* take on values of the degree of L not found statistically significant.

To see the logic of this, if only a few terms in L were required to reduce the residual variance $V_{res}(Data, K, L)$ below some threshold ε, then the residuals must have had low-order structure which was missed by *K*; so, the data categories were not orthogonal in view of the theory and indeed should be tested at lower *df*. On the other hand, if numerous terms were required in L to reduce the residuals below ε, the residuals must have had enough orthogonality to be tested at high *df*.

Stopping Criterion

The one remaining problem is the stopping criterion: How good a fit should we require of the orthogonal functions $\{L(x)\}$, none of which being significant?

We find an answer by again referring to the null hypothesis that the fit is good: Under this hypothesis, we assume that the initial curve-fit by *K* must have removed a statistically significant amount of variance; if so, $V_{cat}(Data) > V_{res}(Data, K)$; in fact, we may write this in parametric terms as,

$$V_{cat}(Data) \cdot F_p \geq V_{res}(Data, K), \qquad (26)$$

in which F_p is variance ratio at the one-tailed *p*-level for the Fisher analysis of variance *F*-test. So, to calculate where to terminate the nomial, we suggest applying this criterion:

$$\text{nomial}\left(Data, K, L_{\max}\right) - 1 \equiv Min\{n(L)\} \text{ s. t. } V_{res}\left(Data, K, L_0\right) \cdot F_p \geq V_{res}\left(Data, K, L_n\right). \quad (27)$$

This just sets a threshold which prevents the creation of nonexistent *df* by the orthogonal function fitting process: If the *n*-th polynomial leaves the *n*-th residuals with significantly less variance than the original *0*-th residuals, then all degrees of freedom (beyond the structure evidently overlooked by *K* in the data and now found in the original residuals) have been discovered.

A similar rationale is used sometimes in factor analysis (as distinguished from principle components analysis). It should be mentioned that the *df* for the significance level of F is not a difficult issue here, because we are looking at residuals after our problematical curve *K* has been fit and gone. The df for computing $V_{res}\left(Data, K, L_n\right)$ from a sum of squares of deviations will be just the *df* of the data, $(N_C - 1)$, minus the polynomial constraints $(n+1)$. This is because, in both cases, the residual variance will have the dimensionality of the data under the null hypothesis that the fit of *K* was good and therefore that L will be ineffective in explaining the data.

So, we have described a process to reduce an arbitrary problem in counting physical parameters as df, to one of counting linearly independent polynomial parameters as df.

The Correction Procedure

In the following, $V_{res}(\cdot)$ refers to a variance between categories and is computed from a sum of squares by dividing by a *df* depending on the number of categories and their linear independence. The sample variance in the original data, as categorized, is $V_{cat}(Data)$ and is a variance among the category means; in context of analysis of variance, this is called the "between treatments" variance.

We restate the problem: To perform a chi square goodness of fit test in the context of a physics experiment, assuming predetermined data categories *C* and significance level *p*, the degrees of freedom *df* may have to be adjusted for the free parameters in the curve to be fit. The proposed answer is:

1. Fit the curve *K* to the categorized data. Record the sum of squares of the residuals (differences or remainders) after the fit, standardizing each difference by the variance in its category. The result will be the value of chi square for the fit:

$$\chi^2_{Data} = \sum_{j=1}^{N_C} SS_j = \sum_{j=1}^{N_C} \frac{\left(\overline{X}_j - K_j\right)^2}{V_j} = \sum_{j=1}^{N_C} \left(\frac{\overline{X}_j - K_j}{sd_j} \right)^2 . \tag{28}$$

The standard deviation *sd* here is the standard deviation of the *j*-th category mean, also called the standard error, which should be estimated from the original data before the fit by *K*. Note that the category variance corresponding to chi square, which will not be used in the goodness of fit test, is obtained by dividing chi square by the *df*: $V_{res}(Data,K) = \chi^2_{Data}/df$, in which we approximate the *df* conventionally by $N_C - 1 - N_P$.

If *K* is a polynomial, count the number N_P of free parameters in the polynomial and compute $df = (N_C - 1) - N_P$: This immediately is the *df* sought. If *K* includes polynomial free parameters (a constant offset, for example), subtract the number of them before performing the goodness of fit test.

Also, establish a stopping criterion by looking up or calculating the Fisher variance ratio value *F* for one-tailed rejection of a null hypothesis test of *df* of $V_{cat}(Data)$ *vs.* *df* of $V_{res}(Data,K)$ at significance level *p*. This $\hat{F}_p$ may be found in statistics references in the context of analysis of variance. The significance test will not actually be performed; however its *df* will set $\hat{F}_p$ for the stopping criterion. One *df* will be that of the data, $N_C - 1$; the other will be that of the data with constraints of the curve fit *K* uncorrected, $N_C - 1 - N_P$. So, find,

$$F_p\left(df_{V(Data)}, df_{V(Data,K)}\right) \equiv \hat{F}_p . \tag{29}$$

Note that in general *K* will have removed considerable variance from the data, so an immediate *F* ratio test with,

$$F = \frac{V_{cat}(Data)}{V_{res}(Data,K)}, \tag{30}$$

easily might exceed $\hat{F}_p$. This will be irrelevant, because (*a*) we are not merely testing whether *K* says something about the data; and, (*b*) we cannot assume that the original variance over the whole set of categories was constant.

2. One sum at a time, progressively fit the residual data with the sum of the terms from an orthogonal polynomial set $\{L(x)\}$ and compute the new residual variance from that fit, $V_{res}(Data,K,L)$. When the nomial equals $N_C - 1$, stop. Or, stop when the stopping criterion is invoked: This will be when the ratio of variances among elements of *L* no longer is below $\hat{F}_p$, namely, when

$$F = \frac{V_{res}\left(Data, K, L_0\right)}{V_{res}\left(Data, K, L_n\right)} \geq \hat{F}_p . \tag{31}$$

For every L yielding a statistically significant fit, there was a linear dependence, so subtract 1 from the nomial for each such fit.

To compute the first and all subsequent values of $V_{res}\left(Data, K, L\right)$, compute the only variance present, namely that between the categories:

$$V_{res}\left(Data, K, L_n\right) \equiv \frac{1}{df_{res}} \sum_{j=1}^{N_C} SS_j , \tag{32}$$

with $df_{res} = N_C - 1 - (n+1)$, for L_n the highest-degree term in a polynomial of n-th degree.

3. Finally, to test the goodness of the fit, perform a standard chi square test with df equal to the nomial. In cases not requiring a correction, this df will equal $N_C - 1$.

Step 2 may be performed expeditiously by doing a parametric multiple regression of the residuals $V_{res}\left(Data, K\right)$ on a convenient set of orthogonal polynomials. A Chebyshev set of coefficients, or any of several others, may be found in [8] and used with PC software.

Application in Synthetic Examples

The value of using synthetic examples is that the underlying, correct functional form of the data will be known. Thus, we can predict how a good chi square goodness of fit test should behave.

To insist on stating the obvious, in the following examples, compared with the introductory ones above, we have a qualitatively different set of data. The next data below are simulated experimental data, with real variance. Uniformly distributed random numbers were used to add the variance.

In the following examples, we shall assume the central limit theorem so as to be able to treat each bin mean as normally distributed. In the first example below, we shall be computing the standard deviation in each bin using 14 -1 degrees of freedom; one degree is lost because the bin mean was estimated from the same 14 data. Knowing each bin mean and standard deviation, we shall compute the standard deviation of each data bin mean (standard error of the mean) and use it, not the corresponding theoretical fit parameter, to standardize each element of the chi square sum.

Fifty Categories in Random Chi Square

Quadratic Fit

Recall from Figure 4 that the quadratic fit with 50 categories visibly was not good. Let's look at the same data categories, but with the data randomized by addition of some variance.

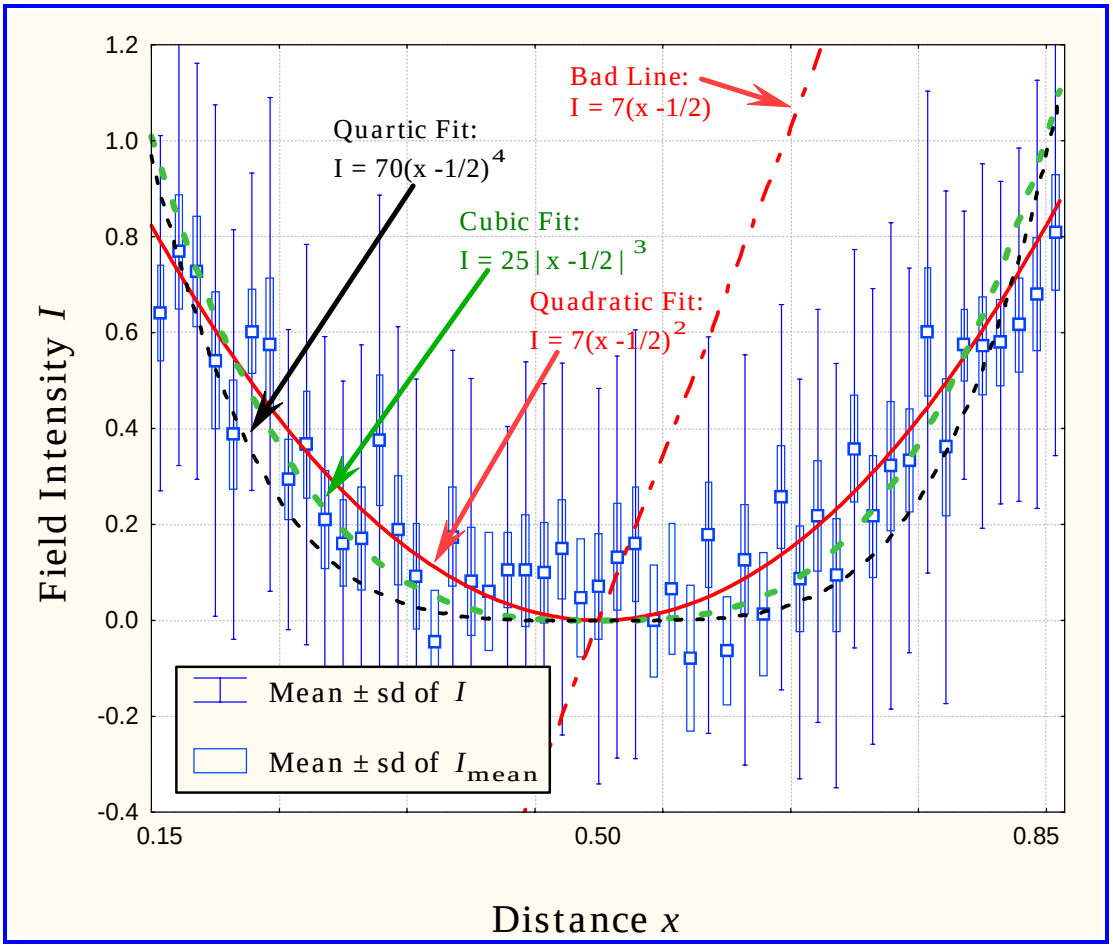

Figure 5. Fifty-category fits by eye. The 700 artificial data of Fig. 4 here are added some random error and again fit with the same categorized quadratic, $I = a(x-b)^2 = 7\left(x-1/2\right)^2$. A purposely bad linear fit also is shown, as is the old cubic fit by eye from Figure 1 above and a new quartic fit.

Computing χ^2_{Data} as in (5) or (28) above yields a value of about 35.8 for the solid line in Figure 5, which represents the quadratic curve. We know from published tables or calculation that $\chi^2_{.001}(13)$ = 34.5 and $\chi^2_{.001}(14)$ = 36.1. We find in the tables, also, that our stopping criterion will be $F_{.001}(49,47) = 2.50$. The quadratic has just two

polynomial parameters, so our correction should allow us immediately to assign a corrected *df* of ($N_C - 1$) - 2 = 47; however, we proceed with the complete correction analysis, to show how it works.

If we computed our correction to *df* = 14 or above, and if no term was significant, and no residual reached the L_0/L_n variance ratio stopping criterion of 2.50, we would have included sufficient *df* to know we could accept the null hypothesis that the fit was good.

The results of these computations are in Table 5. To reduce possible errors on this first real trial of the correction, the table was constructed by repeatedly fitting nonorthogonal polynomials with terms of the form, $a_n x^n$, as described for the correction above. The result was verified by repeatedly running multiple regression on an increasing number of terms of the Chebyshev orthogonal polynomials of the first kind. In practice, of course, only the regression would be necessary.

The low, relatively unchanging $\hat{F}$ ratios for the lowest-order nomials in Table 5 are a good indication that we should test at high *df*.

Table 5. Corrected *df* calculations for the 50-category quadratic in Fig. 5. All but the rightmost two columns were done by curve-fit with nonorthogonal $\sum a_n x^n$ polynomials, term by term. The "MR" columns were done by multiple regression on Chebyshev polynomials.

Raw Nom-ial	Description	Poly Residual SS	Poly Residual Variance	$\hat{F}$	MR Residual SS	MR Residual Variance
-	$\chi^2_{Data} =$	35.82	-	-	-	-
0	$V_{res}(Data, K, L_0)$	0.4725	0.00964	-	0.4725	0.00964
1	$V_{res}(Data, K, L_1)$	0.4695	0.00978	0.985	0.4695	0.00978
2	$V_{res}(Data, K, L_2)$	0.4130	0.00879	1.097	0.4128	0.00878
3	$V_{res}(Data, K, L_3)$	0.4127	0.00897	1.075	0.4127	0.00897
4	$V_{res}(Data, K, L_4)$	0.4082	0.00907	1.063	0.4079	0.00906
8	$V_{res}(Data, K, L_8)$	0.3643	0.00888	1.085	0.3613	0.00881
20	$V_{res}(Data, K, L_{20})$	0.3597	0.01240	0.778	(not done)	(not done)

None of the fits was significant, and all computed values of $\hat{F}$ stayed well below 2.50, so we are justified in testing at $\chi_p^2(20)$ or perhaps higher *df*. We know already that we cannot reject the null hypothesis at *df* = 14, so any higher *df* surely will yield a statistic sufficient to the same conclusion . Further analysis is not required, and according to our corrected chi square test, we accept the fit of the quadratic as good.

It should be mentioned here that a least-squares or other similar fit of a 20-term polynomial technically is not trivial, even on a fast computer. It would be rather amazing that such a polynomial really could be fit optimally to a set of data small enough to approach the number of terms being fit. But, we assume the best and look to the future for better.

Linear Fit

Table 5a shows the same test as Table 5, but for the obviously bad fit of a straight line passing on a steep slope through the middle of the data (Figure 5). Again, the two polynomial parameters should allow immediate correction to *df* = 47; we continue anyway, for demonstration purposes.

Table 5a. Computations as in Table 5 for corrected chi square goodness of fit test of the bad $I = a(x-b) = 7\left(x-1/2\right)$ line in Fig. 5. The stop on the third row makes the nomial = 1; however, the remainder is supplied to show the result of a bad fit. Terms significant in regression are shown in parentheses.

Raw Nomial	Description	Residual SS	Residual Variance	$\hat{F}$
-	$\chi^2_{Data} =$	106.85	-	-
0	$V_{res}\left(Data, K, L_0\right)$	101.58	2.073	-
1	$V_{res}\left(Data, K, L_1\right)$	2.8708	(*L_0,L_1) 0.05981	34.7
2	$V_{res}\left(Data, K, L_2\right)$	0.4128	(*L_0-L_2) 0.00878	236.1
3	$V_{res}\left(Data, K, L_3\right)$	0.4127	(*L_0,L_1) 0.00897	231.1
4	$V_{res}\left(Data, K, L_4\right)$	0.4079	0.009064	228.7
8	$V_{res}\left(Data, K, L_8\right)$	0.3616	0.008819	235.1
12	$V_{res}\left(Data, K, L_{12}\right)$	0.3607	0.009749	212.1

As before, $F_{.001}(49,47) = 2.50$. But, as obvious in Table 5a, the bad linear fit has left too much structure unaccounted and pushes $\hat{F}$ immediately out to 34.7, making the nomial 1; this would imply testing against $\chi_{.001}{}^2(1) = 10.8$. In any case, because $\chi_{.001}{}^2(49)$ is only about 85, the χ_{Data} value of 106.85 quickly allows us to reject the fit of the bad line as not good no matter what the *df*.

Cubic Fit

Once more, we look at a different curve fit to the data in Figure 5: This time, it is the cubic from Figure 1, replotted in Figure 5 and originally rejected as a fit to the data before random variation had been added. We again proceed with the analysis, although two polynomial parameters permit an immediate correction to *df* = 47.

The raw χ^2_{Data} for the cubic fit is found to be 64.25; because $\chi_{.001}{}^2(49)$ is about 85.3, the data value from the cubic might allow the fit to be rejected, if our correction were to reduce the *df* by enough. From published tables or calculation, we find that $\chi_{.001}{}^2(33) = 63.9$ and $\chi_{.001}{}^2(34) = 65.2$; so, if we retain a corrected *df* of less than 34, we will reject the fit of the cubic as not good.

As before, the correction stopping criterion will be $F_{.001}(49,47) = 2.50$. The results of the analysis are in Table 5b.

Table 5b. Computations as in Table 5 for corrected chi square goodness of fit test of the cubic $I = a(|x-b|)^3 = 25|x-1/2|^3$ in Fig. 5. Terms significant in regression are shown in parentheses.

Raw Nomial	Description	Residual SS	Residual Variance	$\hat{F}$
-	$\chi^2_{Data} =$	64.25	-	-
0	$V_{res}(Data, K, L_0)$	0.834828	0.017037	-
1	$V_{res}(Data, K, L_1)$	0.780223	0.016255	1.048
2	$V_{res}(Data, K, L_2)$	0.479143	(*L_0-L_2) 0.010195	1.671
3	$V_{res}(Data, K, L_3)$	0.476474	0.010358	1.645
4	$V_{res}(Data, K, L_4)$	0.395849	0.008797	1.937
5	$V_{res}(Data, K, L_5)$	0.394304	0.008961	1.901
7	$V_{res}(Data, K, L_7)$	0.364182	0.008671	1.965
8	$V_{res}(Data, K, L_8)$	0.360949	0.008804	1.935
12	$V_{res}(Data, K, L_{12})$	0.360192	0.009735	1.750
20	$V_{res}(Data, K, L_{20})$	0.352694	0.012162	1.400

Although Table 5b omits some results, the value of $\hat{F}$ wandered up to the maximum shown at a nomial of 7, then declined gradually to the nomial 20 value shown; this was the limit of the author's PC software. There is no sign the stopping criterion would be invoked, and only one nomial term was dropped because of significance, so the corrected chi square *df* would be $N_C - 1 - 1 - 2 = 46$, the final reduction by two being because the fit was a two-parameter polynomial. We repeat that the immediate correction, because of the two polynomial terms, by subtraction of 2 from 49, would leave a proper corrected *df* of 47; in effect, we pretended there was at least one nonpolynomial parameter to test the procedure. But, $46 >> 34$ still means that the cubic fit to the Figure 5 data would be accepted as good.

Quartic Fit

Let's look once again at a curve fit to the data in Figure 5: This time, it is the quartic, which was fit by eye.

The raw χ^2_{Data} for the quartic fit is found to be 83.01. From published tables or calculation, we find that $\chi_{.001}{}^2(47) = 82.7$ and $\chi_{.001}{}^2(48) = 84.04$; so, if we retain a corrected *df* of less than 48, we will reject the fit of the quartic as not good. The common practice in (24) above of setting *df* at $N_C - 1 - N_P$, or our strictly applied correction procedure, would assign *df* = 47, causing an immediate rejection of the fit. We pretend there is at least one nonpolynomial parameter and pursue our correction here to see how it performs in this interesting marginal case.

As before, the correction stopping criterion will be $F_{.001}(49,47) = 2.50$. The results of the correction analysis is in Table 5c.

Table 5c. Computations for corrected chi square goodness of fit test of the quartic $I = a(x-b)^4 = 70(x-1/2)^4$ in Fig. 5. Terms significant in regression are shown in parentheses.

Raw Nomial	Description	Residual SS	Residual Variance	$\hat{F}$
-	$\chi^2_{Data} =$	83.01	-	-
0	$V_{res}(Data, K, L_0)$	0.824809	(*L_0) 0.016833	-
1	$V_{res}(Data, K, L_1)$	0.766592	0.015971	1.054
2	$V_{res}(Data, K, L_2)$	0.658320	0.014007	1.202
3	$V_{res}(Data, K, L_3)$	0.650998	0.014152	1.189
4	$V_{res}(Data, K, L_4)$	0.407871	(*L_0-L_4) 0.009064	1.857
5	$V_{res}(Data, K, L_5)$	0.406808	0.009246	1.821
6	$V_{res}(Data, K, L_6)$	0.383786	0.008925	1.886
7	$V_{res}(Data, K, L_7)$	0.366298	0.008721	1.930
8	$V_{res}(Data, K, L_8)$	0.361569	0.008819	1.909
12	$V_{res}(Data, K, L_{12})$	0.360525	0.009744	1.728
20	$V_{res}(Data, K, L_{20})$	0.352887	0.012169	1.383

As shown in Table 5c, the value of $\hat{F}$ wandered up to a maximum again at a nomial of 7, but the stopping criterion was not invoked. However, the low-degree analysis was enough to reveal two terms of significant residuals; the pretend-corrected chi square *df* would be $N_C - 1 - 2 - 2 = 45$, and the quartic fit to the Figure 5 data is rejected as not good. Actually, we of course would have rejected the fit at 49 - 2 = 47 *df*, because there were just two polynomial parameters and no others.

Finally, before leaving this example, we should remark that the same quadratic function which failed miserably in Figure 4 easily was shown a good fit in Figure 5. The dispersion in the Figure 5 data clearly weakened the test. In fact, even the cubic of Figure 1 was accepted as a good fit. When the randomness is great enough, almost any fit will be good enough--but, not much confidence in the hypothesis will be added by not rejecting it, either.

We consider next a more realistic simulated example.

Application in a Twenty Category Decay Simulation

Assume there is a theory which predicts that a certain material, initially translucent, will fluoresce if illuminated with light at a certain, short wavelength. Let's consider a hypothetical experiment in which some of this material is enclosed in a spherical detector, a glowball, and irradiated with a pulse of short-wavelength light. We wish to predict the time course of the glow inside the sphere.

The hypothesis is that a 4π Planckian detector will respond as the convolution of the Gaussian light exciting impulse with a decaying negative exponential of the form,

$$E(t;w,a|I_0) = I_0 \int_0^t d\tau \cdot e^{-\frac{1}{2}(\tau/w)^2} e^{-(t-\tau)/a}, \tag{33}$$

which has two free parameters, pulse width *w* and time-constant *a*, the initial intensity I_0 being fixed by theory.

The result of the simulated experiment is given in Figure 6.

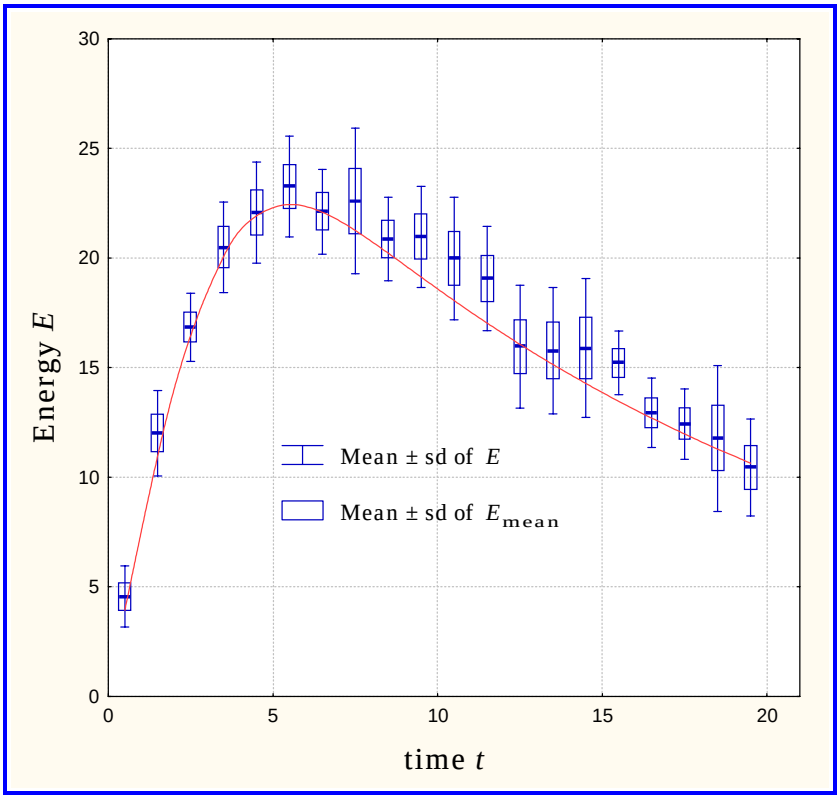

Figure 6. Glowball decay experiment. Time *t* and energy *E* are in arbitrary units. The solid line represents a fit of formula (33) above, with I_0 = 8 fixed by theory and *w* = 2.9 and *a* = 17 fit by eye to the 20 bin means shown.

From published tables or calculation, we find that $\chi^2_{.001}(19) = 43.8$ and $F_{.001}(19,17) = 4.81$. Computing χ^2_{Data}, we get over 75. Therefore, the null hypothesis of a good fit is rejected at once.

The analysis, although unnecessary for this particular decision, is given in Table 6. Looking at the several low-degree values of L in the table, it is obvious they are changing slowly, and, clearly, after computing L_2, it would be very reasonable to conclude that a test at $df = N_C - 1 - 1$ was justified.

Table 6. Glowball decay experiment. Computations for corrected chi square goodness of fit of the curve in Fig. 6, which was text equation (33) fit by eye. Two free parameters, *w* and *a*, were available to fit. Terms in parenthesis were significant in regression.

Raw Nomial	Description	Residual SS	Residual Variance	$\hat{F}$
-	$\chi^2_{Data} =$	75.7	-	-
0	$V_{res}(Data, K, L_0)$	9.551	(*L_0) 0.503	-
1	$V_{res}(Data, K, L_1)$	9.551	0.531	0.947
2	$V_{res}(Data, K, L_2)$	7.631	0.449	1.120
3	$V_{res}(Data, K, L_3)$	6.493	0.406	1.239
4	$V_{res}(Data, K, L_4)$	6.226	0.415	1.212
8	$V_{res}(Data, K, L_8)$	5.432	0.494	1.018
12	$V_{res}(Data, K, L_{12})$	5.559	0.794	0.634

None of the computed values of $\hat{F}$ exceeded even 2, although the L_0 term would not be counted; so, if the initial value of 75.7 for chi square had not been so obvious, we would have performed the corrected test against $\chi^2_{.001}(N_C - 1 - 1) = \chi^2_{.001}(18)$.

But what if theory allowed us to adjust the third parameter, I_0, for a better fit? The result of a three-parameter fit is shown in Figure 7 and is analyzed in Table 7:

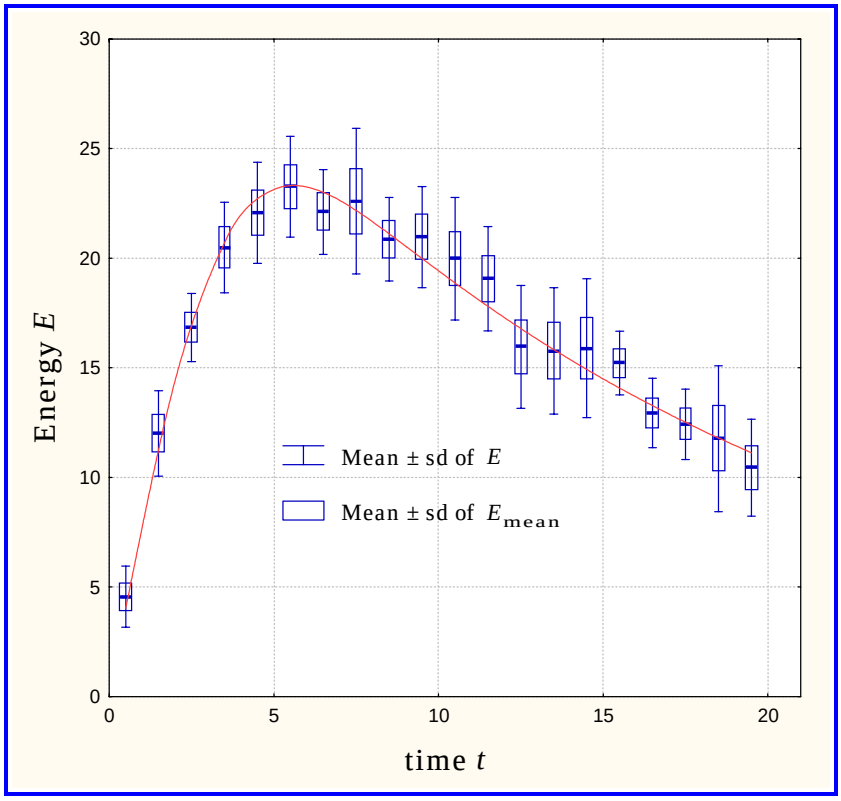

Figure 7. Glowball decay experiment. Analysis with three free parameters. Time *t* and energy *E* are in arbitrary units. The solid line represents a fit of formula (33) above, with I_0 = 8.2, *w* = 2.95, and *a* = 17 as fit by eye again to the same 20 bin means as in Fig. 6.

As shown below in Table 7, χ^2_{Data}, with an additional free parameter, now is just 34.9, which is below $\chi^2_{.001}(19) = 43.8$. Therefore, assuming a correction for free parameters, no immediate decision can be made. Because the lowest *df* allowing acceptance of the fit is 14 ($\chi^2_{.001}(14) = 36.1$), we seek a decision as to whether the *df* might be 14 or more.

The correction analysis is done in Table 7. In that Table, the wandering of the residual variance and the lack of consistent increase of $\hat{F}$ make it quite certain that the critical *F* value of 4.81 never will be reached. Therefore, the fit of the theory using three free parameters may be accepted as good.

Compare the appearance of the two fits in Figures 6 and 7: The eye can hardly decide, whereas the statistics are unequivocal, given the category set and the significance level.

Table 7. Glowball decay data as in Table 6. Computations for corrected chi square goodness of fit of the curve in Fig. 7. Three parameters, I_0, w, and a, were free to fit.

Raw Nomial	Description	Residual SS	Residual Variance	$\hat{F}$
-	$\chi^2_{Data} =$	34.97	-	-
0	$V_{res}(Data, K, L_0)$	9.179	0.4831	-
1	$V_{res}(Data, K, L_1)$	9.173	0.5096	0.948
2	$V_{res}(Data, K, L_2)$	8.868	0.5216	0.926
3	$V_{res}(Data, K, L_3)$	6.622	0.4139	1.167
4	$V_{res}(Data, K, L_4)$	6.300	0.4200	1.150
8	$V_{res}(Data, K, L_8)$	5.433	0.4939	0.978
12	$V_{res}(Data, K, L_{12})$	5.572	0.7960	0.607
13	$V_{res}(Data, K, L_{13})$	5.516	0.9193	0.526
14	$V_{res}(Data, K, L_{14})$	5.385	1.0771	0.449
15	$V_{res}(Data, K, L_{15})$	5.199	1.2998	0.372

Application in Neutrino Oscillation Theory

Muon neutrinos are created in great numbers in the upper atmosphere by cosmic particles. They penetrate the Earth and may be detected both in a downward direction, in which they traverse a short distance of atmosphere, and in an upward direction, in which they traverse the diameter of the Earth. To explain an apparent deficit in upward *vs* downward muon neutrinos, relative to electron neutrinos similarly created, one theory holds that the muon neutrinos oscillate or transform into other types not easily identified.

Data from the Super-Kamiokande water-Cerenkov neutrino detector [9, Table I] are plotted in Figure 8. The "statistical error" plotted is the standard deviation of the assumed-Poisson bin totals, which is to say, $\sqrt{(\text{bin total})}$ for each bin; this is the same as the standard error of the bin mean, as described near (23) above. Neutrino

and antineutrino data are combined. The over-1-GeV events (multiGeV events) were substantially over 1.0 GeV. Higher-energy events are more easily distinguished from background than those of lower energy; this causes most of the separation, even though atmospheric neutrinos tend to average about 1 GeV.

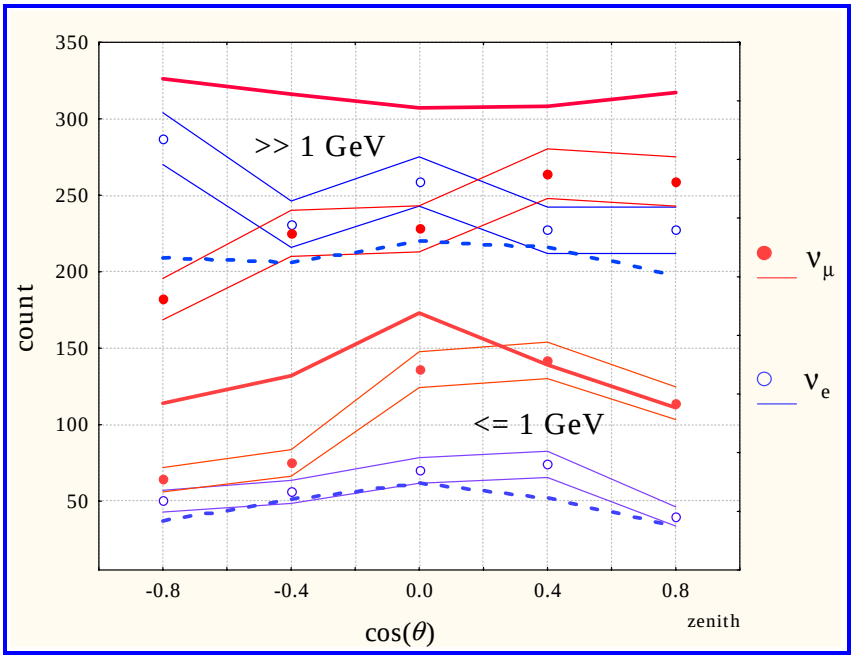

Figure 8. Super-Kamiokande atmospheric neutrino data from [9] after about 535 days online, categorized into five counting bins by cosine of the angle from the zenith. The electron-like vs muon-like events are shown as ν_e and ν_μ, respectively. The heavier single lines connect means of Monte Carlo simulations (theory); the event data means are plotted as points, with the statistical error enclosed in the lighter lines and representing ±1 standard deviation of the bin mean.

The "theory" was obtained by Monte Carlo simulation, because the detector calibrations are very complicated although well understood. The Monte Carlo means ignore any directional bias in neutrino flux; so, if the data fit the theory, we would conclude there was no atmospheric deficit and hence no observable oscillation.

Because Monte Carlo simulation is stochastic, the resulting theory will itself include some randomness (variance), representing uncertainty in the estimates obtained during simulation. In the example [9] chosen, the Monte Carlo standard errors of the means were substantial compared with those of the data, but we chose to ignore them entirely. Treatment of such variances is discussed in an appendix in [9]; in general, if considered truly to measure the uncertainty in the expectancies to be tested, they should be added to the variances of the data categories.

Monte Carlo expectancies enter into the chi square elements the same way as do those of the data, so it may be assumed that the *df* correction proposed here also will reveal dependencies or other artifacts introduced by the Monte Carlo model and not fit by the curve being evaluated. However, if we assume the Central Limit Theorem to apply to the Monte Carlo expectancies (something not so obvious as that Theorem's application to the physical data or instrumentation), the major effect of the stochasticity will be to add to the variance in the residuals, weakening the chi square test and biasing it toward acceptance.

For purposes of chi-square testing, one would recommend avoidance of a Monte Carlo theory which generated nonnormal category expectancies or ones with category variance of the order of that of the data alone. However, running multiple Monte Carlo simulations and averaging the result by bin should correct nonnormality and high variance for almost any model

If the Monte Carlo model was written to model the variances as well as the expectancies, the issue would be how the expectancies were computed--an issue internal to the model. It is unclear what would be the value explicitly of modelling the variance, unless to confirm correctness of the Monte Carlo computer software independent of the physics. The same concept of sufficiency should be used to evaluate a Monte Carlo model as would be used to evaluate a statistic.

In any event, we wish to see how our *df* correction may be applied in tests of the goodness of fit of the neutrino Monte Carlo theory to the data. There are four different fits to test. Summary calculations are in Table 8. $F_{.001}(4,4) = 53.4$

Table 8. Correction of Monte Carlo fit to neutrino data from [9]. Each event type has $N_C = 5$ data bins, so the maximum *df* would be 4, with no correction. All nomial calculations were done by multiple regression on Chebyshev orthogonal polynomials of the first kind.

Event Type	Raw Nomial	Description	Residual SS	Resid. Variance	$\hat{F}$
SubGeV	-	$\chi^2_{Data} =$	34.09	-	-
electron	0	$V_{res}(Data, K, L_0)$	2567.3	641.80	-
	1	$V_{res}(Data, K, L_1)$	1312.8	437.6	1.467
	2	$V_{res}(Data, K, L_2)$	598.51	299.26	2.145
	3	$V_{res}(Data, K, L_3)$	554.41	554.41	1.158
SubGeV	-	$\chi^2_{Data} =$	198.3	-	-
muon	0	$V_{res}(Data, K, L_0)$	5946.8	1486.7	-
	1	$V_{res}(Data, K, L_1)$	1150.70	383.57	3.876
	2	$V_{res}(Data, K, L_2)$	270.63	135.31	10.987
	3	$V_{res}(Data, K, L_3)$	264.23	264.23	5.627
MultiGeV	-	$\chi^2_{Data} =$	12.2	-	-
electron	0	$V_{res}(Data, K, L_0)$	194.8	48.700	-
	1	$V_{res}(Data, K, L_1)$	193.9	64.633	0.753
	2	$V_{res}(Data, K, L_2)$	192.1	96.057	0.507
	3	$V_{res}(Data, K, L_3)$	24.01	24.01	2.028
MultiGeV	-	$\chi^2_{Data} =$	92.1	-	-
muon	0	$V_{res}(Data, K, L_0)$	3327.2	831.8	-
	1	$V_{res}(Data, K, L_1)$	571.60	190.53	4.366
	2	$V_{res}(Data, K, L_2)$	489.03	244.51	3.402
	3	$V_{res}(Data, K, L_3)$	40.129	40.129	20.73

None of the regression coefficients was significant, and our stopping criterion never was invoked, so the goodness of fit criterion will be $\chi^2_{.001}(4) = 18.5$. We thus conclude from Table 8 that the Monte Carlo was a good fit for the multiGeV electron-like events but that there was a statistically significant effect in the other cases, reflecting the expected atmospheric deficit in muon-neutrinos.

IV. Conclusion

We suggest that the proposed correction to *df* is almost trivial conceptually and is not difficult to use in practice.

We believe it has been shown to perform as hoped: It equalizes the meaning of statistical significance for goodness of fit, as applied to theoretical curves with any number of physically important free parameters.

References

[1] R. A. Fisher, *The Design of Experiments*, Edinburgh: Oliver and Boyd (1935).

[2] R. L. Winkler, *Introduction to Bayesian Inference and Decision*, San Francisco: Holt, Rinehart, and Winston (1972).

[3] K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering* (2nd ed), New York: John Wiley (1965).

[4] M. H. De Groot, *Optimal Statistical Decisions*, San Francisco: McGraw-Hill (1970).

[5] R. D. Bock, *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill (1975).

[6] J. R. Blum and J. I. Rosenblatt, *Probability and Statistics*, Philadelphia: W. B. Saunders Co. (1972).

[7] W. Feller, *Introduction to Probability Theory and Its Applications, Volume I & II*, New York: John Wiley & Sons (1966).

[8] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions (Applied Mathematics Series **55**)*, Chapter 22, Washington, DC: U. S. National Bureau of Standards (1964).

[9] G. L. Fogli, *et al*, "Super-Kamiokande atmospheric neutrino data, zenith distributions, and three-flavor oscillations", preprint *BARI-TH/309-98* (August 1998).

Acknowledgements

Levels of significance and multiple regressions were computed using *STATISTICA* 5.1. Chebyshev coefficients were generated, and integrals were evaluated, in *MathCAD* 4.0.