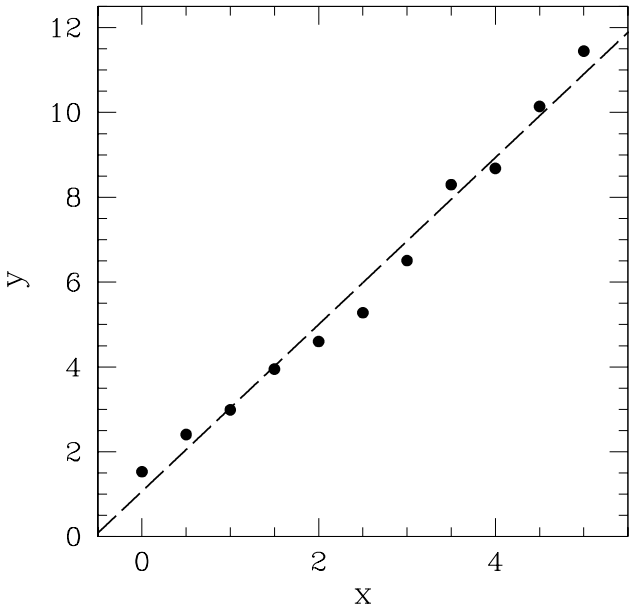
Physics 115/242
Least squares fitting.

Peter Young

(Dated: April 28, 2014)

I. INTRODUCTION

Frequently we are given a set of data points $(x_i, y_i), i = 1, 2, \cdots, N$, through which we would like to fit to a smooth function. The function could be straight line (the simplest case), a higher order polynomial, or a more complicated function. As an example, the data in the figure below seems to follow a linear behavior and we may like to determine the "best" straight line through it. More generally, our fitting function, $y = f(x)$, will have some adjustable parameters and we would like to determine the "best" choice of those parameters.



The definition of "best" is not unique. However, the most useful choice, and the one nearly always taken, is the "least squares" fit, in which one minimizes the sum of the squares of the difference between the observed y-value, y_i, and the fitting function evaluated at x_i, i.e. one minimizes

$$\boxed{\sum_{i=1}^{N} [\, y_i - f(x_i) \,]^2 \ .} \tag{1}$$

The simplest cases, and the only ones to be discussed in detail here, are where the fitting function is a *linear function of the parameters*. We shall call this a *linear model*. Examples are a straight line

$$y = a_0 + a_1 x \tag{2}$$

and an m-th order polynomial

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = \sum_{\alpha=0}^{m} a_\alpha x^m \,, \tag{3}$$

where the parameters to be adjusted are the a_α. (Note that we are *not* requiring that y is a linear function of x, only of the fit parameters a_α.)

An example where the fitting function depends *non*-linearly on the parameters is

$$y = a_0 x^{a_1} + a_2 \,. \tag{4}$$

Linear models are fairly simply because, as we shall see, for a least-squares fit the parameters are determined by *linear* equations, which, in general, have a unique solution which can be found by straightforward methods. However, for fitting functions which are non-linear functions of the parameters, the resulting equations are *non-linear* which may have many solutions or non at all, and so are much less straightforward to solve (see Numerical Recipes, Sec. 15.5, in 2nd. edition, for a discussion of the Levenberg-Marquardt method for fitting to non-linear models). One of the main reasons why least-squares fitting is almost always used (rather than a different definition of the "best" fit) is that the equations are linear for a linear model. This is not true for other types of fit.

Sometimes a non-linear model can be transformed into a linear model by a change of variables. For example if we want to fit to

$$y = a_0 x^{a_1} \,, \tag{5}$$

which has a non-linear dependence on a_1, then taking logs gives

$$\ln y = \ln a_0 + a_1 \ln x \,, \tag{6}$$

which is a *linear* function of the parameters $a_0' = \ln a_0$ and a_1. Fitting a straight line to a log-log plot is a very common procedure in science and engineering. However, later we will include error bars on the data points, and we often assume that the error bars have a Gaussian distribution, and we should note that transforming the data does not exactly take Gaussian errors into Gaussian

errors, though the difference will be small if the errors are "sufficiently small". For the above log transformation this means that $\sigma_i/y_i \ll 1$ (where σ_i is the error in data point y_i), i.e. the *relative* error is much less than unity.

II. FITTING TO A STRAIGHT LINE

To see how least squares fitting works, consider the simplest case of a straight line fit, Eq. (2), for which we have to minimize

$$F(a_0, a_1) = \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i \right)^2 , \tag{7}$$

with respect to a_0 and a_1. Differentiating F with respect to these parameters and setting the results to zero gives

$$\sum_{i=1}^{N} (a_0 + a_1 x_i) = \sum_{i=1}^{N} y_i, \tag{8a}$$

$$\sum_{i=1}^{N} x_i \, (a_0 + a_1 x_i) = \sum_{i=1}^{N} x_i y_i. \tag{8b}$$

We write this as

$$U_{00}\, a_0 + U_{01}\, a_1 = v_0, \tag{9a}$$

$$U_{10}\, a_0 + U_{11}\, a_1 = v_1, \tag{9b}$$

where

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} x_i^{\alpha+\beta},} \quad \text{and} \tag{10}$$

$$\boxed{v_\alpha = \sum_{i=1}^{N} y_i \, x_i^{\alpha} \, .} \tag{11}$$

The matrix notation, while an overkill here, will be convenient later when we do a general polynomial fit. Note that $U_{10} = U_{01}$. (More generally, later on, U will be a symmetric matrix). Equations (9) are two linear equations in two unknowns. These can be solved by eliminating one variable, which immediately gives an equation for the second one. The solution can also be determined from

$$\boxed{a_\alpha = \sum_{\beta=0}^{m} \left(U^{-1}\right)_{\alpha\beta} \, v_\beta,} \tag{12}$$

where the inverse of the 2×2 matrix U is given, according to standard rules, by

$$U^{-1} = \frac{1}{\Delta} \begin{pmatrix} U_{11} & -U_{01} \\ -U_{01} & U_{00} \end{pmatrix} \tag{13}$$

where

$$\boxed{\Delta = U_{00}U_{11} - U_{01}^2,} \tag{14}$$

and we have noted that U is symmetric so $U_{01} = U_{10}$. The solution for a_0 and a_1 is therefore given by

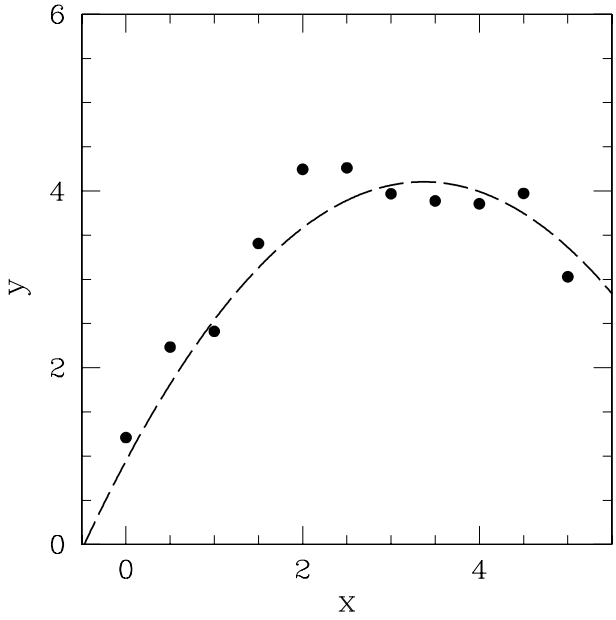
$$\boxed{a_0 = \frac{U_{11}\, v_0 - U_{01}\, v_1}{\Delta},} \tag{15a}$$

$$\boxed{a_1 = \frac{-U_{01}\, v_0 + U_{00}\, v_1}{\Delta}.} \tag{15b}$$

We see that it is straightforward to determine the slope, a_1, and the intercept, a_0, of the fit from Eqs. (10), (11), (14) and (15) using the N data points (x_i, y_i).

III. FITTING TO A POLYNOMIAL

Frequently it may be better to fit to a higher order polynomial than a straight line, as for example in the plot below where the fit is a parabola.



Using the notation for an m-th order polynomial shown in Eq. (3), we need to minimize

$$F(a_0, a_1, \cdots, a_m) = \sum_{i=1}^{N} \left(y_i - \sum_{\alpha=0}^{m} a_\alpha x_i^\alpha \right)^2 \tag{16}$$

with respect to the $N_{\rm fit} = m + 1$ parameters a_α. Setting to zero the derivative of F with respect to a_α gives

$$\sum_{i=1}^{N} x_i^\alpha \left(y_i - \sum_{\beta=0}^{m} a_\beta x_i^\beta \right) = 0, \tag{17}$$

which we write as

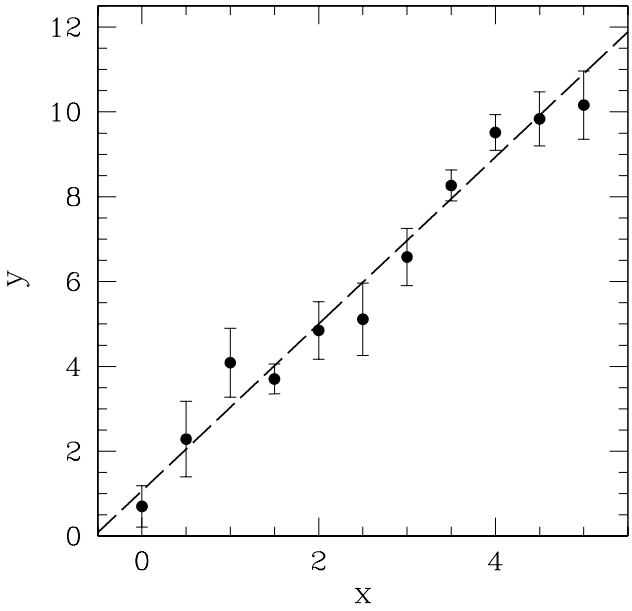
$$\boxed{\sum_{\beta=0}^{m} U_{\alpha\beta}\, a_\beta = v_\alpha,} \tag{18}$$

where $U_{\alpha\beta}$ and v_α have been defined in Eqs. (10) and (11). Eq. (18) represents $m + 1$ *linear* equations, one for each value of α. Their solution is given formally by Eq. (12).

Hence polynomial least squares fits, being linear in the parameters, are also quite straightforward. We just have to solve a set of linear equations, Eq. (18), to determine the optimal fit parameters.

IV. FITTING TO DATA WITH ERROR BARS

Frequently we have an estimate of the uncertainty in the data points, the "error bar". A fit would be considered satisfactory if it goes through the points "within the error bars" (we will discuss at the end of this handout more precisely what this means). An example of data with error bars and a straight line fit is shown in the figure below.



If some points have smaller error bars than other we would like to force the fit to be closer to those points. A suitable quantity to minimize, therefore, is

$$\boxed{\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 ,} \tag{19}$$

called the "chi-squared" function , in which σ_i is the error bar for point i. Note that χ^2 should be thought of as a single variable rather than the square of something called χ. This notation is standard. Assuming a polynomial fit, we proceed exactly as before, and again find that the best parameters are given by the solution of Eq. (18), i.e.

$$\sum_{\beta=0}^{m} U_{\alpha\beta}\, a_\beta = v_\alpha, \tag{20}$$

but where now $U_{\alpha\beta}$ and v_α are given by

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} \frac{x_i^{\alpha+\beta}}{\sigma_i^2},} \quad \text{and} \tag{21}$$

$$\boxed{v_\alpha = \sum_{i=1}^{N} \frac{y_i\, x_i^\alpha}{\sigma_i^2}\,.} \tag{22}$$

The solution of Eqs. (20) can be obtained from the inverse of the matrix U, as in Eq. (12).

Interestingly, the matrix U^{-1} contains additional information. We would like to know the *range* of values of the a_α which provide a suitable fit. The y_i can vary within an amount σ_i, and this

variation will cause a change in the values of the optimal fit parameters a_α. We write

$$\sigma_\alpha^2 = \langle \delta a_\alpha^2 \rangle \tag{23}$$

where $\langle \cdots \rangle$ denotes an average of the different values of the y_i with the appropriate weight, and $\delta a_\alpha = a_\alpha - \langle a_\alpha \rangle$. It turns out that the square of the uncertainty in a_α is just the corresponding diagonal element of U^{-1}, i.e.

$$\boxed{\sigma_\alpha^2 = \left(U^{-1}\right)_{\alpha\alpha}} , \tag{24}$$

where the elements of the matrix U are given by Eq. (21). I emphasize that the uncertainty in a_α is σ_α, not the square or this.

For the case of a straight line fit, the inverse of U is given explicitly in Eq. (13). Using this information, and the values of (x_i, y_i, σ_i) for the data in the above figure, I find that the fit parameters (assuming a straight line fit) are

$$a_0 = 0.840 \pm 0.321, \tag{25}$$

$$a_1 = 2.054 \pm 0.109, \tag{26}$$

in which the error bars on the fit parameters on a_0 and a_1, i.e. σ_0 and σ_1, are determined from Eq. (24). I had generated the data by starting with $y = 1 + 2x$ and adding some noise with zero mean. Hence the fit should be consistent with $y = 1 + 2x$ within the error bars, and it is.

We call U^{-1} the "covariance matrix". Its off-diagonal elements are also useful since they contain information about correlations between the fitted parameters. More precisely defining the covariance of fit parameters α and β, we have

$$\boxed{\mathrm{Cov}(\alpha, \beta) \equiv \langle \delta a_\alpha \, \delta a_\beta \rangle = \left(U^{-1}\right)_{\alpha\beta}} . \tag{27}$$

The correlation coefficient, $r_{\alpha\beta}$, which is a dimensionless parameter lying between -1 and 1 and is a measure of the correlation between δa_α and δa_β, is given by

$$\boxed{r_{\alpha\beta} = \frac{\mathrm{Cov}(\alpha, \beta)}{\sigma_\alpha \sigma_\beta}} . \tag{28}$$

It is all very well to get some fit parameters and error bars but they don't mean much unless the fit really describes the data, which means, roughly speaking, that it **goes through the data within the error bars**. To see if this is the case, we look at the value χ^2, Eq. (19), with the optimal parameters. If the fit goes through the data within about one error bar, then χ^2 should be about N. To be more precise, we have to take into account that we *adjusted several parameters*

to get the *best* fit. It the number of fit parameters, $N_{\rm fit}$ is equal to the number of data points, then we can clearly get the fit to go *exactly* through the data, so χ^2 would be zero. Hence the relevant quantity is not N but rather, it turns out, *the difference* $N_{\rm DOF} = N - N_{\rm fit}$, which is called the **number of degrees of freedom** (DOF). (When fitting to an m-th order polynomial, $N_{\rm fit} = m + 1$ so $N_{\rm DOF} = N - m - 1$.)

For a good fit, we expect that χ^2 should be about $N_{\rm DOF}$. To be precise, the mean value of χ^2 (averaged over many sets of data from the same distribution as the one set of data we actually have) is equal to $N_{\rm DOF}$. This is shown in Eq. (B7b) for the (important) case when the y_i have a Gaussian distribution, but is true in general. In many cases the y_i *are* distributed with a Gaussian distribution, see Appendix A. In this case, χ^2 is the sum of $N_{\rm DOF}$ terms each of which is a Gaussian with zero mean and standard deviation unity, see Appendix B. This resulting distribution is called **the χ^2 distribution**, and the expression for it is given in Eq. (B6). Its standard deviation is $\sqrt{2N_{\rm DOF}}$, see Eq. (B7d). Plots of the χ^2 distribution for different values of $N_{\rm DOF}$ are shown in Fig. 1.

There is a detailed theory, discussed in Appendix B, which converts a value of χ^2 for a given number of degrees of freedom to a **"goodness of fit" parameter, Q**, which is the probability that this value of χ^2, or greater, could occur by chance, assuming that the data points are distributed with a Gaussian distribution. The expression for Q is given by Eq. (B8). A very small value of Q indicates that the fit is very unlikely, and one should then look for another model to fit the data. (Another possibility is that the error bars have been underestimated.) The interested student who would like additional information is referred to the books, such as Numerical Recipes, Secs. 15.1 and 15.2. Even if the original data points are not distributed with a Gaussian distribution, the central limit theorem indicates that the expressions for the variance of χ^2 in Eq. (B7d) and the goodness of fit parameter Q in Eq. (B8) are still valid if the number of degrees of freedom is large enough. (The expression for the mean of χ^2, given in Eq. (B7b) is *always* valid, even if the data doesn't have a Gaussian distribution and $N_{\rm DOF}$ is not large.)

To summarize, a fitting program should provide the following information (assuming that error bars are given on the points):

1. The values of the fitting parameters.

2. Error bars on those parameters.

3. A measure of the goodness of fit.

Appendix A: The Gaussian distribution

Why is the Gaussian distribution special? There is a theorem of statistics, called **the central limit theorem,** which states that, for large N and under rather general conditions, the distribution of the sum of N random variables is Gaussian, even if the distribution of the individual variables is not Gaussian. For a proof see advanced books on probability and statistics, and my handout http://young.physics.ucsc.edu/116C/clt.pdf. A related handout may also be useful: http://young.physics.ucsc.edu/116C/dist_of_sum.pdf.

Note, though, that if N is not large enough for the central limit theorem to apply, the probability of getting a large deviation from the mean is invariably larger than would be expected from a Gaussian distribution. The reason is that the Gaussian falls off very fast at large deviations, and distributions which occur in nature, generically seem to fall off less fast.

Appendix B: The chi-squared distribution and the goodness of fit parameter

The χ^2 distribution for M degrees of freedom is the distribution of the sum of the squares of M random variables with a Gaussian distribution with zero mean and standard deviation unity. To determine this we write the distribution of the M variables x_i as

$$P(x_1, x_2, \cdots, x_M)\, dx_1 dx_2 \cdots dx_M = \frac{1}{(2\pi)^{M/2}}\, e^{-x_1^2/2}\, e^{-x_2^2/2} \cdots e^{-x_M^2/2}\, dx_1 dx_2 \cdots dx_M\,. \tag{B1}$$

Converting to polar coordinates, and integrating over directions, we find the distribution of the radial variable to be

$$\widetilde{P}(r)\, dr = \frac{S_M}{(2\pi)^{M/2}}\, r^{M-1}\, e^{-r^2/2}\, dr\,, \tag{B2}$$

where S_M is the surface area of a unit M-dimensional sphere. To determine S_M we integrate Eq. (B2) over r, noting that $\widetilde{P}(r)$ is normalized to unity, which gives

$$S_M = \frac{2\pi^{M/2}}{\Gamma(M/2)}\,, \tag{B3}$$

where $\Gamma(x)$ is the Euler gamma function defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t}\, dt\,. \tag{B4}$$

From Eqs. (B2) and (B3) we have

$$\widetilde{P}(r) = \frac{1}{2^{M/2-1}\Gamma(M/2)}\, r^{M-1} e^{-r^2/2}\,. \tag{B5}$$

This is the distribution of r but we want the distribution of $\chi^2 \equiv \sum_i x_i^2 = r^2$. To avoid confusion of notation we write X for χ^2, and define the χ^2 distribution for M variables as $P^{(M)}(X)$. We have $P^{(M)}(X)\,dX = \widetilde{P}(r)\,dr$ so the χ^2 distribution for M degrees of freedom is

$$P^{(M)}(X) = \frac{\widetilde{P}(r)}{dX/dr}$$
$$\boxed{= \frac{1}{2^{M/2}\Gamma(M/2)}\, X^{(M/2)-1}\, e^{-X/2} \qquad (X > 0)\,,} \tag{B6}$$

and is zero for $X < 0$. The χ^2 distribution for several value of $M \equiv N_{\rm DOF}$ is plotted in Fig. (1). The mean and variance are given by Eqs. (B7b) and (B7d) below. For large M, according to the central limit theorem, the χ^2 distribution becomes a Gaussian.

Using Eq. (B4) and the property of the gamma function that $\Gamma(n+1) = n\Gamma(n)$ one can show that

$$\int_0^\infty P^{(M)}(X)\,dX = 1\,, \tag{B7a}$$
$$\langle X\rangle \equiv \int_0^\infty X\, P^{(M)}(X)\,dX = \boxed{M\,,} \tag{B7b}$$
$$\langle X^2\rangle \equiv \int_0^\infty X^2\, P^{(M)}(X)\,dX = M^2 + 2M\,, \tag{B7c}$$
$$\langle X^2\rangle - \langle X\rangle^2 = \boxed{2M\,.} \tag{B7d}$$

The goodness of fit parameter is the probability that the specified value of χ^2, or greater, could occur by random chance. From Eq. (B6) it is given by

$$Q = \frac{1}{2^{M/2}\Gamma(M/2)} \int_{\chi^2}^\infty X^{(M/2)-1}\, e^{-X/2}\, dX\,,$$
$$\boxed{= \frac{1}{\Gamma(M/2)} \int_{\chi^2/2}^\infty y^{(M/2)-1}\, e^{-y}\, dy\,,} \tag{B8}$$

which is known as an incomplete gamma function. The area of the shaded area in Fig. (1) is the value of Q for $M \equiv N_{\rm DOF} = 10, \chi^2 = 15$. Note that $Q = 1$ for $\chi^2 = 0$ and $Q \to 0$ for $\chi^2 \to \infty$. If χ^2 per degree of freedom is one, the value of Q is around $1/2$.

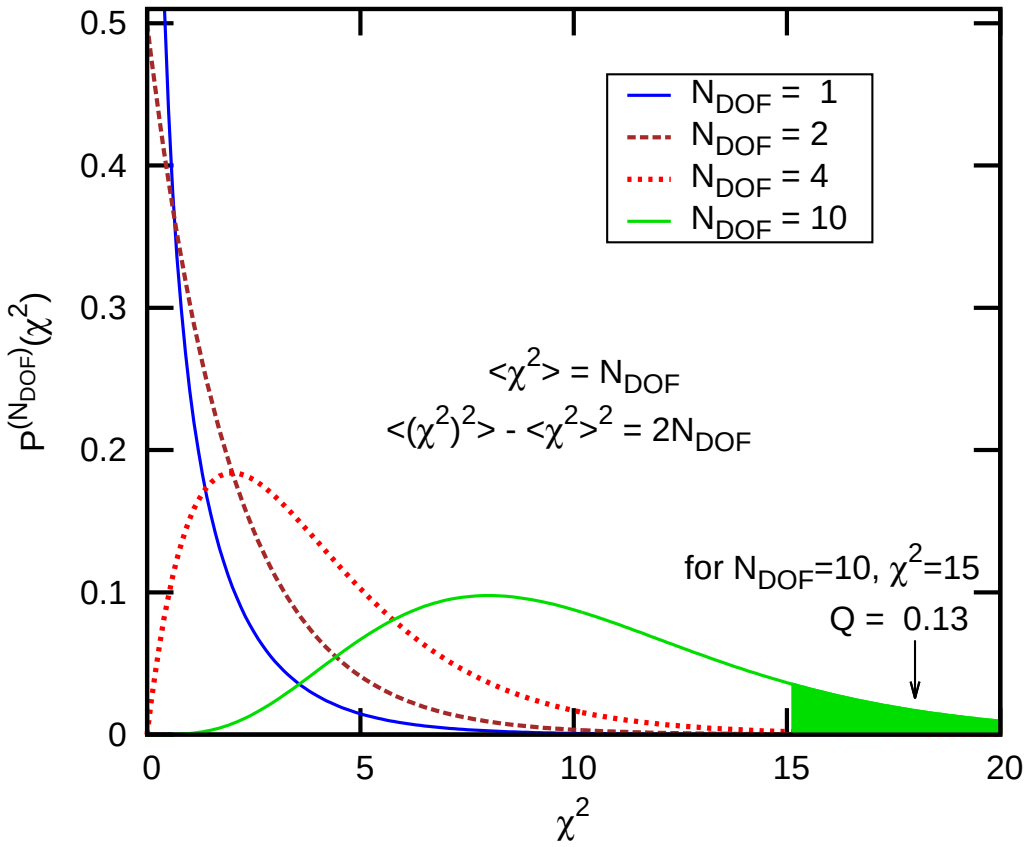


FIG. 1: The χ^2 distribution for several values of $N_{\rm DOF}$ the number of degrees of freedom. The mean and standard deviation depend on $N_{\rm DOF}$ in the way specified. The goodness of fit parameter Q, defined in Eq. (B8), depends on the values of $N_{\rm DOF}$ and χ^2, and is the probability that χ^2 could have the specified value or larger by random chance. The area of the shaded region in the figure is the value of Q for $N_{\rm DOF} = 10, \chi^2 = 15$. Note that the *total* area under each of the curves is unity because they represent probability distributions.