CHI-SQUARE: TESTING FOR GOODNESS OF FIT

In the previous chapter we discussed procedures for fitting a hypothesized function to a set of experimental data points. Such procedures involve minimizing a quantity we called Φ in order to determine best estimates for certain function parameters, such as (for a straight line) a slope and an intercept. Φ is proportional to (or in some cases equal to) a statistical measure called χ^2, or *chi-square*, a quantity commonly used to test whether any given data are well described by some hypothesized function. Such a determination is called a *chi-square test for goodness of fit*.

In the following, we discuss χ^2 and its statistical distribution, and show how it can be used as a test for goodness of fit.[1]

The definition of χ^2

If ν independent variables x_i are each normally distributed with mean μ_i and variance σ_i^2, then the quantity known as *chi-square*[2] is defined by

$$\chi^2 \equiv \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_\nu-\mu_\nu)^2}{\sigma_\nu^2} = \sum_{i=1}^{\nu} \frac{(x_i-\mu_i)^2}{\sigma_i^2} \tag{1}$$

Note that ideally, given the random fluctuations of the values of x_i about their mean values μ_i, each term in the sum will be of order unity. Hence, if we have chosen the μ_i and the σ_i correctly, we may expect that a calculated value of χ^2 will be approximately equal to ν. If it is, then we may conclude that the data are well described by the values we have chosen for the μ_i, that is, by the hypothesized function.

If a calculated value of χ^2 turns out to be much larger than ν, and we have correctly estimated the values for the σ_i, we may possibly conclude that our data are not well-described by our hypothesized set of the μ_i.

This is the general idea of the χ^2 test. In what follows we spell out the details of the procedure.

[1] The chi-square distribution and some examples of its use as a statistical test are also described in the references listed at the end of this chapter.

[2] The notation of χ^2 is traditional and possibly misleading. It is a single statistical variable, and not the square of some quantity χ. It is therefore not *chi squared*, but *chi-square*. The notation is merely suggestive of its construction as the sum of squares of terms. Perhaps it would have been better, historically, to have called it ξ or ζ.

The χ^2 distribution

The quantity χ^2 defined in Eq. 1 has the probability distribution given by

$$f(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} e^{-\chi^2/2} (\chi^2)^{(\nu/2)-1} \tag{2}$$

This is known as the χ^2-*distribution with* ν *degrees of freedom.* ν is a positive integer.[3] Sometimes we write it as $f(\chi^2_\nu)$ when we wish to specify the value of ν. $f(\chi^2)\,d(\chi^2)$ is the probability that a particular value of χ^2 falls between χ^2 and $\chi^2 + d(\chi^2)$.

Here are graphs of $f(\chi^2)$ versus χ^2 for three values of ν:

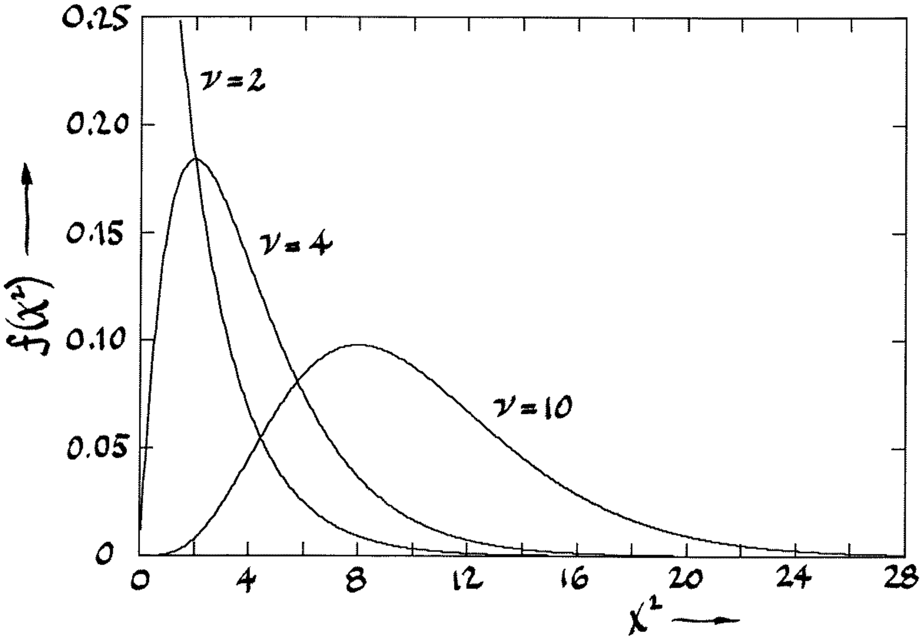


Figure 1 — The chi-square distribution for ν = 2, 4, and 10.

Note that χ^2 ranges only over positive values: $0 < \chi^2 < \infty$.

The mean value of χ^2_ν is equal to ν, and the variance of χ^2_ν is equal to 2ν. The distribution is highly skewed for small values of ν, and becomes more symmetric as ν increases, approaching a Gaussian distribution for large ν, just as predicted by the Central Limit Theorem.

[3] $\Gamma(p)$ is the "Gamma function", defined by $\Gamma(p+1) \equiv \int_0^\infty x^p e^{-x} dx$. It is a generalization of the factorial function to non-integer values of p. If p is an integer, $\Gamma(p+1) = p!$. In general, $\Gamma(p+1) = p\Gamma(p)$, and $\Gamma(1/2) = \sqrt{\pi}$.

How to use χ^2 to test for goodness of fit

Suppose we have a set of N experimentally measured quantities x_i. We want to test whether they are well-described by some set of hypothesized values μ_i. We form a sum like that shown in Eq. 1. It will contain N terms, constituting a sample value for χ^2. In forming the sum, we must use estimates for the σ_i that are *independently* obtained for each x_i.[4]

Now imagine, for a moment, that we could repeat our experiment many times. Each time, we would obtain a data sample, and each time, a sample value for χ^2. If our data were well-described by our hypothesis, we would expect our sample values of χ^2 to be distributed according to Eq. 2, and illustrated by example in Fig. 1. However, we must be a little careful. The expected distribution of our samples of χ^2 will *not* be one of N degrees of freedom, even though there are N terms in the sum, because our sample variables x_i will invariably *not* constitute a set of N independent variables. There will, typically, be at least one, and often as many as three or four, relations connecting the x_i. Such relations are needed in order to make estimates of hypothesized parameters such as the μ_i, and their presence will *reduce* the number of degrees of freedom. With r such relations, or *constraints*, the number of degrees of freedom becomes $\nu = N - r$, and the resulting χ^2 sample will be one having ν (rather than N) degrees of freedom.

As we repeat our experiment and collect values of χ^2, we expect, if our model is a valid one, that they will be clustered about the median value of χ^2_ν, with about half of these collected values being greater than the median value, and about half being less than the median value. This median value, which we denote by $\chi^2_{\nu,0.5}$, is determined by

$$\int_{\chi^2_{\nu,0.5}}^{\infty} f(\chi^2)\, d\chi^2 = 0.5$$

Note that because of the skewed nature of the distribution function, the *median* value of χ^2_ν will be somewhat less than the *mean* (or average) value of χ^2_ν, which as we have noted, is equal to ν. For example, for $\nu = 10$ degrees of freedom, $\chi^2_{10,0.5} \approx 9.34$, a number slightly less than 10.

Put another way, we expect that a single measured value of χ^2 will have a probability of 0.5 of being greater than $\chi^2_{\nu,0.5}$.

[4] In the previous chapter, we showed how a hypothesized function may be fit to a set of data points. There we noted that it may be either impossible or inconvenient to make independent estimates of the σ_i, in which case estimates of the σ_i can be made only by *assuming* an ideal fit of the function to the data. That is, we assumed χ^2 to be equal to its mean value, and from that, estimated uncertainties, or confidence intervals, for the values of the determined parameters. Such a procedure *precludes* the use of the χ^2 test.

We can generalize from the above discussion, to say that we expect a single measured value of χ^2 will have a probability α ("alpha") of being greater than $\chi^2_{\nu,\alpha}$, where $\chi^2_{\nu,\alpha}$ is defined by

$$\int_{\chi^2_{\nu,\alpha}}^{\infty} f(\chi^2)\, d\chi^2 = \alpha$$

This definition is illustrated by the inset in Fig. 2 on page 4–9.

Here is how the χ^2 test works:

(a) We hypothesize that our data are appropriately described by our chosen function, or set of μ_i. This is the hypothesis we are going to test.

(b) From our data sample we calculate a sample value of χ^2 (chi-square), along with ν (the number of degrees of freedom), and so determine χ^2/ν (the normalized chi-square, or the chi-square per degree of freedom) for our data sample.

(c) We choose a value of the significance level α (a common value is .05, or 5 per cent), and from an appropriate table or graph (*e.g.*, Fig. 2), determine the corresponding value of $\chi^2_{\nu,\alpha}/\nu$. We then compare this with our sample value of χ^2/ν.

(d) If we find that $\chi^2/\nu > \chi^2_{\nu,\alpha}/\nu$, we may conclude that either (i) the model represented by the μ_i is a valid one but that a statistically improbable excursion of χ^2 has occurred, or (ii) that our model is so poorly chosen that an unacceptably large value of χ^2 has resulted. (i) will happen with a probability α, so if we are satisfied that (i) and (ii) are the only possibilities, (ii) will happen with a probability $1 - \alpha$. Thus if we find that $\chi^2/\nu > \chi^2_{\nu,\alpha}/\nu$, we are $100 \cdot (1 - \alpha)$ per cent confident in *rejecting* our model. Note that this reasoning breaks down if there is a possibility (iii), for example if our data are *not* normally distributed. The theory of the chi-square test relies on the assumption that chi-square is the sum of the squares of *random normal deviates*, that is, that each x_i is normally distributed about its mean value μ_i. However for some experiments, there may be occasional non-normal data points that are too far from the mean to be real. A truck passing by, or a glitch in the electrical power could be the cause. Such points, sometimes called *outliers*, can unexpectedly increase the sample value of chi-square. It is appropriate to discard data points that are clearly outliers.

(e) If we find that χ^2 is too small, that is, if $\chi^2/\nu < \chi^2_{\nu,1-\alpha}/\nu$, we may conclude only that either (i) our model is valid but that a statistically improbable excursion of χ^2 has occurred, or (ii) we have, too conservatively, over-estimated the values of σ_i, or (iii) someone has given us fraudulent data, that is, data "too good to be true". A too-small value of χ^2 cannot be indicative of a poor model. A poor model can only increase χ^2.

Generally speaking, we should be pleased to find a sample value of χ^2/ν that is near 1, its mean value for a good fit.

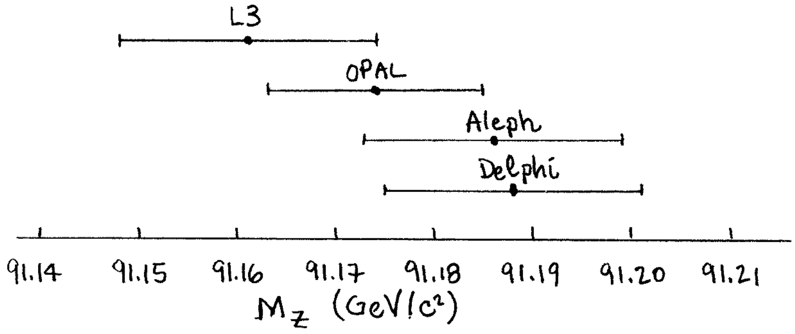
> In the final analysis, we must be guided by our own intuition and judgment. The chi-square test, being of a statistical nature, serves only as an indicator, and cannot be iron clad.

An example

The field of particle physics provides numerous situations where the χ^2 test can be applied. A particularly simple example[5] involves measurements of the mass M_Z of the Z^0 boson by experimental groups at CERN. The results of measurements of M_Z made by four different detectors (L3, OPAL, Aleph and Delphi) are as follows:

Detector	Mass in GeV/c^2
L3	91.161 ± 0.013
OPAL	91.174 ± 0.011
Aleph	91.186 ± 0.013
Delphi	91.188 ± 0.013

The listed uncertainties are estimates of the σ_i, the standard deviations for each of the measurements. The figure below shows these measurements plotted on a horizontal mass scale (vertically displaced for clarity).



Measurements of the Z^0 boson.

The question arises: Can these data be well described by a single number, namely an estimate of M_Z made by determining the weighted mean of the four measurements?

[5] This example is provided by Pat Burchat.

We find the weighted mean $\overline{M}_Z$, and its standard deviation $\sigma_{\overline{M}_Z}$ like this:[6]

$$\overline{M}_Z = \frac{\sum M_i/\sigma_i^2}{\sum 1/\sigma_i^2} \qquad \text{and} \qquad \sigma^2_{\overline{M}_Z} = \frac{1}{\sum 1/\sigma_i^2}$$

to find

$$\overline{M}_Z \pm \sigma_{\overline{M}_Z} = 91.177 \pm 0.006$$

Then we form χ^2:

$$\chi^2 = \sum_{i=1}^{4} \frac{(M_i - \overline{M}_Z)^2}{\sigma_i^2} \approx 2.78$$

We expect this value of χ^2 to be drawn from a chi-square distribution with 3 degrees of freedom. The number is 3 (not 4) because we have used the mean of the four measurements to estimate the value of μ, the true mass of the Z^0 boson, and this uses up one degree of freedom. Hence $\chi^2/\nu = 2.78/3 \approx 0.93$. Now from the graph of α versus χ^2/ν shown in Fig. 2, we find that for 3 degrees of freedom, α is about 0.42, meaning that if we were to repeat the experiments we would have about a 42 per cent chance of finding a χ^2 for the new measurement set larger than 2.78, assuming our hypothesis is correct. We have therefore no good reason to reject the hypothesis, and conclude that the four measurements of the Z^0 boson mass are consistent with each other. We would have had to have found χ^2 in the vicinity of 8.0 (leading to an α of about 0.05) to have been justified in suspecting the consistency of the measurements. The fact that our sample value of $\chi^2/3$ is close to 1 is reassuring.

Using χ^2 to test hypotheses regarding statistical distributions

The χ^2 test is used most commonly to test the nature of a statistical distribution from which some random sample is drawn. It is this kind of application that is described by Evans in his text, and is the kind of application for which the χ^2 test was first formulated.

Situations frequently arise where data can be classified into one of k classes, with probabilities $p_1, p_2, \ldots, p_k$ of falling into each class. If all the data are accounted for, $\sum p_i = 1$. Now suppose we take data by *classifying* it: We *count* the number of observations falling into each of the k classes. We'll have n_1 in the first class, n_2 in the second, and so on, up to n_k in the k^{th} class. We suppose there are a total of N observations, so $\sum n_i = N$.

[6] These expressions are derived on page 2–12.

It can be shown by non-trivial methods that the quantity

$$\frac{(n_1 - Np_1)^2}{Np_1} + \frac{(n_2 - Np_2)^2}{Np_2} + \cdots + \frac{(n_k - Np_k)^2}{Np_k} = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{3}$$

has approximately the χ^2 distribution with $k-r$ degrees of freedom, where r is the number of constraints, or relations used to estimate the p_i from the data. r will always be at least 1, since it must be that $\sum n_i = \sum Np_i = N \sum p_i = N$.

Since Np_i is the mean, or expected value of n_i, the form of χ^2 given by Eq. 3 corresponds to summing, over all classes, the squares of the deviations of the observed n_i from their mean values divided by their mean values.

At first glance, this special form looks different from that shown in Eq. 1, since the variance for each point is replaced by the mean value of n_i for each point. Such an estimate for the variance makes sense in situations involving counting, where the counted numbers are distributed according to the Poisson distribution, for which the mean is equal to the variance: $\mu = \sigma^2$.

Equation 3 forms the basis of what is sometimes called *Pearson's Chi-square Test.* Unfortunately some authors (incorrectly) use this equation to *define* χ^2. However, this form is not the most general form, in that it applies only to situations involving counting, where the data variables are dimensionless.

Another example

Here is an example in which the chi-square test is used to test whether a data sample consisting of the heights of 66 women can be assumed to be drawn from a Gaussian distribution.[7]

We first arrange the data in the form of a *frequency distribution*, listing for each height h, the value of $n(h)$, the number of women in the sample whose height is h (h is in inches):

h	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
$n(h)$	1	0	1	4	6	7	13	8	11	2	7	4	1	0	0	1

We make the hypothesis that the heights are distributed according to the Gaussian distribution (see page 2–6 of this manual), namely that the probability $p(h)\,dh$ that a height falls between h and $h + dh$ is given by

$$p(h)\,dh = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(h-\mu)^2/2\sigma^2}\, dh$$

[7] This sample was collected by Intermediate Laboratory students in 1991.

This expression, if multiplied by N, will give, for a sample of N women, the number of women $n_{th}(h)\,dh$ theoretically expected to have a height between h and $h + dh$:

$$n_{th}(h)\,dh = \frac{N}{\sigma\sqrt{2\pi}}\,e^{-(h-\mu)^2/2\sigma^2}\,dh \tag{4}$$

In our example, $N = 66$. Note that we have, in our table of data above, grouped the data into *bins* (we'll label them with the index j), each of size 1 inch. A useful approximation to Eq. 4, in which dh is taken to be 1 inch, gives the expected number of women $n_{th}(j)$ having a height h_j:

$$n_{th}(j) = \frac{N}{\sigma\sqrt{2\pi}}\,e^{-(h_j-\mu)^2/2\sigma^2} \tag{5}$$

Now the sample mean $\overline{h}$ and the sample standard deviation s are our best estimates of μ and σ. We find, calculating from the data:

$$\overline{h} = 64.9\text{ inches}, \qquad \text{and} \qquad s = 2.7\text{ inches}$$

Using these values we may calculate, from Eq. 5, the number expected in each bin, with the following results:

h	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
$n_{th}(j)$	0.3	0.9	1.8	3.4	5.5	7.6	9.3	9.9	9.1	7.3	5.0	3.1	1.6	0.7	0.3	0.1

In applying the chi-square test to a situation of this type, it is advisable to re-group the data into new bins (classes) such that the expected number occurring in each bin is greater than 4 or 5; otherwise the theoretical distributions within each bin become too highly skewed for meaningful results. Thus in this situation we shall put all the heights of 61 inches or less into a single bin, and all the heights of 69 inches or more into a single bin. This groups the data into a total of 9 bins (or classes), with actual numbers and expected numbers in each bin being given as follows (note the bin sizes need not be equal):

h	≤ 61	62	63	64	65	66	67	68	≥ 69
$n(h)$	6	6	7	13	8	11	2	7	6
$n_{th}(j)$	6.5	5.5	7.6	9.3	9.9	9.1	7.3	5.0	5.8

Now we calculate the value of χ^2 using these data, finding

$$\chi^2 = \frac{(6-6.5)^2}{6.5} + \frac{(6-5.5)^2}{5.5} + \cdots + \frac{(6-5.8)^2}{5.8} = 6.96$$

Since we have grouped our data into 9 classes, and since we have used up three degrees of freedom by demanding (a) that the sum of the n_j be equal to N, (b) that the mean of the hypothesized distribution be equal to the sample mean, and (c) that the variance of the hypothesized distribution be equal to the sample variance, there are 6 degrees of freedom left.[8] Hence $\chi^2/\nu = 6.96/6 \approx 1.16$, leading to an α of about 0.33. Therefore we have no good reason to reject our hypothesis that our data are drawn from a Gaussian distribution function.

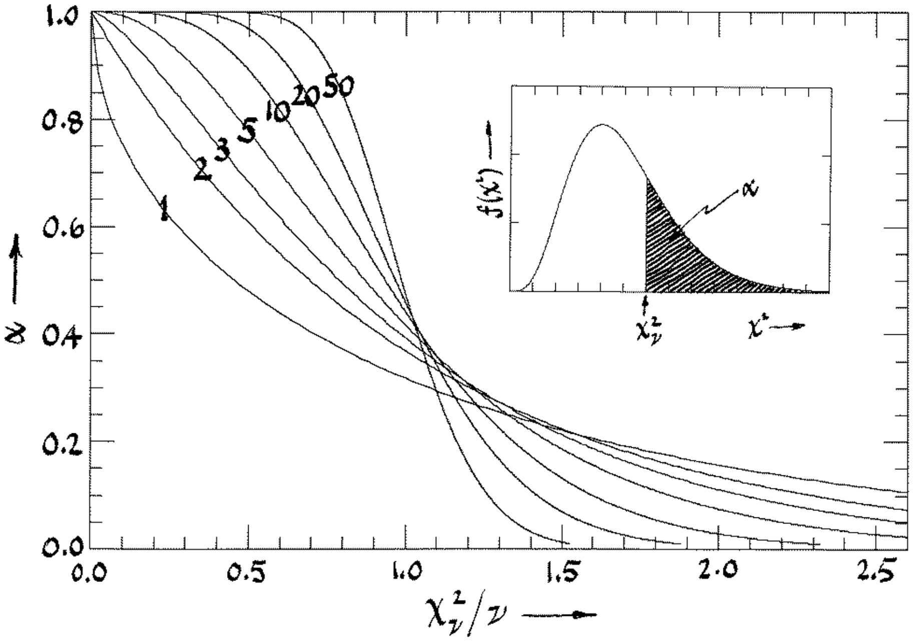


Figure 2 — α versus the normalized chi-square: χ^2_ν/ν. α is the probability that a sample chi-square will be larger than χ^2_ν, as shown in the inset. Each curve is labeled by ν, the number of degrees of freedom.

References

1. Press, William H. et. al., *Numerical Recipes in C—The Art of Scientific Computing*, 2nd Ed. (Cambridge University Press, New York, 1992). Press devotes considerable discussion to the subject of fitting parameters to data, including the use of the chi-square test. Noteworthy are his words of advice, appearing on pages 656–657:

> “To be genuinely useful, a fitting procedure should provide (i) parameters, (ii) error estimates on the parameters, and (iii) a statistical measure of goodness-of-fit. When the third item suggests that the model is an unlikely match to the data, then items (i) and (ii) are probably worthless. Unfortunately, many practitioners of parameter estimation never proceed beyond item (i). They deem a fit acceptable

[8] Note that if we were hypothesizing a Poisson distribution (as in a counting experiment), there would be 7 degrees of freedom (only 2 less than the number of classes). For a Poisson distribution the variance is *equal* to the mean, so there is only 1 parameter to be determined, not 2.

> if a graph of data and model ‘looks good’. This approach is known as *chi-by-eye*. Luckily, its practitioners get what they deserve.”

2. Bennett, Carl A., and Franklin, Norman L., *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, 1954). An excellent discussion of the chi-square distribution function, with good examples illustrating its use, may be found on pages 96 and 620.

3. Evans, Robley D., *The Atomic Nucleus* (McGraw-Hill, 1969). Chapter 27 of this advanced text contains a description of *Pearson’s Chi-square Test*.