Chapter 12

The Chi-Squared Test for a Distribution

By now, you should be reasonably familiar with the notion of limiting distributions. These are the functions that describe the expected distribution of results if an experiment is repeated many times. There are many different limiting distributions, corresponding to the many different kinds of experiments possible. Perhaps the three most important limiting distributions in physical science are the three we have already discussed: the Gauss (or normal) function, the binomial distribution, and the Poisson distribution.

This final chapter focuses on how to decide whether the results of an actual experiment are governed by the expected limiting distribution. Specifically, let us suppose that we perform some experiment for which we believe we know the expected distribution of results. Suppose further that we repeat the experiment several times and record our observations. The question we now address is this: How can we decide whether our observed distribution is consistent with the expected theoretical distribution? We will see that this question can be answered using a simple procedure called the *chi-squared,* or χ^2, *test.* (The Greek letter χ is spelled "chi" and pronounced "kie.")

12.1 Introduction to Chi Squared

Let us begin with a concrete example. Suppose we make 40 measurements $x_1, \ldots, x_{40}$ of the range x of a projectile fired from a certain gun and get the results shown in Table 12.1. Suppose also we have reason to believe these measurements are governed by a Gauss distribution $G_{X,\sigma}(x)$, as is certainly very natural. In this type

Table 12.1. Measured values of x (in cm).

731	772	771	681	722	688	653	757	733	742
739	780	709	676	760	748	672	687	766	645
678	748	689	810	805	778	764	753	709	675
698	770	754	830	725	710	738	638	787	712

of experiment, we usually do not know in advance either the center X or the width σ of the expected distribution. Our first step, therefore, is to use our 40 measurements to compute best estimates for these quantities:

$$\text{(best estimate for } X) = \bar{x} = \frac{\sum x_i}{40} = 730.1 \text{ cm} \tag{12.1}$$

and

$$\text{(best estimate for } \sigma) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{39}} = 46.8 \text{ cm.} \tag{12.2}$$

Now we can ask whether the actual distribution of our results $x_1, \ldots, x_{40}$ is consistent with our hypothesis that our measurements were governed by the Gauss distribution $G_{X,\sigma}(x)$ with X and σ as estimated. To answer this question, we must compute how we would expect our 40 results to be distributed if the hypothesis is true and compare this expected distribution with our actual observed distribution. The first difficulty is that x is a continuous variable, so we cannot speak of the expected number of measurements equal to any one value of x. Rather, we must discuss the expected number in some interval $a < x < b$. That is, we must divide the range of possible values into *bins*. With 40 measurements, we might choose bin boundaries at $X - \sigma$, X, and $X + \sigma$, giving four bins as in Table 12.2.

Table 12.2. A possible choice of bins for the data of Table 12.1. The final column shows the number of observations that fell into each bin.

Bin number k	Values of x in bin		Observations O_k
1	$x < X - \sigma$	(or $x < 683.3$)	8
2	$X - \sigma < x < X$	(or $683.3 < x < 730.1$)	10
3	$X < x < X + \sigma$	(or $730.1 < x < 776.9$)	16
4	$X + \sigma < x$	(or $776.9 < x$)	6

We will discuss later the criteria for choosing bins. In particular, they must be chosen so that all bins contain several measured values x_i. In general, I will denote the number of bins by n; for this example with four bins, $n = 4$.

Having divided the range of possible measured values into bins, we can now formulate our question more precisely. First, we can count the number of measurements that fall into each bin k.[1] We denote this number by O_k (where O stands for "observed number"). For the data of our example, the observed numbers O_1, O_2, O_3, O_4 are shown in the last column of Table 12.2. Next, assuming our measurements are distributed normally (with X and σ as estimated), we can calculate the *expected* number E_k of measurements in each bin k. We must then decide how well the observed numbers O_k compare with the expected numbers E_k.

[1] If a measurement falls exactly on the boundary between two bins, we can assign half a measurement to each bin.

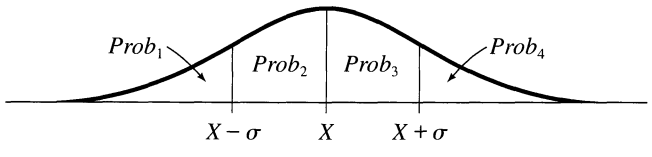

Figure 12.1. The probabilities $Prob_k$ that a measurement falls into each of the bins, $k = 1, 2, 3, 4$, of Table 12.2 are the four areas shown under the Gauss function.

The calculation of the expected numbers E_k is quite straightforward. The *probability* that any one measurement falls in an interval $a < x < b$ is just the area under the Gauss function between $x = a$ and $x = b$. In this example, the probabilities $Prob_1$, $Prob_2$, $Prob_3$, $Prob_4$ for a measurement to fall into each of our four bins are the four areas indicated in Figure 12.1. The two equal areas $Prob_2$ and $Prob_3$ together represent the well-known 68%, so the probability for falling into one of the two central bins is 34%; that is, $Prob_2 = Prob_3 = 0.34$. The outside two areas comprise the remaining 32%; thus $Prob_1 = Prob_4 = 0.16$. To find the expected numbers E_k, we simply multiply these probabilities by the total number of measurements, $N = 40$. Therefore, our expected numbers are as shown in the third column of Table 12.3. That the numbers E_k are not integers serves to remind us that the "expected number" is not what we actually expect in any one experiment; it is rather the expected average number after we repeat our whole series of measurements many times.

Our problem now is to decide how well the expected numbers E_k do represent the corresponding observed numbers O_k (in the last column of Table 12.3). We

Table 12.3. The expected numbers E_k and the observed numbers O_k for the 40 measurements of Table 12.1, with bins chosen as in Table 12.2.

Bin number k	Probability $Prob_k$	Expected number $E_k = NProb_k$	Observed number O_k
1	16%	6.4	8
2	34%	13.6	10
3	34%	13.6	16
4	16%	6.4	6

would obviously not expect *perfect* agreement between E_k and O_k after any finite number of measurements. On the other hand, if our hypothesis that our measurements are normally distributed is correct, we would expect that, in some sense, the deviations

$$O_k - E_k \tag{12.3}$$

would be *small.* Conversely, if the deviations $O_k - E_k$ prove to be *large,* we would suspect our hypothesis is incorrect.

To make precise the statements that the deviation $O_k - E_k$ is "small" or "large," we must decide how large we would *expect* $O_k - E_k$ to be if the measurements really are normally distributed. Fortunately, this decision is easily made. If we imagine repeating our whole series of 40 measurements many times, then the number O_k of measurements in any one bin k can be regarded as the result of a counting experiment of the type described in Chapter 11. Our many different answers for O_k should have an average value of E_k and would be expected to fluctuate around E_k with a standard deviation of order $\sqrt{E_k}$. Thus, the two numbers to be compared are the deviation $O_k - E_k$ and the expected size of its fluctuations $\sqrt{E_k}$.

These considerations lead us to consider the ratio

$$\frac{O_k - E_k}{\sqrt{E_k}}. \tag{12.4}$$

For some bins k, this ratio will be positive, and for some negative; for a few k, it may be appreciably larger than one, but for most it should be of order one, or smaller. To test our hypothesis (that the measurements are normally distributed), it is natural to square the number (12.4) for each k and then sum over all bins $k = 1, \ldots, n$ (here $n = 4$). This procedure defines a number called *chi squared,*

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{12.5}$$

This number χ^2 is clearly a reasonable indicator of the agreement between the observed and expected distributions. If $\chi^2 = 0$, the agreement is perfect; that is, $O_k = E_k$ for all bins k, a situation most unlikely to occur. In general, the individual terms in the sum (12.5) are expected to be of order one, and there are n terms in the sum. Thus, if

$$\chi^2 \lesssim n$$

(χ^2 of order n or less), the observed and expected distributions agree about as well as could be expected. In other words, if $\chi^2 \lesssim n$, we have no reason to doubt that our measurements were distributed as expected. On the other hand, if

$$\chi^2 \gg n$$

(χ^2 significantly greater than the number of bins), the observed and expected numbers differ significantly, and we have good reason to suspect that our measurements were not governed by the expected distribution.

In our example, the numbers observed and expected in the four bins and their differences are shown in Table 12.4, and a simple calculation using them gives

$$\begin{aligned} \chi^2 &= \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k} \\ &= \frac{(1.6)^2}{6.4} + \frac{(-3.6)^2}{13.6} + \frac{(2.4)^2}{13.6} + \frac{(-0.4)^2}{6.4} \\ &= 1.80. \end{aligned} \tag{12.6}$$

Table 12.4. The data of Table 12.1, shown here with the differences $O_k - E_k$.

Bin number k	Observed number O_k	Expected number $E_k = NProb_k$	Difference $O_k - E_k$
1	8	6.4	1.6
2	10	13.6	−3.6
3	16	13.6	2.4
4	6	6.4	−0.4

Because the value of 1.80 for χ^2 is less than the number of terms in the sum (namely, 4), we have no reason to doubt our hypothesis that our measurements were distributed normally.

Quick Check 12.1. Each of the 100 students in a class measures the time for a ball to fall from a third-story window. They calculate their mean $\bar{t}$ and standard deviation σ_t and then group their measurements into four bins, chosen as in the example just discussed. Their results are as follows:

less than $(\bar{t} - \sigma_t)$: 19
between $(\bar{t} - \sigma_t)$ and $\bar{t}$: 30
between $\bar{t}$ and $(\bar{t} + \sigma_t)$: 37
more than $(\bar{t} + \sigma_t)$: 14.

Assuming their measurements are normally distributed, what are the expected numbers of measurements in each of the four bins? What is χ^2, and is there reason to doubt that the measurements *are* distributed normally?

12.2 General Definition of Chi Squared

The discussion so far has focused on one particular example, 40 measurements of a continuous variable x, which denoted the range of a projectile fired from a certain gun. We defined the number χ^2 and saw that it is at least a rough measure of the agreement between our observed distribution of measurements and the Gauss distribution we expected our measurements to follow. We can now define and use χ^2 in the same way for many different experiments.

Let us consider any experiment in which we measure a number x and for which we have reason to expect a certain distribution of results. We imagine repeating the measurement many times (N) and, having divided the range of possible results x into n bins, $k = 1, \ldots, n$, we count the number O_k of observations that actually fall into each bin k. Assuming the measurements really are governed by the expected

distribution, we next calculate the expected number E_k of measurements in the kth bin. Finally, we calculate χ^2 exactly as in (12.5),

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{12.7}$$

The approximate significance of χ^2 is always the same as in our previous example. That is, if $\chi^2 \leqslant n$, the agreement between our observed and expected distributions is acceptable; if $\chi^2 \gg n$, there is significant disagreement.

The procedure for choosing the bins in terms of which χ^2 is computed depends somewhat on the nature of the particular experiment. Specifically, it depends on whether the measured quantity x is continuous or discrete. I will discuss these two situations in turn.

MEASUREMENTS OF A CONTINUOUS VARIABLE

The example discussed in Section 12.1 involved a continuous variable x, and little more needs to be said. The only limiting distribution we have discussed for a continuous variable is the Gauss distribution, but there are, of course, many different distributions that can occur. For example, in many atomic and nuclear experiments, the expected distribution of the measured variable x (actually an energy) is the Lorentzian distribution

$$f(x) \propto \frac{1}{(x - X)^2 + \gamma^2},$$

where X and γ are certain constants. Another example of a continuous distribution, mentioned in Problem 5.6, is the exponential distribution $\frac{1}{\tau} e^{-t/\tau}$, which gives the probability that a radioactive atom (whose expected mean life is τ) will live for a time t.

Whatever the expected distribution $f(x)$, the total area under the graph of $f(x)$ against x is one, and the probability of a measurement between $x = a$ and $x = b$ is just the area between a and b,

$$Prob(a < x < b) = \int_a^b f(x)\, dx.$$

Thus, if the kth bin runs from $x = a_k$ to $x = a_{k+1}$, the expected number of measurements in the kth bin (after N measurements in all) is

$$\begin{aligned} E_k &= N \times Prob(a_k < x < a_{k+1}) \\ &= N \int_{a_k}^{a_{k+1}} f(x)\, dx. \end{aligned} \tag{12.8}$$

When we discuss the quantitative use of the chi-squared test in Section 12.4, we will see that the expected numbers E_k should not be too small. Although there is no definite lower limit, E_k should probably be approximately five or more,

$$E_k \geqslant 5. \tag{12.9}$$

We must therefore choose bins in such a way that E_k as given by (12.8) satisfies this condition. We will also see that the number of bins must not be too small. For instance, in the example of Section 12.1, where the expected distribution was a Gauss distribution whose center X and width σ were not known in advance, the chi-squared test cannot work (as we will see) with less than four bins; that is, in this example we needed to have

$$n \geqslant 4. \tag{12.10}$$

Combining (12.9) and (12.10), we see that we cannot usefully apply the chi-squared test to this kind of experiment if our total number of observations is less than about 20.

MEASUREMENT OF A DISCRETE VARIABLE

Suppose we measure a discrete variable, such as the now-familiar number of aces when we throw several dice. In practice, the most common discrete variable is an integer (such as the number of aces), and we will denote the discrete variable by v instead of x (which we use for a continuous variable). If we throw five dice, the possible values of v are $v = 0, 1, \ldots, 5$, and we do not actually need to group the possible results into bins. We can simply count how many times we got each of the six possible results. In other words, we can choose six bins, each of which contains just one result.

Nonetheless, it is often desirable to group several different results into one bin. For instance, if we threw our five dice 200 times, then (according to the probabilities found in Problem 10.11) the expected distribution of results is as shown in the first two columns of Table 12.5. We see that here the expected numbers of throws giving four and five aces are 0.6 and 0.03, respectively, both much less than the five or so occurrences required in each bin if we want to use the chi-squared test. This difficulty is easily remedied by grouping the results $v = 3$, 4, and 5 into a single bin. This grouping leaves us with four bins, $k = 1, 2, 3, 4$, which are shown with their corresponding expected numbers E_k, in the last two columns of Table 12.5.

Table 12.5. Expected occurrence of v aces ($v = 0, 1, \ldots, 5$) after throwing five dice 200 times.

Result	Expected occurrences	Bin number k	Expected number E_k
No aces	80.4	1	80.4
One	80.4	2	80.4
Two	32.2	3	32.2
Three	6.4	4	7.0
Four	0.6		
Five	0.03		

Having chosen bins as just described, we could count the observed occurrences O_k in each bin. We could then compute χ^2 and see whether the observed and expected distributions seem to agree. In this experiment, we know that the expected distribution is certainly the binomial distribution $B_{5,1/6}(v)$ *provided* the dice are true

(so that p really is $\frac{1}{6}$). Thus, our test of the distribution is, in this case, a test of whether the dice are true or loaded.

In any experiment involving a discrete variable, the bins can be chosen to contain just one result each, provided the expected number of occurrences for each bin is at least the needed five or so. Otherwise, several different results should be grouped together into a single larger bin that does include enough expected occurrences.

OTHER FORMS OF CHI SQUARED

The notation χ^2 has been used earlier in the book, in Equations (7.6) and (8.5); it could also have been used for the sum of squares in (5.41). In all these cases, χ^2 is a sum of squares with the general form

$$\chi^2 = \sum_1^n \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}} \right)^2. \tag{12.11}$$

In all cases, χ^2 is an indicator of the agreement between the observed and expected values of some variable. If the agreement is good, χ^2 will be of order n; if it is poor, χ^2 will be much greater than n.

Unfortunately, we can use χ^2 to test this agreement only if we know the expected values and the standard deviation, and can therefore calculate (12.11). Perhaps the most common situation in which these values are known accurately enough is the kind of test discussed in this chapter, namely, a test of a distribution, in which E_k is given by the distribution, and the standard deviation is $\sqrt{E_k}$. Nevertheless, the chi-squared test is of very wide application. Consider, for example, the problem discussed in Chapter 8, the measurement of two variables x and y, where y is expected to be some definite function of x,

$$y = f(x)$$

(such as $y = A + Bx$). Suppose we have N measured pairs (x_i, y_i), where the x_i have negligible uncertainty and the y_i have known uncertainties σ_i. Here, the expected value of y_i is $f(x_i)$, and we could test how well y fits the function $f(x)$ by calculating

$$\chi^2 = \sum_1^N \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2.$$

All our previous remarks about the expected value of χ^2 would apply to this number, and the quantitative tests described in the following sections could be used. This important application will not be pursued here, because only rarely in the introductory physics laboratory would the uncertainties σ_i be known reliably enough (but see Problem 12.14).

12.3 Degrees of Freedom and Reduced Chi Squared

I have argued that we can test agreement between an observed and an expected distribution by computing χ^2 and comparing it with the number of bins used in

collecting the data. A slightly better procedure, however, is to compare χ^2, not with the number of bins n, but instead with the *number of degrees of freedom,* denoted d. The notion of degrees of freedom was mentioned briefly in Section 8.3, and we must now discuss it in more detail.

In general, the number of degrees of freedom d in a statistical calculation is defined as the number of observed data *minus* the number of parameters computed from the data and used in the calculation. For the problems considered in this chapter, the observed data are the numbers of observations O_k in the n bins, $k = 1, \ldots,$ n. Thus, the number of observed data is just n, the number of bins. Therefore, in the problems considered here,

$$d = n - c,$$

where n is the number of bins and c is the number of parameters that had to be calculated from the data to compute the expected numbers E_k. The number c is often called the number of *constraints,* as I will explain shortly.

The number of constraints c varies according to the problem under consideration. Consider first the dice-throwing experiment of Section 12.2. If we throw five dice and are testing the hypothesis that the dice are true, the expected distribution of numbers of aces is the binomial distribution $B_{5,1/6}(v)$, where $v = 0, \ldots, 5$ is the number of aces in any one throw. Both parameters in this function—the number of dice, five, and the probability of an ace, $\frac{1}{6}$—are known in advance and do not have to be calculated from the data. When we calculate the expected number of occurrences of any particular v, we must multiply the binomial probability by the total number of throws N (in our example, $N = 200$). This parameter *does* depend on the data. Specifically, N is just the sum of the numbers O_k,

$$N = \sum_{k=1}^{n} O_k. \tag{12.12}$$

Thus, in calculating the expected results of our dice experiment, we have to calculate just one parameter (N) from the data. The number of constraints is, therefore,

$$c = 1,$$

and the number of degrees of freedom is

$$d = n - 1.$$

In Table 12.5, the results of the dice experiment were grouped into four bins (that is, $n = 4$), so that experiment had 3 degrees of freedom.

The equation (12.12) illustrates well the curious terminology of constraints and degrees of freedom. Once the number N has been determined, we can regard (12.12) as an equation that "constrains" the values of $O_1, \ldots, O_n$. More specifically, we can say that, because of the constraint (12.12), only $n - 1$ of the numbers $O_1, \ldots, O_n$ are independent. For instance, the first $n - 1$ numbers $O_1, \ldots, O_{n-1}$ could take any value (within certain ranges), but the last number O_n would be completely determined by Equation (12.12). In this sense, only $n - 1$ of the data are *free* to take on independent values, so we say there are only $n - 1$ independent degrees of freedom.

In the first example in this chapter, the range x of a projectile was measured 40

times ($N = 40$). The results were collected into four bins ($n = 4$) and compared with what we would expect for a Gauss distribution $G_{X,\sigma}(x)$. Here, there were *three* constraints and hence only one degree of freedom,

$$d = n - c = 4 - 3 = 1.$$

The first constraint is the same as (12.12): The total number of observations N is the sum of the observations O_k in all the bins. But here there were two more constraints, because (as is usual in this kind of experiment) we did not know in advance the parameters X and σ of the expected Gauss distribution $G_{X,\sigma}(x)$. Thus, before we could calculate the expected numbers E_k, we had to estimate X and σ using the data. Therefore, there were three constraints in all, so in this example

$$d = n - 3. \tag{12.13}$$

Incidentally, this result explains why we had to use at least four bins in this experiment. We will see that the number of degrees of freedom must always be one or more, so, from (12.13), we clearly had to choose $n \geqslant 4$.

The examples considered here will always have at least one constraint (namely, the constraint $N = \sum O_k$, involving the total number of measurements), and there may be one or two more. Thus, the number of degrees of freedom, d, will range from $n - 1$ to $n - 3$ (in our examples). When n is large, the difference between n and d is fairly unimportant, but when n is small (as it often is, unfortunately), there is obviously a significant difference.

Armed with the notion of degrees of freedom, we can now begin to make our chi-squared test more precise. It can be shown (though I will not do so) that the *expected* value of χ^2 is precisely d, the number of degrees of freedom,

$$(\text{expected average value of } \chi^2) = d. \tag{12.14}$$

This important equation does not mean that we really expect to find $\chi^2 = d$ after any one series of measurements. It means instead that if we could repeat our whole series of measurements infinitely many times and compute χ^2 each time, the average of these values of χ^2 would be d. Nonetheless, even after just *one* set of measurements, a comparison of χ^2 with d is an indicator of the agreement. In particular, if our expected distribution was the *correct* distribution, χ^2 would be very unlikely to be a lot larger than d. Turning this statement around, if we find $\chi^2 \gg d$, we can assert that our expected distribution was most unlikely to be correct.

We have *not* proved the result (12.14), but we can see that some aspects of the result are reasonable. For example, because $d = n - c$, we can rewrite (12.14) as

$$(\text{expected average value of } \chi^2) = n - c. \tag{12.15}$$

That is, for any given n, the expected value of χ^2 will be smaller when c is larger (that is, if we calculate more parameters from the data). This result is just what we should expect. In the example of Section 12.1, we used the data to calculate the center X and width σ of the expected distribution $G_{X,\sigma}(x)$. Naturally, because X and σ were chosen to fit the data, we would expect to find a somewhat better agreement between the observed and expected distributions; that is, these two extra constraints would be expected to reduce the value of χ^2. This reduction is just what (12.15) implies.

The result (12.14) suggests a slightly more convenient way to think about our chi-squared test. We introduce a *reduced chi squared* (or *chi squared per degree of freedom*), which we denote by $\tilde{\chi}^2$ and define as

$$\tilde{\chi}^2 = \chi^2/d. \tag{12.16}$$

Because the expected value of χ^2 is d, we see that the

$$(\text{expected average value of } \tilde{\chi}^2) = 1. \tag{12.17}$$

Thus, whatever the number of degrees of freedom, our test can be stated as follows: If we obtain a value of $\tilde{\chi}^2$ of order one or less, then we have no reason to doubt our expected distribution; if we obtain a value of $\tilde{\chi}^2$ much larger than one, our expected distribution is unlikely to be correct.

Quick Check 12.2. For the experiment of Quick Check 12.1, what is the number of degrees of freedom, and what is the value of the reduced chi squared, $\tilde{\chi}^2$?

12.4 Probabilities for Chi Squared

Our test for agreement between observed data and their expected distribution is still fairly crude. We now need a *quantitative* measure of agreement. In particular, we need some guidance on where to draw the boundary between agreement and disagreement. For example, in the experiment of Section 12.1, we made 40 measurements of a certain range x whose distribution should, we believed, be Gaussian. We collected our data into four bins, and found that $\chi^2 = 1.80$. With three constraints, there was only one degree of freedom ($d = 1$), so the reduced chi squared, $\tilde{\chi}^2 = \chi^2/d$, is also 1.80,

$$\tilde{\chi}^2 = 1.80.$$

The question is now: Is a value of $\tilde{\chi}^2 = 1.80$ sufficiently larger than one to rule out our expected Gauss distribution or not?

To answer this question, we begin by supposing that our measurements *were* governed by the expected distribution (a Gaussian, in this example). With this assumption, we can calculate the *probability* of obtaining a value of $\tilde{\chi}^2$ as large as, or larger than, our value of 1.80. Here, this probability turns out to be

$$Prob(\tilde{\chi}^2 \geq 1.80) \approx 18\%,$$

as we will soon see. That is, if our results were governed by the expected distribution, there would be an 18% probability of obtaining a value of $\tilde{\chi}^2$ greater than or

equal to our actual value 1.80. In other words, in this experiment a value of $\tilde{\chi}^2$ as large as 1.80 is not at all unreasonable, so we would have no reason (based on this evidence) to reject our expected distribution.

Our general procedure should now be reasonably clear. After completing any series of measurements, we calculate the reduced chi squared, which we now denote by $\tilde{\chi}_o^2$ (where the subscript o stands for "observed," because $\tilde{\chi}_o^2$ is the value actually observed). Next, assuming our measurements do follow the expected distribution, we compute the probability

$$Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) \tag{12.18}$$

of finding a value of $\tilde{\chi}^2$ greater than or equal to the observed value $\tilde{\chi}_o^2$. If this probability is high, our value $\tilde{\chi}_o^2$ is perfectly acceptable, and we have no reason to reject our expected distribution. If this probability is unreasonably low, a value of $\tilde{\chi}^2$ as large as our observed $\tilde{\chi}_o^2$ is very unlikely (if our measurements were distributed as expected), and our expected distribution is correspondingly unlikely to be correct.

As always with statistical tests, we have to decide on the boundary between what is reasonably probable and what is not. Two common choices are those already mentioned in connection with correlations. With the boundary at 5%, we would say that our observed value $\tilde{\chi}_o^2$ indicates a significant disagreement if

$$Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) < 5\%,$$

and we would reject our expected distribution at the 5% significance level. If we set the boundary at 1%, then we could say that the disagreement is highly significant if $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) < 1\%$ and reject the expected distribution at the 1% significance level.

Whatever level you choose as your boundary for rejection, the level chosen should be stated. Perhaps even more important, you should state the probability $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$, so that your readers can judge its reasonableness for themselves.

The calculation of the probabilities $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ is too complicated to describe in this book. The results can be tabulated easily, however, as in Table 12.6 or in the more complete table in Appendix D. The probability of getting any particular values of $\tilde{\chi}^2$ depends on the number of degrees of freedom. Thus, we will write the probability of interest as $Prob_d(\tilde{\chi} \geq \tilde{\chi}_o^2)$ to emphasize its dependence on d.

The usual calculation of the probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ treats the observed numbers O_k as continuous variables distributed around their expected values E_k according to a Gauss distribution. In the problems considered here, O_k is a discrete variable distributed according to the Poisson distribution.[2] Provided all numbers involved are reasonably large, the discrete character of the O_k is unimportant, and the Poisson distribution is well approximated by the Gauss function. Under these conditions, the tabulated probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ can be used safely. For this reason, we have said the bins must be chosen so that the expected count E_k in each bin is reasonably large (at least five or so). For the same reason, the number of bins should not be too small.

---

[2] I have argued that finding the number O_k amounts to a counting experiment and hence that O_k should follow a Poisson distribution. If the bin k is too large, then this argument is not strictly correct, because the probability of a measurement in the bin is not much less than one (which is one of the conditions for the Poisson distribution, as mentioned in Section 11.1), so we must have a reasonable number of bins.

Table 12.6. The percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ of obtaining a value of $\tilde{\chi}^2$ greater than or equal to any particular value $\tilde{\chi}_o^2$, assuming the measurements concerned are governed by the expected distribution. Blanks indicate probabilities less than 0.05%. For a more complete table, see Appendix D.

	$\tilde{\chi}_o^2$												
d	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2	3	4	5	6
1	100	62	48	39	32	26	22	19	16	8	5	3	1
2	100	78	61	47	37	29	22	17	14	5	2	0.7	0.2
3	100	86	68	52	39	29	21	15	11	3	0.7	0.2	—
5	100	94	78	59	42	28	19	12	8	1	0.1	—	—
10	100	99	89	68	44	25	13	6	3	0.1	—	—	—
15	100	100	94	73	45	23	10	4	1	—	—	—	—

With these warnings, we now give the calculated probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ for a few representative values of d and $\tilde{\chi}_o^2$ in Table 12.6. The numbers in the left column give six choices of d, the number of degrees of freedom ($d = 1, 2, 3, 5, 10, 15$). Those in the other column heads give possible values of the observed $\tilde{\chi}_o^2$. Each cell in the table shows the percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ as a function of d and $\tilde{\chi}_o^2$. For example, with 10 degrees of freedom ($d = 10$), we see that the probability of obtaining $\tilde{\chi}^2 \geq 2$ is 3%,

$$Prob_{10}(\tilde{\chi}^2 \geq 2) \;=\; 3\%.$$

Thus, if we obtained a reduced chi squared of 2 in an experiment with 10 degrees of freedom, we could conclude that our observations differed significantly from the expected distribution and reject the expected distribution at the 5% significance level (though not at the 1% level).

The probabilities in the second column of Table 12.6 are all 100%, because $\tilde{\chi}^2$ is always certain to be greater than or equal to 0. As $\tilde{\chi}_o^2$ increases, the probability of getting $\tilde{\chi}^2 \geq \tilde{\chi}_o^2$ diminishes, but it does so at a rate that depends on d. Thus, for 2 degrees of freedom ($d = 2$), $Prob_d(\tilde{\chi}^2 \geq 1)$ is 37%, whereas for $d = 15$, $Prob_d(\tilde{\chi}^2 \geq 1)$ is 45%. Note that $Prob_d(\tilde{\chi}^2 \geq 1)$ is always appreciable (at least 32%, in fact), so a value for $\tilde{\chi}_o^2$ of 1 or less is perfectly reasonable and never requires rejection of the expected distribution.

The minimum value of $\tilde{\chi}_o^2$ that does require questioning the expected distribution depends on d. For 1 degree of freedom, we see that $\tilde{\chi}_o^2$ can be as large as 4 before the disagreement becomes significant (5% level). With 2 degrees of freedom, the corresponding boundary is $\tilde{\chi}_o^2 = 3$; for $d = 5$, it is closer to 2 ($\tilde{\chi}_o^2 = 2.2$, in fact), and so on.

Armed with the probabilities in Table 12.6 (and Appendix D), we can now assign a quantitative significance to the value of $\tilde{\chi}_o^2$ obtained in any particular experiment. Section 12.5 gives some examples.

Quick Check 12.3. Each student in a large class times a glider on an air track as it coasts the length of the track. They calculate their mean time and standard deviation and then divide their data into six bins. Assuming their measurements

ought to be normally distributed, they calculate the numbers of measurements expected in each bin and the reduced chi squared, for which they get 4.0. If their measurements really were normally distributed, what would have been the probability of getting a value of $\tilde{\chi}^2$ this large? Is there reason to think the measurements were *not* normally distributed?

12.5 Examples

We have already analyzed rather completely the example of Section 12.1. In this section, we consider three more examples to illustrate the application of the chi-squared test.

Example: Another Example of the Gauss Distribution

The example of Section 12.1 involved a measurement for which the results were expected to be distributed normally. The normal, or Gauss, distribution is so common that we consider briefly another example. Suppose an anthropologist is interested in the heights of the natives on a certain island. He suspects that the heights of the adult males should be normally distributed and measures the heights of a sample of 200 men. Using these measurements, he calculates the mean and standard deviation and uses these numbers as best estimates for the center X and width parameter σ of the expected normal distribution $G_{X,\sigma}(x)$. He now chooses eight bins, as shown in the first two columns of Table 12.7, and groups his observations, with the results shown in the third column.

Table 12.7. Measurements of the heights of 200 adult males.

Bin number k	Heights in bin	Observed number O_k	Expected number E_k
1	less than $X - 1.5\sigma$	14	13.4
2	between $X - 1.5\sigma$ and $X - \sigma$	29	18.3
3	between $X - \sigma$ and $X - 0.5\sigma$	30	30.0
4	between $X - 0.5\sigma$ and X	27	38.3
5	between X and $X + 0.5\sigma$	28	38.3
6	between $X + 0.5\sigma$ and $X + \sigma$	31	30.0
7	between $X + \sigma$ and $X + 1.5\sigma$	28	18.3
8	more than $X + 1.5\sigma$	13	13.4

Our anthropologist now wants to check whether these results are consistent with the expected normal distribution $G_{X,\sigma}(x)$. To this end, he first calculates the probability $Prob_k$ that any one man has height in any particular bin k (assuming a normal distribution). This probability is the integral of $G_{X,\sigma}(x)$ between the bin boundaries and is easily found from the table of integrals in Appendix B. The expected number E_k in each bin is then $Prob_k$ times the total number of men sampled (200). These numbers are shown in the final column of Table 12.7.

To calculate the expected numbers E_k, the anthropologist had to use three parameters calculated from his data (the total number in the sample and his estimates for X and σ). Thus, although there are eight bins, he had three constraints; so the number of degrees of freedom is $d = 8 - 3 = 5$. A simple calculation using the data of Table 12.7 gives for his reduced chi squared

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{i=1}^{8}\frac{(O_k - E_k)^2}{E_k} = 3.5.$$

Because this value is appreciably larger than one, we immediately suspect that the islanders' heights do not follow the normal distribution. More specifically, we see from Table 12.6 that, if the islanders' heights were distributed as expected, then the probability $Prob_5(\tilde{\chi}^2 \geq 3.5)$ of obtaining $\tilde{\chi}^2 \geq 3.5$ is approximately 0.5%. By any standards, this value is very improbable, and we conclude that the islanders' heights are very unlikely to be normally distributed. In particular, at the 1% (or highly significant) level, we can reject the hypothesis of a normal distribution of heights.

Example: More Dice

In Section 12.2, we discussed an experiment in which five dice were thrown many times and the number of aces in each throw recorded. Suppose we make 200 throws and divide the results into bins as discussed before. Assuming the dice are true, we can calculate the expected numbers E_k as before. These numbers are shown in the third column of Table 12.8.

Table 12.8. Distribution of numbers of aces in 200 throws of 5 dice.

Bin number k	Results in bin	Expected number E_k	Observed number O_k
1	no aces	80.4	60
2	one ace	80.4	88
3	two aces	32.2	39
4	3, 4, or 5 aces	7.0	13

In an actual test, five dice were thrown 200 times and the numbers in the last column of Table 12.8 were observed. To test the agreement between the observed and expected distributions, we simply note that there are three degrees of freedom (four bins minus one constraint) and calculate

$$\tilde{\chi}^2 = \frac{1}{3}\sum_{k=1}^{4}\frac{(O_k - E_k)^2}{E_k} = 4.16.$$

Referring back to Table 12.6, we see that with three degrees of freedom, the probability of obtaining $\tilde{\chi}^2 \geq 4.16$ is approximately 0.7%, *if* the dice are true. We conclude that the dice are almost certainly not true. Comparison of the numbers E_k and O_k in Table 12.8 suggests that at least one die is loaded in favor of the ace.

Example: An Example of the Poisson Distribution

As a final example of the use of the chi-squared test, let us consider an experiment in which the expected distribution is the Poisson distribution. Suppose we arrange a Geiger counter to count the arrival of cosmic-ray particles in a certain region. Suppose further that we count the number of particles arriving in 100 separate one-minute intervals, and our results are as shown in the first two columns of Table 12.9.

Table 12.9. Numbers of cosmic-ray particles observed in 100 separate one-minute intervals.

Counts ν in one minute	Occurrences	Bin number k	Observations O_k in bin k	Expected number E_k
None	7	1	7	7.5
One	17	2	17	19.4
Two	29	3	29	25.2
Three	20	4	20	21.7
Four	16	5	16	14.1
Five	8	6	11	12.1
Six	1			
Seven	2			
Eight or more	0			
Total	100			

Inspection of the numbers in column two immediately suggests that we group all counts $\nu \geq 5$ into a single bin. This choice of six bins ($k = 1, \ldots, 6$) is shown in the third column and the corresponding numbers O_k in column four.

The hypothesis we want to test is that the number ν is governed by a Poisson distribution $P_\mu(\nu)$. Because the expected mean count μ is unknown, we must first calculate the average of our 100 counts. This value is easily found to be $\bar{\nu} = 2.59$, which gives us our best estimate for μ. Using this value $\mu = 2.59$, we can calculate the probability $P_\mu(\nu)$ of any particular count ν and hence the expected numbers E_k as shown in the final column.

In calculating the numbers E_k, we used two parameters based on the data, the total number of observations (100), and our estimate of μ ($\mu = 2.59$). (Note that because the Poisson distribution is completely determined by μ, we did not have to estimate the standard deviation σ. Indeed, because $\sigma = \sqrt{\mu}$, our estimate for μ automatically gives us an estimate for σ.) There are, therefore, two constraints, which reduces our six bins to four degrees of freedom, $d = 4$.

A simple calculation using the numbers in the last two columns of Table 12.9 now gives for the reduced chi squared

$$\tilde{\chi}^2 = \frac{1}{d} \sum_{k=1}^{6} \frac{(O_k - E_k)^2}{E_k} = 0.35.$$

Because this value is less than one, we can conclude immediately that the agreement between our observations and the expected Poisson distribution is satisfactory. More

specifically, we see from the table in Appendix D that a value of $\tilde{\chi}^2$ as large as 0.35 is very probable; in fact

$$Prob_4(\tilde{\chi}^2 \geq 0.35) \approx 85\%.$$

Thus, our experiment gives us absolutely no reason to doubt the expected Poisson distribution.

The value of $\tilde{\chi}^2 = 0.35$ found in this experiment is actually appreciably less than one, indicating that our observations fit the Poisson distribution very well. This small value does *not,* however, give stronger evidence that our measurements are governed by the expected distribution than would a value $\tilde{\chi}^2 \approx 1$. If the results really are governed by the expected distribution, and if we were to repeat our series of measurements many times, we would expect many different values of $\tilde{\chi}^2$, fluctuating about the average value one. Thus, if the measurements are governed by the expected distribution, a value of $\tilde{\chi}^2 = 0.35$ is just the result of a large chance fluctuation away from the expected mean value. In no way does it give extra weight to our conclusion that our measurements do seem to follow the expected distribution.

If you have followed these three examples, you should have no difficulty applying the chi-squared test to any problems likely to be found in an elementary physics laboratory. Several further examples are included in the problems below. You should certainly test your understanding by trying some of them.

Principal Definitions and Equations of Chapter 12

DEFINITION OF CHI SQUARED

If we make n measurements for which we know, or can calculate, the expected values and the standard deviations, then we define χ^2 as

$$\chi^2 = \sum_1^n \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}\right)^2. \qquad \text{[See 12.11)]}$$

In the experiments considered in this chapter, the n measurements were the numbers, $O_1, \ldots, O_n$, of times that the value of some quantity x was observed in each of n bins. In this case, the expected number E_k is determined by the assumed distribution of x, and the standard deviation is just $\sqrt{E_k}$; therefore,

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k}. \qquad \text{[See (12.7)]}$$

If the assumed distribution of x is correct, then χ^2 should be of order n. If $\chi^2 \gg n$, the assumed distribution is probably incorrect.

DEGREES OF FREEDOM AND REDUCED CHI SQUARED

If we were to repeat the whole experiment many times, the mean value of χ^2 should be equal to d, the number of *degrees of freedom,* defined as

$$d = n - c,$$

where c is the number of *constraints,* the number of parameters that had to be calculated from the data to compute χ^2.

The *reduced* χ^2 is defined as

$$\tilde{\chi}^2 = \chi^2/d. \qquad \text{[See (12.16)]}$$

If the assumed distribution is correct, $\tilde{\chi}^2$ should be of order 1; if $\tilde{\chi}^2 \gg 1$, the data do not fit the assumed distribution satisfactorily.

PROBABILITIES FOR CHI SQUARED

Suppose you obtain the value $\tilde{\chi}_o^2$ for the reduced chi squared in an experiment. If $\tilde{\chi}_o^2$ is appreciably greater than one, you have reason to doubt the distribution on which your expected values E_k were based. From the table in Appendix D, you can find the probability,

$$Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2),$$

of getting a value $\tilde{\chi}^2$ as large as $\tilde{\chi}_o^2$, assuming the expected distribution is correct. If this probability is small, you have reason to reject the expected distribution; if it is less than 5%, you would reject the assumed distribution at the 5%, or significant, level; if the probability is less than 1%, you would reject the distribution at the 1%, or highly significant, level.

Problems for Chapter 12

For Section 12.1: Introduction to Chi Squared

12.1. ★ Each member of a class of 50 students is given a piece of the same metal (or what is said to be the same metal) and told to find its density ρ. From the 50 results, the mean $\bar{\rho}$ and standard deviation σ_ρ are calculated, and the class decides to test whether the results are normally distributed. To this end, the measurements are grouped into four bins with boundaries at $\bar{\rho} - \sigma_\rho$, $\bar{\rho}$, and $\bar{\rho} + \sigma_\rho$, and the results are shown in Table 12.10.

Table 12.10. Observed densities of 50 pieces of metal arranged in four bins; for Problem 12.1.

Bin number k	Range of bin	Observed number O_k
1	less than $\bar{\rho} - \sigma_\rho$	12
2	between $\bar{\rho} - \sigma_\rho$ and $\bar{\rho}$	13
3	between $\bar{\rho}$ and $\bar{\rho} + \sigma_\rho$	11
4	more than $\bar{\rho} + \sigma_\rho$	14