

A Short Introduction
to
Probability and Statistics
in Physics

Grant W. Mason

“Seek simplicity and distrust it.”... Alfred North Whitehead

Copyright ©(2007) A. Nony Mous. All Rights Reserved.

Contents

1	Probability	5
1.1	Introduction	5
1.2	What can go wrong?	6
1.3	Random Variables and Probability Distributions	6
1.4	Example: The Binomial (or Bernoulli) Distribution	8
1.4.1	Permutations and Combinations	9
1.4.2	Flipping a Coin	9
1.5	Moving Beyond the Binomial Distribution: Radioactive Decay and the Poisson Distribution	12
1.5.1	Example: Using the Poisson Distribution to Estimate	16
1.6	$\langle n \rangle$ Gets Big and n Becomes Continuous: The Normal (Gaussian) Distribution	16
1.7	Two More Continuous Probability Distribution Functions: Uniform and Lorentzian (Cauchy)	20
1.8	A Probability Distribution Function in Time	22
1.9	Cumulative Probability Functions	23
1.9.1	Monte Carlo: Cumulative Probability Functions and the Uniform Probability Distribution Work Together	25
2	Statistics: Pure Math Becomes Applied Math	27
2.1	Statistics to the Rescue!	27
2.2	The Central Limit Theorem and the Law of Large Numbers	28
2.3	Estimating Mean and Variance	30
2.4	Probable Error	31
2.5	Propagation of Errors	32
2.5.1	Example of Propagation of Errors	34
2.5.2	A Second Example of Propagation of Errors: Some- times it Matters!	34

2.5.3	Estimating Propagation of Errors	35
2.5.4	Derivation: Estimating the Variance of the Mean . . .	38
2.5.5	Variation of the Mean: Trial of the Pyx	38
2.6	Estimating Other Parameters: The Maximum Likelihood Method	39
2.6.1	Example: Estimating the Time Constant for Radioac- tive Decay	40
2.7	Finding the Parent Distribution: Chi-Squared	42
2.7.1	χ^2	46
2.7.2	χ^2 as a Random Variable	49
2.8	Least Squares Fitting: Minimizing χ^2	51
2.8.1	Example: Least Squares Fitting to a Straight Line . . .	52
2.8.2	Estimating Uncertainties for the Slope and Intercept of a Linear Fit	54

A Rules of Thumb 58

Chapter 1

Probability

1.1 Introduction

Although physicists use probability and statistics, the typical undergraduate curriculum doesn't usually have room for a formal course. Perhaps this document can serve as a brief introduction to the subject.

There is an element of craftsmanship that goes into the practice of being a physicist. The practice of physics is research. Research means working at the boundary between what is known and what is unknown. Professional physicists must learn to troubleshoot their work by recognizing obvious inconsistencies before presenting their work to others.

Theorists work out equations to describe their work. They have to check their work constantly in ways that will ensure that what they have done is correct. They have to check that their units are right, that the prediction is reasonable, that the result reduces to simpler cases in the proper limits, etc., etc. These methods and process of continual checking and skepticism are a kind of craftsmanship. We don't teach it very well in undergraduate courses. The grader is expected to catch any errors, so why bother?

Experimentalists make measurements. Someone once observed that the first measurements in any experiment are wrong. Again a kind of craftsmanship has to be practiced to ensure that experimental measurements are worthy of acceptance. Some of the methods for this kind of craftsmanship are the methods of statistics.

1.2 What can go wrong?

Errors can creep into experimental results in various ways. Imagine an experiment to measure the number of counts in a detector resulting from some kind of radioactive emissions from a sample placed near the detector. The experiment runs all night, the elapsed time being recorded so that the number of counts per second can be reported. The length of the run is rather accurately controlled to be the same each night. To make sure about the result, the experiment is repeated a number of times over the course of a month. Each time, the raw number of counts n_i is entered into the experimental logbook.

The raw numbers of counts are seldom the same.

There are a number of reasons why that might be the case. It's possible that the electrical power went off for an unknown amount of time during the night of one of the runs or perhaps the power supply to the counter malfunctioned. Or perhaps you wrote down the number of counts incorrectly in the logbook. These are simply *blunders* that can happen to anyone and all the more reason to be careful and skeptical of measurements that seem to be completely inconsistent with expectation or other measurements. About all you can do is fix the problem as soon as it is identified . . . and try again. Statistics can't help.

Some errors are *systematic*. These are reproducible errors that arise from faulty calibration or from measurement biased in one direction or another. For example, the counter may have a short "dead time" following each count during which it is incapable of registering a subsequent count. One could correct for such an error by deducting the total dead time from the length of the observing run. Thus, if known and understood, correction can be made for systematic errors. Statistics can sometimes help.

1.3 Random Variables and Probability Distributions

But even when corrected and free of blunders, the number of counts is seldom the same. When repeated experiments give different answers, the measured variable is called a *random variable*. If one were able to repeat an infinite number of measurements, one could find the percentage of times that a given measured value was obtained. These percentages as a function of their

corresponding measured values are together called the *parent probability distribution function* (p.d.f.). The word “parent” is inserted to indicate that it is the ideal distribution obtained from an infinite number of measurements. Any finite number of measurements, particularly a small number, would give an approximate and imperfect representation of the parent distribution.

In our example, the number of counts is a discrete integer. One gets 12 counts or one gets 13 counts, but one does not get 13.7 counts. If plotted, the p.d.f. for such a case is discrete or is sometimes given a staircase appearance in plots. We might represent it as a mathematical function $P(n; \theta)$, where θ represents one or more parameters that themselves may characterize the shape of the distribution. The value of the function $P(n; \theta)$ is the probability of observing n counts, given θ .

Some random variables are continuous in nature rather than discrete. For example, the speed of molecules in a gas can take on a range of continuous possibilities. For a continuous random variable x , the value of the function $P(x; \theta)$ is the probability of an observation occurring between the values x and $x+dx$, given the parameter(s) θ . In this case $P(x)$ itself is more correctly thought of as a probability density and $P(x)dx$ as probability.

One of the important applications of a probability distribution is the calculation of weighted averages called *expectation values*. In general, a weighted average \bar{x} of some set of values x_i weighted by w_i is given by

$$\bar{x} = \frac{\sum_{i=1}^N x_i w_i}{\sum_{i=1}^N w_i}.$$

If we use a probability distribution as the weighting function such that the sum of all probabilities is 1, we define the expectation value $\langle g(x) \rangle$ of some function $g(n)$ or $g(x)$ to be,

$$\langle g(n) \rangle \equiv \sum_{n_i=1}^N g(n) P_i(n_i; \theta)$$

$$\langle g(x) \rangle \equiv \int_{-\infty}^{\infty} g(x) P(x; \theta) dx.$$

The discrete and continuous distributions differ in the kind of sum, Greek (Σ) or Hebrew (\int). The expectation value $\langle n \rangle$ (or $\langle x \rangle$) is of particular importance and is usually called the *mean*. As we shall see, the *variance* $\langle (n - \langle n \rangle)^2 \rangle$ (or $\langle (x - \langle x \rangle)^2 \rangle$) is also very important. Observe that the expectation value is a real number and that $\langle \langle n \rangle \rangle = \langle n \rangle$, so that $\langle (n - \langle n \rangle)^2 \rangle = \langle n^2 \rangle - \langle n \rangle^2$.

We can also define a *cumulative distribution function* (c.d.f) for each of the two cases which sums the total probabilities from one extreme possibility up to some chosen upper limit possibility:

$$P_c(a) = \sum_{n_i=1}^{n_a} P_i(n_i; \theta)$$

$$P_c(a) = \int_{-\infty}^a P(x; \theta) dx.$$

If the sums are extended across all possibilities, they must sum to 1, i.e., the sum of probabilities of all possibilities is 1.

Statistics packages in most major professional quality mathematics applications such as Matlab, Maple and Mathematica provide on demand the values of common distributions and the associated cumulative functions. The cumulative functions provide the area under the parent probability curves beginning at $-\infty$ to some chosen point on the horizontal axis and represent the cumulative probability of getting a random value from the distribution that equals or is less than the chosen value.

Behind every measurement of a random variable lies some parent probability distribution. Usually, we don't actually know what the parent distribution is and infer what it might be and what its parameters are from approximations taken from the incomplete data that we actually have. And that's where statistics comes in.

1.4 Example: The Binomial (or Bernoulli) Distribution

A probability distribution provides an answer to a question. To know what probability distribution applies, one must know what the question is.

Consider an experiment that has just two possible results. Let the two possibilities be designated Y (yea) or N (nay). For example, flipping a coin to get heads or tails or observing radiation in a counter that discharges or doesn't discharge. Let the probability of Y be p and the probability of N be $1 - p$. The parameter p is an example of the parameters symbolized by θ in $P(n; \theta)$.

Question: What is the probability in K trials that Y occurs n times and hence that N occurs $K - n$ times? That is the question.

1.4.1 Permutations and Combinations

To answer the question, we must first learn to count.

Imagine an egg carton (one dozen variety). Take the eggs out and label each of them with one of the first twelve letters of the alphabet: abcdefghijkl. Now take five eggs (randomly chosen) and put them side-by-side so that they are in an order such as “adlgb.” Now put the eggs back with the other seven and do it again. How many distinct orders of five eggs chosen from among twelve will there be? The different orderings are called *permutations*. By the permutations of N different things taken n at a time, $n \leq N$, is meant all the possible distinct ordered arrangements consisting of n things chosen from N different things,

$$Q(N, n) = \frac{N!}{(N - n)!}.$$

For the labeled five eggs, the number of permutations is $Q(12, 5) = 95040$. That's a lot!

Wash the labels from the eggs. Now take five eggs and put them into the carton in some random choice of slots. Observe the pattern of filled slots. How many different patterns can you create NOT counting any permutations of the eggs for a given pattern as separate. (Equivalently, you could put all twelve eggs into the box and remove them five at a time.) The patterns are *combinations*. By the combinations of N different things taken n at a time, $n \leq N$, is meant all the possible selections of n different things chosen from the N given things, without regard to the order of arrangement or selection,

$$C(N, n) = \frac{Q(N, n)}{Q(n, n)} = \frac{N!}{(N - n)!n!}.$$

For the eggs placed into the carton, the number of combinations is $C(12, 5) = 792$.

1.4.2 Flipping a Coin

Consider five flips of a coin to yield a sequence of heads (Y) and tails (N). The total number of possible sequences of heads and tails for the exercise, including zero to five Y 's is $2^5 = 32$. (The total number of possibilities is the number of five-digit binary numbers 00000 to 11111 that count from 0

to 32.) This number is neither a number of permutations nor a number of combinations as described above.

Of these 32 possibilities, how many have 2 Ys? This number *is* a number of combinations. Put 5 eggs in a row in the egg carton and pull out two at a time (the two Ys) and count the number of patterns that you create in the remaining row of eggs (the Ns). The answer is the number of combinations of N things taken n at a time,

$$C(5, 2) = \frac{N!}{n!(N-n)!} = \frac{5!}{2!(5-2)!} = 10.$$

The combinations with two Ys are:

$YYNNN$ $YNYNN$
 $YNNYN$ $YNNNY$
 $NYYNN$ $NYNYN$
 $NYNNY$ $NNYYN$
 $NNYNY$ $NNNYY$

The probability of each Y is p . The probability of getting two Ys is p^2 . The probability of getting a N is $(1-p)$. The probability of getting 3 Ns is $(1-p)^3$. Therefore the probability of getting two Ys and three Ns is $p^2(1-p)^3$.

Each of the 10 combinations containing 2 Ys has this same probability. For flipping an honest coin, $p = 1/2$ and

$$p^2(1-p)^3 = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{1}{32}.$$

But, there are 10 combinations with 2 Ys, so that the total probability for 2 Ys is the number of combinations with 2 Ys times the probability of each one,

$$P_5(2) = C(5, 2)p^2(1-p)^3,$$

or $10/32$ for the case of the coin.

In general, for K trials (e.g., flips of a coin), the probability of obtaining n Ys and, hence, $(K-n)$ Ns is given by,

$$P_K(n) = C(K, n)p^n(1-p)^{K-n}.$$

This expression for P may be familiar to you. It is n -th term in the binomial expansion of $(p + q)^K$, where we have taken $q = 1 - p$. Thus, since $p + q = p + (1 - p) = 1$,

$$(p + q)^K = \sum_{n=0}^K P_K(n) = 1$$

is just a reassuring statement that the probabilities of all of the possibilities sum to 1.

The expectation value $\langle n \rangle$ (or *mean*) that a number n of Y s occurs in K trials (flips of a coin, for example) is

$$\langle n \rangle = \sum_{n=0}^K n P_K(n) = \sum_{n=0}^K n C(K, n) p^n (1 - p)^{K-n}.$$

The mean is a measure of the central tendency, i.e., a measure of the tendency of repeated observations to cluster around a central value.

However, observe that

$$\frac{\partial}{\partial p} (p + q)^K = \frac{\partial}{\partial p} \sum_{n=0}^K C(K, n) p^n q^{K-n} = \sum_{n=0}^K n C(K, n) p^{n-1} q^{K-n} = \frac{1}{p} \langle n \rangle,$$

so that,

$$\langle n \rangle = p \frac{\partial}{\partial p} (p + q)^K = p K (p + q)^{K-1} = p K \quad (1.1)$$

because $p + q = 1$.

The *variance* of a distribution σ^2 is defined to be

$$\sigma^2 = \langle (n - \langle n \rangle)^2 \rangle.$$

Its square root σ is called the *standard deviation* of the distribution. These are measures of what someone has described as the “wildness” of the distribution, i.e., a measure of how much individual values of n differ from the mean $\langle n \rangle$.

$$\langle (n - \langle n \rangle)^2 \rangle = \langle (n^2 - 2n \langle n \rangle + \langle n \rangle^2) \rangle = \langle n^2 \rangle - \langle n \rangle^2.$$

Can we adapt the trick that we used to show that $\langle n \rangle = Kp$ to handle $\langle n^2 \rangle$?

$$\begin{aligned}
\langle n^2 \rangle &= \sum_{n=0}^K n^2 P_K(n) \equiv \sum_{n=0}^K (n(n-1) + n) C(K, n) p^n q^{K-n} \\
&= \left(p^2 \frac{\partial^2}{\partial p^2} + p \frac{\partial}{\partial p} \right) \sum_{n=0}^K C(K, n) p^n q^{K-n} \\
&= \left(p^2 \frac{\partial^2}{\partial p^2} + p \frac{\partial}{\partial p} \right) (p+q)^K = K(K-1)p^2 + Kp,
\end{aligned}$$

Thus,

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = K(K-1)p^2 + Kp - p^2 K^2 = Kp(1-p) = Kpq$$

and

$$\sigma = (Kpq)^{1/2}.$$

The mean of a distribution is a parameter that is a measure of the tendency for repeated values drawn from the parent probability distribution to tend toward a “normal” or central value. On the other hand, the variance is a parameter that indicates how much individual values tend to deviate away from the central. We might have tried to use $\langle (n - \langle n \rangle) \rangle$, but $\langle n \rangle - \langle n \rangle = 0$ doesn’t tell us very much. Hence, we use a definition of the variance that precludes the cancellation of positive and negative variations relative to the mean. See Table 2.1 for a concrete example.

1.5 Moving Beyond the Binomial Distribution: Radioactive Decay and the Poisson Distribution

There is more to physics than just flipping a coin. Can the binomial distribution be adapted to give the answer to a more interesting question? Imagine a radioactive substance that has been observed with a disintegration counter long enough to conclude that the mean number of disintegrations in some period of time is μ and that the overall activity of the sample is not diminishing significantly over time. The question is: What is the probability $P(n; \mu)$ that in a given period of time, such as a second, there are n disintegrations? (Note that n is a pure number, NOT a rate.)

Imagine dividing the period of time into K equal subintervals, where K is a very large number. Referring to Eq. (1.1), in any one subinterval the probability of disintegration p is μ/K and the probability of no disintegrations q in that subinterval is $1 - \mu/K$. K is taken large enough so that there is no more than one disintegration in any subinterval.

The probability of disintegrations in two intervals is

$$\frac{\mu}{K} \frac{\mu}{K} = \left(\frac{\mu}{K}\right)^2$$

and of no disintegrations in two intervals is

$$\left(1 - \frac{\mu}{K}\right) \left(1 - \frac{\mu}{K}\right) = \left(1 - \frac{\mu}{K}\right)^2.$$

Similarly, for disintegrations in n intervals, but *not* in $K - n$ intervals,

$$\left(\frac{\mu}{K}\right)^n \left(1 - \frac{\mu}{K}\right)^{K-n}.$$

The number of combinations that are possible for n counts and *not* $K - n$ is $C(K, n)$ so that,

$$P_K(n) = C(K, n) \left(\frac{\mu}{K}\right)^n \left(1 - \frac{\mu}{K}\right)^{K-n},$$

i.e., the binomial distribution.

But the result is not of much use to us in this latter application because it contains K which is a very large but unknown parameter. Do the expressions have meaning if the large value of K is stretched to be infinite?

There is a remarkable series expansion, called the Stirling series, that is worth noting at this point.

$$n! = n^n e^{-n} \sqrt{2\pi n} \exp\left(\frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^5} + \dots\right)$$

In taking K to be very large and ultimately becoming infinite, we are helped by Stirling's approximation (for large n)

$$\ln n! \approx n \ln n - n + \frac{1}{2} \ln 2\pi n.$$

For sufficiently large n , the last term in the approximation is usually dropped as well. Using Stirling's approximation in this latter form on

$$\ln[C(K, n)] = \ln \left[\frac{K!}{(K-n)!n!} \right]$$

and observing that

$$\ln(K-n) = \ln \left[K \left(1 - \frac{n}{K} \right) \right] = \ln K + \ln \left(1 - \frac{n}{K} \right) \rightarrow \ln K$$

we obtain

$$\lim_{K \rightarrow \infty} C(K, n) = \lim_{K \rightarrow \infty} \frac{K!}{(K-n)!n!} = \frac{K^n}{n!}.$$

We also use a “lookable-uppable” relationship involving the exponential function,

$$\lim_{K \rightarrow \infty} \left(1 + \frac{x}{K} \right)^K = e^x$$

to show that

$$\lim_{K \rightarrow \infty} \left(1 - \frac{\mu}{K} \right)^{K-n} = \lim_{K \rightarrow \infty} \frac{(1 - \mu/K)^K}{(1 - \mu/K)^n} = \frac{e^{-\mu}}{1} = e^{-\mu}.$$

Thus,

$$P(n; \mu) = \lim_{K \rightarrow \infty} C(K, n) \left(\frac{\mu}{K} \right)^n \left(1 - \frac{\mu}{K} \right)^{K-n} = \frac{\mu^n}{n!} e^{-\mu}$$

In this latter form, the distribution is known as the *Poisson distribution*. It is a discrete expression and answers the question: For a system yielding a mean number of counts μ per unit time, what is the probability of obtaining n counts in a particular given unit of time? The arbitrary parameter K of the binomial distribution has disappeared to be replaced by the mean μ , the expectation value $\langle n \rangle = p/K$ has become μ and the standard deviation $\sigma = \langle (n - \langle n \rangle)^2 \rangle^{1/2}$ of the binomial distribution has been replaced by,

$$\sigma = (Kpq)^{1/2} \rightarrow \sigma = \lim_{K \rightarrow \infty} \left[K \frac{\mu}{K} \left(1 - \frac{\mu}{K} \right) \right]^{1/2} = \mu^{1/2}.$$

Figure 1.1 shows Poisson distributions for different values of the parameter μ . As the mean value μ increases, the distributions become more and more symmetric with respect to the peak.

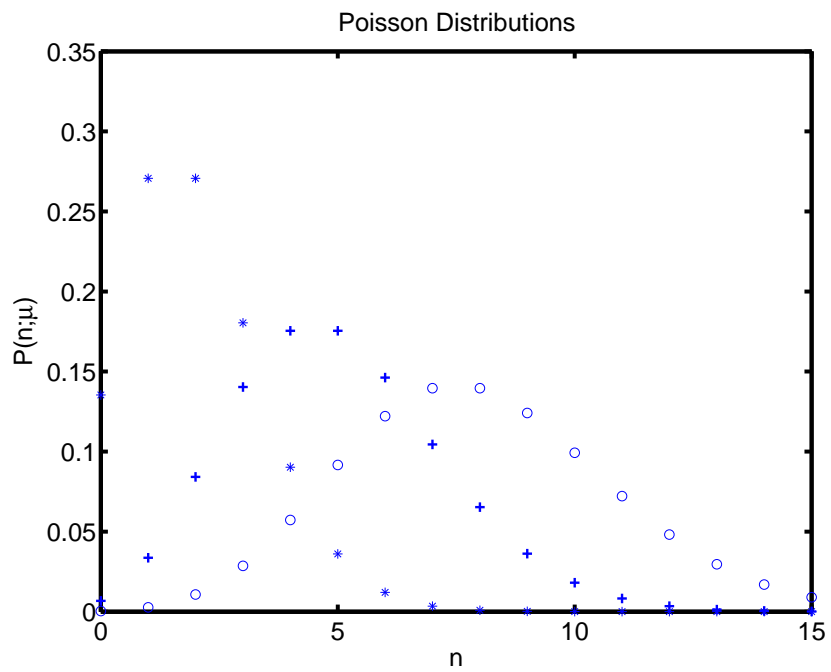


Figure 1.1: Poisson distributions for $\mu = 2, 5, 8$ (left to right). The distributions are defined for discrete values of n .

1.5.1 Example: Using the Poisson Distribution to Estimate

In Table 2.1 we see the results of ten one-minute measurements of counts from a radioactive source. The average is 39541 counts per minute or 659 counts per second. What is the probability of finding an interval of t seconds without any counts? What is the period of time $t_{1/2}$ for which the probability of zero counts is 50%?

If we take $R = 659$ counts per second to be the average count rate, then the mean number of counts in a period of time is $\mu \approx Rt$. The probability of n counts is then given by the Poisson distribution,

$$P(n) = \frac{\mu^n}{n!} e^{-\mu} \approx \frac{(Rt)^n}{n!} e^{-Rt}.$$

Then, $P(0) \approx e^{-Rt}$ and the period of time for which $P(0) = 0.5$ is $t_{1/2} = \ln 2/R \approx 1$ millisecond.

1.6 $\langle n \rangle$ Gets Big and n Becomes Continuous: The Normal (Gaussian) Distribution

The Poisson distribution is *discrete*, i.e., it involves integer values for n . What about continuous random variables?

For large values of n , the Poisson distribution has a maximum peak that falls near (but not exactly) $\langle n \rangle = \mu$. Formally we can find the maximum of the peak by finding the maximum of $P(n)$, or equivalently, of $\ln P(n)$, by taking the first derivative and equating it to zero. To do so, we use Stirling's approximation in the form,

$$\ln n! \approx n \ln n - n + \frac{1}{2} \ln 2\pi n.$$

Thus,

$$\begin{aligned} P(n) &= \frac{\mu^n}{n!} e^{-\mu}, \\ \ln P(n) &= n \ln \mu - \ln n! - \mu \\ &\approx n \ln \mu - n \ln n + n - \frac{1}{2} \ln 2\pi n - \mu, \end{aligned}$$

and,

$$\frac{d \ln P(n)}{dn} \approx \ln \mu - \ln n - \frac{1}{2n}. \quad (1.2)$$

Setting the latter to zero, we see that the maximum (for large n) lies near μ . Given that the maximum does lie near μ , we can use a Taylor's expansion near that value,

$$\ln P(n) = \ln P(\mu) + \frac{(n - \mu)^2}{2!} \left[\frac{d^2}{dn^2} \ln P(n) \right]_{n=\mu} + \dots$$

But, from Eq. (1.2),

$$\frac{d^2 \ln P(n)}{dn^2} \approx -\frac{1}{n} + \frac{1}{2n^2},$$

so, for large n ,

$$\ln P(n) \approx -\frac{1}{2} \ln 2\pi\mu - \frac{(n - \mu)^2}{2\mu}.$$

or, reverting to exponential form,

$$P(n; \mu) = \frac{1}{\sqrt{2\pi\mu}} e^{-(n-\mu)^2/2\mu}.$$

When expressed as a continuous random variable, x , this distribution is known as the *normal* or *Gaussian* distribution,

$$P(x; \mu) = \frac{1}{\sqrt{2\pi\mu}} e^{-(x-\mu)^2/2\mu}.$$

Figure 1.2 shows a Gaussian distribution centered on a mean value of zero and with $\sigma = 1.0$. The distribution is symmetric with respect to its peak.

Figure 1.3 shows a comparison between Poisson and Gaussian distributions when μ becomes large. For large μ , the Poisson distribution is nearly symmetric with respect to its peak and one can see that the Gaussian distribution evolves from a Poisson distribution.

The mean of the distribution remains $\langle x \rangle = \mu$ and the variance also remains μ . In the continuous form it is the familiar bell-shaped curve that characterizes everything from IQs to grade distributions in college courses. Although we have used some approximations along the way, the result is actually normalized as it stands,

$$\int_{-\infty}^{+\infty} P(x) dx = 1.$$

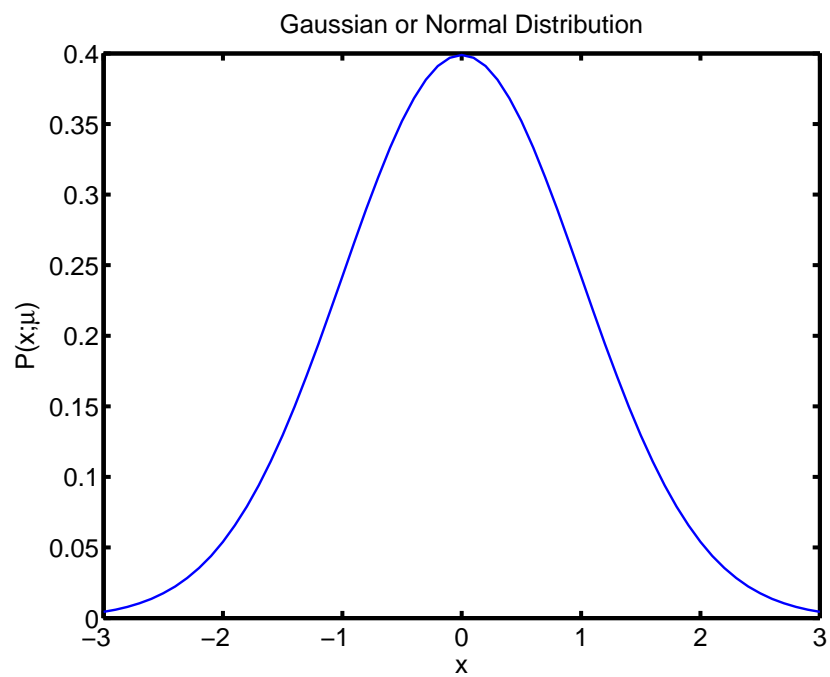


Figure 1.2: A one-parameter Gaussian distribution with $\mu = 0.0$ and $\sigma = 1.0$.

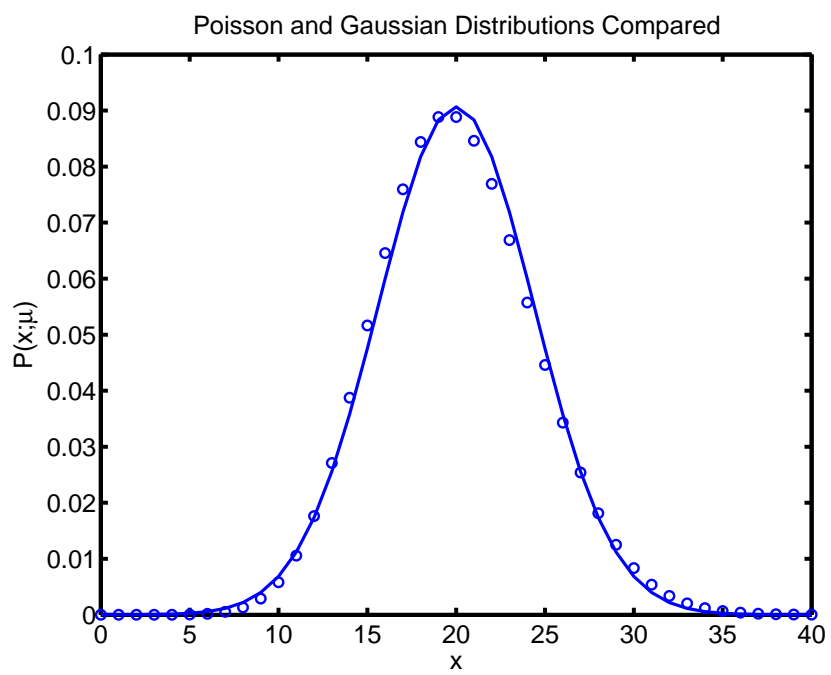


Figure 1.3: Gaussian and Poisson distributions compared for $\mu = 20$ (Poisson) and $\sigma = 4.4$ (Gaussian)

We can also change the Gaussian distribution into a more general *two* parameter distribution by severing the connection between the mean of the distribution μ and the variance of the distribution σ^2 ,

$$P(x; \mu, \Delta) = \frac{e^{-(x-\mu)^2/2\Delta^2}}{(2\pi\Delta^2)^{1/2}} \quad \text{two parameter Gaussian.}$$

Now the mean $\langle x \rangle = \mu$ and the variance $\sigma^2 = \Delta^2$. The normalization remains intact.

However, the importance of the Gaussian distribution in probability and statistics is out of proportion to the way we have derived it.

1.7 Two More Continuous Probability Distribution Functions: Uniform and Lorentzian (Cauchy)

There are two more continuous probability distributions that routinely arise in physics.

The first is the simplest of all, the uniform distribution. If all values between two limits a and b are equally probable, the distribution is said to be uniform. The uniform distribution answers the question: Given that x is between a and b and that all x are equally probably, what is the probability that x is between x and $x + dx$?

We can formally write the distribution,

$$P(x; a, b) = \begin{cases} 1/(b-a) & , \quad a \leq x \leq b \\ 0 & , \quad \text{otherwise.} \end{cases}$$

The distribution is properly normalized, $\int_a^b P(x)dx = 1$. The mean and variance are given by,

$$\langle x \rangle = \frac{1}{2}(b+a) \quad \sigma^2 = \frac{1}{12}(b-a)^2$$

Now imagine the following. Imagine a small radioactive source that randomly decays, emitting some kind of detectable particle. The source radiates uniformly at all angles θ . At some distance D is a wall running perpendicular to a line drawn from the source to the wall. Measure θ from this line. A small

detector can be positioned at various positions $\pm x$ along the wall, measured in such a way that $\tan \theta = x/D$. What is the probability of a decay particle hitting a detector covering the range from x to $x + dx$?

The problem is essentially that of changing from the uniform probability distribution $P_u(\theta)$ to the probability distribution $P_x(x)$ in such a way that

$$P_u(\theta)d\theta = P_x(x)dx.$$

The limits a and b in this case are $-\pi/2$ and $+\pi/2$, so that $P_u(\theta)d\theta = d\theta/\pi$. Since $\theta = \arctan x/D$,

$$d\theta = \frac{dx/D}{1 + (x/D)^2}$$

and,

$$P_u(\theta)d\theta = \frac{1}{\pi} \frac{D}{x^2 + D^2} dx = P_x(x)dx.$$

The probability function $P_x(x)$ is an even function of x and is therefore symmetric about $x = 0$. It has its maximum for $x = 0$, which minimizes the denominator. Further, its half-maximum occurs at $x_{1/2} = \pm D$, so that its full-width, taken at half-maximum is $2D$. Let us denote this measure of the width of the distribution as $\Gamma (= 2D)$. We can displace the distribution to center on a non-zero mean by replacing x with $x - \langle x \rangle$. With these adjustments, we can write the distribution in its more common form,

$$P_x(x; \mu, \Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(x - \mu)^2 + (\Gamma/2)^2}. \quad (1.3)$$

In this form it is known as either the Lorentzian or Cauchy probability distribution.

Again, the importance of the Lorentzian distribution is out of proportion to the very unusual and limited example that we used to derive it. If it had no further application than this example, we would probably consider it just a curiosity. But the general shape of this distribution comes up in a number of circumstances, usually related to the phenomenon of resonance.

If a sample of aluminum containing the nucleus Al^{27} is bombarded with a beam of low energy protons (500 to 1400 Kev), some of the nuclei transmute to Si^{28} and emit gamma rays. However, for certain proton energies, the yield of gamma rays is dramatically increased. Near one of these unusual "resonant" proton energies, we can plot the yield of gamma rays as a function

of proton energy. There are over 30 such energies for Al^{27} . Both experiment and theory (Breit-Wigner) indicate that the yield as a function of energy $Y(E)$ near one of these resonances is given by,

$$Y(E) = F(E) \frac{(\Gamma/2)^2}{(E - E_0)^2 + (\Gamma/2)^2}$$

where E_0 is the resonant energy and $F(E)$ is a slowly varying function of proton energy. Except for the normalization, this is essentially just the Lorentzian probability distribution. When the proton has certain energies, there is an increased probability that a gamma ray will be emitted. The Lorentzian seems to arise naturally here and elsewhere when a phenomenon exhibits resonance behavior.

The Lorentzian of Eq. (1.3) is properly normalized, $\int P(x; \mu, \Gamma) = 1$. The position of the maximum and the value of the mean are both the parameter μ . However, the Lorentzian has an unusual feature. Variance is undefined,

$$\sigma^2 = \int_{-\infty}^{+\infty} x^2 \left[\frac{1}{\pi} \frac{\Gamma/2}{x^2 + (\Gamma/2)^2} \right] dx \rightarrow \infty.$$

This unexpected result comes about because of the relatively slow decrease of the wings of the distribution. Lacking a variance and a standard deviation, we therefore characterize the width of the distribution by the full-width at half-maximum Γ . For a symmetric Gaussian distribution, we could use either measure of width, σ or Γ .

1.8 A Probability Distribution Function in Time

There are situations that we can imagine where a number of “events” that occurs in a specific time is proportional to a number of “parents” of the events. The number of children born in a given period of time is proportional to the number of eligible parents and the amount of time. The number of radioactive decays from a sample of nuclei is proportional to the time elapsed and the number of radioactive nuclei.

We can express this observation as a differential expression,

$$dn = \pm \lambda n dt.$$

Here dn represent a number of events in a short time dt for which the number of parents is n . What we mean by “ dt is small” is that $dn \ll n$ and that during a long period of time the proportionality constant λ is constant. Dividing both sides by n and integrating, we obtain the familiar expression for exponential growth or decay,

$$n = n_0 e^{\pm \lambda t},$$

where n_0 is the initial population at $t = 0$.

Now if we write, for the case of exponential decay,

$$dn = -n_0 \lambda e^{-\lambda t} dt,$$

we observe that dn/n_0 represents a fraction of the original parent population n_0 that decays between t and $t + dt$. Taken as a function of time, these fractions represent probabilities whose overall sum is one, i.e., a probability density function for the continuous variable t ,

$$P(t; \lambda) = \lambda e^{-\lambda t}. \tag{1.4}$$

The distribution function is normalized, $\int_0^\infty P(t; \lambda) dt = 1$. The probability distribution answers the question: Given n_0 , what is the probability of the fraction dn/n_0 occurring between t and $t + dt$?

1.9 Cumulative Probability Functions

Each parent probability distribution has a *cumulative probability function*. For a continuous parent distribution $P(x; \theta)$,

$$P_c(x) = \int_{-\infty}^x P(x'; \theta) dx'.$$

Figure 1.4 shows the cumulative function for the Gaussian distribution shown in Fig. 1.2. The curve represent the cumulative probability of obtaining a value of a random variable from the parent distribution between $-\infty$ and x . Since, $\int_{-\infty}^{+\infty} P(x'; \theta) dx' = 1$, the curve asymptotically approaches 1 as $x \rightarrow +\infty$. The values on the vertical axis of the cumulative function represent the increasing area under the parent distribution curve as x increases from $-\infty$ to some chosen value.

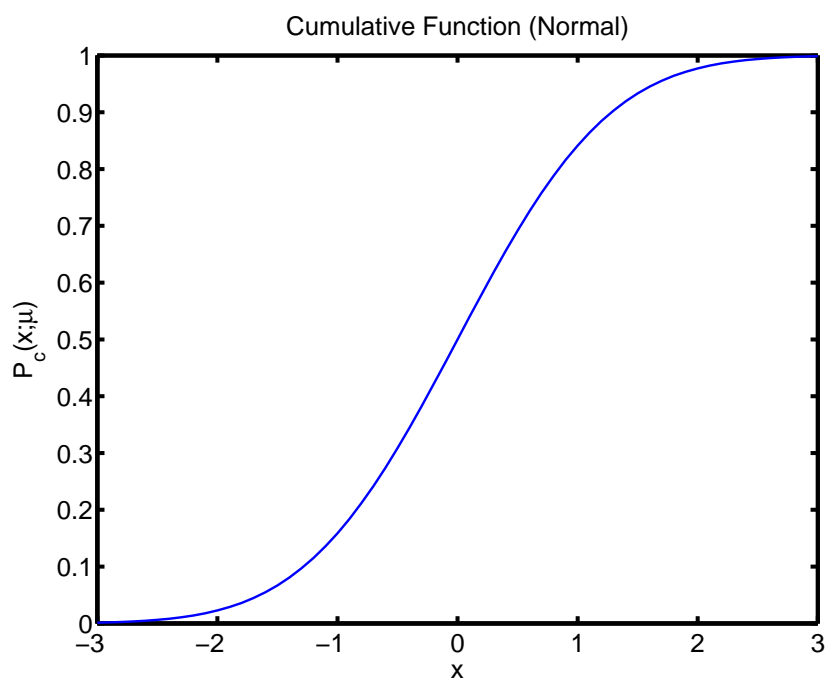


Figure 1.4: The cumulative probability function for the Gaussian distribution of Fig. 1.2. $\mu = 0.0$ and $\sigma = 1.0$.

1.9.1 Monte Carlo: Cumulative Probability Functions and the Uniform Probability Distribution Work Together

Choose a value x and a small increase dx on the horizontal axis of a cumulative probability curve. Run vertical lines until they intersect the curve. See Fig. 1.5. From the intersections, run horizontal lines to the vertical axis. The difference between the two intersect values on the vertical axis represents the increment in area added to the cumulative function when x increases to $x + dx$.

We can turn it around. Choose a random number from a *uniform* distribution of random numbers on the range $(0, 1)$. Every time we get a number in the narrow horizontal sliver (between 0.7 and 0.72 in Fig. 1.5), it would correspond to getting a value from the parent distribution lying between x and $x + dx$. Ranges of random numbers where the cumulative curve is steepest will get more hits for a value of x than those that are relatively flat. In fact, the distribution of x values that we get from using random numbers from a uniform distribution in this way will just be distributed as the parent probability distribution of the cumulative function used!

In so-called “Monte Carlo” computer simulations it is sometimes necessary to generate a series of randomly chosen variables taken from an assumed parent distribution. An algorithm based on the method described above using the cumulative function provides the needed random variables. Most computer languages have built-in random number generators that choose a number on the range from $(0, 1)$ from a uniform distribution.

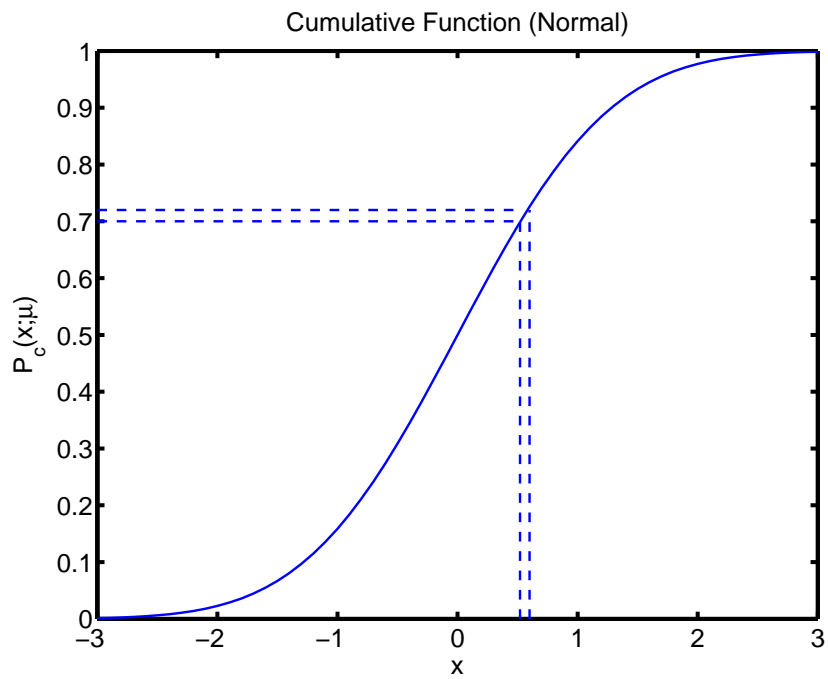


Figure 1.5: The probability for obtaining a value of the random variable between x and $x + dx$ on the horizontal scale is the difference between the values marked by the horizontal lines in a plot of the cumulative probability function.

Chapter 2

Statistics: Pure Math Becomes Applied Math

2.1 Statistics to the Rescue!

There is a problem with all of this. Whenever you wish to use the probability distribution functions, you have to provide the parameters. What shall we use for μ ? You can see the problem by looking at Table 2.1. The table shows a series of recorded counts from a radioactive sample, each taken for one minute. If you were to take the average of just two of the entries for n , then an average of three, until finally you take an average of all ten entries, you will get a different average each time. But the parameter μ is just one number, assumed to be the true mean taken from an infinite number of one minute runs. What we have done is to take just a sample of this infinite number of possibilities. Depending on which of the individual one-minute measurements we include in our average, we see that the average calculated from samples is itself a random variable. We would like to use the average as an estimate for the true mean, but which one should we use and how much does it matter which we choose? And here we come to two of the most important theorems in statistics: The Central Limit Theorem and the Laws of Large Numbers.

Table 2.1: Sample Numerical Data from a Radioactive Counting Experiment

n	$n - \mu_s$	$(n - \mu_s)^2$
39684	143	20449
39328	-213	45369
39498	-43	1849
39728	187	34969
39472	-69	4761
39750	209	43681
39315	-226	51076
39204	-337	113569
39736	195	38025
39696	155	24025
$\mu \approx \mu_s = 39541.1$	$\Sigma = 1$	$\Sigma = 377773$
$\sigma_p(n) \approx \left(\frac{377773}{9}\right)^{1/2} = 204.9 \quad \sigma_s(\mu) \approx 64.8$		

2.2 The Central Limit Theorem and the Law of Large Numbers

In the jargon of statistics, each of the possible minutes that we could have used to make a measurement of the number of decays is a element of a *population*. When we use just a small subset of minutes of this population for our actual measurements, we are taking a *sample* of the population. Of course, we could take lots of different samples and each would have its own average. If we take many samples, we see that the average obtained from each is itself a random variable taken from its own underlying distribution. A plot of the number of times a certain average is obtained versus the value of the average is called a *sampling distribution*. The Central Limit Theorem tells us that a sampling distribution always has significantly less deviation (or “wildness”) than the population it is drawn from. Additionally, the sampling distribution will act more and more like a normal (Gaussian) distribution as the sample size is increased, even when the population itself is not normally distributed! This is a very remarkable result. (If you search the internet for

the Central Limit Theorem, you can find interactive exercises where quite radically nonnormal population distributions nevertheless yield a sampling distribution that is noticeably “normal” in appearance and much less “wild” than random values taken from the population distribution.)

A consequence of the Central Limit Theorem is contained in the relationship between the variance of the population compared to the variance of the sampling distribution,

$$\sigma_s^2 = \frac{\sigma_p^2}{N} \quad (2.1)$$

where N is the sample size, i.e., the number of elements of the “population” in the sample. (We will derive this result in Sec. 2.5.4.)

The significance of the Central Limit Theorem is that we can have considerable confidence that an average taken from even a single large, randomly chosen sample, will be close to the true mean of the population. Without this confidence in the mean, we would be unable to use probability distributions that depend on the mean as a parameter. It is the job of statistics, as opposed to probability theory, to devise ways of estimating the parameters of probability distributions from finite samples. If the methods of statistics were unable to estimate reliably a parameter such as the mean, research in the social sciences and politics would probably have to shut down! Observe particularly that the variance of the sampling distribution get smaller as the sample size N increases and that the total size of the population does not appear in the expression, other than through the assumption that it is much larger than N .

Without deriving the theorems rigorously, perhaps we can best state practical versions of them as Rules of Thumb

1. Law of Large Numbers (Bernoulli’s Law): When a large number of independent events is taken from a parent population, the sample’s average will be close to the true mean of all events, including those that were not observed.
2. Central Limit Theorem: Assume that a variable x has a mean μ and variance σ^2 . If σ^2 is finite, then the distribution of the sample mean μ_s approaches a Gaussian distribution with mean μ and variance σ^2/N as N tends to infinity.

2.3 Estimating Mean and Variance

The variance of the population (of one-minute counts, for example) is defined to be,

$$\sigma_p^2(n) = \langle (n_i - \langle n \rangle)^2 \rangle = \langle (n_i - \mu)^2 \rangle.$$

If we take a sample of N measurements and if we replace μ with an average μ_s calculated from the sample,

$$\mu \approx \mu_s = \frac{1}{N} \sum_{i=1}^N n_i,$$

$$\sigma_p^2 \approx \frac{1}{N-1} \sum_{i=1}^N (n_i - \mu_s)^2.$$

Dividing by $N - 1$ instead of N compensates approximately for using μ_s instead of the true value μ and is justified more quantitatively in standard textbooks on statistics and probability. Basically, if we divided by N , in the case of a single measurement for which $N = 1$ and $n_1 = \mu_s$, we would have the embarrassment that the variance (“uncertainty”) vanishes for an average taken from just one measurement! Putting $N - 1$ in the denominator at least makes the expression indeterminate in this special case. The value of $N - 1$ in the denominator of the expression represents the independent “degrees of freedom” in the expression, but we deduct one of these degrees of freedom from N because the sum of the values n_i has the constraint that $\sum n_i = N\mu_s$, i.e., the values n_i in the complete expression for σ^2 are not completely independent since the expression contains μ_s .

The variance of the sampling distribution is, from Eq.(2.1),

$$\sigma_s^2(\langle n \rangle) \approx \frac{1}{N(N-1)} \sum_{i=1}^N (n_i - \mu_s)^2.$$

We can take as Rules of Thumb,

1. In the case of a single measurement in a counting experiment, $\mu \approx \mu_x = n_1$.
2. In the case of a single measurement in a counting experiment for which the parent distribution is approximately Poisson or Gaussian, $\sigma_p \approx \mu_s^{1/2} = n_1^{1/2}$.

Table 2.2: Probability P of a Given Multiple ϵ of the Standard Deviation σ_p for a Gaussian distribution.

ϵ	0	0.6745	1.0000	1.5000	2.0000	2.5000	3.0000	3.5000	4.0000
P	1	0.5000	0.3173	0.1336	0.0455	0.0124	0.0027	0.00046	0.000063

Table 2.1 contains a sample ($N = 10$) of measurements of one-minute counts of an imagined radioactive substance. For this sample we have, $\mu_s = 39541.1$, $\sigma_p(n) = 204.9$, and, $\sigma_s(\mu_s) = 64.8$. If the parent distribution were Gaussian, the standard deviation σ_p could be calculated from $\sigma_p \approx \mu_s^{1/2} = 190$, which is in approximate agreement with σ_p calculated from,

$$\sigma_p = \left[\frac{1}{9} \sum_{i=1}^{10} (n_i - \mu_s)^2 \right]^{1/2} = 204.9.$$

2.4 Probable Error

We might ask the following question: What is the probability, in the case of a single individual count n_1 , that the deviation from the “true” result n_t will be some multiple ϵ of σ ? If the distribution were Gaussian, we could give an answer to this question. In this case we assume that $n_t = \mu$. See Table 2.2. In particular, the probability of observing a counting rate differing from μ by less than $0.6745\sigma_p$ is 0.5. The *probable error* is defined to be the absolute value of the deviation $|n - n_t|$ such that the probability for the deviation of any single, random observation $|n_1 - n_t|$ taken from the underlying probability distribution is less than 50%.

Data for the experiment of Table 2.1 would then be reported giving the mean $\mu_s = 39541$ counts per minute to indicate the central tendency and $0.67\sigma_s = 44$ to indicate the “wildness”, 39541 ± 44 counts per minute. In plots of the data, “flags” would be attached to the data point to indicate this range of values.

Rules of Thumb:

1. Vertical bars are often drawn attached to displayed data points extending $\pm 0.67\sigma$ to indicate an estimate of the range from the mean within which there is a 50% probability of a measured random variable falling if taken from a distribution that is approximately Gaussian.

2. If taken from a distribution that is approximately Gaussian, it is estimated that there is less than a 1% probability of the measured value falling outside $\pm 3\sigma$ from the measured value (which is taken as an estimate of the mean.)

2.5 Propagation of Errors

As we see in the definition of probable error, the standard deviation σ associated with a probability distribution is used to gauge our certainty (or uncertainty) that a measured value is a good representation of the true value. However, very often the result that we are after is not simply a direct measurement of counts, but rather something calculated from the measured value of counts or calculated from the separate measurement of more than one thing. For example, we may have a background that contributes to the counts we measure when the radioactive source of interest itself is not present. We will have to make measurements both with and without the source present and subtract one number from the other to get our desired result. How do the uncertainties that apply to these separate measurements propagate into the uncertainty of the final answer?

Let's assume that we make two series of measurements, n_i^α and n_i^β from which we calculate a desired result $x_i = f(n_i^\alpha, n_i^\beta)$. (The following argument can be extended to more than two measurements in a straightforward way.) Let us assume that the two measurements have true means μ_α and μ_β , for which x takes on its mean value $\mu_x = f(\mu_\alpha, \mu_\beta)$.

Since individual measurements n_i can be expected to lead to values of x that are close to the mean, we can use a Taylor's expansion in two dimensions to express,

$$x_i \approx \mu_x + (n_i^\alpha - \mu_\alpha) \left(\frac{\partial f}{\partial n_i^\alpha} \right)_{\mu_\alpha, \mu_\beta} + (n_i^\beta - \mu_\beta) \left(\frac{\partial f}{\partial n_i^\beta} \right)_{\mu_\alpha, \mu_\beta} + \dots$$

But, by definition,

$$\sigma_x^2 = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2.$$

If we combine these two expressions,

$$\sigma_x^2 \approx \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \left[(n_i^\alpha - \mu_\alpha) \left(\frac{\partial f}{\partial n_i^\alpha} \right)_{\mu_\alpha, \mu_\beta} + (n_i^\beta - \mu_\beta) \left(\frac{\partial f}{\partial n_i^\beta} \right)_{\mu_\alpha, \mu_\beta} + \dots \right]^2$$

$$\begin{aligned} &\approx \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \left[(n_i^\alpha - \mu_\alpha)^2 \left(\frac{\partial f}{\partial n_i^\alpha} \right)_{\mu_\alpha, \mu_\beta}^2 + (n_i^\beta - \mu_\beta)^2 \left(\frac{\partial f}{\partial n_i^\beta} \right)_{\mu_\alpha, \mu_\beta}^2 \right] \\ &+ \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \left[2(n_i^\alpha - \mu_\alpha)(n_i^\beta - \mu_\beta) \left(\frac{\partial f}{\partial n_i^\alpha} \right)_{\mu_\alpha, \mu_\beta} \left(\frac{\partial f}{\partial n_i^\beta} \right)_{\mu_\alpha, \mu_\beta} \dots \right] \end{aligned}$$

We define the *covariance*,

$$\sigma_{\alpha\beta}^2 = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N [2(n_i^\alpha - \mu_\alpha)(n_i^\beta - \mu_\beta)]$$

and observe that

$$\begin{aligned} \sigma_\alpha^2 &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (n_i^\alpha - \mu_\alpha)^2 \\ \sigma_\beta^2 &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (n_i^\beta - \mu_\beta)^2 \end{aligned}$$

to arrive at the conclusion,

$$\sigma_x^2 \approx \sigma_\alpha^2 \left(\frac{\partial f}{\partial n^\alpha} \right)_{\mu_\alpha, \mu_\beta}^2 + \sigma_\beta^2 \left(\frac{\partial f}{\partial n^\beta} \right)_{\mu_\alpha, \mu_\beta}^2 + 2\sigma_{\alpha\beta}^2 \left(\frac{\partial f}{\partial n^\alpha} \right)_{\mu_\alpha, \mu_\beta} \left(\frac{\partial f}{\partial n^\beta} \right)_{\mu_\alpha, \mu_\beta} .$$

We can expect that the first two terms will dominate because the variance arises from a sum of strictly positive terms whereas we can expect that if the measurements n_i^α and n_i^β are completely uncorrelated, positive and negative terms will tend to drive the covariance term towards zero. For that reason, we usually take

$$\sigma_x^2 \approx \sigma_\alpha^2 \left(\frac{\partial f}{\partial n^\alpha} \right)_{\mu_\alpha, \mu_\beta}^2 + \sigma_\beta^2 \left(\frac{\partial f}{\partial n^\beta} \right)_{\mu_\alpha, \mu_\beta}^2 ,$$

or, more generally, as a Rule of Thumb,

$$\sigma_x^2 \approx \sum_\alpha \sigma_\alpha^2 \left(\frac{\partial f}{\partial n^\alpha} \right)_{\mu_\alpha}^2 . \quad (2.2)$$

2.5.1 Example of Propagation of Errors

Let us assume for sake of example that

$$x_i = f(n_i^\alpha, n_i^\beta) = \frac{an_i^\alpha}{n_i^\beta}.$$

If we take samples n_i to calculate x_i , we get a distribution of values for x_i that has a variance. Then,

$$\left(\frac{\partial f}{\partial n_i^\alpha}\right)_{\mu_\alpha, \mu_\beta} = \frac{a}{\mu_\beta}, \quad \left(\frac{\partial f}{\partial n_i^\beta}\right)_{\mu_\alpha, \mu_\beta} = -\frac{a\mu_\alpha}{(\mu_\beta)^2}$$

yields

$$\sigma_x^2 = \sigma_\alpha^2 \left(\frac{a^2}{(\mu_\beta)^2}\right) + \sigma_\beta^2 \left(\frac{a^2(\mu_\alpha)^2}{(\mu_\beta)^4}\right)$$

or, in a more symmetric form,

$$\frac{\sigma_x^2}{\mu_x^2} = \frac{\sigma_\alpha^2}{\mu_\alpha^2} + \frac{\sigma_\beta^2}{\mu_\beta^2},$$

where we have used,

$$\mu_x = \frac{a\mu_\alpha}{\mu_\beta}.$$

In a similar fashion we can compute the entries in Table 2.3.

2.5.2 A Second Example of Propagation of Errors: Sometimes it Matters!

Global warming is in the news. Some have expressed a concern that global warming may be weakening ocean circulation in the Atlantic Ocean that moderates the climate in Great Britain and Europe. A measure of the circulation that has been monitored for over 50 years is the circulation across the 25° N latitude. The circulation is measured in Sverdrups (1 Sv = 10⁶ m³/s). In 1957 the measured circulation was reported as 22.9 ± 6 Sverdrups, but more recently (2004) has been reported to be 14.8 ± 6 Sverdrups[4]. If one were to report the change as 8.1 ± 6 Sverdrups as certainly is tempting, one would conclude that the change has been significant and would so report.

Table 2.3: Propagation of Errors in Some Simple Cases.

f	σ_x^2
$x_i = an_i^\alpha \pm bn_i^\beta$	$\sigma_x^2 = a^2\sigma_\alpha^2 + b^2\sigma_\beta^2$
$x_i = \pm an_i^\alpha n_i^\beta$	$\frac{\sigma_x^2}{\mu_x^2} = \frac{\sigma_\alpha^2}{\mu_\alpha^2} + \frac{\sigma_\beta^2}{\mu_\beta^2}$
$x_i = \pm \frac{an_i^\alpha}{n_i^\beta}$	$\frac{\sigma_x^2}{\mu_x^2} = \frac{\sigma_\alpha^2}{\mu_\alpha^2} + \frac{\sigma_\beta^2}{\mu_\beta^2}$
$x_i = a(n_i^\alpha)^{\pm b}$	$\sigma_x^2 = b^2 \frac{\sigma_\alpha^2}{\mu_\alpha^2}$
$x_i = ae^{\pm n_i^\alpha}$	$\frac{\sigma_x^2}{\mu_x^2} = b^2 \sigma_\alpha^2$
$x_i = a \ln(\pm bn_i^\alpha)$	$\sigma_x^2 = \frac{a^2 \sigma_\alpha^2}{\mu_\alpha^2}$

But, if the two measurements were purely uncorrelated and the reported errors are purely statistical, the uncertainties in the two measurements should combine as (see Table 2.3 for a quantity computed from a difference),

$$\sigma^2 = \sigma_1^2 + \sigma_2^2,$$

so that our result should better be reported as 8.1 ± 8.5 Sverdrups. In the first (incorrect) estimate of errors, the effect would be judged to be significant, but in the latter (correct) way, the effect is judged to be within statistical uncertainty and, hence, not significant. So, it can matter how one handles the propagation of errors!

2.5.3 Estimating Propagation of Errors

Of course, these expressions for the variances are riddled with values of μ that are not known exactly. *If* the distribution were Gaussian, we can estimate $\mu \approx \mu_s$. Sometimes we only have one measurement in our sample, for which case we have to make do with $\mu \approx n$. In that case, for the first three entries in Table 2.3 we can estimate,

$$\sigma(an_\alpha \pm bn_\beta) \approx (a^2 n_\alpha + b^2 n_\beta)^{1/2}$$

$$\sigma(an_\alpha n_\beta) \approx a[n_\alpha n_\beta (n_\alpha + n_\beta)]^{1/2}$$

and,

$$\sigma\left(a\frac{n_\alpha}{n_\beta}\right) \approx a\left(\frac{n_\alpha}{n_\beta^2} + \frac{n_\alpha^2}{n_\beta^3}\right)^{1/2}.$$

EXAMPLE: Let's imagine a radioactive counting experiment. Let us imagine that the background count is n_β in time t_β . With the radioactive source present, we get n_α counts in time t_α . We imagine that t_α and t_β are precisely known and that $T = t_\alpha + t_\beta$. For the rates, we have $R_\alpha = n_\alpha/t_\alpha$ and $R_\beta = n_\beta/t_\beta$. Our quantity of interest is the net rate,

$$R = R_\alpha - R_\beta = \frac{n_\alpha}{t_\alpha} - \frac{n_\beta}{t_\beta}.$$

Identifying $a = t_\alpha^{-1}$ and $b = t_\beta^{-1}$, we use the first formula in Table 2.3 to obtain,

$$\sigma(R) = \left(\frac{n_\alpha}{t_\alpha^2} + \frac{n_\beta}{t_\beta^2}\right)^{1/2}.$$

We may then ask: What is the best way to apportion T so as to minimize the uncertainty in the answer. We have,

$$\sigma^2(R) = \frac{R_\alpha}{t_\alpha} + \frac{R_\beta}{T - t_\alpha}.$$

Taking the derivative of $\sigma^2(R)$ with respect to t_α and setting to zero, we obtain,

$$\frac{t_\alpha}{t_\beta} = \left(\frac{R_\alpha}{R_\beta}\right)^{1/2}.$$

Imagine that we measure the activity of a radioactive source, including a background, R_α . Let us imagine that the background rate R_β is roughly 10% of R_α . Imagine that we make a run of ten minutes with the source present. Effectively, we take a sample of 10 individual minutes from the population of possible minutes and obtain an average by dividing the total number of counts over that period by 10. Let us imagine that the result is 39541 counts per minute. See Table 2.1. We should therefore make a background run of 3.2 minutes to minimize the variance in the net activity that we are calculating for the source. Let us imagine getting 3571 counts per minute for the background. We have

$$R = 39541 - 3571 = 35970$$

and

$$\sigma = \left(\frac{39541}{10} + \frac{3571}{3.1} \right)^{1/2} = 71,$$

which is close to the standard variation $\sigma_s = 65$ of the sample distribution for the ten one-minute samples in Table 2.1.

EXAMPLE: We observe what appears to be a small effect that slightly changes the counts compared to background in a detector. Is the effect real? This is a very important question in research.

If the distribution of counts is a Gaussian one, we can turn to Table 2.2 which gives the probability that an observation of $n_i - \mu$ exceeds various multiples of σ . The probability of a measurement of $n_i - \mu$ exceeding 3σ is 0.0027, i.e. less than 1%. Equivalently, the probability that $n_i - \mu$ is less than 3σ is 99%. It is pretty safe that if we get a value of $n_i - \mu$ exceeding 3σ , it is probably (99%) real and not just a statistical aberration (1%).

Let n_α be the counts in condition α and n_β be the counts in condition β . Assume that both counts were take in the same length of time t . Then,

$$\text{Rate of possible effect} = x = \frac{|n_\alpha - n_\beta|}{t}.$$

From Table 2.3 ,

$$\begin{aligned} \sigma_x &= \left(\frac{1}{t^2} \sigma_\alpha^2 + \frac{1}{t^2} \sigma_\beta^2 \right)^{1/2} \\ &\approx \frac{1}{t} (n_\alpha + n_\beta)^{1/2}. \end{aligned}$$

The effect is probably real (99% confidence), if

$$\frac{|n_\alpha - n_\beta|}{t} \geq 3\sigma_x,$$

i.e.,

$$|n_\alpha - n_\beta| \geq 3(n_\alpha + n_\beta)^{1/2}.$$

For example, if you use the two extreme values from Table 2.1, $n_\alpha = 39204$ and $n_\beta = 39750$, the inequality is not satisfied and one could not with confidence attribute a real effect as a cause for the difference in counts.

2.5.4 Derivation: Estimating the Variance of the Mean

The variance of the sampling distribution σ_s represents the variance of the mean. In Eq.(2.1) we noted that for a sample size of N and a parent distribution with variance σ_p , the variance of the mean is given by

$$\sigma_s^2 = \frac{\sigma_p^2}{N}.$$

Here's a derivation of that important result.

When we calculate the mean $\langle x_i \rangle$ from a series of measurements x_i , each data point is obtained from a parent distribution with variance σ_i . From Eq. (2.2), we can think of that uncertainty in each point propagating through to become the uncertainty in the mean.

If $\sigma_i = \sigma_p$ is the same for all measurements,

$$\frac{\partial \mu}{\partial x_i} = \frac{\partial}{\partial x_i} \left[\frac{1}{N} \sum x_i \right] = \frac{1}{N}.$$

Thus,

$$\sigma_s^2 = \sum \left[\sigma_p^2 \left(\frac{1}{N} \right)^2 \right] = \frac{\sigma_p^2}{N}.$$

If the σ_i were unequal,

$$\frac{\partial \mu}{\partial x_i} = \frac{\partial}{\partial x_i} \frac{\sum(x_i/\sigma_i^2)}{\sum(1/\sigma_i^2)} = \frac{1/\sigma_i^2}{\sum(1/\sigma_i^2)}.$$

and,

$$\sigma_s^2 = \frac{1}{\sum(1/\sigma_i^2)}$$

2.5.5 Variation of the Mean: Trial of the Pyx

An ignorance of the relationship $\sigma_s^2 = \sigma_p^2/N$ may have cost the monarchs of England a ton of money[5]! The early kings of England began issuing money in the form of gold coins. The king and his barons provided gold to the London Mint, an independent organization, and the mint struck the coins. About the year 1150 the king and barons realized that the mass of the coins could not be controlled precisely and so they devised a method to keep the mint honest by establishing an allowable variability in the coins. Otherwise,

if the mint systematically produced “guineas” below the intended weight of 128 grains, it could illegally convert the difference in gold to its own benefit. But if the allowable variability were too generous, the mint could produce a surplus of coins that were too heavy, then later, collect these coins, restrike them, and keep the difference in gold. A lot hinged on the allowed variability.

The allowable deviation that was chosen for each coin was $1/400$ of its weight. For a guinea, that amounted to $\sigma_p = 0.32$ grains. A hundred coins were taken as the standard and placed for safe keeping in a wooden box, called a pyx. A run of 100 coins was then taken periodically from the mint and compared in the aggregate to this standard, thus obtaining a mean for a sample of $N = 100$. Since there were a hundred coins, the watchdogs took $\sigma_s = \sigma_p$ and allowed a variability of 100×0.32 grains in the weight of the 100 coins. The Trial of the Pyx was an exercise that was performed for 600 years with this allowable variability until 1730 when the French mathematician Abraham de Moivre showed that the standard deviation $\sigma_s = \sigma_p/\sqrt{N}$. The standard deviation for the mean should have been 0.032 grains rather than 0.32 grains and the variability for 100 coins should better have been taken to be 3.2 grains rather than 32 grains. We trust that the mint did not take advantage of this misunderstanding.

2.6 Estimating Other Parameters: The Maximum Likelihood Method

We have been thinking of probability distributions that are functions of some variable x as well as certain fixed parameters that we have denoted by θ . But if one makes a series of measurements x_i , the data become known “constants”. If we consider the parameters themselves as the unknowns, can we then turn things around and find the parameters?

Let us imagine that we take a series of N measurements of x (or n if the variable is discrete rather than continuous). Each x_i is taken from some parent probability distribution function that is a function of parameters, θ_k . Let the probability of each x_i be denoted by $\phi_i dx_i$. Then, assuming each measurement is independent of the others, the probability of the series of measurements is

$$P(x_1, x_2, \dots, x_N; \theta_k) dx_1 dx_2 \dots dx_N = \phi_1 \phi_2 \dots \phi_N dx_1 dx_2 \dots dx_N,$$

or, in the case of discrete probability distributions,

$$P(n_1, n_2, \dots, n_N; \theta_k) = \phi_1 \phi_2 \dots \phi_N.$$

In both cases $P(x_i)$ can also be thought of as a function of the parameters θ_k . Seen as a function of θ_k , we define (in both cases) the *likelihood function* $L(\theta_1, \theta_2, \dots, \theta_k)$,

$$L(\theta_1, \theta_2, \dots, \theta_k) = P(x_1, x_2, \dots, x_N; \theta_1, \theta_2, \dots, \theta_k) \geq 0,$$

whether x_i represents a set of measurements of a continuous or of a discrete variable.

One can certainly imagine that there would be possible sets of values of θ_k that would make the actual observation of the particular set of x_i improbably small. But, given that one did actually observe the set of x_i , what was the most *likely* set of the θ_k . This would almost certainly be the set of θ_k that makes the likelihood function take on its maximum possible value, i.e., using the basic mini-max approach of standard calculus, the set of θ_k that satisfy the equations,

$$\frac{\partial L}{\partial \theta_k} = 0, \quad k = 1, \dots, K.$$

There are no obvious guarantees that the set of equations actually has a unique solution, but insofar as it does, the so-called *maximum likelihood method* has a kind of intuitive simplicity to it. The flip side is that to make it work, you must know the individual probability distributions for the x_i to form the likelihood function and these may not be known.

2.6.1 Example: Estimating the Time Constant for Radioactive Decay

Imagine that we have a radioactive sample containing only a single radioactive substance that decays exponentially in time with a single decay constant λ . We observe the sample and register the times of a series of individual decays from the sample, $t_1, t_2, \dots, t_N \leq T$. The total time of the observation is T . The probability distribution function that governs the decays is Eq.(1.4)

$$P_\infty(t; \lambda) = \lambda e^{-\lambda t}$$

if the total time of observation is infinite. In our case, the observations only extend to T . To keep our probability distribution function properly normalized for this period of time,

$$\int_0^T P_T(t; \lambda) = 1$$

requires that we divide $P_\infty(t; \lambda)$ by a factor $(1 - e^{-\lambda T})$. With this minor correction, the likelihood function becomes

$$L(t_1, t_2, \dots, t_N; \lambda) = \left(\frac{\lambda}{1 - e^{-\lambda T}} \right)^N \prod_{i=1}^N e^{-\lambda t_i}.$$

Because of the product of factors, it is convenient to find the value of λ that maximizes $\ln L$, rather than L itself, by satisfying,

$$\begin{aligned} 0 &= \frac{\partial \ln L(t_i; \lambda)}{\partial \lambda} \\ &= \frac{\partial}{\partial \lambda} \left(N \ln \lambda - N \ln(1 - e^{-\lambda T}) - \sum_{i=1}^N \lambda t_i \right), \end{aligned}$$

for which,

$$\sum_{i=1}^N t_i + \frac{NT}{e^{\lambda T} - 1} - \frac{N}{\lambda} = 0.$$

The equation requires a numerical or iterative solution in general, but simplifies if $e^{\lambda T} \gg 1$ to yield

$$\frac{1}{\lambda} \equiv \tau = \frac{\sum t_i}{N}.$$

The likely ‘‘time constant’’ for the decay is just the average of the set of times that form the data set for the experiment.

There is certainly some uncertainty associated with this value of τ . Can we make an estimate of the uncertainty in λ ?

We could if the likelihood function were Gaussian as a function of λ ,

$$L(\lambda) = L_0 e^{-(\lambda - \lambda_0)^2 / 2\sigma^2}.$$

Then,

$$\ln L(\lambda) = \ln L_0 - \frac{(\lambda - \lambda_0)^2}{2\sigma^2},$$

and,

$$-\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2} = \frac{1}{\sigma^2},$$

i.e.,

$$\sigma = \left[-\frac{\partial^2 \ln L(t_i; \lambda)}{\partial \lambda^2} \right]^{-1/2}.$$

If the likelihood function is not known to be Gaussian, the expression becomes either an estimate of σ or a “figure of merit.”

2.7 Finding the Parent Distribution: Chi-Squared

Having made estimates of mean and standard deviation from a sample of measurements x_i , we can sometimes use these parameters to construct the parent distribution *if* we know what basic form the underlying parent distribution is expected to take. But what if we don’t know the parent distribution and have to conjecture what it is? What confidence do we have that our conjecture is a good one?

If we make enough measurements (truly a lot), we could simply keep count of the frequency with which we obtain each value x_i . We could plot these frequencies versus the possibilities and look at the result. If we have taken enough measurements, we should see the form of the underlying parent distribution displayed in the plot. But if we have only a small sample of measurements, the “visual method” is less convincing.

To illustrate the problem, consider throwing a pair of dice and observing the outcome. The possibilities for each throw of two dice are “2”, . . . , “12”. The parent distribution for the outcomes for honest dice is shown as the discrete possibilities on the solid line in Fig. 2.1. A “7” is most probable and “2” or “12” are equally least probable. Also shown in the figure as asterisks are the number of times that each possibility actually occurred in a series of 108 throws. The results of the series are also recorded as Series 1 in Table 2.4.

Focusing on how many times a “3” came up on the dice, we see that it happened eight times, but “11” came up twice, even though a “3” and an “11” have equal probabilities. What confidence do we have that the parent distribution that we have drawn is the one that is responsible for

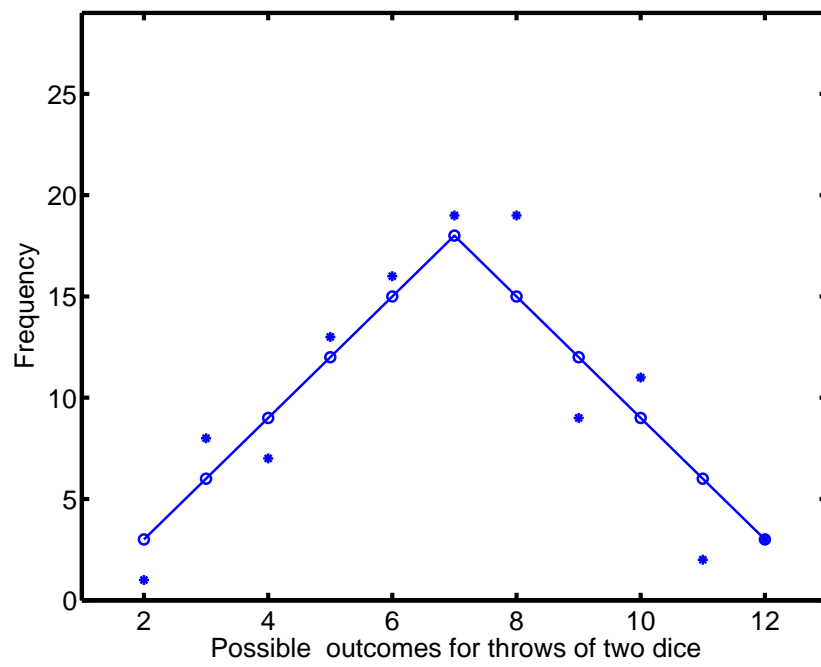


Figure 2.1: A Single Series of 108 Dice Throws

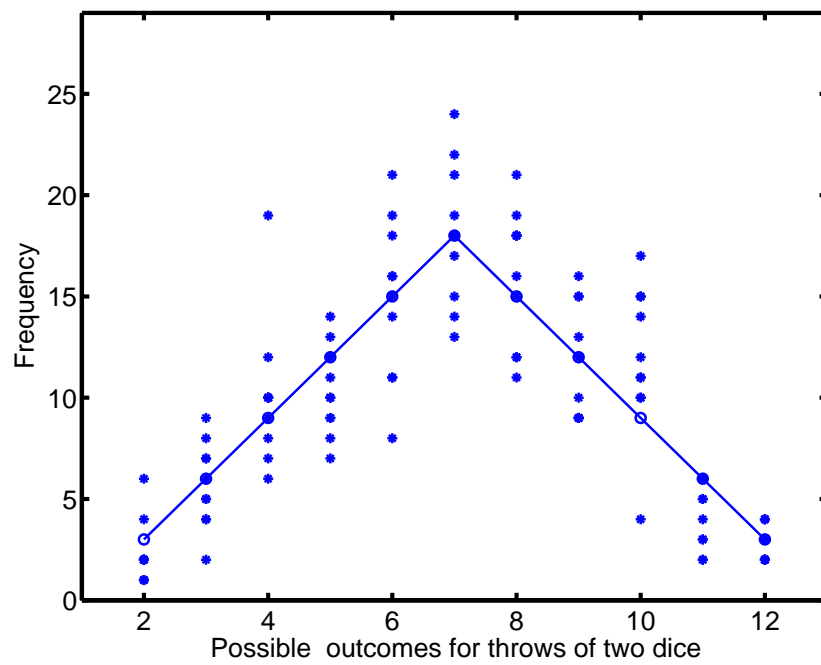


Figure 2.2: Ten Series of Dice Throws

Table 2.4: Outcomes for Ten Series of 108 Dicethrows

	2	3	4	5	6	7	8	9	10	11	12
Series 1	1	8	7	13	16	19	19	9	11	2	3
Series 2	4	7	10	10	11	22	18	9	10	3	4
Series 3	2	5	12	9	21	15	12	9	15	5	3
Series 4	1	9	9	7	19	17	15	15	12	2	2
Series 5	1	4	19	10	11	18	18	10	11	4	2
Series 6	2	2	6	9	14	24	12	15	14	6	4
Series 7	2	5	10	11	16	13	16	13	15	5	2
Series 8	2	7	10	12	15	21	11	12	10	6	2
Series 9	6	6	8	8	8	18	18	9	17	6	4
Series 10	2	4	10	14	18	14	21	16	4	3	2
Expected	3	6	9	12	15	18	15	12	9	6	3

the outcomes in our series of 108 throws? Table 2.4 shows the outcomes for ten series of 108 throws. None are identical. However, it is plausible, but not certain, that the values shown by the asterisks originate from the parent probability distribution given. What confidence can we derive from the actual data points themselves?

Let us denote the frequency of each observed outcome as $f_k(n_i)$, where $n_i = \text{"2"}, \dots, \text{"12"}$. The index k keeps track of the particular series. Let $P(n_i)$ be the parent distribution. Then, the expected frequency for each outcome in a series of N throws is $NP(n_i)$. If the probability $P(n_3)$ of a "3" is 0.055556, we expect a "3" to come up six times in 108 throws. In fact, a "3" came up six times in only one of the ten series. In Fig. 2.2, we see the frequencies from each of the ten series plotted against the possible outcomes.

The frequencies themselves, therefore, are random variables, varying from series to series. The probability of a "3" is $p = 0.055556$ and the probability of NOT "3" is $q = 1 - p$. The frequencies for a given outcome satisfy the conditions that should result in a Poisson distribution. If we divide the number of times each frequency is obtained for each value of n_i by the total number of instances (the ten series), our frequencies become percentages (probabilities).

For example, for n_3 , $f_3(n_3)$ and $f_7(n_3)$ are both equal to 5 in Table 2.4.

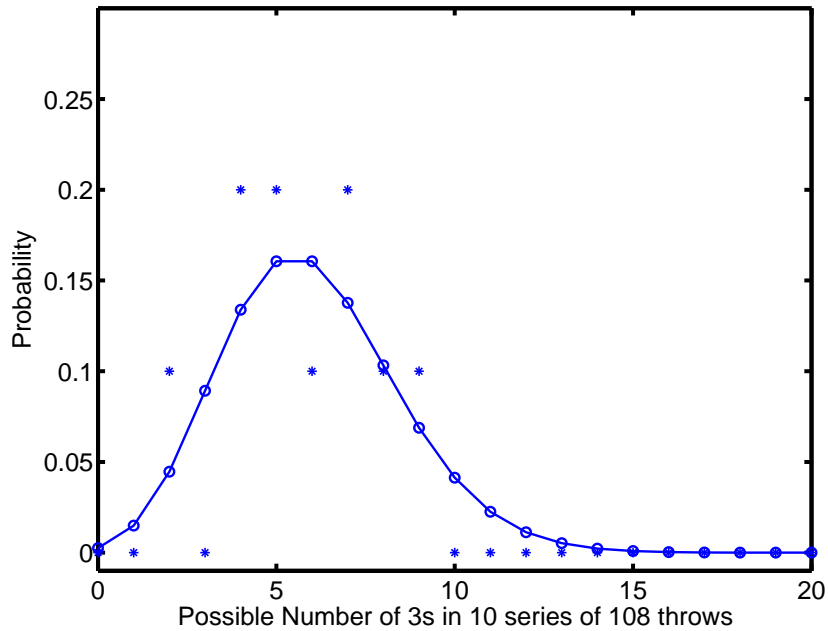


Figure 2.3: Frequency of 3s for 10 Series of Throws

In ten series, the frequency of getting a “3” exactly 5 times was 20% of the times. (See Fig. 2.3.) From the Poisson distribution, the actual probability of getting five “3”s when the actual expectation is six is ≈ 0.16 .

In Fig. 2.3 we show the percentages of the frequencies of “3”s for 10 series of 108 throws (shown as asterisks) compared to the Poisson probability distribution with μ taken to be the expected value of six. It is plausible at least that the asterisks are taken from the Poisson distribution, but one would have to run many more than ten series to use the “visual method” and be convinced.

2.7.1 χ^2

Building on our experience from the example of the dice, let us imagine that we only have one series ($k = 1$) of N measurements. Let us imagine that there are M possibilities n_i that arise in our measurements and that these are taken from a parent distribution that we can only conjecture to be $P(n_i)$. In the series of measurements, the possibilities are observed to recur with frequencies $f_1(n_i)$. In the example of the dice, $M = 11$ and $N = 108$.

Although we are imagining that we do not know the parent distribution with confidence, we do know that the distribution of frequencies should be a Poisson distribution, just as it was for the dice. For a Poisson distribution, the variance σ^2 equals the mean μ . For a given value of one of the possibilities n_i , the mean of the Poisson distribution of frequencies should be approximately $NP(n_i)$. In the example of the dice, $N = 108$ and the probability of a “3” was 0.055556, to give $NP(n_3) = 6$ which is approximately where the Poisson distribution in Fig. 2.3 has its peak. Each frequency will have a different value for $NP(n_i)$.

With this in mind, define a function χ^2 as follows,

$$\begin{aligned}\chi^2 &\equiv \sum_{i=1}^M \frac{[f_1(n_i) - NP(n_i)]^2}{\sigma^2(f)} \approx \sum_{i=1}^M \frac{[f_1(n_i) - NP(n_i)]^2}{NP(n_i)} \\ &= \sum_{i=1}^M \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}\end{aligned}$$

What might we expect χ^2 to be if our conjectured parent distribution were about right? The value of χ^2 vanishes if the observed frequencies $f_1(n_i)$ exactly equal the expected values $NP(n_i)$. But that is too much to hope for in actual observation of a random variable. What we really expect is that each difference $[f_1(n_i) - NP(n_i)]$ should lie within about one standard deviation of the expected value, so that the fraction,

$$\frac{[f_1(n_i) - NP(n_i)]^2}{\sigma^2(f)}$$

should be about 1 in each term and the sum of 1s should be M if the parent probability distribution were the correct one.

If we were to divide our χ^2 by M , we expect to have a value of about 1. However, in practice we divide by a number ν called the *number of degrees of freedom*. The conclusion that the sum be M hinges on the assumption that every term in the sum comes from completely equivalent and independent observations. If you think of Poisson and Gaussian distributions as examples of possible parent distributions, you will realize that they are really families of distributions, each differing from each other by mean μ and standard deviation σ . In order to compute χ^2 , we must choose the parent probability distribution to which we want to make comparison. We can only decide from

Table 2.5: Chi-Squared for Ten Series of 108 Dicethrows

Series	1	2	3	4	5	6	7	8	9	10
χ_r^2	0.76	0.62	1.07	1.08	1.69	1.13	0.68	0.26	1.65	1.13

the data that we actually have in hand. For each distribution parameter (μ , σ , etc.) or other constraint determined from the data, we lose some degree of independence in the terms of the sum. In a sense, each time we add a parameter computed from the data into the assumed probability distribution, we lose independence of one term in our sum. The number of “degrees of freedom” is taken to be the number of terms in our sum minus the number of independent parameters n_p computed from the data, $\nu = M - n_p$. We then define a “reduced” χ_r^2 ,

$$\chi_r^2 \equiv \frac{1}{\nu} \sum_{i=1}^M \frac{[f_1(n_i) - NP(n_i)]^2}{\sigma^2(f)}$$

whose value we expect to be about 1 for a “good” choice of the parent probability distribution.

The test is not absolutely conclusive. It does not allow one to accept or reject an assumed parent probability distribution in a completely objective way. But the further χ_r^2 drifts from a value of 1.0, either too small or too large, the less confidence one has that the data derive from the conjectured parent probability distribution.

Are the values in Table 2.4 plausible representations of the parent probability distribution shown in Fig. 2.1? In Table 2.5 we show the values of χ_r^2 for each of the ten series of our dice throwing example. There is one constraint among the outcomes, namely that $\sum n_i = 108$, so that the number of degrees of freedom is the number of possible outcomes (11) minus 1, i.e., $\nu = 10$.

There does seem to be a tendency for values near 1.0, but there are a few questionable outliers as well. We therefore take as a

Rule of Thumb

1. For χ^2 tests,

$$\chi_r^2 = \frac{1}{\nu} \sum_{i=1}^M \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}} \approx 1.$$

2.7.2 χ^2 as a Random Variable

From the summary of our dice throwing example in Table 2.5, we conclude that χ_r^2 is itself a random variable. Is there an accessible parent probability distribution for χ^2 itself that would allow us to draw a conclusion about the probability of particular values of χ_r^2 ? The derivation of such a distribution goes beyond our current scope, but statistics textbooks tell us that if the parent probability distribution is *Gaussian*, then

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

and, probably more importantly, its associated cumulative function is,

$$P_c(\chi^2, \nu) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \int_{-\infty}^{\chi^2} (x^2)^{(\nu-2)/2} e^{-x^2/2} dx^2.$$

Here the $\Gamma(\nu)$ function is defined for integer or half-integer arguments as,

$$\Gamma(1) = 1, \quad \Gamma(1/2) = \sqrt{\pi}, \quad \Gamma(\nu + 1) = \nu \Gamma(\nu).$$

Thus, for half-integer values $\nu/2$,

$$\Gamma(\nu + 1) = \nu(\nu - 1) \dots (3/2)(\sqrt{\pi}/2),$$

and for integer values of $\nu/2$,

$$\Gamma(\nu + 1) = \nu!.$$

Figure 2.4 shows a chi-squared distribution for $\nu = 10$.

The cumulative function is particularly important because it tells us the probability that, for a given number of degrees of freedom, we should get a value of χ^2 that is equal to or smaller than the one we have, again assuming that the underlying parent probability distribution of the data is Gaussian. Any computer with a statistical package of consequence will return values for these functions if asked with the proper syntax!

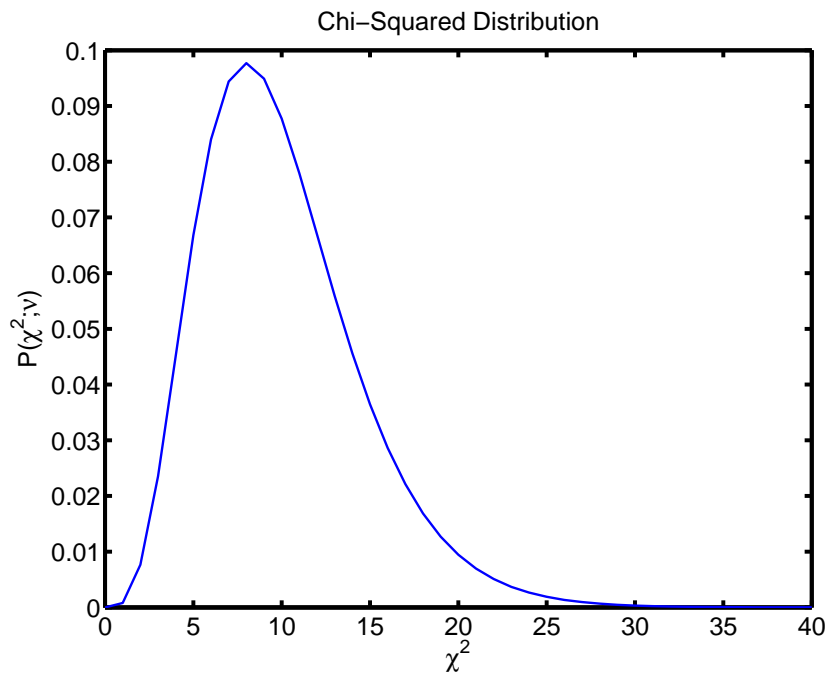


Figure 2.4: The chi squared distribution with $\nu = 10$. Observe that the distribution peaks at $\chi^2 = \nu$. The cumulative function $P_c(\chi^2, \nu)$ is the area under the curve from $-\infty$ to a chosen value of χ^2 and represents the cumulative probability of obtaining a χ^2 at or less than the value chosen.

For example, if we are willing to accept that the tent-shaped probability distribution underlying the dice throws approximates a Gaussian distribution, we can estimate probabilities associated with the values of χ^2 that we obtained for the different series of dice rolls. We obtained values of χ_r^2 exceeding 1.5 for two of the series. What is the probability that should happen? For ten degrees of freedom, the value of χ^2 (not reduced) that corresponds to $\chi_r^2 = 1.5$ is 15.0. Asking our computer for the value of $1.0 - P_c(15.0, 10)$ we get 0.1321. Thus, we expect to get a value of χ_r^2 exceeding 1.5 in 13% of the series. It actually happened twice (20% of the series).

What is the probability of getting our smallest value of 0.26 or less? The computer tells us $P_c(2.6, 10) = 0.0107$. There is about a one percent chance of getting $\chi_r^2 < 0.26 \dots$ and we got one (10% of the series)! This should give us some respect for the tricks that data can play on us if we are not careful!

What is the probability of getting $0.5 < \chi_r^2 < 1.5$? $P_c(15.0, 10) - P_c(5.0, 10) = 0.76$. We actually obtained χ_r^2 in this range in 7 out of 10 series (70%).

2.8 Least Squares Fitting: Minimizing χ^2

It may have occurred to you that there is another way to play the game with χ^2 . If we just knew the family of the parent probability distribution, we could find out what choice of parameters minimizes χ^2 and, hence, the particular member of the family that seems to best describe the data!

Even more generally, we may want to simply find a functional form to “fit the data.” Imagine that we have a set of N data points $y_i(x_i)$ to which we have ascribed uncertainties σ_i . Let $P(x; \theta_k)$ be a “family” of functions characterized by the set of parameters θ_k . A parent probability distribution could be such a function. Then,

$$\chi^2 \equiv \sum_{i=1}^N \frac{[y_i(x_i) - P(x_i; \theta_k)]^2}{\sigma_i^2}$$

becomes a function of θ_k . Because we are essentially trying to minimize the numerators in this expression, the method is called *least squares fitting*.

We could find the values of the parameters that minimize χ^2 by solving the set of equations,

$$\frac{\partial \chi^2}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{i=1}^N \frac{[y_i(x_i) - P(x_i; \theta_k)]^2}{\sigma_i^2} = 0.$$

In general, the equations thus derived do not lend themselves to analytic solution except in simple cases. In practice, numerical methods are used to find the set of θ_k by searching for minima of χ^2 in the multi-dimensional hyperspace of θ_k . Practitioners of this art often have their own customized algorithms and tricks for seeking out the solutions in the least amount of computer time.

The algorithms have to deal with local minima that may appear falsely to the unwary computer code to be the sought after *absolute* minimum. The codes also have to guard against finding solutions θ_k that are simply ruled out by the physical situation. In some cases where even the family of probability functions is not known, one must try different functional forms for $P(x)$ and compare values of χ^2 for the separate cases.

2.8.1 Example: Least Squares Fitting to a Straight Line

Fitting to a straight line or a simple polynomial can be done without resorting to searching in the hyperspace of parameters for a minimum. The exercise is a bit more general than one might immediately suppose when one realizes that if

$$n = n_0 e^{-\lambda t},$$

or,

$$n = n_0 t^\gamma,$$

then,

$$\ln n = -\lambda t + \ln n_0,$$

and,

$$\ln n = \gamma \ln t + \ln n_0$$

are straight line forms.

Let us imagine that we count radioactive decays of some source. We make ten measurements in ten one-second time periods that are continuous. We plot the results at the midpoint of each one-second time interval, i.e. at 0.5 sec, 1.5 sec, 2.5 sec, etc. Data for the experiment are shown in Table 2.6.

Over this short time span the number of counts appear to trend linearly downward with time (rather than exponentially), but the data points themselves do not fit nicely on a single straight line. What, then, is the *best* straight line $n = mt + n_0$ that represents these data? We will consider the

Table 2.6: Decay of a Radioactive Source

Time (sec), t_i	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Counts, n_i	33	27	20	22	24	15	14	18	15	11
Uncertainty, σ_i	5.7	5.2	4.5	4.7	4.9	3.9	3.7	4.2	3.9	3.3

data to be given and treat the slope m and intercept n_0 as unknowns that are chosen to minimize χ^2 .

Let us begin by defining for this assumed functional dependence,

$$\chi^2 = \sum \frac{(n_i - mt_i - n_0)^2}{\sigma_i^2}.$$

We will need estimates for σ_i . We can reasonably assume that the parent distribution for the number of counts in each time subinterval is a Poisson distribution since the probability of observing n_i is some probability p and for NOT observing n_i it is $1 - p$. That being the case, with only a single count taken in each time subinterval, we will have to do the best that we can by taking $\mu \approx n_i$ and $\sigma_i = \mu^{1/2} \approx n_i^{1/2}$. These are the values shown in Table 2.6.

To minimize χ^2 , we solve the equations,

$$\frac{\partial \chi^2}{\partial n_0} = -2 \sum \frac{(n_i - mt_i - n_0)}{\sigma_i^2} = 0$$

$$\frac{\partial \chi^2}{\partial m} = -2 \sum \frac{t_i(n_i - mt_i - n_0)}{\sigma_i^2} = 0$$

which can be rearranged into the forms,

$$n_0 \sum \frac{1}{\sigma_i^2} + m \sum \frac{t_i}{\sigma_i^2} = \sum \frac{n_i}{\sigma_i^2}$$

$$n_0 \sum \frac{t_i}{\sigma_i^2} + m \sum \frac{t_i^2}{\sigma_i^2} = \sum \frac{n_i t_i}{\sigma_i^2}.$$

These are two linear equations that have a solution for n_0 and m if the determinant of the coefficients Δ does not vanish,

$$\Delta = \begin{vmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_j \frac{t_j}{\sigma_j^2} \\ \sum_i \frac{t_i}{\sigma_i^2} & \sum_j \frac{t_j^2}{\sigma_j^2} \end{vmatrix} \neq 0.$$

The solutions are,

$$n_0 = \frac{1}{\Delta} \begin{vmatrix} \sum_i \frac{n_i}{\sigma_i^2} & \sum_j \frac{t_j}{\sigma_j^2} \\ \sum_i \frac{n_i t_i}{\sigma_i^2} & \sum_j \frac{t_j^2}{\sigma_j^2} \end{vmatrix} = \left(\left(\sum_i \frac{n_i}{\sigma_i^2} \right) \left(\sum_j \frac{t_j^2}{\sigma_j^2} \right) - \left(\sum_i \frac{n_i t_i}{\sigma_i^2} \right) \left(\sum_j \frac{t_j}{\sigma_j^2} \right) \right) / \Delta \quad (2.3)$$

$$m = \frac{1}{\Delta} \begin{vmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_j \frac{n_j}{\sigma_j^2} \\ \sum_i \frac{t_i}{\sigma_i^2} & \sum_j \frac{n_j t_j}{\sigma_j^2} \end{vmatrix} = \left(\left(\sum_i \frac{1}{\sigma_i^2} \right) \left(\sum_j \frac{n_j t_j}{\sigma_j^2} \right) - \left(\sum_i \frac{t_i}{\sigma_i^2} \right) \left(\sum_j \frac{n_j}{\sigma_j^2} \right) \right) / \Delta. \quad (2.4)$$

Given our data, these are the values of the parameters n_0 and m that minimize χ^2 and are presumably those that give the best linear fit to our data. Figure 2.5 shows the data points plotted with their most probable error flags and the subsequent least squares fit line. The minimum value of χ_r^2 is 0.55 for eight degrees of freedom (ten data points less two parameters).

2.8.2 Estimating Uncertainties for the Slope and Intercept of a Linear Fit

How have the uncertainties in the values of the individual data points propagated into the values of n_0 and m ?

If we think of n_0 and m as functions of n_k , we can use Eq. (2.2),

$$\sigma_{n_0}^2 \approx \sum_k \sigma_k^2 \left(\frac{\partial n_0}{\partial n_k} \right)_{\mu^k}^2$$

$$\sigma_m^2 \approx \sum_k \sigma_k^2 \left(\frac{\partial m}{\partial n_k} \right)_{\mu^k}^2.$$

Using Eq. (2.3) and Eq. (2.4) we get,

$$\sigma_{n_0}^2 \approx \frac{1}{\Delta^2} \sum_k \frac{1}{\sigma_k^2} \left(\sum_j \frac{t_j^2}{\sigma_j^2} - t_k \sum_j \frac{t_j}{\sigma_j^2} \right)^2$$

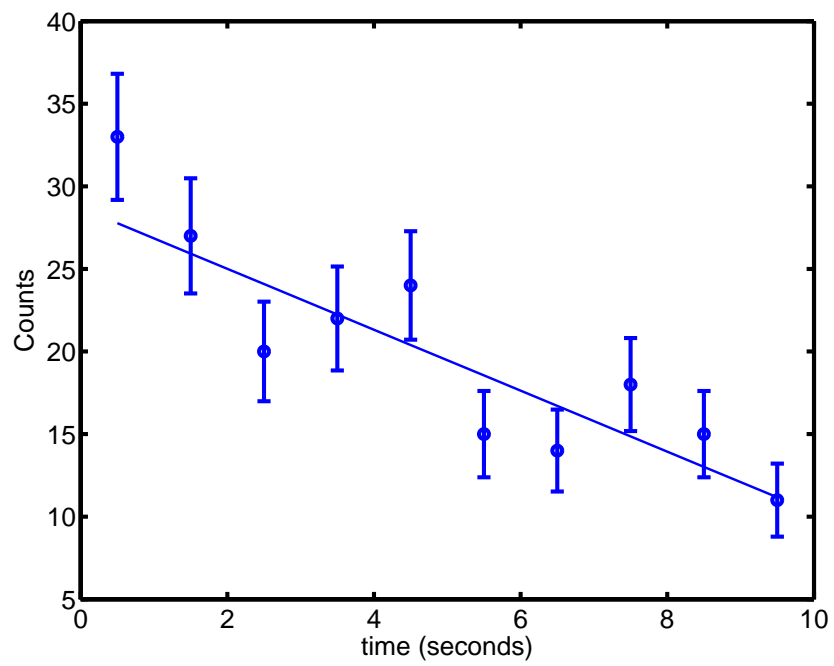


Figure 2.5: Linear decay of a radioactive source. χ^2 is minimized to yield the “best” linear fit to the data.

$$\sigma_m^2 \approx \frac{1}{\Delta^2} \sum_k \frac{1}{\sigma_k^2} \left(t_k \sum_j \frac{1}{\sigma_j^2} - \sum_j \frac{t_j}{\sigma_j^2} \right)^2.$$

In particular,

$$\sigma_{n0}^2 \approx \frac{1}{\Delta^2} \sum_j \sum_k \sum_\ell \frac{1}{\sigma_j^2} \frac{1}{\sigma_k^2} \frac{1}{\sigma_\ell^2} (t_j^2 t_\ell^2 - t_j t_k t_\ell^2 - t_j^2 t_k t_\ell + t_j t_k^2 t_\ell).$$

The indices j , k , and ℓ are dummy indices and can be exchanged in any given term, so,

$$\begin{aligned} \sigma_{n0}^2 &\approx \frac{1}{\Delta^2} \sum_j \sum_k \sum_\ell \frac{1}{\sigma_j^2} \frac{1}{\sigma_k^2} \frac{1}{\sigma_\ell^2} (t_j^2 t_\ell^2 - t_j t_k t_\ell^2) \\ &= \frac{1}{\Delta^2} \sum_\ell \frac{1}{\sigma_\ell^2} (\Delta) t_\ell^2 \\ &= \frac{1}{\Delta} \sum_\ell \frac{t_\ell^2}{\sigma_\ell^2}. \end{aligned}$$

Similarly,

$$\sigma_m^2 \approx \frac{1}{\Delta} \sum_\ell \frac{1}{\sigma_\ell^2}.$$

If the uncertainties in the data points are instrumental, some estimate must be made of the σ_ℓ s based on the idiosyncracies of the particular measuring device and protocol for the measurements. However, if the uncertainties are purely statistical, we can go one step further and estimate that $\sigma_i^2 \approx n_i$. In this latter case,

$$\begin{aligned} \sigma_{n0} &\approx \frac{1}{\Delta} \sum_\ell \frac{t_\ell^2}{n_\ell} \\ \sigma_m^2 &\approx \frac{1}{\Delta} \sum_\ell \frac{1}{n_\ell}. \end{aligned}$$

Once we have the values of $\sigma_{n0} = 3.08$ and $\sigma_m = 0.48$ for the data of Table 2.6, we multiply each by 0.67 to get the probable errors. We would then report the results as $n_0 = 28.7 \pm 2.08$ and $m = -1.84 \pm 0.32$.

Bibliography

- [1] Philip R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, (McGraw-Hill Book Company, New York, 1969).
- [2] Niels Arley and K. Rander Buch, *Introduction to the Theory of Probability and Statistics*, (John Wiley & Sons, Inc., New York, 1950).
- [3] Emilio Segre, *Nuclei and Particles*, (W. A. Benjamin, Inc., New York, 1964).
- [4] H. L. Bryden, H. R. Longworth, S. A. Cunningham, *Nature* **438**, 655 (2005).
- [5] Howard Wainer, *American Scientist*, **95**, 249 (2007).

Appendix A

Rules of Thumb

1. Law of Large Numbers (Bernoulli's Law): When a large number of independent events is taken from a parent population, the sample's average will be close to the true mean of all events, including those that were not observed.
2. Central Limit Theorem: Assume that a variable x has a mean μ and variance σ^2 . If σ^2 is finite, then the distribution of the sample mean \bar{x} approaches a Gaussian distribution with mean μ and variance σ^2/N as N tends to infinity.
3. In the case of a single measurement in a counting experiment, $\mu \approx \mu_x = n_1$.
4. In the case of a single measurement in a counting experiment for which the parent distribution is approximately Poisson or Gaussian, $\sigma_p \approx \mu_s^{1/2} = n_1^{1/2}$.
5. Vertical bars are often drawn attached to displayed data points extending $\pm 0.67\sigma$ to indicate an estimate of the range from the mean within which there is a 50% probability of a measured random variable falling if taken from a distribution that is approximately Gaussian.
6. If taken from a distribution that is approximately Gaussian, it is estimated that there is less than a 1% probability of the measured value falling outside $\pm 3\sigma$ from the measured value (which is taken as an estimate of the mean.)

7. Propagation of Errors:

$$\sigma_x^2 \approx \sum_{\alpha} \sigma_{\alpha}^2 \left(\frac{\partial f}{\partial n^{\alpha}} \right)_{\mu_{\alpha}}^2$$

8. For χ^2 tests,

$$\chi_r^2 = \frac{1}{\nu} \sum_{i=1}^M \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}} \approx 1.$$