

Statistics—A Survival Guide

Michael A. Marsh
Institute for Advanced Computer Studies
University of Maryland
mmarsh@umiacs.umd.edu

August 4, 2005

Contents

1	Introduction	1
2	Counting Statistics	2
2.1	Binomial Distribution	2
2.1.1	Expectation Values and Statistical Measures	2
2.2	Poisson Distribution	4
2.3	Gaussian Distribution	6
3	Repeated Measurements	8
3.1	Probability Distribution Functions	10
3.2	Fitting and Binned vs. Unbinned Data	11
3.2.1	Maximum Likelihood Fitting	11
3.2.2	Least Squares Fitting	15
4	Measurement Uncertainties	16
4.1	Statistical Uncertainty	16
4.2	Systematic Uncertainty	17
4.3	Propagation of Uncertainties	18
4.3.1	Uncorrelated	18
4.3.2	Correlated	19
5	Data Selection and the Pitfalls of Tuning	20
6	Additional References	21
7	Acknowledgments	21

1 Introduction

This is not intended to be a long formal treatise on statistics. Rather, it’s designed to provide a general overview of some of the basic methods of statistical analysis that should be useful for many practical applications. We’ll see where uncertainties arise in experiments and how to deal with them when reporting results.

Often, the term “error” will be used when what’s really meant is “uncertainty.” In part, this reflects the fact that, when propagated into a result, we’re stating to what extent we think we might be in error. Really, though, it’s because “error” has half the syllables.

I’ve tried to provide fairly complete mathematical derivations. For the most part, you can skip these and only consider the numbered equations.

2 Counting Statistics

Most of the time, when measuring something in an experiment, we're really counting. This is because we're generally sorting measurements into *bins*, or collections of values that are between two specific values. For example, if we're timing a series of events, we might round the times to the nearest second, so that the bin centered at 5 seconds really represents the range $[4.5, 5.5)$ seconds. Consequently, it is usually counting statistics that plays the largest role in our statistical analyses.

For the moment, we're going to forget about binning, and only consider measurements in which we count a single type of event. We'll return to the subject of bins and measurements taken from a range of values later.

2.1 Binomial Distribution

Let's say we're trying to count the number of times we see something occur. For instance, we might be counting the number of corrupted packets on a noisy link.¹ For any packet travelling on a particular link, there's some probability p that the packet will be corrupted due to line noise. A process for which each observation has the same independent probability p of generating an event is called a *Bernoulli process*. If one hundred packets have been transmitted, we'd expect to see, on average, $100p$ corrupted packets.

The average value will often be fractional, and consequently we would never be able to observe this number of events in any actual experiment. More generally, we don't expect to see a number of corrupted packets equal to the average, we expect to see a number of corrupted packets *near* the average. If we see d corrupted packets in total, and we observe a total number of packets n ($n = 100$ in our example), then the probability of the specific pattern of corrupted packets is $p^d \cdot (1 - p)^{n-d}$. There are, however, a potentially large number of corruption patterns that will result in d corrupted packets. In fact, this number is $\binom{n}{d} = \frac{n!}{d!(n-d)!}$. Each is equally likely, so the probability of observing *some* d corrupted packets given n opportunities is

$$P[d; n, p] = \binom{n}{d} p^d \cdot (1 - p)^{n-d}. \quad (1)$$

The maximum of this probability over possible measurements occurs at (if np is an integer) or near (if not) the average.

Note that the total probability of observing between 0 and n corrupted packets is

$$\sum_{d=0}^n \binom{n}{d} p^d \cdot (1 - p)^{n-d} = (p + (1 - p))^n = 1.$$

Because the probability of observing a particular number of events is just a term in the binomial expansion of $(p + (1 - p))^n$, we call the probability distribution in (1) the *binomial distribution*. Note that any probability distribution must have the property that the sum (or integral) over all possible outcomes is 1, since *some* outcome will occur. We refer to such a distribution as being *normalized*, and any normalized distribution can in principle serve as a probability distribution.

2.1.1 Expectation Values and Statistical Measures

We don't observe probability distributions, we observe measurements or events. Often we will make several observations, and with enough of them the probability distribution should become evident. The connection between a set of observations and an underlying distribution is typically made using *statistics*. Each statistic characterizes some aspect of the distribution. Two common statistics are *mean* and *variance*, which describe the average observed value and the degree to which the data are spread about the average. One advantage of using these statistics rather than attempting to recover the probability distribution is that both mean and variance characterize a wide range of common distributions, and knowing these statistics is often as valuable as knowing the distribution itself.

The mean and variance of a probability distribution are calculable properties, and are defined as *expectation values* of the distribution. For a given distribution, we'll write the expectation of a function $f[d]$ as $E[f[d]]$, which for a

¹Here we assume a constant source of noise that corrupts packets independently and with the same probability. We say that the packet corruption events are *independent identically distributed*.

(normalized) discrete distribution $P[d]$ we define as

$$E[f[d]] = \sum_d P[d] \cdot f[d]. \quad (2)$$

The sum is performed over all possible values of d . For example, the binomial distribution is only defined for values of d in the range $[0, n]$. In terms of expectation values, the mean of the distribution is

$$\mu = E[d]. \quad (3)$$

The other quantity we will use is the *variance* σ^2 , defined as the expected distance from the mean *in quadrature*:

$$\sigma^2 = E[(d - E[d])^2] = E[(d - \mu)^2]. \quad (4)$$

The reason for computing the expectation in quadrature is that if we take the expectation of $d - E[d]$, we'll end up with $E[d] - E[d] = 0$, so we need a function $f[d]$ where values below the mean do not cancel out values above the mean. We could use other functions, but this particular definition is the most useful, as it closely relates to a natural definition of the *width* of a distribution. In particular, we often use the *standard deviation* σ as the width of a distribution, which is simply the square root of the variance. It is also worth noting that

$$E[(d - E[d])^2] = E[d^2] - (E[d])^2 = E[d^2] - \mu^2,$$

which is often easier to calculate. For continuous distributions, the mean and variance are defined identically, but the expectation becomes an integral rather than a sum:

$$\sigma^2 = E[f[x]] = \int P[x] \cdot f[x] \cdot dx. \quad (5)$$

Why are the mean and variance interesting statistics? Many probability distributions have a single “peak” with rapidly decreasing “tails” to either side. The mean of the distribution is often in this central peak, at or near the point where the probability distribution reaches its maximum. Consequently, the mean is often a good measure of the *most likely* observation. The square-root of the variance, the standard deviation, gives the size of the peaked region; the majority of observations are expected to fall within one standard deviation of the mean for many distributions. Thus the standard deviation is often a good measure of the *uncertainty* in an observation, since it's likely that a single observation is no more than one standard deviation from the actual mean of the distribution. Note that there are also uncertainties in the mean and standard deviation (or variance) that are dependent on the number of observations made rather than on the underlying probability distribution.

Let's calculate the mean and variance for the binomial distribution. First, the mean

$$\begin{aligned} \mu = E[d] &= \sum_{d=0}^n \binom{n}{d} p^d (1-p)^{n-d} d \\ &= \sum_{d=1}^n \binom{n}{d} p^d (1-p)^{n-d} d && \text{the } d=0 \text{ term vanishes} \\ &= \sum_{d=1}^n n \binom{n-1}{d-1} p^d (1-p)^{n-d} && \binom{n}{d} = \frac{n}{d} \binom{n-1}{d-1} \\ &= \sum_{d=1}^n np \binom{n-1}{d-1} p^{d-1} (1-p)^{(n-1)-(d-1)} && p^d = pp^{d-1} \\ &= np \sum_{c=0}^{m} \binom{m}{c} p^c (1-p)^{m-c} && c = d-1, m = n-1 \end{aligned}$$

Note that the last term is just the binomial expansion of $(p + (1-p))^m$, which is equal to 1, so

$$\mu = E[d] = np. \quad (6)$$

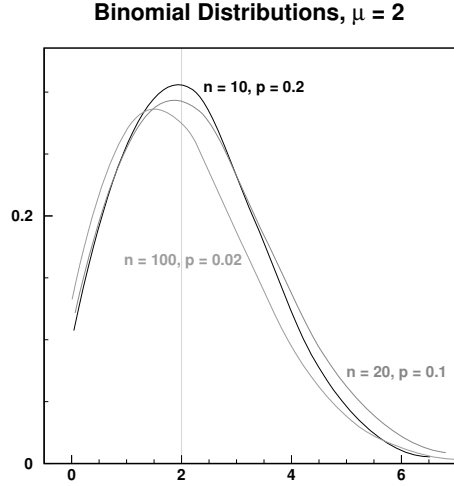


Figure 1: Binomial distributions with mean 2 for varying numbers of measurements and probabilities.

To calculate the variance, we'll use the simpler form that only requires calculating $E[d^2]$:

$$\begin{aligned}
 E[d^2] &= \sum_{d=0}^n \binom{n}{d} p^d (1-p)^{n-d} d^2 \\
 &= \sum_{d=1}^n \binom{n}{d} p^d (1-p)^{n-d} d^2 && \text{the } d=0 \text{ term vanishes} \\
 &= \sum_{d=1}^n n \binom{n-1}{d-1} p^d (1-p)^{n-d} d && \binom{n}{d} = \frac{n}{d} \binom{n-1}{d-1} \\
 &= np \sum_{c=0}^{m} \binom{m}{c} p^c (1-p)^{m-c} (c+1) && c = d-1, m = n-1 \\
 &= np \sum_{c=0}^m \binom{m}{c} p^c (1-p)^{m-c} c + np && \text{by the normalization of } P[c; m, p] \\
 &= np \cdot mp + np && \text{by the above result} \\
 &= n(n-1)p^2 + np \\
 &= n^2p^2 - np^2 + np = \mu^2 + np(1-p)
 \end{aligned}$$

Since $\sigma^2 = E[d^2] - \mu^2$,

$$\sigma^2 = E[(d - \mu)^2] = np(1-p). \quad (7)$$

The standard deviation of the binomial distribution is thus $\sqrt{np(1-p)} = \sqrt{\mu(1-p)}$.

Several examples of binomial distributions are shown in Figure 1. All of the distributions have the same mean μ , but note how the standard deviation σ changes as n increases. In particular, σ tends towards $\sqrt{\mu}$ as n gets larger, since $1-p \approx 1$ as $p = \mu/n$ becomes very small.

2.2 Poisson Distribution

Combinatorial operators are generally painful to work with. Consequently, we'd like to simplify (1) to something a little "friendlier." When p is very small and n is very large, we can make several approximations. The expected

number of (packet corruption) events is just $\mu = np$, the mean of a binomial distribution.² Simplifying (1):

$$\begin{aligned}
P[d; n, p] &= \binom{n}{d} \cdot p^d \cdot (1-p)^{n-d} \\
&= \frac{n!}{d!(n-d)!} \cdot p^d \cdot (1-p)^{n-d} && \text{definition of combinatoric} \\
&\approx \frac{n^n e^{-n} \sqrt{2\pi n}}{d!(n-d)^{n-d} e^{-(n-d)} \sqrt{2\pi(n-d)}} \cdot p^d \cdot (1-p)^{n-d} && \text{Stirling's approximation of } x! \\
&= \frac{n^{n+1/2} e^{-d}}{d! (n-d)^{n-d+1/2}} \cdot p^d (1-p)^{n-d} \\
&= \frac{n^{n+1/2} e^{-d}}{d! n^{n-d+1/2} (1-\frac{d}{n})^{n-d+1/2}} \cdot p^d (1-p)^{n-d} && (n-d)^a = n^a (1-\frac{d}{n})^a \\
&\approx \frac{n^d e^{-d}}{d! e^{-d+d^2/n-d/2n}} \cdot p^d (1-p)^{n-d} && (1-\frac{d}{n})^a \approx e^{-ad/n} \\
&\approx \frac{n^d}{d!} \cdot p^d (1-p)^{n-d} && e^{-d+d^2/n-d/2n} \approx e^{-d} \\
&= \frac{\mu^d}{d!} \cdot (1-p)^{n-d} && n^d p^d = \mu^d \\
&\approx \frac{\mu^d}{d!} \cdot e^{-np} && (1-p)^{n-d} \approx e^{-np+dp} \approx e^{-np}
\end{aligned}$$

$$P[d; \mu] = \frac{\mu^d \cdot e^{-\mu}}{d!}. \quad (8)$$

We call (8) the *Poisson distribution*, and this describes most processes where a small number of events are observed. An alternate derivation of the Poisson distribution exists for events that are not discretized, and hence the Poisson distribution is more general than simply the large- n limit of the binomial distribution.

There are several things to observe about the Poisson distribution. The mean is

$$\begin{aligned}
E[d] &= \sum_{d=0}^{\infty} \frac{\mu^d e^{-\mu}}{d!} d \\
&= \sum_{d=1}^{\infty} \frac{\mu^d e^{-\mu}}{d!} d && \text{the } d=0 \text{ term vanishes} \\
&= \sum_{d=1}^{\infty} \frac{\mu \mu^{d-1} e^{-\mu}}{(d-1)!} && d! = d \cdot (d-1)! \\
&= \mu \sum_{c=0}^{\infty} \frac{\mu^c e^{-\mu}}{c!} && c = d-1
\end{aligned}$$

$$E[d] = \mu, \quad (9)$$

Though the Poisson distribution is only defined for integer values, the mean μ might not be an integer. The variance

²This is, in fact, the case we would expect on most links. Line noise is typically low enough that large numbers of packet corruptions are rare.

Comparison of Binomial and Poisson, $\mu=2$

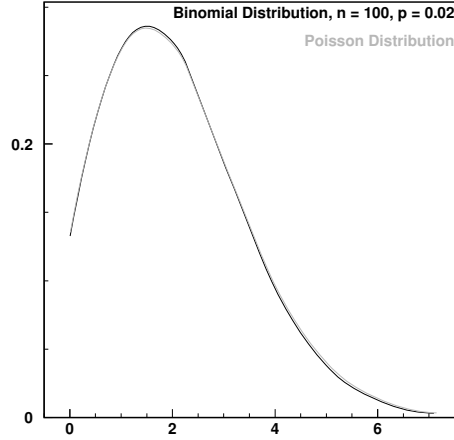


Figure 2: A binomial distribution and Poisson distribution with equal means.

of the Poisson distribution is

$$\begin{aligned}
 E[d^2] - \mu^2 &= \sum_{d=0}^{\infty} \frac{\mu^d e^{-\mu}}{d!} d^2 - \mu^2 \\
 &= \sum_{d=1}^{\infty} \frac{\mu \mu^{d-1} e^{-\mu}}{(d-1)!} d - \mu^2 && \text{the } d=0 \text{ term vanishes and } d! = d \cdot (d-1)! \\
 &= \mu \sum_{c=0}^{\infty} \frac{\mu^c e^{-\mu}}{c!} (c+1) - \mu^2 && c = d-1 \\
 &= \mu \left(\sum_{c=0}^{\infty} \frac{\mu^c e^{-\mu}}{c!} c + 1 \right) - \mu^2 && \text{by the normalization of } P[c; \mu] \\
 &= \mu^2 + \mu - \mu^2 && \text{by the above result} \\
 \sigma^2 &= E[(d - \mu)^2] = \mu, && (10)
 \end{aligned}$$

so the standard deviation of a Poisson-distributed counting measurement is $\sqrt{\mu}$. As with the binomial distribution, of which this is a special case, the probability of measuring a particular d drops off the farther d differs from the mean μ . A comparison of a binomial distribution and a Poisson distribution is shown in Figure 2, demonstrating graphically that the Poisson distribution is a limiting case of a binomial distribution. Poisson distributions of various means are shown in Figure 3. Note that as the mean increases, the asymmetry of the distribution decreases.

When we measure d events from a rare ($p \ll 1$) Bernoulli process, we assume that this is the mean of a Poisson distribution. As described in Section 2.1.1, the standard deviation represents the uncertainty in the measurement, and is \sqrt{d} . Putting these together, we would report the measurement as $d \pm \sqrt{d}$. For many experiments, this is the defining source of uncertainties.

2.3 Gaussian Distribution

The Poisson distribution still includes $d!$, which becomes more time-consuming to compute as d grows larger. Growth of d generally follows growth of μ , since the Poisson distribution drops off rapidly as d diverges from μ , so we can make the substitution $d = \mu + z$ for $z \ll \mu$ as μ becomes large.

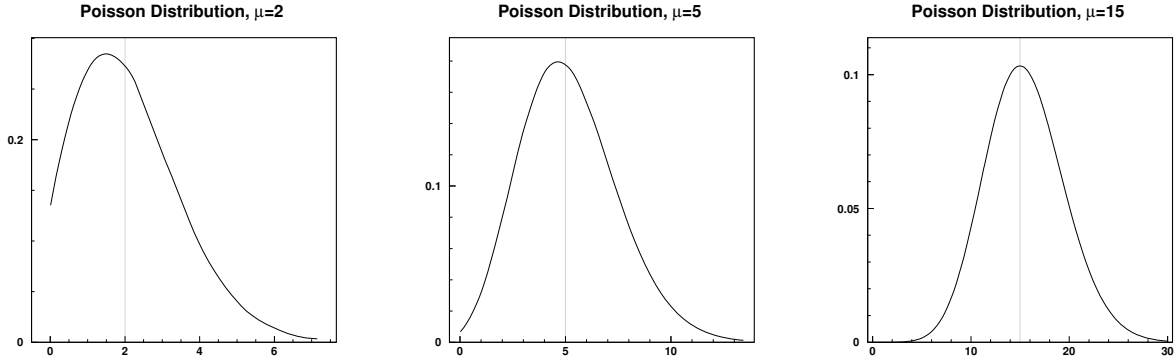


Figure 3: Poisson distributions with varying means.

We can now simplify (8):

$$\begin{aligned}
 P[z; \mu] &= \frac{\mu^{\mu+z} e^{-\mu}}{(\mu+z)!} && d = \mu + z \\
 &\approx \frac{\mu^{\mu+z} e^{-\mu} e^{\mu+z}}{(\mu+z)^{\mu+z} \sqrt{2\pi(\mu+z)}} && \text{Stirling's approximation of } x! \\
 &= \frac{\mu^{\mu+z} e^z}{(\mu+z)^{\mu+z} \sqrt{2\pi(\mu+z)}} \\
 \ln P[z; \mu] &\approx (\mu+z) \ln \mu + z - (\mu+z) \ln(\mu+z) - \ln \sqrt{2\pi(\mu+z)} && \text{taking natural log of both sides} \\
 &= (\mu+z) \ln \left(\frac{\mu}{\mu+z} \right) + z - \ln \sqrt{2\pi(\mu+z)} && \ln a - \ln b = \ln(a/b) \\
 &= -(\mu+z) \ln \left(1 + \frac{z}{\mu} \right) + z - \ln \sqrt{2\pi(\mu+z)} && \ln(a/b) = -\ln(b/a) \\
 &= -(\mu+z) \left[\frac{z}{\mu} - \frac{1}{2} \left(\frac{z}{\mu} \right)^2 + O\left(\frac{z^3}{\mu^3} \right) \right] + z - \ln \sqrt{2\pi(\mu+z)} && \text{Taylor expansion of } \ln \left(1 + \frac{z}{\mu} \right) \\
 &\approx -z - \frac{z^2}{2\mu} + \frac{z^3}{2\mu^2} + z - \ln \sqrt{2\pi(\mu+z)} \\
 &\approx -\frac{z^2}{2\mu} - \ln \sqrt{2\pi(\mu+z)} \\
 P[z; \mu] &\approx \frac{e^{-z^2/2\mu}}{\sqrt{2\pi(\mu+z)}} && \text{exponentiating both sides} \\
 &\approx \frac{e^{-z^2/2\mu}}{\sqrt{2\pi\mu}} && \sqrt{\mu+z} \approx \sqrt{\mu}
 \end{aligned}$$

This is simply a *Gaussian distribution* with $\sigma^2 = \mu$

$$P[d; \mu, \sigma = \sqrt{\mu}] = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(d-\mu)^2/2\sigma^2} \quad (11)$$

The Gaussian distribution is sometimes called the *normal distribution*. Demonstrating that the mean and variance of this distribution are indeed μ and σ^2 respectively is left to the motivated reader, though we note that a Gaussian is a continuous distribution, so expectations must be calculated as integrals.³ If the final approximation in the derivation

³There are hard ways to do this and easy ways. For the mean, I recommend a change of variables that exposes the symmetry of the integral. For the variance, a properly set up integration by parts is one “simple” way to compute the integral.

Comparison of Poisson and Gaussian, $\mu=40$

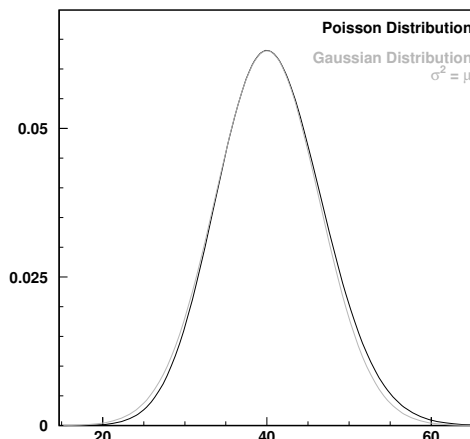


Figure 4: A Poisson distribution and a Gaussian distribution with equal means.

($\sqrt{\mu + z} \approx \sqrt{\mu}$) seems troubling, note that the exponential in the numerator causes the probability distribution to go to zero much faster than the increase in the denominator could. Also note that the constant $1/\sigma\sqrt{2\pi}$ is needed so that the distribution is normalized. We could have focused exclusively on the functional dependence on z in the derivation and computed the normalization constant afterwards.

Again, we compare a Gaussian and Poisson distributions (Figure 4), and see that the Poisson distribution indeed approaches a Gaussian with $\sigma^2 = \mu$ as the mean increases. The shape of a Gaussian distribution remains fundamentally unchanged; only the mean and width vary as its parameters change.

For large counting measurements, we assume that a measured value d is drawn from a Gaussian distribution with mean d and standard deviation \sqrt{d} . As before, this implies that our measurement is reported as $d \pm \sqrt{d}$. Note that for a Gaussian the distribution is symmetric around the mean. This is not in general true for Poisson or binomial distributions. Because symmetric uncertainties are easier to work with, we often treat measurements as Gaussian-distributed even when the total number of measured events is small. This is especially true at the earlier stages of an analysis, when precision is less important than general characterization of the data.

3 Repeated Measurements

An experiment seldom consists of performing a single count. In fact, in many experiments the process of counting is almost irrelevant, but we'll address that in a bit. For now, we'll examine what happens when you perform multiple counts of a particular process.

Returning to our noisy link, we might count the number of corrupted packets that occur for each one thousand packets transmitted, and repeat this counting a number of times. We can simulate this with the following simple program, which we'll call `throws.pl`:

```
#!/usr/bin/perl

$exe = 'basename $0';
chomp $exe;

die "Usage: $exe <nthrows> <probability> [<nmeas>]\n" unless ( @ARGV >= 2 );

($nthrows, $prob, $nmeas) = @ARGV;
if ( ! defined $nmeas )
```

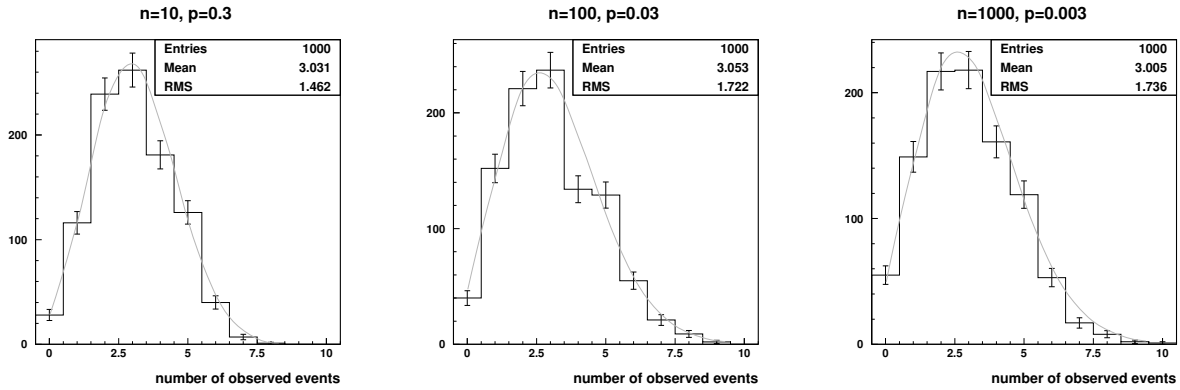


Figure 5: Simulated experiments for different numbers of observations and probabilities, but with a common mean. The horizontal axis gives the number of events observed in a single experiment, and the vertical axis gives the number of experiments producing that number of events, along with one standard-deviation (ie, square-root) error bars. For each figure, one thousand experiments are run. The theoretical (scaled binomial) distribution is overlaid for comparison.

```

{
    $nmeas = 1;
}

for ( $n = 0 ; $n < $nmeas ; ++$n )
{
    $total = 0;
    for ( $i = 0 ; $i < $nthrows ; ++$i )
    {
        $p = rand(1);
        ++$total if ( $p < $prob );
    }
    print "$n $total\n";
}

```

This program takes as its first argument the number of observations in an experiment; that is, the total number of packets transmitted. The second argument gives the probability of observing an event for each observation. The mean number of observed events should then be the product of these two arguments. The third (optional) argument allows us to perform multiple experiments, each producing a different count.

What do we expect to see when we run this program? According to the previous section, we should see a binomial distribution of the counts, since the probability of each recorded measurement is given by the binomial distribution. When the number of observations is large, the distribution should have a Poisson or Gaussian shape, depending on the mean. For a small mean ($\mu = 3$), several examples are shown in Figure 5 as *histograms*. One thousand experiments are performed for a given probability and number of observations. The result of an experiment (the count) determines a *bin* of the histogram in which to place the event. The final content of a bin (vertical axis) is thus the number of experiments that yielded the bin's value (horizontal axis). As the number of observations is increased, the distribution looks more Poisson-like. In particular, the standard deviation (called "RMS", for "root-mean-squared", in the figures) approaches $\sqrt{\mu} = 1.732$.

For comparison, Figure 6 shows a simulation for $\mu = 30$. As we would expect, this distribution is Gaussian in shape. In order to accentuate the shape, we have widened the bins to include five different results in each. This increases the statistics in each bin, so the proportional variance in each (\sqrt{d}/d) is smaller, reducing random fluctuations from bin to bin. Selecting the appropriate binning for data is often an important part of data analysis, and we will return

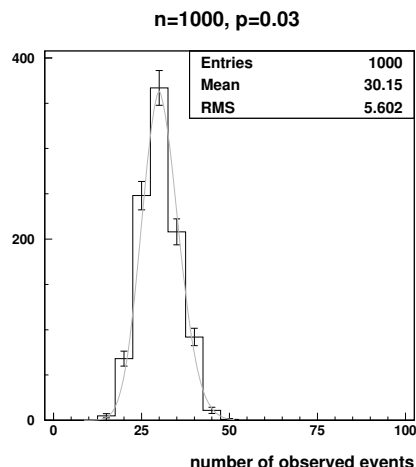


Figure 6: Simulated experiments for a large expected mean.

to this later.

3.1 Probability Distribution Functions

The distributions we've looked at so far are all specific examples of a more general class of *probability distribution functions*, or PDFs. The fundamental property of a PDF for a process is that the area under a segment of the function is the probability that a measurement will fall in that segment. Another way to look at this is that as the number of measurements increases, the sample distribution will qualitatively more closely resemble the PDF. The distribution in Figure 6, for example, is approximately a Gaussian with mean 30 and width $\sqrt{30}$ multiplied by 1000. Since the peak of a Gaussian (at its mean) is $1/\sigma\sqrt{2\pi}$, we expect this particular histogram to have a maximum bin content of

$$\frac{1000 \text{ measurements} \times 5 \text{ counts/bin}}{5.48 \text{ counts} \times \sqrt{2\pi}} = 364 \text{ measurements/bin}$$

which is very close to the actual maximum of 367. Note that we also had to scale by the size of a bin (effectively integrating a roughly constant function over the bin), because a bin is the unit of measurement of the histogram, but not of the horizontal axis. Note also that we have chosen the binning so that the expected mean lies at the center of a bin. Had the mean been on the edge between bins, values very close to the mean would have been spread between the two bins.

PDFs can have many different shapes, and do not necessarily represent binomially derived shapes. For example, in radioactive decays there is a constant probability of any atom decaying in any finite interval. The number of decays is thus proportional to the number of atoms that have not yet decayed. As atoms decay, there are fewer left un-decayed, so in the next finite interval we will see fewer decays. The result of this is a decreasing, or *decaying*, exponential.

Again, we can simulate this, and the results are shown in Figure 7. This simulates the decay of carbon-10, which has a half-life of 9.3s. Counts are recorded every second, with the cumulative time recorded on the horizontal axis. The content of each bin is a count, and so is subject to counting statistics. In order to highlight the effects of counting statistics, three different experiments are shown, corresponding to different initial population sizes. For each time slice in the histogram, we represent the count as a point (centered in the bin) and an *error bar*. These error bars extend \sqrt{d} above and below the data point, indicating the uncertainty of the measurement. Note that for larger populations we have both (proportionally) smaller error bars and a smoother distribution. In fact, if our error bars shrank proportionally but the distribution did not become closer to the theoretical curve, that would indicate a problem with the theory. We'll address this when we discuss fitting data to curves in the next section.

Another point regarding PDFs is that they do not need to come from some concrete analytical theory. In addition to combining different distributions in order to account for multiple effects, it is also common to determine a suitable analytic PDF and determine afterwards what could give rise to it. Slightly less common, PDFs might be determined

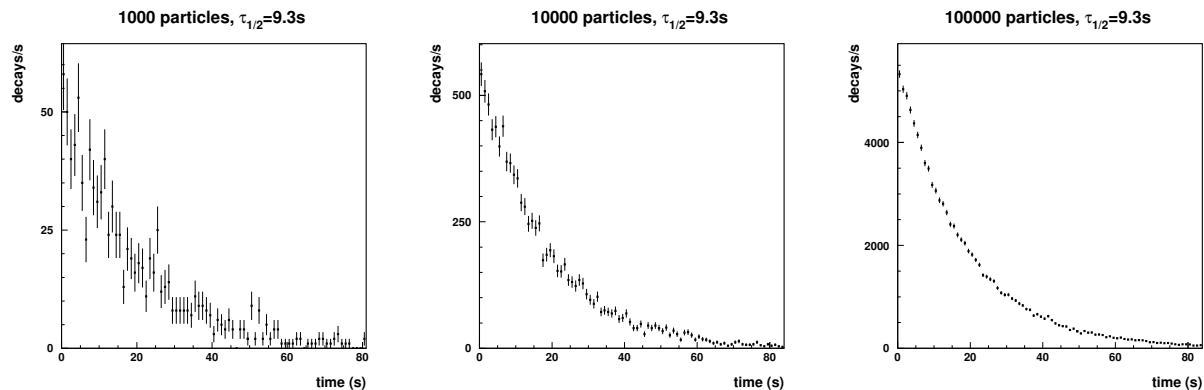


Figure 7: Simulation of ^{10}C decay for three different initial populations. Smaller populations may be thought of as enlargements of appropriate tail regions from larger-population distributions, though all three distributions are from independent simulations.

empirically from other measurements or simulations as histograms which are then scaled to have unit area. Empirical PDFs add a further *systematic uncertainty* (see Section 4.2) when used in an analysis.

3.2 Fitting and Binned vs. Unbinned Data

Sometimes we need to extract more information from data than statistics such as mean and variance. For instance, two competing hypotheses might predict the same statistical measures but with different distributions, or the specific details of a distribution might provide additional information applicable to other cases. Sometimes the data is distributed according to a function that cannot be normalized (such as a line or a parabola), and consequently there are no meaningful statistics because expectation values cannot be computed. In these cases, we must try to *fit* the data to a distribution.

Up to now, when we've looked at data, it has always been grouped into bins (except for the radioactive decay example, the measurements have been integral and hence inherently binned, though Figure 6 has a coarser binning). There are times when this is not ideal, in particular when there is relatively little data. What do we mean by little data? If binning the data would result in either a large number of empty bins or a very small number of bins that are so wide that they obscure the distribution of the data, then we are likely better off leaving the data *unbinned*. The choice of binning data or leaving it unbinned will have some impact on how we perform the fitting.

In some cases, fitting can be reduced to an analytic formula. For example, you might have used linear regression in an introductory physics course. In general, such a formulation is either not possible or not worth the effort. Instead, the function *parameters* are varied while trying to maximize or minimize some statistic for the given data. We will not discuss these variation techniques—fitting is generally done using some data analysis software package, and our goal with respect to fitting is to understand what goes into and comes out of one of these packages.

3.2.1 Maximum Likelihood Fitting

Fitting using the method of maximum likelihood is a general technique that works with both binned and unbinned data. We'll begin by considering unbinned data.

A given PDF tells us the probability of finding a data point at a particular value. Consider a PDF $f[x; \alpha]$, where $\alpha = \{\alpha_1, \dots, \alpha_m\}$ is a set of m parameters describing the shape of f . For example, for a Gaussian $\alpha = \{\mu, \sigma\}$. If f is normalized, that is

$$\int f[x; \alpha] dx = 1,$$

then the probability of a particular measurement x_i is $f[x_i, \alpha]$.

With multiple data points, we can form a combined probability

$$\mathcal{L} = \prod_i f[x_i, \boldsymbol{\alpha}]. \quad (12)$$

Because the data are given and the parameters are to be determined, we call this the *likelihood* that the parameters are correct for the data. As the parameters are varied, \mathcal{L} will change, and for some set of parameter values it will obtain its maximum value, or the *maximum likelihood*. If f incorporates all of our theoretical assumptions (perhaps including bounds on the parameters), then the parameters that yield the maximum likelihood are, by definition, the most likely set of parameters that correspond to the data provided.

Often, instead of directly maximizing \mathcal{L} , we'll instead minimize the related quantity $L = -\ln \mathcal{L}$. If the parameters obey a Gaussian distribution (as is often the case), then their fit values α_i are the means of Gaussians with widths σ_i . Varying only one parameter α_i changes the likelihood L from its minimum value L_{\min} such that

$$L - L_{\min} = \frac{(\alpha_i - \langle \alpha_i \rangle)^2}{2\sigma_i^2}$$

where $\langle \alpha_i \rangle$ is the fit value of the parameter.

Significance Values We define the *significance* s of a measurement x with mean μ and standard deviation σ as

$$s = \left| \frac{x - \mu}{\sigma} \right|, \quad (13)$$

so the above difference relates the fit likelihood to the significance as

$$L - L_{\min} = \frac{s^2}{2}. \quad (14)$$

From this, we can see that determining the range corresponding to s standard deviations is done by varying α_i on either side of its mean value until the difference in log-likelihoods is $s^2/2$. If this definition of significance seems odd, keep in mind that it is the significance of the *distance from the mean* or alternatively the significance of the *difference from a prediction*. If we are trying to verify a prediction then we want the significance to be small (not significantly different), but if we are trying to invalidate a prediction (such as the absence of an effect) then we want the significance to be large (significantly different).

Standard Deviations of Parameters The standard deviation is given by

$$\sigma_i^2 = 2(\alpha_i - \langle \alpha_i \rangle)^2 \quad \text{s.t. } L[\alpha_i] - L_{\min} = \frac{1}{2}. \quad (15)$$

Note that we can calculate the standard deviation on either side of the mean. In some cases, the parameter does not obey a Gaussian distribution, and the true distribution might be *asymmetric*, leading to two different values for σ_i , which we might label σ_i^+ (for values above the mean) and σ_i^- (for values below).

Non-Gaussian distributions also motivate calculating widths for values of s other than 1. In particular, say you are interested in the 99% confidence interval. That is, you want to be 99% certain that the fit value is within a particular range. The limits of this range can be determined as the values of α_i where $s = 2.58$.

Confidence intervals for various significances are shown in Table 1. Here we have both two-sided (inclusive) confidence intervals and one-sided (exclusive) confidence intervals. The latter are useful for setting upper or lower limits on a parameter. For instance, if $L[\alpha_i = 0] - L_{\min} = 1.34 = 1.64^2/2$, then we are 95% certain that the parameter α_i is non-zero. In some fields, claiming the observation of a new effect requires a significance of at least 5, which excludes about 3×10^{-5} % of the probability distribution.

s	Confidence Interval (%)	
	Two-Sided	One-Sided
0.67	50	75
1	68.26	84.13
1.15	75	87.5
1.28	80	90
1.64	90	95
1.96	95	97.5
2	95.44	97.72
2.33	98	99
2.58	99	99.5
3	99.74	99.87
3.29	99.9	99.95
3.89	99.99	99.995
4	99.9937	99.99685

Table 1: Significance Values and Confidence Intervals.

Trial # (i)	1	2	3	4	5	6	7	8	9	10
Result (d_i)	4	1	4	1	1	5	7	4	0	3

Table 2: Ten events from a binomially distributed counting experiment with $n = 1000$ and $p = 0.003$.

Uncertainty Contours We can, of course, vary multiple fit parameters from their optimal values. If we plot one parameter along the horizontal axis and another along the vertical axis, we can trace *contours* of constant ΔL , say corresponding to $s = 1$, $s = 2$, etc. Typically, these contours will describe ellipses (for Gaussian distributions). If the parameters are uncorrelated, their contours will have semi-major and semi-minor axes aligned horizontally and vertically. If they are correlated, the axes will be rotated according to the degree of correlation. We will discuss parameter correlations further in Section 4.3.

An important observation regarding unbinned likelihood fits is that we have no way to measure the quality, or *goodness*, of the fit. The parameters will obtain their optimal values, but we have no objective way to evaluate the fit other than whether the fit converges at all and how well the fit result matches the data in a subjective sense. Subjectively judging a fit is often not difficult when there is only a single independent variable. Often, however, we want to fit data to a function of two or more variables, at which point visualization becomes considerably more difficult.

An Example of Likelihood Fitting Let's consider the data that generated Figure 5, $n = 1000$, but without binning the data. Specifically, we'll consider the first ten events, shown in Table 2. We are going to fit this data to a Poisson distribution using the method of maximum likelihood.

The first thing to note is that the product in \mathcal{L} becomes a sum when we compute the negative natural logarithm L . Specifically,

$$L = \sum_i \left[\mu - d_i \ln \mu + \ln(d_i!) \right]$$

We then note that as the parameter μ changes, terms dependent only on the data do not change. Since we're interested in the point of smallest *relative* L , we can safely omit these data-dependent terms. In fact, the only terms that matter are those with some dependence on one or more parameters. This allows us to simplify our definition of L :

$$L = \sum_i \left[\mu - d_i \ln \mu \right] = n\mu - \ln \mu \sum_i d_i \tag{16}$$

As an aside, note that L asymptotically approaches infinity as μ approaches zero unless $\sum_i d_i = 0$.

In this simple example, we can actually solve for $\langle \mu \rangle$ (the optimal value of μ) directly, rather than scanning through different values of μ for the one that minimizes L . The optimal value of μ will be the one for which $dL/d\mu = 0$,

which is

$$\langle \mu \rangle = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{10} \sum_{i=1}^{10} d_i = \frac{30}{10} = 3$$

To find the uncertainty in $\langle \mu \rangle$, we need to find μ such that

$$\mu \cdot n - \ln \mu \cdot \sum_i d_i = \frac{s^2}{2} + \langle \mu \rangle \cdot n - \ln \langle \mu \rangle \cdot \sum_i d_i$$

for $s = 1$. By trial and error, we find that the values of μ for which this holds are 1.44 and 5.41, or $\sigma_{\langle \mu \rangle}^- = 1.56$ and $\sigma_{\langle \mu \rangle}^+ = 2.41$. We would typically write this as

$$\langle \mu \rangle = 3.00_{-1.56}^{+2.41}$$

It should not be surprising that our uncertainties aren't symmetric; since the amount of data we have is small, we'd expect to be far from the Gaussian limit. (See, for example, Figures 3 and 5.)

Note the number of significant figures quoted. There are no hard and fast rules governing this, other than that the number of significant figures should be sufficient to express the uncertainty (that is, not rounded to 0), and the uncertainty should not be expressed with an overly large number of significant figures. Three significant figures for the uncertainty is reasonable, though had $\sigma_{\langle \mu \rangle}^-$ been closer to either 1.5 or 1.6 I'd likely have used two significant figures, and had both uncertainties been closer to integer values I'd have used one. The asymmetry in the uncertainties is lost in this particular case, where to one significant figure we'd write

$$\langle \mu \rangle = 3 \pm 2,$$

though this too is a perfectly acceptable way to quote the fit result.

Likelihood Fitting of Binned Data When the number of data points becomes large, we typically bin the data so that it is easier to work with. Strictly speaking, this reduces the amount of information that we have, so if you're going to use the method of maximum likelihood, you're generally better off using it with unbinned data, if possible.

Sometimes we have no choice, and we want to fit binned data using the maximum likelihood. To do this, we use not the likelihood as defined in (12), but rather a *ratio of likelihoods*. We begin by defining a set of *predictions* based on the binning of the data and the current values of the fit parameters, which we'll call $\nu[x; \alpha]$. x is the set of central values for the bins, so a bin i from 0 to 1 would have a value $x_i = 0.5$. For each bin center x_i , there is an associated element $\nu_i[x_i; \alpha]$ of ν . Typically, ν_i will be defined by a function $f[x_i; \alpha]$ to which you're fitting. Each bin i , centered at x_i , has contents y_i .

The best that we could possibly do in a fit is for the predicted values ν_i to exactly equal y_i , so we normalize our likelihood to this and define the likelihood ratio as

$$\lambda = \frac{\prod_{i=1}^N \mathbf{P}[y_i; \nu_i[x_i; \alpha]]}{\prod_{i=1}^N \mathbf{P}[y_i; y_i]} = \prod_{i=1}^N \frac{\mathbf{P}[y_i; \nu_i[x_i; \alpha]]}{\mathbf{P}[y_i; y_i]}, \quad (17)$$

where $\mathbf{P}[y, \nu]$ is the Poisson distribution and N is the number of bins. In contrast to the unbinned fit, normalizing to the data distribution allows us to fit using a function that is *not* a probability distribution, for example a polynomial, and cannot itself be normalized.

As before, we're going to minimize the negative natural logarithm of this likelihood. However, we observe that as the number of samples increases, the quantity $-2 \ln \lambda$ is approximately distributed according to a χ^2 distribution. What we will minimize is thus

$$-2 \ln \lambda = 2 \sum_{i=1}^N [\nu_i - y_i + y_i \ln(y_i/\nu_i)] \quad (18)$$

Note that if $\nu_i = y_i$ for all bins, $-2 \ln \lambda = 0$. While we could, in principle, omit the y_i term, this would result in a quantity no longer distributed as a χ^2 , and with a minimum different from 0. Because we have multiplied the log likelihood by 2, the distance from the minimal value is s^2 for purposes of calculating confidence intervals and parameter uncertainties.

Goodness of Fit In order to evaluate the goodness of a fit, we need more than the minimal $-2 \ln \lambda$. We also need the *degrees of freedom*, which can be thought of as the number of dials that we can turn while minimizing $-2 \ln \lambda$.

For N bins and m fit parameters (that is, $\alpha = \{\alpha_1, \dots, \alpha_m\}$), the degrees of freedom n is defined as

$$n = \begin{cases} N - m - 1 & \text{if the sum of bin contents was fixed} \\ N - m & \text{if each bin can be viewed as Poisson-distributed} \end{cases} \quad (19)$$

The first case occurs when an experiment consists of measuring a particular number of events, which results in correlations between bins (robbing Peter to pay Paul, so to speak).

We will not go into detail about how goodness of fit is actually calculated, except to say that it is the integral of a χ^2 distribution for n degrees of freedom from the minimal fit $-2 \ln \lambda$ to infinity. If we call this integral p , what it tells us is that for any other experiment we perform for the same amount of data distributed according to the fit function, there is a probability p that we would obtain a *larger* value for $-2 \ln \lambda$. p is often called the *probability of χ^2* or the *confidence level* of the fit. Note that this is different than a confidence interval, which relates not to the quality of a fit but to the spread of parameter values.

How do we interpret the confidence level? This is, to an extent, subjective, but generally a confidence level of at least 5% is considered a successful fit. On the other hand, a confidence level *above* 90% is often considered *bad*. In particular, a very high confidence level is often an indication of *over-parameterization*. For instance, if you've fit your data to a line and the confidence level is very high, you should try fitting the data to a constant. If the confidence level of this second fit is acceptable, it is preferred over the linear fit. Another example would be a set of data that appears to have two Gaussians superimposed, one smaller than and offset from the other. If a fit to a single Gaussian yields a reasonable confidence level, it is generally considered that this single-Gaussian fit is the "correct" one.

A quick and easy way to judge the quality of a fit is to look at the ratio $(-2 \ln \lambda)/n$, often called the *reduced χ^2* . If this ratio is close to 1, the fit is reasonably good. As the number of degrees of freedom increases, the quality of the fit decreases more rapidly for a reduced χ^2 greater than 1.

Due to the subjective nature of what constitutes a "good" fit, it is common practice to include either the confidence level of a fit or $-2 \ln \lambda$ (often just called χ^2) and n . This allows the reader to make an independent decision regarding the goodness of the fit. The reduced χ^2 is sometimes quoted instead, but this should be avoided as it does not convey sufficient information to determine a meaningful goodness of fit.

3.2.2 Least Squares Fitting

Likelihood fitting of binned data requires bin contents to have Poisson or Gaussian (that is, square-root) uncertainties. If this requirement is not met, we must use a different fitting technique. The most common of these is *the method of least squares*.⁴

Consider data distributed in N bins, where the center of bin i is x_i , the bin's content is y_i , and the uncertainty in y_i is σ_i . We wish to fit this to a function $f[x; \alpha]$. When all parameters are independent (that is, uncorrelated), the quantity to minimize is

$$X[\alpha] = \sum_{i=1}^N \frac{(y_i - f[x_i; \alpha])^2}{\sigma_i^2} \quad (20)$$

It should be clear from this why we refer to this as least squares fitting. One way to regard this is that we are trying to find the set of parameters α for which the difference between the data and the fit function is least significant. When parameters are correlated, this becomes slightly more complicated. A decent fitting package will determine the correlation effects from the data and account for them properly during the fit.

We will not go into detail about least squares fitting here, as the fitting will generally be done by a statistical analysis package. What is worth noting is that confidence intervals and parameter uncertainties can be obtained from the significance

$$s^2 = X[\alpha] - X_{min} \quad (21)$$

and the goodness of fit is calculated in the same way as for binned likelihood fitting, where X is the lower limit of the χ^2 integral.

⁴Some disciplines, such as physics, refer to this as χ^2 fitting because the quantity to be minimized approximately follows a χ^2 distribution. This is not universally accepted terminology, and is one example of why physics publications often make mathematicians cringe.

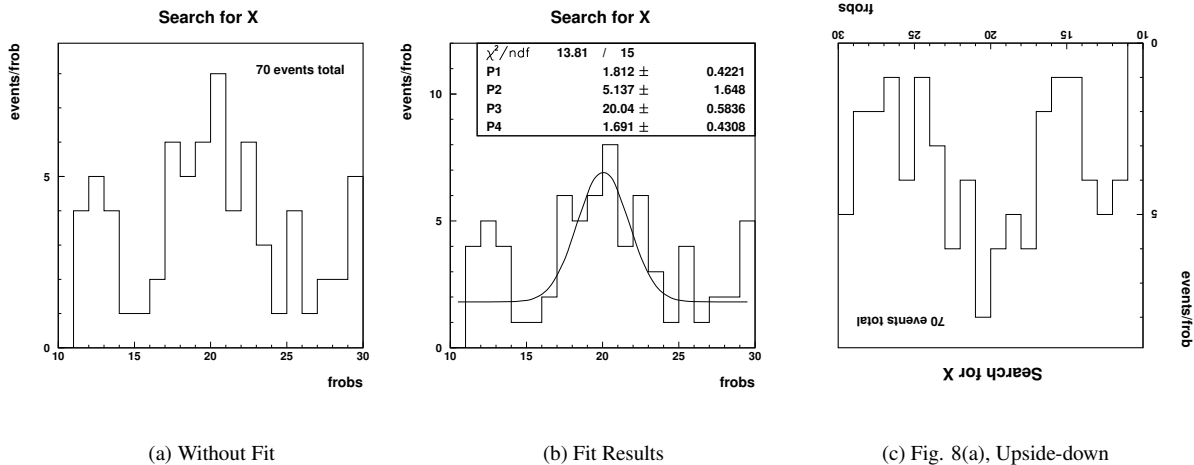


Figure 8: An experiment with a weak Gaussian signal over a constant background.

4 Measurement Uncertainties

We will now consider various sources of uncertainty in measurements. Understanding measurement uncertainties is crucial for understanding the results of any experiment. There are two basic sources of uncertainty: statistical and systematic. These are frequently kept separate, often through quoting a final result, since they are independent and represent different limits on the knowledge of a result.

4.1 Statistical Uncertainty

Statistical uncertainty has been the primary focus of our discussion so far. It is the measure of *precision* in our data; that is, to what tolerance we can measure something. Statistical uncertainty can generally be reduced as more measurements are made, as we have seen in Figure 7. That is, as the sample size increases, there is less variability in the statistics with which we characterize the data.

In particular, consider an experiment in which we are measuring the throughput of a file-swarming system. For a particular configuration, we measure the time required for every participant to receive the complete file. If we perform one such measurement, we get a value that tells us little about what we should expect for another trial. As we perform more and more measurements, a clearly defined mean and variance should emerge, from which we can more reliably characterize the expected performance of the system in that configuration. This assumes, of course, that there is some well-behaved distribution to completion times. We might in fact find that each new measurement is reasonably likely to alter the mean appreciably, in which case predicting the behavior of the system is probably futile.

When the statistical uncertainty in a measurement is large, it indicates that the experiment was to some extent insufficient for measuring the desired effect. This is often the case when a new effect emerges. Because we don't *a priori* know the size of the effect (otherwise it wouldn't be research), the first experiment to see it will often not have enough data to measure it conclusively. This generally leads to the establishment of a *limit*, rather than an actual measurement, and the limit is determined by a one-sided confidence interval.

A more concrete example should help to clarify this. We're going to consider two separate processes. One process, which we're trying to measure, is the *signal*. We expect it to appear as a Gaussian, if it appears at all. The other process is the *background*. We understand the background well enough to know that it should appear to be constant near the presumed signal, and we know how the level of the constant background relates to some interesting scale. The latter allows us to interpret the size of the signal in a meaningful way. The result of an experiment is shown in Figure 8.

Figure 8(a) shows the data in bins of width 1 "frob". We see what appears to be something like a Gaussian centered around 20 frobs. Fitting the data to a constant plus a Gaussian yields Figure 8(b). The first parameter is the constant, and the second is essentially the amplitude of the Gaussian. The mean and width are the third and fourth parameters. The confidence level of the fit is 54%, which is perfectly reasonable, but the significance of the Gaussian is only 3.12

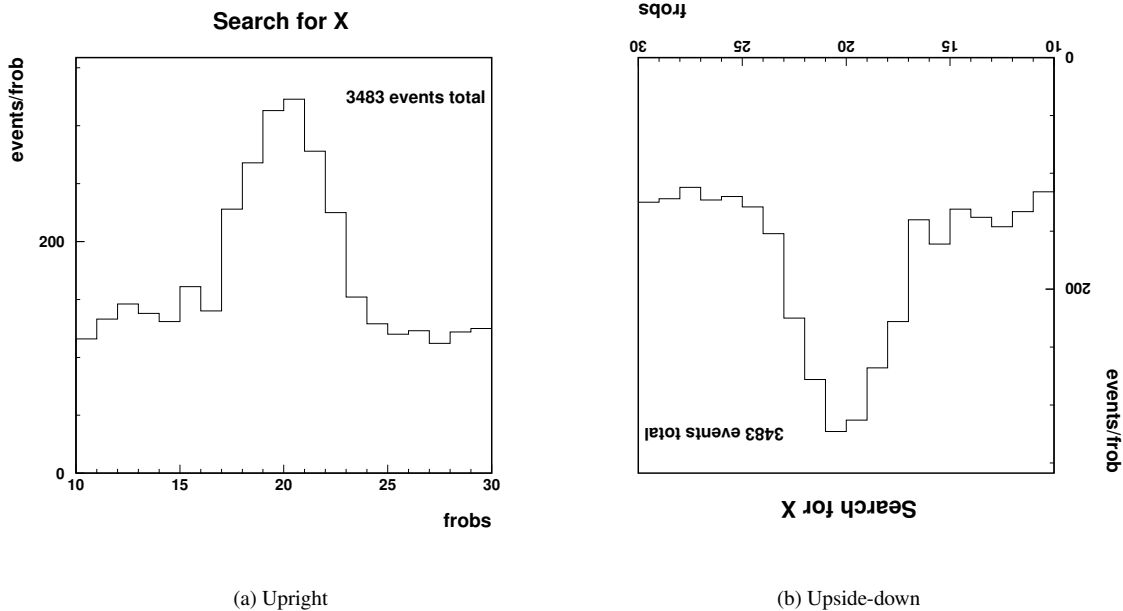


Figure 9: Data with a statistically stronger signal. Note that the dip is prominent to the eye.

assuming Gaussian errors on the parameter; the actual significance is 3.38. This is less than what we would consider acceptable for a new effect, so we want to set a limit. In particular, we think there’s an effect, but we can only say with certainty that it’s *less than* some value. If we want a limit with 95% confidence, we need (from Table 1) the value of the second parameter with significance 1.64, which is an *upper limit* of 7.95. This tells us that, at this scale, we are 95% confident that the effect has an amplitude less than 7.95.

Our brains evolved to pick out shapes sticking up above more-or-less level surfaces, so we have a tendency to see “peaks” where there are none. This can be very misleading when looking at data, since it leads us to see things that aren’t there. We’re not so good, however, at picking out dips. Figure 8(c) illustrates an easy way to separate our overzealous built-in pattern matching from genuinely significant effects. In this case, the data looks more like two shapes, centered around 27 and 15, over a constant background. Had the apparent Gaussian in Figure 8(a) been more statistically significant, we would instead see a pronounced dip below a constant background. This is illustrated in Figure 9.

4.2 Systematic Uncertainty

Where statistical uncertainty measures the precision of our experiment, systematic uncertainty measures the *accuracy*; that is, how biased our result is based on the experimental technique. Since most of our time is generally focused on statistical uncertainties, it’s very easy to lose sight of the systematic uncertainties. Of course, for some experiments systematics are the only source of uncertainty, making their importance considerably more obvious.

Systematic uncertainties arise from assumptions about our tools and methods. For example, a mis-calibrated ruler adds a systematic error (in the real sense of error, not as a synonym for uncertainty), and any ruler adds a systematic uncertainty of roughly half the smallest gradation.

Identifying assumptions can be difficult. Assigning them meaningful uncertainties can be considerably more so. Consider an experiment in which we measure the performance of a distributed system, such as a peer-to-peer network. The latencies, bandwidths available and used, processor time consumed, and so on are all affected by the specific environment in which the experiment was performed. If we perform one large experiment, taking many measurements over the course of a few minutes (or even a few hours), we can gather a reasonable statistical sample. However, we have to assume that the conditions under which the experiment was run are representative of conditions generally. We either need to be able to estimate the impact of changing conditions, or we need to perform more experiments under

different conditions.

Even if we perform multiple experiments in differing conditions, we still don't necessarily know how likely those conditions are in general. If we can somehow determine the likelihood of a specific environment, we can potentially produce a PDF that describes the environment (most likely with several independent variables). Each experiment would contribute data for a fit of the PDF's characteristic parameters, and the square root of the PDF's variance would tell us the appropriate systematic uncertainty for our result.

If the above sounds daunting, it's no surprise. A more common, though less rigorous, method for assigning a systematic uncertainty to an assumption is to vary the assumption over a range of "reasonable" conditions. This might mean trying to find (or engineer) a period of high network utilization, for example. We can calculate results independently under each set of conditions, and conservatively state that the largest difference in results is the systematic uncertainty for that assumption.

Sometimes altering the assumptions is not possible. In these cases it is up to the experimenter to be creative in determining a credible uncertainty.

It is often not crucial to determine *all* of the systematic uncertainties, but rather only those uncertainties that will *dominate* the final result. If one assumption clearly produces a much smaller uncertainty than another, it can safely be neglected.

4.3 Propagation of Uncertainties

We are rarely so fortunate that a single count or a single data fit produces the quantity that interests us. Each intermediate result carries with it an associated uncertainty, and when these results are combined so too must their uncertainties be combined. Even with a single measurement, if we have multiple systematic uncertainties, these must be combined into a single value. How to combine uncertainties, how to *propagate* them through additional calculations, is the subject of this section.

4.3.1 Uncorrelated

We'll begin with the simplest case, that of uncertainties from independent sources. Moreover, we'll consider systematic uncertainties, where we have independent measures of uncertainty for the *same* quantity. Consider k sources of systematic uncertainty. For each source i we assign an uncertainty σ_i . The total uncertainty σ is then given by

$$\sigma^2 = \sum_{i=1}^k \sigma_i^2. \quad (22)$$

That is, we simply sum the individual uncertainties in quadrature. Given a statistical and systematic uncertainty, we can obtain the total uncertainty in exactly the same way, though it's often useful to list them separately, as this highlights whether the uncertainty in the result is dominated by paucity of data or limitations in the methodology.

The more complicated case is when we have a result that is a function of multiple intermediate results. In particular, consider a final result R given by

$$R = F[r_1, \dots, r_k], \quad (23)$$

and each r_i is an intermediate result with uncertainty δr_i . The total uncertainty δR is then

$$(\delta R)^2 = \sum_{i=1}^k \left(\frac{\partial F}{\partial r_i} \right)^2 (\delta r_i)^2. \quad (24)$$

The use of δQ for the uncertainty in a quantity Q is merely a convenience to avoid an over-proliferation of subscripts. You could equally well use σ_Q instead of δQ , and we'll switch back and forth between the two occasionally.

A few special cases are worth noting. In order to measure very small effects, such as the time it takes to compute a square root, we often perform some large number N of these effects and measure the total cumulative time $t \pm \delta t$. The time T to perform one calculation is then $T = t/N$, and the associated uncertainty is $\delta T = \delta t/N$.

The sum S of two values x and y has an uncertainty $\delta S = \sqrt{(\delta x)^2 + (\delta y)^2}$. Their average A has an uncertainty $\delta A = \frac{1}{2} \sqrt{(\delta x)^2 + (\delta y)^2}$. Their difference D (either $x - y$ or $y - x$) has an uncertainty $\delta D = \sqrt{(\delta x)^2 + (\delta y)^2}$, which is identical to the uncertainty of their sum.

If R can be written as

$$R = F[r_1, \dots, r_k] = C \cdot \prod_{i=1}^k r_i^{n_i} \quad (25)$$

for n_i positive or negative, then the uncertainty is given by

$$\left(\frac{\delta R}{R}\right)^2 = \sum_{i=1}^k n_i^2 \left(\frac{\delta r_i}{r_i}\right)^2 \quad (26)$$

4.3.2 Correlated

Correlated variables make the situation considerably more complex. When two variables are correlated, they vary together. For positive correlations, when one variable increases, so does the other. For negative correlations, when one variable increases the other decreases. We say that these variables share a *covariance*.

The definition of covariance is similar to that of variance, in that it's an expectation value. However, it's an expectation value of a multivariate function over two variables. Specifically, the covariance of variables x and y over a distribution $P[x, y]$ is

$$V_{xy} = E[(x - E[x])(y - E[y])]. \quad (27)$$

We can see from this that the variance of a variable x is just its covariance with itself, $\sigma_x^2 = V_{xx}$, and that $V_{yx} = V_{xy}$. If an ordering is imposed on the variables, the set of covariances defines a *covariance matrix* V . The diagonal elements of V are just the variances of the corresponding variables, and V is symmetric.

Correlation is an easier concept than covariance. The correlation C_{xy} between two variables x and y is 0 if they are uncorrelated, 1 if they are perfectly correlated, and -1 if they are perfectly anti-correlated. In principle, C_{xy} can take any value in the range $[-1, 1]$, and might depend on x , y , or other variables. $C_{xx} = 1$ for any variable x , and $C_{yx} = C_{xy}$. Covariance and correlation are related as follows:

$$V_{xy} = C_{xy} \sigma_x \sigma_y. \quad (28)$$

We need one more piece of notation before we can write down the formula for propagating uncertainties with correlations. For a function $F[r_1, \dots, r_k]$, we will write (∂F) to denote the column vector of partial derivatives, so

$$(\partial F) = \begin{pmatrix} \frac{\partial F}{\partial r_1} \\ \vdots \\ \frac{\partial F}{\partial r_k} \end{pmatrix}. \quad (29)$$

The transpose $(\partial F)^T$ is the corresponding row vector. This is not standard notation, but it greatly simplifies the expression for the propagated uncertainty. If $R = F[r_1, \dots, r_k]$, we can write the uncertainty in R as

$$(\delta R)^2 = (\partial F)^T \cdot V \cdot (\partial F). \quad (30)$$

Calculating covariances is decidedly non-trivial. During fitting, they will typically be approximated from the data as a part of the regular fit procedure, and often you will be able to extract the correlations or covariances between the fit parameters. The covariances in a fit are computed from approximate correlations called sample, or Pearson product-moment, correlation coefficients.

When combining quantities analytically, it is usually simpler to ensure that your quantities are independent. For example, if you are taking the ratio of two measured quantities, it matters critically whether they are dependent or not. To make our example concrete, say

$$r = \frac{n}{d}. \quad (31)$$

A ratio defined by two separate measurements has an uncertainty

$$\delta r = r \sqrt{\frac{1}{n} + \frac{1}{d}}. \quad (32)$$

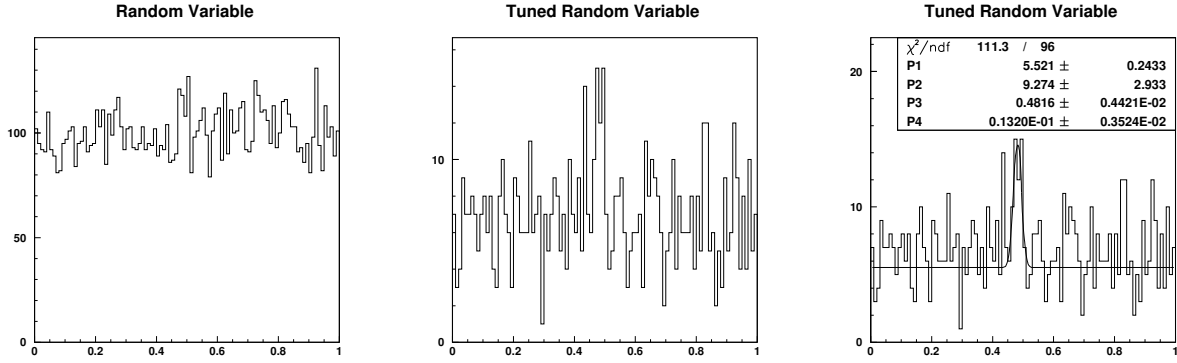


Figure 10: Tuning of data selection criteria in order to create an artificial signal.

If the ratio is actually an efficiency, so that the data in n are a subset of the data in d , the measurements are correlated, and the uncertainty is

$$\delta r = r \sqrt{\frac{1}{n} - \frac{1}{d}}. \quad (33)$$

This is reasonably easy to compute if we replace the denominator d with $n + b$, where the two populations n and b are independent. The reader is encouraged to verify this as a simple, but non-trivial, exercise in propagating uncertainties.

5 Data Selection and the Pitfalls of Tuning

Data is often not represented by a single variable. A signal might show up in one variable, but other variables might help distinguish between signal and background. Some experiments involve a huge number of variables, and some selection criteria must be applied to these variables in order to drive down the effect of the background while preserving as much of the signal as possible. My PhD thesis analysis involved over 400 variables, though not all of them were ultimately used.

One thing an experimenter must be careful of in these situations is *tuning* data selection criteria to accentuate the signal. Selection criteria generally come from some well-established properties, such as a temperature range in which the signal might be present but outside of which it cannot be. Tuning refers to varying these criteria based not on what's known *a priori* but on the characteristics of the data itself.

Let's look at a specific example. Consider a sample of 10,000 data points. Each point comprises five different variables distributed uniformly at random between 0 and 1. One of these variables is shown in the left-hand picture of Figure 10. We're going to create an artificial signal in this variable with a significance of at least 3 centered near 0.5, even though no such signal is actually present. To do this, we select a small area around 0.5 and look at the other four variables for any deviation from a uniform distribution; this gives us a handle with which to tune the data.

Using only three of the remaining variables, we're able to apply very mildly tuned selection criteria to produce the distribution (again over the first variable) in the center plot of Figure 10. A peaked shape starts to emerge from the data at this point. With more effort we could probably have produced an even cleaner-looking "signal".

A fit using a Gaussian over a constant background is shown in the right-hand plot. According to the symmetric uncertainties in the amplitude of the Gaussian (parameter 2), the significance of this signal is 3.16. The fit converged without problems, and the confidence level of the fit is 13.6%, which isn't bad at all for a non-existent effect.

There are a couple of lessons to draw from this. The first is that if you cannot derive useful data selection criteria from first principles, then you should derive them from a simulation or an independent data sample not contributing to the result. Never use the data to determine selection criteria. The second lesson is that you should always list your selection criteria as part of your experimental methodology, and be skeptical of any results presented without this information provided. You should also look for some credible rationale for each selection criterion applied.

6 Additional References

I've purposely avoided many details in this note, partly for length and partly to avoid overwhelming the reader. This should, however, provide you with a sufficient background in experimental statistics to properly analyze your own experimental data in the vast majority of cases. There are many topics which have been omitted entirely, among them:

- determining distributions to which data should be fit,
- selecting appropriate binning for data,
- auto-correlation and hysteresis in data,
- more sophisticated statistical methods and tests,
- and many mathematical details.

Much of what's covered here is also covered more rigorously in the Particle Data Group's Review of Particle Properties (<http://pdg.lbl.gov>). In particular, the reviews on Probability, Statistics, and Monte Carlo techniques are worth looking at for more details. The intended audience for these review articles is particle physicists, but the discussions are more mathematical than physical in nature.

The book *Probability and Statistics*, by Morris H. DeGroot and Mark J. Schervish, comes highly recommended as a general introduction that does justice to the topic. Another reference is *Data Reduction and Error Analysis for the Physical Sciences*, by Philip R. Bevington and D. Keith Robinson, which has a practical bent. Wikipedia (<http://en.wikipedia.org>), a free community-produced encyclopedia, has a number of articles on statistics.

There are a number of software packages that provide data analysis and fitting functionality. The figures and analyses in this note were produced with the PAW program, which is part of the CERNLIB programming suite (<http://wwwasd.web.cern.ch/wwwasd/cernlib/>). This is a very powerful program, but is geared towards the particle physics community. The primary effect of this is that, while there is a powerful macro facility and the ability to write programs that are interpreted at run-time, the interface inherits much of its feel from VMS and the programming is in a slightly restricted version of FORTRAN 77. Neither is especially difficult to learn, however. Another popular statistical analysis package is S, and its open-source doppelganger R.

7 Acknowledgments

This work was supported by DoD contract MDA90402C0428 and NSF ITR Award CNS-0426683.

I'd like to thank Stephen Pappas for numerous discussions on statistics as well as his comments on this note. A number of graduate students also read and provided feedback on an earlier draft: Vijay Gopalakrishnan, Dave Levin, Christian Lumezanu, Ruggero Morselli, and Rob Sherwood.

Finally, I'd like to acknowledge the systems group in the Cornell University Computer Science department. My near-continual harping on the lack of proper statistical analysis in systems papers led to discussions about presenting a statistics primer for computer scientists. While this never came to fruition, the idea of planning a lecture or series of lectures on statistics motivated me to write this note.