Introductory Notes on Probability and Statistics

Prof. Dave Goldberg

February 10, 2014

Contents

Background

These lecture notes aren't for a class. Rather, they're meant to serve two purposes:

1. They're intended as introductory (or review) material for students thinking about doing research in physics or astronomy. They are by no means comprehensive, but they should give students a working knowledge of the vocabulary, analytic tools and (provided I'm not too lazy) additional resources to adequately work on theoretical, observational, or experimental research.
2. They are intended as a worksheet for me. I am in the process of thinking about a forthcoming book on how randomness rules the universe and while my books are non-technical, I've found it useful to work things out mathematically before trying to describe them conceptually. That way, if I have any misconceptions of my own, I can identify them early.

This is a work in progress. At some point fairly early on, I'll be posting this on my blog. Comments and additions are appreciated.

1 Random Processes

There is often a disconnect between the work that we assign to students in physics classes and the skills required to be a working scientist. One of the biggest is an understanding of statistical analysis. On homework and exams we normally ask students to do some calculation with a fixed and well-determined analytic answer, but in the world of scientific inquiry we not only want to know the answer, but to have some sense of how likely that answer is to be correct.

The physical world – and our understanding of it – are all governed by **randomness**, and a big part of making a scientific prediction is to explore the space of probabilities in an effort to prove or disprove (or, more accurately, to support more or less) a physical theory.

We're going to build up our understanding of probability and statistics from the ground up. Naturally, I won't derive every relation, or even discuss every statistical test. Rather, I'll give enough of a sense of the derivations such that you could (with a term or two of calculus under your belt) derive the rest for yourself – or to google what you still don't understand.

We begin, naturally, with games of chance.

1.1 Binomial Probability

Gambling provides some of the most obvious examples of random processes. Slot machines, roulette wheels, and even (though the system is dramatically more complicated) horse racing provides insight into how random events *work*. Let's start with the simplest random system of them all: a coin flip.

A coin flip isn't really random. It's deterministic, at least in principle.

If you knew the exact weight of a coin, the air pressure and flow, the layout of all of the tables, the elasticity of the floor, and the exact force applied on the coin itself – all to insane detail – and had a powerful enough computer to do the calculations, you could figure out what side a coin would come up on any given flip.

In the absence of all of that, our laziness ensures that a coin flip is as random as it needs to be. Our idealized coin (which is going to form the basis of our first foray into randomness) will have the following properties:

1. The **probability**, p, of a coin coming up heads is $50\% = 0.5$.
2. This probability remains the same regardless of the results of the previous flip, the previous 10 flips, or the previous hundred flips. The formal way of saying this is that the flips are **statistically independent**.

Over time, we're going to relax both of these rules (especially the fairness of our coin).

I also should say something about what the word probability means. That's why I've put it in boldface. I'll generally take the **frequentist** interpretation of probability[1] which goes something like this:

> *Over a very long run, the frequency of outcomes of an experiment will approach a fixed ratio.*

or, as Aristotle put it even more simply (although not particularly quantitatively):

> *The probable is that which for the most part happens.*[2]

You'll note that in my "definition," I didn't necessarily assume that $p = 0.5$. I could certainly imagine circumstances where I don't know the weighting of a coin, and a frequentist would argue that the way to figure out if a coin is fair is to flip it a huge number of times.

[1] We'll encounter another interpretation, the Bayesian interpretation, in due course.

[2] *Rhetoric*, Book 1, Chapter 2.

How many times do I need to flip a coin before I can confidently compute p? If I flip it twice and it comes up heads both times, is it fair to say that $p = 1$? No. And we'll see why shortly.

In some sense, the frequentist interpretation requires us to imagine flipping a coin a billion times or more, counting up the outcomes, and determining p from that. This couldn't be done in practice, even if we had the patience. The wear on the coin would change its behavior over time and our calculation would try to hit a moving target.

Instead, we have to imagine the completely ridiculous situation of a billion of you flipping a billion nearly identical coins in a billion nearly identical universes and recording *those* outcomes. Suffice it to say, the frequentist interpretation requires at least some measure of abstraction.

Quantum mechanics is governed by a frequentist view. Measurement of the spin of an electron, or any other system that isn't in a particular eigenstate, should produce a truly random result. A large set of parallel universes with identical initial conditions would yield a particular outcome (spin-up, for instance) with a well-defined frequency.

Coins aren't the only game in town (so to speak). A fair die gives each outcome $p = 1/6$ of the time. In the US, a roulette wheel has 38 numbers (18 black numbers, 18 red numbers, and 0 and 00, both in green), meaning the probability of any particular outcome (33 black, for instance) is $p = 1/38$. The probability of a particular color is $p = 9/19$.

We don't particularly care what generates our random numbers, only that they are random. So for now, we're going to stick with coins.

1.1.1 Joint Probabilities

Enough philosophizing. You probably had an idea of how coin flips worked before we started with all of this. Instead, let's focus on calculating practical probabilities of more complicated series of events.

Suppose we flip a fair coin twice. Each flip has two possible outcomes: heads (H) or tails (T). There are exactly 4 different outcomes we can get:

$$\text{TT TH HT HH}$$

each with a 25% probability, since the probability of any *given* series of two independent events is:

$$P = p_1 \cdot p_2 \tag{1}$$

where P is known as the **joint probability** of the two events.

Since the probability of a heads and a tails are both 0.5 we get:

$$P = (0.5)(0.5) = 0.25$$

I hope you don't mind me being pedantic early on.

We could extend this to any number, N, flips, and at the end of it, we'd have a total of:

$$2^N$$

different series of possible flips, each with

$$P = \frac{1}{2^N}$$

The probability of flipping 1 head is 50%. The probability of 10 in a row is about 0.1%. 20 in a row is approximately 0.0001%, about 1 chance in a million. While *any* series of 20 flips is a one in a million probability, we only tend to notice those that have a distinct pattern.

We'll save a detailed discussion of this for later, but if a friend pulls out a coin and flips 20 heads in a row, you should at least *consider* the possibility that it's not a fair coin.

On the other hand, you shouldn't be surprised to see apparently "non-random" looking runs in a very long string of coin flips. Flip a million times and you're quite likely to see a string of 20 in a row.

More generally, if the probability of one of two possible outcomes (H, as I've been describing it) is p, then the probability of getting the other outcome (T) is $1-p$, and thus getting a *specific* sequence of m heads and $N-m$ tails is:

$$P = p^m (1-p)^{N-m} \tag{2}$$

1.1.2 Micro- and Macro- States

It's fairly unusual for us to care about the specific sequence of coin flips, the **microstate** of a system. Instead, we are generally concerned with the **macrostate**, the total number of heads and tails that we get.

For our 2 flip example, the cases TH and HT would normally be grouped together, and we'd simply want to know how many times we get two heads. To figure that out, we need to determine all of the permutations. I'll spare you the derivation, but if you have N events, and m of them come up heads, then the number of combinations that produce the same number of heads is:

$$\binom{N}{m} \equiv \frac{N!}{m!(N-m)!} \tag{3}$$

also known as "N choose m."

You can check, and you'll find that the degeneracy (as we'd call it in quantum mechanics) for $m = 1; N = 2$ is 2. You could, of course, plug in any pair of numbers you'd like.

Thus, the probability of a particular outcome (m heads) can be computed from the product of equation 2 and 3:

$$P(m) = p^m (1-p)^{N-m} \frac{N!}{m!(N-m)!} \tag{4}$$

which is a little cumbersome under normal circumstances, but for convenience, let's just plot up a weighted coin ($p = 0.6$) that we flip 10 times. Graphically, this looks like Fig. 1.

You will note that the most likely outcome corresponds to 6 heads. This shouldn't be a surprise. After all, the coin was weighted to come up heads 60% of the time, and we flipped 10 times.

That said, we can imagine another scenario one in which we work for a polling firm. Imagine that (unbeknownst to us), candidate A has a 60-40 lead over her opponent in a town's mayoral election. We decide to poll 10 people at random. In practice, just selecting the 10 people "at random" is difficult enough (as different groups tend to have different likelihoods of voting and different response rates), but we'll sweep that under the rug.

While most of the time (about 63%) candidate A wins the poll with 6 or more in favor, the two tie 20% of the time, and candidate B wins the poll about 17% of the time. Bear in mind, that this is a race in which the eventual winner (candidate A) has a *real* 20 point lead over her opponent.

The point is that "large enough" a sample depends sensitively on what you're trying to find out.

1.2 The PDF

The plot in Figure 1 is known as the Probability Distribution. It tells you how likely any given outcome is.

In this case (the repeated flipping of a potentially weighted coin), the possible outcomes are discrete: you can't ever get a fraction of a head. We're going to label the probability distribution as p_i. Appropriately

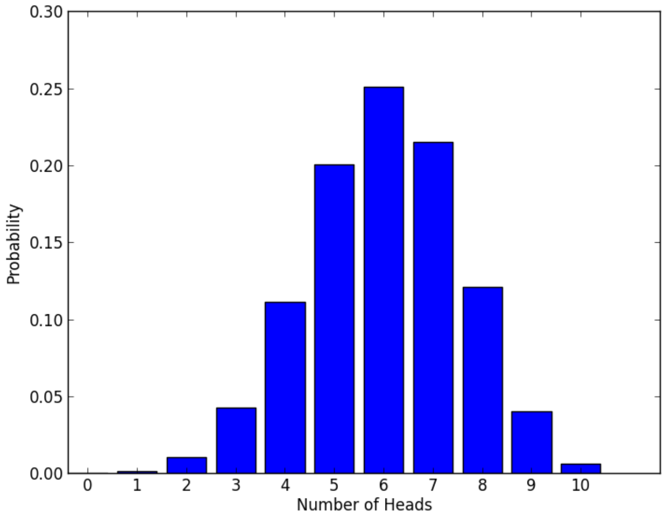

Figure 1: The Probability Distribution of 10 flips of a weighted (p=0.6) coin.

normalized, these probabilities have the property:

$$\sum_i p_i = 1 \tag{5}$$

These probabilities only make sense if they are all positive.

Not every random process falls so neatly into evenly spaced integer outcomes. There are also continuous probability distributions. For instance, if we're trying to predict the weather for tomorrow, we could be could try to figure out probabilities to hundredths of a degree if we like.

A continuous probability is known as a **Probability Density Function** (PDF). It is written as $f(x)$, such that:

$$f(x)\Delta x$$

is the probability of finding some outcome between the values of x and $x + \Delta x$.

This too has a normalization:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{6}$$

As we will see shortly, for large numbers, you can typically approximate a discrete distribution as a continuous one.[3]

One final note before we actually compute something from our PDFs. We sometimes want to know the probability of a particular value coming up that's *less than* some fiducial value. For instance: I flip 10 coins. What's the probability of getting 5 or fewer heads?

This is known as the **Cumulative Probability Distribution (CDF)** and it's computed as:

$$F(x) \equiv \int_{-\infty}^{x} f(x')dx' \tag{7}$$

[3] You may also note that we are primarily considering 1-d PDFs in this set of notes. This work does, indeed, extend to more dimensions, but most of the mean features are illustrated in 1-d.

1.2.1 Example: A Uniform Distribution

Suppose a variable is selected from a **uniform probability distribution** between 0 and 2. Thus,

$$f(x) = \frac{1}{2}$$

The cumulative probability in that range is:

$$F(x) = \int_0^x \frac{1}{2} dx' = \frac{x}{2}$$

By necessity, $F(0) = 0$, and $F(2) = 1$.

1.3 Mean, Standard Deviation, and All That

Now that we've gotten our definitions out of the way, let's see what we can do with our probabilities and PDFs. I'm going to concentrate on discrete distributions for the time being, but it should be fairly obvious at this point that a continuous distribution will look the same, except with an integral, rather than a sum.

Suppose we have a set of outcomes, $\{x_i\}$, each with a probability, P_i. We can compute the **mean** or, more accurately in this case **expectation value** (a term that you've likely seen in quantum mechanics) as:

$$\mu = \langle x \rangle = \sum_i P_i x_i \tag{8}$$

We need to make *some* distinction between the mean (normally given by μ) and the expectation value. The mean represents the average of some measured set of numbers. For instance, if we actually flipped a coin a million times in sets of ten, and averaged the number of heads per group, that would be a mean. In the frequentist interpretation, we're really taking the expectation value because we're omniscient and *know* the intrinsic probability of getting heads.

For the most part, however, I'm going to use the expected mean, μ, and the expectation value, $\langle x \rangle$, of a distribution interchangeably.

Those brackets surrounding the x tell us to take the expectation value of what's inside, even if it's a complicated function. So:

$$\langle Q(x) \rangle = \sum_i P_i Q(x_i) \tag{9}$$

The next highest "moment" of the distribution function is known as the variance:

$$\begin{aligned} \text{var}(x) &= \langle (x-\mu)^2 \rangle \qquad (10) \\ &= \langle x^2 \rangle - 2\mu\langle x \rangle + \mu^2 \\ &= \langle x^2 \rangle - \mu^2 \end{aligned}$$

where the second line simplifies to the third since $\langle x \rangle = \mu$.

Dimensionally, the variance has units of x^2, so to compare it meaningfully to the mean, we take the square root and get the **standard deviation**, σ:

$$\sigma^2 = \text{var}(x) \tag{11}$$

1.3.1 Example: Coin Flips

Let's consider these various statistics for our weighted ($p = 0.6$) coin. For a single flip, the mean is:

$$\begin{aligned} \mu_1 &= p \cdot 1 + (1-p)0 \\ &= 0.6 \cdot 1 + (1-0.6) \cdot 0 \\ &= 0.6 \end{aligned}$$

The mean, (like the expectation value in quantum mechanics) need not correspond to a realizable quantity.

The standard deviation can be computed as:

$$\begin{aligned} \sigma_1 &= \sqrt{p(1-\mu_1)^2 + (1-p)(0-\mu_1)^2} \\ &= \sqrt{0.6 \cdot (1-0.6)^2 + (1-0.6)(0-0.6)^2} \\ &\simeq 0.49 \end{aligned}$$

The standard deviation represents something like the spread in measurement that you might expect. After 1 flip, we'd expect to get 0.6 ± 0.5 heads. And of course, by necessity, we do.

Let's consider what happens if we apply the same analysis to 10 flips. I'll spare you the lengthy addition, but we get:

$$\mu_{10} = 6 \quad ; \quad \sigma_{10} \simeq 1.55$$

The mean, unsurprisingly, is 6. With a coin weighted 60% for heads, we *expect* 6 heads after 10 flips. That's the frequentist interpretation in a nutshell.

The standard deviation for 10 flips is fractionally much smaller than for 1 flip, only 16% of the total, compared to 50% of the total for 1 flip. We haven't shown this yet, but it'll be a general sort of rule that fractional errors scale as:

$$\frac{\sigma_N}{\mu_N} \propto \frac{1}{\sqrt{N}}$$

The longer you flip, the closer your measured average will approach the "true" value.

1.4 A few other stats

The mean and the standard deviations are the big 2, but there are other statistics that you might want to apply to your probability distribution. While the mean is based on x^1, and the variance (and hence standard deviation) are based on x^2, we can also compute the **skewness**:

$$\gamma_1 \equiv \frac{(x-\mu)^3}{\sigma^3} \tag{12}$$

which measures how lopsided a distribution is, and the **kurtosis**:

$$\gamma_2 \equiv \frac{(x-\mu)^4}{\sigma^4} - 3 \tag{13}$$

which measures the boxiness of the distribution. The 3 is subtracted off so that a Gaussian (the gold standard of PDFs which we'll encounter very shortly) will have zero kurtosis. Skewed and kurtote distributions are illustrated in Fig. 2.

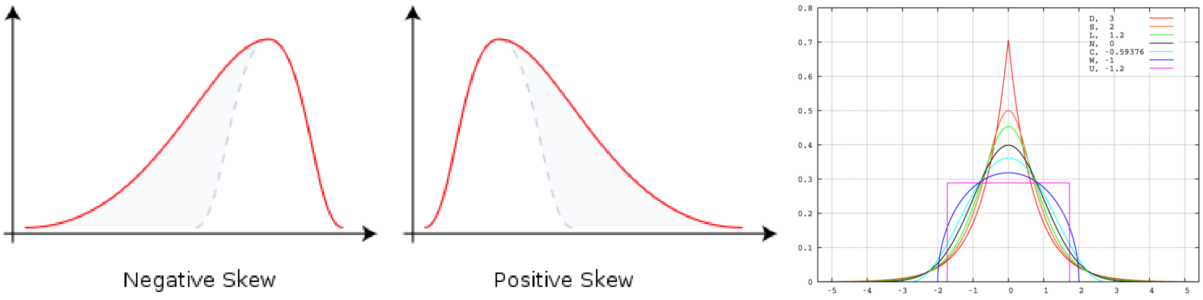

Figure 2: From wikipedia, skewed (left) and kurtote (right) distributions.

There are other statistics as well that don't come directly from measuring moments. The **median** is the value for which $F(x) = 0.5$, while the **mode** is the most popular value. For our $N = 10$ weighted ($p = 0.6$) coin flip example, we'd get 6 for both measures.

1.5 The Large Number Limit

Even with just 10 flips, the probability distribution begins to look like a fairly continuous function. We can ask the question: what happens to the probability as the number of flips, N becomes arbitrarily large?

In equation (4), we saw that after N flips, the probability of getting exactly m heads is:

$$P_N(m) = p^m (1-p)^{N-m} \frac{N!}{m!(N-m)!}$$

For reasons that will become clear in a moment, let's consider the natural log of this probability:

$$\ln(P) = m\ln(p) + (N-m)\ln(1-p) + \ln(N!) - \ln(m!) - \ln((N-m)!)$$

The natural log of a factorial is an odd term to encounter. Fortunately, for a large argument, there is a way to estimate this known as **Stirling's Approximation**:

$$\ln(n!) \simeq n\ln(n) - n \tag{14}$$

And thus, the natural log of the probability distribution is:

$$\begin{aligned} \ln(P) &= m\ln(p) + (N-m)\ln(1-p) + N\ln(N) - N - m\ln(m) + m - (N-m)\ln(N-m) + (N-m) \\ &= m\ln(p) + (N-m)\ln(1-p) + N\ln(N) - m\ln(m) - (N-m)\ln(N-m) \\ &= m\ln(p) + (N-m)\ln(1-p) + (N-m)\ln(N) + m\ln(N) - m\ln(m) - (N-m)\ln(N-m) \\ &= (N-m)\ln\left(\frac{N(1-p)}{N-m}\right) + m\ln\left(\frac{Np}{m}\right) \end{aligned} \tag{15}$$

To go any further, it's important to realize that the probability will peak when $m = Np$ (as we've already seen). Near that peak, the argument in each of the natural logs is approximately 1, and the probability will drop off very quickly far away from the peak. Consider the Taylor expansion of the natural log near 1:

$$\ln(1+x) \simeq x - \frac{x^2}{2}$$

Taking only the second set of terms on the right in equation (15), we note:

$$\begin{aligned} \ln\left(\frac{Np}{m}\right) &= \ln\left(1 + \frac{Np-m}{m}\right) \\ &\simeq \frac{Np-m}{m} - \frac{1}{2}\left(\frac{Np-m}{m}\right)^2 \end{aligned}$$

The linear terms end up canceling exactly with the similar first term on the right hand side of equation (15). Expanding out the second order terms we get:

$$\ln(P) = -\frac{1}{2}\left[(N-m)\left(\frac{Np-m}{N-m}\right)^2 + m\left(\frac{Np-m}{m}\right)^2\right]$$

or, combining the terms, we get:

$$\ln(P) = -\frac{1}{2}\left[N\frac{(Np-m)^2}{m(N-m)}\right]$$

This form works fine, but the function drops off far away from $m = Np$. Thus, we can simplify the dominator by taking:

$$m = Np \quad ; \quad N - m = N(1-p)$$

and thus:

$$P = \exp\left(-\frac{1}{2}\frac{(Np-m)^2}{Np(1-p)}\right) \tag{16}$$

1.6 Properties of the Normal Distribution

This is outstanding! This is exactly the form of a so-called **Gaussian** or **Normal** distribution. We haven't worried overly about normalization, but a normalized Gaussian model looks like:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{17}$$

where μ is the peak of the distribution, and σ represents the standard deviation.

By comparison with the analytic form of a Gaussian, for N coin flips:

$$\mu_N = pN \tag{18}$$

$$\sigma_N = \sqrt{Np(1-p)} \tag{19}$$

Just to check, for $N = 10$ and $p = 0.6$ (the example we've been using), we get $\mu = 6$ and $\sigma = 1.55$, to within 1% of what we calculated using the exact expression.

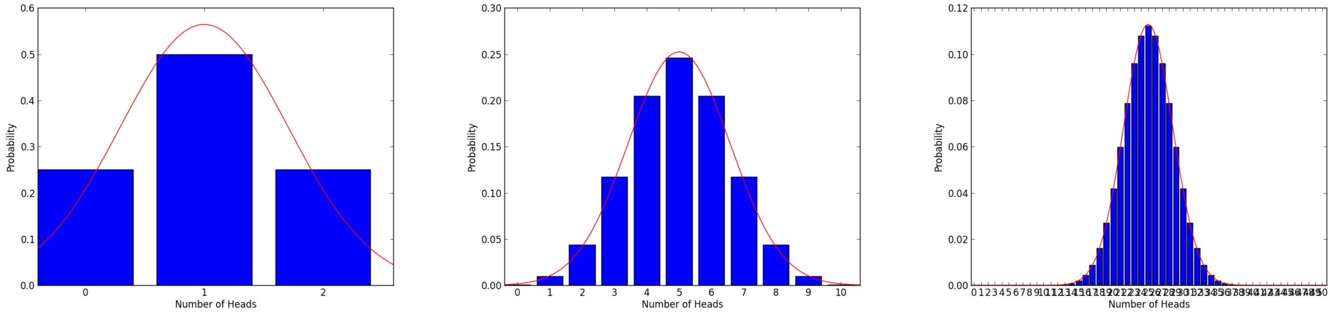

Figure 3: The exact (blue bar graph) and the large number continuous limit (red) of 2 coin flips (left), 10 flips (middle), and 50 flips (right).

Moreover, the standard deviation clearly scales as:

$$\sigma \propto \sqrt{N}$$

as I said it would.

The Gaussian has a lot of nice properties, besides the fact that binomial and other distribution functions tend to look a lot like it. For one thing, it has some well defined probability intervals.

The $1-\sigma$ range is defined as:

$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx \simeq 0.68$$

while the $2-\sigma$ range is:

$$\int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dx \simeq 0.95$$

Physicists will frequently talk about "2-sigma errorbars" meaning that there is only a 5% probability of randomly generating a number outside of that range. In the 10 coin flip example, this means that there is roughly a 5% change that we'll get as many as 10 heads or as few as 2 or less.

In observational astronomy, the significance of a signal is extremely important. If a radio telescope observes ten thousand points on the sky, then even the meager 2.5% probability of observing a $2-\sigma$ "peak" by pure chance will result in 250 false detections.

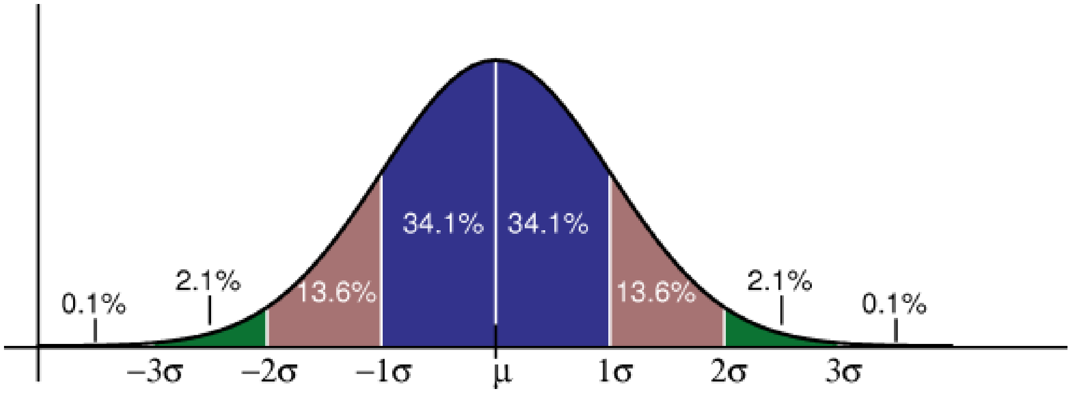

Figure 4: A Gaussian distribution. Credit: http://introcs.cs.princeton.edu/java/11gaussian/

Particle physicists frequently require $5-\sigma$ detections, at which point there is only a 1 in 2 million probability that μ is outside the detected range.

The Gaussian (or normal) distribution is useful, but there is a danger. Many young scientists assume that *all* PDFs are Gaussian, and that simply isn't the case. Look to Wall Street. Huge firms have collapsed because "unlikely" events (at the $5-\sigma$ level) turned out to occur far more frequently than expected. Why? The probability of outlier events were far more likely than Gaussian statistics would predict.

1.7 Parameter Estimation (part 1)

Normally you don't actually *know* the PDF of a variable. Part of science is *estimating* parameters of a distribution, based on random, discrete realizations.

Let's suppose that we're measuring a new physical parameter that has a true (but unknown) value, μ (that, for now could be literally any real number), and our experiment has an unknown random, Gaussian error, σ. If we run the experiment once, then we will generate a random variable via the PDF: $\mathcal{N}(\mu,\sigma)$ (which is just the shorthand for the PDF I just described).

Suppose I now make many measurements: $\{x_i\}$. How can I estimate μ? How certain am I of the answer?

The PDF describes the distribution of values that we are likely to get from a single experiment, the **random error** around the true value. In our polling example, supposing 53% of respondents support a candidate and the underlying p for that candidate is 0.55. That 2% difference is an error in my measurement; one I can't know about unless I ask more people.

1.7.1 Estimating the True Value

Let's suppose I run my experiment a single time. What is the best estimator, $\hat{\mu}$? One approach is to minimize the **residual** between my estimate and the unknown true value:

$$\begin{aligned}\langle\hat{\mu}-\mu\rangle &= \int_{-\infty}^{\infty}(\hat{\mu}(x_1)-\mu)f(x_1)dx_1\\ &= \int_{-\infty}^{\infty}\hat{\mu}(x_1)f(x_1)dx_1-\mu\end{aligned}$$

If, and only if, $\hat{\mu}=x_1$ will we get an expected residual of zero since the integral on the right hand side of the above equation will then become the *definition* of the mean, canceling the other term. We are equally likely to be high or low. This is known as an **unbiased** estimator.

It's a fairly straightforward exercise to show that if we have many measurements, an unbiased estimate must take the form:

$$\hat{\mu} = \sum w_i x_i$$

where

$$\sum w_i = 1$$

How can we figure out what weighting function, w_i is best? And how good of a result will it be? One way to figure out the size of a typical residual is to take the sum of the square of the residuals:

$$\begin{aligned}\langle (\hat{\mu} - \mu)^2 \rangle &= \langle \hat{\mu}^2 - 2\hat{\mu}\mu + \mu^2 \rangle \\ &= \langle \hat{\mu}^2 \rangle - 2\mu\langle \hat{\mu} \rangle + \mu^2 \\ &= \langle \hat{\mu}^2 \rangle - \mu^2\end{aligned}$$

where I've exploited the fact that the expectation value of a constant is just the constant itself.

How can we minimize this function? Let's not make it too complicated. Let's suppose we're estimating from two values from two runs of an experiment, and thus: $w_2 = 1 - w_1$. Thus:

$$\hat{\mu} = w_1 x_1 + (1 - w_1) x_2$$

so

$$\langle \hat{\mu}^2 \rangle = w_1^2 \langle x_1^2 \rangle + 2w_1(1 - w_1)\langle x_1 x_2 \rangle + (1 - w_1)^2 \langle x_2^2 \rangle$$

Since x_1 and x_2 are independent:

$$\langle x_1 x_2 \rangle = \langle x_1 \rangle \langle x_2 \rangle = \mu^2$$

where the last part of the relation comes from the fact that both are drawn from a distribution of mean, μ.

Likewise, as we've shown (generally; not just for Gaussians):

$$\sigma^2 = \langle x^2 \rangle - \mu^2$$

so

$$\langle x^2 \rangle = \sigma^2 + \mu^2$$

So:

$$\begin{aligned}\langle \hat{\mu}^2 \rangle &= w_1^2(\sigma^2 + \mu^2) + 2w_1(1 - w_1)\mu^2 + (1 - w_1)^2(\sigma^2 + \mu^2) \\ &= (w_1^2 + 1 - 2w_1 + w_1^2)\sigma^2 + (w_1^2 + 2w_1 - 2w_1^2 + 1 - 2w_1 + w_1^2)\mu^2 \\ &= (2w_1^2 - 2w_1 + 1)\sigma^2 + \mu^2\end{aligned}$$

How do we minimize the variance? By taking the derivative of this expression with respect to w_1 and setting it to 0.

$$\frac{d(2w_1^2 - 2w_1 + 1)}{dw_1} = 4w_1 - 2 = 0$$

No surprise, $w_1 = 1/2$.

We can compute this in general and find that for fixed errorbars (important!) the best way to estimate the true value, μ is:

$$\hat{\mu} = \frac{1}{N} \sum_i x_i \tag{20}$$

which is just the mean of the measured value.

What about the error in the estimate

$$\sigma_{\hat{\mu}} = \sqrt{\langle (\hat{\mu} - \mu)^2 \rangle}$$

?

Computing:

$$\begin{aligned}
\langle(\hat{\mu}-\mu)^2\rangle &= \langle\hat{\mu}^2\rangle - \mu^2 \\
&= \frac{1}{N^2}\left(\sum_i \langle x_i^2\rangle + 2\sum_i\sum_{j>i}\langle x_i x_j\rangle\right) - \mu^2 \\
&= \frac{1}{N^2}\left(N\left[\sigma^2+\mu^2\right] + N(N-1)\mu^2\right) - \mu^2 \\
&= \frac{1}{N^2}\left(N\sigma^2 + N\mu^2 + N^2\mu^2 - N\mu^2\right) - \mu^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

Thus, the uncertainty in the mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{N}} \tag{21}$$

Root N once again!

1.7.2 Estimating the Uncertainty

We're not quite done. While we know taking many measurements of a parameter will reduce the uncertainty of the final parameter, we don't know the actual value of the uncertainty unless we know σ ahead of time (which we may know under some circumstances).

$$\hat{\sigma}^2 \equiv s^2 = \frac{1}{N-1}\sum_i (x_i - \hat{\mu})^2 \tag{22}$$

Obviously, I haven't derived this, though the approach is similar to deriving the estimator for the mean. The only thing that should strike you is the $N-1$ in the denominator rather than N (as with the mean). Why? Because if $N = 1$, you don't have enough information to estimate the standard deviation at all.

1.7.3 Example: Numbers from a Normal Distribution

Let's do an example, but I'm going to keep things simple. Suppose we're trying to experimentally measure a parameter that has a true value of $\mu = 0$, with an (also unknown) experimental error of $\sigma = 1$. This was generated by a computer, so technically, the outcome is **pseudorandom**.

There are a number of ways that computers can use to generate pseudorandom numbers, but most of them are based on something along the lines of:

$$Y_{n+1} = mY_n \text{mod } n$$

where n is a large prime number, and m is a much larger number for which n is not a factor. Appropriately chosen, the number:

$$y = \frac{Y_n}{n}$$

will be a uniform deviate from 0 to 1. If you know how to invert the cumulative distribution function:

$$F(x) = y$$

you can generate numbers drawn from any "random" PDF that you like.

Important Warning: This is not really a way to get random numbers! If you always select the same initial seed, Y_0, then you will always run through the same series of "random" numbers. Some algorithms choose the seed by doing something like looking at the computer clock, but you should *know* where this comes from before you do random calculations.

After 10 trials, I get the following measurements

$$[-0.181, -2.608, -0.857, 0.761, 0.848, 0.616, -1.669, -1.147, -0.477, 0.889]$$

Adding everything together, we get an estimate of the mean as:

$$\hat{\mu} = -0.38$$

It may seem like we're way off the true value (or you may be surprised that more than half of the terms are negative), but you shouldn't be surprised by the value.

The expected error in the mean is:

$$\sigma_{\hat{\mu}} = \frac{1}{\sqrt{10}} = 0.32$$

You should expect that you will often differ from the expected value by a standard deviation.

By the way, estimating the standard deviation of the actual distribution:

$$s = \hat{\sigma} = 1.2$$

which again, is close, but not exact.

Try to verify this on your own!

1.8 Poisson Statistics

Not every distribution is Gaussian. I've already mentioned the uniform distribution function, but for time-series analyses, often the **Poisson Distribution** shows up.

These processes include things like radioactive decay, the rate of photons striking a detector (and thus the bias in CCDs), and other measures of discrete events. During a storm, lightning may strike every minute or so, but there's no saying when exactly, the next strike will hit (see, e.g. `Back to the Future`).

The Poisson distribution arises when there is a probability of a discrete event occurring over a short period of time:

$$\frac{dt}{\tau}$$

where τ is a characteristic time for the process, the inverse rate at which we expect the event to occur.

We may want to calculate the probability, P_0, that no events have occurred (no lightning strikes) after time, t. The subscript refers to the number of events, and since we are talking about the probability of no events at all, the subscript is zero. Thus, we have the limits of $p_0(0) = 1$, since we can say for certain that no events have yet occurred. Moreover, we may immediately write the differential equation:

$$\frac{dp_0}{dt} = -p_0 \frac{1}{\tau}$$

since a *first* event occurring during some time is proportional to the probability that no events have yet occurred.

This is easy to solve:

$$P_0(t) = \exp(-t/\tau) \tag{23}$$

where the subscript "0" represents the probability that there are no events at all.

But what if we want to consider the probability of *exactly 1* lightning strike having occurred after time, t? We then have the relation:

$$\frac{dp_1}{dt} = p_0 \frac{1}{\tau} - p_1 \frac{1}{\tau}$$

We subtract the probability of getting a second strike after already having 1 from the probability of getting a first strike in that interval.

More generally:

$$\frac{dp_n}{dt} = \frac{1}{\tau}(p_{n-1} - p_n) \tag{24}$$

This can be solved with a fun trick. Consider:

$$\frac{d}{dt}\left(p_n \exp(t/\tau)\right) = \frac{dp_n}{dt}\exp(t/\tau) + \frac{1}{\tau} p_n \exp(t/\tau)$$

Substituting this into equation 24, we see:

$$\frac{d}{dt}\left(p_n \exp(t/\tau)\right) = \frac{p_{n-1}}{\tau}\exp(t/\tau)$$

For $n = 1$, we have $p_{n-1} = \exp(t/\tau)$, so:

$$\frac{d}{dt}\left(p_1 \exp(t/\tau)\right) = \frac{1}{\tau}$$

and thus:

$$p_1 = \frac{t}{\tau}\exp(-t/\tau)$$

The pattern, more generally, yields:

$$p_n(t) = \left(\frac{t}{\tau}\right)^n \frac{1}{n!}\exp(-t/\tau) \tag{25}$$

Remember, this is the probability of getting *exactly* n lightning strikes (or photons counted, or nuclear decays) after a time t.

Plotting this out, we see something interesting:

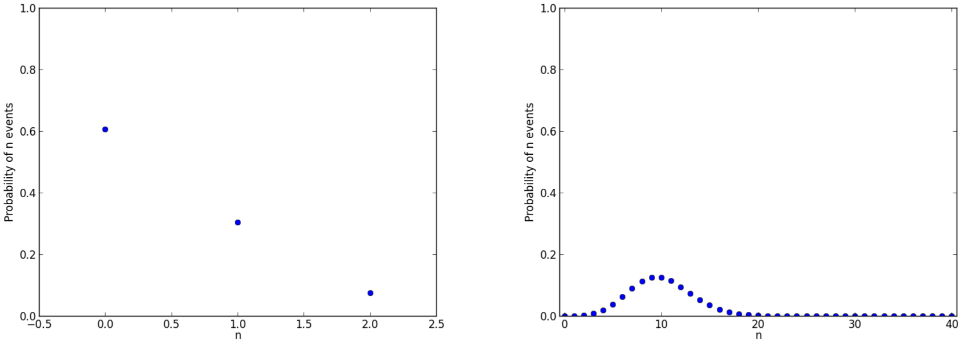

Figure 5: The probability of n events after 0.5τ (left panel), and 10τ (right panel). The latter looks a lot like a Gaussian.

Wait long enough and *everything* looks like a Gaussian.[4]

[4] This is a result known as the "Central Limit Theorem," and as written it's a bit of an exaggeration. The scatter in the distribution needs to be well-defined and definite, but let's hold off on those concerns for now.

After any given time, t, we can compute various quantities in (what is now) the normal way. I'm going to go out on a limb and *define*:

$$\mu \equiv \frac{t}{\tau}$$

which yields:

$$\begin{aligned}
\langle n \rangle &= \sum_n p_n(t) n \\
&= \sum_{n=0}^{\infty} n \left(\frac{t}{\tau}\right)^n \frac{1}{n!} \exp(-t/\tau) \\
&= \exp(-\mu) \sum_{n=0}^{\infty} \mu^n \frac{1}{(n-1)!} \\
&= \exp(-\mu) \sum \mu \frac{\mu^{n-1}}{(n-1)!} \\
&= \mu = \frac{t}{\tau}
\end{aligned}$$

Since the last sum is simply the series expansion of $\exp(\mu)$.

Thus, for a Poisson distribution:

$$\langle n \rangle = \frac{t}{\tau} \tag{26}$$

as you might guess.

More interestingly:

$$\langle (n-\mu)^2 \rangle = \frac{t}{\tau} = \mu \tag{27}$$

and thus.

$$\sigma_n = \sqrt{n} \tag{28}$$

Square root of n again!

1.8.1 Example: Photon Noise

Suppose we are observing a galaxy image which (for simplicity), produces detector counts of $r = 2s^{-1}$ within some aperture, *on average*. Further, suppose the sky background produces $b = 20s^{-1}$. We're throwing a lot of observational uncertainty under the rug, but this model will work for now.

How long do we need to observe?

First, note that the number of photons striking our detector will be given by a Poisson distribution with a mean of:

$$\mu = (r+b)t$$

The longer we observe, the more photons we collect. However, the expected *signal* is given by:

$$S = \mu - bt = rt$$

The *noise* comes from both the sky and the galaxy:

$$\sigma = \sqrt{(r+b)t}$$

and thus the so called "signal to noise" (or S/N) is:

$$\begin{aligned}
\frac{S}{\sigma} &= \frac{rt}{\sqrt{(r+b)t}} \\
&= \frac{r}{\sqrt{r+b}} \sqrt{t}
\end{aligned}$$

The longer we observe, the greater the signal-to-noise. In this case, a 1 second observation yields a S/N ratio of 0.43 – not a detection. It's not until about 5.5 seconds that we have a S/N of 1, meaning that the observed galaxy rises above the noise about 16% of the time. After 22 seconds, we have a S/N ratio of 2, and so on.

1.9 The Algebra of Random Variables

We've talked a fair amount about the probability distribution function generated by a single variable, but if we're going to do multiple measurements of the same system (for instance), then it is worth considering how a joint probability works. For instance, consider the relation:

$$z = x + y$$

where both x and y are random variables with their own distinct probability distribution functions, $f_x(x)$ and $f_y(y)$. How do we compute $f_z(z)$? Conceptually, it's nothing more than the product of the two probabilities, but we need to integrate over all possible combinations of x and y. Mathematically:

$$z = x + y \rightarrow f_z(z) = \int f_x(x) f_y(y = z - x) dx \tag{29}$$

and likewise for any combination of two or more variables. For instance:

$$z = xy \rightarrow f_z(z) = \int f_x(x) f_y(y = z/x) dx$$

Combining three variables produces a double integral and so forth; in practice you can simply combine them two at a time.

To make things concrete, let's do a couple of examples.

1.9.1 Example: Sum of Two Uniformly Distributed Functions

Consider two variables, x and y, each uniformly distributed in the range $[0, 1]$. Their sum, z is necessarily distributed in the range, $[0, 2]$, but the distribution won't be uniform.

The integral over x needs to be defined in a piecemeal fashion simply because both x and y need to be within $[0, 1]$:

$$f(z) = \int_{\max(0,z-1)}^{\min(1,2-z)} f_x(x) f_y(y = z - x) dx$$

In that range, both PDFs are simply 1, simplifying the expression dramatically, and producing:

$$z = x + y \rightarrow f(z) = \begin{cases} z, & \text{if } z < 1. \\ 2 - z, & \text{if } z \geq 1. \end{cases} \tag{30}$$

This is illustrated in Fig. 6.

We could compute all sorts of properties of $f_x(x)$ and $f_z(z)$. For isntance, it is fairly easy to show that:

$$\mu_x = \mu_y = 0.5$$

and

$$\mu_z = 1$$

which is also the median ($F_z(z) = 0.5$) and mode (peak).

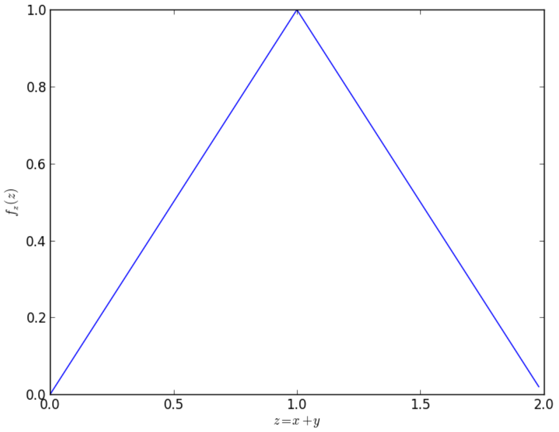

Figure 6: The probability distribution function (PDF) of the sum of two uniform variables, each defined over $[0, 1]$.

Computing the standard deviation, we find:

$$\begin{aligned}\sigma_x^2 = \sigma_y^2 &= \int_0^1 f_x(x)(x-\mu_x)^2 dx \\ &= \int_0^1 (x-0.5)^2 dx \\ &= \frac{1}{3}(x-0.5)^3\Big|_0^1 \\ &= \frac{1}{12}\end{aligned}$$

I won't bore you with the details, but plugging in the form for z we get:

$$\sigma_z^2 = \frac{1}{6}$$

and thus:

$$\sigma_z = \sqrt{2}\sigma_x$$

Adding random deviates produces a function with a larger standard deviation than the originals (by $\sqrt{2}$, of course), but if we subsequently *divide* by 2 to take the average, the standard deviation of the mean *drops* by $1/\sqrt{2}$.

We could imagine adding up lots and lots of uniform variables (and dividing by the number of values to get the mean). In practice, this would involve ever higher power law series with more and more complex piecemeal definitions. I've done the exercise for the average of 5 uniform deviates in Fig. 7. This distribution starts to look startlingly like a Gaussian.

1.9.2 Example: Sum of Two Gaussians

The sums of Gaussians are even easier and don't need to be defined piecemeal, since the limits of integration are $[-\infty, \infty]$. Thus:

$$z = x + y \rightarrow f_z(z) = \frac{1}{2\pi\sigma_x\sigma_y}\int dx \exp\left(-\frac{(x-\mu_x)^2}{2\sigma^2}\right)\exp\left(-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}\right)$$

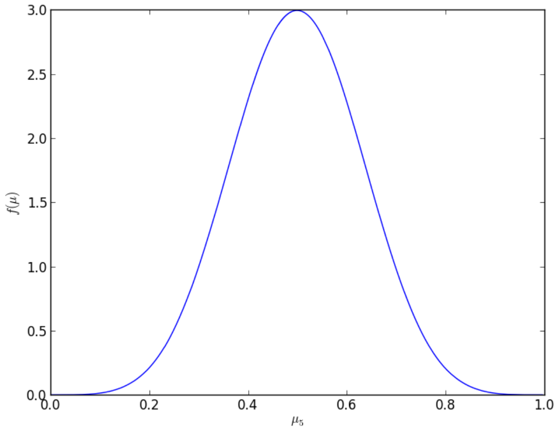

Figure 7: The average of 5 uniform deviates defined over $[0, 1]$.

which looks ugly, but simplifies beautifully:

$$z = x + y \rightarrow f_z(z) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_x^2 + \sigma_y^2}} \exp\left(-\frac{(z - (\mu_x + \mu_y))^2}{2(\sigma_x^2 + \sigma_y^2)}\right) \tag{31}$$

Put another way, the sum of two Gaussians produces a Gaussian with:

$$\mu_z = \mu_x + \mu_y \tag{32}$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 \tag{33}$$

The standard deviation adds in quadrature, so adding two terms with equal σ's produces a sum with a standard deviation $\sqrt{2}$ larger than the original. The average of the two terms produces a standard deviation $1/\sqrt{2}$ as large as the standard deviation in the originals. After this, I'll try not to keep pointing out how frequently $\sqrt{n}$ shows up, but it's worth noting that we've proven it formally for Gaussians and uniform distributions.

1.9.3 Some Rules of Thumb

We are starting to see that lots of distributions end up resembling Gaussians if you either wait long enough, or if you combine enough data. As a result, we can start to exploit many of the relations as found from Gaussians.

Starting simple, consider a number (or set of numbers), x drawn from a distribution with an expectation, μ, and a Gaussian noise, σ. Now, consider a function:

$$z = x + b$$

where b is known with certainty. Under those circumstances, it's clear that the uncertainty in y is simply:

$$z = x + b \ \rightarrow \ \sigma_z = \sigma_x \tag{34}$$

Of course, it is often the case that a monotonic function is a bit more complicated. For instance:

$$z = Cx$$

In this case, we can start to get a sense of the errors by noting:

$$dz = Cdx$$

and thus:

$$z = Cx \;\rightarrow\; \sigma_z = C\sigma_x \tag{35}$$

We could even extend this further, though with the understanding the resulting errors are not necessarily Gaussian. For instance, consider the power law relation:

$$z = x^n$$

Under those circumstances:

$$dz = nx^{n-1}dx$$

or

$$\frac{dz}{z} = n\frac{dx}{x}$$

or, roughly speaking:

$$z = x^n \;\rightarrow\; \sigma_z = n\frac{\mu_z}{\mu_x}\sigma_x \tag{36}$$

but as noted above, the errors won't necessarily be Gaussian.

We've already seen that for the sum of two independent numbers, x and y, each drawn from a Gaussian, we get:

$$z = x + y \;\rightarrow\; \sigma_z^2 = \sigma_x^2 + \sigma_y^2 \tag{37}$$

And similarly:

$$z = x - y \;\rightarrow\; \sigma_z^2 = \sigma_x^2 + \sigma_y^2 \tag{38}$$

Doesn't matter whether you add or subtract. The error propagates the same way. This is why subtracting two large (random) numbers from one another can be so noisy.

2 Conditional Probabilities

2.1 Why priors are necessary

Thus far, we've focused on the idea that some random process is governed by a discrete probability distribution or PDF, $f(x)$, and we've tried to figure out the frequency of outcomes and their statistical properties. We've touched only briefly on model fitting and inference. Inference and random realization are closely related, but the relation is a nuanced one.

Let's consider a very simple, and by now very familiar, example: coin flips.

Suppose I assume that a coin is fair, and thus the probability, p of heads is 0.5. I then flip once, twice, 10 times, 20 times, and get heads every time. Our previous work allows me to make the following statement:

> *Supposing the coin is fair, there is a 1 in a million probability of getting a string of 20 heads in a row in one go.*

Does that mean the coin isn't fair? Not necessarily. You could imagine scenarios where I have other reasons to believe that the coin is fair (I may have done other tests, or had the assurance of a numismatist), and I simply believe the other evidence has much less than a 1 in a million chance of misleading me.

On the other hand, the situation is quite different if the coin flipping is done by a shady carnival barker.

In other words, our conclusions about whether or not the coin is fair is based largely on our **a priori** assumptions about the probability. This is the foundation of **Bayes' Theorem**, named after Reverend Thomas Bayes.

2.2 Bayes' Theorem by Example

Let's keep with the coin-flip example and consider the following scenario: I have a jar of 100 coins. 1 of those coins has two heads, and the other 99 are perfectly normal, and evenly balanced. I then draw a coin at random from the jar and flip it. What is the probability of getting heads?

To aid us, I'm going to define two different processes, each with a "true" and a "false."

- A: We flip heads
- B: We choose the two headed coin from the jar.

We know a great deal about this process already. For instance, we know that:

$$P(B) = 0.01$$

by definition.

We also know something about the relationship of these two probabilities. For instance, we know:

$$P(A|B) = 1$$

This expression reads as "The probability of A *given* B," and in our case, it simply means that provided we chose the two-headed coin, we necessarily get a heads when we flip.

We can also say:

$$P(A|\neg B) = 0.5$$

where "$\neg B$" (read "not B") is the condition that we *didn't* chose the two-headed coin. Our probability of flipping heads is thus 50%.

We can even compute the joint probability, $A \cap B$, which is simply the probability that both A *and* B are true. What is the probability that we both chose the 2-headed coin *and* we flip heads? Well, the former is 1% and the latter is 100% provided that we've already chosen the 2-headed coin.

Or in other words:

$$P(A \cap B) = P(A|B)P(B) = 0.01$$

We only need one more step to actually *derive* Bayes' Theorem, and that's to realize that the joint probability distribution doesn't care which term you call "A" and which term you call "B":

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{39}$$

This is a very powerful relation. It allows us to create **conditional probability tables** for any situation. For instance, suppose we want to compute:

$$\begin{aligned} P(A \cap \neg B) &= P(A|\neg B) \cdot P(\neg B) \\ &= 0.5 \cdot 0.99 \\ &= 0.495 \end{aligned}$$

or

$$\begin{aligned} P(\neg A \cap \neg B) &= P(\neg A|\neg B)P(\neg B) \\ &= 0.5 \cdot 0.99 \\ &= 0.495 \end{aligned}$$

For our coin-flip example, we have:

	A	¬A
B	0.01	0
¬B	0.495	0.495

If you add up all of the cells, you should find that the total probability of something happening ($P(A \cap B) + P(A \cap \neg B) + P(\neg A \cap B) + P(\neg A \cap \neg B)$) adds up to exactly 1.

What's more, we can figure out the overall probability of flipping heads:

$$P(A) = P(A \cap B) + P(A \cap \neg B) = 0.505$$

A bit more than half, as you might have expected.

2.3 Bayes' Theorem

Bayes' Theorem itself comes from exploiting the relationships between the various conditional probabilities. We can re-write equation 39 to get Bayes' theorem immediately:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{40}$$

where the A's and B's can be swapped or replaced by their converses.

In our coin example, we know $P(A|B)$ (It's simple, and given by the problem), but not $P(B|A)$ (In English, "What is the probability that you've picked the 2-headed coin given that you just flipped heads?").

We can use:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{1 \cdot 0.01}{0.505} \simeq 0.0198$$

Flipping a coin and finding heads (under these specific circumstances) implies a 2% **posterior probability** that you've picked the two-headed coin to begin with.

You may note that **Bayesian Statistics** appears to be counter to the frequentist approach that we started with. The frequentist approach, you'll recall, was based on imagining running an experiment an infinite number of times to converge on a true probability (or model). The Bayesian approach, on the other hand, gives a way of determining a probability for a particular model given the observations to date. You can then update the prior probability once you've run additional experiments.

2.3.1 Example: The Higgs Boson

You may have heard of a pretty significant discovery in the last few years: the Higgs Boson. One of the problems with "discovering" a new particle is that there isn't a single unambiguous event that lands on the PI's desk. Instead, there are many data channels, where the signal rises significantly above the noise. This can happen by random chance, though the greater the deviation from the noise, the clearer the signal.

For instance (and I'm going to dramatically simplify things here), suppose there was an experiment studying a single channel at 125 GeV. Even if there were no Higgs (or no Higgs at that mass), there would be some background level, $\mu \pm \sigma$.

Suppose we ran the experiment and then found that the actual signal was $\mu + 2\sigma$. Have we discovered the Higgs?

As we saw earlier, the odds of getting a $+2\sigma$ measurement (or more) are roughly 2.5% (the plus and minus side are symmetric), sometimes known as the **p-value** (the probability of getting as extreme a case as we've gotten under the assumption that there is no true signal). But even so, we can't say:

> *We are* 97.5% *certain that we've discovered the Higgs.*

Instead, we have to say:

> *If there is no Higgs, there is only a* 2.5% *chance of getting a signal at this level by pure chance.*

Those two statements seem similar, but they aren't. The difference depends entirely on your priors, how likely you thought it was that the Higgs would show up at this particular mass, and that it existed in the first place.

Suppose, optimistically, that prior to running the experiment, you thought that there was a 10% chance that the Higgs both existed and that it had a mass of 125 GeV. Let's suppose further that *if* the Higgs exists at this mass, we'll always get a "detection" (not necessarily the case in reality). On the other hand, if the Higgs doesn't exist, we'll only get a detection 2.5% of the time.

In other words, we have the following propositions:

- A: The Higgs exists
- B: We get a detection

We've already specified that:

$$P(A) = 0.1$$

and that

$$P(B|A) = 1$$

The only complication is computing $P(B)$, the probability of getting a detection. It is:

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\neg A)P(\neg A) \\ &= 1 \cdot 0.1 + 0.025 \cdot 0.9 \\ &= .122 \end{aligned}$$

Either the Higgs exists or it doesn't, and we know the probability of getting a detection either way.

Thus:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{1 \cdot 0.1}{0.122} \\ &= 80\% \end{aligned}$$

An 80% probability that the Higgs exists, rather than 97.5%

In reality, though, this is a huge overestimate. There were hundreds of possible channels, and the very real possibility that the Higgs doesn't exist at all. We might conclude that a better prior probability is something like $P(A) \sim 10^{-4}$ (just a guess for illustration purposes). Re-run with those numbers, and you'd conclude that there's 0.4% probability that the signal represents a true Higgs detection.

It's for that reason that particle physicists tend to demand 5σ detections for announcing a discovery. The odds of getting a 5σ signal by shear chance is about 1 part in a million (the p-value), significantly lower than the prior probability of a Higgs.

But again, setting these priors requires intelligent guesswork. There is no ironclad way to make the initial assessment.

2.4 False Positives and False Negatives

In the discussion of the Higgs discovery, and in many science experiments, drug treatments, and surveys, there is an approximate relationship between one test and another. In the case of the Higgs, we could write it as:

$$B \to A$$

Or, in plain English, "*If* we get a positive detection, *then* the Higgs exists at the relevant mass."

But as we've seen, this causal relationship doesn't hold. We might get a positive detection, but it could be noise; there may be no Higgs. On the other hand (though I set the probability of this as zero in my particular example), we might have gotten a negative detection despite the existence of a Higgs.

In the literature, these two sources of error have names.

1. **False Negatives** (or "Type I errors," as they are sometimes called), in which the test gives a negative despite the true condition being positive. The probability of a false negative can be written as:

   $$P(A|\neg B)$$

   For the Higgs example, this is the probability that there is a Higgs given that we get a negative detection.

2. **False Positives** (or "Type II errors"). This is the reverse of the previous. We get a positive outcome from a test, despite the fact that the underlying condition is negative. Mathematically, the probability is:

   $$P(\neg A|B)$$

2.4.1 Example: Disease Screening

Consider a rare, but deadly disease that affects 1 person in a million. In a routine screening (that is, the screening itself is independent of whether you have the disease), a diagnostic test reveals that you have the disease.

Should you be concerned?

That depends, of course, on the nature of the test. Suppose that the test reveals a positive for 99% of sick patients. Under those conditions it would be marketed (honestly) as "99% accurate."

However, suppose further that 1 healthy patient in 100 gets a positive reading on the diagnostic test. This, too, might be marketed as a 99% accurate test, but consider the consequences.

To do so, let's first define our terms mathematically:

- A: You have the disease
- B: You get a positive reading on the screening.

We know:

1. $P(A) = 10^{-6}$
2. $P(B|A) = 0.99$
3. $P(B|\neg A) = 0.01$

So, for example:

$$\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)} \\
&= \frac{0.99 \cdot 10^{-6}}{0.99 \cdot 10^{-6} + 0.01 \cdot 0.999999} \\
&= 0.0001
\end{aligned}$$

The probability that you have the disease *given that you got a positive reading* is only 0.01%! Thus, the probability of a false positive is 99.99%. This is just a consequence of the disease itself being so rare.

On the other hand, false negatives are only 1% likely. We don't need to do any additional calculations. We were just given that to start.

2.5 Continuous Probabilities

So far we've only concentrated on discrete probabilities with a binomial condition. Either A is true or false. But there's nothing preventing us from considering the joint probability of many or even a continuous distribution.

Consider two random variables, x and y, each with their own probability distribution function, $f_X(x)$ and $f_Y(y)$. We can write, with no additional work:

$$f_X(x|Y = y) = \frac{f_Y(y|X = x)f_X(x)}{f_Y(y)} \tag{41}$$

or, supposing that one probability is discrete, $P(A)$, and the other is continuous, $f(x)$,

$$f(x|A) = \frac{P(A|x)f(x)}{P(A)} \tag{42}$$

2.5.1 Example: Coin Flipping Revisited

Let's revisit the issue of determining whether a coin is fair or not. In one extreme case, we might give a prior probability distribution of:

$$f(p) = \delta(p - 0.5)$$

which is a fancy way of saying that we believe that a coin is fair and no amount of data will ever persuade us otherwise.

In the other extreme, we may consider ourselves totally agnostic. We might suppose:

$$f(p) = 1$$

We have no idea, prior to flipping, whether the coin is weighted, and if so how. Indeed, we think all probabilities are equally likely.

I flip a coin once and it comes up heads. We call that event "A". What is the posterior probability distribution of the weighting of the coin?

To solve that, we note:

$$P(A|p) = p$$

Since $P(A)$ is the probability of flipping a heads, and p is the weighting of the coins under consideration, the relationship is tautological. The probability of getting heads is, by definition, p.

Further, the total probability:

$$\begin{aligned} P(A) &= \int_0^1 P(A|p)f(p)dp \\ &= \int_0^1 p \cdot 1 \cdot dp \\ &= \left.\frac{p^2}{2}\right|_0^1 \\ &= \frac{1}{2} \end{aligned}$$

As you might have expected. Thus:

$$f(p|A) = 2p$$

Suppose I flipped twice, and got 1 head, 1 tails. This is event A. What is the posterior probability distribution of p in this case?

Again, we need

$$P(A|p) = 2p(1 - p)$$

which we've computed before. Likewise, we need to know $P(A)$:

$$\begin{aligned} P(A) &= \int_0^1 P(A|p)f(p)dp \\ &= 2\int_0^1 p(1-p)\cdot 1\cdot dp \\ &= 2\left(\frac{p^2}{2}-\frac{p^3}{3}\right)\Big|_0^1 \\ &= \frac{1}{3} \end{aligned}$$

So:

$$P(p|A) = 6p(1-p)$$

Graphically, we can illustrate our posterior PDF after a particular series:

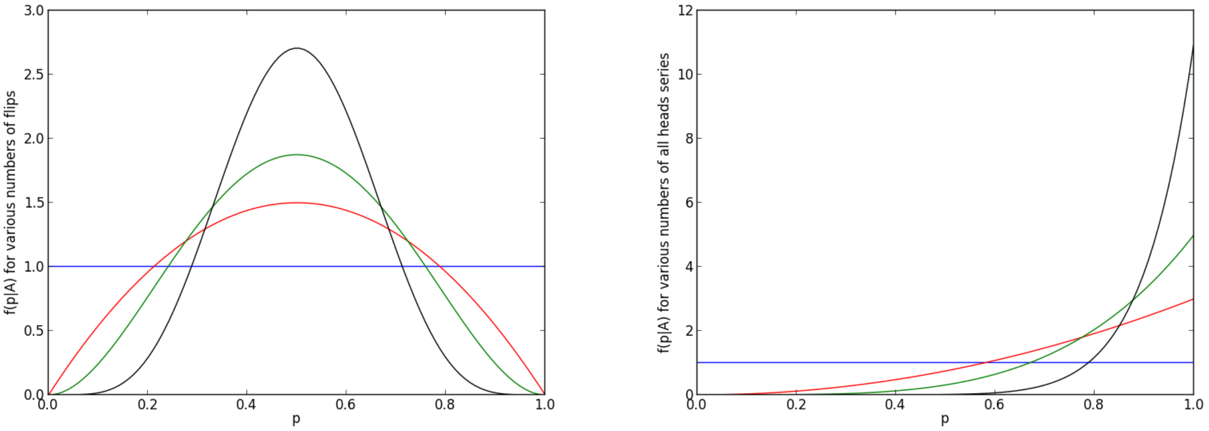

Figure 8: The posterior probability distribution after 2 (red) ,4 (green) and 10 flips (black) of a coin. The left panel indicates the PDF if the observed heads and tails are split evenly, while the right panel is for the case where all heads come up. These results depend sensitively on your prior probability (shown in blue).

You should use Bayesian statistics with caution. You will get a very different result depending on your Bayesian prior. You can easily imagine running into trouble if either a) You assume that you know a value perfectly, or b) you assume that you know so little that you set your prior to be uniform over an infinite range.

No matter what you assume, however, there's an important truth to all of this. No matter how many experiments you perform, the width of your posterior PDF (or equivalently, the probability of a Type II error) will get smaller and smaller, but will never vanish entirely. Put another way, while we've always measured gravity to point downwards on earth, there is always the possibility that we've just had an exceptionally lucky run and that one day gravity could reverse itself. In other words, we can't *prove* a physical theory, we can only **falsify** it by showing that it's contradicted in at least one case.

3 Model Fitting

3.1 The Likelihood Function

We're finally ready to start fitting models and quantifying our certainty about those fits. For the moment, we're going to make the following assumptions:

1. The data, $\{x_i\}$ is drawn from a Gaussian random distribution with mean, μ, and standard deviation, σ. Put another way, there is a true value μ, with a measurement error, σ.
2. When performing the experiment, we don't know anything about the true value, μ, and thus the prior probability is:

   $$f(\mu) = const$$

   over the range $[-\infty, \infty]$.

For the moment, let's consider a measurement of only a single value. The posterior probability of a particular model value, $\hat{\mu}$ is:

$$f(\hat{\mu}|x_1) = \frac{f(x_1|\hat{\mu})f(\hat{\mu})}{f(x_1)}$$

Since the prior probability distribution function is a constant, $f(x_1) = const$ as well. Thus, for a single measurement:

$$f(\hat{\mu}|x_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_1-\hat{\mu})^2}{2\sigma^2}\right)$$

This is known as the **Likelihood Function**, $L(\hat{\mu})$, which is maximized for the best fit model.

For two or more measurements, the likelihood function is simply the product of the individual likelihoods. This is just a consequence of the probability of independent events:

$$L(\hat{\mu}) = \frac{1}{(2\pi)^{N/2}\sigma^N} \prod \exp\left(-\frac{(x_i-\hat{\mu})^2}{2\sigma^2}\right) \tag{43}$$

This generalizes even more. The errors for the various measurements need not be identical:

$$L(\hat{\mu}) = \frac{1}{(2\pi)^{N/2}} \prod \frac{1}{\sigma_i} \exp\left(-\frac{(x_i-\hat{\mu})^2}{2\sigma_i^2}\right) \tag{44}$$

3.1.1 Maximum Likelihood

As written, the Likelihood function is symmetric around the maximum likelihood value. Thus, the value of $\hat{\mu}$ that maximizes the likelihood is also the best estimate for μ:

$$\langle\hat{\mu}\rangle = \mu$$

If we're trying to maximize the likelihood, then finding the $\hat{\mu}$ that minimizes the *negative log* of the likelihood function will accomplish the same thing:

$$\mathcal{L} = -\ln(L) = \frac{N}{2}\ln(2\pi) + \sum_i \ln(\sigma_i) + \frac{1}{2}\sum \frac{(x_i-\hat{\mu})^2}{\sigma_i^2}$$

Since $\hat{\mu}$ only appears in the last term, we can maximize the likelihood by minimizing:

$$X^2 = \sum_i \frac{(x_i-\hat{\mu})^2}{\sigma_i^2} \tag{45}$$

This is known as “Chi-squared minimization.”

A note of caution: I should note that while people use X^2 minimization (or equivalently, the form of the likelihood function above) under just about every circumstance imaginable, it’s only strictly applicable under the assumptions that the prior is uniform over an infinite range and that the errors are Gaussian. They aren’t necessarily.

3.2 Minimizing X^2

Suppose you have a parameter that you’re trying to estimate. As a result, we try to minimize X^2. For a simple function, this is fairly straightforward. Minimizing a function involves simply taking the derivative of the parameter:

$$\frac{dX^2}{d\hat{\mu}} = 2\sum_i \frac{(x_i-\hat{\mu})}{\sigma_i^2}$$

and setting to zero, which yields:

$$\begin{aligned}\sum_i \frac{\hat{\mu}}{\sigma_i^2} &= \sum_i \frac{{x_i}^2}{\sigma_i} \\ \hat{\mu} &= \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}\end{aligned}$$

I’ll leave it as an exercise to show that in the limit of every error being equal, the best estimate for the mean is simply the average of the values. Otherwise, we get the relation:

$$w_i = \frac{1/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{46}$$

Unsurprisingly, greatest weight is given to datapoints with the smallest error.

3.2.1 Reduced χ^2

The X^2 function has some interesting properties. For instance, suppose we knew ahead of time what the true value of μ was. In that case, we’d expect:

$$\begin{aligned}\langle X^2\rangle &= \sum_i \frac{\langle (x_i-\mu)2\rangle}{\sigma_i^2} \\ &= \sum_i \frac{\sigma_i^2}{\sigma_i^2} \\ &= N\end{aligned}$$

In other words, under all of the caveats that I listed before, a good model should produce a X^2 of about N.

In fact, it’s a little more complicated than that. After all, we *don’t* know the true mean of the distribution. Suppose we only have 1 datapoint. By definition, $\hat{\mu} = x_1$ under those circumstances, and thus the $X^2 = 0$. That doesn’t mean we have a perfect model. It simply means that we have no residuals.

I’ll shortly introduce models that have more than 1 parameter, so let’s generalize and say that we have N datapoints and $m < N$ parameters to our model. We define a **reduced chi square** as:

$$\chi^2 \equiv \frac{X^2}{N-m} \tag{47}$$

For "good fits" $\chi^2 \simeq 1$.

Very important note: Let's suppose you ran an analysis and found that $\chi^2 = 0.02$. You might congratulate yourself on getting an awesome fit as this number is much less than 1. Don't. It almost certainly means that you have dramatically overestimated your errorbars. This is very dangerous. If you overestimate your errorbars then all sorts of bad fits will be counted as good ones. Indeed, oftentimes people will scale their errorbars to produce $\chi^2 = 1$. This is illustrated in Fig. 9.

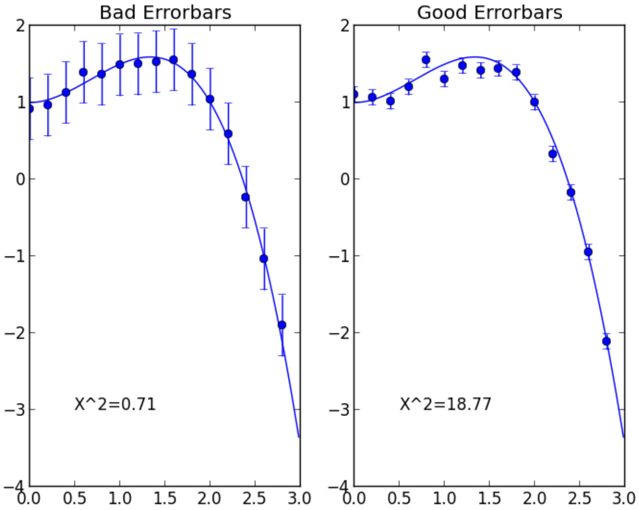

Figure 9: In the left panel, we fit a model with errorbars that are far too large. If you see something like this, you know immediately that you've overestimated. By comparison, in the right panel, the errors seem reasonable. With 15 points fit, the X^2 is approximately 19, within the range of expectations. Likewise, in the left panel, all 15 points are "within 1-σ", which is extremely unlikely. 6 of the 15 (40%) are outliers in the right panel. Remember that we *expect* around 32% of points to be outside the 1-σ range. This is a good example of a case where we can be reasonably confident of good errorbars.

I won't prove it here, but the X^2 (with a capital) distribution, can be shown to satisfy a $\mathcal{N}(N-m, \sqrt{2(N-m)})$ normal distribution for a well-constructed model. Thus, provided the errorbars are accurate, we'd expect a minimized χ^2 to have a value of:

$$\chi^2_{min} = 1 \pm \sqrt{\frac{2}{N-m}} \tag{48}$$

Much more than that and you should seriously look at your uncertainties.

Warning: All of this discussion about getting a "low" value of χ^2 is predicated on the assumption that the underlying model is a sound one. If the data is drawn from a process that will produce a sine curve and you're trying to fit it to a straight line, then your minimum χ^2 will be quite large and you should probably interpret that as an indication that your model isn't a good one.

3.3 Function Fitting

So far, I've concentrated on estimating a single number. In many situations, we're not trying to fit a number, but rather, a function with one or more parameters. For a domain, $\{x_i\}$ (measured with certainty), we might imagine some model

$$f(x; \{\theta_n\})$$

where $\{\theta_n\}$ represents some set of parameters to be determined.

Our data, $\{y_i\}$ is meant to match the function as closely as possible. Thus, we minimize:

$$X^2 = \sum_i \frac{(y_i - f(x_i; \{\theta_n\}))^2}{\sigma_i^2} \tag{49}$$

where I'll omit the explicit reliance on the parameters for now.

To minimize X^2, we simply take the derivative with respect to each parameter and set to zero. Thus:

$$\sum_i \frac{y_i \frac{df(x_i)}{d\theta_n}}{\sigma_i^2} = \sum_i \frac{f(x_i)\frac{df(x_i)}{d\theta_n}}{\sigma_i^2}$$

As it stands, this is not a terribly helpful form, but it turns out to be very simple under certain circumstances.

3.3.1 Example: Linear Regression

Consider a dataset that is drawn from a linear model of the form:

$$f(x_i) = mx_i + b$$

with an errorbar, σ_i associated with each point. For concreteness, I've generated a random dataset of exactly this form in Fig. 10.

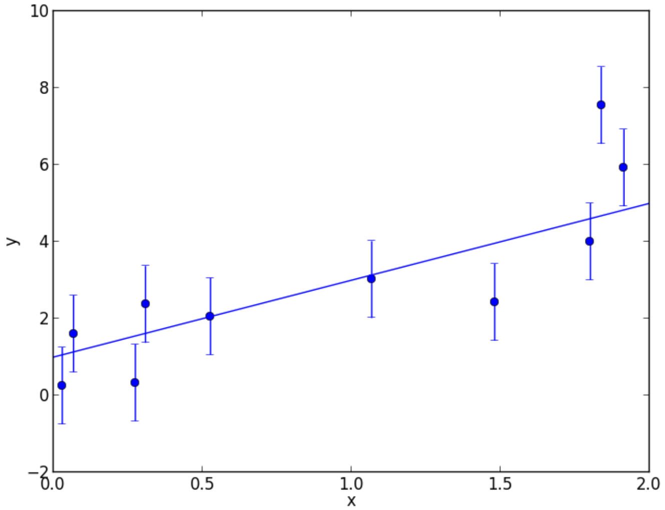

Figure 10: A randomly generated model following a $y = 2x + 1$ curve with $\sigma = 1$ Gaussian random errors. The solid line shows the true model.

In this particular case, I've chosen a slope of $m = 2$ and an intercept of $b = 1$ with a constant error associated with each point of $\sigma = 1$. You will also note that 3 (30%) of the points do not intersect the true curve at the $1 - \sigma$ level. This is as expected. For a very large sample, only 68% of points should be within $1 - \sigma$ of the true value. If you have significantly fewer or more than this, warning bells should go off.

To minimize X^2, we take the derivative of the function with respect to m and b. I'll do a slightly simplified form of linear fitting for now since the errors are constant:

$$X^2 = \frac{1}{\sigma^2}\sum_i (y_i - \hat{m}x_i - \hat{b})^2$$

Thus:

$$\begin{aligned} \frac{dX^2}{d\hat{m}} &= -\frac{1}{\sigma^2}\sum_i x_i(y_i - \hat{m}x_i - \hat{b}) \\ 0 &= \sum_i x_i y_i - \hat{m}\sum_i x_i x_i - \hat{b}\sum_i x_i \\ 0 &= S_{xy} - \hat{m}S_{xx} - \hat{b}S_x \end{aligned}$$

where I've noted that the derivative of X^2 with respect to $\hat{m}$ should be zero, and I've defined:

$$S_{xy} = \sum_i x_i y_i$$

and others similarly.

As we have 1 equation and 2 unknowns, we can't solve this uniquely. However, we have another derivative:

$$\begin{aligned} \frac{dX^2}{d\hat{b}} &= -\frac{1}{\sigma^2}\sum_i (y_i - \hat{m}x_i - \hat{b}) \\ 0 &= S_y - \hat{m}S_x - N\hat{b} \end{aligned}$$

Our constraints can be rewritten as:

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & N \end{pmatrix}\begin{pmatrix} \hat{m} \\ \hat{b} \end{pmatrix} - \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix} = 0$$

solving

$$\begin{pmatrix} \hat{m} \\ \hat{b} \end{pmatrix} = \frac{1}{NS_{xx} - S_x S_x}\begin{pmatrix} N & -S_x \\ -S_x & S_{xx} \end{pmatrix}\begin{pmatrix} S_{xy} \\ S_y \end{pmatrix} \tag{50}$$

where I've simply inverted a 2x2 matrix in the normal way. Writing it out explicitly:

$$\hat{m} = \frac{NS_{xy} - S_x S_y}{NS_{xx} - S_x S_x} \tag{51}$$

$$\hat{b} = \frac{-S_x S_{xy} + S_{xx} S_y}{NS_{xx} - S_x S_x} \tag{52}$$

It's worth noting that even if the errorbars *aren't* uniform, the same approach will work, but with:

$$S_{xy} = \sum \frac{x_i y_i}{\sigma_i^2}$$

and the other sums redefined similarly, along with:

$$N \to S = \sum_i \frac{1}{\sigma_i^2}$$

Plugging this in, we find (for our dataset):

$$\hat{m} = 2.54 \ \ ; \ \ \hat{b} = 0.58$$

We plot the best fit line in Fig. 11.

Note that the minimum X^2 is actually on the high side: 13.4. This is not surprising, since $N - m = 8$ for our model and thus, the standard deviation in X^2 is $\sqrt{2(N-m)} = 4$.

I haven't computed the errors in our parameter estimates, but it is worth noting that they are highly correlated with one another.

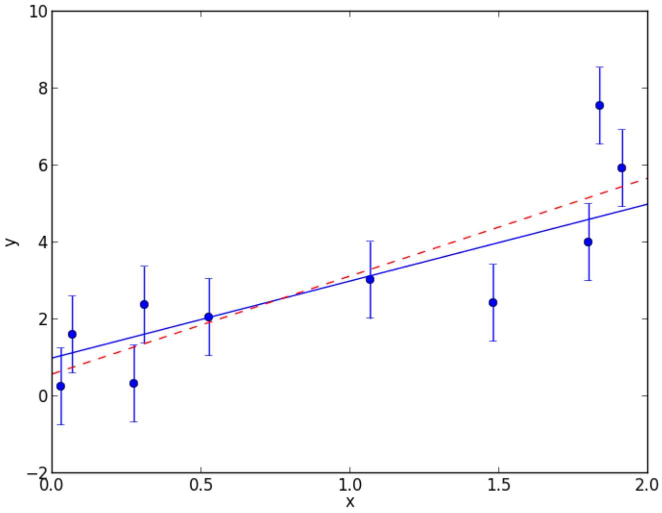

Figure 11: As above, but with the best fit line drawn in.

3.3.2 Polynomial Fits

We can extend the general reasoning in our linear fit to any sort of finite polynomial series:

$$f(x) = \sum_n c_n x^n$$

Thus:

$$\frac{df}{dc_n} = x^n$$

And we will simply get a set of m linear equations at the end of all of this. As with the linear fitting, we'll end up needing to invert an $m \times m$ matrix, but that's the only additional complication.

3.3.3 A Quick note on Nonlinear Functions

While I've described a very specific technique for minimizing X^2 under a very particular set of conditions (Gaussian errors, linear functions, etc.) you should remember that finding your optimal parameters is simply a subset of a very general problem: minimization of a function. You plug in parameters and get a X^2 out. Repeat infinitely until you get the lowest possible number.

For smoothly varying X^2 surfaces, this may mean performing some sort of steepest descents in an m dimensional space.

On the other hand, the surface may be bumpy. In other words, a local minimum of X^2 (around some value) won't necessarily correspond to a global minimum. This requires more clever methods (beyond the scope of these notes). You should consider looking into techniques like "Simulated Annealing" and other random approaches. Good luck!

3.4 Errors

So far, all of our errors have been generated by assuming that a number is generated via a Gaussian distributed around the true value. In a general sense, we've found that the errors scale as something like:

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

But this is not the only type of errors. Broadly speaking, they fall into two categories (Fig. 12):

1. **Random errors**
   Scatter around the true value of a system. A system with small random errors is said to be **precise**.
2. **Systematic Errors**
   An offset from the true value. No matter how many additional measurements you make, you can't get rid of a systematic error. A system with small systematic error is said to be **accurate**.

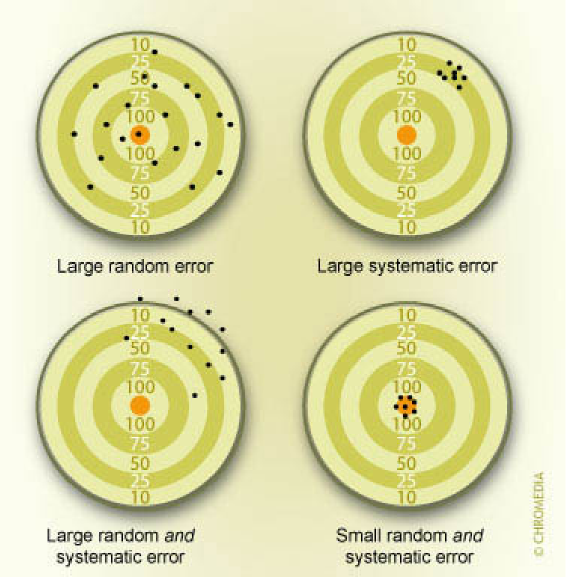
10
25
50
75
100
Large random error
Large systematic error
Large random *and* systematic error
Small random *and* systematic error
© CHROMEDIA

Figure 12: The difference between random and systematic error. Credit: chromedia.org. The upper left is accurate, but not precise. The upper right is precise, but not accurate. The lower left is neither accurate nor precise. The lower right is both accurate and precise.

3.4.1 Estimating Parameter Errors

Because systematic errors are tough (impossible?) to determine from the dataset itself, we're going to focus on random errors.

Suppose we're trying to estimate a single parameter, θ, of a function, f, and a dataset from which we can compute a likelihood function, which you'll recall was defined as:

$$\begin{aligned} L(\theta) &= \frac{1}{(2\pi)^{N/2}} \prod \frac{1}{\sigma_i} \exp\left(-\frac{(y_i - f(x_i;\theta))^2}{2\sigma_i^2}\right) \\ &= \frac{1}{(2\pi)^{N/2}} \left(\prod \frac{1}{\sigma_i}\right) \exp\left[-\sum \frac{(y_i - f(x_i;\theta))^2}{2\sigma_i^2}\right] \\ &= \frac{1}{(2\pi)^{N/2}} \left(\prod \frac{1}{\sigma_i}\right) \exp\left[-\frac{X^2}{2}\right] \end{aligned} \quad (53)$$

We earlier defined:

$$\mathcal{L} = -\ln(L) = const + \frac{X^2}{2}$$

Now, let us suppose that we find the parameter, $\hat{\theta}$, which maximizes the likelihood in equation **??**. The likelihood takes on some value, L_{max}, but if:

$$X^2 = X^2(\hat{\theta}) + 1$$

then the likelihood will drop to $1/\sqrt{e}$ of its peak value. This is the *definition* of the half-width, σ of a Gaussian.

Near the minimum value of the log-likelihood, $\mathcal{L}$, we can model the function as:

$$\mathcal{L} = \mathcal{L}_{min} + \frac{1}{2}(\theta - \theta_0)^2 \frac{\partial^2 \mathcal{L}}{\partial \theta^2}$$

To find the $1 - \sigma$ error range for our parameter, we're looking for:

$$\frac{1}{2}(\theta - \theta_0)^2 \frac{\partial^2 \mathcal{L}}{\partial \theta^2} = \frac{1}{2}$$

or

$$\langle (\theta - \theta_0)^2 \rangle = \left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2} \right)^{-1}$$

The term on the left is simply the variance of the parameter, θ, and thus we've figured out a way to determine the uncertainty of a parameter:

$$\sigma_\theta = \left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2} \right)^{-1/2} \tag{54}$$

3.4.2 Example: The Mean

We've developed a fairly sophisticated approach to computing the uncertainty of a parameter. All we need to do is to compute the log-likelihood function (simply $X^2/2$ for the Gaussian cases we've explored) and take the second derivative with respect to the parameter in question.

Let's suppose we measure a bunch of numbers, $\{x_i\}$ drawn from some mean, μ, errors, σ_i, associated with each. We know how to compute X^2, and thus $\mathcal{L}$:

$$\mathcal{L} = \sum_i \frac{(x_i - \hat{\mu})^2}{2\sigma_i^2}$$

The only parameter is $\theta = \hat{\mu}$. So, taking the first derivative:

$$\frac{\partial \mathcal{L}}{\partial \hat{\mu}} = -\sum \frac{(x_i - \hat{\mu})}{\sigma_i^2}$$

And thus the second derivative is:

$$\frac{\partial^2 \mathcal{L}}{\partial \hat{\mu}^2} = \sum_i \frac{1}{\sigma_i^2}$$

and so the uncertainty in the mean is:

$$\sigma_\mu = \left(\sum_i \frac{1}{\sigma_i^2} \right)^{-1/2}$$

This might look a bit more familiar if we take an example in which the errors on the datapoints are equal:

$$\begin{aligned} \sigma_\mu &= \left(\sum_i \frac{1}{\sigma^2}\right)^{-1/2} \\ &= \left(\frac{N}{\sigma^2}\right)^{-1/2} \\ &= \frac{\sigma}{\sqrt{N}} \end{aligned}$$

Which is exactly what we've seen before!

Of course, prior to this, we had no idea how to compute the errorbars on the mean if the errors *weren't* equal.

3.4.3 Example: A Straight Line

Now let's consider the errors associated with our straight line example from earlier. I'm going to warn you in advance that this approach is not perfect since the errors in the slope and intercept (the two parameters of a straight line) are correlated. But we'll deal with that in a little bit.

We have the likelihood function:

$$\mathcal{L} = \sum_i \frac{(mx_i + b - y_i)^2}{2\sigma_i^2}$$

And so:

$$\frac{\partial^2 \mathcal{L}}{\partial m^2} = \sum_i \frac{x_i^2}{\sigma_i^2} = S_{xx}$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial b^2} = \sum_i \frac{1}{\sigma_i^2} \equiv S$$

For our particular case, we find, $S = 10$ and $S_{xx} = 14.1$. This yields:

$$\sigma_m = 0.27 \;\; ; \;\; \sigma_b = 0.32$$

Given our estimated values from earlier (2.54 and 0.58, respectively), and comparing them to the "true" parameters of the line (2 and 1), our slope is about $2-\sigma$ from the true value, while the intercept is about $1.4-\sigma$ from the true value. Unlikely, but no cause for alarm.

It turns out, though, that I'm overestimating the errors, and that's because I haven't taken in to account the correlations between them.

3.5 The Fisher Matrix

If all of our parameters are independent, we found the relationship:

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_n^2} = \frac{1}{\sigma_n^2}$$

It turns out, though, that the parameters for our line are dependent on one another. Another way of saying this is that the errors are **correlated**. Or, to put it another way:

$$\sigma_{bm} = \langle (\hat{b} - b)(\hat{m} - m) \rangle \neq 0$$

If they were statistically independent then if we overestimate $\hat{m}$ then we're equally likely to overestimate and underestimate $\hat{b}$.

In order to figure out their dependence, we need to compute the **Fisher Information Matrix**:

$$\mathcal{F}_{nm} = \frac{\partial^2 \mathcal{L}}{\partial \theta_n \partial \theta_m} \tag{55}$$

You will note that $1/\mathcal{F}_{nn}$ was our previous estimate for σ_n^2.

The inverse of the Fisher Matrix yields not only the errors, but also the correlation between them:

$$\sigma_{nm} = (\mathcal{F})^{-1}_{nm} \tag{56}$$

where if $n = m$, this is simply a shorthand:

$$\sigma_{nn} = \sigma_n^2$$

and otherwise, it's a measure of the correlation between the two parameters.

We can get a measure of the significance of the correlation between the errors (or between any two variables, really), by computing the **Pearson Correlation Coefficient**:

$$\rho_{nm} = \frac{\sigma_{nm}}{\sigma_n \sigma_m} \tag{57}$$

By construction, the Pearson Coefficient is dimensionless and has a range of -1 (perfectly anti-correlated) to +1 (perfectly correlated).

By the way, it can be shown in general (from the general properties of 2x2 matrix inversion) that:

$$\frac{1}{\mathcal{F}_{nn}} \geq (\mathcal{F})^{-1}_{nn} \tag{58}$$

To put it another way, we've over-estimated the errors by assuming previously that they were uncorrelated.

3.5.1 Example: Parameterizing a Line

We can finally compute the correlated errors on the parameters for a straight line model (again, assuming uncorrelated Gaussian errors).

We've already seen:

$$\mathcal{F}_{11} = \frac{\partial^2 \mathcal{L}}{\partial m^2} = S_{xx}$$

and

$$\mathcal{F}_{22} = \frac{\partial^2 \mathcal{L}}{\partial b^2} = S$$

But now we need to compute the symmetric, off-diagonal terms:

$$\mathcal{F}_{12} = \frac{\partial^2 \mathcal{L}}{\partial m \partial b} = S_x$$

Thus the full Fisher Matrix is:

$$\mathcal{F}_{nm} = \begin{pmatrix} S_{xx} & S_x \\ S_x & S \end{pmatrix}$$

Inverting, we get:

$$\sigma_{nm} = (\mathcal{F})^{-1}_{nm} = \frac{1}{S S_{xx} - S_x^2} \begin{pmatrix} S & -S_x \\ -S_x & S_{xx} \end{pmatrix}$$

I've plotted the errors in Fig. 13. As you'll note, the two parameters are anti-correlated. If you overestimate the slope, you're likely to underestimate the intercept, and vice-versa.

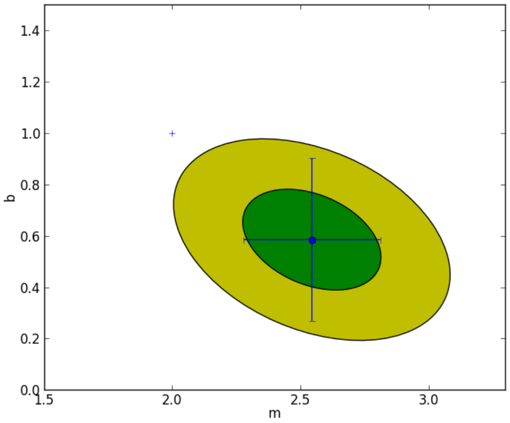

Figure 13: The estimate of the slope and intersept of 10 datapoints with random errors. The plus sign indicates the two parameters, while the dot indicates the χ^2 minimum for this realization. The drawn errorbars are $1-\sigma$ (incorrectly) assuming that they are uncorrelated. The green ellipse represents the $1-\sigma$ limit, while yellow indicates $2-\sigma$.

Note: For linear functions (like the one we're working with), we'll get an error ellipse. On the other hand, while you can model most functions as linear over a small range (hence the usefulness of Taylor series) most functions will produce a much more complicated banana-shaped errorbar. This requires actually exploring all of the combinations of parameters that produce $X^2/2 = X^2_{min}/2 + 1$.

3.6 Covariance

Ideally, we'd like to reduce all of the simplifying assumptions we've made so far, but that would expand these notes far too much. However, there's one situation that will arise a lot in research: correlated errorbars.

In general, we write the covariance matrix between two data elements as:

$$C_{ij} = \langle (y_i - \mu_i)(y_j - \mu_j) \rangle \tag{59}$$

where μ_i is the expected value of element i with the true model. For uncorrelated errors, C_{ij} is a diagonal matrix, with elements, $1/\sigma_i^2$.

It's fairly straightforward to re-write X^2 including covariant errors:

$$X^2 = \sum_{i,j} (y_i - f_i) \left(C^{-1}\right)_{ij} (y_j - f_j) \tag{60}$$

where f_i is a fiducial model of element, i. With a diagonal covariance matrix, this produces our original X^2. Likewise, we simply have a likelihood function:

$$L = Const \times \exp\left(-\frac{X^2}{2}\right) \tag{61}$$

3.7 The KS Test

We've focused so far on trying measure particular parameters of a random system, with the idea that for a good fit, X^2 will be small. However, we could ask the more general question:

How likely that our measured data was drawn from a particular Probability Distribution Function?

To get a handle on *that*, we introduce (quite without derivation) the **Kolmogorov-Smirnov** (KS) test.

Suppose we have a set of data, $\{x_i\}$, and a presumed cumulative distribution function $F(x)$. Do they match?

Rather than simply define things mathematically, let's do this by example. Suppose that we have reason to believe that a set of numbers are drawn from a Gaussian distribution with a mean of zero, and a standard deviation of 1: $\mathcal{N}(0,1)$.

Unbeknownst to us, the actual data was drawn uniformly from $[-1,1]$.

The theoretical cumulative distribution function is related to the "error function," and looks like:

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-x'^2}2dx'$$

while the sample cumulative distribution function is given by:

$$S_N(x)=\frac{1}{N}\sum_i H(x_i-x)$$

where $H(x_i-x)$ is the "Heaviside Step Function."

In Fig. 14, I plot both the theoretical and measured cumulative functions (along with the "correct" theoretical distribution).

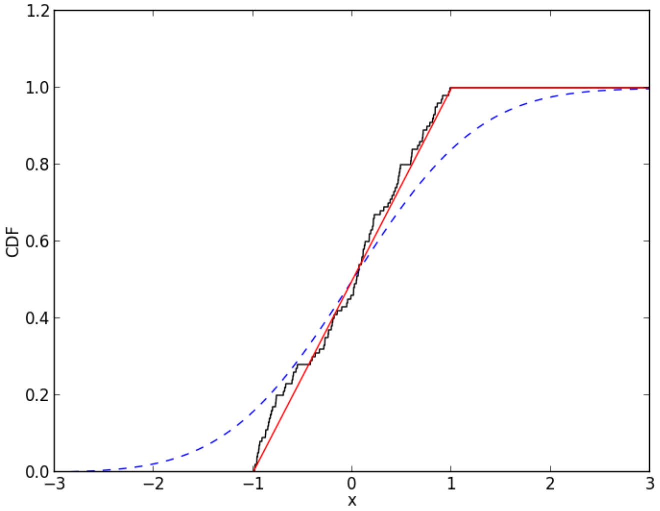

Figure 14: In black, the Cumulative Distribution Function, CDF, of 100 points randomly selected with a uniform distribution between -1 and 1. In dashed blue, the CDF for a Normal distribution with $\sigma=1$. In red, the theoretical CDF for a [-1,1] uniform distribution.

By eye, you will notice that the measured cumulative distribution function seems to fit a uniform distribution more closely than the normal distribution. By definition, all CDFs go to 0 at $-\infty$ and 1 at ∞, but in the middle, differences between the curves can be quite significant. Further, we don't much care how the x-axis is spaced. We only care about the differences between points. In particular, in comparing a measured versus theoretical distribution, we want to measure:

$$D_n=\max|S_N(x)-F(x)| \tag{62}$$

For our distributions:

$$D_{N,Gauss} = 0.164 \quad ; \quad D_{N,Uniform} = 0.083$$

The KS test itself tells us what the probability of getting a value of D_N as large as we have found or larger. There are lookup tables in general, but so long as N is more than 20 or so, you can use an asymptotic relation:

$$Pr(D_N > z/\sqrt{N}) = 2\sum_{m=1}^{\infty}(-1)^{m-1}e^{-2m^2z^2} \tag{63}$$

where in this case, $z_{gauss} = 1.64$ and $z_{uniform} = 0.83$.

Thus we find:

$$Pr(D_N > 1.64)_{Gauss} = 0.009$$

and

$$Pr(D_N > 0.83)_{uniform} = 0.496$$

While we can't be certain that the distribution was drawn from a $[-1, 1]$ uniform distribution, we can be fairly certain it wasn't drawn from a $\mathcal{N}(0, 1)$ Gaussian distribution.

There are other variants on the KS test, and other ways of checking distributions, but this is a fairly good start.

4 Closing

This is by no means a comprehensive discussion. I give a few references below if you'd like to read further, but everything above falls under the heading of "stuff every physicist should know." I have completely ignored non-parametric reconstructions, evaluating the errors in a distribution from bootstrap methods, a formal discussion of hypothesis testing, tests of fit, and many other topics. As I mentioned above, this is a work in progress, so should any additional topics occur to you (the hypothetical reader), please suggest them.

Acknowledgements

This is an ongoing project. I'd like to thank Austen Groener, Travis Hoppe and Gordon Richards for providing useful feedback on early drafts.

References

This is intended as a general bibliography for further reading. I will add annotations as they occur to me.

1. Feigelson, Eric D. & Babu, G. Jogesh, *Modern Statistical Methods for Astronomy with R Applications.* Cambridge, UK: Cambridge University Press, 2012. Print. I can't personally attest to the quality of this, but Gordon Richards speaks highly of it.

2. Kendall, Maurice G., Alan Stuart, J. K. Ord, Steven F. Arnold, and Anthony Hagan. *Kendall's advanced theory of statistics.* 6th ed. London: Edward Arnold, 1994. Print. This is a standard text in advanced statistics. I'd master everything else first before proceeding to this one.

3. Lupton, Robert. *Statistics in theory and practice.* Princeton, NJ: Princeton University Press, 1993. Print. I used this book as a grad student, and continue to refer to it today. There is a lot of great discussion, including hypothesis testing, KS tests, and many other topics.

4. Press, William H., Teukolsky, Saul A., Vetterling, William T. & Flannery, Brian P. *Numerical Recipes: The Art of Scientific Computing.* 3rd Edition. Cambridge, UK: Cambridge University Press, 2007. Print. When I say things like, "solving numerically in the usual way," this book describes the "usual way."