Notes on Statistical Methods

1060-710: Mathematical and Statistical Methods for Astrophysics[*]

Fall 2009

Contents

[*]

Thursday, October 22, 2009

1 Probability

Statistical analysis and statistical inference is based on probabilistic statements, such as

1. "A card drawn from this deck has a 25% chance of being a spade"
2. "If the true luminosity of this star is L, there is a 67% chance that I will measure a luminosity between $L - \Delta L$ and $L + \Delta L$"
3. "If I measure the position of this electron, there is a 12.5% chance I will find it in this octant"
4. "There is a 40% chance it will rain tomorrow"
5. "The Hubble constant is 90% likely to lie between 70 and 76 km/sec/Mpc"

There are subtle philosophical distinctions among the different kinds of probability, whether they describe true quantum mechanical uncertainty, ignorance of the inner workings of a physical system, outcomes of hypothetical experiments, or best estimates about unknown parameters. They all obey the same quantitative rules, however, and I prefer to take the approach used to deal with temperature in elementary thermodynamics: we know more or less what we mean by probability, and we won't examine the definition too closely unless we need to. I encourage you to read Chapter Two of Gregory for a more careful consideration of probability as a quantitative measure of plausibility.

1.1 Probability of Logical Propositions

The most basic thing we can assign a probability to is a logical proposition. For instance, the probability that a fair die lands on three is 1/6. The probability that an atom decays within its half-life is 50%. We typically use capital letters to label propositions, so we talk about things like $p(A)$.

Logical propositions can be combined; using the notation of Gregory, the basic operations are

- Negation. $\overline{A}$ is true if A is false, and vice-versa. In words, we can think of $\overline{A}$ as "not A".
- Intersection. A, B is true if A and B are both true. In words, this is "A and B".
- Union $A + B$ is true if either A or B (or both) is true. In words, this is "A or B". Note the unfortunate aspect of Gregory's notation that $+$ is to be read as "or" rather than "and".

Another important quantity is the conditional probability $p(A|B)$, the probability that A is true *if* B is true, often referred to as the probability of A given B.

The basic rules of probability, which sort of follow from common sense, are

- $p(A) + p(\overline{A}) = 1$; A and $\overline{A}$ are exhaustive, mutually exclusive alternatives, i.e., either A or $\overline{A}$ is true, but not both, so the probability is 100% that either one or the other is true.
- $p(A, B) = p(A|B)\, p(B) = p(B)\, p(B|A)$; again, this is sort of self-apparent. If B is true, for which the odds are $p(B)$, then the odds of A also being true are $p(A|B)$. Note that if these are independent propositions, i.e., $p(A|B) = p(A|\overline{B}) = p(A)$, this means $p(A, B) = p(A)p(B)$.
- $p(A + B) = p(A) + p(B) - p(A, B)$

Identities like this are easier to see with so-called *truth tables*, where you make a list of all of the possible combinations of truth and falsehood for different propositions:

A	B	A, B	$A, \overline{B}$	$\overline{A}, B$	$\overline{A}, \overline{B}$	$A + B$
T	T	T	F	F	F	T
T	F	F	T	F	F	T
F	T	F	F	T	F	T
F	F	F	F	F	T	F

You can see that

$$p(A) = p(A, B) + p(A, \overline{B}) \tag{1.1a}$$

$$p(B) = p(A, B) + p(\overline{A}, B) \tag{1.1b}$$

$$p(A + B) = p(A, B) + p(A, \overline{B}) + p(\overline{A}, B) \ , \tag{1.1c}$$

from which $p(A + B) = p(A) + p(B) - p(A, B)$ follows by simple algebra.

An important rule of probability is that if $\{A_k\}$ form a mutually exclusive exhaustive set of alternatives,

$$\sum_k p(A_k) = 1 \tag{1.2}$$

We can also examine the truth table, and see that

$$p(A|B) = \frac{p(A, B)}{p(A, B) + p(\overline{A}, B)} = \frac{p(A, B)}{p(B)} \tag{1.3}$$

This is another very important property, in the form

$$p(A, B) = p(A|B)p(B) \ . \tag{1.4}$$

Of course, it also works the other way, so

$$p(B|A)p(A) = p(A, B) = p(A|B)p(B) \tag{1.5}$$

If we divide through by A we find

$$P(B|A) = \frac{p(A|B)p(B)}{p(A)} \tag{1.6}$$

this seemingly trivial result is known as Bayes's Theorem, and has remarkably deep consequences.

1.2 Bayesian and Frequentist Applications of Probability

So far, we've been sort of casual about just what these probabilities mean. Now it's time to think more carefully about what we use them for.

In the classical or frequentist approach to statistics, probability is to be thought of as the relative frequency of observable events. If I perform an experiment which, under certain conditions B, has a probability $p(A|B)$ of having outcome A, then if I do the experiment a large number $\mathcal{N}$ of times, and the conditions really are B, roughly $\mathcal{N}p(A|B)$ of those experiments will give the result A and $\mathcal{N}[1-p(A|B)]$ of them will give the result $\overline{A}$. We refer to $p(A|B)$ as the *likelihood function*.

In the Bayesian approach, probabilities can also represent our degree of confidence that a certain set of conditions B exist. Typically, one considers an experiment into which one enters with prior expectation $p(B)$ about those conditions. One then does the experiment and gets the result A, and as a result, the probability one assigns to condition B becomes the posterior probability $p(B|A)$. If we have modelled the experiment and know $p(A|B)$ and $p(A|\overline{B})$, we can use Bayes's theorem to find the posterior:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} = \frac{p(A|B)p(B)}{p(A|B)p(B) + p(A|\overline{B})p(\overline{B})} \tag{1.7}$$

1.2.1 Example: Disease Testing

To give a concrete, simple example of this, consider a disease that a fraction α of the population is known to have. There is a test for the disease that has a false positive rate of β and a false negative rate of γ. (One might be told that the test is "99% accurate", which means either the false positive or false negative rate is 1%.) This means the likelihood is

$p(+|S) = 1-\gamma$	$p(-|S) = \gamma$
$p(+|\overline{S}) = \beta$	$p(-|\overline{S}) = 1-\beta$

where $+$ indicates a positive test result, $-$ a negative test result, S (for sick) a patient with the disease and $\overline{S}$ a patient without the disease.

Suppose that I take the test and get a positive result. I might understandably want to know the odds that I actually have the disease, and I might naïvely think they are 99% if the test is 99% accurate. But that 99% is a frequentist statement, and it is not a statement about the probability that a given person has the disease. From a frequentist point of view, someone either has the disease or they don't; the probabilistic thing is the outcome of the test. So the 99% (or in general $1-\beta$) is not the odds that someone who tests positive has the disease; it is the odds that someone without disease will test negative. I.e., it's $p(-|\overline{S})$, not $p(S|+)$.

To find $p(S|+)$, we need to use Bayes's Theorem, and we need to assign a prior probability $p(S)$. Assuming that I don't have any special risk factors, my odds of having the disease are the same as the fraction of the population which has it, i.e, $p(S) = \alpha$ and $p(\overline{S}) = 1-\alpha$. Then I can say

$$p(S|+) = \frac{p(+|S)p(S)}{p(+)} = \frac{p(+|S)p(S)}{p(+|S)p(S) + p(+|\overline{S})p(\overline{S})} = \frac{(1-\gamma)\alpha}{(1-\gamma)\alpha + \beta(1-\gamma)} \tag{1.8}$$

So if $\alpha = 0.001$ and $\beta = \gamma = 0.01$,

$$p(S|+) = \frac{0.99 * 0.001}{0.99 * 0.001 + 0.999 * 0.01} = \frac{0.00099}{0.00999 + 0.00099} \tag{1.9}$$

or about 9%. Another way to see it is that

$$p(S|+) = \frac{p(S,+)}{p(S,+) + p(\overline{S},+)} \tag{1.10}$$

while

$$p(+|S) = \frac{p(S,+)}{p(S,+) + p(S,-)} \tag{1.11}$$

1.3 Probability Distributions of Discrete Variables

One important application of the idea of probabilities for different propositions is to the case of a random variable which can take one of a discrete set of values, like the number rolled on a die. Then the alternatives corresponding to different values are obviously mutually exclusive, and if you sum the probability over all possible values, you have to get unity:

$$\sum_I p(I) = 1 \tag{1.12}$$

Note that the probabilities of different possible values need not be equal. For example, if I and J are the results of individual die rolls, so that

$$p(I) = \frac{1}{6} \qquad \text{for } I = 1, 2, \ldots, 6 \tag{1.13a}$$

$$p(J) = \frac{1}{6} \qquad \text{for } J = 1, 2, \ldots, 6 \tag{1.13b}$$

and the die rolls are independent so that $p(I, J) = p(I)p(J)$, then their sum $K = I + J$ will be distributed according to

$$p(K) = \sum_I \sum_J \delta_{K,I+J} p(I)p(J) = \frac{6 - |K-7|}{36} \qquad \text{for } K = 2, 3, \ldots, 12 \ . \tag{1.14}$$

1.4 Probability Distributions of Continuous Variables

When talking about the probability for a continuous random variable to take on a certain value, things get a little more subtle, because one specific value is a set of measure zero. If we want to talk about sensible alternatives, we should think about alternatives that represent the variable falling into different ranges. For example, we can talk about the probability for x to lie between x_0 and $x_0 + dx$:

$$p(x_0 \le x < x_0 + dx) \tag{1.15}$$

In the limit $dx \to 0$, we expect this to be infinitesimal, so we can define the *probability density function* (pdf)

$$f(x_0) = \lim_{dx\to 0} \frac{p(x_0 \le x < x_0 + dx)}{dx} \ . \tag{1.16}$$

There are many notations for the pdf, including $p(x)$, $p(x)$, and even pdf(x). Sometimes it will be useful to stress its nature as a density by writing

$$\frac{dP}{dx}(x_0) = f(x_0) \ ; \tag{1.17}$$

note that by $\frac{dP}{dx}$ we don't actually mean a derivative of some function $P(x)$, although of course we could define a function such that $f(x)$ is its derivative. In face, that function is kind of useful, since it's the probability that x is at or below a certain value:

$$p(x \le x_0) = \int_{-\infty}^{x_0} f(x)\, dx =: f(x) \tag{1.18}$$

This function $f(x)$ is called the *cumulative density function* or cdf.

The normalization (1.12) becomes, in the case of a continuous random variable,

$$p(-\infty < x < \infty) = F(\infty) = \int_{-\infty}^{\infty} f(x)\, dx = 1 \ . \tag{1.19}$$

The pdf finally gives us a chance to write down a precise formula for the expectation value considered previously. This is the average value of some function of x so it's

$$\langle A(x) \rangle = \int_{-\infty}^{\infty} A(x)\, f(x)\, dx \tag{1.20}$$

Note that we can also consider a change of variables, where we ask the about the pdf of some y which is a function of x. (If it's a good coördinate change, $y(x)$ will be a monotonic function in the interval we're interested in.) This is then

$$f(y_0) = f(y(x_0)) = \lim_{dy\to 0} \frac{p(y_0 \le y(x) < y_0 + dy)}{dy} \tag{1.21}$$

Now, for infinitesimal dx, we can write

$$y(x_0 + dx) = y(x_0) + y'(x_0)\, dx = y_0 + y'(x_0)\, dx \tag{1.22}$$

so if x is between x and x_0, y is between y_0 and $y'(x_0)\, dx$. To get the relationship between $f(y_0)$ and $f(x_0)$, we need to consider two different cases: $y'(x_0) > 0$ and $y'(x_0) < 0$.

If $y'(x_0) > 0$**:** Then $y_0 + y'(x_0)\, dx > y_0\, dx$ and

$$\begin{aligned} f(x_0)\, dx &= p(x_0 < x < x_0 + dx) = p(y_0 < y < y_0 + y'(x_0)\, dx) = f(y_0)\, y'(x_0)\, dx \\ &= f(y_0)\ |y'(x_0)|\ dx \end{aligned} \tag{1.23}$$

and

$$f(y(x_0)) = \frac{f(x_0)}{|y'(x_0)|} \tag{1.24}$$

If $y'(x_0)<0$: Then $y_0+y'(x_0)\,dx=y_0+\lvert y'(x_0)\rvert\;dx<y_0$ and

$$\begin{aligned} f(x_0)\,dx &= p(x_0<x<x_0+dx)=p(y_0-\lvert y'(x_0)\rvert\;dx<y<y_0) \\ &= f(y_0-\lvert y'(x_0)\rvert\;dx)\,y'(x_0)\,dx=f(y_0)\;\lvert y'(x_0)\rvert\;dx \end{aligned} \tag{1.25}$$

and once again

$$f(y(x_0))=\frac{f(x_0)}{\lvert y'(x_0)\rvert} \tag{1.26}$$

This is of course just the chain rule of calculus applied to a change of variables, and is easy to remember if we write it in the notation

$$f(y)=\frac{dp}{dy}=\frac{dp}{dx}\left\lvert\frac{dx}{dy}\right\rvert=\frac{f(x)}{y'(x)}\ . \tag{1.27}$$

1.4.1 Example: Inclination

As an example of the importance of a change of variables, consider the inclination ι between an arbitrary direction (say the normal to the orbital plane of a binary star system) and our line of sight. This is uniformly distributed in $\chi=\cos\iota$, from $\chi=-1$ to $\chi=1$, so

$$f(\chi)=\frac{dp}{d\chi}=\frac{1}{2}\qquad -1\le\chi\le 1 \tag{1.28}$$

If we change variables to ι, we have to use

$$\frac{d\chi}{d\iota}=\frac{d}{d\iota}\cos\iota=-\sin\iota \tag{1.29}$$

to find

$$f(\iota)=\frac{dp}{d\iota}=\frac{dp}{d\chi}\left\lvert\frac{d\chi}{d\iota}\right\rvert=\frac{1}{2}\sin\iota\qquad 0\le\iota\le\pi \tag{1.30}$$

which is not uniform in ι.

Of course, this can also be extended to a probability density in the inclination and an azimuthal angle ψ, and leads to a direction uniformly distributed over the sphere:

$$\frac{d^2p}{d^2\Omega}=\frac{d^2p}{\sin\iota\,d\iota\,d\psi}=\frac{d^2p}{d\chi\,d\psi}=\frac{1}{4\pi} \tag{1.31}$$

Tuesday, October 27, 2009

2 Statistical Inference

2.1 Example: Bayesian and Frequentist Upper Limits

The difference between the Bayesian and frequentist approaches is illustrated by the statement

"Our experiment set an upper limit of x_{UL} on the value of x at the 90% confidence level."

That turns out to mean rather different things when made in a Bayesian and a frequentist context.

For simplicity, assume that the output of the experiment is a single measurement, which results in a number y. Our understanding of the underlying theory and the experimental setup tells us the likelihood function $p(y|x)$. Note that this is a density in y but *not* x. It tells us that the probability that y will lie in a range of values given a specific value of x is

$$p(y_1 < y < y_2|x) = \int_{y_1}^{y_2} p(y|x)\, dy \; . \tag{2.1}$$

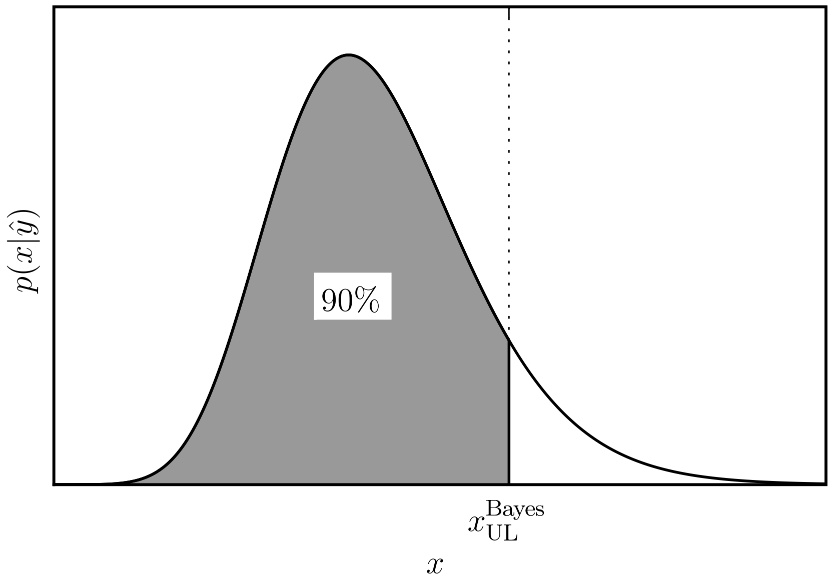
Now suppose we run the experiment and get the actual number $\hat{y}$. How would a Bayesian and a frequentist calculate a 90% upper limit on x?

2.1.1 Bayesian Upper Limit

The Bayesian 90% upper limit statement means what you'd think it means: given the result $\hat{y}$, you are 90% confident that the true value x is below $x_{\mathrm{UL}}^{\mathrm{Bayes}}$. We can write this in terms of the posterior probability distribution $p(x|y)$

$$p(x < x_{\mathrm{UL}}^{\mathrm{Bayes}}|\hat{y}) = \int_{-\infty}^{x_{\mathrm{UL}}^{\mathrm{Bayes}}} p(x|\hat{y})\, dx = 90\% \tag{2.2}$$

Plotting the posterior pdf, x_{UL} is defined such that 90% of the area under the posterior $p(x|\hat{y})$ lies in the region $x < x_{\mathrm{UL}}$:



This is probably what you think of when you hear a 90% upper limit statement:

"Given that the result of the measurement was $\hat{y}$, we estimate a 90% probability that the true value of x is x_{UL} or lower."

Of course, it uses the posterior probability $p(x|y)$ rather than the likelihood $p(y|x)$, so we have to use Bayes's theorem to evalate it:

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} = \frac{p(y|x)\,p(x)}{\int_{-\infty}^{\infty} p(y|x')\,p(x')\,dx'} \ , \tag{2.3}$$

and that in turn means we need to know the prior $p(x)$.

2.1.2 Frequentist Upper Limit

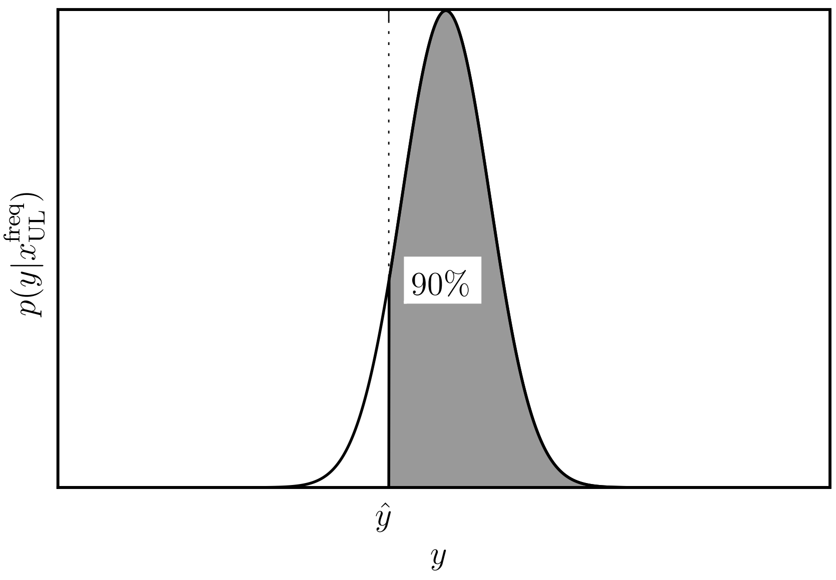
In the frequentist approach, we can't estimate the probability that $x < x_{\mathrm{UL}}$, because we don't talk about probabilities for physical quantities to have particular values. (After all x has some fixed value, even if we don't know it.) The probabilities we can discuss are those of the outcome of an experiment including some random measurement error. The definition of the frequentist 90% upper limit is actually

$$p(y > \hat{y}|x_{\mathrm{UL}}^{\mathrm{freq}}) = \int_{\hat{y}}^{\infty} p(y|x_{\mathrm{UL}}^{\mathrm{freq}})\,dy = 90\% \tag{2.4}$$

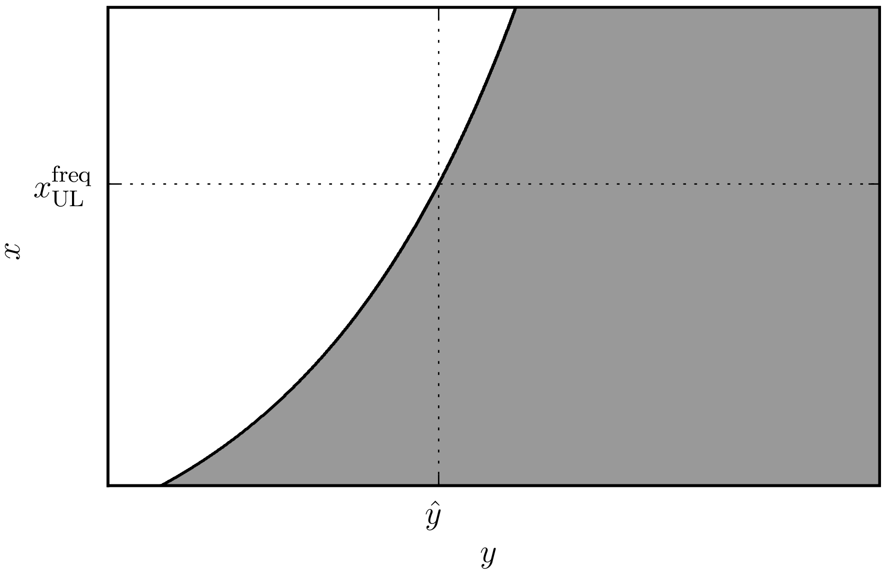
or more accurately

$$p(y > \hat{y}|x) = \int_{\hat{y}}^{\infty} p(y|x)\,dy > 90\% \qquad \text{if } x > x_{\mathrm{UL}}^{\mathrm{freq}} \ . \tag{2.5}$$

That means that the upper limit $x_{\mathrm{UL}}^{\mathrm{freq}}$ is defined such that, if the actual value of x is right at $x_{\mathrm{UL}}^{\mathrm{freq}}$, 90% of the area under the likelihood function $p(y|x_{\mathrm{UL}}^{\mathrm{freq}})$ lies in the region $y > \hat{y}$, i.e., we would have expected a y value higher than the one we saw 90% of the time:



The idea is that, while we can't assign probabilities to different values of x, we can think about how unlikely it would be to make a y measurement as low as $\hat{y}$ if the true value of x were large. For each possible value of x, we can find the range of y values which we'd expect to find 90% of the time; those are shaded on this plot:



At any x, the y values that fall in the unshaded region would be expected only 10% of the time. The range of x values ($x > x_{\rm UL}^{\rm freq}$) excluded at the 90% confidence level is those for which the actual measured $\hat{y}$ falls in the 10th percentile or lower among expected y values. The statement in words is:

"If the true value of x were above the upper limit $x_{\rm UL}^{\rm freq}$, we would expect to get our actual result $\hat{y}$ or lower in less than 10% of the experiments."

It's almost never stated that way, but that's what "90% frequentist upper limit" means.

2.2 Consequences of Choice of Priors in Bayesian Analysis

Let's return to the Bayesian interpretation of our experiment, and consider how the choice of prior $p(x)$ impacts the construction of the posterior

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} = \frac{p(y|x)\,p(x)}{\int_{-\infty}^{\infty} p(y|x')\,p(x')\,dx'} \ . \tag{2.6}$$

(Recall that x is the underlying physical parameter and y is the quantity returned by the experiment.) Notice that since we typically focus on the x dependence, we can write

$$p(x|\hat{y}) = \frac{p(\hat{y}|x)\,p(x)}{p(\hat{y})} \propto p(y|x)\,p(x) \tag{2.7}$$

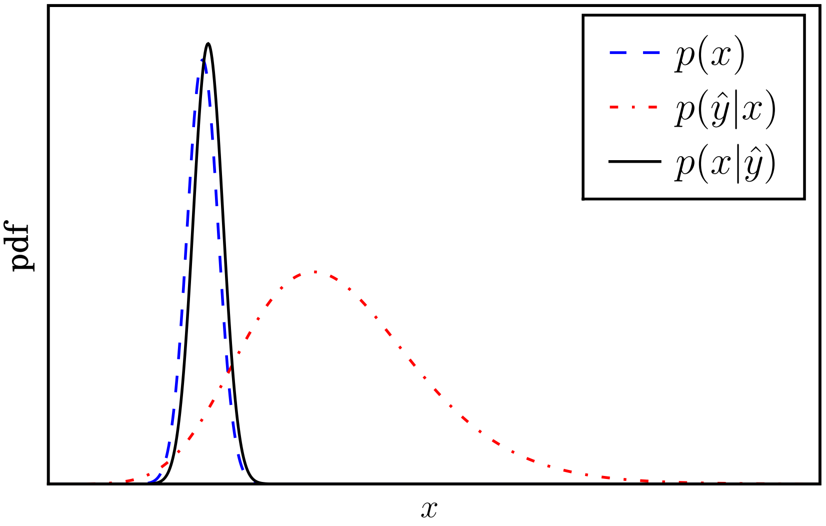
where the normalization $1/p(\hat{y})$ is independent of x and can be set by requiring

$$\int_{-\infty}^{\infty} p(x|\hat{y})\,dx = 1 \tag{2.8}$$

so schematically,

$$(\text{posterior}) = (\text{likelihood}) \times (\text{prior}) \times (\text{normalization}) \tag{2.9}$$

But the big question is what we use for $p(x)$. It is supposed to reflect our prior knowledge, but sometimes it can be dangerous to take that too literally. For example, suppose our prior expectations have already pretty tightly constrained x compared to the sensitivity of the experiment. Then we can get a scenario like this



where the posterior looks a lot like the prior. We'd conclude that our state of knowledge after the experiment is basically what it was before, and was bound to be so regardless of the outcome of the experiment. That might be true, but it's not the most informative description of the experimental results. It's also the sort of thing which critics of Bayesian statistics often suspect: letting our prior expectations be part of the analysis can make the results conform to those expectations.

So sometimes you may choose to be as ignorant as possible about the prior. One seemingly obvious way to assume we know nothing is to let any value of x be equally likely, and choose the prior

$$p(x) = \text{constant} \ . \tag{2.10}$$

There's a formal problem, in that we should normalize $p(x)$ so that

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{2.11}$$

and you can't do that if x is literally a constant for all x. There's an easy workaround, though. You can take

$$p(x) = \begin{cases} \frac{1}{X} & |x| < \frac{X}{2} \\ 0 & |x| > \frac{X}{2} \end{cases} \ , \tag{2.12}$$

which is constant over some range of values of width X and then just choose X much larger than any other relevant scale in the problem. If the likelihood is well-behaved, the posterior will remain well-defined in the limit $X \to \infty$. In this case[1]

$$p(x|\hat{y}) = \frac{p(\hat{y}|x)\,p(x)}{\int_{-\infty}^{\infty} p(\hat{y}|x')\,p(x')\,dx'} = \frac{p(\hat{y}|x)\,\Theta(X-2\,|x|)\,X^{-1}}{\int_{-X/2}^{X/2} p(\hat{y}|x')\,X^{-1}\,dx'} \overset{X\to\infty}{\longrightarrow} \frac{p(\hat{y}|x)}{\int_{-\infty}^{\infty} p(\hat{y}|x')\,dx'} \tag{2.13}$$

so the posterior is just a constant normalization factor times the likelihood, and the Bayesian approach starts to look a lot like the Frequentist one.

There is another problem with the approach of choosing a uniform prior: the prior $p(x)$ and the posterior $p(x|y)$ are both densities the physical quantity x. That means if we make a coördinate change, the prior will not remain constant. For concreteness, suppose x is a quantity which is physically constrained to be positive, so that the uniform prior is actually

$$p(x) = \begin{cases} \frac{1}{X} & 0 < x < X \\ 0 & x > X \end{cases} , \tag{2.14}$$

which would lead to a posterior

$$p(x|\hat{y}) = \frac{p(\hat{y}|x)}{\int_0^{\infty} p(\hat{y}|x')\,dx'} \propto p(\hat{y}|x) \ . \tag{2.15}$$

We could also define $\xi = \ln x$, which is allowed to range from $-\infty$ to ∞ as x ranges from 0 to ∞. Well, since $p(x)$ is a density in x, $p(\xi)$ is not just $p(x = e^{\xi})$. Rather,

$$p(\xi) = \frac{dP}{d\xi} = \frac{dx}{d\xi}\frac{dP}{dx} = e^{\xi} p(x = e^{\xi}) \tag{2.16}$$

So

$$\text{If } p(x) = \text{constant then } p(\xi) \propto e^{\xi} \tag{2.17}$$

and conversely

$$\text{If } p(\xi) = \text{constant then } p(x) \propto \frac{1}{x} \ . \tag{2.18}$$

On the other hand, the likelihood $p(y|x)$ is *not* a density in x, so it is unchanged by a chance of coördinates:

$$p(y|\xi) = p(y|x = e^{\xi}) \ . \tag{2.19}$$

So the main moral is that what it means to use a uniform prior depends on how you parametrize the relevant physical quantities.

Thursday, October 29, 2009

2.3 Joint Probability Distributions

Just as we can consider the joint probability for multiple propositions to be true, $p(A, B)$, we can also consider the joint pdf for multiple variables. For two variables, the pdf $p(x, y)$ is defined as

$$p(x_0, y_0) = \lim_{dx,dy\to\infty} \frac{p(x_0 < x < x_0 + dx, y_0 < y < y_0 + dy)}{dx\,dy} \ . \tag{2.20}$$

[1] We use the Heavyside step function $\Theta(\xi)$, which is 1 when $\xi > 0$ and 0 when $\xi < 0$

By the standard probability sum rule for disjoint alternatives, the probability that (x,y) lies in some region R is

$$p((x,y)\in R)=\iint\limits_R p(x,y)\,dx\,dy \tag{2.21}$$

and the normalization of the joint pdf is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\,dx\,dy=1 \tag{2.22}$$

Since $p(x,y)$ is a density in x and y, it is sometimes useful to write it as $\frac{d^2p}{dx\,dy}$, to help us remember that if we change variables to $u=u(x,y)$ and $v=v(x,y)$, the pdf transforms like

$$p(u,v)=\frac{d^2p}{du\,dv}=\left\|\frac{\partial(u,v)}{\partial(x,y)}\right\|^{-1}\frac{d^2p}{dy\,dy}=\left\|\frac{\partial(u,v)}{\partial(x,y)}\right\|^{-1}p(x,y) \tag{2.23}$$

where we have used the usual Jacobian determinant

$$\left\|\frac{\partial(u,v)}{\partial(x,y)}\right\|=\left|\det\begin{pmatrix}\frac{\partial u}{\partial x} & \frac{\partial u}{\partial y}\\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y}\end{pmatrix}\right| \tag{2.24}$$

from advanced calculus. This is needed to ensure that

$$p((u,v)\in R)=\iint\limits_R p(u,v)\,du\,dv \tag{2.25}$$

since the area element transforms like

$$du\,dv=\left\|\frac{\partial(u,v)}{\partial(x,y)}\right\|\,dx\,dy \tag{2.26}$$

The generalization to a joint pdf of many variables $p(\{x_\alpha\})$ is straightforward.

Note that the joint pdf can be written in terms of the conditional pdf; the product rule from probability translates into

$$p(x,y)=p(x|y)\,p(y)\ . \tag{2.27}$$

The factors of dx and dy work out because $p(x|y)$ is a density in x (but not y), $p(y)$ is a density in y, and $p(x,y)$ is a density in both.

If x and y are independent, this simplifies, because

$$p(x|y)=p(x)\qquad \text{if } x \text{ and } y \text{ are independent} \tag{2.28}$$

so that

$$p(x,y)=p(x)\,p(y)\qquad \text{if } x \text{ and } y \text{ are independent}\ . \tag{2.29}$$

In general, though, we can find the pdf of one variable by integrating out (another word for this is *marginalizing*) over the other one:

$$p(x)=\int_{-\infty}^{\infty}p(x,y)\,dy\ . \tag{2.30}$$

This can be seen by applying (2.21) to the region

$$x_0 < x < x_0 + dx \tag{2.31a}$$

$$-\infty < y < \infty \ . \tag{2.31b}$$

Another special case is when y is completely determined by x, i.e., $y = y(x)$. This deterministic situation is represented by the conditional pdf

$$p(y|x) = \delta(y - y(x)) \qquad \text{if } y \text{ is determined by } x \ ; \tag{2.32}$$

in that case

$$p(x, y) = \delta(y - y(x))\, p(x) \qquad \text{if } y \text{ is determined by } x \ . \tag{2.33}$$

In general, though, the relationship represented by $p(x, y)$ is more complicated.

2.3.1 Nuisance Parameters

The outcome of a measurement may depend both on physical parameters of interest and also on other parameters we don't care about, known as *nuisance parameters.* We generally want to marginalize the relevant joint probability distributions over the nuisance parameters.

As a concrete example, return to our toy experiment which returns a single number y, but let that number depend now on the physical quantity of interest x but also on a nuisance parameter λ, so that our experiment is modelled by the likelihood function $p(y|x, \lambda)$.

From a Bayesian perspective, we can still apply Bayes's theorem, but now it gives us a joint posterior pdf on x and the nuisance parameter λ:

$$p(x, \lambda|y) = \frac{p(y|x, \lambda)\, p(x, \lambda)}{p(y)} \tag{2.34}$$

where now we need to know the joint prior $p(x, \lambda)$. The normalization is given by

$$p(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y|x, \lambda)\, p(x, \lambda)\, dx\, d\lambda \ . \tag{2.35}$$

To marginalize, we just calculate

$$p(x|y) = \int_{-\infty}^{\infty} p(x, \lambda|y)\, d\lambda \tag{2.36}$$

We can also marginalize at the level of the likelihood, in terms of the conditional probability

$$p(\lambda|x) = \frac{p(x, \lambda)}{p(x)} \tag{2.37}$$

we can do this by calculating

$$p(y|x) = \int_{-\infty}^{\infty} p(y|x, \lambda)\, p(\lambda|x)\, d\lambda \tag{2.38}$$

This is particularly convenient if x and λ are independent (e.g., if λ represents an unknown property of the instrument which is uncorrelated with the physical observable of interest) so that $p(\lambda|x) = p(\lambda)$.

2.3.2 Example: Calibration Uncertainty

Suppose y is an estimator of x, with an associated statistical error of σ (assumed to be known), but that the overall calibration is uncertain, so that we're actually estimating the product of x and an unknown λ:

$$p(y|x,\lambda) = \frac{e^{-(y-\lambda x)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \; ; \tag{2.39}$$

Suppose that λ is known to be close to 1, but with an uncertainty σ_λ, so that

$$p(\lambda|x) = p(\lambda) = \frac{e^{-(\lambda-1)^2/2\sigma_\lambda^2}}{\sigma_\lambda\sqrt{2\pi}} \; . \tag{2.40}$$

With these pdfs, we can actually do the marginalization analytically:

$$p(y|x) = \frac{1}{2\pi\sigma\sigma_\lambda}\int_{-\infty}^{\infty} \exp\left(-\frac{(y-\lambda x)^2}{2\sigma^2} - \frac{(\lambda-1)^2}{2\sigma_\lambda^2}\right) d\lambda \; . \tag{2.41}$$

Since the argument of the exponential is

$$\begin{aligned}
-\frac{(y-\lambda x)^2}{2\sigma^2} - \frac{(\lambda-1)^2}{2\sigma_\lambda^2} &= -\frac{(\sigma^2+x^2\sigma_\lambda^2)\lambda^2 - 2(\sigma^2+xy\sigma_\lambda^2)\lambda + (\sigma^2+y^2\sigma_\lambda^2)}{2\sigma^2\sigma_\lambda^2} \\
&= -\frac{\sigma^2+x^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2}\left(\lambda - \frac{\sigma^2+xy\sigma_\lambda^2}{\sigma^2+x^2\sigma_\lambda^2}\right)^2 + \frac{(\sigma^2+xy\sigma_\lambda^2)^2}{2\sigma^2\sigma_\lambda^2(\sigma^2+x^2\sigma_\lambda^2)} - \frac{\sigma^2+y^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2} \\
&= -\frac{\sigma^2+x^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2}\lambda'^2 + \frac{(\sigma^2+xy\sigma_\lambda^2)^2 - (\sigma^2+x^2\sigma_\lambda^2)(\sigma^2+y^2\sigma_\lambda^2)}{2\sigma^2\sigma_\lambda^2(\sigma^2+x^2\sigma_\lambda^2)} \\
&= -\frac{\sigma^2+x^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2}\lambda'^2 - \frac{(x^2+y^2-2xy)\sigma^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2(\sigma^2+x^2\sigma_\lambda^2)} \\
&= -\frac{\sigma^2+x^2\sigma_\lambda^2}{2\sigma^2\sigma_\lambda^2}\lambda'^2 - \frac{(y-x)^2}{2(\sigma^2+x^2\sigma_\lambda^2)}
\end{aligned} \tag{2.42}$$

we can evaluate the integral and get

$$p(y|x) = \frac{e^{-(y-x)^2/2(\sigma^2+x^2\sigma_\lambda^2)}}{\sqrt{2\pi(\sigma^2+x^2\sigma_\lambda^2)}} \tag{2.43}$$

This is still a Gaussian, but now instead of σ, its width is $\sqrt{\sigma^2+x^2\sigma_\lambda^2}$. We can think of σ as the statistical error and $x\sigma_\lambda$ as the systematic error associated with the calibration.

Tuesday, November 3, 2009

3 "Maximum Posterior" Parameter Estimation

3.1 Single-Parameter Case

Suppose we've done an experiment and, based on the data D collected (we've been calling this $\hat{y}$, but we'd like to generalize beyond the case of a single number), the posterior pdf for some parameter x is $p(x|D)$. How do we distill that result to an estimate of x and our uncertainty of x? Well, we could use the expectation value to talk about a mean value

$$\langle x \rangle_D = \int_{-\infty}^{\infty} x\, p(x|D)\, dx \tag{3.1}$$

and a variance

$$\left\langle (x - \langle x \rangle_D)^2 \right\rangle_D \tag{3.2}$$

but that's sometimes harder to calculate than it is to state in the abstract. In particular, when we generalize to the case of many parameters, we can quickly end up with multi-dimensional integrals that are computationally expensive to evaluate.

Another thing we could consider is which value of x is most likely given the results of the experiment, i.e., the x which maximizes the posterior pdf. We can call this $\hat{x}$, defined by

$$\forall x : p(\hat{x}|D) \geq p(x|D)\ . \tag{3.3}$$

This is often called the "maximum likelihood estimate", although in this case we're actually maximizing the posterior rather than the likelihood. (There's an equivalent frequentist approach which works with the likelihood, and is equivalent to assuming a uniform prior in x.) To define an uncertainty Δx associated with this, we could do something like requiring that x be so likely to fall in the interval $\hat{x} - \Delta x < x < \hat{x} + \Delta x$ according to the posterior. But again, this would require integrating over the posterior, which we can't always do. But note that this would be an integral near the maximum of the posterior, which is motivation for an approximation: expand the posterior about its maximum.

Now, it's actually not such a good idea to expand $p(x|D)$ itself about $x = \hat{x}$ because we know $p(x|D)$ can't ever be negative, and in fact we expect it to be close to zero if we go far away from the maximum. Instead, what we want to do is expand its logarithm,

$$L(x) = \ln p(x|D)\ . \tag{3.4}$$

taking advantage of the fact that as $p(x|D) \to 0$, $L(x) \to -\infty$. So we write the Taylor series as

$$L(x) = L(\hat{x}) + (x - \hat{x})\, L'(\hat{x}) + \frac{(x - \hat{x})^2}{2}\, L''(\hat{x}) + \dots \tag{3.5}$$

Since $x = \hat{x}$ is a maximum of $p(x|D)$ and therefore of $L(x)$, we know that $L'(\hat{x}) = 0$ and $L''(\hat{x}) < 0$, so to lowest non-trivial order

$$L(x) \approx L(\hat{x}) - \frac{(x - \hat{x})^2}{2}\, [-L''(\hat{x})] \tag{3.6}$$

and

$$p(x|D) \approx p(\hat{x}|D) \exp\left(\frac{-(x - \hat{x})^2}{2[\sqrt{-1/L''(\hat{x})}]^2} \right)\ , \tag{3.7}$$

so the posterior is approximated near its maximum by a Gaussian of width $\sqrt{-1/L''(\hat{x})}$. If it were actually equal to a Gaussian, the expectation value of x would be $\hat{x}$ and the expected variance of x would be $-1/L''(\hat{x})$. So $\sqrt{-1/L''(\hat{x})}$ is an estimate in the "one-sigma error" associated with the estimate $\hat{x}$.

3.2 Multiple-Parameter Case

Now consider the case where the model has multiple parameters $\{x_\alpha\}$; writing the vector of parameters as $\mathbf{x}$, the expansion of the log-posterior about its maximum at $\mathbf{x} = \hat{\mathbf{x}}$ is

$$\ln p(\mathbf{x}|D) = L(\mathbf{x}) \approx L(\hat{\mathbf{x}}) - \frac{1}{2} \sum_\alpha \sum_\beta \left(- \frac{\partial^2 L}{\partial x_\alpha \partial x_\beta} \right)_{\mathbf{x}=\hat{\mathbf{x}}} (x_\alpha - \hat{x}_\alpha)(x_\beta - \hat{x}_\beta) \; . \tag{3.8}$$

If we define a matrix $\mathbf{F}$ with elements

$$F_{\alpha\beta} = - \frac{\partial^2 L}{\partial x_\alpha \partial x_\beta} \tag{3.9}$$

we can write this approximation to the posterior as

$$p(\mathbf{x}|D) \approx p(\hat{\mathbf{x}}|D) \exp \left(- \frac{1}{2} (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}} \, \mathbf{F} \, (\mathbf{x} - \hat{\mathbf{x}}) \right) \tag{3.10}$$

The matrix $\mathbf{F}$ is called the *Fisher information matrix*, and its inverse provides a measure of the variance and covariance of the $\{x_\alpha\}$:

$$\left\langle (\mathbf{x} - \hat{\mathbf{x}}) \, (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}} \right\rangle \approx \mathbf{F}^{-1} \tag{3.11}$$

(this would be exact if the posterior really were Gaussian).

Near the maximum, curves of constant posterior will be ellipsoids in the $\{x_\alpha\}$ space. You can see that by noting that since the Fisher matrix $\mathbf{F}$ is a real, symmetric matrix, it has a full set of real eigenvalues $\{f_\alpha\}$ with orthonormal eigenvectors $\{\mathbf{u}_\alpha\}$, and we can write it as

$$\mathbf{F} = \sum_\alpha \mathbf{u}_\alpha \, f_\alpha \mathbf{u}_\alpha^{\mathrm{T}} \tag{3.12}$$

and so

$$p(\mathbf{x}|D) \approx p(\hat{\mathbf{x}}|D) \exp \left(- \sum_\alpha \frac{f_\alpha}{2} (\xi_\alpha)^2 \right) \tag{3.13}$$

where

$$\xi_\alpha = \mathbf{u}_\alpha^{\mathrm{T}} (\mathbf{x} - \hat{\mathbf{x}}) \tag{3.14}$$

the equation

$$\sum_\alpha \frac{f_\alpha}{2} (\xi_\alpha)^2 = \text{constant} \tag{3.15}$$

defines an ellipsoid in the $\{\xi_\alpha\}$ coördinates with axes proportional to $1/\sqrt{f_\alpha}$, and the $\{\xi_\alpha\}$ are just a different set of coördinates rotated relative to $\{x_\alpha - \hat{x}_\alpha\}$. These error ellipses mean that it's important to use the diagonal elements of the inverse Fisher matrix for error estimates rather than just taking one over the diagonal elements of the Fisher matrix itself. I.e., the error estimate for x_α is

$$\left\langle (x_\alpha - \hat{x}_\alpha)^2 \right\rangle \approx (F^{-1})_{\alpha\alpha} \neq \frac{1}{F_{\alpha\alpha}} \tag{3.16}$$

3.3 Example: Scale Parameter

Return to the example we considered last week of the likelihood

$$p(y|x,\lambda) = \frac{e^{-(y-\lambda x)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \tag{3.17}$$

with a prior of

$$p(\lambda) = \frac{e^{-(\lambda-1)^2/2\sigma_\lambda^2}}{\sigma_\lambda\sqrt{2\pi}} \tag{3.18}$$

on the scale parameter λ. If we assume a uniform prior on x, the joint posterior for x and λ after a measurement $\hat{y}$ will be

$$p(x,\lambda|\hat{y}) \propto e^{-(y-\lambda x)^2/2\sigma^2} e^{-(\lambda-1)^2/2\sigma_\lambda^2} \tag{3.19}$$

where we won't bother working out the proportionality constant because it's independent of x and λ. The log-posterior

$$L(x,\lambda) = \ln p(x,\lambda|\hat{y}) = -\frac{(\hat{y}-\lambda x)^2}{2\sigma^2} - \frac{(\lambda-1)^2}{2\sigma_\lambda^2} + \text{const} \tag{3.20}$$

is maximized by

$$\hat{x} = \hat{y} \tag{3.21a}$$

$$\hat{\lambda} = 1 \tag{3.21b}$$

; the second partial derivatives are

$$\frac{\partial^2 L}{\partial x^2} = -\frac{\lambda^2}{\sigma^2} \tag{3.22a}$$

$$\frac{\partial^2 L}{\partial x \partial \lambda} = -\frac{2\lambda x - \hat{y}}{\sigma^2} \tag{3.22b}$$

$$\frac{\partial^2 L}{\partial \lambda^2} = -\frac{x^2}{\sigma^2} - \frac{1}{\sigma_\lambda^2} \tag{3.22c}$$

so the Fisher matrix is

$$\mathbf{F} = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{\hat{y}}{\sigma^2} \\ \frac{\hat{y}}{\sigma^2} & \frac{\hat{y}^2}{\sigma^2} + \frac{1}{\sigma_\lambda^2} \end{pmatrix} = \sigma^{-2}\sigma_\lambda^{-2} \begin{pmatrix} \sigma_\lambda^2 & \hat{y}\sigma_\lambda^2 \\ \hat{y}\sigma_\lambda^2 & \hat{y}^2\sigma_\lambda^2 + \sigma^2 \end{pmatrix} \tag{3.23}$$

so the posterior is approximated near its maximum by

$$p(x,\lambda|\hat{y}) \approx p(x=\hat{y},\lambda=1|\hat{y}) \exp\left[-\frac{1}{2\sigma^2\sigma_\lambda^2} \begin{pmatrix} x-\hat{y} & \lambda-1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \sigma_\lambda^2 & \hat{y}\sigma_\lambda^2 \\ \hat{y}\sigma_\lambda^2 & \hat{y}^2\sigma_\lambda^2 + \sigma^2 \end{pmatrix} \begin{pmatrix} x-\hat{y} & \lambda-1 \end{pmatrix}\right] . \tag{3.24}$$

The inverse Fisher matrix is

$$\mathbf{F}^{-1} = \begin{pmatrix} \sigma^2 + \hat{y}^2\sigma_\lambda^2 & -\hat{y}\sigma_\lambda^2 \\ -\hat{y}\sigma_\lambda^2 & \sigma_\lambda^2 \end{pmatrix} . \tag{3.25}$$

Note that in particular

$$(F^{-1})_{xx} = \sigma^2 + \hat{y}^2\sigma_\lambda^2 \tag{3.26}$$

while

$$\frac{1}{F_{xx}} = \sigma^2 \; ; \tag{3.27}$$

the difference is equivalent to the difference between marginalizing over the nuisance parameter λ and just assuming it takes on its most likely value.

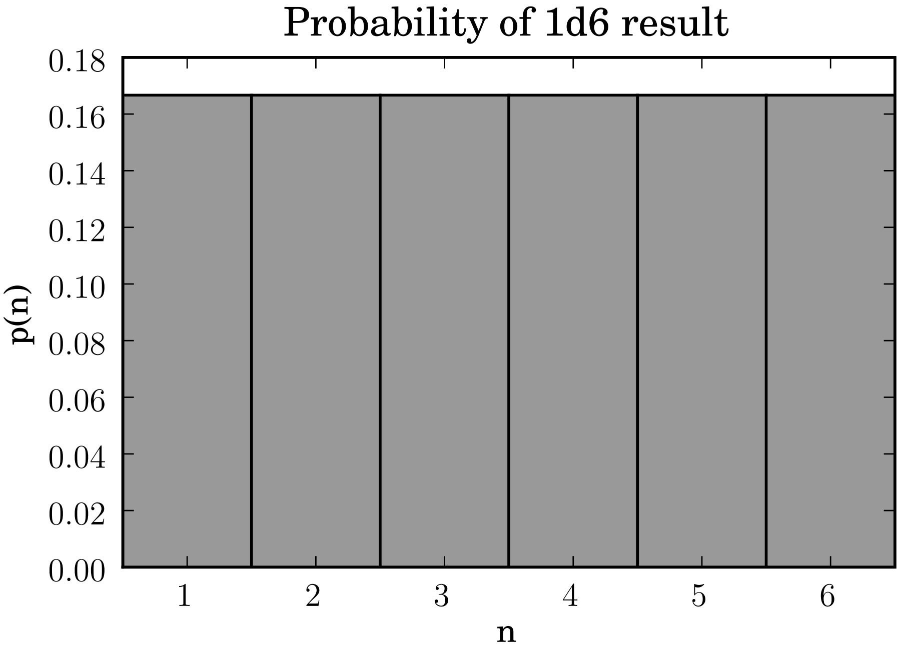
Thursday, November 5, 2009

4 The Central Limit Theorem

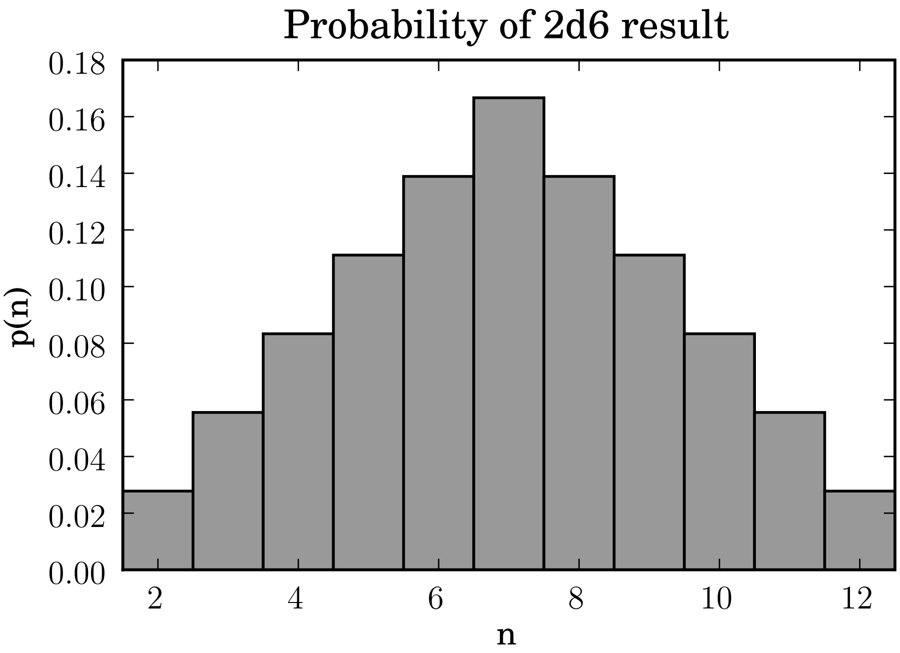
4.1 Anecdotal Evidence

Having seen that a general pdf can be approximated by a Gaussian near its maximum, we now examine a principle that often causes physical quantities to have Gaussian pdfs. This is the *Central Limit Theorem*, which says that the average of many independent, identically distributed variables has a distribution which is approximately Gaussian.

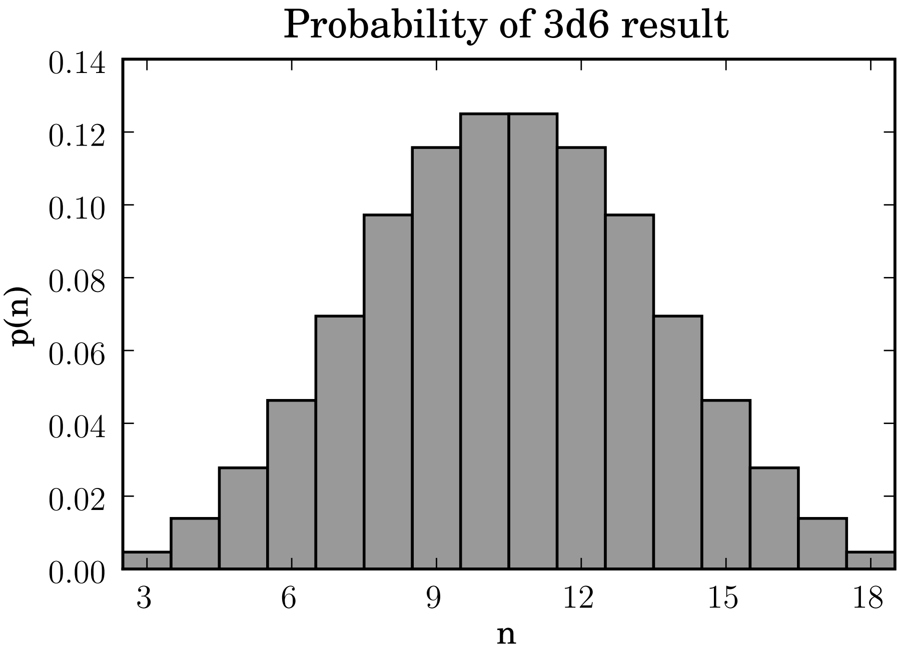
To see a simple example of how the distribution of a sum of random quantities approaches a Gaussian, consider rolling a fair six-sided die. The results are uniformly distributed:



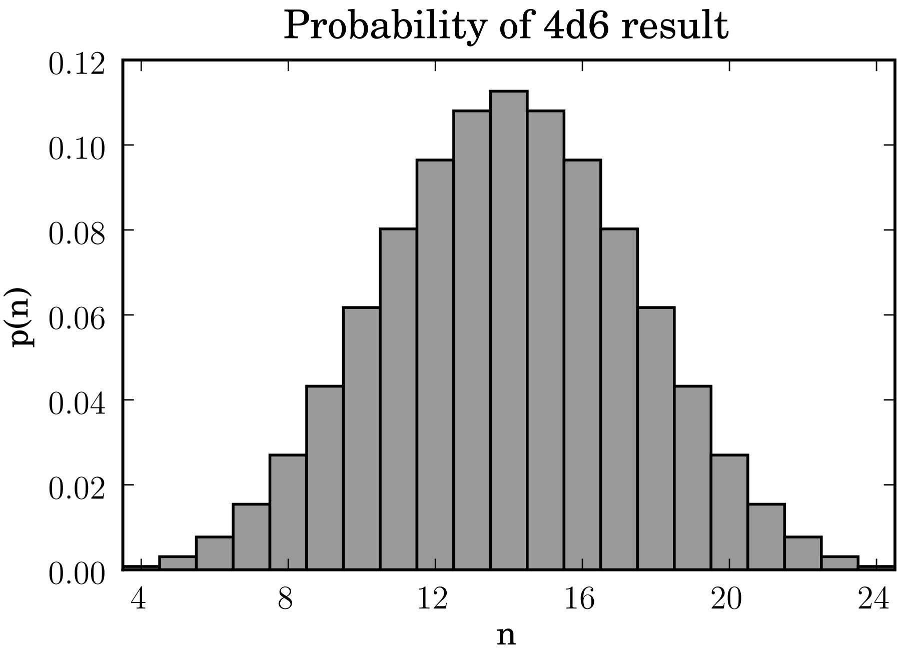
If we roll two dice and add the results, we get a non-uniform distribution, with results close to 7 being most likely, and the probability distribution declining linearly from there:

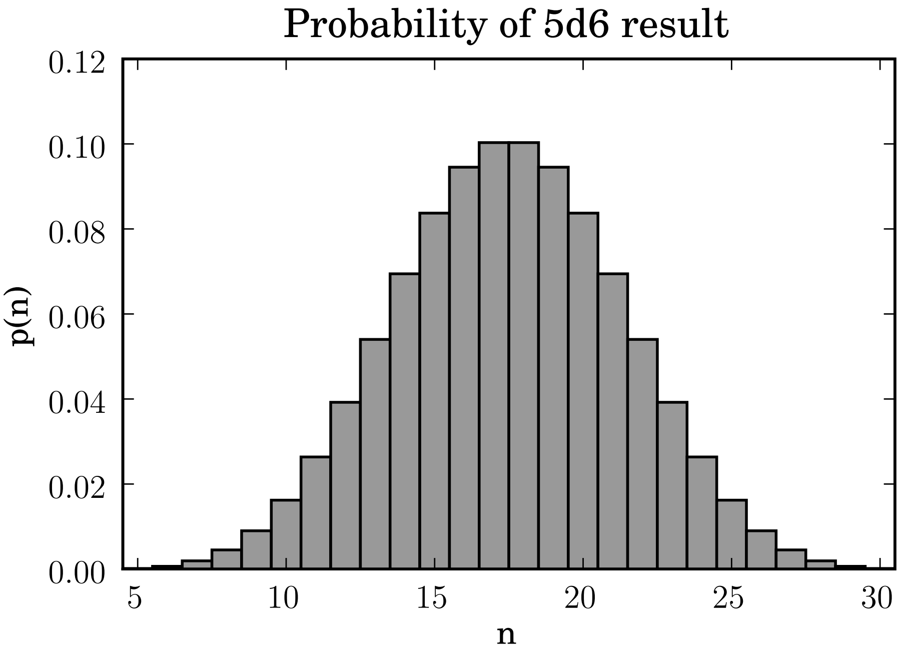


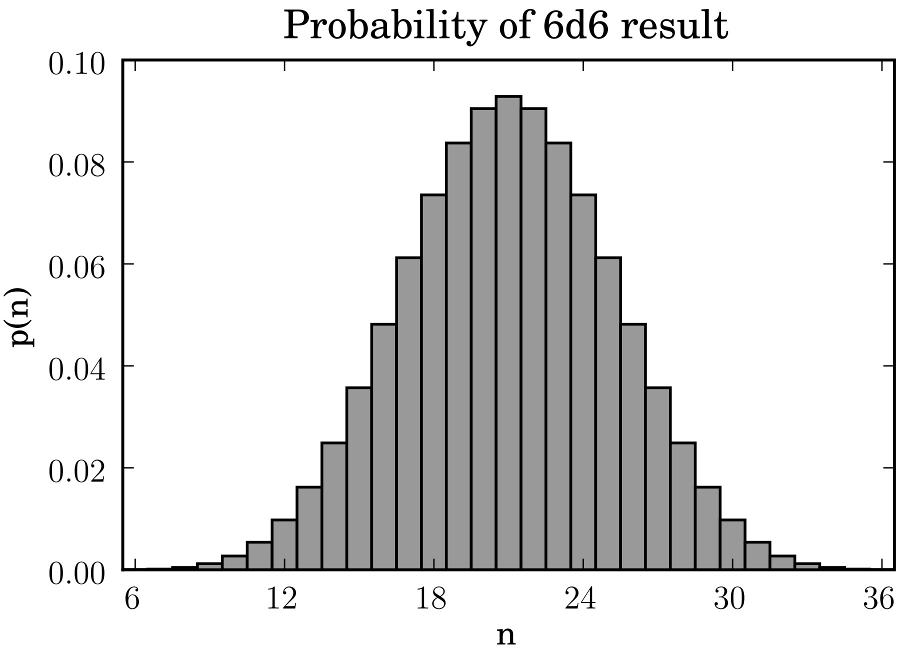
If we add three dice, the distribution begins to take on the shape of a “bell curve” and in fact such a random distribution is used to approximate the distribution of human physical properties in some role-playing games:



Adding more and more dice produces histograms that look more and more like a Gaussian:


Probability of 4d6 result
0.12
0.10
0.08
0.06
0.04
0.02
0.00
p(n)
4
8
12
16
20
24
n


Probability of 5d6 result
0.12
0.10
0.08
0.06
0.04
0.02
0.00
p(n)
5
10
15
20
25
30
n



4.2 Averages of Identical Variables

You considered on the homework the example of N identically distributed random variables $\{x_k\}$ with expectation values

$$\langle x_k \rangle = \mu \qquad \text{and} \qquad \langle (x_k - \mu)(x_n - \mu) \rangle = \delta_{kn}\, \sigma^2 \tag{4.1}$$

and found that the average

$$\overline{x} = \frac{1}{N} \sum_{k=0}^{N-1} x_k \tag{4.2}$$

had expected mean

$$\langle \overline{x} \rangle = \mu \tag{4.3}$$

and variance

$$\left\langle (\overline{x} - \mu)^2 \right\rangle = \frac{\sigma^2}{N} \ . \tag{4.4}$$

This is easy to see if we define

$$z_k = \frac{x_k - \mu}{\sigma} \tag{4.5}$$

so that

$$\langle z_k \rangle = 0 \qquad \text{and} \qquad \langle z_k z_n \rangle = \delta_{kn} \ . \tag{4.6}$$

Then if we define

$$Z = \sum_{k=0}^{N-1} z_k = \frac{1}{\sigma} \sum_{k=0}^{N-1} (x_k - \mu) = N \frac{\overline{x} - \mu}{\sigma} \tag{4.7}$$

we find $\langle Z \rangle = 0$ and

$$\left\langle Z^2 \right\rangle = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} \langle z_k z_n \rangle = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} \delta_{kn} = N = N^2 \frac{\langle (\overline{x} - \mu)^2 \rangle}{\sigma^2} \ . \tag{4.8}$$

If we want to think about adding a bunch of copies of something and taking the limit as $N \to \infty$, we should choose our quantity so that the expected mean and variance remain well behaved in that limit, so we define

$$\mathcal{Z} = \frac{Z}{\sqrt{N}} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} z_k = \frac{1}{\sqrt{N}\sigma} \sum_{k=0}^{N-1} (x_k - \mu) \tag{4.9}$$

which obeys

$$\langle \mathcal{Z} \rangle = 0 \qquad \text{and} \qquad \left\langle \mathcal{Z}^2 \right\rangle = 1 \ . \tag{4.10}$$

If we define $\mathcal{X} = \frac{X}{\sqrt{N}} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k$ we find

$$\langle \mathcal{X} \rangle = \mu\sqrt{N} \qquad \text{and} \qquad \left\langle (\mathcal{X} - \mu\sqrt{N})^2 \right\rangle = \sigma^2 \ . \tag{4.11}$$

4.3 PDF of a Sum of Random Variables

If we consider two independent random variables x_0 and x_1 with (not necessarily identical) pdfs $p_0(x_0)$ and $p_1(x_1)$, so that their joint pdf is

$$p(x_0, x_1) = p_0(x_0) p_1(x_1) \tag{4.12}$$

and write their sum as

$$X = x_0 + x_1 \tag{4.13}$$

we can ask what the pdf $p(X)$ is. One way to approach this[2] is to consider the joint pdf $p(X, x_1)$. We can do this by changing variables from x_0 to $X = x_0 + x_1$; since we're actually changing from (x_0, x_1) to (X, x_1), we can treat x_1 as a constant[3] and since $dX = dx_0$ in that case,

$$p(X, x_1) = p_0(X - x_1) p_1(x_1) \ . \tag{4.14}$$

if we then marginalize over x_1, we find

$$p(X) = \int_{-\infty}^{\infty} p_0(X - x_1)\, p_1(x_1)\, dx_1 \tag{4.15}$$

and so we see that *the pdf of a sum of variables is the convolution of their individual pdfs.*

Of course, since convolutions map onto products under the Fourier transform, that means the Fourier transform of the pdf of a sum of variables is the product of the Fourier transforms

[2] An alternative, slick shortcut is to write

$$p(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(X - [x_0 + x_1])\, p(x_0, x_1)\, dx_0\, dx_1$$

[3] Alternatively, we can consider the Jacobian determinant

$$\left\| \frac{\partial(X, x_1)}{\partial(x_0, x_1)} \right\| = \left| \det \begin{pmatrix} \left(\frac{\partial X}{\partial x_0}\right)_{x_1} & \left(\frac{\partial X}{\partial x_1}\right)_{x_0} \\ \left(\frac{\partial x_0}{\partial x_0}\right)_{x_1} & \left(\frac{\partial x_0}{\partial x_1}\right)_{x_0} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right| = 1$$

.

of their individual pdfs. The Fourier transform of a pdf is a very handy thing, known as the *characteristic function* for that random variable:

$$\Phi_x(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi x\xi}\, p(x)\, dx \tag{4.16}$$

4.3.1 Properties of the Characteristic Function

- $\Phi_x(\xi) = \left\langle e^{-i2\pi x\xi} \right\rangle$. This is often given as the definition of the characteristic function, but it seems more natural to think of the Fourier transform of the pdf.

- $\Phi_x(0) = 1$. This is apparent from the normalization:

$$\Phi_x(0) = \int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{4.17}$$

- If we Taylor expand the exponential, we get the moments of the distribution, i.e., the expectation values of powers of x:

$$\Phi_x(\xi) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(-i2\pi)^n}{n!}\, x^n\, \xi^n\, p(x)\, dx = \sum_{n=0}^{\infty} \frac{(-i2\pi)^n}{n!}\, \langle x^n \rangle\, \xi^n \tag{4.18}$$

- If $p(x)$ is a Gaussian,

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{4.19}$$

  then

$$\Phi_x(\xi) = \exp\left(-i2\pi\mu\xi - \frac{(2\pi\xi)^2}{2\sigma^{-2}} \right) \tag{4.20}$$

  which we get as usual by completing the square in the integral over x.

- $$\Phi_{ax}(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi ax\xi}\, p(ax)\, d(ax) = \int_{-\infty}^{\infty} e^{-i2\pi ax\xi}\, p(x)\, dx = \Phi_x(a\xi) \tag{4.21}$$

4.4 Proof of the Central Limit Theorem

We are now ready to show that if the $\{x_j\}$ are independent identically distributed variables, whatever the pdf, $\mathcal{Z}$ is Gaussian-distributed in the limit $N \to \infty$. We do this by showing that the characteristic function $\Phi_{\mathcal{Z}}(\xi)$ becomes the characteristic function for a Gaussian in the limit of large N.

First, note that since $Z = \sum_{k=0}^{N-1} z_k$,

$$\Phi_Z(\xi) = [\Phi_z(\xi)]^N \tag{4.22}$$

(since the pdf $p(Z)$ can be made by convolving together N copies of $p(z)$). Now, since $\mathcal{Z} = Z/\sqrt{N}$,

$$\Phi_{\mathcal{Z}}(\xi) = \Phi_Z(\xi/\sqrt{N}) = [\Phi_z(\xi/\sqrt{N})]^N\ . \tag{4.23}$$

For large N, we can Taylor expand

$$\Phi_z(\xi/\sqrt{N}) = 1 - \frac{i2\pi\xi\,\langle z_k\rangle}{\sqrt{N}} - \frac{(2\pi\xi)^2\,\langle (z_k)^2\rangle}{2N} + \mathcal{O}(N^{-3/2}) \tag{4.24}$$

where $\mathcal{O}(N^{-3/2})$ represents terms which go to zero for large N at least as fast as $N^{-3/2}$. Since we've constructed the $\{z_k\}$ so that $\langle z_k\rangle = 0$ and $\langle (z_k)^2\rangle = 1$, this becomes

$$\Phi_z(\xi/\sqrt{N}) = 1 - \frac{(2\pi\xi)^2}{2N} + \mathcal{O}(N^{-3/2}) \tag{4.25}$$

and so

$$\Phi_{\mathcal{Z}}(\xi) = \left(1 - \frac{(2\pi\xi)^2}{2N} + \mathcal{O}(N^{-3/2})\right)^N = \left(1 - \frac{(2\pi\xi)^2}{2N}\right)^N + \mathcal{O}(N^{-1/2}) \tag{4.26}$$

and

$$\lim_{N\to\infty} \Phi_{\mathcal{Z}}(\xi) = \lim_{N\to\infty} \left(1 - \frac{(2\pi\xi)^2}{2N}\right)^N = \exp\left(-\frac{(2\pi\xi)^2}{2}\right) . \tag{4.27}$$

But this is just the characteristic function for a Gaussian of zero mean and unit variance, so

$$p(\mathcal{Z}) \overset{N\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} e^{-\mathcal{Z}^2/2} \tag{4.28}$$

and since

$$\mathcal{Z} = \frac{X - N\mu}{\sigma\sqrt{N}} \tag{4.29}$$

that means that for large N,

$$p(X) \approx \frac{1}{\sigma\sqrt{2\pi N}} \exp\left(\frac{-(X-N\mu)^2}{2N\sigma^2}\right) \tag{4.30}$$

Tuesday, November 10, 2009

5 Probabilities for Discrete Numbers of Events

5.1 Binomial Distribution

Consider a random event that has a probability of α of occurring in a given trial (e.g., detection of a simulated signal by an analysis pipeline, where α is the efficiency), so that

$$p(1|\alpha, 1) = p(Y|\alpha) = \alpha \tag{5.1a}$$

$$p(0|\alpha, 1) = p(N|\alpha) = 1 - \alpha \tag{5.1b}$$

We write $p(k|\alpha, n)$ as the probability that if we do n trials, we will find a "yes" result in k of them. For n trials, there are 2^n possible sequences of yes and no results. The probability of a particular sequence of k yes and $n-k$ no results is $\alpha^k(1-\alpha)^{n-k}$, and the number of such sequences for a given k and n is "n choose k", $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, so the probability of exactly k "yes" results in n trials is

$$p(k|\alpha, n) = \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \tag{5.2}$$

You are currently exploring the consequences of this distribution in problem set 8.

5.2 Poisson Distribution

Now consider a process occurring over continuous time such that:

- The average rate of events per unit time is r;
- the number of events that occurs in a time interval $t_0 < t < t_1$ depends only on the duration $t_1 - t_0$.

If we want to calculate the pdf $p(k|r,T)$ that exactly k events occur in time T, we can derive this as the limit of the binomial distribution. (See Section 4.7 of Gregory for an alternative derivation.)

Subdivide the interval into N small intervals of duration $\delta t = T/N$. The average number of events in one of these intervals will be $r\,\delta t$. If we choose $\delta t \ll r^{-1}$, so that the expected number of events is very small, it will be very unlikely that we see any events at all, and phenomenally unlikely that we see more than one event in a time δt. So to lowest order in $r\,\delta t$ we expect

$$p(0|r,\delta t) = 1 - r\,\delta t + \mathcal{O}([r\,\delta t]^2) \tag{5.3a}$$

$$p(1|r,\delta t) = r\,\delta t + \mathcal{O}([r\,\delta t]^2) \tag{5.3b}$$

$$\sum_{k=2}^{\infty} p(k|r,\delta t) = \mathcal{O}([r\,\delta t]^2) \tag{5.3c}$$

But in the limit $\delta t \to 0$ this is just a single trial with a probability $r\,\delta t$ of a "yes" result. Since each infinitesimal interval of time is independent of each other, the probability of k events in the N trials is then given by the binomial distribution

$$\begin{aligned} p(k|r,T) &= \lim_{\delta t\to 0} p(k|r\delta t, T/\delta t) = \lim_{N\to\infty} p(k|rT/N, N) = \lim_{N\to\infty} \frac{N!}{(N-k)!k!}\left(\frac{rT}{N}\right)^k \left(1-\frac{rT}{N}\right)^{N-k} \\ &= \frac{(rT)^k}{k!} \lim_{N\to\infty}\left(1-\frac{rT}{N}\right)^N \frac{N!}{(N-k)!}\,(N-rT)^{-k} \end{aligned} \tag{5.4}$$

Now,

$$\frac{N!}{(N-k)!} = N(N-1)\ldots(N-k+1) = \prod_{\ell=0}^{k-1} N-\ell \tag{5.5}$$

and of course

$$(N-rT)^{-k} = \prod_{\ell=0}^{k-1} \frac{1}{N-rT} \tag{5.6}$$

so

$$\frac{N!}{(N-k)!}\,(N-rT)^{-k} = \prod_{\ell=0}^{k-1} \frac{N-\ell}{N-rT} = \prod_{\ell=0}^{k-1} \frac{1-\ell/N}{1-rT/N} \tag{5.7}$$

but for finite k this is the product of a finite number of things, each of which goes to 1 as $N \to \infty$, so

$$p(k|r,T) = \frac{(rT)^k}{k!} \lim_{N\to\infty}\left(1-\frac{rT}{N}\right)^N = \frac{(rT)^k}{k!} e^{-rT}\ . \tag{5.8}$$

This is the *Poisson distribution.* It's easy to check that it's normalized, i.e.,

$$\sum_{k=0}^{\infty} p(k|r,T) = e^{-rT} \sum_{k=0}^{\infty} \frac{(rT)^k}{k!} = e^{-rT} e^{rT} = 1 \ . \tag{5.9}$$

Note that as a consistency check we can go back and verify our assumptions (5.3):

$$p(0|r,\delta t) = e^{-r\delta t} = 1 - r\,\delta t + \mathcal{O}([r\,\delta t]^2) \tag{5.10a}$$

$$p(1|r,\delta t) = r\,\delta t\, e^{-r\delta t} = r\,\delta t + \mathcal{O}([r\,\delta t]^2) \tag{5.10b}$$

$$\sum_{k=2}^{\infty} p(k|r,\delta t) = 1 - p(0|r,\delta t) - p(1|r,\delta t) = \mathcal{O}([r\,\delta t]^2) \tag{5.10c}$$

5.2.1 Application: Estimating Event Rates

Suppose we have observed a Poisson process, with an unknown event rate r, for a time T, and seen $\hat{k}$ events. We can use Bayes's theorem to construct a posterior pdf on the rate r:

$$p(r|k,T) = \frac{p(k|r,T)\,p(r)}{p(k|T)} \tag{5.11}$$

Uniform Prior If we use a uniform prior, $p(r) = \text{constant}$, on the rate, we get a posterior

$$p(r|\hat{k},T,\text{uniform prior}) = \frac{T(rT)^{\hat{k}}\,e^{-rT}}{\hat{k}!} \ . \tag{5.12}$$

Exercise: check that $p(r|\hat{k},T,\text{uniform prior})$ is normalized, and that it has a maximum at $\hat{k}/T$.

Jeffreys Prior On the other hand, a uniform prior doesn't make all that much sense for an event rate. It means that a priori, we think it's equally likely for the rate to be between 1 and 2 per hour as it is to be between 3600 and 3601 per hour, i.e., between 1 and 1.00028 per second. If we really know nothing about the rate, it might be more reasonable to say we expect it to be equally likely to be between 1 and 2 per hour as it is to be between 1 and 2 per second. This is a Jeffreys prior, uniform in $\ln r$ rather than in r, i.e.,

$$\frac{dp}{d\ln r} = r\frac{dp}{dr} = r\,p(r) = \text{constant} \tag{5.13}$$

so that

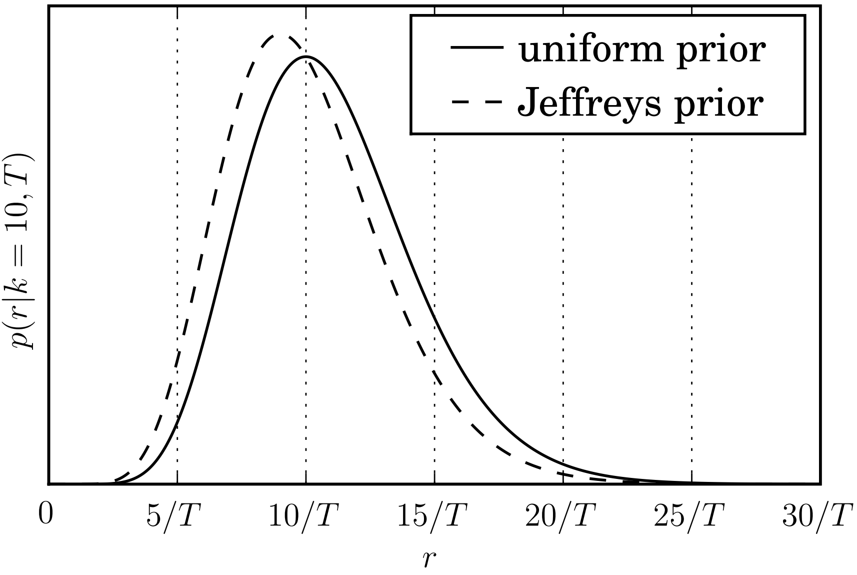
$$p(r) \propto \frac{1}{r} \ . \tag{5.14}$$

Starting from this prior, we get the normalized posterior

$$p(r|\hat{k},T,\text{Jeffreys prior}) = \frac{T(rT)^{\hat{k}-1}\,e^{-rT}}{(\hat{k}-1)!} \ . \tag{5.15}$$

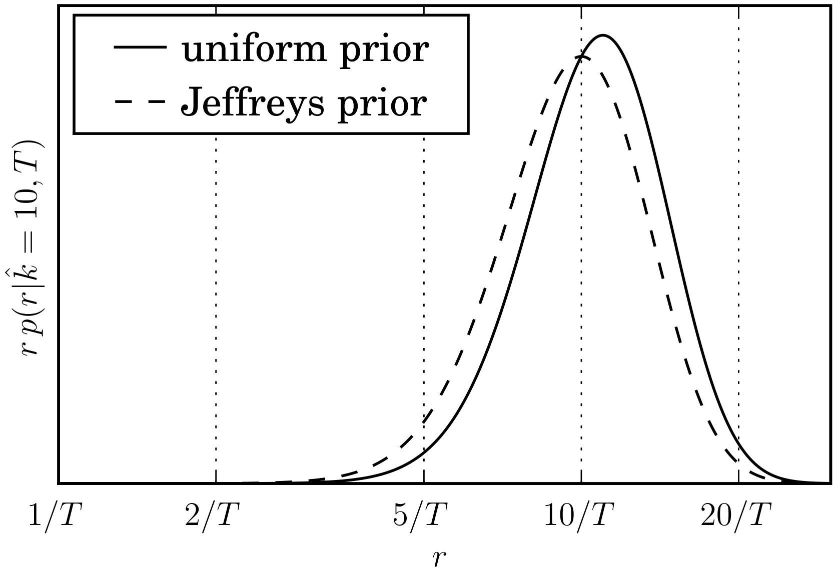
This seems like kind of a funny result: the posterior given $\hat{k}$ events and a Jeffreys prior is the same as the posterior given $\hat{k}-1$ events and a uniform prior. In particular, it's peaked at $(\hat{k}-1)/T$ rather than $\hat{k}/T$.

Comparison of Posteriors We can see this by plotting the two posteriors for the specific value $\hat{k} = 10$:



The posterior from the uniform prior is peaked at $r = 10/T$ while that for the Jeffreys prior is peaked at $r = 9/T$.

The explanation is that, in the perspective that led us to the Jeffreys prior, we should not be looking at the pdf $p(r|\hat{k}, T)$, which is a density in r. We chose a prior which was a uniform density in $\ln r$, and by the same logic we should be interested in the posterior per logarithmic rate interval, $p(\ln r|\hat{k}, T) = r\, p(r|\hat{k}, T)$,[4] and then plot r on a log scale:



[4] Note that the normalization on this is

$$\int_{-\infty}^{\infty} p(\ln r|\hat{k}, T)\, d\ln r = \int_{0}^{\infty} p(\ln r|\hat{k}, T)\, \frac{dr}{r} = 1$$

.

when plotted this way, we see that the posterior corresponding to the Jeffreys prior,

$$p(\ln r|\hat{k}, T, \text{Jeffreys prior}) = r\frac{T(rT)^{\hat{k}-1}\, e^{-rT}}{(\hat{k}-1)!} = \frac{(rT)^{\hat{k}}\, e^{-rT}}{(\hat{k}-1)!} \tag{5.16}$$

is indeed peaked at $r = 10/T$, while

$$p(\ln r|\hat{k}, T, \text{uniform prior}) = r\frac{T(rT)^{\hat{k}}\, e^{-rT}}{\hat{k}!} = \frac{(rT)^{\hat{k}+1}\, e^{-rT}}{\hat{k}!} \tag{5.17}$$

is peaked at $r = 11/T$.

The moral of the story is to be careful about identifying the variable in which your pdfs are densities!