Notes on Statistical Methods

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

Fall 2010

Contents

[*]

Tuesday, October 5, 2010

1 Fundamentals of Probability and Statistics

See Gregory, Chapters 1 and 2

In science, especially observational science, our state of knowledge is often incomplete. While a lot of standard models are described in a deterministic way (given these conditions, this *will* happen; given this piece of data, this *is* the underlying physical situation), there are often cases where it is impractical or impossible to model every element of the problem deterministically, or where we would like to make useful statements based on incomplete information. These sorts of statements involve taking a logical proposition (I will roll a six on this six-sided die, the mass of Jupiter lies in the following range, I will measure a temperature within 0.1 degrees of the true temperature, etc), and rather than stating definitively true or false, instead assigning a number between 0 and 1 to that proposition, which can be thought of as our degree of certainty or belief in that statement, or the *probability* of it being true. For statements regarding the outcome of repeatable experiments, the probability can also be thought of as the expected fraction of trials (in the limit of infinite trials) in which the statement is true.

Some examples of probabilistic statements:

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 it's often useful 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 Bayesian and Frequentist Interpretations of Probability

A distinction can be made between two kinds of probability:

- In the *frequentist* approach, probabilities are assigned to the possible outcomes of random processes. We describe a system according to certain models and parameters, and then predict the outcome of experiments and measurements, assigning probabilities to different alternatives. It's called "frequentist" because if we prepare a large ensemble of identical systems (with the same parameters), the frequency with which a given measurement outcome is observed should be approximately equal to the probability assigned to it.

- In the *Bayesian* approach, a probability is a number reflecting our degree of certainty in a proposition. A typical situation of interest is one in which we've done a measurement and observed a specific outcome, and use that information to assign probabilities to e.g., different parameter values. This approach has been considered philosophically problematic, because the parameters of a system have specific values, even if those are unknown. However, it is closer to what one actually does as an observational scientist.

Note that these two approaches are the probabilistic analogues of two procedures that are used in deterministic situations:

- *Deductive reasoning* says that if certain models and parameter values hold, a system will behave in a specific way. E.g., if I know the masses and initial conditions in a planetary system, I can predict the trajectories.

- *Inductive reasoning*[1] uses measurements, e.g., the trajectory of a system, to determine the underlying fundamental parameter values.

When things are deterministic, you can use observations to infer parameter values with absolute certainty, or use models and parameters to predict evolution exactly. The difference in a probabilistic situation is that, instead of a statement about parameter values or trajectories being true (probability=1) or false (probability=0), it is assigned a "probability" between 0 and 1.

1.2 Rules of Probability

Whatever the interpretation of our "probabilities", they are numbers assigned to logical propositions. Logical propositions behave like sets, and you can think of them as being partitions of the set of all possible arbitrarily-detailed "outcomes". We formally refer to

[1] This is not quite the standard definition of inductive reasoning, but it's useful to give it a name.

propositions with letters like A and B, so that $P(A)$ is the probability that A is true, $P(B)$ is the probability that B is true, etc.[2]

You can combine logical propositions with the same sorts of operations that are used for combining sets:

- Negation. $\overline{A}$ is true if A is false, and vice-versa. In words, we can think of $\overline{A}$ as "not A". (Other notations include A' and $\neg A$.)

- Intersection. A, B is true if A and B are both true. In words, this is "A and B". (Other notations include $A \cap B$.) The advantage of the comma is that $P(A, B)$ is the probability that both A and B are true.

- Union $A + B$ is true if either A or B (or both) is true. In words, this is "A or B". (Other notations include $A \cup B$.) 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. It is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{1.1}$$

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

$$P(A) = P(A, B) + P(A, \overline{B}) \tag{1.2a}$$

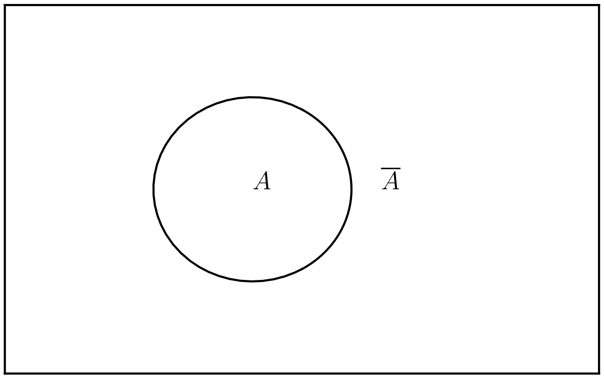
$$P(B) = P(A, B) + P(\overline{A}, B) \tag{1.2b}$$

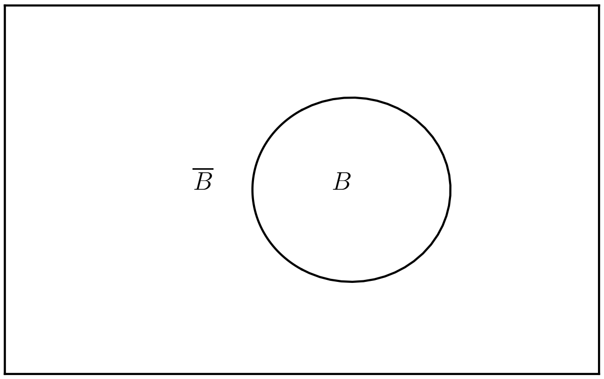
$$P(A + B) = P(A, B) + P(A, \overline{B}) + P(\overline{A}, B) \ , \tag{1.2c}$$

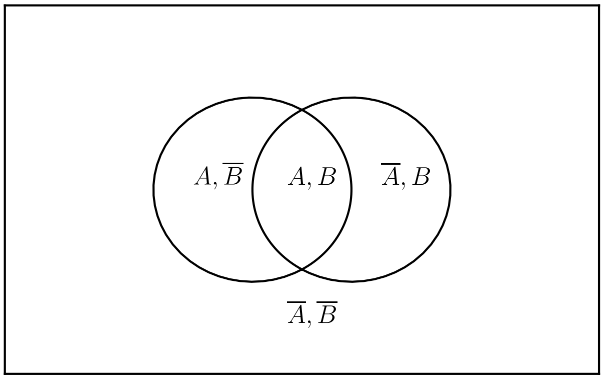
[2] I will, at least in this section, use a capital P to refer to the probability of a proposition, which is slightly different from Gregory's convention of using a lowercase p.

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

Alternatively, we can illustrate (1.2) using Venn diagrams:







The alternative $A+B$ is made up of the disjoint alternatives A, B, $A, \overline{B}$, and $\overline{A}, B$, i.e.,

$$A + B = A, B + A, \overline{B} + \overline{A}, B \tag{1.3}$$

so the probability of $A+B$ is the sum of those three probabilities.

1.3 Bayes's Theorem

Because the intersection operation is symmetric (i.e., A, B is the same as B, A), we can write $P(A, B)$ two different ways, in terms of two different conditional probabilities:

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

If we divide through by A we find

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1.5}$$

this seemingly trivial result is known as Bayes's Theorem, and has remarkably deep consequences. It turns out that we often would like to know one conditional probability and would like to know the opposite one. Suppose we have a hypothesis H which may be either true or false, and we have done an experiment which returned a result in a particular range, which we write as D for "data". We can model the experiment and write the *likelihood function* $P(D|H)$, i.e., the probability that the experiment would turn out a certain way if our hypothesis were true. But the question we really want to answer as scientists is this: given that we observed the data D, what is the likelihood that our hypothesis H is true? I.e., what is $P(H|D)$. Well, Bayes's theorem tells us that

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{1.6}$$

or, written more descriptively.

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \ . \tag{1.7}$$

Now, of course, the tricky part is assigning a value to $P(H)$, the *prior probability* of our hypothesis being true, without the knowledge gained from our observation. $P(H|D)$ is called the *posterior probability* of the hypothesis being true, given the additional information about the outcome D of the observation. We also have to work out the term in the denominator, $P(D)$, the overall probability of making the observation we did, whether or not the hypothesis is true. But this is straightforward, since

$$P(D) = P(D,H) + P(D,\overline{H}) = P(D|H)P(H) + P(D,\overline{H})P(\overline{H}) \tag{1.8}$$

So if we have the experiment modelled so that we know $P(D|H)$ and $P(D|\overline{H})$, and we know the prior $P(H)$, we can calculate $P(\overline{H}) = 1 - P(H)$, and the denominator is no problem.

1.3.1 Example: Disease Testing

The classic example application of Bayes's theorem is a test for a disease. Suppose that one one-thousandth of the population has a disease. There is a test that can detect the disease, but it has a 2% false positive rate (on average one out of fifty healthy people will test positive) and as 1% false negative rate (on average one out of one hundred sick people will test negative). The question we ultimately want to answer is: if someone gets a positive test result, what is the probability that they actually have the disease. Note, it is not 98%!

The statement of the problem tells us the following probabilities for a randomly chosen individual:

$$P(\text{sick}) = 0.001 \tag{1.9}$$

$$P(\text{pos}|\text{healthy}) = 0.02 \tag{1.10}$$

$$P(\text{neg}|\text{sick}) = 0.01 \tag{1.11}$$

and Bayes's theorem lets us construct the probability we want:

$$P(\text{sick}|\text{pos}) = \frac{P(\text{pos}|\text{sick})P(\text{sick})}{P(\text{pos})} \tag{1.12}$$

The factor $P(\text{pos}|\text{sick})$ in the denominator, the probability that a sick person will test positive, is

$$P(\text{pos}|\text{sick}) = 1 - P(\text{neg}|\text{sick}) = 0.99 \tag{1.13}$$

(Note that it is **NOT** $1 - P(\text{pos}|\text{healthy})$.) The denominator is the overall probability of getting a positive test result, either because you're sick and get an accurate test result, or because you're healthy and get a false positive. This is

$$P(\text{pos}) = P(\text{pos}|\text{sick})P(\text{sick}) + P(\text{pos}|\text{healthy})P(\text{healthy}) = (0.99)(0.001) + (0.02)(0.999) \tag{1.14}$$

so that

$$P(\text{sick}|\text{pos}) = \frac{(0.99)(0.001)}{(0.99)(0.001) + (0.02)(0.999)} = \frac{.00099}{.00099 + .01998} = \frac{0.0099}{0.02097} \approx 0.04721 \tag{1.15}$$

So only about 4.7% of people who test positive have the disease. It's a lot more than one in a thousand, but a lot less than 98% or 99%.

Approach Considering a Hypothetical Population *S*ome of the arguments in this section are adapted from
`http://yudkowsky.net/rational/bayes`
which gives a nice explanation of Bayes's theorem.

The standard treatment of Bayes's Theorem and the Law of Total Probability can be sort of abstract, so it's useful to keep track of what's going on by considering a hypothetical population which tracks the various probabilities. So, assume the probabilities arise from a population of 100,000 individuals. Of those, one one-one-thousandth, or 100, have the disease. The other 99,900 do not. The 2% false positive rate means that of the 99,900 healthy individuals, 2% of them, or 1,998, will test positive. The other 97,902 will test negative. The 1% false negative rate means that of the 100 sick individuals, one will test negative and the other 99 will test positive. So let's collect this into a table:

	Positive	Negative	Total
Sick	99	1	100
Healthy	1,998	97,902	99,900
Total	2,097	97,903	100,000

(As a reminder, if we choose a *sample* of 100,000 individuals out of a larger population, we won't expect to get exactly this number of results, but the 100,000-member population is a useful conceptual construct.)

Translating from numbers in this hypothetical population, we can confirm that it captures the input information:

$$P(\text{sick}) = \frac{100}{100,000} = .001 \tag{1.16a}$$

$$P(\text{positive}|\text{healthy}) = \frac{1,998}{99,900} = .02 \tag{1.16b}$$

$$P(\text{negative}|\text{sick}) = \frac{1}{100} = .01 \tag{1.16c}$$

But now we can also calculate what we want, the conditional probability of being sick given a positive result. That is the fraction of the total number of individuals with positive test results that are in the "sick *and* positive" category:

$$P(\text{sick}|\text{positive}) = \frac{99}{2,097} \approx .04721 \tag{1.17}$$

or about 4.7%.

1.4 Probability Distributions

See Gregory, Chapters 5

1.4.1 Discrete Random Variables and Probability Mass Functions

We now return to the concept of a random variable. For simplicity, we look first at a discrete random variable X. If x is a specific value which that variable could take, the statement $X = x$ is a logical proposition which can be assigned a probability $P(X = x)$. The probability distribution or *probability mass function* (pmf)

$$p_X(x) = P(X = x) \tag{1.18}$$

gives us all the information we need about the random variable X. We will often write this as simply $p(x)$ if the random variable in question is clear from the context.

Since there is a complete set of mutually exclusive alternatives in which X takes on each of its possible values, the pmf $p(x)$ obeys the *normalization condition*

$$\sum_x p(x) = \sum_x P(X = x) = 1 \ . \tag{1.19}$$

We can use the pmf to define the expectation value (which we used last week)

$$\langle X \rangle = \sum_x x\, p(x) \tag{1.20}$$

or more generally for some function $g()$,

$$\langle g(X) \rangle = \sum_x g(x)\, p(x) \tag{1.21}$$

If there are multiple random variables, say X and Y, we can define a joint probability distribution

$$p(x, y) = P(X = x, Y = y) \tag{1.22}$$

which obeys or more generally for some function $g()$,

$$\langle g(X,Y)\rangle = \sum_x \sum_y g(x,y)\, p(x,y) \tag{1.23}$$

and

$$\sum_x \sum_y p(x,y) = 1 \tag{1.24}$$

Sometimes we know the joint pmf $p(x,y)$ but are not interested in the behavior of Y. In that case, we can *marginalize* over the possible values of Y to obtain

$$\begin{aligned} p_X(x) = P(X=x) = P(X=x,[Y=y_1]+[Y=y_2]+\ldots) &= \sum_y P(X=x,Y=y) \\ &= \sum_y p(x,y) \end{aligned} \tag{1.25}$$

A parameter over which we wish to marginalize is often called a *nuisance parameter*.

1.4.2 Continuous Random Variables and Probability Density Functions

In the case of a continuous random variable X, which can take on a range of values, the probability of having any one value x is vanishingly small. The probability

$$P(x < X < x+dx) \tag{1.26}$$

that X lies in an interval of width dx should be proportional to the width of the interval, so that

$$f_X(x) = \lim_{dx\to 0} \frac{P(x<X<x+dx)}{dx} \tag{1.27}$$

(also known as $f(x)$) is well-behaved. This is called the *probability density function*. It's sometimes useful to emphasize the density nature by writing $f(x)$ as $\frac{dP}{dx}$. This is to be thought of as a density, not a derivative.

Since

$$f(x)\,dx \approx P(x<X<x+dx) \tag{1.28}$$

we can translate the expressions involving discrete random variables into the corresponding expressions involving continuous random variables by replacing sums with integrals. Specifically

- Expectation value

$$\langle g(X)\rangle = \int_{-\infty}^{\infty} g(x)\,p(x)\,dx \tag{1.29}$$

- Normalization

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{1.30}$$

- Joint probability density

$$f(x,y) \equiv \frac{d^2P}{dx\,dy} = \lim_{dx,dy\to 0} \frac{P(x<X<x+dx, y<Y<y+dy)}{dx\,dy} \tag{1.31}$$

- Marginalization

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \tag{1.32}$$

Note that a probability density function has the appropriate units for a density. If X has units of energy, then $f(x) = \frac{dP}{dx}$ has units of one over energy, because probability is dimensionless.

Thursday, October 7, 2010

1.4.3 Change of Variables in Probability Distributions

Imagine you have a random variable X and another random variable $Y = h(X)$ whose value is given by acting on the random value of X with the deterministic function $h()$. How can we determine the probability distribution for Y from the probability distribution for X?

Well, if they're discrete random variables, things are pretty straightforward:

$$p_Y(h(x)) = P(Y = h(x)) = P(X = x) = p_X(x) \tag{1.33}$$

The pmf for Y has the same value as the pmf for X; you just have to evaluate it at the appropriate value.

Things get more interesting, though, for continuous random variables, since the pdf is a density, and not the probability of a specific value. You can look at section 1.4 of last year's notes for a more detailed treatment, but the appropriate transformation is suggested by the notation:

$$\frac{dP}{dy} = \frac{\frac{dP}{dx}}{\left|\frac{dy}{dx}\right|} \tag{1.34}$$

i.e.,

$$f_Y(h(x)) = \frac{f_X(x)}{|h'(x)|} \tag{1.35}$$

Why the absolute value? Because the probability density for X and Y is defined to be positive, even if the transformation is such that Y decreases with increasing X. (Basically, it's a property of the way densities transform.)

You can also perform this sort of transformation on a joint probability density. Suppose you have N random variables $\{X_i\}$ from which you can determine the values of N random variables $\{Y_i\}$. Then the transformation uses the Jacobian determinant:

$$\frac{d^N P}{d^N y} = \left(\frac{d^N P}{d^N x}\right) \bigg/ \left|\det\left\{\frac{\partial y_i}{\partial x_j}\right\}\right| \tag{1.36}$$

You can see this is the right thing to do because the Jacobian determinant is used to transform the measure of a multiple integral:

$$d^N y = \left|\det\left\{\frac{\partial y_i}{\partial x_j}\right\}\right| d^N x \tag{1.37}$$

and these probability densities are meant to be put under multiple integrals. Written in more standard notation, if we define

$$\mathbf{x} \equiv \{x_i\}, \qquad \mathbf{y} \equiv \{y_i\} \tag{1.38}$$

and

$$J_{\mathbf{yx}}(\mathbf{x}) = \det \left\{ \frac{\partial y_i}{\partial x_j} \right\} \tag{1.39}$$

then

$$f_{\mathbf{Y}}(\mathbf{h}(\mathbf{x})) = \frac{f_{\mathbf{X}}(\mathbf{x})}{|J_{\mathbf{yx}}(\mathbf{x})|} \tag{1.40}$$

1.5 Mathematical Interlude: the Error Function and the Gamma Function

See Arfken & Weber, Chapter 8, especially Section 8.1

When we examine specific probability distributions next week, it'll be useful to have at our disposal a couple of definite integrals. The Gamma function is defined for $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{1.41}$$

along with the incomplete Gamma functions

$$\gamma(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt \tag{1.42a}$$

$$\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1} e^{-t} dt \tag{1.42b}$$

Meanwhile, the error function and complementary error function are defined, respectively, as

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-x^2} dx \tag{1.43a}$$

$$\operatorname{erfc}(a) = \frac{2}{\sqrt{\pi}} \int_a^\infty e^{-x^2} dx \tag{1.43b}$$

You can use integration by parts[3] to show that

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \; ; \tag{1.44}$$

This, together with

$$\Gamma(1) = \int_0^\infty e^{-t} dt = -e^{-t}\Big|_0^\infty = -0 + 1 = 1 \tag{1.45}$$

is enough to imply that for any non-negative integer n,

$$\Gamma(n+1) = n! \tag{1.46}$$

In a sense, the integral form (1.41) is an extension of the factorial function to non-integer values.

Also of some interest is

$$\Gamma(1/2) = \int_0^\infty \frac{e^{-t}}{\sqrt{t}} dt \tag{1.47}$$

[3] or parametric differentiation; see last year's math notes

If we change variables from t to $x = \sqrt{t}$, so that $dt = d(x^2) = 2x\,dx$, we see that we can also write

$$\Gamma(1/2) = 2\int_0^\infty e^{-x^2} dx = \int_{-\infty}^\infty e^{-x^2} dx = \sqrt{\pi}\ \mathrm{erf}(\infty) = \sqrt{\pi}\ \mathrm{erfc}(0)\ . \tag{1.48}$$

This is a very famous integral, which can be done by a cool trick, using conversion from Cartesian to polar coördinates in the plane:

$$\begin{aligned}\Big(\Gamma(1/2)\Big)^2 &= \left(\int_{-\infty}^\infty e^{-x^2} dx\right)\left(\int_{-\infty}^\infty e^{-y^2} dy\right) = \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-x^2-y^2} dx\,dy \\ &= \int_0^{2\pi}\int_0^\infty e^{-r^2} r\,dr\,d\phi = 2\pi\int_0^\infty e^{-u}\frac{du}{2} = \pi\end{aligned} \tag{1.49}$$

(where in the last step we've used the substitution $u = r^2$; $du = 2r\,dr$) so

$$\Gamma(1/2) = \sqrt{\pi} \tag{1.50}$$

and therefore

$$\mathrm{erf}(\infty) = \mathrm{erfc}(0) = 1 \tag{1.51}$$

This also means that

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) \tag{1.52}$$

Tuesday, October 12, 2010 – Midterm Exam on Mathematical Methods

Thursday, October 14, 2010

1.6 Some Specific Probability Distributions

See Gregory, Chapter 5

Before we delve into frequentist and Bayesian applications of probability distributions, it's useful to consider some specific random variables, the sort of physical situations to which they're relevant, and what their probability distributions look like.

1.6.1 The Binomial Distribution (discrete)

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

$$p(1|\alpha, 1) = P(Y|\alpha) = \alpha \tag{1.53a}$$

$$p(0|\alpha, 1) = P(N|\alpha) = 1 - \alpha \tag{1.53b}$$

We define the random variable K as the total number of "yes" events we find in n trials. We write $p(k|\alpha, n)$ is the probability mass function for this random variable. It's the 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

$$p(k|\alpha, n) \equiv b(k|\alpha, n) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k} \tag{1.54}$$

You are currently exploring the consequences of this distribution in problem set 5. In particular, you show that the mean is

$$\langle K \rangle = n\alpha \tag{1.55}$$

and the variance is

$$\langle K^2 \rangle - \langle K \rangle^2 = n\alpha(1-\alpha) \tag{1.56}$$

1.6.2 The Poisson Distribution (discrete)

Consider a random process in which a discrete number of events will occur in some interval, and that number has an expected value of λ. We usually think of this as some process with a rate r in units of inverse time, and then λ is the rate times the observing time. For example, popcorn kernels popping, clicks on a Geiger counter, or gamma-ray bursts observed. But the interval could also be in space, e.g., we could be counting the number of cosmic rays collected in a detector of a certain area, or the number of galaxies within a certain redshift range found in a patch of sky of a given solid angle. The number K of events is a random variable with a probability mass function

$$P(K=k) = p(k|\lambda) \tag{1.57}$$

Such a process is called a *Poisson process* if we can sub-divide the interval into smaller intervals (in time, space, sky position, or whatever) and then the number of events in each sub-interval is an independent random variable with the same properties (but a smaller rate, obviously).

We can calculate the pmf for the *Poisson random variable* K as follows: subdivide the interval into some large number N of identically-sized sub-intervals. Each one has an expected number of events of λ/N. If we choose N large enough, we can make this number very very small. This means that in any one sub-interval, the odds are pretty good that there will be no events. There is some small chance (of order λ/N) of seeing one event, and a vanishingly small chance (of order $[\lambda/N]^2$) of seeing two or more events in this sub-interval:

$$p\left(0\,|\lambda/N\right) = 1 - \lambda/N + \mathcal{O}\left([\lambda/N]^2\right) \tag{1.58a}$$

$$p\left(1\,|\lambda/N\right) = \lambda/N + \mathcal{O}\left([\lambda/N]^2\right) \tag{1.58b}$$

$$\sum_{k=2}^{\infty} p\left(k\,|\lambda/N\right) = \mathcal{O}\left([\lambda/N]^2\right) \tag{1.58c}$$

but this is basically a single trial which can have a yes (there is an event) or no (there is not an event) result, and the total number K of events in the larger interval can be approximated by a binomial random variable with N trials and a probability for success of λ/N for each trial. That means

$$\begin{aligned} p(k|\lambda) &= \lim_{N\to\infty} b(k|\lambda/N, N) = \lim_{N\to\infty} \frac{N!}{(N-k)!k!}\left(\frac{\lambda}{N}\right)^k \left(1-\frac{\lambda}{N}\right)^{N-k} \\ &= \frac{(\lambda)^k}{k!} \lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^N \frac{N!}{(N-k)!} \left(N-\lambda\right)^{-k} \end{aligned} \tag{1.59}$$

Now,

$$\frac{N!}{(N-k)!} = N(N-1)\dots(N-k+1) = \prod_{\ell=0}^{k-1}(N-\ell) \tag{1.60}$$

and of course

$$(N-\lambda)^{-k} = \prod_{\ell=0}^{k-1}\frac{1}{N-\lambda} \tag{1.61}$$

so

$$\frac{N!}{(N-k)!}\,(N-\lambda)^{-k} = \prod_{\ell=0}^{k-1}\frac{N-\ell}{N-\lambda} = \prod_{\ell=0}^{k-1}\frac{1-\ell/N}{1-\lambda/N} \tag{1.62}$$

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

$$p(k|r,T) = \frac{(\lambda)^k}{k!}\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N = \frac{(\lambda)^k}{k!}e^{-\lambda}\ . \tag{1.63}$$

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

$$\sum_{k=0}^{\infty}p(k|\lambda) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!} = e^{-\lambda}e^{\lambda} = 1\ . \tag{1.64}$$

Note that as a consistency check we can go back and verify our assumptions (1.58):

$$p(0|\lambda/N) = e^{-\lambda/N} = 1-\lambda/N+\mathcal{O}([\lambda/N]^2) \tag{1.65a}$$

$$p(1|\lambda/N) = \frac{\lambda}{N}\,e^{-\lambda/N} = \frac{\lambda}{N}+\mathcal{O}([\lambda/N]^2) \tag{1.65b}$$

$$\sum_{k=2}^{\infty}p(k|\lambda/N) = 1-p(0|\lambda/N)-p(1|\lambda/N) = \mathcal{O}([\lambda/N]^2) \tag{1.65c}$$

By considering the limiting form of the binomial distribution, we can find the mean and variance of the Poisson distribution:

$$\langle K\rangle = \lambda \tag{1.66a}$$

$$\left\langle K^2\right\rangle - \langle K\rangle^2 = \lambda \tag{1.66b}$$

1.6.3 The Exponential Distribution (continuous)

Let's consider further a Poisson process. The Poisson distribution gives the pmf for the total number of events in an interval, which is a discrete random variable. Now consider another question. Suppose we are observing a Poisson process with an event rate r. Let's assume the intervals are in time, so that r has units of inverse time, and the number of events in a time Δt will be a Poisson random variable with parameter $r\,\Delta t$. Now suppose we start watching at a given time and see how long we have to wait for the next event. This waiting time T will itself be a random variable, with a probability density function $f_T(t|r)$ which depends on the rate r. Note that this is a *continuous* random variable. We can actually work out the pdf from our knowledge of the Poisson process. Consider the probability that T is longer than some value t:

$$P(T > t) = \int_t^\infty f_T(t'|r)\, dt' \tag{1.67}$$

This is the probability that in the interval of length t, beginning when we start watching, there are no events. But we know how to write the probability that there are no events from a Poisson process in an interval of a given length. It is the probability that the corresponding Poisson random variable (which has parameter rt) will take on the value 0:

$$P(K = 0) = p(0|rt) = \frac{(rt)^0}{0!} e^{-rt} = e^{-rt} \tag{1.68}$$

Equating the two expressions for this probability gives

$$\int_t^\infty f_T(t'|r)\, dt' = e^{-rt} \tag{1.69}$$

We can differentiate both sides with respect to t (*not* t') and find

$$- f_T(t|r) = -r\, e^{-rt} \tag{1.70}$$

which gives us the pdf for T, the *exponential distribution*:

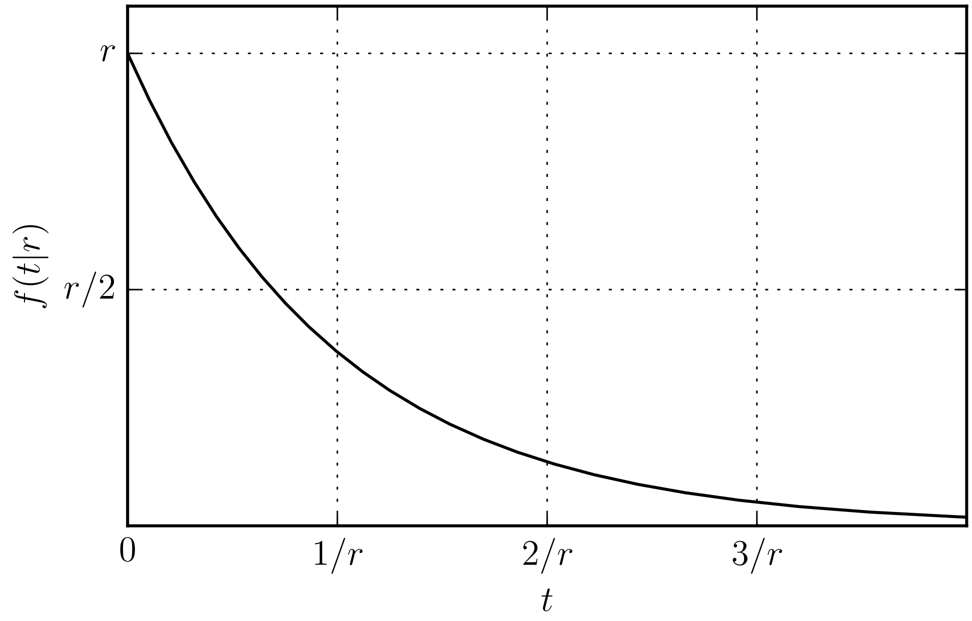
$$f(t|r) = r\, e^{-rt} \qquad t \geq 0 \tag{1.71}$$

It's straightforward to show that $f(t|r)$ is normalized, and that the mean and variance are

$$\langle T \rangle = \frac{1}{r} \tag{1.72a}$$

$$\langle T^2 \rangle - \langle T \rangle^2 = \frac{1}{r^2} \tag{1.72b}$$

The pdf looks like this:



Note that this derivation made use of an integral of the pdf. For a general continuous random variable, we define the *cumulative distribution function*

$$F_X(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(x')\, dx' \tag{1.73}$$

The derivative of the cdf is the pdf:

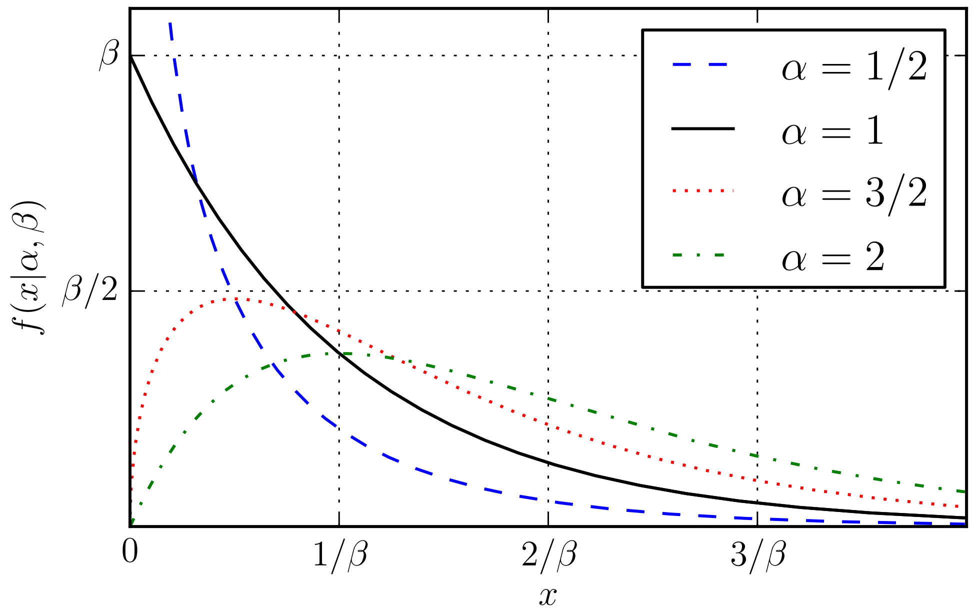
$$\frac{dF_X}{dx}(x) = f_X(x) \tag{1.74}$$

1.6.4 The Gamma Distribution (continuous)

The plot of the exponential distribution above has the same shape regardless of the value of the parameter r; changing its value just changes the scales of the axes. It is, however, one member of a family of distributions known as the *Gamma distribution*, with pdf

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\beta x} \qquad x \geq 0 \tag{1.75}$$

(Gregory uses the other parametrization, defining $\theta = 1/\beta$.) Here's the shape of the pdf for different choices of α.



It's a straightforward exercise to show that if X is a gamma-distributed random variable with parameters α and β then the mean is

$$\langle X \rangle = \frac{\alpha}{\beta} \tag{1.76}$$

and the variance is

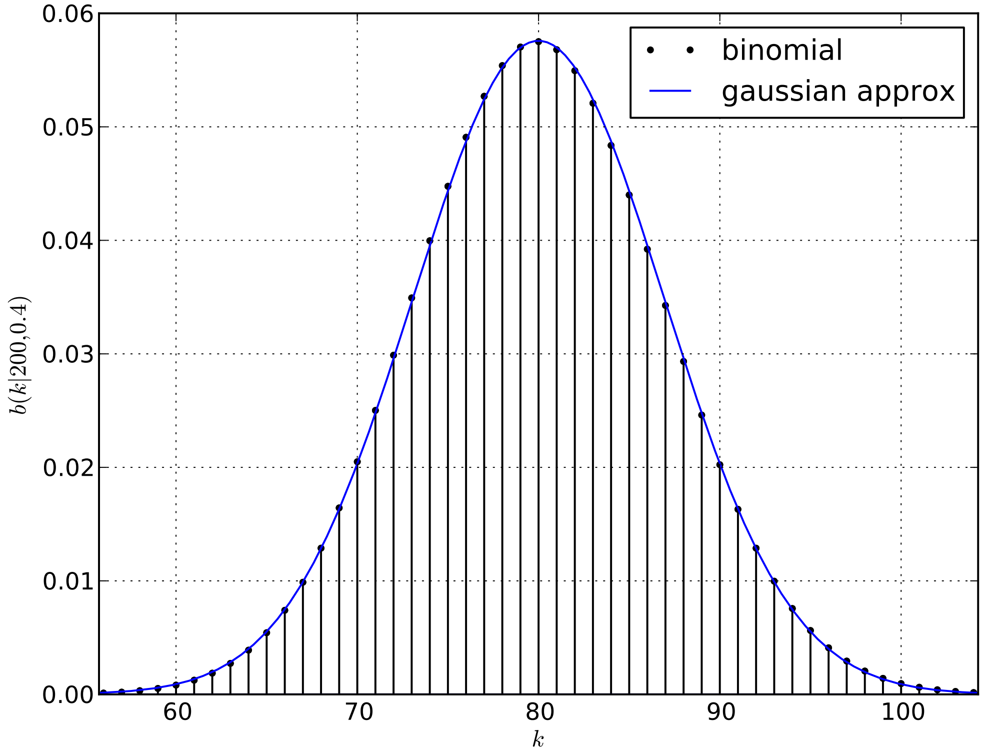
$$\langle X^2 \rangle - \langle X \rangle^2 = \frac{\alpha}{\beta^2} \tag{1.77}$$

An exponential distribution is a special case of the Gamma distribution with $\alpha = 0$ and $\beta = r$.

1.6.5 The Gaussian (aka Normal) Distribution (continuous)

If we consider the pmfs for the binomial distribution in the case where the number of trials n is large:[4]

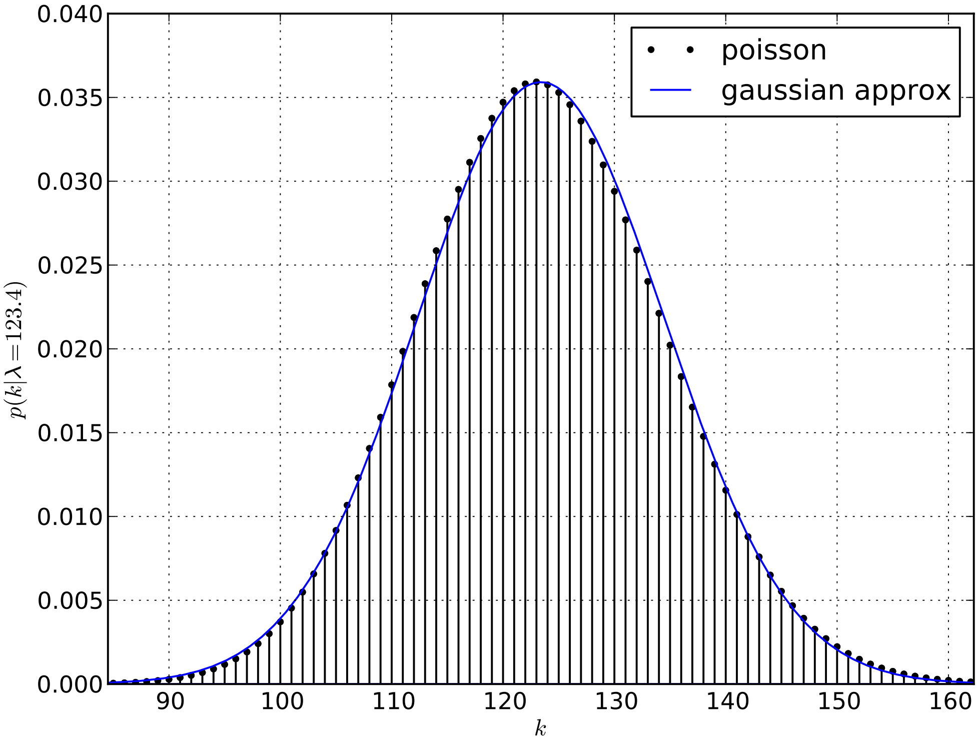
```
> ipython -pylab
from scipy.special import *
from scipy import comb # this gives the binomial coefficient
n = 200
alpha = 0.4
k = arange(n+1)
mu = n*alpha
sigma = sqrt(mu*(1-alpha))
kcont = linspace(mu-4*sigma,mu+4*sigma,1000)
pmf = comb(n,k) * alpha**k * (1-alpha)**(n-k)
pgauss = exp(-0.5*((kcont-mu)/sigma)**2)/(sigma*sqrt(2.0*pi))
stem(k,pmf,linefmt='k-',markerfmt='k.',basefmt='')
plot(k,pmf,'k.',label='binomial')
plot(kcont,pgauss,'b-',label='gaussian approx')
legend()
xlabel(r'$k$')
ylabel(r'$b(k|%d,%.1f)$' % (n,alpha))
xlim([mu-3.5*sigma,mu+3.5*sigma])
grid(True)
savefig('binomgauss.eps',bbox_inches='tight')
```



[4] actually, we need $n\alpha$ and $n(1-\alpha)$ both to be large

or the Poisson distribution where the expected number of events λ is large:[5]

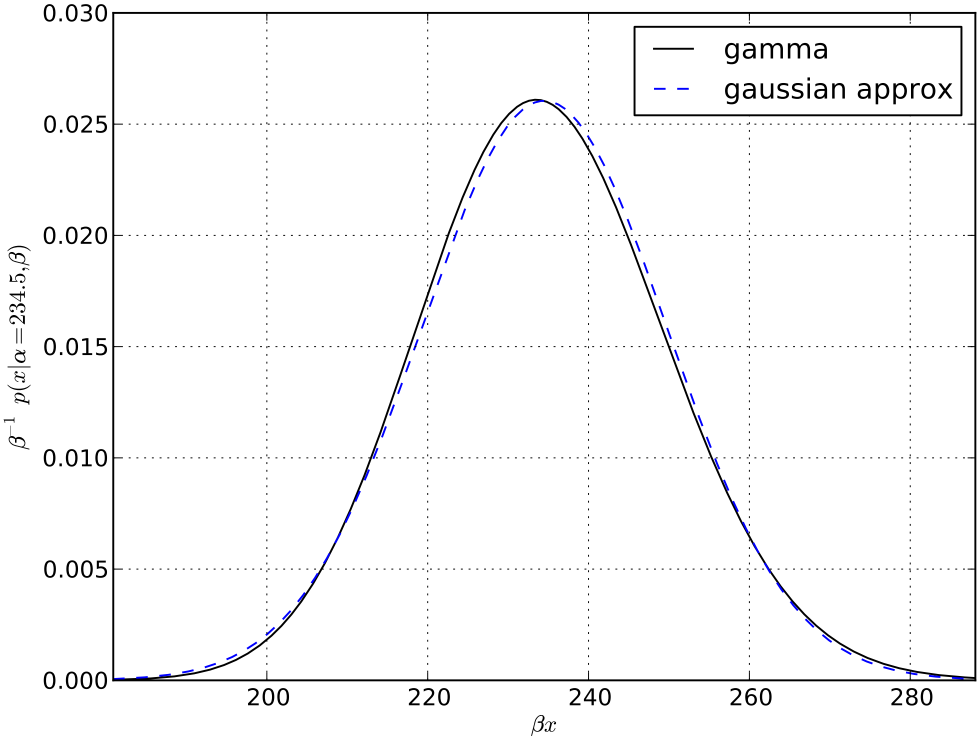
```
> ipython -pylab
from scipy.special import *
lam = 123.4
mu = lam
sigma = sqrt(lam)
k = arange(floor(mu-4*sigma),ceil(mu+4*sigma)+1)
kcont = linspace(mu-4*sigma,mu+4*sigma,1000)
pmfln = k * log(lam) - lam - gammaln(k+1)
pmf = exp(pmfln)
pgauss = exp(-0.5*((kcont-mu)/sigma)**2)/(sigma*sqrt(2.0*pi))
figure()
stem(k,pmf,linefmt='k-',markerfmt='k.',basefmt='')
plot(k,pmf,'k.',label='poisson')
plot(kcont,pgauss,'b-',label='gaussian approx')
legend()
xlabel(r'$k$')
ylabel(r'$p(k|\lambda=%.1f)$' % (lam))
xlim([mu-3.5*sigma,mu+3.5*sigma])
grid(True)
savefig('poissgauss.eps',bbox_inches='tight')
```



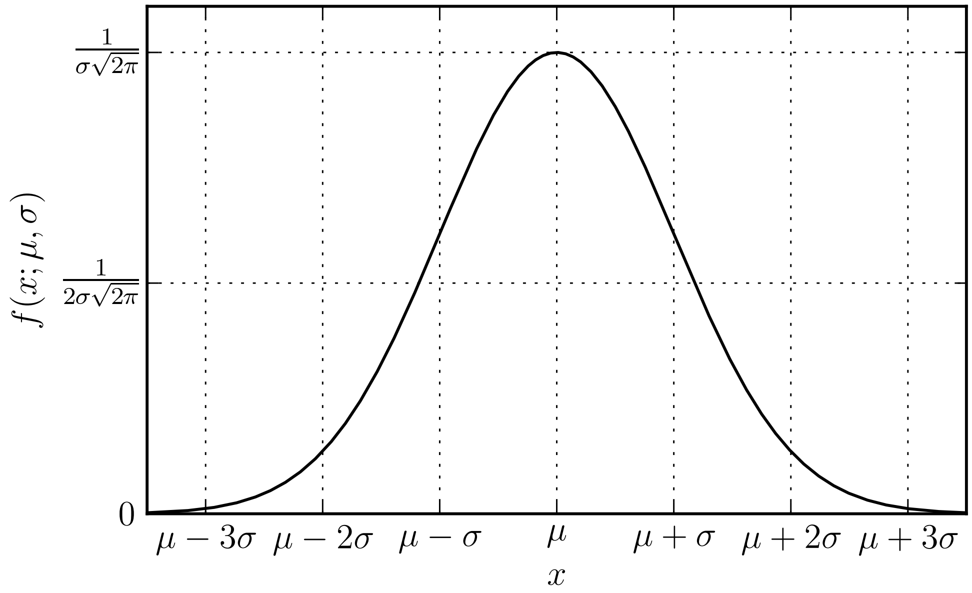
or the Gamma distribution where the parameter α is large

[5] We calculate the natural log of the pmf first because terms like λ^k and $e^{-\lambda}$ can separately overflow or underflow when k and λ are large.

```
> ipython -pylab
from scipy.special import *
alpha = 234.5
mu = alpha
sigma = sqrt(alpha)
betax = linspace(floor(mu-4*sigma),ceil(mu+4*sigma),1000)
pdfln = (alpha-1) * log(betax) - betax - gammaln(alpha)
pdf = exp(pdfln)
pgauss = exp(-0.5*((betax-mu)/sigma)**2)/(sigma*sqrt(2.0*pi))
figure()
plot(betax,pdf,'k-',label='gamma')
plot(betax,pgauss,'b--',label='gaussian approx')
legend()
xlabel(r'$\beta x$')
ylabel(r'$\beta^{-1}\,p(x|\alpha=%.1f,\beta)$' % (alpha))
xlim([mu-3.5*sigma,mu+3.5*sigma])
grid(True)
savefig('gammagauss.eps',bbox_inches='tight')
```



they all have a very similar shape, which approximates the shape of the Gaussian distribution:



The pdf for this distribution is

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \tag{1.78}$$

In the coming weeks, we will go into the reasons for this, which include

- Taylor expansion of the log of the pdf $\ln f(x)$ about its maximum; the quadratic expansion of $\ln f(x)$ leads to a Gaussian form for $f(x)$.

- The Central Limit Theorem, which states that the sum of a large number of independent and identically distributed random variables will always be approximated by a Gaussian.

The integrals we calculated last week can be used to show that (1.78) is normalized, and that the mean is μ and the variance is σ^2.

Given a Gaussian random variable X we can always define a related variable

$$Z = \frac{X-\mu}{\sigma} \tag{1.79}$$

which follows the standard normal distribution:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \tag{1.80}$$

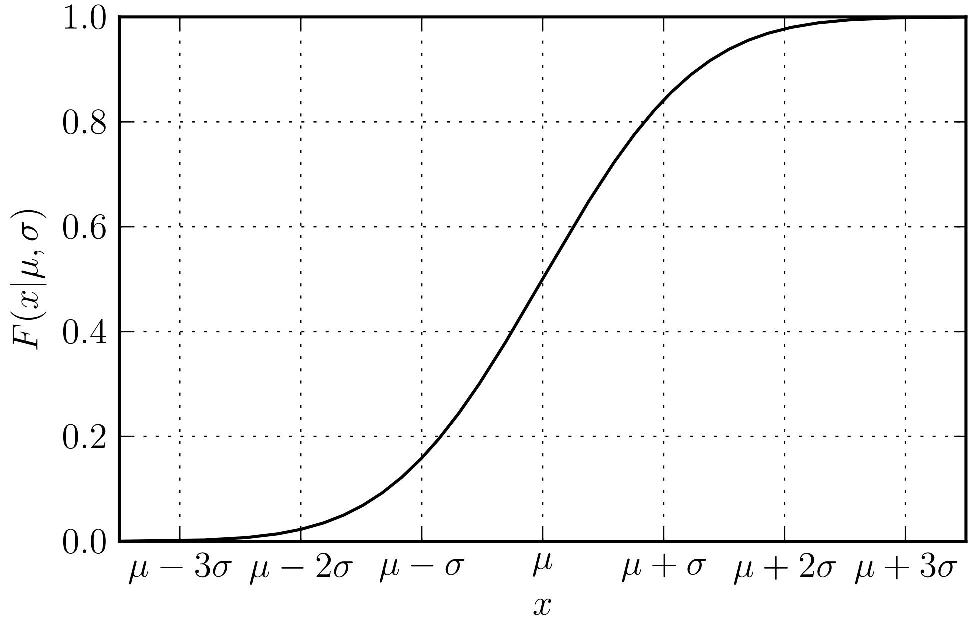
Note that the cdf is

$$\begin{aligned} F_Z(z) = P(Z < z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\zeta^2/2}\, d\zeta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{z/\sqrt{2}} e^{-\zeta^2/2}\, d\zeta \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_{0}^{z/\sqrt{2}} e^{-t^2}\, dt = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) \end{aligned} \tag{1.81}$$

and the cdf of the original Gaussian random variable is

$$F_X(x) = P(X < x) = P\left(Z < \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \tag{1.82}$$

which looks like this:



1.6.6 The Log-Normal Distribution (continuous)

A distribution closely related to the Gaussian distribution is the log-normal distribution. A log-normally distributed random variable is just one whose logarithm is normally distributed. I.e., if Y is log-normally distributed with parameters μ and σ, then $X = \ln Y$ is Gaussian with mean μ and variance σ^2. (Note that μ and σ^2 are *not* the mean and variance of Y.)

We can work out the pdf by a change of variables:

$$f_Y(y) = \frac{dP}{dy} = \frac{dx}{dy}\frac{dP}{dx} = \frac{1}{y}\,f_X(\ln y) = \frac{1}{\sigma\sqrt{2\pi}}\,\frac{e^{-(\ln y-\mu)^2/(2\sigma^2)}}{y} \tag{1.83}$$

1.6.7 The Chi-Square Distribution (continuous)

A chi-square (χ^2) random variable with ν degrees of freedom is just the sum of the squares of ν independent standard normal random variables:

$$X = \sum_{k=1}^{\nu} {Z_k}^2 \tag{1.84}$$

We can deduce the form of the pdf by starting with the joint pdf for the ν independent standard normal random variables:

$$f_{\mathbf{Z}}(\{z_k\}) = (2\pi)^{\nu/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{\nu} z_k^2\right) = \frac{d^\nu P}{d^\nu z} \tag{1.85}$$

Now, let's change variables to ν-dimensional spherical coördinates $\{r, \phi_1, \ldots, \phi_{\nu-1}\}$. We can get Jacobian of this transformation from

$$d^\nu z = r^{\nu-1}\, dr\, d^{\nu-1}\Omega = r^{\nu-1} \frac{d^{\nu-1}\Omega}{d^{\nu-1}\phi}\, dr\, d^{\nu-1}\phi \tag{1.86}$$

so that

$$f_{R,\mathbf{\Phi}}(r, \phi_1, \ldots, \phi_{\nu-1}) = r^{\nu-1} \frac{d^{\nu-1}\Omega}{d^{\nu-1}\phi} (2\pi)^{\nu/2} \exp\left(-r^2/2\right) \tag{1.87}$$

The geometrical factor[6]

$$\frac{d^{\nu-1}\Omega}{d^{\nu-1}\phi} \tag{1.88}$$

will depend on the angles $\{\phi_1, \ldots, \phi_{\nu-1}\}$, but we don't care about its exact form, just that it does not depend on r. This is because we are only interested in the PDF for $X = R^2$, and can marginalize over all of the angular coördinates to get

$$f_R(r) = r^{\nu-1} e^{-r^2/2} (2\pi)^{\nu/2} \int \cdots \int \frac{d^{\nu-1}\Omega}{d^{\nu-1}\phi}\, d^{\nu-1}\phi \propto r^{\nu-1} e^{-r^2/2} \tag{1.89}$$

We would need to evaluate the angular integral to get the proportionality constant, but we'll get that by requiring the pdf to be normalized, in a moment.

To get the PDF for the chi-square random variable $X = R^2$ we note

$$f_X(x) = \frac{dP}{dx} = \frac{dx}{dr}\frac{dP}{dr} = \frac{1}{2\sqrt{x}} f_R(\sqrt{x}) \propto x^{-1/2} x^{(\nu-1)/2} e^{-x/2} = x^{\frac{\nu}{2}-1}\, e^{-x/2} \tag{1.90}$$

If we compare this to (1.75) we see that X follows a Gamma distribution with $\alpha = \nu/2$ and $\beta = 1/2$, which allows us to write down the pdf including the normalization constant:

$$f(x|\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\frac{\nu}{2}-1}\, e^{-x/2} \qquad x \geq 0 \tag{1.91}$$

It also tells us that the mean is

$$\langle X \rangle = \nu \tag{1.92}$$

and the variance is

$$\left\langle X^2 \right\rangle - \langle X \rangle^2 = 2\nu \tag{1.93}$$

An exponential distribution is a special case of the Gamma distribution with $\alpha = 0$ and $\beta = r$.

Tuesday, October 19, 2010

2 Parameter Estimation

Now that we've laid some of the ground rules for calculating probabilities for the outcome of an observation given some parameter values, and for turning that likelihood function, if desired, into a posterior probability for one or more parameters, let's explore some methods for using the results of observations to make statements about parameters and models.

[6] For $\nu = 3$, it is $\sin\theta$ so that $d^2\Omega = \sin\theta\, d\theta\, d\phi$.

2.1 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 $f(y|x)$. 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

$$P(y_1 < {\color{blue}Y} < y_2|x) = \int_{y_1}^{y_2} f(y|x)\,dy\ . \tag{2.1}$$

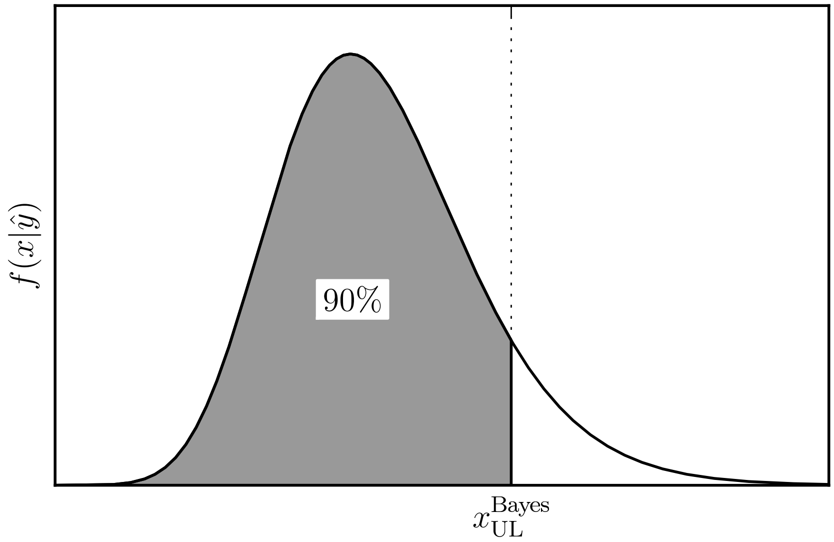
Now suppose we run the experiment and get the actual number $\hat{y}$. How would a Bayesian and a frequentist calculate a 90% upper limit on x?

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 $f(x|y)$

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

Plotting the posterior pdf, x_{UL} is defined such that 90% of the area under the posterior $f(x|\hat{y})$ 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 $f(x|y)$ rather than the likelihood $f(y|x)$, so we have to use Bayes's theorem to evalate it:

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

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

2.1.2 Frequentist Upper Limit

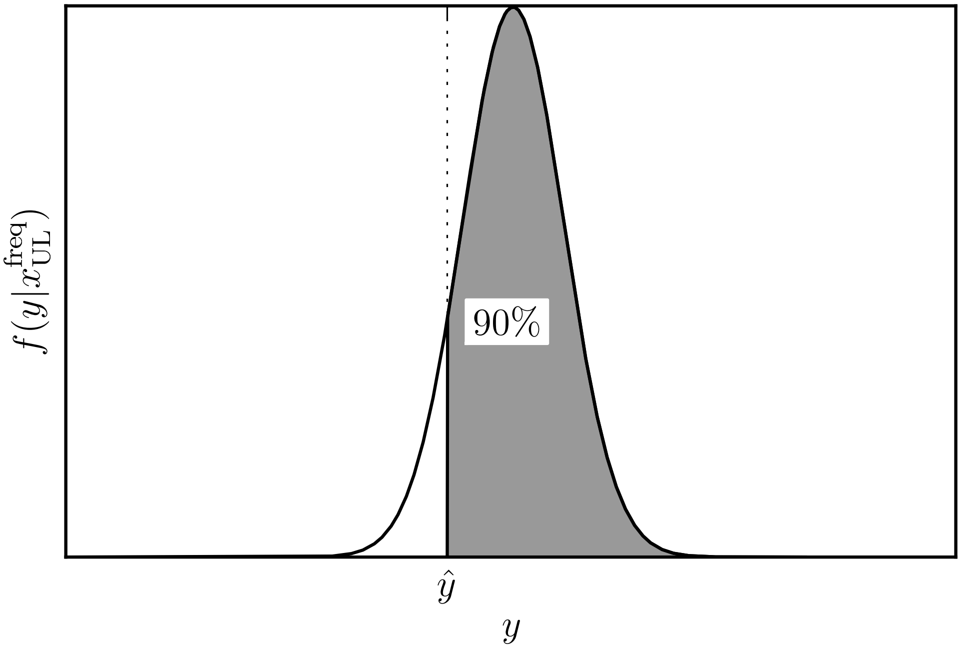
In the frequentist approach, we can't estimate the probability that $x < x_{\mathrm{UL}}$, because we don't talk about probabilities for physical quantities to have particular values. (After all x has some fixed value, even if we don't know it.) The probabilities we can discuss are those of the outcome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

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

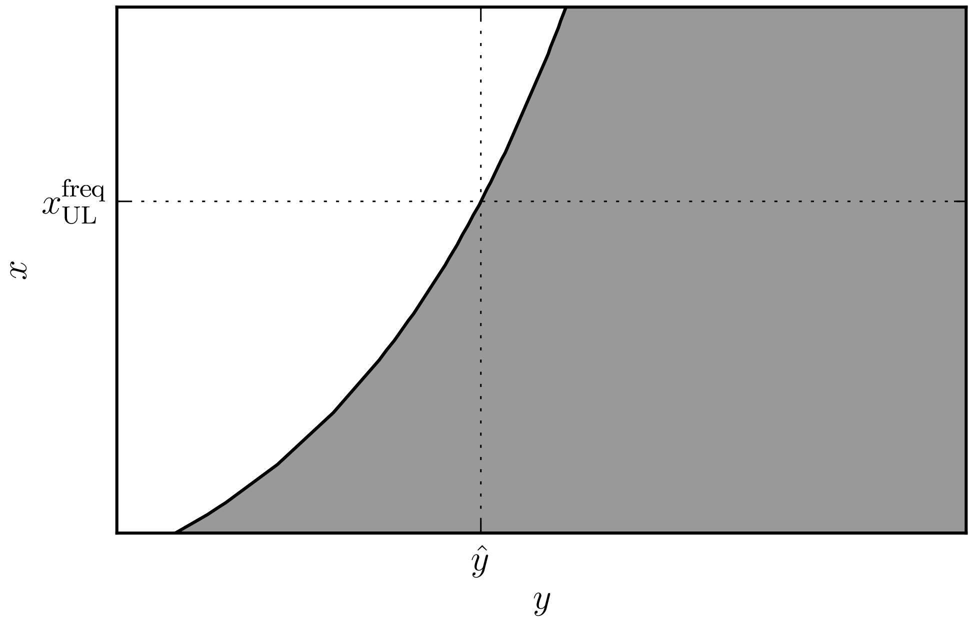
or more accurately

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

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 $f(y|x_{\mathrm{UL}}^{\mathrm{freq}})$ 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 ($x > x_{\mathrm{UL}}^{\mathrm{freq}}$) 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 $x_{\mathrm{UL}}^{\mathrm{freq}}$, 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.

2.1.3 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 $f(x)$ impacts the construction of the posterior

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

(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

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

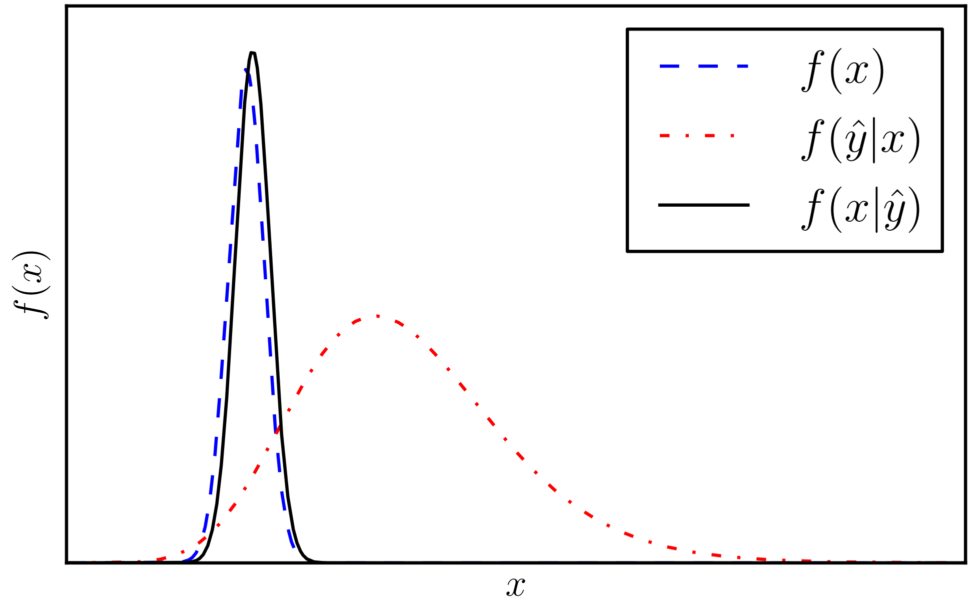
where the normalization $1/f(\hat{y})$ is independent of x and can be set by requiring

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

so schematically,

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

But the big question is what we use for $f(x)$. 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

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

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

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

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

$$f(x) = \begin{cases} \frac{1}{x_{\text{max}}} & |x| < \frac{x_{\text{max}}}{2} \\ 0 & |x| > \frac{x_{\text{max}}}{2} \end{cases} \ , \tag{2.12}$$

which is constant over some range of values of width $x_{\max}$ and then just choose $x_{\max}$ 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 $x_{\max} \to \infty$. In this case[7]

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

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 $f(x)$ and the posterior $f(x|y)$ 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

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

which would lead to a posterior

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

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

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

So

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

and conversely

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

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

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

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

2.2 "Maximum Posterior" Parameter Estimation

2.2.1 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

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

for some parameter x is $f(x|D)$. 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

$$\langle X \rangle_D = \int_{-\infty}^{\infty} x\, f(x|D)\, dx \tag{2.20}$$

and a variance

$$\left\langle (X - \langle X \rangle_D)^2 \right\rangle_D \tag{2.21}$$

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

$$\forall x : f(\hat{x}|D) \geq f(x|D) \ . \tag{2.22}$$

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 $f(x|D)$ itself about $x = \hat{x}$ because we know $f(x|D)$ 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,

$$L(x) = \ln f(x|D) \ . \tag{2.23}$$

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

$$L(x) = L(\hat{x}) + (x - \hat{x})\, L'(\hat{x}) + \frac{(x-\hat{x})^2}{2}\, L''(\hat{x}) + \ldots \tag{2.24}$$

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

$$L(x) \approx L(\hat{x}) - \frac{(x-\hat{x})^2}{2}\, [-L''(\hat{x})] \tag{2.25}$$

and

$$f(x|D) \approx f(\hat{x}|D) \exp\left(\frac{-(x-\hat{x})^2}{2[\sqrt{-1/L''(\hat{x})}]^2} \right) \ , \tag{2.26}$$

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 $\hat{x}$.

2.2.2 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

$$\ln f(\mathbf{x}|D) = L(\mathbf{x}) \approx L(\hat{\mathbf{x}}) - \frac{1}{2}\sum_\alpha \sum_\beta \left(-\frac{\partial^2 L}{\partial x_\alpha \partial x_\beta}\right)_{\mathbf{x}=\hat{\mathbf{x}}} (x_\alpha - \hat{x}_\alpha)(x_\beta - \hat{x}_\beta) \ . \tag{2.27}$$

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

$$F_{\alpha\beta} = -\frac{\partial^2 L}{\partial x_\alpha \partial x_\beta} \tag{2.28}$$

we can write this approximation to the posterior as

$$f(\mathbf{x}|D) \approx f(\hat{\mathbf{x}}|D) \exp\left(-\frac{1}{2}\left(\mathbf{x} - \hat{\mathbf{x}}\right)^{\mathrm{T}} \mathbf{F} \left(\mathbf{x} - \hat{\mathbf{x}}\right)\right) \tag{2.29}$$

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

$$\left\langle (\mathbf{X} - \hat{\mathbf{x}})\, (\mathbf{X} - \hat{\mathbf{x}})^{\mathrm{T}} \right\rangle \approx \mathbf{F}^{-1} \tag{2.30}$$

(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

$$\mathbf{F} = \sum_\alpha \mathbf{u}_\alpha\, f_\alpha \mathbf{u}_\alpha^{\mathrm{T}} \tag{2.31}$$

and so

$$f(\mathbf{x}|D) \approx f(\hat{\mathbf{x}}|D) \exp\left(-\sum_\alpha \frac{f_\alpha}{2} (\xi_\alpha)^2\right) \tag{2.32}$$

where

$$\xi_\alpha = \mathbf{u}_\alpha^{\mathrm{T}} (\mathbf{x} - \hat{\mathbf{x}}) \tag{2.33}$$

the equation

$$\sum_\alpha \frac{f_\alpha}{2} (\xi_\alpha)^2 = \text{constant} \tag{2.34}$$

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

$$\left\langle (x_\alpha - \hat{x}_\alpha)^2 \right\rangle \approx (F^{-1})_{\alpha\alpha} \neq \frac{1}{F_{\alpha\alpha}} \tag{2.35}$$

Thursday, October 21, 2010

3 Hypothesis Testing

3.1 Chi-Squared Testing

Let's turn now to another frequentist method, chi-squared testing. Chi-squared tests take many forms, but they're all in some way associated with goodness-of-fit, i.e., are the data consistent with the model. Actually, what they have in common is that they all involve constructing a statistic which, if the model is correct, obeys–exactly or approximately–a chi-squared distribution. (See Section 1.6.7.) The further the data are from the model prescription, the higher this chi-squared statistic will be, and we can use the cumulative distribution function to say how unlikely it is that the model would lead to a χ^2 that high.

3.1.1 Models Without Free Parameters

The simplest application is to test the validity of a single model. Suppose that we make n measurements, and the model $\mathcal{M}$ predicts that those n random variables $\{Y_i | i = 1, \ldots, n\}$, will each be Gaussian distributed, with means $\{\mu_i\}$ and standard deviations $\{\sigma_i\}$. Then we can construct random variables $Z_i = \frac{Y_i - \mu_i}{\sigma_i}$ which should be normally distributed. The sum of their squares is

$$X = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{\sigma_i^2} \tag{3.1}$$

and its pdf is the pdf for a Gamma distribution with $\alpha = n/2$ and $\beta = 1/2$, i.e.,

$$f_X(x|\mathcal{M}) = \frac{1}{2^{n/2}\Gamma(n/2)}\, x^{\frac{n}{2}-1}\, e^{-x/2} \tag{3.2}$$

If we perform the measurement and obtain a χ^2 value of $\hat{x}$, we can ask about the probability that X is $\hat{x}$ or higher, i.e., that we would get a χ^2 at least that high from a random data set generated by the model. This is

$$P(X \geq \hat{x}) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_{\hat{x}}^{\infty} x^{\frac{n}{2}-1}\, e^{-x/2}\, dx = \frac{1}{\Gamma(n/2)} \int_{\hat{x}/2}^{\infty} u^{\frac{n}{2}-1}\, e^{-u/2}\, du = \frac{\Gamma(n/2, \hat{x}/2)}{\Gamma(n/2)} \tag{3.3}$$

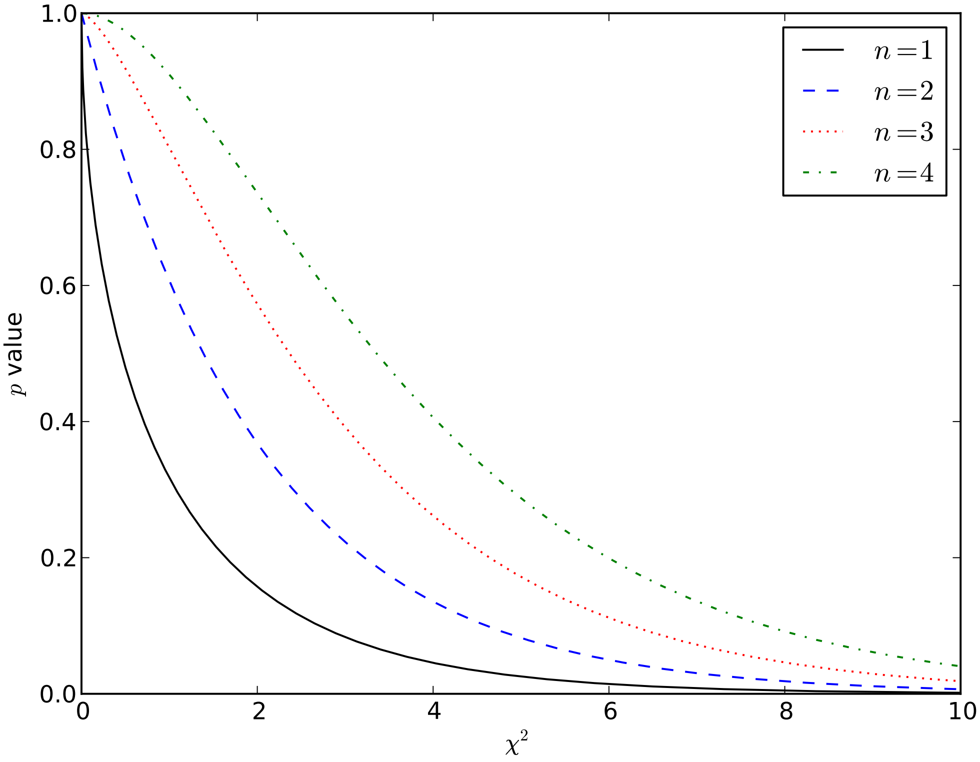
where we have performed a change of variables from x to $u = x/2$ to convert the integral into an upper incomplete Gamma function. $P(X \geq \hat{x})$ is called the "p-value" corresponding to the chi-squared value $\hat{x}$. We can obtain this in SciPy with the help of `scipy.special.gammaincc`:[8]

```
> ipython -pylab
from scipy.special import gammaincc
chisq = linspace(0,10,1000)
linestyles = ['k-','b--','r:','g-.']
figure()
for i in xrange(len(linestyles)):
```

[8] Note that the incomplete Gamma functions in scipy already have the factor of $1/\Gamma(\alpha)$ included in them.

```
  ndof = i + 1
  pval = gammaincc(0.5*ndof,0.5*chisq)
  plot(chisq,pval,linestyles[i],label=(r'$n=%d$'%ndof))

legend()
xlabel(r'$\chi^2$')
ylabel(r'$p$ value')
savefig('chisqpval.eps',bbox_inches='tight')
```



So for example if we fit 20 data points and obtain a χ^2 of 31.2, the p-value is 5.3% which I calculated from

```
> ipython
from scipy.special import gammaincc
ndof = 20
chisq = 31.2
pval = gammaincc(0.5*ndof,0.5*chisq)
resfile = open('notes_stats_pval.tex','w')
resfile.write('\\newcommand{\\chisq}{%.1f}\n' % chisq)
resfile.write('\\newcommand{\\pval}{%.1f\\%%}\n' % (100*pval))
resfile.close()
```

That means that there is only a 5.3% chance that such a high χ^2 would be found from data consistent with the model.

Since the mean of a chi-squared distribution with n degrees of freedom is n and its variance is $2n$, a useful quantity to look at is the χ^2 per degree of freedom, which for moderate-sized n should be $1 \pm \sqrt{2/n}$.

Note that we can also represent this problem in matrix notation, defining column vectors

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} ; \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ; \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \tag{3.4}$$

and a matrix

$$\boldsymbol{\sigma}^2 = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_1^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} = \left\langle (\mathbf{Y} - \boldsymbol{\mu}) \, (\mathbf{Y} - \boldsymbol{\mu}^{\mathrm{T}} \right\rangle \tag{3.5}$$

Then

$$f_{\mathbf{Y}}(\mathbf{y}|\mathcal{M}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\sigma}^2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{y} - \boldsymbol{\mu})\right) \tag{3.6}$$

and we can define

$$\mathbf{Z} = \boldsymbol{\sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \tag{3.7}$$

which is a column vector of n independent random variables with joint pdf

$$f_{\mathbf{Z}}(\mathbf{z}|\mathcal{M}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}\right) \tag{3.8}$$

and then the chi-squared statistic constructed from $\mathbf{Y}$ is

$$X = \mathbf{Z}^{\mathrm{T}} \, \mathbf{Z} = (\mathbf{Y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{Y} - \boldsymbol{\mu}) \tag{3.9}$$

Note that nothing about this matrix construction actually required $\boldsymbol{\sigma}^2$ to be diagonal. (Off-diagonal elements σ_{ij} correspond to correlations between the Y_ks.) As long as

$$\left\langle (\mathbf{Y} - \boldsymbol{\mu}) \, (\mathbf{Y} - \boldsymbol{\mu})^{\mathrm{T}} \right\rangle = \boldsymbol{\sigma}^2 \ , \tag{3.10}$$

which is a real symmetric matrix, is positive definite, we can take both its inverse and its square root.

3.1.2 Models With Free Parameters

Now consider the case where the model $\mathcal{M}$ has m parameters $\{\lambda_\alpha|\alpha = 1, \dots, m\}$, and we want to find the best possible "fit" for these parameters, and use any redundant information to test the validity of the model. Given a set of parameters, represented as a column vector $\boldsymbol{\lambda}$, the likelihood function

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\lambda}, \mathcal{M}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\sigma}^2}} \exp\left(-\frac{1}{2}[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\lambda})]^{\mathrm{T}}\boldsymbol{\sigma}^{-2}[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\lambda})]\right) \tag{3.11}$$

Now, if $m = n$ (i.e., there are as many parameters as data points), we can find a unique "best fit" to the data by solving the n equations

$$y_i = \mu_i(\boldsymbol{\lambda}) \qquad (\text{MLE when } m = n) \tag{3.12}$$

for the m unknowns $\{\lambda_\alpha\}$. This will clearly maximize the likelihood function (3.11).

More generally, with the situation described by (3.11), in which the model predicts $\mathbf{Y}$ to be $\boldsymbol{\mu}(\boldsymbol{\lambda})$ plus a Gaussian error, the likelihood is maximized by minimizing the χ^2

$$[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda})]^{\mathrm{T}} \boldsymbol{\sigma}^{-2} [\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda})] \; . \tag{3.13}$$

This is also known as a least-squares fit, since in the case where the errors are uncorrelated, it means finding the parameters which minimize the normalized squared deviation from the model

$$\sum_{i=1}^{n} \frac{[y_i - \mu_i(\boldsymbol{\lambda})]^2}{\sigma_i^2} \tag{3.14}$$

If $\hat{\boldsymbol{\lambda}}(\mathbf{y})$ are the parameter values which accomplish this for a particular set of observations $\mathbf{y}$, the residual χ^2 is

$$\left[\mathbf{y} - \boldsymbol{\mu}\left(\hat{\boldsymbol{\lambda}}(\mathbf{y})\right)\right]^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \left[\mathbf{y} - \boldsymbol{\mu}\left(\hat{\boldsymbol{\lambda}}(\mathbf{y})\right)\right] \tag{3.15}$$

If we model this whole procedure, the residual chi-squared for an experiment is a random variable

$$X_{\text{red}} = \left[\mathbf{Y} - \boldsymbol{\mu}\left(\hat{\boldsymbol{\lambda}}(\mathbf{Y})\right)\right]^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \left[\mathbf{Y} - \boldsymbol{\mu}\left(\hat{\boldsymbol{\lambda}}(\mathbf{Y})\right)\right] \tag{3.16}$$

It is standard practice to assume this obeys a chi-squared distribution with $n-m$ degrees of freedom. Depending on the model, this may just be an approximation (just as the assumption of Gaussian errors may be an approximation in a more general situation). We can show, however, that it is exactly true when the modelled expectation value $\boldsymbol{\mu}(\boldsymbol{\lambda})$ is a linear function of the parameters. This is the case, for example, when the different observations occur at values t_i of some variable, and the model is a superposition of known functions of the t_i with unknown coëfficients:

$$\mu_i = \sum_{\alpha} \lambda_\alpha g_\alpha(t_i) \qquad \text{(example of superposition model)} \tag{3.17}$$

the functions $\{g_\alpha\}(t)$ might be polynomials, or Fourier modes, or spherical harmonics, or whatever.

So, let's assume that the functional relationship $\boldsymbol{\mu}(\boldsymbol{\lambda})$ is something linear

$$\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbf{A}\boldsymbol{\lambda} \tag{3.18}$$

then the χ^2 is

$$(\mathbf{y} - \mathbf{A}\boldsymbol{\lambda})^{\mathrm{T}} \boldsymbol{\sigma}^{-2} (\mathbf{y} - \mathbf{A}\boldsymbol{\lambda}) \tag{3.19}$$

If we differentiate this with respect to each of the $\{\lambda_\alpha\}$ and set the result to zero, we get m maximum likelihood equations which can be combined into a column vector

$$\mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \left[\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\lambda}}(\mathbf{y})\right] = \mathbf{0}_{m\times 1} \tag{3.20}$$

which means

$$\left[\mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \mathbf{A}\right] \hat{\boldsymbol{\lambda}}(\mathbf{y}) = \mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \mathbf{y} \tag{3.21}$$

Now, the matrix

$$\left[\mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} \mathbf{A}\right] \tag{3.22}$$

is an $m \times m$ symmetric matrix. For the maximum likelihood estimate to exist, it has to be invertible. If it is, we have

$$\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right] = \mathbf{1}_{m\times m} \tag{3.23}$$

and the maximum likelihood estimate for the parameters is

$$\hat{\boldsymbol{\lambda}}(\mathbf{y}) = \left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{y} \tag{3.24}$$

and the corresponding expectation values for the data are

$$\mu\left(\hat{\boldsymbol{\lambda}}(\mathbf{y})\right) = \mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{y} \tag{3.25}$$

Note that we may be tempted to simplify this with something involving the matrix inverse of $\mathbf{A}$, but that doesn't exist because $\mathbf{A}$ is *not* a square matrix. What we can do, though, is define the matrix

$$\mathbf{P} = \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} \tag{3.26}$$

so that

$$\mu\left(\hat{\boldsymbol{\lambda}}(\mathbf{y})\right) = \boldsymbol{\sigma}\mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{y} \tag{3.27}$$

In fact, since

$$\mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{A} = \boldsymbol{\sigma}^{-1}\mathbf{A} \tag{3.28}$$

we see that for *any* $\boldsymbol{\lambda}$,

$$\mathbf{P}\boldsymbol{\sigma}^{-1}\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{A}\boldsymbol{\lambda} = \boldsymbol{\sigma}^{-1}\mathbf{A}\boldsymbol{\lambda} = \boldsymbol{\sigma}^{-1}\boldsymbol{\mu}(\boldsymbol{\lambda}) \tag{3.29}$$

The matrix $\mathbf{P}$ is not only symmetric, but it's a projection matrix, since

$$\begin{aligned}\mathbf{P}\,\mathbf{P} &= \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} \\ &= \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} = \mathbf{P}\end{aligned} \tag{3.30}$$

This $n \times n$ matrix is a projector onto an m-dimensional subspace, since

$$\operatorname{tr}\mathbf{P} = \operatorname{tr}\left(\boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1}\right) = \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\right) = \operatorname{tr}\mathbf{1}_{m\times m} = m \tag{3.31}$$

That means that it has m eigenvectors with unit eigenvalue and $n - m$ eigenvectors with zero eigenvalue. Let $\{\mathbf{u}_i | i = 1, \dots, m\}$ be a set of m orthonormal eigenvectors with unit eigenvalue and $\{\mathbf{u}_i | i = m+1, \dots, n\}$ be a set of $n - m$ orthonormal eigenvectors with unit eigenvalue, so that

$$\mathbf{P} = \sum_{i=1}^{m}\mathbf{u}_i\,\mathbf{u}_i^{\mathrm{T}} \tag{3.32}$$

and

$$\mathbf{1}_{n\times n} - \mathbf{P} = \sum_{i=m+1}^{n}\mathbf{u}_i\,\mathbf{u}_i^{\mathrm{T}} \tag{3.33}$$

are projection operators onto the two orthogonal subspaces.

Now, let $\boldsymbol{\lambda}_{\text{true}}$ be the column vector of true, unknown parameters. Then

$$\mathbf{Z} = \boldsymbol{\sigma}^{-1}\left[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})\right] \tag{3.34}$$

is a vector of n independent standard normal random variables. Since $\{\mathbf{u}_i | i = 1, \ldots, n\}$ is an orthonormal basis, we can also construct independent standard normal random variables $\{\mathcal{Z}_i | i = 1, \ldots, n\}$ where

$$\mathcal{Z}_i = \mathbf{u}_i^{\text{T}}\mathbf{Z} = \mathbf{u}_i^{\text{T}}\boldsymbol{\sigma}^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})\right) \tag{3.35}$$

If we take the sum of the squares of the last $n - m$ of these random variables, that will obey a chi-squared distribution with $n - m$ degrees of freedom:

$$\begin{aligned} X_{n-m} &= \sum_{i=m+1}^{n} [\mathcal{Z}_i]^2 = [\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})]^{\text{T}}\boldsymbol{\sigma}^{-1}\left(\sum_{i=m+1}^{n} \mathbf{u}_i\,\mathbf{u}_i^{\text{T}}\right)\boldsymbol{\sigma}^{-1}[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})] \\ &= [\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})]^{\text{T}}\boldsymbol{\sigma}^{-1}\left(\mathbf{1}_{n\times n} - \mathbf{P}\right)\boldsymbol{\sigma}^{-1}[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})] \end{aligned} \tag{3.36}$$

But by (3.29) and (3.27),

$$\begin{aligned} \left(\mathbf{1}_{n\times n} - \mathbf{P}\right)\boldsymbol{\sigma}^{-1}[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})] &= \left(\mathbf{1}_{n\times n} - \mathbf{P}\right)\boldsymbol{\sigma}^{-1}\mathbf{Y} = \boldsymbol{\sigma}^{-1}\mathbf{Y} - \mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{Y} \\ &= \boldsymbol{\sigma}^{-1}\left[\mathbf{Y} - \mu\left(\hat{\boldsymbol{\lambda}}(\mathbf{Y})\right)\right] \end{aligned} \tag{3.37}$$

so

$$X_{n-m} = \left[\mathbf{Y} - \mu\left(\hat{\boldsymbol{\lambda}}(\mathbf{Y})\right)\right]^{\text{T}}\sigma^{-2}\left[\mathbf{Y} - \mu\left(\hat{\boldsymbol{\lambda}}(\mathbf{Y})\right)\right] = X_{\text{red}} \tag{3.38}$$

which means this quantity we've constructed, which is χ^2-distributed with $n - m$ degrees of freedom (in the case where the modelled expectation values are linear in the parameters, as described in (3.18)), is indeed the residual chi-squared defined in (3.16).

Tuesday, October 26, 2010

3.2 Odds Ratio and Bayes Factor

One of the problems about using a frequentist test like a chi-squared test to assess the validity of a model is that you can always make the fit better by adding more parameters to the model. In the extreme case, if you have as many model parameters as data points, you can make the fit perfect. But clearly a model which is "overtuned" in this way is scientifically unsatisfying.

Bayesian statistics offers a natural way to compare models, which automatically penalizes models that use too many parameters to fine-tune themselves to match a data set. This is known as the odds ratio.

Consider Bayes's theorem in the context of a model $\mathcal{M}$ with parameters $\boldsymbol{\lambda}$. Given an observation $\mathbf{y}$, we can construct the posterior pdf for the parameters $\boldsymbol{\lambda}$ as follows

$$f(\boldsymbol{\lambda}|\mathbf{y}, \mathcal{M}) = \frac{f(\mathbf{y}|\boldsymbol{\lambda}, \mathcal{M})f(\boldsymbol{\lambda}|\mathcal{M})}{f(\mathbf{y}|\mathcal{M})} \tag{3.39}$$

which is sometimes abbreviated as

$$(\text{posterior}) = \frac{(\text{likelihood})(\text{prior})}{(\text{evidence})} \tag{3.40}$$

So far we've just treated the denominator as a normalization factor

$$f(\mathbf{y}|\mathcal{M}) = \int d\lambda\, f(\mathbf{y}|\boldsymbol{\lambda},\mathcal{M}) f(\boldsymbol{\lambda}|\mathcal{M}) \tag{3.41}$$

but we will now see how it gets the name "evidence". Note that it is the overall probability to get the observed result $\mathbf{y}$ given the model $\mathcal{M}$, marginalizing over the parameters $\boldsymbol{\lambda}$.

Now, consider the case where $\mathcal{M}$ is one of a number of possible models, and we'd like to construct a posterior probability $P(\mathcal{M}|\mathbf{y})$ that $\mathcal{M}$ is the correct model. Well, since we have a way to calculate $f(\mathbf{y}|\mathcal{M})$, we can try using Bayes's theorem:

$$P(\mathcal{M}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathcal{M})P(\mathcal{M})}{f(\mathbf{y})} \tag{3.42}$$

The right-hand side has a couple of things that are harder to get a handle on: the prior probability $P(\mathcal{M})$ of $\mathcal{M}$ being the correct model, and the overall pdf $f(\mathbf{y})$ which requires somehow marginalizing over all possible models. The usual way around this is to consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, and to calculate the ratio of their posteriors, known as the odds ratio

$$\mathcal{O}_{12} = \frac{P(\mathcal{M}_1|\mathbf{y})}{P(\mathcal{M}_2|\mathbf{y})} = \frac{f(\mathbf{y}|\mathcal{M}_1)P(\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_2)P(\mathcal{M}_2)} = \left(\frac{f(\mathbf{y}|\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_2)}\right)\left(\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}\right) = \left(\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}\right)\mathcal{B}_{12} \tag{3.43}$$

So the factor of $f(\mathbf{y})$ has cancelled out, and the odds ratio $\mathcal{O}_{12}$ is the ratio of prior probabilities for each model times something known as the *Bayes factor*

$$\mathcal{B}_{12} = \frac{f(\mathbf{y}|\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_2)} \tag{3.44}$$

which is the ratio of the "evidence" in each of the models. It represents how our relative confidence in the two probabilities has changed with the measurement $\mathbf{y}$. If each model has some parameters, the Bayes factor can be written as

$$\mathcal{B}_{12} = \frac{\int d\lambda_1\, f(\mathbf{y}|\boldsymbol{\lambda}_1,\mathcal{M}_1)\, f(\boldsymbol{\lambda}_1|\mathcal{M}_1)}{\int d\lambda_2\, f(\mathbf{y}|\boldsymbol{\lambda}_2,\mathcal{M}_2)\, f(\boldsymbol{\lambda}_2|\mathcal{M}_2)} \tag{3.45}$$

To see how the Bayes factor penalizes modes for over-tuning, consider a simple case where there are two models: $\mathcal{M}_0$, which has no parameters and $\mathcal{M}_1$, which has a parameter λ. If we measure data $\mathbf{y}$, the Bayes factor comparing the two models is

$$\mathcal{B}_{10} = \frac{\int_{-\infty}^{\infty} d\lambda\, f(\mathbf{y}|\lambda,\mathcal{M}_1)\, f(\lambda|\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_0)} \tag{3.46}$$

To get a handle on what the marginalization of the parameter λ does, as compared with the maximization done by the frequentist method, let's make some simplifying assumptions. First let's assume the likelihood $f(\mathbf{y}|\lambda,\mathcal{M}_1)$, seen as a function of λ, can be approximated as a Gaussian about the maximum likelihood value $\hat{\lambda}$:

$$f(\mathbf{y}|\lambda,\mathcal{M}_1) \approx f(\mathbf{y}|\hat{\lambda},\mathcal{M}_1)\, e^{-(\lambda-\hat{\lambda})/2\sigma_\lambda^2} \tag{3.47}$$

We'll also assume that this is sharply peaked compared to the prior $f(\lambda|\mathcal{M}_1)$ and therefore we can replace λ in the argument of the prior with $\hat{\lambda}$, and

$$\begin{aligned} \int_{-\infty}^{\infty} d\lambda\, f(\mathbf{y}|\lambda,\mathcal{M}_1)\, f(\lambda|\mathcal{M}_1) &\approx f(\mathbf{y}|\hat{\lambda},\mathcal{M}_1)\, f(\hat{\lambda}|\mathcal{M}_1) \int_{-\infty}^{\infty} d\lambda\, e^{-(\lambda-\hat{\lambda})/2\sigma_\lambda^2} \\ &= f(\mathbf{y}|\hat{\lambda},\mathcal{M}_1)\, f(\hat{\lambda}|\mathcal{M}_1)\, \sigma_\lambda\sqrt{2\pi} \end{aligned} \tag{3.48}$$

We can then approximate the Bayes factor as

$$\mathcal{B}_{10} = \frac{f(\mathbf{y}|\hat{\lambda},\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_0)} \frac{\sigma_\lambda\sqrt{2\pi}}{[f(\hat{\lambda}|\mathcal{M}_1)]^{-1}} \tag{3.49}$$

The first factor is the ratio of the likelihoods between the best-fit version of model $\mathcal{M}_1$ and the parameter-free model $\mathcal{M}_0$. That's basically the end of the story in frequentist model comparison, and we can see that if $\mathcal{M}_0$ is included as a special case of $\mathcal{M}_1$, this ratio will always be greater or equal to one, i.e., the tunable model will always be able to find a higher likelihood than the model without that tunable parameter. But in Bayesian model comparison, there is also the second factor:

$$\frac{\sigma_\lambda\sqrt{2\pi}}{[f(\hat{\lambda}|\mathcal{M}_1)]^{-1}} \qquad \text{“Occam factor”} \tag{3.50}$$

This is called the *Occam factor* because it implements Occam's razor, the principle that, all else being equal, simpler explanations will be favored over more complicated ones. Because the prior $f(\lambda|\mathcal{M}_1)$ is normalized, $[f(\hat{\lambda}|\mathcal{M}_1)]^{-1}$ is a measure of the width of the prior, i.e., how much parameter space the tunable model has available to it. In particular, if the prior is uniform over some range:

$$f(\lambda|\mathcal{M}_1) = \begin{cases} \frac{1}{\lambda_{\text{max}}-\lambda_{\text{min}}} & \lambda_{\text{min}} < \lambda < \lambda_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{3.51}$$

then the Occam factor becomes

$$\frac{\sigma_\lambda\sqrt{2\pi}}{\lambda_{\text{max}} - \lambda_{\text{min}}} \tag{3.52}$$

because we assumed the likelihood function was narrowly peaked compared to the prior, the Occam factor is always less than one, and the tunable model must have a large enough increase in likelihood over the simpler model in order to overcome this.

Thursday, October 28, 2010

4 Sums of Random Variables

4.1 Mean and Variance

Consider a situation where there are two random variables X_1 and X_2, and we construct a new random variable which is their sum,

$$T = X_1 + X_2 \ . \tag{4.1}$$

If the expectation values of the random variables are

$$\mu_1 = \langle X_1 \rangle \tag{4.2a}$$
$$\mu_2 = \langle X_2 \rangle \tag{4.2b}$$

then the linearity of the expectation value operation means that the expectation value of their sum is

$$\mu_T = \langle T \rangle = \langle X_1 \rangle + \langle X_2 \rangle = \mu_1 + \mu_2 \tag{4.3}$$

If the random variables ave standard deviations σ_1 and σ_2 and covariance $\operatorname{Cov}(X_1, X_2)$, so that

$$\left\langle (X_1 - \mu_1)^2 \right\rangle = \sigma_1^2 \tag{4.4a}$$
$$\left\langle (X_2 - \mu_2)^2 \right\rangle = \sigma_2^2 \tag{4.4b}$$
$$\langle (X_1 - \mu_1)(X_2 - \mu_2) \rangle = \operatorname{Cov}(X_1, X_2) \tag{4.4c}$$

then the variance of their sum is

$$\begin{aligned} \sigma_T^2 &= \left\langle (T - \mu_T)^2 \right\rangle = \left\langle (X_1 + X_2 - \mu_1 - \mu_2)^2 \right\rangle = \left\langle ([X_1 - \mu_1] + [X_2 - \mu_2])^2 \right\rangle \\ &= \left\langle (X_1 - \mu_1)^2 \right\rangle + 2 \left\langle (X_1 - \mu_1)(X_2 - \mu_2) \right\rangle + \left\langle (X_2 - \mu_2)^2 \right\rangle = \sigma_1^2 + \sigma_2^2 + 2 \operatorname{Cov}(X_1, X_2) \end{aligned} \tag{4.5}$$

In particular, if X_1 and X_2 are independent or otherwise uncorrelated, then the variance of their sum is equal to the sum of their variances:

$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 \qquad \text{if } X_1 \text{ and } X_2 \text{ uncorrelated} \tag{4.6}$$

Note that this means the standard deviations are "added in quadrature":

$$\sigma_T = \sqrt{\sigma_1^2 + \sigma_2^2} \qquad \text{if } X_1 \text{ and } X_2 \text{ uncorrelated} \tag{4.7}$$

This is the standard way in which uncorrelated random errors are combined.

One more property, which we will state now without proving, is that if X_1 and X_2 are both Gaussian random variables, their sum is also a Gaussian random variable.

4.2 Identical Random Variables (Random Samples)

Considered now the example of N independent, identically distributed (iid) random variables $\{X_i\}$ with expectation values

$$\langle X_i \rangle = \mu \qquad \text{and} \qquad \langle (X_i - \mu)(X_j - \mu) \rangle = \delta_{ij}\, \sigma^2 \tag{4.8}$$

This is known as a *random sample*, and it can be used to estimate the properties of the underlying distribution. If we construct the sum

$$T = \sum_{i=1}^{N} X_i \tag{4.9}$$

then an extension of the results for a pair of random variables shows that its mean is

$$\mu_T = \langle T \rangle = \sum_{i=1}^{N} \mu = N\mu \tag{4.10}$$

and its variance is

$$\sigma_T^2 = \left\langle (T - \mu_T)^2 \right\rangle = \sum_{i=1}^{N} \sigma^2 = N\sigma^2 \tag{4.11}$$

so its standard deviation is

$$\sigma_T = \sqrt{N}\, \sigma\ . \tag{4.12}$$

If we take the average of the N random variables

$$\overline{X} = \frac{1}{N} \sum_{k=0}^{N-1} X_i = \frac{T}{N}\ , \tag{4.13}$$

which is itself a random variable, we can see

$$\mu_{\overline{X}} = \langle \overline{X} \rangle = \frac{\langle T \rangle}{N} = \mu \tag{4.14}$$

and

$$\sigma_{\overline{X}}^2 = \operatorname{Var}(\overline{X}) = \operatorname{Var}\left(\frac{T}{N}\right) = \frac{\operatorname{Var}(T)}{N^2} = \frac{\sigma_T^2}{N^2} = \frac{N\sigma^2}{N} \tag{4.15}$$

which means

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}} \tag{4.16}$$

I.e., if you take the average of N iid random variables, the standard deviation is $1/\sqrt{N}$ times theie individual standard deviation.

4.2.1 Biased and Unbiased Estimators

We can use combinations of the data in a random sample to estimate information about the underlying data. For instance, because

$$\langle \overline{X} \rangle = \mu \tag{4.17}$$

we say the sample average $\overline{X}$ is an *unbiased estimator* of the underling mean (expectation value) of a data point in the sample.

To estimate the underlying variance σ^2, we could construct something like

$$\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2 \tag{4.18}$$

which would indeed have an expectation value of σ^2. This works fine if we already know what μ is, but often we only have the random sample itself to work with, so we need to use $\overline{X}$ to estimate μ. If we examine

$$\sum_{i=1}^{N}(X_i-\overline{X})^2 \tag{4.19}$$

we will see that its expectation value is not $N\sigma^2$ but rather $(N-1)\sigma^2$. If we divided (4.19) by N we would get a *biased estimator* of the variance, with expectation value $\frac{N-1}{N}\sigma^2$. To see this construct

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})^2 \ ; \tag{4.20}$$

Its expectation value is

$$\langle S^2 \rangle = \frac{1}{N-1}\sum_{i=1}^{N}\langle (X_i-\overline{X})^2 \rangle \tag{4.21}$$

The expectation value inside the sum is

$$\begin{aligned} \langle (X_i-\overline{X})^2 \rangle &= \left\langle \left[(X_i-\mu)-(\overline{X}-\mu)\right]^2 \right\rangle \\ &= \langle (X_i-\mu)^2 \rangle - 2\langle (X_i-\mu)(\overline{X}-\mu) \rangle + \langle (\overline{X}-\mu)^2 \rangle \end{aligned} \tag{4.22}$$

Now, we already know the first term (the variance of X_i) is σ^2 and the last term (the variance of $\overline{X}$) is σ^2/N. The cross term involves

$$\langle (X_i-\mu)(\overline{X}-\mu) \rangle = \frac{1}{N}\sum_{j=1}^{N}\langle (X_i-\mu)(X_j-\mu) \rangle = \frac{1}{N}\sum_{j=1}^{N}\delta_{ij}\sigma^2 = \sigma^2/N \tag{4.23}$$

so

$$\langle S^2 \rangle = \frac{1}{N-1}\sum_{i=1}^{N}\left(\sigma^2 - 2\frac{\sigma^2}{N} + \frac{\sigma^2}{N}\right) = \frac{N}{N-1}\left(1-\frac{1}{N}\right)\sigma^2 = \sigma^2 \tag{4.24}$$

as advertized.

4.3 PDF of a Sum of Random Variables

If we consider two independent random variables X_1 and X_2 with (not necessarily identical) pdfs $f_1(x_1)$ and $f_2(x_2)$, so that their joint pdf is

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \tag{4.25}$$

and write their sum as

$$T = X_1 + X_2 \tag{4.26}$$

we can ask what the pdf $f_T(t)$ is. One way to approach this[9] is to consider the joint pdf $p(t, x_2)$. We can do this by changing variables from x_1 to $t = x_1 + x_2$; since we're actually changing from $\{x_1, x_2\}$ to $\{t, x_2\}$, we can treat x_2 as a constant[10] and since $dt = dx_1$ in that case,

$$f(t, x_2) = f_1(t - x_2) f_2(x_2) \ . \tag{4.27}$$

if we then marginalize over x_2, we find

$$f(t) = \int_{-\infty}^{\infty} f(t - x_2)\, f_2(x_2)\, dx_2 \tag{4.28}$$

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

$$\Phi_X(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi x\xi}\, f_X(x)\, dx \tag{4.29}$$

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:

$$\Phi_X(0) = \int_{-\infty}^{\infty} f_X(x)\, dx = 1 \tag{4.30}$$

[9] An alternative, slick shortcut is to write

$$f_T(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t - [x_1 - x_2])\, f(x_1, x_2)\, dx_1\, dx_2$$

[10] Alternatively, we can consider the Jacobian determinant

$$\left\| \frac{\partial(t, x_2)}{\partial(x_1, x_2)} \right\| = \left| \det \begin{pmatrix} \left(\frac{\partial t}{\partial x_1}\right)_{x_2} & \left(\frac{\partial t}{\partial x_1}\right)_{x_1} \\ \left(\frac{\partial x_1}{\partial x_1}\right)_{x_2} & \left(\frac{\partial x_1}{\partial x_1}\right)_{x_1} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right| = 1$$

.

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

$$\Phi_X(\xi) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(-i2\pi)^n}{n!} \, x^n \, \xi^n \, f_x(x) \, dx = \sum_{n=0}^{\infty} \frac{(-i2\pi)^n}{n!} \, \langle X^n \rangle \, \xi^n \tag{4.31}$$

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

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{4.32}$$

  then

$$\Phi_x(\xi) = \exp\left(-i2\pi\mu\xi - \frac{(2\pi\xi)^2}{2\sigma^{-2}}\right) \tag{4.33}$$

  which we get as usual by completing the square in the integral over x. Note that (4.33) can be used to show that the sum of two Gaussian random variables is itself a Gaussian.

-

$$\Phi_{aX}(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi ax\xi} \, f_{aX}(ax) \, d(ax) = \int_{-\infty}^{\infty} e^{-i2\pi ax\xi} \, f_X(x) \, dx = \Phi_x(a\xi) \tag{4.34}$$

4.4 The Central Limit Theorem

We are now in a position to prove the *Central Limit Theorem*, which says that the average of many independent, identically distributed variables has a distribution which is approximately Gaussian. Note that we have seen manifestations in this by examining the large-n forms of the binomial and Poisson distributions (which represent the sum of many *Bernoulli random variables* which can take on the values 0 and 1), and of the Gamma distribution (which, in its application as the chi-squared distribution, represents the sum of many χ^2 random variables with one degree of freedom each).

Given N iid random variables $\{X_i\}$, we can construct for each of them a corresponding random variable

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{4.35}$$

An individual Z_i will not be a Gaussian rv, but it will have zero mean and unit variance.

$$\langle Z_i \rangle = 0 \qquad \text{and} \qquad \mathrm{Var}(Z_i) = 1 \tag{4.36}$$

If we construct the random variable

$$\mathcal{Z} = \frac{T_Z}{\sqrt{N}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu) = \frac{T - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma}\sqrt{N} \tag{4.37}$$

this will also have mean 0 and variance 1. Now we can show that whatever the pdf for the underlying random variables, $\mathcal{Z}$ is Gaussian-distributed in the limit $N \rightarrow \infty$. 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 $T_Z = \sum_{k=0}^{N-1} Z_k$,

$$\Phi_{T_Z}(\xi) = [\Phi_Z(\xi)]^N \tag{4.38}$$

(since the pdf $f_{T_Z}(t_z)$ can be made by convolving together N copies of $f_Z(z_i)$). Now, since $\mathcal{Z} = T_Z/\sqrt{N}$,

$$\Phi_{\mathcal{Z}}(\xi) = \Phi_{T_Z}(\xi/\sqrt{N}) = [\Phi_Z(\xi/\sqrt{N})]^N \ . \tag{4.39}$$

For large N, we can Taylor expand

$$\Phi_Z(\xi/\sqrt{N}) = 1 - \frac{i2\pi\xi \langle Z_k \rangle}{\sqrt{N}} - \frac{(2\pi\xi)^2 \langle (Z_k)^2 \rangle}{2N} + \mathcal{O}(N^{-3/2}) \tag{4.40}$$

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

$$\Phi_Z(\xi/\sqrt{N}) = 1 - \frac{(2\pi\xi)^2}{2N} + \mathcal{O}(N^{-3/2}) \tag{4.41}$$

and so

$$\Phi_{\mathcal{Z}}(\xi) = \left(1 - \frac{(2\pi\xi)^2}{2N} + \mathcal{O}(N^{-3/2})\right)^N = \left(1 - \frac{(2\pi\xi)^2}{2N}\right)^N + \mathcal{O}(N^{-1/2}) \tag{4.42}$$

and

$$\lim_{N\to\infty} \Phi_{\mathcal{Z}}(\xi) = \lim_{N\to\infty} \left(1 - \frac{(2\pi\xi)^2}{2N}\right)^N = \exp\left(-\frac{(2\pi\xi)^2}{2}\right) \ . \tag{4.43}$$

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

$$f_{\mathcal{Z}}(\mathfrak{z}) \overset{N\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} e^{-\mathfrak{z}^2/2} \tag{4.44}$$

and since

$$\mathcal{Z} = \frac{T - N\mu}{\sigma\sqrt{N}} \tag{4.45}$$

that means that for large N,

$$f_T(t) \approx \frac{1}{\sigma\sqrt{2\pi N}} \exp\left(\frac{-(t-N\mu)^2}{2N\sigma^2}\right) \tag{4.46}$$