The Gaussian distribution

Now, finally, we're ready to think about Gaussian distributions. For a Gaussian distribution with arithmetic mean μ and standard deviation σ, the normalized probability distribution is

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

assuming that x can range from $-\infty$ to $+\infty$.

The way we have written P should be read as "the probability density $P(x)$ given μ and σ". Please note that "probability density" means that the probability of x being between, say x_0 and $x_0 + dx$ (with dx being an infinitesimal) is $P(x_0)dx$. To integrate to 1, therefore, it must be that $P(x)$ has units of $1/x$, given that dx has the same units as x. That's why part of the prefactor is $1/\sigma$ (recall that σ has the same units as x).

This distribution has a lot of wonderful properties: it is symmetric, its arithmetic mean, median, and mode are always the same as each other, all moments are well defined and finite, and there are straightforward analytic expressions for all of those moments. People will often quote significances in units of σ; a 5σ result, for example. In doing so, they are using shorthand for "the probability that a draw from a Gaussian is at $+5\sigma$ or more beyond the mean" (or something similar). But why should we use it?

In fact, the Gaussian distribution crops up so often in limiting cases that it is commonly called the "normal" distribution. That, in fact, is why so many statistical tests assume Gaussian distributions.

But how can that be? There are plenty of distributions that are definitely *not* Gaussian. Our die-rolling experiment provides an example. If the die is fair, then after many rolls we expect the relative probabilities of 1 through 6 all to equal 1/6. Nothing peaked about that. Other very common and useful probability distributions are also not Gaussian. As an example of another distribution, if (a) the probability of a count in one time interval is independent of the probability of a count in the next time interval, and (b) if the probability of a count in a very short time interval is proportional to the duration of that interval, then if we expect m counts in some time, the probability of actually seeing d counts is given by the Poisson distribution:

$$P(d) = \frac{m^d}{d!} e^{-m} \ . \tag{2}$$

As yet another example, suppose that you have a source which is intrinsically steady. That is, in a given amount of time T you would always expect m counts. However, in a given measurement time T you actually see d counts, determined by the Poisson distribution above (this type of statistical variation is called Poisson variation). If you now compute the power spectrum of a data set consisting of many such measurements, then if you normalize your power spectrum such that the average is P_0, the probability of getting a power between P

and $P+dP$ is $\frac{1}{P_0}e^{-P/P_0}dP$. There are plenty of other examples of useful, common, probability distributions that arise in astronomical data sets that are *not* Gaussian.

Thus it sounds as if, despite the aesthetic beauty and analytic convenience of Gaussians, we're out of luck. But the Gaussian-favoring statistician has an ace up her sleeve: the *central limit theorem*.

In one standard form of this theorem, we suppose that we have a probability distribution $P(x)$. $P(x)$ can be anything as long as its variance is not infinite. Thus $P(x)$ could be weirdly asymmetric, multimodal, spiky, or whatever. We imagine that we select x with probability $P(x)$ (said another way, we *draw* x from the distribution $P(x)$), and do this n times, independently. Then we take the arithmetic mean of the n values of x that we obtained. The central limit theorem says that in the limit $n \rightarrow \infty$, the probability distribution of the arithmetic mean approaches a normal distribution with the same average μ as the original distribution, and with a standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of the original distribution.

I am sorry to say that I do not know of a simple, short proof of the central limit theorem. However, for completeness I give at the end a straightforward but lengthy proof.

To test this out, please now read the notes on the coding assignment for this class, and perform the analyses described there. What do you notice from the plots? How do the arithmetic mean and standard deviation of your distributions compare with what you would expect from the central limit theorem?

This is the reason that Gaussian distributions play such a prominent role in statistics. For small numbers of counts, we don't necessarily expect a Gaussian. For example, if the average number of counts in a bin is 1, and if the Poisson distribution is the right distribution, then the actual distribution of the number of counts doesn't look very Gaussian (feel free to plot this if you like). But as your average number of counts goes up, the distribution looks more and more Gaussian. Given that many analysis packages *assume* that the distribution is Gaussian (e.g., anything that has χ^2 assumes this), some analysis packages will *automatically* group bins of data so that there are enough counts that Gaussians are decent approximations. Enough people are used to this type of analysis that they think it is *necessary* to do such grouping. But it isn't. There is a more rigorous way, which we'll discuss in the classes on Bayesian statistics.

Proof of the central limit theorem

This section is a proof of the central limit theorem. It is **optional**; read it if you are interested in details, but for most purposes all you need to know is that the theorem is true, so just read the statement of the theorem below.

The theorem was apparently first proven by Laplace in 1810, but here we reproduce very

closely a proof given at the Wolfram MathWorld site
http://mathworld.wolfram.com/CentralLimitTheorem.html.

Let $p(x)$ be a probability distribution in x with mean μ and a finite standard deviation σ. Let X be a random variable defined as the average of N samples of x from $p(x)$:

$$X \equiv \frac{1}{N} \sum_{i=1}^{N} x_i \ . \tag{3}$$

Then the central limit theorem says that as $N \to \infty$, the probability distribution of X, $P(X)$, tends to a Gaussian with mean μ and standard deviation $\sigma/\sqrt{N}$. Note the capital letters here, which distinguish $P(X)$ (the probability distribution of X) from $p(x)$ (the probability distribution of x).

Consider the Fourier transform of $P(X)$, with respect to a frequency f (some references call this an inverse Fourier transform):

$$P_X(f) = \int_{-\infty}^{\infty} e^{2\pi i f X} P(X) dX \ . \tag{4}$$

When we Taylor-expand the exponential, this becomes

$$P_X(f) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(2\pi i f X)^n}{n!} P(X) dX \ . \tag{5}$$

Because the integral is over X, we can take the parts not involving X out of the integral:

$$P_X(f) = \sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \int_{-\infty}^{\infty} X^n P(X) dX \ . \tag{6}$$

But for a normalized probability distribution $P(X)$, so that $\int_{-\infty}^{\infty} P(X)dX = 1$, the integral in the above equation is just the expectation value of X^n, or $\langle X^n \rangle$, so we find

$$P_X(f) = \sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \langle X^n \rangle \ . \tag{7}$$

Recalling that

$$X = N^{-1}(x_1 + x_2 + \ldots + x_N) \ , \tag{8}$$

this means that

$$\begin{aligned} \langle X^n \rangle &= \langle N^{-n}(x_1 + x_2 + \ldots + x_N)^n \rangle \\ &= \textstyle\int_{-\infty}^{\infty} N^{-n}(x_1 + x_2 + \ldots + x_N)^n p(x_1)p(x_2) \cdots p(x_N) dx_1 \cdots dx_N \ . \end{aligned} \tag{9}$$

Thus we can write

$$\begin{aligned} P_X(f) &= \textstyle\sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \langle X^n \rangle \\ &= \textstyle\sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \int_{-\infty}^{\infty} N^{-n}(x_1 + \ldots + x_N)^n p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \\ &= \textstyle\int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \left[\frac{2\pi i f(x_1 + \ldots + x_N)}{N} \right]^n \frac{1}{n!} p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \ . \end{aligned} \tag{10}$$

Note that the sum through the $1/n!$ factor is just the Taylor series for another exponential, so we can write this as

$$P_X(f) = \int_{-\infty}^{\infty} e^{2\pi i f (x_1 + \ldots + x_N)/N} p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \ . \tag{11}$$

The exponential is of course the product of the exponential of the individual terms in the exponents, so

$$P_X(f) = \left[\int_{-\infty}^{\infty} e^{2\pi i f x_1/N} p(x_1) dx_1\right] \times \cdots \times \left[\int_{-\infty}^{\infty} e^{2\pi i f x_N/N} p(x_N) dx_N\right] \ . \tag{12}$$

But all of the x_i's are drawn from the same probability distribution $p(x)$, so this becomes

$$P_X(f) = \left[\int_{-\infty}^{\infty} e^{2\pi i f x/N} p(x) dx\right]^N \ . \tag{13}$$

Now we'll Taylor-expand the exponent once more, but this time we will keep only the first few terms:

$$P_X(f) = \left\{\int_{-\infty}^{\infty} \left[1 + \left(\frac{2\pi i f}{N}\right) x + \frac{1}{2}\left(\frac{2\pi i f}{N}\right)^2 x^2 + \mathcal{O}(N^{-3})\right] p(x) dx\right\}^N \ . \tag{14}$$

Using $\langle x\rangle = \int_{-\infty}^{\infty} x p(x) dx$ and similarly for $\langle x^2\rangle$, we get

$$\begin{aligned} P_X(f) \ &= \left[1 + \tfrac{2\pi i f}{N}\langle x\rangle - \tfrac{(2\pi f)^2}{2N^2}\langle x^2\rangle + \mathcal{O}(N^{-3})\right]^N \\ &= \exp\left\{N \ln\left[1 + \tfrac{2\pi i f}{N}\langle x\rangle - \tfrac{(2\pi f)^2}{2N^2}\langle x^2\rangle + \mathcal{O}(N^{-3})\right]\right\} \ . \end{aligned} \tag{15}$$

In the second step we just used the identity $Y^N = \exp(N \ln Y)$.

Now we use Taylor series again: $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \ldots$. Remember that we are thinking about the limit $N \to \infty$, which means that our logarithm is indeed of an argument that is 1 plus a small quantity. With this approximation, and expanding to the square of that small quantity, we get

$$P_X(f) = \exp\left[N\left[\frac{2\pi i f}{N}\langle x\rangle - \frac{(2\pi f)^2}{2N^2}\langle x^2\rangle - \frac{1}{2}\frac{(2\pi i f)^2}{N^2}\langle x\rangle^2 + \mathcal{O}(N^{-3})\right]\right\} \ . \tag{16}$$

Simplifying the exponent, and regrouping terms, we get

$$\begin{aligned} P_X(f) \ &= \exp\left[2\pi i f\langle x\rangle - \tfrac{(2\pi f)^2(\langle x^2\rangle - \langle x\rangle^2)}{2N} + \mathcal{O}(N^{-2})\right] \\ &\approx \exp\left[2\pi i f\mu - \tfrac{(2\pi f)^2\sigma^2}{2N}\right] \ , \end{aligned} \tag{17}$$

because $\langle x\rangle = \mu$ and $\langle x^2\rangle - \langle x\rangle^2 = \sigma^2$.

Taking the Fourier transform again,

$$\begin{aligned} P(X) &= \int_{-\infty}^{\infty} e^{-2\pi i f X} P_X(f) df \\ &= \int_{-\infty}^{\infty} e^{2\pi i f(\mu - X) - (2\pi f)^2 \sigma^2 / 2N} df \ . \end{aligned} \tag{18}$$

This integral is of the form

$$\int_{-\infty}^{\infty} e^{iaf - bf^2} df = e^{-a^2/4b} \sqrt{\pi/b} \ , \tag{19}$$

so after we substitute $a = 2\pi(\mu - X)$ and $b = (2\pi\sigma)^2/2N$ we get finally

$$P(X) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} e^{-(\mu - X)^2/2(\sigma/\sqrt{N})^2} \ . \tag{20}$$

Q.E.D. (at last!)