Is it Gaussian?

Many statistical approaches, including the ubiquitous χ^2, *assume* that the relevant probability distribution is Gaussian. In this class we'll talk about what that means, and the circumstances in which it is or is not appropriate to use.

Before talking about Gaussians per se, let's discuss some fundamental aspects of probability distributions. To do that, when we need something specific we'll use the following data set, which I obtained by virtually rolling dice and then sorting the numbers in increasing order:
1,1,2,3,3,4,5,6,6,6.

A properly normalized probability distribution $P(\mathbf{x})$, where $\mathbf{x}$ indicates the parameters (written here as a vector, i.e., there could be multiple parameters), has the property that

$$\int P(\mathbf{x}) d\mathbf{x} = 1 \ , \tag{1}$$

where the integral is over all possible values of $\mathbf{x}$. For any parameters that can only take on a set of discrete values, the integral is replaced by a sum.

For our specific case, let x represent the number on the die, so that the full set of possibilities is $x = 1, 2, 3, 4, 5, 6$. We know that for a fair die, $P(x = 1) = 1/6$, $P(x = 2) = 1/6, \ldots, P(x = 6) = 1/6$. But our particular data don't have that distribution. Instead, for this data set, $P(x = 1) = 2/10$, $P(x = 2) = 1/10$, $P(x = 3) = 2/10$, $P(x = 4) = 1/10$, $P(x = 5) = 1/10$, and $P(x = 6) = 3/10$.

Clearly we retain all of the information if we just list the data points. But often we want a quick look at the data, and for that purpose we might want to characterize it in different ways. Here are some of those ways, and please keep in mind that many of these only apply to a *one-dimensional* probability distribution:

The "average".—Often we'd like a single best value to describe a distribution. The average is a good choice... except that there are many different types of average! Here are some examples:

1. The median. This is the value such that half the values are below the median, and half the values are above. In our specific example, the median is 3.5 because half of the ten values are below this, and half of the ten values are above this. If we have a continuous distribution $P(x)$, then the median value $x_{\rm median}$ is the solution to

$$\int_{x_{\rm min}}^{x_{\rm median}} P(x) dx = 0.5 \ . \tag{2}$$

   Here $x_{\rm min}$ is the minimum possible value of x. The median is a good measure of the average if you want to avoid being biased by outliers. For example, suppose you

compute the arithmetic mean (see below) of the personal wealth of the people in your small town, and the answer is \$100 million. What a rich community! But maybe Bill Gates lives in your small town, and in reality most people are dirt poor. The median would give a better idea of how the typical person is doing.

2. The mode. This is the single most common value in your data. In our case, 6 appears 3 times, which is more than any other number, so it is the mode. For a continuous distribution, it's the peak of that distribution, so $x_{\rm mode}$ is such that the largest value of $P(x)$ is at $P(x_{\rm mode})$.

3. The mean. Here we usually talk about the arithmetic mean, but there are other variants. Examples:

   (a) The arithmetic mean. For a set of discrete values, you just add them up and divide by the total number of values: in our case the sum is 1+1+2+3+3+4+5+6+6+6=37, and there are 10 values, so the arithmetic mean is 37/10=3.7. For a continuous distribution, the arithmetic mean is $\langle x\rangle = \int_{x_{\rm min}}^{x_{\rm max}} xP(x)dx$. Note again that this requires that $P(x)$ is normalized so that $\int_{\rm x_{min}}^{x_{\rm max}} P(x)dx = 1$. This is also our first example of a *moment* of the probability distribution $P(x)$; it is the first moment, because the thing multiplying $P(x)$ in the integral is x^1.

   (b) The geometric mean. This is the nth root of the product of the n measurements. In our case, the geometric mean is $(1*1*2*3*3*4*5*6*6*6)^{1/10} = 3.08$. This type of mean isn't used a lot in probability and statistics, but it does enter in some physical processes (e.g., some problems in radiative transfer).

   (c) The harmonic mean. This is the reciprocal of the arithmetic mean of the reciprocals of the n measurements. In our case, the harmonic mean is $10/(1/1+1/1+1/2+1/3+1/3+1/4+1/5+1/6+1/6+1/6) = 2.43$. Again, this doesn't enter much in statistics, but it does tend to put greater weight on smaller values, which can be useful in other types of radiative transfer (e.g., it is related to the Rosseland mean opacity).

That's all very well, but even if you have carefully selected one of these measures, you have limited information. For example, the following distributions have the same median, mode, and arithmetic mean: (1) ten 3's, (2) three 1's, four 3's, and three 5's, (3) one 1, two 2's, four 3's, two 4's, and one 5. They are clearly different, however, so it would be good to have a way to distinguish them.

The variance.—This is a measure of the spread of the numbers. To get to the definition, we can define the second moment of the distribution, which for a continuous probability function is

$$\langle x^2\rangle = \int x^2 P(x)dx\ . \tag{3}$$

To reiterate, this formula is only valid if $P(x)$ has been normalized such that $\int P(x)dx = 1$. This is therefore the average of x^2 over the probability distribution (and as always if we have a discrete probability distribution, we sum rather than integrating). For our sample data set, $\langle x^2 \rangle = (1/10)(1^2 + 1^2 + 2^2 + 3^2 + 3^2 + 4^2 + 5^2 + 6^2 + 6^2 + 6^2) = 17.3$. But note that this really isn't what we want. You could imagine, for example, some tight distribution with a large arithmetic mean (say, 100), such that $\langle x^2 \rangle$ is large; that wouldn't tell us what we want to know, which is how much the data are spread. What we'd really like to know, therefore, is the average of the square of the deviation from the mean:

$$\begin{aligned} \langle (x - \langle x \rangle)^2 \rangle &= \int (x - \langle x \rangle)^2 P(x) dx \\ &= \int x^2 P(x) dx - 2 \int x \langle x \rangle P(x) dx + \int \langle x \rangle^2 P(x) dx \\ &= \langle x^2 \rangle - 2 \langle x \rangle \int x P(x) dx + \langle x \rangle^2 \int P(x) dx \\ &= \langle x^2 \rangle - 2 \langle x \rangle^2 + \langle x \rangle^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{4}$$

This is the *variance* of the distribution, and its square root is the *standard deviation* (note that the variance can never be negative, so a square root is okay!); often the standard deviation is represented by σ, and often the arithmetic mean is represented by μ. Note that the standard deviation has the same units as the mean. For our specific case, $\sigma^2 = 17.3 - (3.7)^2 = 3.61$, and therefore the standard deviation is a pleasingly exact $\sigma = 1.9$.

So now we have two measures of the distribution. Of course, these don't capture every aspect of the distribution. For example, there are many distributions that have the same mean and standard deviation but are asymmetric in different ways. To deal with this there is a quantity called the skewness, which can be written using our previous notation as

$$\gamma_1 = \left(\langle x^3 \rangle - 3\mu\sigma^2 - \mu^3 \right) / \sigma^3 \, . \tag{5}$$

We could then go to the fourth moment and define something called the kurtosis, which can be thought of as a measure of how peaked the distribution is, and so on. However, we need to keep in mind that (1) the original full distribution contains all of the information, so (2) if we are using mean, standard deviation, and so on to characterize the distribution, then we are being concise in a way that could throw away some information.

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

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \, , \tag{6}$$

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:

$$P(d) = \frac{m^d}{d!} e^{-m} \ . \tag{7}$$

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 prob-

ability $P(x)$ (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 $n \to \infty$, 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 next three classes.

Proof of the central limit theorem

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 $p(x)$:

$$X \equiv \frac{1}{N} \sum_{i=1}^{N} x_i \; . \tag{8}$$

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):

$$P_X(f) = \int_{-\infty}^{\infty} e^{2\pi i f X} P(X) dX \ . \tag{9}$$

When we Taylor-expand the exponential, this becomes

$$P_X(f) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(2\pi i f X)^n}{n!} P(X) dX \ . \tag{10}$$

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

$$P_X(f) = \sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \int_{-\infty}^{\infty} X^n P(X) dX \ . \tag{11}$$

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

$$P_X(f) = \sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \langle X^n \rangle \ . \tag{12}$$

Recalling that

$$X = N^{-1}(x_1 + x_2 + \ldots + x_N) \ , \tag{13}$$

this means that

$$\begin{aligned} \langle X^n \rangle &= \langle N^{-n}(x_1 + x_2 + \ldots + x_N)^n \rangle \\ &= \textstyle\int_{-\infty}^{\infty} N^{-n}(x_1 + x_2 + \ldots + x_N)^n p(x_1) p(x_2) \cdots p(x_N) dx_1 \cdots dx_N \ . \end{aligned} \tag{14}$$

Thus we can write

$$\begin{aligned} P_X(f) &= \textstyle\sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \langle X^n \rangle \\ &= \textstyle\sum_{n=0}^{\infty} \frac{(2\pi i f)^n}{n!} \int_{-\infty}^{\infty} N^{-n}(x_1 + \ldots + x_N)^n p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \\ &= \textstyle\int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \left[\frac{2\pi i f (x_1 + \ldots + x_N)}{N} \right]^n \frac{1}{n!} p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \ . \end{aligned} \tag{15}$$

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

$$P_X(f) = \int_{-\infty}^{\infty} e^{2\pi i f (x_1 + \ldots + x_N)/N} p(x_1) \cdots p(x_N) dx_1 \cdots dx_N \ . \tag{16}$$

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

$$P_X(f) = \left[\int_{-\infty}^{\infty} e^{2\pi i f x_1/N} p(x_1) dx_1 \right] \times \cdots \times \left[\int_{-\infty}^{\infty} e^{2\pi i f x_N/N} p(x_N) dx_N \right] \ . \tag{17}$$

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

$$P_X(f) = \left[\int_{-\infty}^{\infty} e^{2\pi i f x/N} p(x) dx\right]^N . \tag{18}$$

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

$$P_X(f) = \left\{\int_{-\infty}^{\infty} \left[1 + \left(\frac{2\pi i f}{N}\right) x + \frac{1}{2}\left(\frac{2\pi i f}{N}\right)^2 x^2 + \mathcal{O}(N^{-3})\right] p(x) dx\right\}^N . \tag{19}$$

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

$$\begin{aligned} P_X(f) &= \left[1 + \tfrac{2\pi i f}{N}\langle x\rangle - \tfrac{(2\pi f)^2}{2N^2}\langle x^2\rangle + \mathcal{O}(N^{-3})\right]^N \\ &= \exp\left\{N \ln\left[1 + \tfrac{2\pi i f}{N}\langle x\rangle - \tfrac{(2\pi f)^2}{2N^2}\langle x^2\rangle + \mathcal{O}(N^{-3})\right]\right\} . \end{aligned} \tag{20}$$

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

$$P_X(f) = \exp\left[N\left[\frac{2\pi i f}{N}\langle x\rangle - \frac{(2\pi f)^2}{2N^2}\langle x^2\rangle - \frac{1}{2}\frac{(2\pi i f)^2}{N^2}\langle x\rangle^2 + \mathcal{O}(N^{-3})\right]\right] . \tag{21}$$

Simplifying the exponent, and regrouping terms, we get

$$\begin{aligned} P_X(f) &= \exp\left[2\pi i f\langle x\rangle - \tfrac{(2\pi f)^2(\langle x^2\rangle - \langle x\rangle^2)}{2N} + \mathcal{O}(N^{-2})\right] \\ &\approx \exp\left[2\pi i f\mu - \tfrac{(2\pi f)^2\sigma^2}{2N}\right] , \end{aligned} \tag{22}$$

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

Taking the Fourier transform again,

$$\begin{aligned} P(X) &= \int_{-\infty}^{\infty} e^{-2\pi i f X} P_X(f) df \\ &= \int_{-\infty}^{\infty} e^{2\pi i f(\mu - X) - (2\pi f)^2\sigma^2/2N} df . \end{aligned} \tag{23}$$

This integral is of the form

$$\int_{-\infty}^{\infty} e^{iaf - bf^2} df = e^{-a^2/4b}\sqrt{\pi/b} , \tag{24}$$

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

$$P(X) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} e^{-(\mu - X)^2/2(\sigma/\sqrt{N})^2} . \tag{25}$$

Q.E.D. (at last!)