Probability Distributions: Averages Etc.

In this lecture we will 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 $(1\times1\times2\times3\times3\times4\times5\times6\times6\times6)^{1/10} \approx 3.08$. 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 $\sigma = 1.9$.

[By the way, if you have n samples from a distribution and you want to estimate the variance of the underlying distribution, then for technical reasons you would need to multiply the value above by $n/(n-1)$, otherwise your estimate will be biased. But here we are simply computing the variance among the numbers in the sample.]

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.