Probability and Statistics

As you will learn throughout your careers, analysis of data is fundamental to astronomy, and statistics is fundamental to analysis of data. Statistics, and the probability theory that underlies it, is necessary to answer questions such as: what is the value and uncertainty of the redshift to that galaxy? Do my data give convincing evidence of a detection? and so on. For a live example of this, look at Figure 1, which shows a recent cosmic ray spectrum. Could the high point be significant? If it is, then we need to think about what it would mean. For this particular case, some people hope that the excess comes from dark matter decay or annihilation, which would be a Nobel Prize winning discovery. It is therefore very important to do our analysis correctly!

It is, of course, not possible to give a full summary of probability and statistics in a short set of notes, or one lecture! Fortunately, there are many online resources you can consult, and classes you can take. I will also mention in the spring of 2019 I ran a "Practical Astrostatistics" class for our astronomy undergraduates. The website is http://www.astro.umd.edu/∼miller/teaching/astr288a/, and it has lecture notes and computational exercises on data sets (some synthetic, some actually taken from the astronomical literature). From those lectures I will excerpt my prime goal for the students in the class:

"My intent in this course is to get you to ask yourself two questions before you perform any statistical analysis. First:
How *would* I perform this analysis if I had unlimited time and resources?
and then second:
How *can* I perform my analysis, with the least loss of accuracy and precision, given my finite time and resources?"

Your ability to ask and answer those questions will evolve as you gain knowledge and experience, but you should always *think* about the answers you get, and should be wary of using "black box" analysis tools whose assumptions you don't know; maybe those assumptions don't work in your particular case!

We will now go over some of the basic rules of probability, and in the following section we will list some common sins of statistical analysis.

1. Basic Probability

Suppose you perform an experiment or make an observation. You are looking for some particular outcome. For instance, you might roll a die and look for occurrences of the number

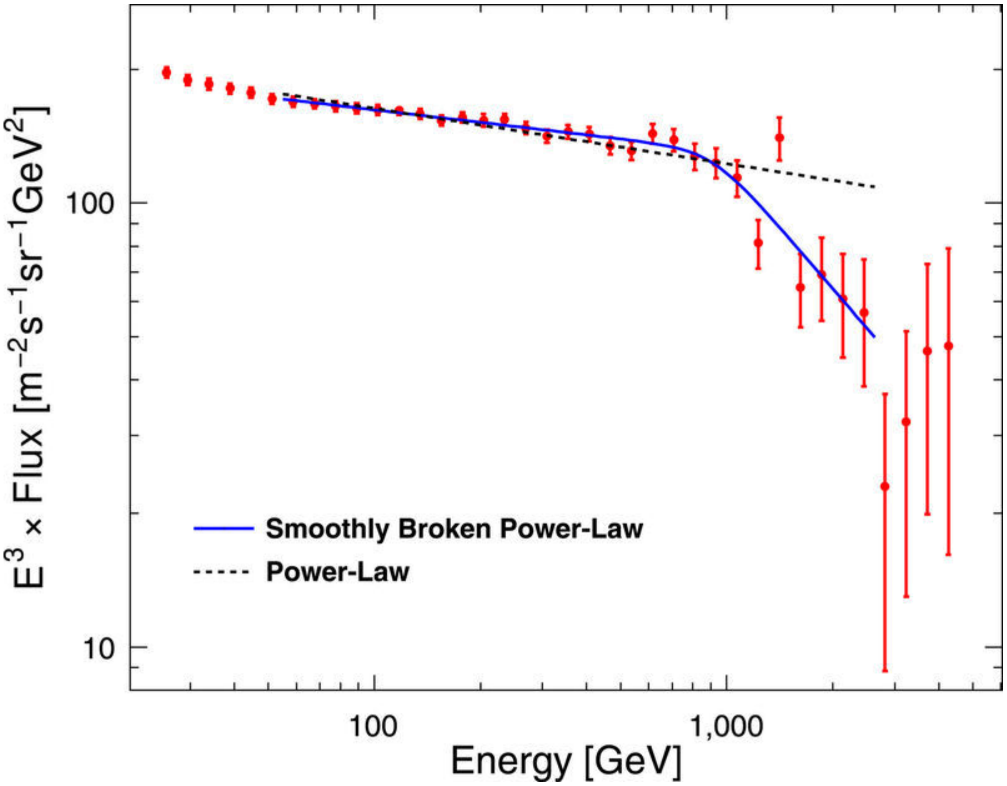

Fig. 1.— Energy spectrum of cosmic ray electrons plus positrons, as reported by the DArk Matter Particle Explorer (DAMPE) collaboration. The red points with associated uncertainties are the data; the black dashed line is the best fit of a single power law (note that both the horizontal and vertical axes are logarithmic, which means that power laws, which are of the form Flux $\propto$ Energy$^{\alpha}$, appear as straight lines); and the blue line is the best fit of a smoothly broken power law. An example of a statistical question is: is the single high point at the spectrum around 1,400 GeV energy significant? The stakes are high; some researchers think that such an excess might be a signature of dark matter. We need to make sure that our analyses are reliable!

4. We can then define the probability of the given outcome as

$$P = \text{Fraction of ways the specified outcome can happen, among all possible outcomes}\ . \tag{1}$$

One way to restate this is that if you were to do the experiment over and over again, then the ratio of times that you got your specified outcome to the total number of experiments tends to approach closer and closer to P with more and more experiments. For example, suppose the die we are rolling is fair. Then there are 6 equally likely outcomes, of which 4 is one, thus the probability is $P(4) = 1/6$. If instead the die is loaded, so that on average 4 comes up in 9 of 10 rolls, then $P(4) = 9/10$.

Put in a more axiomatic way:

- A probability is a number between 0 and 1 inclusive: $0 \leq P(A) \leq 1$ for any outcome A. 0 means impossibility, and 1 means certainty.
- If you have two or more non-overlapping outcomes $A, B, C, \ldots$ then the probability that any of them happen is $P(A) + P(B) + P(C) + \ldots$ If you add up the probabilities of all the possible nonoverlapping outcomes, you get 1.

From these it follows that:

1. The probability that an outcome A occurs, plus the probability that it does not occur, is always 1: $P(A) + P(\text{not } A) = 1$.
2. If you have two outcomes A and B, the probability that either or both happens is the sum of their individual probabilities minus the probability that both occur: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.
3. If you have two outcomes A and B, the probability that both occur is the product of the probability of one with the probability that the other happens *given* that the first one happened: $P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$, where for example $P(A|B)$ is read as "the probability of A given that B happened". If the probabilities are independent, this reduces to $P(A)P(B)$.

We can always figure out the theoretical probability of some outcome by enumerating all possibilities. For example, if I flip a fair coin, so that heads (H) and tails (T) are equally likely, then how probable is it that in two flips of the coin I get one head and one tail? The four mutually exclusive possibilities are: HH, HT, TH, TT. All four have a probability of $(1/2) \times (1/2) = 1/4$, and two of them (HT and TH) give the desired outcome, so the probability is $1/4 + 1/4 = 1/2$.

But when the number of possibilities is large, we need some kind of shortcut. For example, in a hundred flips of a fair coin, how likely is it that you will get 36 heads? There is no way that you would want to list out all $2^{100} = 1,267,650,600,228,229,401,496,703,205,376$ possibilities for the 100 flips. How can you save yourself some effort?

1.1. Permutations and combinations

When counting possibilities, we need to know exactly what we mean. For instance, are we interested in a *particular* sequence of heads and tails, say HHTTHHT, or in the number of ways that you could get the same number of heads and tails (respectively four and three in that example) but in any order? The first possibility leads to the concept of permutations, the second to the concept of combinations.

As an example of the use of permutations, consider the following problem. We have three balls, numbered 1, 2, and 3. We also have three slots labeled 1, 2, and 3. If the balls are put randomly into the slots, what is the probability that the order matches, i.e., that ball 1 is in slot 1, ball 2 is in slot 2, and ball 3 is in slot 3? We can list out all the possibilities by making the first number the number of the ball in slot 1, the second the number of the ball in slot 2, and the third the number of the ball in slot 3. We then have 123, 132, 213, 231, 312, and 321 as options, for 6 total. Of these, only 123 has the correct order, so the probability is 1/6. But what if we have more balls and slots, say 5, 10, or 20? Brute force counting rapidly becomes unworkable, but we can reason it out. Suppose we take five slots and balls as an example:

1. Any of the five balls can go in the first slot.
2. Given that one ball has already been placed, any of the other four can go in the second slot.
3. There are now three balls that can go in the third slot.
4. This leaves two balls that can go in the fourth slot.
5. And finally we have only one ball remaining, which must go in the fifth slot.

This tells us that the total number of ways to arrange balls in slots is $5\times4\times3\times2\times1 = 120$. The shorthand for this is the factorial symbol: $5! \equiv 5 \times 4 \times 3 \times 2 \times 1$. By convention $0! = 1$, then $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$, and so on. The numbers grow large very quickly. For example, $20! = 2,432,902,008,176,640,000$. This tells you that if you have only three balls and slots you could get the right order by random chance a decent fraction of the time, but with 20 balls and 20 slots there is no practical likelihood of getting the right order randomly.

In contrast, in combinations you do not care about the order, just the membership in the set. For example, suppose that you have three balls in a sack: red, green, and blue. If you take out two of them, what is the probability that they will be the red and blue ones? Again we can enumerate all the possibilities: RG, GR, RB, BR, GB, BG. Of these, only RB and BR satisfy our criterion, so the probability is 2/6 or 1/3. What if we have four balls, red, green, blue, and yellow, and we want to know the probability of taking out red, blue, and yellow in some order? Or what if we had 8 objects and we wanted to know in how many ways we could pick out some specific 5 in any order?

This is where we get to combinations. Suppose we have n objects and we want to pick out m of them. We have n choices for the first object, $n-1$ for the second, $n-2$ for the third, and so on down to $n-m+1$ for the mth object. In fact, we can write this as $n!/(n-m)!$ because

$$n \cdot (n-1) \cdot (n-2) \ldots (n-m+1) = \frac{n \cdot (n-1) \cdot (n-2) \ldots 2 \cdot 1}{(n-m) \cdot (n-m-1) \cdot (n-m-2) \ldots 2 \cdot 1} \ . \qquad (2)$$

But we don't care about the order of the objects. For each set of m objects there are $m!$ ways to arrange them. Since the order is irrelevant, we must divide by $m!$. This tells us that the number of ways to choose m objects from n is

$$\binom{n}{m} = \frac{n!}{(n-m)!m!} \qquad (3)$$

and this is read as "n choose m". Notice that there is a symmetry: $\binom{n}{m} = \binom{n}{n-m}$.

Examples

1. Madame Zelda claims to have mystical psychic powers that allow her to predict the outcome of flipped coins. She offers a demonstration, writing in advance her predictions for a succession of 10 flips. She gets 8 of the 10 right. Assuming that the coin is fair and that she hasn't cheated, how likely would it be that she could get at least that many by blind chance? Put another way, if she then suggested that you hand over your entire bank account so that she can use her powers in investments, would you do it?

Answer:

Here we are asking: assuming complete randomness, how likely is the actual outcome? If it is very unlikely, we can discount the hypothesis that the guesses were random chance (but then we would have to determine whether Madame Zelda was actually psychic, or was using some magician's trick).

In Madame Zelda's case, our hypothesis is that every guess has a 50% chance of being right. Therefore every sequence (hit, miss, miss, hit, hit, ...) is equally probable. What we need to determine is how many outcomes will give her 8 or more hits by chance, and how many total outcomes there are. The ratio of favorable outcomes to total outcomes is, by definition, the probability.

Since there are 10 flips and for each there are two possibilities (hit or miss), there are $N_{\text{tot}} = 2^{10} = 1024$ total outcomes. We want to know the number of ways in which she could get 8 *or more* right in any order, so we need to sum the number of ways she could get 8 right (which is $\binom{10}{8}$), with the number of ways she could get 9 right (which is $\binom{10}{9}$), with the number of ways she could get all 10 right (which is $\binom{10}{10}$). We find that $\binom{10}{8} = 10!/(8!2!) = 10 \times 9/2 = 45$, $\binom{10}{9} = 10!/(9!1!) = 10/1 = 10$, and $\binom{10}{10} = 10!/(10!0!) = 1$. There are therefore $N_{\text{favor}} = 45 + 10 + 1 = 56$ favorable outcomes. The chance probability is therefore 56/1024=5.5%. Mildly impressive, but not enough to convince yourself that she is worth a major investment!

2. Suppose you flip a coin 100 times and it comes up with heads 64 of them. Do you have good evidence that the coin is not fair?

Answer:

Again we want to test the hypothesis that the coin *is* fair. There are 2^{100} possible outcomes. Of these, you have $\binom{100}{64}$ in which you get exactly 64 heads, $\binom{100}{65}$ in which you get exactly 65, and so on. Therefore, we add up all the cases in which we get 64 or more heads, and divide by 2^{100} to determine the probability. Note that, as in the case with Madame Zelda, we are *not* just determining the number of cases with *exactly* 64 heads. Why? Because the probability of hitting some exact number gets smaller and smaller as the number of flips increases, and if we did this ratio we would get an unreasonably small number. Summing over all the possibilities with 64 heads or more indicates the probability that the imbalance would be what we see or larger.

In any case, $\binom{100}{64} = 1.98 \times 10^{27}$, $\binom{100}{65} = 1.10 \times 10^{27}$, $\binom{100}{66} = 5.8 \times 10^{26}$, $\binom{100}{67} = 2.95 \times 10^{26}$, and so on. Notice that the numbers are getting smaller rather rapidly, by about a factor of 2 each time. Therefore, additional terms don't contribute that much, and to rough accuracy the total is about 4×10^{27}. In comparison, $2^{100} = 1.27 \times 10^{30}$, so the ratio gives us a probability of $4 \times 10^{27}/1.27 \times 10^{30} \approx 0.003$. There are only 3 chances in 1000 that a fair coin would give this number of heads or more in 100 flips. You have some reason to be suspicious.

Note an interesting aspect of the last two problems: Madame Zelda hit on 80% of the flips, and here we have heads on only 64% of the flips, yet in the 64/100 case the chance probability is significantly less despite the lower fraction. Why is that? It is a consequence of the law of large numbers. Specifically, the more trials you do, the closer the *fraction* of a given outcome will be to the actual probability. As an extreme example of this, note that if you flip a fair coin twice, there is a 25% chance of getting two heads, so that's no big deal despite a 100% success rate. On the other hand, getting all heads a million times in a row is vanishingly improbable if the coin is fair. Something to keep in mind when you read the results of polls is that unless they have asked a large and representative sample of people, the poll results don't mean much.

1.2. The binomial theorem

The previous examples both involved a 50% probability, but what if the fraction is different? For example, suppose that some experiment has only two, mutually exclusive, outcomes: A and B. The probability of A is a, and the probability of B is b. If you do n trials, what is the probability that you will see A m times?

To answer this, we use the binomial theorem. The key is to consider the quantity

$$(a+b)^n \ . \tag{4}$$

Since A and B are mutually exclusive, $b = 1 - a$, meaning that the quantity in parentheses is 1 and thus $1^n = 1$. However, just as with flipping coins, we note that $(a+b)^n$ can be used to write out every possible outcome in sequence. For example,

$$(a+b)^2 = (a+b)(a+b) = aa + ab + ba + bb \ . \tag{5}$$

Similarly,

$$(a+b)^3 = (a+b)(a+b)(a+b) = aaa + aab + aba + abb + baa + bab + bba + bbb \ . \tag{6}$$

If we now say that the order doesn't matter, so that $ab = ba$ and $aab = aba = baa$ as examples, we can group those two cases as

$$(a+b)^2 = a^2 + 2ab + b^2 = \binom{2}{2}a^2 + \binom{2}{1}ab + \binom{2}{0}b^2 \tag{7}$$

and

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3 = \binom{3}{3}a^3 + \binom{3}{2}a^2b + \binom{3}{1}ab^2 + \binom{3}{0}b^3 \ . \tag{8}$$

More generally,

$$(a+b)^n = \binom{n}{n}a^n + \binom{n}{n-1}a^{n-1}b + \ldots + \binom{n}{1}ab^{n-1} + \binom{n}{0}b^n \ . \tag{9}$$

Remember that this is all guaranteed to add up to 1 because $b = 1 - a$. Therefore, the first term is the probability that in n trials A happens n times, the second term is the probability that in n trials A happens $n-1$ times (in any order), and so on. We have been using the special case in which $a = b = 0.5$. As a result, $a^n = a^{n-1}b = a^{n-2}b^2 = \ldots = ab^{n-1} = b^n$. These all factor out, which is why we were able to ignore them previously (although actually we didn't; note that 0.5^{10}, which is this factor, is what we multiplied by to get the probability).

1.3. Discrete and continuous probability distributions

Let us now return to the idea of probability itself. Our examples so far have been for *discrete* distributions, i.e., ones in which the thing you are measuring can take on only a finite number of values. Examples include a flip of a coin (heads or tails), a roll of a die (1, 2, 3, 4, 5, 6), and many others.

But there are plenty of cases in which the quantity of interest could take on a *continuous* set of values. For example, what is the probability distribution of the high temperature tomorrow as measured at some precise point? Temperature measurements, in principle, could be anything; for example, they don't have to come in integer numbers of degrees even though that's how they're often represented.

This raises what appears to be a problem. When we think of discrete outcomes, each outcome can have a nice, finite probability. For example, the probability of getting a 4 in a roll of a fair die is 1/6. But what is the probability that tomorrow's high temperature at some exact point will be 283.3953817233 K? Well, it's basically zero. So we need another way to represent the probability.

That way uses calculus. In the case of a continuous probability distribution, we talk about the probability *density* $P(x)$ for some variable x (e.g., maybe x is the temperature). This is defined such that

$$P(x_0)dx \tag{10}$$

is the probability that x is between x_0 and $x_0 + dx$. Note that, like dx itself, $P(x)dx$ is infinitesimal, but $P(x)$ can be perfectly finite. Because the probability that *something* will be measured is always 100%, it must be that the integral of $P(x)$ over all possible values of x is 1; if the minimum possible value of x is $x_{\rm min}$ and the maximum possible value is $x_{\rm max}$, then

$$\int_{x_{\rm min}}^{x_{\rm max}} P(x)dx = 1\ . \tag{11}$$

For example, let's say that the probability density for the temperature T is $P(T) = (1/8)\mathrm{K}^{-1}$ for a temperature between $T_{\rm min} = 283$ K and $T_{\rm max} = 291$ K, and zero otherwise.

Then

$$\int_{283\ \mathrm{K}}^{291\ \mathrm{K}} P(T)dT = 1\ . \tag{12}$$

Note that because dT has the same units as T, i.e., Kelvin, $P(T)$ must have units of inverse Kelvin so that the integral is 1.

What this means is that if your distribution is *discrete*, the probabilities must *sum* to 1 over the possible outcomes, whereas if your distribution is *continuous*, the probabilities must *integrate* to 1 over the possible outcomes. This pattern (sum for discrete, integrate for continuous) appears in many contexts, and is indicative of the close relation between sums and integrals.

2. Statistical sins and Gaussians

For this set of notes we have two goals: to acquaint you with some common errors made by astronomers when they do statistical analysis, and to give you a brief introduction to the mean, median, mode, variance, standard deviation, and Gaussians. These notes are largely taken from my "Practical Astrostatistics" class.

Mean, median, mode, variance, standard deviation, and Gaussians

For the following, let's consider 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.

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 the three most common:

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 $P(x)$, such that the probability of finding a value between x and $x+dx$ is $P(x)dx$, then the median value $x_{\rm median}$ is the solution to

$$\int_{x_{\rm min}}^{x_{\rm median}} P(x)dx = 0.5\ . \tag{13}$$

Here $x_{\rm min}$ is the minimum possible value of x. Note here that it is important that the probability distribution $P(x)$ is *normalized* so that $\int_{x_{\rm min}}^{x_{\rm max}} P(x)dx = 1$; this is necessary because the probability of measuring *some* value is always 1!

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 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 $\langle x\rangle \equiv \int_{x_{\rm min}}^{x_{\rm max}} xP(x)dx$. Note again that this requires that $P(x)$ is normalized so that $\int_{\mathrm{x}_{\rm min}}^{x_{\rm max}} P(x)dx = 1$. 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.

   By the way, this is called the "arithmetic mean" because there are two other types of mean that are used more rarely: (1) the "geometric mean", which is the Nth root of the product of the N values (e.g., if your values are 1, 2, and 3, the geometric mean is $(1 \times 2 \times 3)^{1/3} \approx 1.82$) and (2) the "harmonic mean", which is the reciprocal of the sum of the reciprocals (e.g., if your values are again 1, 2, and 3, the harmonic mean is $1/(1/1 + 1/2 + 1/3) \approx 0.55$). If someone says just "the mean" it's a good bet they are talking about the arithmetic mean.

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

$$\langle x^2\rangle = \int x^2 P(x)dx \; . \tag{14}$$

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:

$$\begin{aligned}
\langle (x - \langle x \rangle)^2 \rangle &= \int (x - \langle x \rangle)^2 P(x) dx \\
&= \int x^2 P(x) dx - 2 \int x \langle x \rangle P(x) dx + \int \langle x \rangle^2 P(x) dx \\
&= \langle x^2 \rangle - 2 \langle x \rangle \int x P(x) dx + \langle x \rangle^2 \int P(x) dx \\
&= \langle x^2 \rangle - 2 \langle x \rangle^2 + \langle x \rangle^2 \\
&= \langle x^2 \rangle - \langle x \rangle^2
\end{aligned} \tag{15}$$

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. 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.

2.1. 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

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \ , \tag{16}$$

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". Recall 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 $1/\sigma$ (note 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 when we pick a value from a Gaussian distribution, the value 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.

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 continuous 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 $P(x)$. What we mean by that is that we want to pick x such that the probability that we select a value between x and $x + dx$ equals $P(x)dx$ (said another way, we *draw* x from the distribution $P(x)$). We 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.

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 counts are independent and random, then the actual distribution of the number of counts doesn't look very Gaussian. 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; that way is Bayesian statistical analysis. I strongly encourage you to learn about the principles of Bayesian analysis, which will serve you well as you analyze astronomical data sets.

2.2. Some Statistical Sins

Ignoring systematics.—There's a saying that in astronomy 3σ happens half the time. That's a little tongue-in-cheek, but the reason this is said (when really 3σ should happen 0.3% of the time; here "3σ" refers to three standard deviations beyond the mean in a Gaussian) is that it is very rare indeed that we understand our instruments and contaminating effects perfectly. Maybe that bump in the light curve of your source was a flare, but maybe it was just a cosmic ray that hit your detector. Maybe the detector had a nonlinear response to some photons, or perhaps its calibration isn't perfectly understood. There are also cases in which contaminating sources can intervene. For example, in 1989 a remarkable discovery was announced: an active galaxy had a clear periodic signal in its X-ray emission, with a period of $\sim$12,100 seconds. Revolutionary! But it turned out to be an accreting white dwarf along the line of sight. Not so revolutionary. Think twice before you rewrite physics...

Not estimating "trials" correctly.—In an otherwise featureless spectrum you see an intriguing bump, which you excitedly calculate to have a statistical probability of 10^{-3}. Wow! Have you just discovered unobtanium? Maybe, but did you take into account that you have 1000 spectral bins and thus that there were 1000 chances to have a bump that is improbable at the 10^{-3} level? Many times people will not account correctly for the number of "trials" they perform, and thus they overestimate the significance of the effect. This can be insidious, in the sense that it may not be obvious how many trials are being performed. For example, in the last several years it has become popular to see planar structures in the distribution of satellite galaxies. One well-publicized result notes that 15 out of 29 satellites of Andromeda are in a plane with a thickness of about 10 kpc... and 13 of the 15 are orbiting in the same direction. Amazing! But you'd be equally amazed if the structure made an "S" shape or something like that. Here the possible flaw is that the sequence is (1) see something that looks interesting, then (2) calculate the probability that exactly that thing should happen. This is *a posteriori* statistics, and is well-represented by Feynman's "license-plate fallacy": isn't it remarkable that yesterday the car parked next to mine had a license plate that read HSX 495? That exact license plate, out of all possibilities!

Null hypothesis testing.—This is a little tricky, and it has some relation to the issue of trials. In an introductory statistics class you are often told that this is *the* way to test a hypothesis. That is, you have a model, and you determine how likely it is that you would see some data if your model is correct. For instance, your model might be that a signal is constant, and you use some statistical approach to determine whether the data are consistent with your model. If you judge the model to be inconsistent with the data at some significance level, then you reject the model at that significance level. This may sound reasonable, but the reason it is tricky is that this approach compares, in a nebulous way, a specific model with *all other models combined.* It could easily be that no other specific model does better than your model, which would mean that you were incorrect to reject your model.

For example, let's say that my null hypothesis is that I have a fair coin, which will give heads and tails with equal probability. I flip it two million times (evidently I don't have anything else to do...) and get exactly one million heads. But then I discover that the probability of getting exactly one million heads in two million flips, given a fair coin, is about 10^{-3}. I therefore reject my null hypothesis that the coin is fair. Clearly, that would be the wrong conclusion; in fact, no other probability of heads does better, so no other specific model does better.

That's why in Bayesian statistics (which I recommend looking at very closely) there is an insistence on doing model comparison between *precisely specified* models. That being said, here is a respect in which I differ somewhat from Bayesian orthodoxy; I think it's not a bad idea to have *some* way to determine whether the model you're considering is an adequate fit to the data, in an absolute sense.

Thinking that you need to bin.—It is very common in statistical analyses to assume a Gaussian distribution for some quantity. Many tools require that assumption (e.g., this underlies the calculation of χ^2, if you've heard of that test). But people usually understand that when one has a small number of points, the distribution will typically *not* be Gaussian. So they take their data and group it so that there are larger number of data points per group, and thus so that the statistics are closer to Gaussian. I have, incredibly, had Ph.D. scientists tell me that this *improves* the precision of the resulting statistical inference. No, no, no! By grouping data you lose track of where in the group the data originated, so you are guaranteed to lose information. Now, it could be that the information you lose is of negligible importance, or that it is computationally infeasible to use all the data in their original form, but if you are somehow forced to bin you should do so with eyes open.

Confirmation bias and the elimination of "outliers".—It's easy to want certain results from an analysis. But because we do know that glitches occur, sometimes an observation or a point in that observation might not really be representative of the source. As a result, we can be tempted to try to identify those "outliers" and eliminate them, to get "clean" data. But beware! This leads to an astro-statistics version of confirmation bias, by which we reinforce our prejudices when we see something we like, and dismiss evidence that contradicts our prior conclusions.

Subtracting a background rather than modeling it.—Suppose that you're looking for an excess above a background; maybe there is some overall sky glow, and you're looking for evidence of a dim high-redshift galaxy. Or, maybe a source has some constant level of emission and you're interested in whether it has flared in a particular time. A common and incorrect procedure is to *subtract* the constant level from the emission when either assessing the case for the existence of the source or flare, or determining the parameters and their uncertainties for the phenomenon. Indeed, at least until recently this was automatic in the analysis package XSPEC, which is standard in the X-ray community. Why is this wrong? Because fluctuations in the data depend on the *total* intensity or number of counts. Suppose, for example, that we're in the Gaussian regime, where if the average number of counts in some interval is N, we'd expect $N \pm \sqrt{N}$ in a particular observation. Then if (for instance) we use χ^2 statistics, $\sqrt{N}$ is what we use for the standard deviation of the data. If the background has 99% of the counts, then if we subtract the background then we erroneously conclude that the fluctuation level (and the standard deviation we use in our χ^2 analysis) is $\sqrt{0.01N}$, or only 1/10 of the correct value. The right procedure is to include a model of the background as part of the overall modeling of your data.

Using a black box code.—We have finite time and thus we naturally focus our personal resources on a limited set of things. But when we do statistical analyses, this can come back to bite us. Someone points us to a particular statistical package, which is used for our type of analysis. Yay! These can save us a lot of effort; for example, who wants to spend a huge time writing their own code from scratch to interpret data from a particular instrument? But the analysis performed by the package will usually make certain assumptions, and those won't always be valid. The XSPEC example above is a case in point: if you just stuff your data into the code and ask for an answer, it does things (like background subtraction) that are actually wrong, and you'd never know. It is your responsibility to determine the assumptions used in any package you employ, and to understand the consequences of those assumptions.

Not thinking about whether your answers make sense.—Try actually *looking* at your data! Do the conclusions you drew from your analysis pass the gut check test? If not, think again. For example, it can easily be that you do an analysis, estimate parameters, and end up with some clear conclusions, but actually your model doesn't fit the data. Or, you can do something in a formally right way that leads to an answer that is actually absurd. As an example, many years ago I saw a paper in which the authors computed a correlation coefficient between two quantities, call them A and B. They concluded that the two are stunningly well-correlated; the coefficient was 0.9997! But they had a graph of the quantities, and it was pretty much a scatter plot. What's going on??? It turns out that they had a log-log plot, and because there was one very high point and one very low point, a straight line fit beautifully (basically, it's a line between two points). But they didn't comment on it. Remember, you are the master of the statistics; statistics shouldn't boss you around!

Practice problems

Because probability and statistics might not be that familiar, here we will proceed as we did with vectors: lots of practice problems! Of course you can find many more if you search. Good luck!

1. Calculate, using enumeration, how probable it is that in three flips of a fair coin you will get exactly two heads.

Answer: The possible outcomes, with two-head cases in bold, are HHH, **HHT**, **HTH**, HTT, **THH**, THT, TTH, and TTT. There are therefore three outcomes, out of eight total, have two heads. Thus the probability is 3/8.

2. Calculate, using enumeration, how probable it is that in four flips of a fair coin you will get exactly one head.

3. A strangely-shaped object has two possible outcomes if rolled: "A" and "B". A comes up with probability 0.3, and B comes up with probability 0.7. If you roll the object 5 times, how probable is it that there will be exactly 4 "B"s?

Answer: Using the binomial theorem, with $a = 0.3$ and $b = 0.7$, we see that the probability of 1 "A" and 4 "B"s in 5 rolls is $\binom{5}{1}ab^4 = 5(0.3)(0.7)^4 = 0.36015$.

4. Another strangely-shaped object also has as its only possible rolling outcomes "A" and "B", but this time A comes up with probability 0.8 and B comes up with probability 0.2 If you roll the object 10 times, how probable is it that A shows up exactly 7 times?

5. For a third strangely-shaped object, A comes up with probability 0.6 and B comes up with probability 0.4. If you roll the object 20 times, how probable is it that A shows up *at least* 18 times?

6. Suppose that some quantity x can be anywhere between 0 and 1, but not outside that range. The probability density for x is $P(x) \propto x$. What is the probability that a measurement will find $0 \leq x \leq 1/2$?

Answer: First we need to find the normalized probability density. We know that the integral over all possibilities must be 1, so if the proportionality constant is C, so that $P(x) = Cx$, it must be that

$$1 = \int_0^1 P(x)dx = \int_0^1 Cxdx = C[x^2/2]|_0^1 = C[1/2 - 0] = C/2 \; . \tag{17}$$

Therefore, $C = 2$ and the normalized probability density is $P(x) = 2x$.

Now we can answer our question. The probability that x is between 0 and 1/2 is

$$p(0 \leq x \leq 1/2) = \int_0^{1/2} P(x)dx = \int_0^{1/2} 2xdx = x^2|_0^{1/2} = (1/2)^2 - 0 = 1/4 \; . \tag{18}$$

7. For this problem x can still be anywhere between 0 and 1, but not outside that range. However, the probability density for x is now $P(x) \propto x^2$. What is the probability that a measurement will find $1/2 \leq x \leq 1$?

8. Suppose that x can be anywhere between 1 and 2, but not outside that range. As before, $P(x) \propto x^2$. What is the probability that a measurement will find $1 \leq x \leq 3/2$?

9. Calculate the median, mode, and arithmetic mean of the following set of numbers: 1,1,1,2,2,3,4,5,6,7,8.

Answer: the mode is the most common number, which is 1 in this case. The median is the value such that half are above and half are below; that's 3 for us. The arithmetic mean is $(1 + 1 + 1 + 2 + 2 + 3 + 4 + 5 + 6 + 7 + 8)/11 = 3.64$.

10. Calculate the median, mode, and arithmetic mean of the following set of numbers: 1,2,2,3,3,3,4,4,4,4,5.

11. Calculate the variance of the set of numbers 1,1,1,2,2,3,4,5,6,7,8.

Answer: recall that the variance is $\langle x^2 \rangle - \langle x \rangle^2$, where $\langle x \rangle$ is the arithmetic mean and $\langle x^2 \rangle$ is the sum of the squares of the values divided by the number of values (for a discrete distribution) or $\int x^2 P(x) dx$ (for a continuous distribution). Our case is a discrete distribution, so $\langle x^2 \rangle = (1^2 + 1^2 + 1^2 + 2^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 6^2 + 8^2)/11 \approx 19.1$. We found earlier that $\langle x \rangle = 3.64$ for this distribution, so the variance is $19.1 - 3.64^2 \approx 5.85$.

12. Calculate the standard deviation σ for the same set of numbers.

Answer: the standard deviation is the square root of the variance, so $\sigma = (5.85)^{1/2} \approx 2.42$.

13. Calculate the variance and standard deviation for the set of numbers 1,2,2,3,3,3,4,4,4,4,5.

14. If you have some experience with integration, show that the standard deviation of the Gaussian distribution (Equation 16) is indeed the σ used in the exponential.