Chapter 4

Distributions

From *Probability, For the Enthusiastic Beginner* (Draft version, March 2016)
David Morin, morin@physics.harvard.edu

At the beginning of Section 3.1, we introduced the concepts of *random variables* and *probability distributions*. A random variable is a variable that can take on certain numerical values with certain probabilities. The collection of these probabilities is called the *probability distribution* for the random variable. A probability distribution specifies how the total probability (which is always 1) is distributed among the various possible outcomes.

In this chapter, we will discuss probability distributions in detail. In Section 4.1 we warm up with some examples of discrete distributions, and then in Section 4.2 we discuss continuous distributions. These involve the *probability density*, which is the main new concept in this chapter. It takes some getting used to, but we'll have plenty of practice with it. In Sections 4.3–4.8 we derive and discuss a number of the more common and important distributions. They are, respectively, the uniform, Bernoulli, binomial, exponential, Poisson, and Gaussian (or normal) distributions.

Parts of this chapter are a bit mathematical, but there's no way around this if we want to do things properly. However, we've relegated some of the more technical issues to Appendices B and C. If you want to skip those and just accept the results that we derive there, that's fine. But you are strongly encouraged to at least take a look at Appendix B, where we derive many properties of the number e, which is the most important number in probability and statistics.

4.1 Discrete distributions

In this section we'll give a few simple examples of discrete distributions. To start off, consider the results from Example 3 in Section 2.3.4, where we calculated the probabilities of obtaining the various possible numbers of Heads in five coin flips. We found:

$$\begin{aligned} P(0) &= \frac{1}{32}, \quad P(1) = \frac{5}{32}, \quad P(2) = \frac{10}{32}, \\ P(3) &= \frac{10}{32}, \quad P(4) = \frac{5}{32}, \quad P(5) = \frac{1}{32}. \end{aligned} \tag{4.1}$$

These probabilities add up to 1, as they should. Fig. 4.1 shows a plot of $P(n)$ versus n. The random variable here is the number of Heads, and it can take on the values of 0 through 5, with the above probabilities.

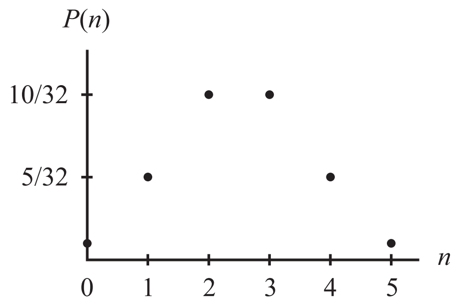


Figure 4.1: The probability distribution for the number of Heads in five coin flips.

As we've done in Fig. 4.1, the convention is to plot the random variable on the horizontal axis and the probability on the vertical axis. The collective information, given either visually in Fig. 4.1 or explicitly in Eq. (4.1), is the probability distribution. A probability distribution simply tells you what all the probabilities are for the values that the random variable can take. Note that $P(n)$ in the present example is nonzero only if n takes on one of the *discrete* values, 0, 1, 2, 3, 4, or 5. It's a silly question to ask for the probability of getting 4.27 Heads, because n must of course be an integer. The probability of getting 4.27 Heads is trivially zero. Hence the word "discrete" in the title of this section.

Another simple example of a discrete probability distribution is the one for the six possible outcomes of the roll of one die. The random variable in this setup is the number on the top face of the die. If the die is fair, then all six numbers have equal probabilities, so the probability for each is $1/6$, as shown in Fig. 4.2.

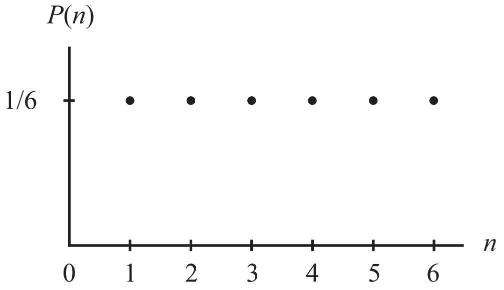


Figure 4.2: The probability distribution for the roll of one die.

What if the die isn't fair? For example, what if we make the "1" face heavier than the others by embedding a small piece of lead in the center of that face, just below the surface? The die is then more likely to land with the "1" face pointing down. The "6" face is opposite the "1," so the die is more likely to land with the "6" pointing up. Fig. 4.2 will therefore be modified by raising the "6" dot and lowering the other five dots; the sum of the probabilities must still be 1, of course. P_2 through P_5 are all equal, by symmetry. The exact values of all the probabilities depend in a complicated way on how the mass of the lead weight compares with the mass of the die, and also on the nature of both the die and the table on which the die is rolled (how much friction, how bouncy, etc.).

As mentioned at the beginning of Section 3.1, a random variable is assumed to take on *numerical* values, by definition. So the outcomes of Heads and Tails for a single coin flip technically aren't random variables. But it still makes sense to plot the probabilities as shown in Fig. 4.3, even though the outcomes on the horizontal axis aren't associated with a random variable. Of course, if we define a random variable to be the number of Heads, then the "Heads" in the figure turns into a 1, and the "Tails" turns into a 0. In most situations, however, the outcomes take on numerical values right from the start, so we can officially label them as random variables. But even if they don't, we'll often take the liberty of still referring to the thing being plotted on the horizontal axis of a probability distribution as a random variable.

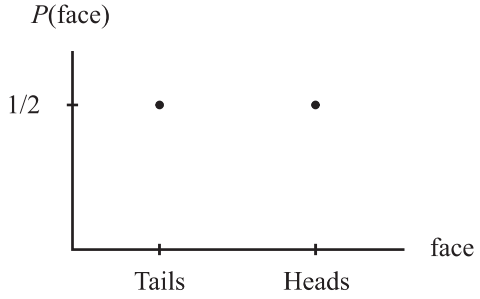


Figure 4.3: The probability distribution for a single coin flip.

4.2 Continuous distributions

4.2.1 Motivation

Probability distributions are fairly straightforward when the random variable is discrete. You just list (or plot) the probabilities for each of the possible values of the random variable. These probabilities will always add up to 1. However, not everything comes in discrete quantities. For example, the temperature outside your house takes on a continuous set of values, as does the amount of water in a glass. (We'll ignore the atomic nature of matter!)

In finding the probability distribution for a continuous random variable, you might think that the procedure should be exactly the same as in the discrete case. That is, if our random variable is the temperature at a particular location at noon tomorrow, then you might think that you simply have to answer questions of the form: What is the probability that the temperature at noon tomorrow will be 70° Fahrenheit?

Unfortunately, there is something wrong with this question, because it is too easy to answer. The answer is that the probability is *zero*, because there is simply no chance that the temperature at a specific time (and a specific location) will be *exactly* 70°. If it's 70.1°, that's not good enough. And neither is 70.01°, nor even 70.00000001°. Basically, since the temperature takes on a continuous set of values (and hence an infinite number of possible values), the probability of a specific value occurring is $1/\infty$, which is zero.[1]

However, even though the above question ("What is the probability that the temperature at noon tomorrow will be 70°?") is a poor one, that doesn't mean we should throw in the towel and conclude that probability distributions don't exist for continuous random variables. They do in fact exist, because there *are* some useful questions we can ask. These useful questions take the general form of: What is the probability that the temperature at a particular location at noon tomorrow lies somewhere between 69° and 71°? This question has a nontrivial answer, in the sense that it isn't automatically zero. And depending on what the forecast is for tomorrow, the answer might be something like 20%.

We can also ask: What is the probability that the temperature at noon lies somewhere between 69.5° and 70.5°? The answer to this question is smaller than the answer to the previous one, because it involves a range of only one degree instead of two degrees. If we assume that inside the range of 69° to 71° the temperature is equally likely to be found anywhere (which is a reasonable approximation although undoubtedly not exactly correct), and if the previous answer was 20%, then the present answer is (roughly) 10%, because the range is half the size.

The point here is that the smaller the range, the smaller the chance that the temperature lies in that range. Conversely, the larger the range, the larger the chance that the temperature lies in that range. Taken to an extreme, if we ask for the probability that the temperature at noon lies somewhere between −100° and 200°, then the answer is exactly equal to 1 (ignoring liquid nitrogen spills, forest fires, and such things!).

In addition to depending on the size of the range, the probability also of course depends on where the range is located on the temperature scale. For example, the probability that the temperature at noon lies somewhere between 69° and 71° is undoubtedly different from the probability that it lies somewhere between 11° and 13°. Both ranges have a span of two degrees, but if the given day happens to be in late summer, the temperature is much more likely to be around 70° than to be sub-freezing (let's assume we're in, say, Boston). To actually figure out the probabilities, many different pieces of data would have to be considered. In the present temperature example, the data would be of the meteorological type. But if we were interested in the probability that a random person is between 69 and 71 inches tall, then we'd need to consider a whole different set of data.

The lesson to take away from all this is that if we're looking at a random variable that can take on a continuous set of values, the probability that this random variable falls into a given range depends on three things. It depends on:

[1] Of course, if you're using a digital thermometer that measures the temperature to the nearest tenth of a degree, then it *does* make sense to ask for the probability that the thermometer reads, say, 70.0 degrees. This probability is generally nonzero. This is due to the fact that the *reading* on the digital thermometer is a *discrete* random variable, whereas the *actual temperature* is a *continuous* random variable.

1. the location of the range,
2. the size of the range,
3. the specifics of the situation we're dealing with.

The third of these is what determines the *probability density*, which is a function whose argument is the location of the range. We'll now discuss probability densities.

4.2.2 Probability density

Consider the plot in Fig. 4.4, which gives a hypothetical probability distribution for the temperature example we've been discussing. This plot shows the probability distribution on the vertical axis, as a function of the temperature T (the random variable) on the horizontal axis. We have chosen to measure the temperature in Fahrenheit. We're denoting the probability distribution by[2] $\rho(T)$ instead of $P(T)$, to distinguish it from the type of probability distribution we've been talking about for discrete variables. The reason for this new notation is that $\rho(T)$ is a probability *density* and not an actual probability. We'll talk about this below. When writing the functional form of a probability distribution, we'll denote probability *densities* with lowercase letters, like the ρ in $\rho(T)$ or the f in $f(x)$. And we'll denote actual *probabilities* with uppercase letters, like the P in $P(n)$.

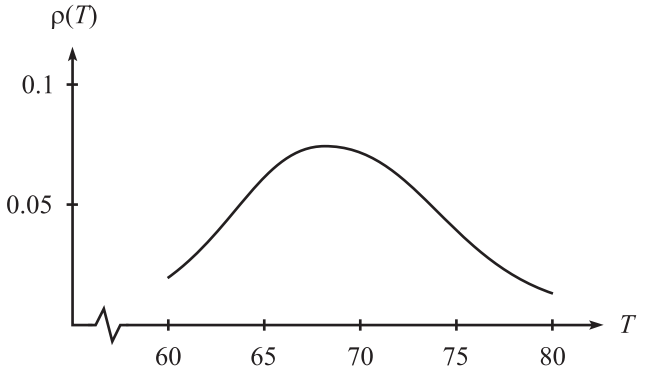


Figure 4.4: A hypothetical probability distribution for the temperature.

We haven't yet said exactly what we mean by $\rho(T)$. But in any case, it's clear from Fig. 4.4 that the temperature is more likely to be near 70° than near 60°. The following definition of $\rho(T)$ allows us to be precise about what we mean by this.

[2] As mentioned at the beginning of Section 3.1, a random variable is usually denoted with an uppercase letter, while the actual values are denoted with lowercase letters. So we should technically be writing $\rho(t)$ here. But since an uppercase T is the accepted notation for temperature, we'll use T for the actual value.

- Definition of the probability density function, $\rho(T)$:

  $\rho(T)$ is the function of T that, when multiplied by a small interval ΔT, gives the probability that the temperature lies between T and $T + \Delta T$. That is,

$$P(\text{temp lies between } T \text{ and } T + \Delta T) = \rho(T) \cdot \Delta T. \qquad (4.2)$$

Note that the lefthand side contains an actual probability P, whereas the righthand side contains a probability *density*, $\rho(T)$. The latter needs to be multiplied by a range of T (or whatever quantity we're dealing with) in order to obtain an actual probability. The above definition is relevant to any continuous random variable, of course, not just temperature.

Eq. (4.2) might look a little scary, but a few examples should clear things up. From Fig. 4.4, it looks like $\rho(70^\circ)$ is about 0.07. So if we pick $\Delta T = 1^\circ$, we find that the probability of the temperature lying between 70° and 71° is about

$$\rho(T) \cdot \Delta T = (0.07)(1) = 0.07 = 7\%. \qquad (4.3)$$

If we instead pick a smaller ΔT, say 0.5°, we find that the probability of the temperature lying between 70° and 70.5° is about $(0.07)(0.5) = 3.5\%$. And if we pick an even smaller ΔT, say 0.1°, we find that the probability of the temperature lying between 70° and 70.1° is about $(0.07)(0.1) = 0.7\%$.

Similarly, we can apply Eq. (4.2) to any other value of T. For example, it looks like $\rho(60^\circ)$ is about 0.02. So if we pick $\Delta T = 1^\circ$, we find that the probability of the temperature lying between 60° and 61° is about $(0.02)(1) = 2\%$. And as above, we can pick other values of ΔT too.

Note that, in accordance with Eq. (4.2), we have been using the value of ρ at the *lower* end of the given temperature interval. That is, when the interval was 70° to 71°, we used $\rho(70^\circ)$ and then multiplied this by ΔT. But couldn't we just as well use the value of ρ at the *upper* end of the interval? That is, couldn't the righthand side of Eq. (4.2) just as well be $\rho(T + \Delta T) \cdot \Delta T$? Indeed it could. But as long as ΔT is small, it doesn't matter much which value of ρ we use. They will both give essentially the same answer. See the second remark below.

Remember that *three* inputs are necessary when finding the probability that the temperature lies in a specified range. As we noted at the end of Section 4.2.1, the first input is the value of T we're concerned with, the second is the range ΔT, and the third is the information encapsulated in the probability density function, $\rho(T)$, evaluated at the given value of T. The latter two of these three quantities are the two quantities that are multiplied together on the righthand side of Eq. (4.2). Knowing only one of these isn't enough to give you a probability.

To recap, there is a very important difference between the probability distribution for a continuous random variable and that for a discrete random variable. For a continuous variable, the probability distribution consists of a *probability density.* But for a discrete variable, it consists of *actual probabilities.* We plot a *density* for a continuous distribution, because it wouldn't make sense to plot actual probabilities, since they're all zero. This is true because the probability of obtaining *exactly* a particular value is zero, since there is an infinite number of possible values.

Conversely, we plot *actual probabilities* for a discrete distribution, because it wouldn't make sense to plot a density, since it consists of a collection of infinite spikes. This is true because on a die roll, for example, there is a 1/6 chance of obtaining a number between, say, 4.9999999 and 5.0000001. The probability density at the outcome of 5, which from Eq. (4.2) equals the probability divided by the interval length, is then (1/6)/(0.0000002), which is huge. And the interval can be made arbitrarily small, which means that the density is arbitrarily large. To sum up, the term "probability distribution" applies to both continuous and discrete variables, whereas the term "probability density" applies only to continuous variables.

Remarks:

1. $\rho(T)$ is a function of T, so it depends on what units we're using to measure T. We used Fahrenheit above, but what if we instead want to use Celsius? Problem 4.1 addresses this issue (but you will need to read Section 4.2.3 first).

2. Note the inclusion of the word "small" in the definition of the probability density in Eq. (4.2). The reason for this word is that we want $\rho(T)$ to be (roughly) constant over the specified range. If ΔT is small enough, then this is approximately true. If $\rho(T)$ varied greatly over the range of ΔT, then it wouldn't be clear which value of $\rho(T)$ we should multiply by ΔT to obtain the probability. The point is that if ΔT is small enough, then all of the $\rho(T)$ values are roughly the same, so it doesn't matter which one we pick.

   An alternative definition of the density $\rho(T)$ is

   $$P\big(\text{temp lies between } T - (\Delta T)/2 \text{ and } T + (\Delta T)/2\big) = \rho(T) \cdot \Delta T. \qquad (4.4)$$

   The only difference between this definition and the one in Eq. (4.2) is that we're now using the value of $\rho(T)$ at the midpoint of the temperature range, instead of the left-end value we used in Eq. (4.2). Both definitions are equally valid, because they give essentially the same result for $\rho(T)$, provided that ΔT is small. Similarly, we could use the value of $\rho(T)$ at the right end of the temperature range.

   How small do we need ΔT to be? The answer to this will be evident when we talk about probability in terms of area in Section 4.2.3. In short, we need the change in $\rho(T)$ over the span of ΔT to be small compared with the values of $\rho(T)$ in that span.

3. The probability density function involves only (1) the value of T (or whatever) we're concerned with, and (2) the specifics of the situation at hand (meteorological data in the above temperature example, etc.). The density is completely independent of the arbitrary value of ΔT that we choose. This is how things work with any kind of density.

   For example, consider the mass density of gold. This mass density is a property of the gold itself. More precisely, it is a function of each point in the gold. For pure gold, the density is constant throughout the volume, but we could imagine impurities that would make the mass density be a varying function of position, just as the above probability density is a varying function of temperature. Let's call the mass density $\rho(\mathbf{r})$, where $\mathbf{r}$ signifies the possible dependence of ρ on the location of a given point within the volume. (The position of a given point can be described by the vector pointing from the origin to the point. And vectors are generally denoted by boldface letters like $\mathbf{r}$.) Let's call the small volume we're concerned with ΔV. Then the mass in the small volume ΔV is given by the product of the density and the volume, that is, $\rho(\mathbf{r}) \cdot \Delta V$. This is directly analogous to the fact that the probability in the above temperature example is given by the product of the probability density and the temperature span, that is, $\rho(T) \cdot \Delta T$. The correspondence among the various quantities is

$$\begin{array}{rcl} \text{Mass in } \Delta V \text{ around location } \mathbf{r} & \Longleftrightarrow & \text{Prob that temp lies in } \Delta T \text{ around } T \\ \rho(\mathbf{r}) & \Longleftrightarrow & \rho(T) \\ \Delta V & \Longleftrightarrow & \Delta T. \quad \clubsuit \end{array} \tag{4.5}$$

4.2.3 Probability equals area

The graphical interpretation of the product $\rho(T) \cdot \Delta T$ in Eq. (4.2) is that it is the area of the rectangle shown in Fig. 4.5. This is true because ΔT is the base of the rectangle, and $\rho(T)$ is the height.

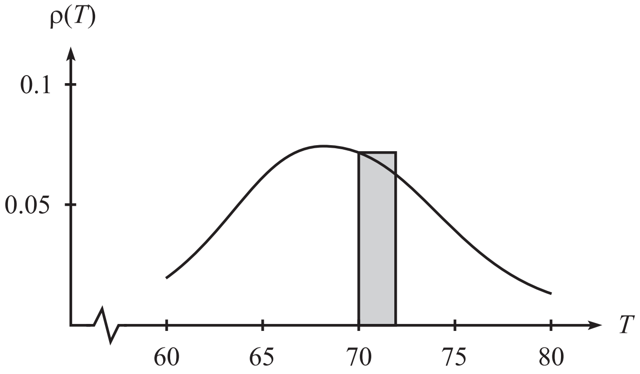


Figure 4.5: Interpretation of the product $\rho(T) \cdot \Delta T$ as an area.

We have chosen ΔT to be 2° in the figure. With this choice, the area of the rectangle, which equals $\rho(70°)\cdot(2°)$, gives a reasonably good approximation to the probability that the temperature lies between 70° and 72°. But it isn't exact, because $\rho(T)$ isn't constant over the 2° interval. A better approximation to the probability that the temperature lies between 70° and 72° is achieved by splitting the 2° interval into two intervals of 1° each, and then adding up the probabilities of lying in each of these two intervals. These two probabilities are approximately equal to $\rho(70°)\cdot(1°)$ and $\rho(71°) \cdot (1°)$, and the two corresponding rectangles are shown in Fig. 4.6.

But again, the sum of the areas of these two rectangles is still only an approximate result for the true probability that the temperature lies between 70° and 72°, because $\rho(T)$ isn't constant over the 1° intervals either. A better approximation is achieved by splitting the 1° intervals into smaller intervals, and then again into even smaller ones. And so on. When we get to the point of having 100 or 1000 extremely thin rectangles, the sum of their areas will essentially be the area shown in Fig. 4.7. This area is the correct probability that the temperature lies between 70° and 72°. So in retrospect, we see that the rectangular area in Fig. 4.5 exceeds the true probability by the area of the tiny triangular-ish region in the upper righthand corner of the rectangle.

We therefore arrive at a more precise definition (compared with Eq. (4.2)) of the probability density, $\rho(T)$:

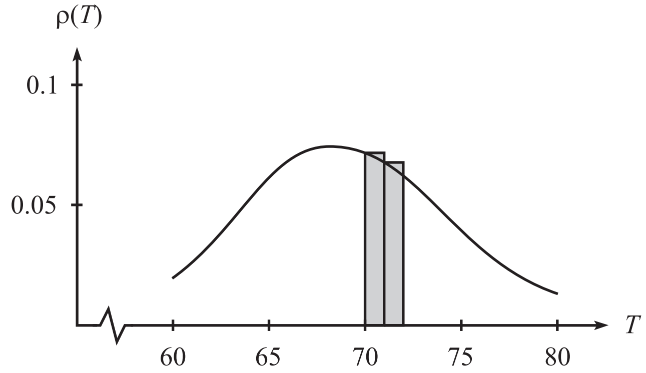


Figure 4.6: Subdividing the area, to produce a better approximation to the probability.

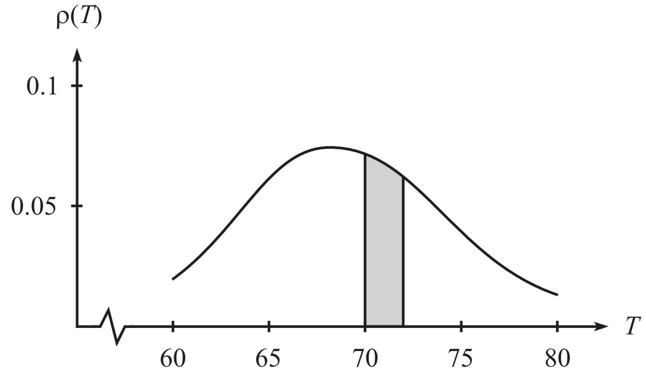


Figure 4.7: The area below the curve between 70° and 72° equals the probability that the temperature lies between 70° and 72°.

- Improved definition of the probability density function, $\rho(T)$:

  $\rho(T)$ is the function of T for which the area under the $\rho(T)$ curve between T and $T + \Delta T$ gives the probability that the temperature (or whatever quantity we're dealing with) lies between T and $T + \Delta T$.

This is an exact definition, and there is no need for ΔT to be small, as there was in the definition in Eq. (4.2). The difference is that the present definition involves the exact area, whereas Eq. (4.2) involved the area of a rectangle (via simple multiplication by ΔT), which was only an approximation. But technically the only thing we need to add to Eq. (4.2) is the requirement that we take the $\Delta T \to 0$ limit. That makes the definition rigorous.

The total area under any probability density curve must be 1, because this area equals the probability that the temperature (or whatever) takes on some value between $-\infty$ and $+\infty$, and because every possible result is included in the $-\infty$ to $+\infty$ range. However, in any realistic case, the density is essentially zero outside a specific finite region. So there is essentially no contribution to the area from the parts of the plot outside that region. There is therefore no need to go to ±∞. The total area under each of the curves in the above figures, including the tails on either side which we haven't bothered to draw, is indeed equal to 1 (at least roughly; the curves were drawn by hand).

Given a probability density function $f(x)$, the *cumulative distribution function* $F(x)$ is defined to be the probability that X takes on a value that is less than or equal to x. That is, $F(x) = P(X \leq x)$. For a continuous distribution, this definition implies that $F(x)$ equals the area under the $f(x)$ curve from $-\infty$ up to the given x value. A quick corollary is that the probability $P(a < x \leq b)$ that x lies between two given values a and b is equal to $F(b) - F(a)$. For a discrete distribution, the definition $F(x) = P(X \leq x)$ still applies, but we now calculate $P(X \leq x)$ by forming a discrete sum instead of finding an area. Although the cumulative distribution function can be very useful in probability and statistics, we won't use it much in this book.

We'll now spend a fair amount of time in Sections 4.3–4.8 discussing some common types of probability distributions. There is technically an infinite number of possible distributions, although only a hundred or so come up frequently enough to have names. And even many of these are rather obscure. A handful, however, come up again and again in a variety of settings, so we'll concentrate on these. They are the uniform, Bernoulli, binomial, exponential, Poisson, and Gaussian (or normal) distributions.

4.3 Uniform distribution

We'll start with a very simple continuous probability distribution, one that is uniform over a given interval, and zero otherwise. Such a distribution might look like the one shown in Fig. 4.8. If the distribution extends from x_1 to x_2, then the value of $\rho(x)$ in that region must be $1/(x_2 - x_1)$, so that the total area is 1.

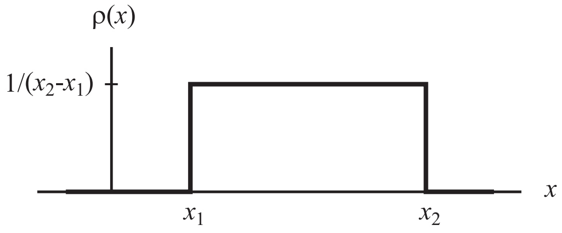


Figure 4.8: A uniform distribution.

This type of distribution could arise, for example, from a setup where a rubber ball bounces around in an empty rectangular room. When it finally comes to rest, we measure its distance x from a particular one of the walls. If you initially throw the ball hard enough, then it's a pretty good approximation to say that x is equally likely to take on any value between 0 and L, where L is the length of the room in the relevant direction. In this setup, the x_1 in Fig. 4.8 equals 0 (so we would need to shift the rectangle to the left), and the x_2 equals L.

The random variable here is X, and the value it takes is denoted by x. So x is what we plot on the horizontal axis. Since we're dealing with a continuous distribution, we plot the probability *density* (not the probability!) on the vertical axis. If L equals 10 feet, then outside the region $0 < x < 10$, the probability density $\rho(x)$ equals zero. Inside this region, the density equals the total probability divided by the total interval, which gives 1 per 10 feet, or equivalently 1/10 per foot. If we want to find the actual probability that the ball ends up between, say, $x = 6$ and $x = 8$, then we just multiply $\rho(x)$ by the interval length, which is 2 feet. The result is (1/10 per foot)(2 feet), which equals $2/10 = 1/5$. This makes sense, of course, because the 2-foot interval is 1/5 of the total distance.

A uniform density is easy to deal with, because the area under a given part of the curve (which equals the probability) is simply a rectangle. And the area of a rectangle is just the base times the height, which is the interval length times the density. This is exactly the product we formed above. When the density isn't uniform, it can be very difficult sometimes to find the area under a given part of the curve.

Note that the larger the region of nonzero $\rho(x)$ in a uniform distribution, the smaller the value of $\rho(x)$. This follows from the fact that the total area under the density "curve" (which is just a straight line segment in this case) must equal 1. So if the base becomes longer, the height must become shorter.

4.4 Bernoulli distribution

We'll now consider a very simple discrete distribution, called the Bernoulli distribution. This is the distribution for a process in which only two possible outcomes, 1 and 0, can occur, with probabilities p and $1 - p$, respectively. (They must add up to 1, of course.) The plot of this probability distribution is shown in Fig. 4.9. It is common to call the outcome of 1 a success and the outcome of 0 a failure. A special case of a Bernoulli distribution is the distribution for a coin toss, where the probabilities for Heads and Tails (which we can assign the values of 1 and 0, respectively) are both equal to 1/2.

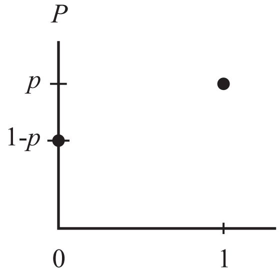


Figure 4.9: A Bernoulli distribution takes on the values 1 and 0 with probabilities p and $1 - p$.

The Bernoulli distribution is the simplest of all distributions, with the exception of the trivial case where only one possible outcome can occur, which therefore has a probability of 1. The uniform and Bernoulli distributions are simple enough that there isn't much to say. In contrast, the distributions in the following four sections (binomial, exponential, Poisson, and Gaussian) are a bit more interesting, so we'll have plenty to say about them.

4.5 Binomial distribution

The binomial distribution, which is discrete, is an extension of the Bernoulli distribution. The binomial distribution is defined to be the probability distribution for the total number of successes that arise in an arbitrary number of independent and identically distributed Bernoulli processes. An example of a binomial distribution is the probability distribution for the number of Heads in, say, five coin tosses, which we discussed in Section 4.1. We could just as well pick any other number of tosses.

In the case of five coin tosses, each coin toss is a Bernoulli process. When we put all five tosses together and look at the total number of successes (Heads), we get a binomial distribution. Let's label the total number of successes as k. In this specific example, there are $n = 5$ Bernoulli processes, with each one having a $p = 1/2$ probability of success. The probability distribution $P(k)$ is simply the one we plotted earlier in Fig. 4.1, where we counted the number of Heads.

Let's now find the binomial distribution associated with a general number n of independent Bernoulli trials, each with the same probability of success, p. So our goal is to find the value of $P(k)$ for all of the different possible values of the total number of successes, k. The possible values of k range from 0 up to the number of trials, n.

To calculate the binomial distribution (for given n and p), we first note that p^k is the probability that a *specific set* of k of the n Bernoulli processes all yield success, because each of the k processes has a p probability of yielding success. We then need the other $n - k$ processes to *not* yield success, because we want *exactly* k successes. This happens with probability $(1 - p)^{n-k}$, because each of the $n - k$ processes has a $1 - p$ probability of yielding failure. The probability that a specific set of k processes (and no others) all yield success is therefore $p^k \cdot (1 - p)^{n-k}$. Finally, since there are $\binom{n}{k}$ ways to pick a specific set of k processes, we see that the probability that exactly k of the n processes yield success is

$$\boxed{P(k) = \binom{n}{k} p^k (1-p)^{n-k}} \qquad \text{(binomial distribution)} \tag{4.6}$$

This is the desired binomial distribution. Note that this distribution depends on two parameters – the number n of Bernoulli trials and the probability p of success in each trial. If you want to make these parameters explicit, you can write the Binomial distribution $P(k)$ as $B_{n,p}(k)$. That is,

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.7}$$

But we'll generally just use the simple $P(k)$ notation.

In the special case of a binomial distribution generated from n coin tosses, we have $p = 1/2$. So Eq. (4.6) gives the probability of obtaining k Heads as

$$P(k) = \frac{1}{2^n}\binom{n}{k}. \tag{4.8}$$

To recap: In Eq. (4.6), n is the total number of Bernoulli processes, p is the probability of success in each Bernoulli process, and k is the total number of successes in the n processes. (So k can be anything from 0 to n.) Fig. 4.10 shows the binomial distribution for the cases of $n = 30$ and $p = 1/2$ (which arises from 30 coin tosses), and $n = 30$ and $p = 1/6$ (which arises from 30 die rolls, with a particular one of the six numbers representing success).

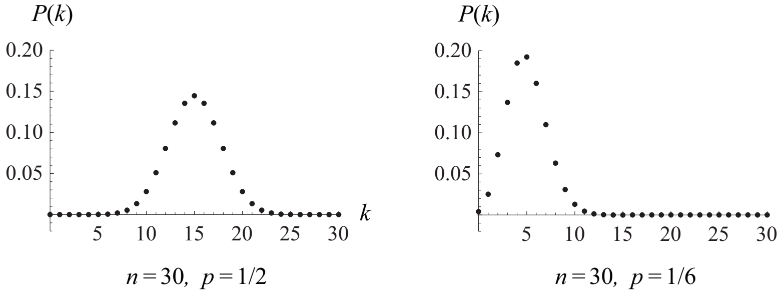


Figure 4.10: Two binomial distributions with $n = 30$ but different values of p.

Example (Equal probabilities): Given n, for what value of p is the probability of zero successes equal to the probability of one success?

Solution: In Eq. (4.6) we want $P(0)$ to equal $P(1)$. This gives

$$\begin{aligned} \binom{n}{0}p^0(1-p)^{n-0} &= \binom{n}{1}p^1(1-p)^{n-1} \\ \Longrightarrow\quad 1\cdot 1\cdot(1-p)^n &= n\cdot p\cdot(1-p)^{n-1} \\ \Longrightarrow\quad 1-p = np \quad&\Longrightarrow\quad p = \frac{1}{n+1}\,. \end{aligned} \tag{4.9}$$

This $p = 1/(n+1)$ value is the special value of p for which various competing effects cancel. On one hand, $P(1)$ contains an extra factor of n from the $\binom{n}{1}$ coefficient, which arises from the fact that there are n different ways for one success to happen. But on the other hand, $P(1)$ also contains a factor of p, which arises from the fact that one success *does* happen. The first of these effects makes $P(1)$ larger than $P(0)$, while the second makes it smaller.[3] The effects cancel when $p = 1/(n+1)$. Fig. 4.11 shows the plot for $n = 10$ and $p = 1/11$.

The $p = 1/(n+1)$ case is the cutoff between the maximum of $P(k)$ occurring when k is zero or nonzero. If p is larger than $1/(n+1)$, as it is in both plots in Fig. 4.10

[3] Another effect is that $P(1)$ is larger because it contains one fewer factor of $(1-p)$. But this effect is minor when p is small, which is the case if n is large, due to the $p = 1/(n+1)$ form of the answer.

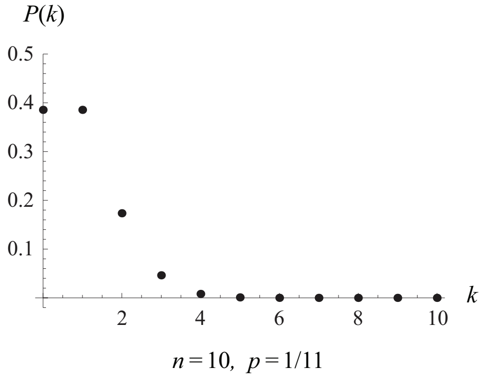


Figure 4.11: $P(0)$ equals $P(1)$ if $p = 1/(n+1)$.

above, then the maximum occurs at a nonzero value of k. That is, the distribution has a bump. On the other hand, if p is smaller than $1/(n+1)$, then the maximum occurs at $k = 0$. That is, the distribution has its peak at $k = 0$ and falls off from there.

Having derived the binomial distribution in Eq. (4.6), there is a simple double check that we can perform on the result. Since the number of successes, k, can take on any integer value from 0 to n, the sum of the $P(k)$ probabilities from $k = 0$ to $k = n$ must equal 1. The $P(k)$ expression in Eq. (4.6) does indeed satisfy this requirement, due to the binomial expansion, which tells us that

$$\big(p + (1-p)\big)^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.10}$$

This is just Eq. (1.21) from Section 1.8.3, with $a = p$ and $b = 1 - p$. The lefthand side of Eq. (4.10) is simply $1^n = 1$. And each term in the sum on the righthand side is a $P(k)$ term from Eq. (4.6). So Eq. (4.10) becomes

$$1 = \sum_{k=0}^{n} P(k), \tag{4.11}$$

as we wanted to show. You are encouraged to verify this result for the probabilities in, say, the left plot in Fig. 4.10. Feel free to make rough estimates of the probabilities when reading them off the plot. You will find that the sum is indeed 1, up to the rough estimates you make.

The task of Problem 4.4 is to use Eq. (3.4) to explicitly demonstrate that the expectation value of the binomial distribution in Eq. (4.6) equals pn. In other words, if our binomial distribution is derived from n Bernoulli trials, each having a probability p of success, then we should expect a total of pn successes (on average, if we do a large number of sets of n trials). This must be true, of course, because a fraction p of the n trials yield success, on average, by the definition of p for the given Bernoulli process.

Remark: We should emphasize what is meant by a probability distribution. Let's say that you want to experimentally verify that the left plot in Fig. 4.10 is the correct probability distribution for the total number of Heads that show up in 30 coin flips. You of course can't do this by flipping a coin just once. And you can't even do it by flipping a coin 30 times, because all you'll get from that is just one number for the total number of Heads. For example, you might obtain 17 Heads. In order to experimentally verify the distribution, you need to perform *a large number of sets of 30 coin flips,* and you need to record the total number of Heads you get in each 30-flip set. The result will be a long string of numbers such as $13, 16, 15, 16, 18, 14, 11, 17, \ldots$. If you then calculate the fractions of the time that each number appears, these fractions should (roughly) agree with the probabilities shown in Fig. 4.10. The longer the string of numbers, the better the agreement, in general. The main point here is that the distribution does't say much about *one particular* set of 30 flips. Rather, it says what the expected distribution of outcomes is for a *large number* of sets of 30 flips. ♣

4.6 Exponential distribution

In Sections 4.6–4.8 we'll look at three probability distributions (exponential, Poisson, and Gaussian) that are a bit more involved than the three we've just discussed (uniform, Bernoulli, and binomial). We'll start with the exponential distribution, which takes the general form,

$$\rho(t) = Ae^{-bt}, \tag{4.12}$$

where A and b are quantities that depend on the specific situation at hand. We will find below in Eq. (4.26) that these quantities must be related in a certain way in order for the total probability to be 1. The parameter t corresponds to whatever the random variable is. The exponential distribution is a continuous one, so $\rho(t)$ is a probability density. The most common type of situation where this distribution arises is the following.

Consider a repeating event that happens completely randomly in time. By "completely randomly" we mean that there is a uniform probability that the event happens at any given instant (or more precisely, in any small time interval of a given length), independent of what has already happened. That is, the process has no "memory." The exponential distribution that we'll eventually arrive at (after a lot of work!) in Eq. (4.26) gives the probability distribution for the *waiting time* until the next event occurs. Since the time t is a continuous quantity, we'll need to develop some formalism to analyze the distribution. To ease into it, let's start with the slightly easier case where time is assumed to be discrete.

4.6.1 Discrete case

Consider a process where we roll a hypothetical 10-sided die once every second. So time is discretized into 1-second intervals. It's actually not necessary to introduce time here at all. We could simply talk about the number of iterations of the process. But it's easier to talk about things like the "waiting time" than the "number of iterations you need to wait for." So for convenience, we'll discuss things in the context of time.

If the die shows a "1," we'll consider that a success. The other nine numbers represent failure. There are two reasonable questions we can ask: What is the average waiting time (that is, the expectation value of the waiting time) between successes? And what is the probability distribution of the waiting times between successes?

Average waiting time

It is fairly easy to determine the average waiting time. There are 10 possible numbers on the die, so on average we can expect 1/10 of them to be 1's. If we run the process for a long time, say, an hour (which consists of 3600 seconds), then we can expect about 360 1's. The average waiting time between successes is therefore $(3600 \text{ seconds})/360 = 10$ seconds.

More generally, if the probability of success in each trial is p, then the average waiting time is $1/p$ (assuming that the trials happen at 1-second intervals). This can be seen by the same reasoning as above. If we perform n trials of the process, then pn of them will yield success, on average. The average waiting time between successes is the total time (n) divided by the number of successes (pn):

$$\text{Average waiting time} = \frac{n}{pn} = \frac{1}{p}. \tag{4.13}$$

Note that the preceding reasoning gives us the average waiting time, without requiring any knowledge of the actual probability distribution of the waiting times (which we will calculate below). Of course, once we *do* know what the probability distribution is, we should be able to calculate the average (the expectation value) of the waiting times. This is the task of Problem 4.7.

Distribution of waiting times

Finding the probability distribution of the waiting times requires a little more work than finding the average waiting time. For the 10-sided die example, the question we're trying to answer is: What is the probability that if we consider two successive 1's, the time between them will be 6 seconds? Or 30 seconds? Or 1 second? And so on. Although the *average* waiting time is 10 seconds, this certainly doesn't mean that the waiting time will always be 10 seconds. In fact, we will find below that the probability that the waiting time is exactly 10 seconds is quite small.

Let's be general and say that the probability of success in each trial is p (so $p = 1/10$ in our present setup). Then the question is: What is the probability, $P(k)$, that we will have to wait exactly k iterations (each of which is 1 second here) to obtain the next success?

To answer this, note that in order for the next success to happen on the kth iteration, there must be failure (which happens with probability $1 - p$) on the first $k - 1$ iterations, and then success on the kth one. The probability of this happening is

$$\boxed{P(k) = (1-p)^{k-1}p} \qquad \text{(geometric distribution)} \tag{4.14}$$

This is the desired (discrete) probability distribution for the waiting time. This distribution goes by the name of the *geometric distribution*, because the probabilities form a geometric progression, due to the increasing power of the $(1-p)$ factor. The geometric distribution is the discrete version of the exponential distribution that we'll arrive at in Eq. (4.26) below.

Eq. (4.14) tells us that the probability that the next success comes on the very next iteration is p, the probability that it comes on the second iteration is $(1 - p)p$, the probability that it comes on the third iteration is $(1 - p)^2 p$, and so on. Each probability is smaller than the previous one by the factor $(1 - p)$. A plot of the distribution for $p = 1/10$ is shown in Fig. 4.12. The distribution is maximum at $k = 1$ and falls off from that value. Even though $k = 10$ is the average waiting time, the probability of the waiting time being *exactly* $k = 10$ is only $P(10) = (0.9)^9(0.1) \approx 0.04 = 4\%$.

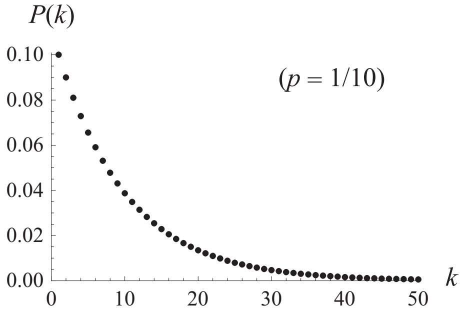


Figure 4.12: The geometric distribution with $p = 1/10$.

If p is large (close to 1), the plot of $P(k)$ starts high (at p, which is close to 1) and then falls off quickly, because the factor $(1-p)$ is close to 0. On the other hand, if p is small (close to 0), the plot of $P(k)$ starts low (at p, which is close to 0) and then falls off slowly, because the factor $(1 - p)$ is close to 1.

As a double check on the result in Eq. (4.14), we know that the next success has to eventually happen *sometime*, so the sum of all the $P(k)$ probabilities must be 1. These $P(k)$ probabilities form a geometric series whose first term is p and whose ratio is $1 - p$. The general formula for the sum of a geometric series with first term a and ratio r is $a/(1 - r)$, so we have

$$\begin{aligned} P(1) + P(2) + P(3) + \cdots &= p + p(1-p) + p(1-p)^2 + \cdots \\ &= \frac{p}{1-(1-p)} \\ &= 1, \end{aligned} \tag{4.15}$$

as desired. As another check, we can verify that the expectation value (the average) of the waiting time for the geometric distribution in Eq. (4.14) equals $1/p$, as we already found above; see Problem 4.7.

You are encouraged to use a coin to experimentally "verify" Eq. (4.14) (or equivalently, the plot analogous to Fig. 4.12) for the case of $p = 1/2$. Just flip a coin as many times as you can in ten minutes, each time writing down a 1 if you get Heads and a 0 if you get Tails. Then make a long list of the waiting times between the 1's. Then count up the number of one-toss waits, the number of two-toss waits, and so on. Then divide each of these numbers by the total number of waits (not the total number of tosses!) to find the probability of each waiting length. The results should be (roughly) consistent with Eq. (4.14) for $p = 1/2$. In this case, the probabilities in Eq. (4.14) for $k = 1, 2, 3, 4, \dots$ are $1/2, 1/4, 1/8, 1/16, \dots$.

4.6.2 Rates, expectation values, and probabilities

Let's now consider the case where time is a continuous quantity. That is, let's assume that we can have a "successful" event at *any* instant, not just at the evenly-spaced 1-second marks as above. A continuous process whose probability is uniform in time can be completely described by just *one* number – the average rate of success, which we'll call λ. We generally won't bother writing the word "average," so we'll just call λ the "rate." Before getting into the derivation of the continuous exponential distribution in Section 4.6.3, we'll need to talk a little about rates.

The rate λ can be determined by counting the number of successful events that occur during a long time interval, and then dividing by this time. For example, if 300 (successful) events happen during 100 minutes, then the rate λ is 3 events per minute. Of course, if you count the number of events in a different span of 100 minutes, you will most likely get a slightly different number, perhaps 313 or 281. But in the limit of a very long time interval, you will find essentially the same rate, independent of which specific long interval you use.

If the rate λ is 3 events per minute, you can alternatively write this as 1 event per 20 seconds, or 1/20 of an event per second. There is an infinite number of ways to write λ, and it's personal preference which one you pick. Just remember that you have to state the "per time" interval you're using. If you just say that the rate is 3, that doesn't mean anything.

What is the expectation value of the number of events that happen during a time t? This expected number simply equals the product λt, from the definition of λ. If the expected number were anything other than λt, then if we divided it by t to obtain the rate, we wouldn't get λ. If you want to be a little more rigorous, consider a very large number n of intervals with length t. The total time in these intervals is nt. This total time is very large, so the number of events that happen during this time is (approximately) equal to $(nt)\lambda$, by the definition of λ. The expected number of events in each of the n intervals with length t is therefore $nt\lambda/n = \lambda t$, as above. So we can write

$$\boxed{\text{(Expected number of events in time } t) = \lambda t} \tag{4.16}$$

In the above setup where λ equals 3 events per minute, the expected number of events that happen in, say, 5 minutes is

$$\lambda t = (3 \text{ events per minute})(5 \text{ minutes}) = 15 \text{ events.} \tag{4.17}$$

Does this mean that we are guaranteed to have exactly 15 events during a particular 5-minute span? Absolutely not. We can theoretically have any number of events, although there is essentially zero chance that the number will differ significantly from 15. (The probability of obtaining the various numbers of events is governed by the Poisson distribution, which we'll discuss in Section 4.7.) But the *expectation value* is 15. That is, if we perform a large number of 5-minute trials and then calculate the average number of events that occur in each trial, the result will be close to 15.

A trickier question to ask is: What is the probability that *exactly one* event happens during a time t? Since λ is the rate, you might think that you can just multiply λ by t, as we did above, to say that the probability is λt. But this certainly can't be correct, because it would imply a probability of 15 for a 5-minute interval in the above setup. This is nonsense, because probabilities can't be larger than 1. If we instead pick a time interval of 20 seconds (1/3 of a minute), we obtain a λt value of 1. This doesn't have the fatal flaw of being larger than 1, but it has another issue, in that it says that exactly one event is *guaranteed* to happen during a 20-second interval. This can't be correct either, because it's certainly possible for zero (or two or three, etc.) events to occur. We'll figure out the exact probabilities of these numbers in Section 4.7.

The strategy of multiplying λ by t to obtain a probability doesn't seem to work. However, there is one special case where it *does* work. If the time interval is extremely small (let's call it ϵ, which is a standard letter to use for something that is very small), then it *is* true that the probability of *exactly one* event occurring during the ϵ time interval is *essentially* equal to $\lambda\epsilon$. We're using the word "essentially" because, although this statement is technically not true, it becomes arbitrarily close to being true in the limit where ϵ approaches zero. In the above example with $\lambda = 1/20$ events per second, the statement, "λt is the probability that exactly one event happens during a time t," is a lousy approximation if $t = 20$ seconds, a decent approximation if $t = 2$ seconds, and a very good approximation if $t = 0.2$ seconds. And it only gets better as the time interval gets smaller. We'll explain why in the first remark below.

We can therefore say that if $P_\epsilon(1)$ stands for the probability that exactly one event happens during a small time interval ϵ, then

$$\boxed{P_\epsilon(1) \approx \lambda\epsilon} \qquad \text{(if } \epsilon \text{ is very small)} \tag{4.18}$$

The smaller ϵ is, the better this approximation is. Technically, the condition in Eq. (4.18) is really "if $\lambda\epsilon$ is very small." But we'll generally be dealing with "normal" sized λ's, so $\lambda\epsilon$ being small is equivalent to ϵ being small. When we deal with continuous time below, we'll actually be taking the $\epsilon \to 0$ limit. In this mathematical limit, the "$\approx$" sign in Eq. (4.18) becomes an exact "=" sign. To sum up:

- If t is very small, then λt is both the expected number of events that happen during the time t *and* (essentially) the probability that exactly one event happens during the time t.

- If t *isn't* very small, then λt is only the expected number of events.

Remarks:

1. We claimed above that λt equals the probability of *exactly one* event occurring, only if t is very small. The reason for this restriction is that if t *isn't* small, then there is the possibility of *multiple* events occurring during the time t. We can be explicit about this as follows. Since we know from Eq. (4.16) that the expected number of events during

any time t is λt, we can use the expression for the expectation value in Eq. (3.4) to write

$$\lambda t = P_t(0) \cdot 0 + P_t(1) \cdot 1 + P_t(2) \cdot 2 + P_t(3) \cdot 3 + \cdots, \tag{4.19}$$

where $P_t(k)$ is the probability of obtaining exactly k events during the time t. Solving for $P_t(1)$ gives

$$P_t(1) = \lambda t - P_t(2) \cdot 2 - P_t(3) \cdot 3 + \cdots. \tag{4.20}$$

We see that $P_t(1)$ is smaller than λt due to the $P_t(2)$ and $P_t(3)$, etc., probabilities. So $P_t(1)$ isn't equal to λt. However, if all of the probabilities of multiple events occurring ($P_t(2)$, $P_t(3)$, etc.) are very small, then $P_t(1)$ is *essentially* equal to λt. And this is exactly what happens if the time interval is very small. For small times, there is hardly any chance of the event even occurring *once*. So it is even less likely that it will occur *twice*, and even less likely for three times, etc.

We can be a little more precise about this. The following argument isn't completely rigorous, but it should convince you that if t is very small, then $P_t(1)$ is essentially equal to λt. If t is very small, then assuming we don't know yet that $P_t(1)$ *equals* λt, we can still say that it should be roughly *proportional* to λt. This is true because if an event has only a tiny chance of occurring, then if you cut λ in half, the probability is essentially cut in half. Likewise if you cut t in half. This proportionality then implies that the probability that exactly two events occur is essentially proportional to $(\lambda t)^2$. We'll see in Section 4.7 that there is actually a factor of $1/2$ involved here, but that is irrelevant in the present argument. The important point is the *quadratic* nature of $(\lambda t)^2$. If λt is sufficiently small, then $(\lambda t)^2$ is negligible compared with λt. Likewise for $P_t(3) \propto (\lambda t)^3$, etc. We can therefore ignore the scenarios where multiple events occur. So with $t \to \epsilon$, Eq. (4.20) becomes

$$P_\epsilon(1) \approx \lambda\epsilon - \cancel{P_\epsilon(2)} \cdot 2 - \cancel{P_\epsilon(3)} \cdot 3 + \cdots, \tag{4.21}$$

in agreement with Eq. (4.18). As mentioned above, if $\lambda\epsilon$ is small, it is because ϵ is small, at least in the situations we'll be dealing with.

2. Imagine drawing the λ vs. t "curve." We have put "curve" in quotes because the curve is actually just a straight horizontal line, since we're assuming a constant λ. If we consider a time interval Δt, the associated area under the curve equals $\lambda\Delta t$, because we have a simple rectangular region. So from Eq. (4.18), this area gives the probability that an event occurs during a time Δt, *provided* that Δt is very small. This might make you think that λ can be interpreted as a probability distribution, because we found in Section 4.2.3 that the area under a distribution curve gives the probability. However, the λ "curve" *cannot* be interpreted as a probability distribution, because this area-equals-probability result holds *only for very small* Δt. The area under a distribution curve has to give the probability for *any* interval on the horizontal axis. The λ "curve" doesn't satisfy this property. The total area under the λ "curve" is infinite (because the straight horizontal line extends for all time), whereas actual probability distributions must have a total area of 1.

3. Since only *one* quantity, λ, is needed to describe everything about a random process whose probability is uniform in time, any other quantity we might want to determine must be able to be written in terms of λ. This will become evident below. ♣

4.6.3 Continuous case

In the case of discrete time in Section 4.6.1, we asked two questions: What is the average waiting time between successes? And what is the probability distribution of the waiting times between successes? We'll now answer these two questions in the case where time is a continuous quantity.

Average waiting time

As in the discrete case, the first of the two questions is fairly easy to answer. Let the average rate of success be λ, and consider a large time t. We know from Eq. (4.16) that the average total number of events that occur during the time t is λt. The average waiting time (which we'll call τ) is the total time divided by the total number of events, λt. That is,

$$\tau = \frac{t}{\lambda t} \quad \Longrightarrow \quad \boxed{\tau = \frac{1}{\lambda}} \qquad \text{(average waiting time)} \tag{4.22}$$

We see that the average waiting time is simply the reciprocal of the rate at which the events occur. For example, if the rate is 5 events per second, then the average waiting time is $1/5$ of a second, which makes sense. This would of course be true in the nonrandom case where the events occur at exactly equally spaced intervals of $1/5$ second. But the nice thing is that Eq. (4.22) holds even for the random process we're discussing, where the intervals aren't equally spaced.

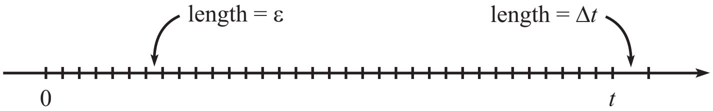
It makes sense that the rate λ is in the denominator in Eq. (4.22), because if λ is small, the average waiting time is large. And if λ is large, the average waiting time is small. And as promised in the third remark above, τ depends on λ.

Distribution of waiting times

Now let's answer the second (more difficult) question: What is the probability distribution of the waiting times between successes? Equivalently, what is the probability that the waiting time from a given event to the next event is between t and $t + \Delta t$, where Δt is small? To answer this, we'll use the same general strategy that we used in the discrete case in Section 4.6.1, except that now the time interval between iterations will be a very small time ϵ instead of 1 second. We will then take the $\epsilon \to 0$ limit, which will make time continuous.

The division of time into little intervals is summarized in Fig. 4.13. From time zero (which is when we'll assume the initial event happens) to time t, we'll break up time into a very large number of very small intervals with length ϵ (which means that there are t/ϵ of these intervals). And then the interval of Δt sits at the end. Both ϵ and Δt are assumed to be very small, but they need not have anything to do with each other. ϵ exists as a calculational tool only, while Δt is the arbitrarily-chosen small time interval that appears in Eq. (4.2).

In order for the next success (event) to happen between t and $t + \Delta t$, there must be failure during every one of the t/ϵ intervals of length ϵ shown in Fig. 4.13, and then there must be success between t and $t + \Delta t$. From Eq. (4.18), the latter happens with probability $\lambda\,\Delta t$, because Δt is assumed to be very small. Also, Eq. (4.18) says



number of intervals = t/ε

Figure 4.13: Dividing time into little intervals.

that the probability of success in any given small interval of length ϵ is $\lambda\epsilon$, which means that the probability of failure is $1 - \lambda\epsilon$. And since there are t/ϵ of these intervals, the probability of failure in all of them is $(1 - \lambda\epsilon)^{t/\epsilon}$. The probability that the next success happens between t and $t + \Delta t$, which we'll label as $P(t, \Delta t)$, is therefore

$$P(t, \Delta t) = \left((1 - \lambda\epsilon)^{t/\epsilon}\right)(\lambda\,\Delta t). \tag{4.23}$$

The reasoning that led to this equation is in the same spirit as the reasoning that led to Eq. (4.14). See the first remark below.

It's now time to use one of the results from Appendix C, namely the approximation given in Eq. (7.14), which says that for small a we can write[4]

$$(1 + a)^n \approx e^{na}. \tag{4.24}$$

This works for negative a as well as positive a. Here e is Euler's number, which has the value of $e \approx 2.71828$. (If you want to know more about e, there's plenty of information in Appendix B!) For the case at hand, a comparison of Eqs. (4.23) and (4.24) shows that we want to define $a \equiv -\lambda\epsilon$ and $n \equiv t/\epsilon$, which yields $na = (t/\epsilon)(-\lambda\epsilon) = -\lambda t$. Eq. (4.24) then gives $(1 - \lambda\epsilon)^{t/\epsilon} \approx e^{-\lambda t}$, so Eq. (4.23) becomes

$$P(t, \Delta t) = e^{-\lambda t}\lambda\,\Delta t. \tag{4.25}$$

The probability distribution (or density) is obtained by simply erasing the Δt, because Eq. (4.2) says that the density is obtained by dividing the probability by the interval length. We therefore see that the desired probability distribution for the waiting time between successes is

$$\boxed{\rho(t) = \lambda e^{-\lambda t}} \qquad \text{(exponential distribution)} \tag{4.26}$$

This is known as the *exponential distribution*. This name is appropriate, of course, because the distribution decreases exponentially with t. As promised in the third remark on page 201, the distribution depends on λ (along with t, of course). In the present setup involving waiting times, it is often more natural to work in terms of the average waiting time τ than the rate λ, in which case the preceding result becomes (using $\lambda = 1/\tau$ from Eq. (4.22))

$$\boxed{\rho(t) = \frac{e^{-t/\tau}}{\tau}} \qquad \text{(exponential distribution)} \tag{4.27}$$

[4] You are strongly encouraged to read Appendix C at this point, if you haven't already. But if you want to take Eq. (4.24) on faith, that's fine too. However, you should at least verify with a calculator that it works fairly well for, say, $a = 0.01$ and $n = 200$.

In the notation of Eq. (4.12), both A and b are equal to $1/\tau$ (or λ). So they are in fact related, as we noted right after Eq. (4.12).

Fig. 4.14 shows plots of $\rho(t)$ for a few different values of the average waiting time, τ. The two main properties of each of these curves are the starting value at $t = 0$ and the rate of decay as t increases. From Eq. (4.27), the starting value at $t = 0$ is $e^0/\tau = 1/\tau$. So the bigger τ is, the smaller the starting value. This makes sense, because if the average waiting time τ is large (equivalently, if the rate λ is small), then there is only a small chance that the next event will happen right away.

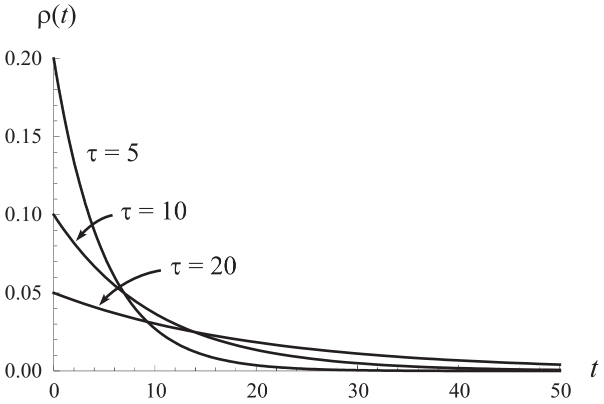


Figure 4.14: Examples of exponential distributions with different values of the average waiting time τ.

How fast do the curves decay? This is governed by the denominator of the exponent in Eq. (4.27). For every τ units that t increases by, $\rho(t)$ decreases by a factor of $1/e$. This can be seen by plugging a time of $t + \tau$ into Eq. (4.27), which gives

$$\rho(t+\tau) = \frac{e^{-(t+\tau)/\tau}}{\tau} = \frac{(e^{-t/\tau}\cdot e^{-1})}{\tau} = \frac{1}{e}\cdot\frac{e^{-t/\tau}}{\tau} = \frac{1}{e}\,\rho(t). \tag{4.28}$$

So $\rho(t + \tau)$ is $1/e$ times as large as $\rho(t)$, and this holds for any value of t. A few particular values of $\rho(t)$ are

$$\rho(0) = \frac{1}{\tau}, \qquad \rho(\tau) = \frac{1}{e\tau}, \qquad \rho(2\tau) = \frac{1}{e^2\tau}, \qquad \rho(3\tau) = \frac{1}{e^3\tau}, \tag{4.29}$$

and so on. If τ is large, the curve takes longer to decrease by a factor of $1/e$. This is consistent with Fig. 4.14, where the large-τ curve falls off slowly, and the small-τ curve falls off quickly. To sum up, if τ is large, the $\rho(t)$ curve starts off low and then decays slowly. And if τ is small, the curve starts off high and then decays quickly.

Example (Same density): Person A measures a very large number of waiting times for a process with $\tau = 5$. Person B does the same for a process with $\tau = 20$. To their surprise, they find that for a special value of t, they both observe (roughly) the same number of waiting times that fall into a given small interval around t. What is this special value of t?

Solution: The given information tells us that the probability densities for the two processes are equal at the special value of t. Plugging the τ values of 5 and 20 into Eq. (4.27) and setting the results equal to each other gives

$$\frac{e^{-t/5}}{5} = \frac{e^{-t/20}}{20} \implies \frac{20}{5} = e^{t/5-t/20} \implies \ln\left(\frac{20}{5}\right) = t\left(\frac{1}{5} - \frac{1}{20}\right)$$
$$\implies \ln 4 = t\left(\frac{15}{100}\right) \implies t = 9.24. \qquad (4.30)$$

This result agrees (at least to the accuracy of a visual inspection) with the value of t where the $\tau = 5$ and $\tau = 20$ curves intersect in Fig. 4.14.

Although it might seem surprising that there exists a value of t for which the densities associated with two different values of τ are equal, it is actually fairly clear, due to the following continuity argument. For small values of t, the $\tau = 5$ process has a larger density (because the events happen closer together), while for large values of t, the $\tau = 20$ process has a larger density (because the events happen farther apart). Therefore, by continuity, there must exist a particular value of t for which the densities are equal. But it takes the above calculation to find the exact value.

Remarks:

1. In comparing Eq. (4.23) with Eq. (4.14), we see in retrospect that we could have obtained Eq. (4.23) by simply replacing the first p in Eq. (4.14) with $\lambda\epsilon$ (because $\lambda\epsilon$ is the probability of success at each intermediate step), the second p with $\lambda\,\Delta t$ (this is the probability of success at the last step), and $k - 1$ with t/ϵ (this is the number of intermediate steps). But you might find these replacements a bit mysterious without the benefit of the reasoning preceding Eq. (4.23).

2. The area under each of the curves in Fig. 4.14 must be 1. The waiting time has to be *something*, so the sum of all the probabilities must be 1. The proof of this fact is very quick, but it requires calculus, so we'll relegate it to Problem 4.8(a). (But note that we did demonstrate this for the discrete case in Eq. (4.15).) Likewise, the expectation value of the waiting time must be τ, because that's how τ was defined. Again, the proof is quick but requires calculus; see Problem 4.8(c). (The demonstration for the discrete case is the task of Problem 4.7.)

3. We've been referring to $\rho(t)$ as the probability distribution of the waiting times from one event to the next. However, $\rho(t)$ is actually the distribution of the waiting times from *any point in time* to the occurrence of the next event. That is, you can start your stopwatch at any time, not just at the occurrence of an event. If you go back through the above discussion, you will see that nowhere did we use the fact that an event actually occurred at $t = 0$.

   However, beware of the following incorrect reasoning. Let's say that an event happens at $t = 0$, but that you don't start your stopwatch until, say, $t = 1$. The fact that the next event after $t = 1$ doesn't happen (on average) until $t = 1 + \tau$ (from the previous paragraph) seems to imply that the average waiting time from $t = 0$ is $1 + \tau$. But it better not be, because we know from above that it's just τ. The error here is that we forgot about the scenarios where the next event after $t = 0$ happens *between* $t = 0$ and $t = 1$. When these events are included, the average waiting time, starting at $t = 0$, ends up correctly being τ. (The demonstration of this fact requires calculus.) In short, the waiting time from $t = 1$ is indeed τ, but the next event (after the $t = 0$ event) might have already happened before $t = 1$.

4. In a sense, the curves for all of the different values of τ in Fig. 4.14 are really the same curve. They're just stretched or squashed in the horizontal and vertical directions. The general form of the curve described by the expression in Eq. (4.27) is shown in Fig. 4.15.

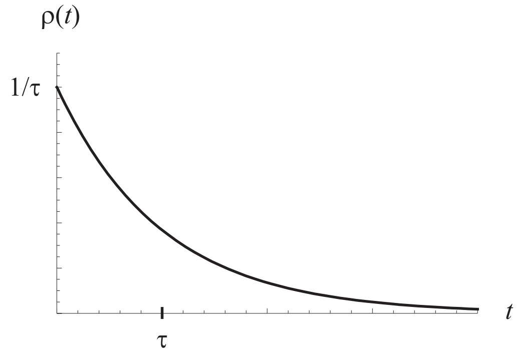


Figure 4.15: The general form of the exponential distribution.

As long as we change the scales on the axes so that τ and $1/\tau$ are always located at the same positions, then the curves will look the same for any τ. For example, as we saw in Eq. (4.29), no matter what the value of τ is, the value of the curve at $t = \tau$ is always $1/e$ times the value at $t = 0$. Of course, when we plot things, we usually keep the scales fixed, in which case the τ and $1/\tau$ positions move along the axes, as shown in Fig. 4.16 (these are the same curves as in Fig. 4.14). But by suitable uniform stretching/squashing of the axes, the curve in Fig. 4.15 can be turned into any of the curves in Fig. 4.16.

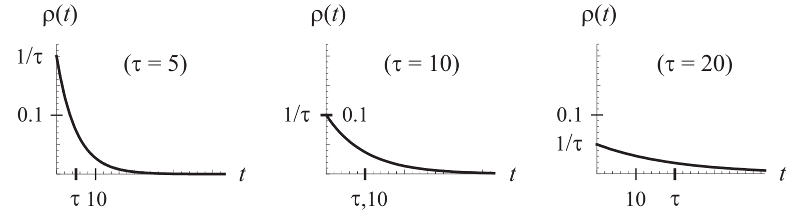


Figure 4.16: These curves can be obtained from the curve in Fig. 4.15 by suitable stretching/squashing of the axes.

5. The fact that any of the curves in Fig. 4.16 can be obtained from any of the other curves by stretching and squashing the two directions by inverse (as you can verify) factors

implies that the areas under all of the curves are the same. (This is consistent with the fact that all of the areas must be 1.) To see how these inverse factors work together to keep the area constant, imagine the area being broken up into a large number of thin vertical rectangles, stacked side by side under the curve. The stretching and squashing of the curve does the same thing to each rectangle. All of the widths get stretched by a factor of f, and all of the heights get squashed by the same factor of f (or $1/f$, depending on your terminology). So the area of each rectangle remains the same. The same thing must then be true for the area under the whole curve.

6. Note that the distribution for the waiting time is a discrete distribution in the case of discrete time (see Eq. (4.14)), and a continuous distribution in the case of continuous time (see Eq. (4.27)). Although these facts make perfect sense, one should be careful about extrapolating to a general conclusion. In the Poisson discussion in the following section, we'll encounter a discrete distribution in the case of continuous time. ♣

4.7 Poisson distribution

The goal of this section is to derive the Poisson probability distribution,

$$\boxed{P(k) = \frac{a^k e^{-a}}{k!}} \qquad \text{(Poisson distribution)} \tag{4.31}$$

The parameter a depends on the situation at hand, and k is the value of the random variable, which is the number of events that occur in a certain region of time (or space, or whatever), as we'll discuss below. Since k is an integer (because it is the number of events that occur), the Poisson distribution is a discrete one. A common type of situation where this distribution arises is the following.

As with the exponential distribution in the previous section, consider a repeating event that happens completely randomly in time. We will show that the probability distribution of the *number of events that occur during a given time interval* takes the form of the above Poisson distribution. Whereas the exponential distribution deals with the *waiting time* until the next event, the Poisson distribution deals with the *number of events* in a given time interval. As in the case of the exponential distribution, our strategy for deriving the Poisson distribution will be to first consider the case of discrete time, and then the case of continuous time.

4.7.1 Discrete case

Consider a process that is repeated each second (so time is discretized into 1-second intervals), and let the probability of success in each trial be p (the same for all trials). For example, as in Section 4.6.1, we can roll a hypothetical 10-sided die once every second, and if the die shows a "1," then we consider that a success. The other nine numbers represent failure. As in Section 4.6.1, it isn't actually necessary to introduce time here. We could simply talk about the number of iterations of the process, as we will in the balls-in-boxes example below.

The question we will answer here is: What is the probability distribution of the number of successes that occur in a time interval of n seconds? In other words, what is the probability, $P(k)$, that exactly k events happen during a time span of

n seconds? It turns out that this is *exactly* the same question that we answered in Section 4.5 when we derived the binomial distribution in Eq. (4.6). So we can just copy over the reasoning here. We'll formulate things in the language of rolls of a die, with a "1" being a success. But the setup could be anything with a probability p of success.

The probability that a *specific set* of k of the n rolls all yield a 1 equals p^k, because each of the k rolls has a p probability of yielding a 1. We then need the other $n-k$ rolls to *not* yield a 1, because we want *exactly* k 1's. This happens with probability $(1-p)^{n-k}$, because each of the $n-k$ rolls has a $1-p$ probability of being something other than a 1. The probability that a specific set of k rolls (and no others) all yield success is therefore $p^k \cdot (1-p)^{n-k}$. Finally, since there are $\binom{n}{k}$ ways to pick a specific set of k rolls, we see that the probability that exactly k of the n rolls yield a 1 is

$$\boxed{P(k) = \binom{n}{k} p^k (1-p)^{n-k}} \tag{4.32}$$

This distribution is exactly the same as the binomial distribution in Eq. (4.6), so there's nothing new here. But there will indeed be something new when we discuss the continuous case in Section 4.7.2.

Example (Balls in boxes): Let n balls be thrown randomly into b boxes. What is the probability, $P(k)$, that a given box has exactly k balls in it?

Solution: This is a restatement of the problem we just solved. Imagine randomly throwing one ball each second into the boxes, and consider a particular box. (As mentioned above, the time interval of one second is irrelevant. All that matters is that we perform n iterations of the process, sooner or later.) If a given ball ends up in that box, we'll call that a success. For each ball, this happens with probability $1/b$, because there are b boxes. So the p in the above discussion equals $1/b$. Since we're throwing n balls into the boxes, we're simply performing n iterations of a process that has a probability $p = 1/b$ of success. Eq. (4.32) is therefore applicable, and with $p = 1/b$ it gives the probability of obtaining exactly k successes (that is, exactly k balls in a particular box) as

$$P(k) = \binom{n}{k} \left(\frac{1}{b}\right)^k \left(1 - \frac{1}{b}\right)^{n-k}. \tag{4.33}$$

We've solved the problem, but let's now see if our answer makes sense. As a concrete example, consider the case where we have $n = 1000$ balls and $b = 100$ boxes. On average, we expect to have $n/b = 10$ balls in each box. But many (in fact, most) of the boxes will have other numbers of balls. In theory, the number k of balls in a particular box can take on any value from 0 to $n = 1000$. But intuitively we expect most of the boxes to have *roughly* 10 balls (say, between 5 and 15 balls). We certainly don't expect many boxes to have 2 or 50 balls.

Fig. 4.17 shows a plot of the $P(k)$ in Eq. (4.33), for the case where $n = 1000$ and $b = 100$. As expected, it is peaked near the average value, $n/b = 10$, and it becomes negligible a moderate distance away from $k = 10$. There is very little chance of having fewer than 3 or more than 20 balls in a given box; Eq. (4.33) gives $P(2) \approx 0.2\%$ and $P(21) \approx 0.1\%$. We've arbitrarily chopped off the plot at $k = 30$ because the probabilities between $k = 30$ (or even earlier) and $k = 1000$ are indistinguishable from zero. But technically all of these probabilities are nonzero. For example, $P(1000) = (1/100)^{1000}$, because if $k = 1000$ then all of the 1000 balls need to end up in the given box, and each one ends up there with probability $1/100$. The resulting probability of 10^{-2000} is utterly negligible.

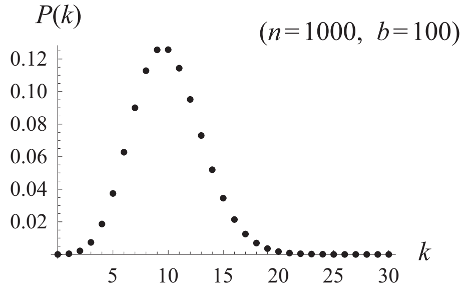


Figure 4.17: The probability distribution for the number of balls in a given box, if $n = 1000$ balls are thrown into $b = 100$ boxes.

4.7.2 Continuous case

As with the exponential distribution in Section 4.6.3, we'll now consider the case where time is continuous. That is, we'll assume that we can have a successful event at *any* instant, not just at the evenly-spaced 1-second marks, as we assumed above. As in Section 4.6.3, such a process can be completely described by just *one* number – the average rate of events, which we'll again call λ. Eq. (4.18) tells us that $\lambda\epsilon$ is the probability that exactly one event occurs in a very small time interval ϵ. The smaller the ϵ, the smaller the probability that the event occurs. We're assuming that λ is constant in time, that is, the event is just as likely to occur at one time as any other.

Our goal here is to answer the question: What is the probability, $P(k)$, that exactly k events occur during a given time span of t? To answer this, we'll use the same general strategy that we used above in the discrete case, except that now the time interval between iterations will be a very small time ϵ instead of 1 second. We will then take the $\epsilon \to 0$ limit, which will make time continuous. The division of time into little intervals is summarized in Fig. 4.18. We're dividing the time interval t into a very large number of very small intervals with length ϵ. There are t/ϵ of these intervals, which we'll label as n. There is no need to stick a Δt interval on the end, as there was in Fig. 4.13.

Compared with the discrete case we addressed above, Eq. (4.18) tells us that the probability of exactly one event occurring in a given small interval of length ϵ is now $\lambda\epsilon$ instead of p. So we can basically just repeat the derivation preceding Eq. (4.32), which itself was a repetition of the derivation preceding Eq. (4.6). You're probably getting tired of it by now!

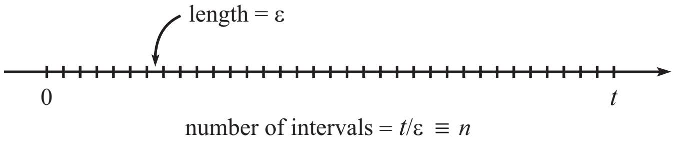


Figure 4.18: Dividing time into little intervals.

The probability that a *specific set* of k of the n little intervals all yield exactly one event each equals $(\lambda\epsilon)^k$, because each of the k intervals has a $\lambda\epsilon$ probability of yielding one event. We then need the other $n-k$ intervals to *not* yield an event, because we want *exactly* k events. This happens with probability $(1-\lambda\epsilon)^{n-k}$, because each of the $n-k$ intervals has a $1-\lambda\epsilon$ chance of yielding zero events. The probability that a specific set of k intervals (and no others) all yield an event is therefore $(\lambda\epsilon)^k \cdot (1-\lambda\epsilon)^{n-k}$. Finally, since there are $\binom{n}{k}$ ways to pick a specific set of k intervals, we see that the probability that exactly k of the n intervals yield an event is

$$P(k) = \binom{n}{k}(\lambda\epsilon)^k(1-\lambda\epsilon)^{n-k}. \tag{4.34}$$

This is simply Eq. (4.32) with p replaced by $\lambda\epsilon$.

Now it's time to have some mathematical fun. Let's see what Eq. (4.34) reduces to in the $\epsilon \to 0$ limit, which will give us the desired continuous-time limit. Note that $\epsilon \to 0$ implies that $n \equiv t/\epsilon \to \infty$. The math here will be a little more involved than the math that led to the exponential distribution in Eq. (4.26).

If we write out the binomial coefficient and expand things a bit, Eq. (4.34) becomes

$$P(k) = \frac{n!}{(n-k)!\,k!}(\lambda\epsilon)^k(1-\lambda\epsilon)^n(1-\lambda\epsilon)^{-k}. \tag{4.35}$$

Of the various letters in this equation, n is huge, ϵ is tiny, and λ and k are "normal," not assumed to be huge or tiny. λ is determined by the setup, and k is the number of events we're concerned with. (We'll see below that the relevant k's are roughly the size of the product $\lambda t = \lambda n\epsilon$.) In the $\epsilon \to 0$ limit (and hence $n \to \infty$ limit), we can make three approximations to Eq. (4.35):

- First, in the $n \to \infty$ limit, we can say that

$$\frac{n!}{(n-k)!} \approx n^k, \tag{4.36}$$

  at least in a multiplicative sense (we don't care about an additive sense). This follows from the fact that $n!/(n-k)!$ is the product of the k numbers from n down to $n-k+1$. And if n is large compared with k, then all of these k numbers are essentially equal to n (multiplicatively). Therefore, since there are k of them, we simply get n^k. You can verify this for, say, the case of $n = 1{,}000{,}000$ and $k = 10$. The product of the 10 numbers from $1{,}000{,}000$ down to $999{,}991$ equals $1{,}000{,}000^{10}$ to within an error of 0.005%

- Second, we can apply the $(1+a)^n \approx e^{na}$ approximation from Eq. (7.14) in Appendix C, which we already used once in the derivation of the exponential distribution; see the discussion following Eq. (4.24). We can use this approximation to simplify the $(1-\lambda\epsilon)^n$ term. With $a \equiv -\lambda\epsilon$, Eq. (7.14) gives

$$(1-\lambda\epsilon)^n \approx e^{-n\lambda\epsilon}. \tag{4.37}$$

- Third, in the $\epsilon \to 0$ limit, we can use the $(1+a)^n \approx e^{na}$ approximation again, this time to simplify the $(1-\lambda\epsilon)^{-k}$ term. The result is

$$(1-\lambda\epsilon)^{-k} \approx e^{k\lambda\epsilon} \approx e^0 = 1, \tag{4.38}$$

because for any fixed values of k and λ, the $k\lambda\epsilon$ exponent becomes infinitesimally small as $\epsilon \to 0$. Basically, in $(1-\lambda\epsilon)^{-k}$ we're forming a finite power of a number that is essentially equal to 1. Note that this reasoning doesn't apply to the $(1-\lambda\epsilon)^n$ term in Eq. (4.37), because n isn't a fixed number. It changes with ϵ, in that it becomes large as ϵ becomes small.

In the $\epsilon \to 0$ and $n \to \infty$ limits, the "≈" signs in the approximations in the preceding three equations turn into exact "=" signs. Applying these three approximations to Eq. (4.35) gives

$$\begin{aligned} P(k) &= \frac{n!}{(n-k)!\,k!}(\lambda\epsilon)^k(1-\lambda\epsilon)^n(1-\lambda\epsilon)^{-k} \\ &= \frac{n^k}{k!}(\lambda\epsilon)^k e^{-n\lambda\epsilon}\cdot 1 \\ &= \frac{1}{k!}(\lambda\cdot n\epsilon)^k e^{-\lambda\cdot n\epsilon} \\ &= \frac{1}{k!}(\lambda t)^k e^{-\lambda t}, \end{aligned} \tag{4.39}$$

where we have used $n \equiv t/\epsilon \Longrightarrow n\epsilon = t$ to obtain the last line. Now, from Eq. (4.16) λt is the average number of events that are expected to occur in the time t. Let's label this average number of events as $a \equiv \lambda t$. We can then write Eq. (4.39) as

$$\boxed{P(k) = \frac{a^k e^{-a}}{k!}} \qquad \text{(Poisson distribution)} \tag{4.40}$$

where a is the average number of events in the time interval under consideration. If you want, you can indicate the a value by writing $P(k)$ as $P_a(k)$.

Since a is the only parameter left on the righthand side of Eq. (4.40), the distribution is completely specified by a. The individual values of λ and t don't matter. All that matters is their product $a \equiv \lambda t$. This means that if we, say, double the time interval t under consideration and also cut the rate λ in half, then a remains unchanged; so we have exactly the same distribution $P(k)$. Although it is clear that doubling t and halving λ yields the same *average* number of events (since the average equals the product λt), it might not be intuitively obvious that the entire $P(k)$ *distribution* is the same. But the result in Eq. (4.40) shows that this is indeed the case.

The Poisson distribution in Eq. (4.40) gives the probability of obtaining exactly k events during a period of time for which the expected number is a. Since k is a discrete variable (being the integer number of times that an event occurs), the Poisson distribution is a discrete distribution. Although the Poisson distribution is derived from a *continuous* process (in that the time t is continuous, which means that an event can happen at any time), the distribution itself is a *discrete* distribution, because k must be an integer. Note that while the *observed* number of events k must be an integer, the *average* number of events a need not be.

Remark: Let's discuss this continuous/discrete issue a little further. In the last remark in Section 4.6.3, we noted that the exponential distribution for the waiting time, t, is a discrete distribution in the case of discrete time, and a continuous distribution in the case of continuous time. This seems reasonable. But for the Poisson distribution, the distribution for the number of events, k, is a discrete distribution in the case of discrete time, and also (as we just noted) a discrete distribution in the case of continuous time. It is simply always a discrete distribution, because the random variable is the number of events, k, which is discrete. The fact that time might be continuous is irrelevant, as far as the discreteness of k goes. The difference in the case of the exponential distribution is that *time itself* is the random variable (because we're considering waiting times). So if we make time continuous, then by definition we're also making the random variable continuous, which means that we have a continuous distribution. ♣

Example (Number of shoppers): On average, one shopper enters a given store every 15 seconds. What is the probability that in a given time interval of one minute, zero shoppers enter the store? Four shoppers? Eight shoppers?

Solution: The given average time interval of 15 seconds tells us that the average number of shoppers who enter the store in one minute is $a = 4$. Having determined a, we simply need to plug the various values of k into Eq. (4.40). For $k = 0, 4$, and 8 we have

$$\begin{aligned}
P(0) &= \frac{4^0 e^{-4}}{0!} = 1 \cdot e^{-4} \approx 0.018 \approx 2\%,\\
P(4) &= \frac{4^4 e^{-4}}{4!} = \frac{32}{3} \cdot e^{-4} \approx 0.195 \approx 20\%,\\
P(8) &= \frac{4^8 e^{-4}}{8!} = \frac{512}{315} \cdot e^{-4} \approx 0.030 = 3\%. \qquad (4.41)
\end{aligned}$$

We see that the probability that four shoppers enter the store in a given minute is about 10 times the probability that zero shoppers enter. The probabilities quickly die off as k gets larger. For example, $P(12) \approx 0.06\%$.

The above results are a subset of the information contained in the plot of $P(k)$ shown in Fig. 4.19. Note that $P(3) = P(4)$. This is evident from the above expression for $P(4)$, because if we cancel a factor of 4 in the numerator and denominator, we end up with $4^3 e^{-4}/3!$ which equals $P(3)$. See Problem 4.10 for more on this equality.

Remember that when finding $P(k)$, the only parameter that matters is a. If we modify the problem by saying that on average one shopper enters the store every 15 *minutes*, and if we change the time interval to one *hour* (in which case a again equals 4), then all of the $P(k)$ values are exactly the same as above.

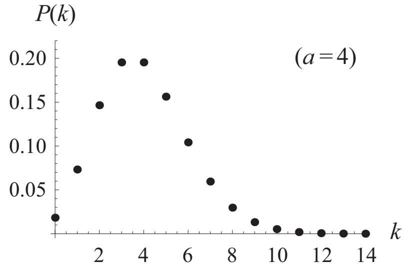


Figure 4.19: The Poisson distribution with $a = 4$.

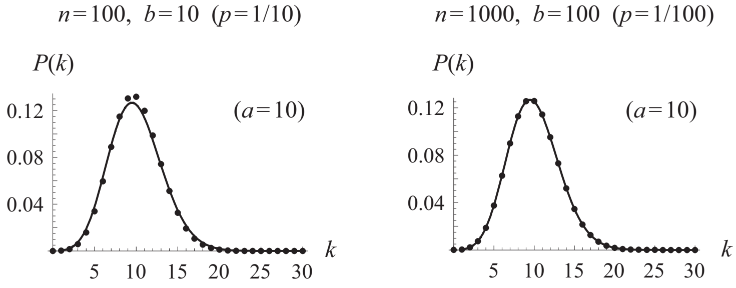
Example (Balls in boxes, again): Although Eq. (4.40) technically holds only in the limit of a continuous process, it still provides a very good approximation for discrete processes, as long as the numbers involved are fairly large. Consider the balls-in-boxes example in Section 4.7.1. With $n = 1000$ and $b = 100$, the average number of balls in a box is $a = n/b = 10$. Since b is fairly large, we expect that the Poisson distribution in Eq. (4.40) with $a = 10$ will provide a good approximation to the exact binomial distribution in Eq. (4.33) with $n = 1000$ and $b = 100$, or equivalently Eq. (4.32) with $n = 1000$ and $p = 1/b = 1/100$.

Let's see how good the approximation is. Fig. 4.20 shows plots for two different sets of n and b values: $n = 100$, $b = 10$; and $n = 1000$, $b = 100$. With these values, both plots have $a = 10$. The dots in the second plot are a copy of the dots in Fig. 4.17. In both plots we have superimposed the exact discrete binomial distribution (the dots) and the Poisson distribution (the curves).[5] Since the plots have the same value of a, they have the same Poisson curve. In the right plot, the points pretty much lie on the curve, so the approximate Poisson probabilities in Eq. (4.40) are essentially the same as the exact binomial probabilities in Eq. (4.33). In other words, the approximation is a very good one.

However, in the left plot, the points lie slightly off the curve. The average $a = n/b$ still equals 10, so the Poisson curve is exactly the same as in the right plot. But the exact binomial probabilities in Eq. (4.33) are changed from the $n = 1000$ and $b = 100$ case. The Poisson approximation doesn't work as well here, although it's still reasonably good. The condition under which the Poisson approximation is a good one turns out to be the very simple relation, $p \equiv 1/b \ll 1$. See Problem 4.14.

The Poisson distribution in Eq. (4.40) works perfectly well for small a, even $a < 1$. It's just that in this case, the plot of $P(k)$ doesn't have a bump, as it does in Figs. 4.19 and 4.20. Instead, it starts high and then falls off as k increases. Fig. 4.21 shows the plot of $P(k)$ for various values of a. We've arbitrarily decided to cut off the plots at $k = 20$, even though they technically go on forever. Since we are assuming that time is continuous, we can theoretically have an arbitrarily large number of

[5] We've drawn the Poisson distribution as a continuous curve (the $k!$ in Eq. (4.40) can be extrapolated to non-integer values of k), because it would be difficult to tell what's going on in the figure if we plotted two sets of points nearly on top of each other. But you should remember that we're really only concerned with integer values of k, since the k in Eq. (4.40) is the number of times something occurs. We've plotted the whole curve for visual convenience only.



Dots = exact binomial result
Curves = Poisson approximation (both plots have a=10)

Figure 4.20: Comparison between the exact binomial result and the Poisson approximation.

events in any given time interval, although the probability will be negligibly small. In the plots, the probabilities are effectively zero by $k = 20$, except in the $a = 15$ case.

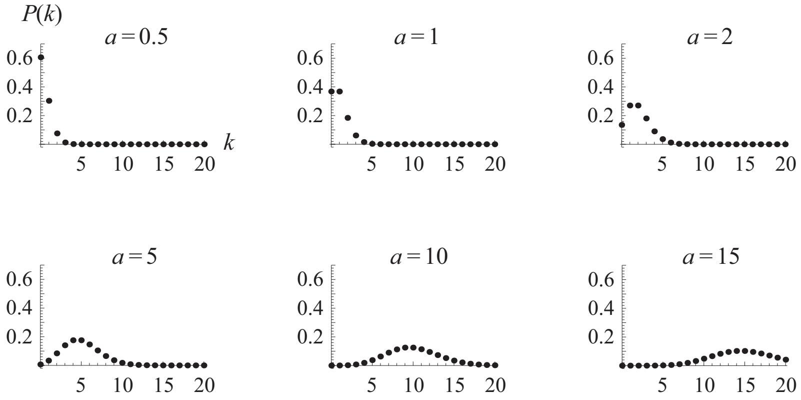


Figure 4.21: The Poisson distribution for various values of a.

As a increases, the bump in the plots (once it actually becomes a bump) does three things: (1) it shifts to the right, because it is centered near $k = a$, due to the result in Problem 4.10, (2) it decreases in height, due to the result in Problem 4.11, and (3) it becomes wider, due to the result in Problem 4.13. The last two of these properties are consistent with each other, in view of the fact that the sum of all the probabilities must equal 1, for any value of a.

Eq. (4.40) gives the probability of obtaining zero events as $P(0) = e^{-a}$. If $a = 0.5$ then $P(0) = e^{-0.5} \approx 0.61$. This agrees with the first plot in Fig. 4.21. Likewise, if $a = 1$ then $P(0) = e^{-1} \approx 0.37$, in agreement with the second plot. If a is large then the $P(0) = e^{-a}$ probability goes to zero, in agreement with the bottom three plots. This makes sense; if the average number of events is *large*, then it is very *unlikely* that we will obtain zero events. In the opposite extreme, if a is very small (for example, $a = 0.01$), then the $P(0) = e^{-a}$ probability is very close to 1.

This again makes sense; if the average number of events is very *small*, then it is very *likely* that we will obtain zero events.

To make it easier to compare the six plots in Fig. 4.21, we have superimposed them in Fig. 4.22. Although we have drawn these Poisson distributions as continuous curves to make things clearer, remember that the distribution applies only to integer values of k.

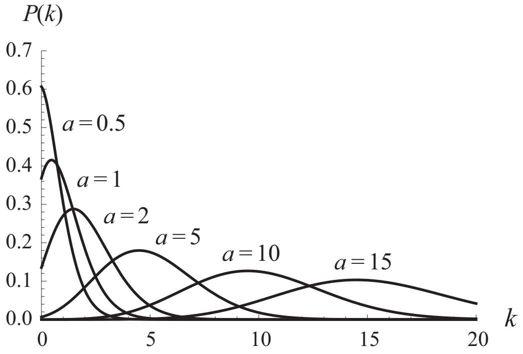


Figure 4.22: Superimposing the plots in Fig. 4.21, drawn as continuous curves.

Problems 4.9 through 4.13 cover various aspects of the Poisson distribution, namely: the fact that the total probability is 1, the location of the maximum, the value of the maximum, the expectation value, and the variance.

4.8 Gaussian distribution

The Gaussian probability distribution (also known as the "normal distribution" or the "bell curve") is the most important of all the probability distributions. The reason, as we will see in Chapter 5, is that in the limit of large numbers, many other distributions reduce to a Gaussian. But for now, we'll just examine the mathematical properties of the Gaussian distribution. The distribution is commonly written in either of the following forms:

$$\boxed{f(x) = \sqrt{\frac{b}{\pi}}\, e^{-b(x-\mu)^2} \quad \text{or} \quad \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}} \tag{4.42}$$

If you want to explicitly indicate the parameters that appear, you can write the distribution as $f_{\mu,b}(x)$ or $f_{\mu,\sigma}(x)$. The Gaussian distribution is a *continuous* one. That is, x can take on a continuum of values, like t in the exponential distribution, but unlike k in the binomial and Poisson distributions. The Gaussian probability distribution is therefore a probability *density*. As mentioned at the beginning of Section 4.2.2, the standard practice is to use lowercase letters (like the f in $f(x)$) for probability densities, and to use uppercase letters (like the P in $P(k)$) for actual probabilities.

The second expression in Eq. (4.42) is obtained from the first by letting $b = 1/2\sigma^2$. The first expression is simpler, but the second one is more common. This is due to the fact that the standard deviation, which we introduced in Section 3.3, turns out simply to be σ. Hence our use of the letter σ here. Note that b (or σ) appears twice in the distribution – in the exponent and in the prefactor. These two appearances conspire to make the total area under the distribution equal to 1. See Problem 4.22 for a proof of this fact.

The quantities μ and b (or μ and σ) depend on the specific situation at hand. Let's look at how these quantities affect the shape and location of the curve. We'll work mainly with the first form in Eq. (4.42) here, but any statements we make about b can be converted into statements about σ by replacing b with $1/2\sigma^2$.

Mean

Let's consider μ first. Fig. 4.23 shows the plots of two Gaussian distributions, one with $b = 2$ and $\mu = 6$, and the other with $b = 2$ and $\mu = 10$. The two functions are

$$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-2(x-6)^2} \qquad \text{and} \qquad f(x) = \sqrt{\frac{2}{\pi}}\, e^{-2(x-10)^2}. \tag{4.43}$$

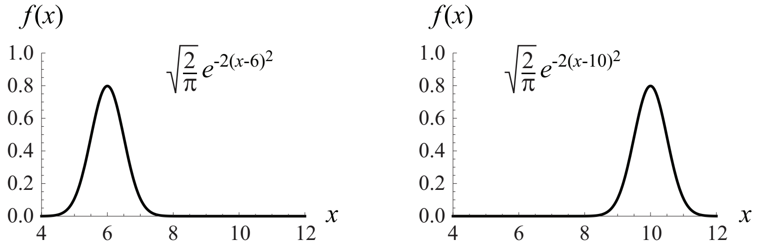


Figure 4.23: Gaussian distributions with different means.

It is clear from the plots that μ is the location of the maximum of the curve. Mathematically, this is true because the $e^{-b(x-\mu)^2}$ exponential factor has an exponent that is either zero or negative (because a square is always zero or positive). So this exponential factor is always less than or equal to 1. Its maximum value occurs when the exponent is zero, that is, when $x = \mu$. The peak is therefore located at $x = \mu$. If we increase μ (while keeping b the same), the whole curve just shifts to the right, keeping the same shape. This is evident from the figure.

Because the curve is symmetric around the maximum, μ is also the mean (or expectation value) of the distribution:

$$\text{Mean} = \mu. \tag{4.44}$$

Since we used the letter μ for the mean throughout Chapter 3, it was a natural choice to use μ the way we did in Eq. (4.42). Of course, for the same reason, it would also have been natural to use μ for the mean of the exponential and Poisson distributions. But we chose to label those means as τ and a, so that there wouldn't be too many μ's floating around in this chapter.

Height

Now let's consider b. Fig. 4.24 shows the plots of two Gaussian distributions, one with $b = 2$ and $\mu = 6$, and the other with $b = 8$ and $\mu = 6$. The two functions are

$$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-2(x-6)^2} \qquad \text{and} \qquad f(x) = \sqrt{\frac{8}{\pi}}\, e^{-8(x-6)^2}. \tag{4.45}$$

Note that the scales on both the x and y axes in Fig. 4.24 are different from those in Fig. 4.23. The first function here is the same as the first function in Fig. 4.23.

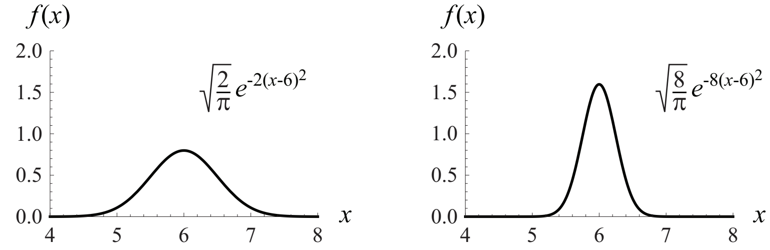


Figure 4.24: Gaussian distributions with different values of b. Both the heights and the widths differ.

It is clear from the plots that b affects both the height and width of the curve. Let's see how these two effects come about. The effect on the height is easy to understand, because the height of the curve (the maximum value of the function) is simply $\sqrt{b/\pi}$. This is true because when x equals μ (which is the location of the maximum), the $e^{-b(x-\mu)^2}$ factor equals 1, in which case the value of $\sqrt{b/\pi}\, e^{-b(x-\mu)^2}$ is just $\sqrt{b/\pi}$. (By the same reasoning, the second expression in Eq. (4.42) gives the height in terms of σ as $1/\sqrt{2\pi\sigma^2}$.) Looking at the two functions in Eq. (4.45), we see that the ratio of the heights is $\sqrt{8/2} = 2$. And this is indeed the ratio we observe in Fig. 4.24. To summarize:

$$\text{Height} = \sqrt{\frac{b}{\pi}} = \sqrt{\frac{1}{2\pi\sigma^2}}\,. \tag{4.46}$$

Width in terms of *b*

Now for the width. We see that the second function in Fig. 4.24 is both taller and narrower than the first. (But it has the same midpoint, because we haven't changed μ.) The factor by which it is shrunk in the horizontal direction appears to be about 1/2. And in fact, it is exactly 1/2. It turns out that the width of a Gaussian curve is proportional to $1/\sqrt{b}$. This means that since we increased b by a factor of 4 in constructing the second function, we decreased the width by a factor of $1/\sqrt{4} = 1/2$. Let's now show that the width is in fact proportional to $1/\sqrt{b}$.

But first, what do we mean by "width"? A vertical rectangle has a definite width, but a Gaussian curve doesn't, because the "sides" are tilted. We could arbitrarily define the width to be how wide the curve is at a height equal to half the maximum height. Or instead of half, we could say a third. Or a tenth. We can define it however we want, but the nice thing is that however we choose to define it, the above "proportional to $1/\sqrt{b}$" result will still hold, as long as we pick one definition and stick with it for whatever curves we're looking at. Similarly, if we want to work with the second expression in Eq. (4.42), then since $1/\sqrt{b} \propto \sigma$, the width will be proportional to σ, independent of the specifics of our arbitrary definition.

The definition we'll choose here is: The width of a curve is the width at the height equal to $1/e$ (which happens to be about 0.37) times the maximum height (which is $\sqrt{b/\pi}$). This $1/e$ choice is a natural one, because the x values that correspond to this height are easy to find. They are simply $\mu \pm 1/\sqrt{b}$, because the first expression in Eq. (4.42) gives

$$\begin{aligned} f(\mu \pm 1/\sqrt{b}) &= \sqrt{b/\pi}\, e^{-b[(\mu \pm 1/\sqrt{b})-\mu]^2} \\ &= \sqrt{b/\pi}\, e^{-b(\pm 1/\sqrt{b})^2} \\ &= \sqrt{b/\pi}\, e^{-b/b} \\ &= \sqrt{\frac{b}{\pi}} \cdot \frac{1}{e}, \end{aligned} \tag{4.47}$$

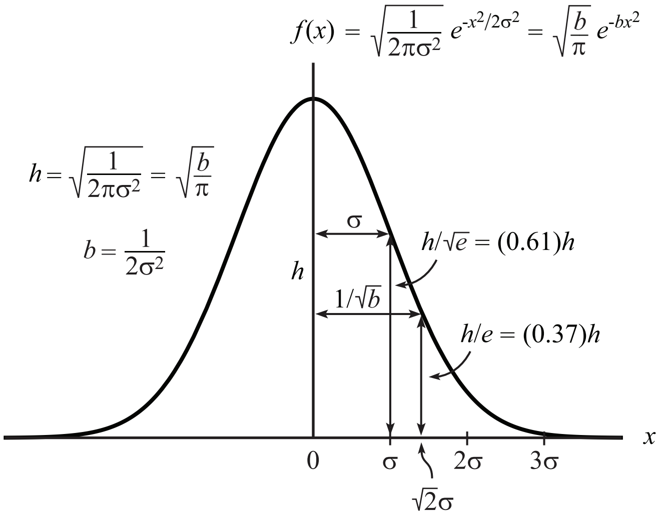
as desired. Since the difference between $\mu + 1/\sqrt{b}$ and $\mu - 1/\sqrt{b}$ equals $2/\sqrt{b}$, the width of the Gaussian curve (by our arbitrary definition) is $2/\sqrt{b}$. So $1/\sqrt{b}$ is half of the width, which we'll call the "half-width". (The term "half-width" can also refer to the full width of the curve at half of the maximum height. We won't use that meaning here.) Again, any other definition of the width would also yield the $\sqrt{b}$ in the denominator. That's the important part. The 2 in the numerator doesn't have much significance. The half-width is shown below in Fig. 4.25, following the discussion of the width in terms of σ.

Width in terms of σ

When working with the second form in Eq. (4.42) (which is the more common of the two), the default definition of the width is the width at the height equal to $1/\sqrt{e}$ times the maximum height. This definition (which is different from the above $1/e$ definition) is used because the values of x that correspond to this height are simply $x \pm \sigma$. This is true because if we plug $x = \mu \pm \sigma$ into the second expression in Eq. (4.42), we obtain

$$\begin{aligned} f(\mu \pm \sigma) &= \sqrt{1/2\pi\sigma^2}\, e^{-[(\mu \pm \sigma)-\mu]^2/2\sigma^2} \\ &= \sqrt{1/2\pi\sigma^2}\, e^{-(\pm\sigma)^2/2\sigma^2} \\ &= \sqrt{1/2\pi\sigma^2}\, e^{-1/2} \\ &= \sqrt{\frac{1}{2\pi\sigma^2}} \cdot \frac{1}{\sqrt{e}}. \end{aligned} \tag{4.48}$$

The factor of $1/\sqrt{e}$ here equals $1/\sqrt{2.718} \approx 0.61$, which is larger than the $1/e \approx 0.37$ factor in our earlier definition. This is consistent with the fact that the $x = \mu \pm \sigma$ points (where the height is $1/\sqrt{e} \approx 0.61$ times the maximum) are closer to the center than the $x = \mu \pm 1/\sqrt{b} = \mu \pm \sqrt{2}\,\sigma$ points (where the height is $1/e \approx 0.37$ times the maximum). This is summarized in Fig. 4.25; we have chosen $\mu = 0$ for convenience.



$f(x) = \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{-x^2/2\sigma^2} = \sqrt{\frac{b}{\pi}}\, e^{-bx^2}$
$h = \sqrt{\frac{1}{2\pi\sigma^2}} = \sqrt{\frac{b}{\pi}}$
$b = \frac{1}{2\sigma^2}$
σ
$h/\sqrt{e} = (0.61)h$
h
$1/\sqrt{b}$
$h/e = (0.37)h$
x
0
σ
2σ
3σ
$\sqrt{2}\sigma$

Figure 4.25: Different definitions of the half-width, in terms of b and σ.

Although the $x = \mu \pm \sigma$ points yield a nice value of the Gaussian distribution ($1/\sqrt{e}$ times the maximum), the *really* nice thing about the $x = \mu \pm \sigma$ points is that they are one *standard deviation* from the mean μ. It can be shown (with calculus, see Problem 4.23) that the standard deviation (defined in Eq. (3.40)) of the Gaussian distribution given by the second expression in Eq. (4.42) is simply σ. This is why the second form in Eq. (4.42) is more widely used than the first. And for the same reason, people usually choose to (arbitrarily) define the half-width of the Gaussian curve to be σ instead of the $1/\sqrt{b} = \sqrt{2}\,\sigma$ half-width that we found earlier. That is, they're defining the width by looking at where the function is $1/\sqrt{e}$ times the maximum, instead of $1/e$ times the maximum. As we noted earlier, any such definition is perfectly fine; it's a matter of person preference. The critical point is that the width is proportional to σ (or $1/\sqrt{b}$). The exact numerical factor involved is just a matter of definition.

As mentioned on page 153, it can be shown numerically that about 68% of the total area (probability) under the Gaussian curve lies between the points $\mu \pm \sigma$. In other words, you have a 68% chance of obtaining a value of x that is within one standard deviation from the mean μ. We used the word "numerically" above, because although the areas under the curves (or the discrete sums) for all of the other distributions we've dealt with in the chapter can be calculated in closed form, this isn't true for the Gaussian distribution. So when finding the area under the Gaussian curve, you always need to specify the numerical endpoints of your interval, and then you can use a computer to calculate the area (numerically, to whatever accuracy you want). It can likewise be shown that the percentage of the total area that is within two standard deviations from μ (that is, between the points $\mu \pm 2\sigma$) is about

95%. And the percentage within three standard deviations from μ is about 99.7%. These percentages are consistent with a visual inspection of the shaded areas in Fig. 4.26. The percentages rapidly approach 100%. The percentage within five standard deviations from μ is about 99.99994%.

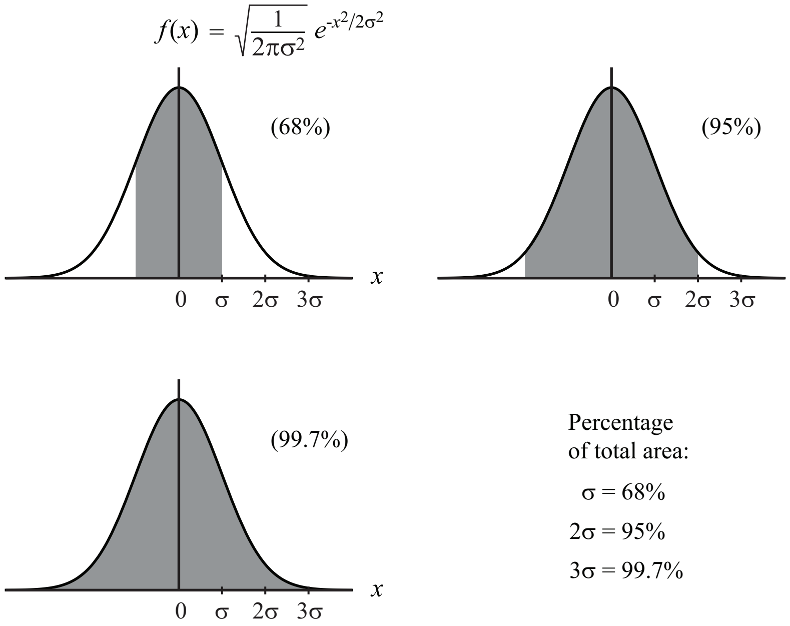


Figure 4.26: Areas under a Gaussian distribution within σ, 2σ, and 3σ from the mean.

Remarks:

1. The Gaussian distribution is a continuous one, because x can take on any value. The distribution applies (either exactly or approximately) to a nearly endless list of processes with continuous random variables such as length, time, light intensity, affinity for butternut squash, etc.

   We'll find in Sections 5.1 and 5.3 that the Gaussian distribution is a good approximation to the binomial and Poisson distributions if the numbers involved are large. In these cases, only integer values of x are relevant, so the distribution is effectively discrete. You can still draw the continuous curve described by Eq. (4.42), but it is relevant only for integer values of x.

2. We mentioned near the beginning of this section that the value of the prefactor in the expressions in Eq. (4.42) makes the total area under the distribution curve be equal to 1. Problem 4.22 gives a proof of this, but for now we can at least present an argument that explains why the prefactor must be proportional to $1/\sigma$ (or equivalently, to $\sqrt{b}$). Basically, since the width of the curve is proportional to σ (as we showed above), the height must be proportional to $1/\sigma$. This is true because if you increase σ by a factor of, say, 10 and thereby stretch the curve by a factor of 10 in the horizontal direction, then you also have to squash the curve by a factor of 10 in the vertical direction, if you want to keep the area the same. (See the fifth remark on page 206.) A factor of $1/\sigma$ in the prefactor accomplishes this. But note that this reasoning tells us only that

the prefactor is *proportional* to $1/\sigma$, and not what the constant of proportionality is. It happens to be $1/\sqrt{2\pi}$.

3. Two parameters are needed to describe the Gaussian distribution: μ and σ (or μ and b). This should be contrasted with the Poisson distribution, where only one parameter, a, is needed. Similarly, the exponential distribution depends on only the one parameter λ (or τ). In the Poisson case, not only does the width determine the height, but it also determines the location of the bump. In contrast, the Gaussian mean μ need not have anything to do with σ (or b). ♣

4.9 Summary

In this chapter we learned about probability distributions. In particular, we learned:

- A *probability distribution* is the collective information about how the total probability (which is always 1) is distributed among the various possible outcomes of the random variable.

- A probability distribution for a continuous random variable is given in terms of a *probability density.* To obtain an actual probability, the density must be multiplied by an interval of the random variable. More generally, the probability equals the area under the density curve.

We discussed six specific probability distributions:

- 1. *Uniform:* (Continuous) The probability density is uniform over a given span of random-variable values, and zero otherwise. The uniform distribution can be described by two parameters: the mean and the width, or alternatively the endpoints of the nonzero region. These two parameters then determine the height.

- 2. *Bernoulli:* (Discrete) The random variable can take on only two values, 1 and 0, with probabilities p and $1 - p$. An example with $p = 1/2$ is a coin toss with Heads = 1 and Tails = 0. The Bernoulli distribution is described by one parameter: p.

- 3. *Binomial:* (Discrete) The random variable is the number k of successes in a collection of n Bernoulli processes. An example is the total number of Heads in n coin tosses. The distribution takes the form,

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.49}$$

  The number k of successes must be an integer, of course. The binomial distribution is described by two parameters: n and p.

- 4. *Exponential:* (Continuous) This is the probability distribution for the waiting time t until the next event, for a completely random process. We derived this by taking the continuum limit of the analogous discrete result, which was the *geometric distribution* given in Eq. (4.14). The exponential distribution takes the form,

$$\rho(t) = \frac{e^{-t/\tau}}{\tau}, \tag{4.50}$$

where τ is the average waiting time. Equivalently, $\rho(t) = \lambda e^{-\lambda t}$, where $\lambda = 1/\tau$ is the average rate at which the events happen. The exponential distribution is described by one parameter: τ (or λ).

- 5. *Poisson:* (Discrete) This is the probability distribution for the number of events that happen in a given region (of time, space, etc.), for a completely random process. We derived this by taking the continuum limit of the analogous discrete result, which was simply the binomial distribution. The Poisson distribution takes the form,

$$P(k) = \frac{a^k e^{-a}}{k!}, \tag{4.51}$$

where a is the expected number of events in the given region. The number k of observed events must be an integer, of course. But a need not be. The Poisson distribution is described by one parameter: a.

- 6. *Gaussian:* (Continuous) This distribution takes the form,

$$f(x) = \sqrt{\frac{b}{\pi}}\, e^{-b(x-\mu)^2} \quad \text{or} \quad \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}. \tag{4.52}$$

σ is the *standard deviation* of the distribution. About 68% of the probability is contained in the range from $\mu - \sigma$ to $\mu + \sigma$. The width of the distribution is proportional to σ (and to $1/\sqrt{b}$). The Gaussian distribution is described by two parameters: μ and σ (or μ and b).

4.10 Exercises

See **www.people.fas.harvard.edu/~djmorin/book.html** for a supply of problems without included solutions.

4.11 Problems

Section 4.2: Continuous distributions

4.1. **Fahrenheit and Celsius** $*$

Fig. 4.4 shows the probability density for the temperature, with the temperature measured in Fahrenheit. Draw a reasonably accurate plot of the same probability density, but with the temperature measured in Celsius. (The conversion from Fahrenheit temperature F to Celsius temperature C is $C = (5/9)(F - 32)$. So it takes a ΔF of $9/5 = 1.8$ degrees to create a ΔC of 1 degree.)

4.2. **Expectation of a continuous distribution** $*$ *(calculus)*

The expectation value of a *discrete* random variable is given in Eq. (3.4). Given a *continuous* random variable with probability density $\rho(x)$, explain why the expectation value is given by the integral $\int x\rho(x)\,dx$.

Section 4.3: Uniform distribution

4.3. **Variance of the uniform distribution** $*$ *(calculus)*

Using the general idea from Problem 4.2, find the variance of a uniform distribution that extends from $x = 0$ to $x = a$.

Section 4.5: Binomial distribution

4.4. **Expectation of the binomial distribution** $**$

Use Eq. (3.4) to explicitly demonstrate that the expectation value of the binomial distribution in Eq. (4.6) equals pn. This must be true, of course, because a fraction p of the n trials yield success, on average, by the definition of p. *Hint*: The goal is to produce the result of pn, so try to factor a pn out of the sum in Eq. (3.4). You will eventually need to use an expression analogous to Eq. (4.10).

4.5. **Variance of the binomial distribution** $***$

As we saw in Problem 4.4, the expectation value of the binomial distribution is $\mu = pn$. Use the technique in either of the solutions to that problem to show that the variance of the binomial distribution is $np(1-p) \equiv npq$ (in agreement with Eq. (3.33)). *Hint*: The form of the variance in Eq. (3.34) works best. When finding the expectation value of k^2 (or really K^2, where K is the random variable whose value is k), it is easiest to find the expectation value of $k(k-1)$ and then add on the expectation value of k.

4.6. **Hypergeometric distribution** $***$

(a) A box contains N balls. K of them are red, and the other $N - K$ are blue. (K here is just a given number, not a random variable.) If you draw n balls *without replacement*, what is the probability of obtaining exactly k red balls? The resulting probability distribution is called the *hypergeometric distribution*.

(b) In the limit where N and K are very large, explain in words why the hypergeometric distribution reduces to the binomial distribution given in Eq. (4.6), with $p = K/N$. Then demonstrate this fact mathematically. What exactly is meant by "N and K are very large"?

Section 4.6: Exponential distribution

4.7. **Expectation of the geometric distribution** $**$

Verify that the expectation value of the geometric distribution in Eq. (4.14) equals $1/p$. (This is the waiting time we found by an easier method in Eq. (4.13).) The calculation involves a math trick, so you should do Problem 3.1 before solving this one.

4.8. **Properties of the exponential distribution** $**$ *(calculus)*

(a) By integrating the exponential distribution in Eq. (4.27) from $t = 0$ to $t = \infty$, show that the total probability is 1.

(b) What is the *median* value t? That is, for what value t_{med} are you equally likely to obtain a t value larger or smaller than t_{med}?

(c) By using the result from Problem 4.2, show that the expectation value is τ, as we know it must be.

(d) Again by using Problem 4.2, find the variance.

Section 4.7: Poisson distribution

4.9. **Total probability** $*$

Show that the sum of all of the probabilities in the Poisson distribution given in Eq. (4.40) equals 1, as we know it must. *Hint*: You will need to use Eq. (7.7) in Appendix B.

4.10. **Location of the maximum** $**$

For what (integer) value of k is the Poisson distribution $P(k)$ maximum?

4.11. **Value of the maximum** $*$

For large a, what approximately is the height of the bump in the Poisson $P(k)$ plot? You will need the result from the previous problem. *Hint*: You will also need to use Stirling's formula, given in Eq. (2.64) in Section 2.6.

4.12. **Expectation of the Poisson distribution** $**$

Use Eq. (3.4) to verify that the expectation value of the Poisson distribution equals a. This must be the case, of course, because a is defined to be the expected number of events in the given interval.

4.13. **Variance of the Poisson distribution** $**$

As we saw in Problem 4.12, the expectation value of the Poisson distribution is $\mu = a$. Use the technique in the solution to that problem to show that the variance of the Poisson distribution is a (which means that the standard deviation is $\sqrt{a}$). *Hint*: When finding the expectation value of k^2, it is easiest to find the expectation value of $k(k-1)$ and then add on the expectation value of k.

4.14. **Poisson accuracy** $**$

In the "balls in boxes, again" example on page 213, we saw that in the right plot in Fig. 4.20, the Poisson distribution is an excellent approximation to the exact binomial distribution. But in the left plot, it is only a so-so approximation. What parameter(s) determine how good the approximation is?

To answer this, we'll define the "goodness" of the approximation to be the ratio of the Poisson expression $P_{\text{P}}(k)$ in Eq. (4.40) to the exact binomial expression $P_{\text{B}}(k)$ in Eq. (4.32), with both functions evaluated at the expected value of k, namely $a = pn$, which we'll assume is an integer. (We're using Eq. (4.32) instead of Eq. (4.33), just because it's easier to work with. The expressions are equivalent, with $p \leftrightarrow 1/b$.) The closer the ratio $P_{\text{P}}(pn)/P_{\text{B}}(pn)$ is to 1, the better the Poisson approximation is. Calculate this ratio. You will need to use Stirling's formula, given in Eq. (2.64). You may assume that n is large (because otherwise there wouldn't be a need to use the Poisson approximation).

4.15. **Bump or no bump** $*$

In Fig. 4.21, we saw that $P(0) = P(1)$ when $a = 1$. (This is the cutoff between the distribution having or not having a bump.) Explain why this is consistent with what we noted about the binomial distribution (namely, that $P(0) = P(1)$ when $p = 1/(n + 1)$) in the example in Section 4.5.

4.16. **Typos** $*$

A hypothetical writer has an average of one typo per 50 pages of work. (Wishful thinking, perhaps!) What is the probability that there are no typos in a 350-page book?

4.17. **Boxes with zero balls** $*$

You randomly throw n balls into 1000 boxes and note the number of boxes that end up with zero balls in them. If you repeat this process a large number of times and observe that the average number of boxes with zero balls is 20, what is n?

4.18. **Twice the events** $**$

(a) Assume that on average, the events in a random process happen a times, where a is large, in a given time interval t. With the notation $P_a(k)$ representing the Poisson distribution, use Stirling's formula (given in Eq. (2.64)) to produce an approximate expression for the probability $P_a(a)$ that exactly a events happen during the time t.

(b) Consider the probability that exactly *twice* the number of events, $2a$, happen during *twice* the time, $2t$. What is the ratio of this probability to $P_a(a)$?

(c) Consider the probability that exactly *twice* the number of events, $2a$, happen during the *same* time, t. What is the ratio of this probability to $P_a(a)$?

4.19. **P(0) the hard way** $***$

For a Poisson process with a expected events, Eq. (4.40) gives the probability of having zero events as

$$P(0) = \frac{a^0 e^{-a}}{0!} = e^{-a} = 1 - \left(a - \frac{a^2}{2!} + \frac{a^3}{3!} - \cdots\right), \tag{4.53}$$

where we have used the Taylor series for e^x given in Eq. (7.7). With the above grouping of the terms, the sum in parentheses must be the probability of having *at least one* event, because when this is subtracted from 1, we obtain the probability of zero events. Explain why this is the case, by accounting for the various multiple events that can occur. You will want to look at the remark in the solution to Problem 2.3 first. The task here is to carry over that reasoning to a continuous Poisson process.

4.20. **Probability of at least 1** $**$

A million balls are thrown at random into a billion boxes. Consider a particular one of the boxes. What (approximately) is the probability that *at least one* ball ends up in that box? Solve this by:

(a) using the Poisson distribution in Eq. (4.40); you will need to use the approximation in Eq. (7.9),

(b) working with probabilities from scratch; you will need to use the approximation in Eq. (7.14).

Note that since the probability you found is very small, it is also approximately the probability of obtaining *exactly one* ball in the given box, because multiple events are extremely rare; see the discussion in the first remark in Section 4.6.2.

4.21. **Comparing probabilities** $***$

(a) A hypothetical 1000-sided die is rolled three times. What is the probability that a given number (say, 1) shows up all three times?

(b) A million balls are thrown at random into a billion boxes. (So from the result in Problem 4.20, the probability that exactly one ball ends up in a given box is approximately $1/1000$.) If this process (of throwing a million balls into a billion boxes) is performed three times, what (approximately) is the probability that exactly one ball lands in a given box all three times? (It can be a different ball each time.)

(c) A million balls are thrown at random into a billion boxes. This process is performed a *single* time. What (approximately) is the probability that exactly three balls end up in a given box? Solve this from scratch by using a counting argument.

(d) Solve part (c) by using the Poisson distribution.

(e) The setups in parts (b) and (c) might seem basically the same, because both setups involve three balls ending up in the given box, and there is a $1/b = 1/10^9$ probability that any given ball ends up in the given box. Give an intuitive explanation for why the answers differ.

Section 4.8: Gaussian distribution

4.22. **Area under a Gaussian curve** $**$ *(calculus)*

Show that the area (from $-\infty$ to ∞) under the Gaussian distribution, $f(x) = \sqrt{b/\pi}\, e^{-bx^2}$, equals 1. That is, show that the total probability equals 1. (We have set $\mu = 0$ for convenience, since μ doesn't affect the total area.) There is a very sneaky way to do this. But since it's completely out of the blue, we'll give a hint: Calculate the *square* of the desired integral by multiplying it by the integral of $\sqrt{b/\pi}\, e^{-by^2}$. Then make use of a change of variables from Cartesian to polar coordinates, to convert the Cartesian double integral into a polar double integral.

4.23. **Variance of the Gaussian distribution** $**$ *(calculus)*

Show that the variance of the second Gaussian expression in Eq. (4.42) equals σ^2 (which means that the standard deviation is σ). You may assume that $\mu = 0$ (because μ doesn't affect the variance), in which case the expression for the variance in Eq. (3.19) becomes $E(X^2)$. And then by the reasoning in Problem 4.2, this expectation value is $\int x^2 f(x)\, dx$. So the task of this problem is to evaluate this integral. The straightforward method is to use integration by parts.

4.12 Solutions

4.1. **Fahrenheit and Celsius**

A density is always given in terms of "something per something else." In the temperature example in Section 4.2, the "units" of probability density were probability per Fahrenheit degree. These units are equivalent to saying that we need to multiply the density by a certain number of Fahrenheit degrees (the ΔT) to obtain a probability; see Eq. (4.2). Analogously, we need to multiply a mass density (mass per volume) by a volume to obtain a mass.

If we want to instead write the probability density in terms of probability per *Celsius* degree, we can't simply use the same function $\rho(T)$ that appears in Fig. 4.4. Since there are 1.8 Fahrenheit degrees in each Celsius degree, the correct plot of $\rho(T)$ is shown in Fig. 4.27.

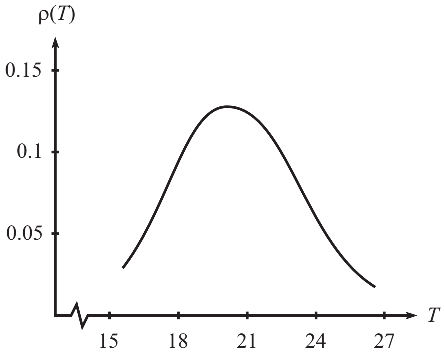


Figure 4.27: Expressing Fig. 4.4 in terms of Celsius instead of Fahrenheit.

This plot differs from Fig. 4.4 in three ways. First, since the peak of the curve in Fig. 4.4 was at about 68 degrees Fahrenheit, it is now *shifted* and located at about (5/9)(68 – 32) = 20 degrees Celsius in Fig. 4.27.

Second, compared with Fig. 4.4, the curve in Fig. 4.27 is *contracted* by a factor of 1.8 in the horizontal direction due to the conversion from Fahrenheit to Celsius. The span is only about 11 Celsius degrees, compared with a span of about 20 Fahrenheit degrees in Fig. 4.4. This follows from the fact that each Celsius degree is worth 1.8 Fahrenheit degrees.

Third, since the area under the entire curve in Fig. 4.27 must still be 1, the curve must also be *expanded* by a factor of 1.8 in the vertical direction. So the maximum value is about 0.13, compared with the maximum value of about 0.07 in Fig. 4.4.

Remark: These contraction and expansion countereffects cause the probabilities we calculate here to be consistent with ones we calculated in Section 4.2. For example, we found in Eq. (4.3) that the probability of the temperature falling between 70 °F and 71 °F is about 7%. Now, 70 °F and 71 °F correspond to 21.11 °C and 21.67 °C, as you can show using $C = (5/9)(F - 32)$. So the probability of the temperature falling between 21.11 °C and 21.67 °C had better also be 7%. It's the same temperature interval; we're just describing it in a different way. And indeed, from the Celsius plot, the value of the density near 21° is about 0.12. Therefore, the probability of falling between 21.11 °C and 21.67 °C, which equals the density times the interval, is $(0.12)(21.67 - 21.11) = 0.067 \approx 7\%$, in agreement with the Fahrenheit calculation (up to the rough readings we made from the plots). If we had forgotten to expand the height of the curve by the factor of 1.8 in Fig. 4.27, we would have obtained only about half of this probability, and therefore a different answer to exactly the same question (asked in a different language). That wouldn't be good. ♣

4.2. Expectation of a continuous distribution

For a general probability density $\rho(x)$, the probability associated with a span dx around a given value of x is $\rho(x)\,dx$; this is true by the definition of the probability density. Now, the expectation value of a *discrete* random variable is given in Eq. (3.4). To extract from this expression the expectation value of a *continuous* random variable, we can imagine dividing up the x axis into a very large number of little intervals dx. The probabilities p_i in Eq. (3.4) get replaced with the various $\rho(x)\,dx$ probabilities. And the outcomes x_i in Eq. (3.4) get replaced with the various values of x.

In making these replacements, we're pretending that all of the x values in a tiny interval dx are equal to the value at, say, the midpoint (call it x_i). This x_i then occurs with probability $p_i = \rho(x_i)\,dx$. We therefore have a discrete distribution that in the $dx \to 0$ limit is the same as the original continuous distribution. The discreteness of our approximate distribution allows us to apply Eq. (3.4) and say that the expectation value equals

$$\text{Expectation value} = \sum p_i x_i = \sum \big(\rho(x_i)\,dx\big)x_i. \qquad (4.54)$$

In the $dx \to 0$ limit, this discrete sum turns into the integral,

$$\text{Expectation value} = \int \Big(\rho(x)\,dx\Big)x = \int x\rho(x)\,dx, \qquad (4.55)$$

as desired. This is the general expression for the expectation value of a continuous random variable. The limits of the integral are technically $-\infty$ to ∞, although it is often the case that $\rho(x) = 0$ everywhere except in a finite region. For example, the density $\rho(t)$ for the exponential distribution is zero for $t < 0$, and it becomes negligibly small for $t \gg \tau$, where τ is the average waiting time.

The above result generalizes to the expectation value of things other than x. For example, the same reasoning shows that the expectation value of x^2 (which is relevant when calculating the variance) equals $\int x^2\rho(x)\,dx$. And the expectation value of x^7 (if you ever happened to be interested in such a quantity) equals $\int x^7\rho(x)\,dx$.

4.3. **Variance of the uniform distribution**

First solution: Since the nonzero part of the distribution has length a on the x axis, the value of the distribution in that region must be $1/a$, so that the total area is 1. We'll use the $E(X^2) - \mu^2$ form of the variance in Eq. (3.34), with $\mu = a/2$ here. Our task is therefore to calculate $E(X^2)$. From the last comment in the solution to Problem 4.2, this equals

$$E(X^2) = \int_0^a x^2\rho(x)\,dx = \int_0^a x^2 \cdot \frac{1}{a}\,dx = \frac{1}{a}\cdot\frac{x^3}{3}\bigg|_0^a = \frac{a^2}{3}. \tag{4.56}$$

The variance is then

$$\text{Var}(X) = E(X^2) - \mu^2 = \frac{a^2}{3} - \left(\frac{a}{2}\right)^2 = \frac{a^2}{12}. \tag{4.57}$$

The standard deviation is therefore $a/(2\sqrt{3}\,) \approx (0.29)a$.

Second solution: Let's shift the distribution so that it is nonzero from $x = -a/2$ to $x = a/2$. This shift doesn't affect the variance, which is now simply $E(X^2)$, because $\mu = 0$. So

$$\begin{aligned}\text{Var}(X) = E(X^2) &= \int_{-a/2}^{a/2} x^2\rho(x)\,dx = \frac{1}{a}\cdot\frac{x^3}{3}\bigg|_{-a/2}^{a/2} \\ &= \frac{1}{3a}\left(\left(\frac{a}{2}\right)^3 - \left(-\frac{a}{2}\right)^3\right) = \frac{a^2}{12}.\end{aligned} \tag{4.58}$$

Third solution: We can use the $E\big[(X-\mu)^2\big]$ form of the variance in Eq. (3.19), with the original $0 < x < a$ span. This gives

$$\begin{aligned}\text{Var}(X) \equiv E\big[(X - a/2)^2\big] &= \int_0^a (x - a/2)^2\rho(x)\,dx \\ &= \frac{1}{a}\int_0^a (x^2 - ax + a^2/4)\,dx \\ &= \frac{1}{a}\left(\frac{a^3}{3} - a\frac{a^2}{2} + \frac{a^2}{4}a\right) = \frac{a^2}{12}.\end{aligned} \tag{4.59}$$

4.4. **Expectation of the binomial distribution**

First solution: The $k = 0$ term doesn't contribute anything to the sum in Eq. (3.4), so we can start with the $k = 1$ term. The sum goes up to $k = n$. Plugging the probabilities from Eq. (4.6) into Eq. (3.4), we obtain an expectation value of

$$\sum_{k=1}^{n} k\cdot P(k) = \sum_{k=1}^{n} k\cdot\binom{n}{k}p^k(1-p)^{n-k}. \tag{4.60}$$

If the factor of k weren't on the righthand side, we would know how to evaluate this sum; see Eq. (4.10). So let's get rid of the k and create a sum that looks like Eq. (4.10).

The steps are the following.

$$\begin{aligned}
&\sum_{k=1}^{n} k \cdot P(k) \\
&= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad \text{(expanding the binomial coeff.)} \\
&= pn \sum_{k=1}^{n} k \cdot \frac{(n-1)!}{k!(n-k)!} p^{k-1} (1-p)^{n-k} \qquad \text{(factoring out } pn\text{)} \\
&= pn \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \qquad \text{(canceling the } k\text{)} \\
&= pn \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \qquad \text{(rewriting)} \\
&= pn \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j} (1-p)^{(n-1)-j} \qquad \text{(defining } j \equiv k-1\text{)} \\
&= pn(p + (1-p))^{n-1} \qquad \text{(using the binomial expansion)} \\
&= pn \cdot 1, \qquad (4.61)
\end{aligned}$$

as desired. Note that in the sixth line, the sum over j goes from 0 to $n-1$, because the sum over k went from 1 to n.

Even though we know that the expectation value has to be pn (as mentioned in the statement of the problem), it's nice to see that the math does in fact work out.

Second solution: Here is another (sneaky) way to obtain the expectation value. This method uses calculus. The binomial expansion tells us that

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}. \qquad (4.62)$$

This relation is identically true for arbitrary values of p (and q), so we can take the derivative with respect to p to obtain another valid relation:

$$n(p+q)^{n-1} = \sum_{k=1}^{n} k \binom{n}{k} p^{k-1} q^{n-k}. \qquad (4.63)$$

If we now multiply both sides by p and then set q to equal $1-p$ (the relation is true for all values of q, in particular this specific one), we obtain

$$np(1)^{n-1} = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \implies np = \sum_{k=1}^{n} k \cdot P(k), \qquad (4.64)$$

as desired.

4.5. **Variance of the binomial distribution**

First solution: As suggested in the statement of the problem, let's find the expectation value of $k(k-1)$. Since we've already done a calculation like this in Problem 4.4, we won't list out every step here as we did in Eq. (4.61). The $k = 0$ and $k - 1$ terms don't contribute anything to the expectation value of $k(k-1)$, so we can start the sum with the $k=2$ term. We have (with $j \equiv k-2$ in the 5th line)

$$\begin{aligned}
&\sum_{k=2}^{n} k(k-1)\cdot P(k) \\
&= \sum_{k=2}^{n} k(k-1)\cdot \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&= p^2 n(n-1)\sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!}p^{k-2}(1-p)^{n-k} \\
&= p^2 n(n-1)\sum_{k=2}^{n} \binom{n-2}{k-2}p^{k-2}(1-p)^{(n-2)-(k-2)} \\
&= p^2 n(n-1)\sum_{j=0}^{n-2} \binom{n-2}{j}p^{j}(1-p)^{(n-2)-j} \\
&= p^2 n(n-1)\big(p+(1-p)\big)^{n-2} \\
&= p^2 n(n-1)\cdot 1. \qquad (4.65)
\end{aligned}$$

The expectation value of k^2 equals the expectation value of $k(k-1)$ plus the expectation value of k. The latter is just pn, from Problem 4.4. So the expectation value of k^2 is

$$p^2 n(n-1) + pn. \qquad (4.66)$$

To obtain the variance, Eq. (3.34) tells us that we need to subtract off $\mu^2 = (pn)^2$ from this result. The variance is therefore

$$\begin{aligned}
\left(p^2 n(n-1) + pn\right) - p^2n^2 &= \left(p^2n^2 - p^2 n + pn\right) - p^2n^2 \\
&= pn(1-p) \\
&\equiv npq, \qquad (4.67)
\end{aligned}$$

as desired. The standard deviation is then $\sqrt{npq}$.

Second solution: Instead of taking just one derivative, as we did in the second solution in Problem 4.4, we'll take two derivatives here. Starting with the binomial expansion,

$$(p+q)^n = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}, \qquad (4.68)$$

we can take two derivatives with respect to p to obtain

$$n(n-1)(p+q)^{n-2} = \sum_{k=2}^{n} k(k-1)\binom{n}{k}p^{k-2}q^{n-k}. \qquad (4.69)$$

If we now multiply both sides by p^2 and then set q to equal $1-p$, we obtain

$$\begin{aligned}
p^2 n(n-1)(1)^{n-1} &= \sum_{k=2}^{n} k(k-1)\binom{n}{k}p^{k}(1-p)^{n-k} \\
\Longrightarrow\quad p^2 n(n-1) &= \sum_{k=2}^{n} k(k-1)\cdot P(k). \qquad (4.70)
\end{aligned}$$

The expectation value of $k(k-1)$ is therefore $p^2n(n-1)$, in agreement with Eq. (4.65) in the first solution. The solution proceeds as above.

4.6. **Hypergeometric distribution**

(a) There are $\binom{N}{n}$ possible sets of n balls (drawn without replacement), and all of these sets are equally likely to be drawn. We therefore simply need to count the number of sets that have exactly k red balls. There are $\binom{K}{k}$ ways to choose k red balls from the K red balls in the box. And there are $\binom{N-K}{n-k}$ ways to choose the other $n-k$ balls (which we want to be blue) from the $N-K$ blue balls in the box. So the number of sets that have exactly k red balls is $\binom{K}{k}\binom{N-K}{n-k}$. The desired probability of obtaining exactly k balls when drawing n balls without replacement is therefore

$$P(k) = \frac{\dbinom{K}{k}\dbinom{N-K}{n-k}}{\dbinom{N}{n}} \qquad \text{(Hypergeometric distribution)} \qquad (4.71)$$

Remark: Since the number of red balls, k, that you draw can't be larger than either K or n, we see that $P(k)$ is nonzero only if $k \le \min(K,n)$. Likewise, the number of blue balls, $n-k$, that you draw can't be larger than either $N-K$ or n. So $P(k)$ is nonzero only if $n-k \le \min(N-K,n) \Longrightarrow n - \min(N-K,n) \le k$. Putting these inequalities together, we see that $P(k)$ is nonzero only if

$$n - \min(N-K,n) \le k \le \min(K,n). \qquad (4.72)$$

If both K and $N-K$ are larger than n, then this reduces to the simple relation, $0 \le k \le n$. ♣

(b) If N and K are small, then the probabilities of drawing red/blue balls change after each draw. This is true because you aren't replacing the balls, so the ratio of red and blue balls changes after each draw.

However, if N and K are large, then the "without replacement" qualifier is inconsequential. The ratio of red and blue balls remains essentially unchanged after each draw. Removing one red ball from a set of a million red balls is hardly noticeable. The probability of drawing a red ball at each stage therefore remains essentially fixed at the value K/N. Likewise, the probability of drawing a blue ball at each stage remains essentially fixed at the value $(N-K)/N$. If we define the red-ball probability as $p \equiv K/N$, then the blue-ball probability is $1-p$. We therefore have exactly the setup that generates the binomial distribution, with red corresponding to success, and blue corresponding to failure. Hence we obtain the binomial distribution in Eq. (4.6).

Let's now show mathematically that the hypergeometric distribution in Eq. (4.71) reduces to the binomial distribution in Eq. (4.6). Expanding the binomial coefficients in Eq. (4.71) gives

$$P(k) = \frac{\dfrac{K!}{k!(K-k)!} \cdot \dfrac{(N-K)!}{(n-k)!((N-K)-(n-k))!}}{\dfrac{N!}{n!(N-n)!}}. \qquad (4.73)$$

If $K \gg k$, then we can say that

$$\frac{K!}{(K-k)!} = K(K-1)(K-2)\cdots(K-k+1) \approx K^k. \qquad (4.74)$$

This is true because all of the factors here are essentially equal to K, in a multiplicative sense. (The "$\gg$" sign in $K \gg k$ means "much greater than" in a multiplicative, not additive, sense.) We can make similar approximations to $(N-K)!/((N-K)-(n-k))!$ and $N!/(N-n)!$, so Eq. (4.73) becomes

$$P(k) \approx \frac{\dfrac{K^k}{k!}\cdot\dfrac{(N-K)^{n-k}}{(n-k)!}}{\dfrac{N^n}{n!}} = \frac{n!}{k!(n-k)!}\left(\frac{K}{N}\right)^k\left(\frac{N-K}{N}\right)^{n-k}$$
$$= \binom{n}{k}p^k(1-p)^{n-k}, \qquad (4.75)$$

where $p \equiv K/N$. This is the desired binomial distribution, which gives the probability of k successes in n trials, where the probability of success in each trial takes on the fixed value of p.

We made three approximations in the above calculation, and they relied on the three assumptions,

$$(1)\ K \gg k, \quad (2)\ N-K \gg n-k, \quad (3)\ N \gg n. \qquad (4.76)$$

In words, these three assumptions are: (1) the number of red balls you draw is much smaller than the total number of red balls in the box, (2) the number of blue balls you draw is much smaller than the total number of blue balls in the box, and (3) the total number of balls you draw is much smaller than the total number of balls in the box. (The third assumption follows from the other two.) These three assumptions are what we mean by "N and K are very large."

4.7. **Expectation of the geometric distribution**

From Eq. (4.14), the probability that we need to wait just one iteration for the next success is p. For two iterations it is $(1-p)p$, for three iterations it is $(1-p)^2p$, and so on. The expectation value of the number of iterations (that is, the waiting time) is therefore

$$1\cdot p + 2\cdot(1-p)p + 3\cdot(1-p)^2p + 4\cdot(1-p)^3p + \cdots. \qquad (4.77)$$

To calculate this sum, we'll use the trick we introduced in Problem 3.1 and write the sum as a geometric series starting with p, plus another geometric series starting with $(1-p)p$, and so on. And we'll use the fact that the sum of a geometric series with first term a and ratio r is $a/(1-r)$. The expectation value in Eq. (4.77) then becomes

$$\begin{array}{r} p + (1-p)p + (1-p)^2p + (1-p)^3p + \cdots \\ (1-p)p + (1-p)^2p + (1-p)^3p + \cdots \\ (1-p)^2p + (1-p)^3p + \cdots \\ (1-p)^3p + \cdots \\ \vdots \end{array} \qquad (4.78)$$

This has the correct number of each type of term. For example, the $(1-p)^2p$ term appears three times. The first line above is a geometric series that sums to $a/(1-r) = p/(1-(1-p)) = 1$. The second line is also a geometric series, and it sums to $(1-p)p/(1-(1-p)) = 1-p$. Likewise the third line sums to $(1-p)^2p/(1-(1-p)) = (1-p)^2$. And so on. The sum of the infinite number of lines in Eq. (4.79) therefore equals

$$1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots. \qquad (4.79)$$

But this itself is a geometric series, and it sums to $a/(1-r) = 1/(1-(1-p)) = 1/p$, as desired.

4.8. **Properties of the exponential distribution**

(a) The total probability equals the total area under the distribution curve. And this area is given by the integral of the distribution. The integral of $e^{-t/\tau}/\tau$ equals $-e^{-t/\tau}$, as you can verify by taking the derivative (and using the chain rule). The desired integral is therefore

$$\int_0^\infty \frac{e^{-t/\tau}}{\tau}\,dt = -e^{-t/\tau}\Big|_0^\infty = -e^{-\infty} - (-e^{-0}) = 1, \tag{4.80}$$

as desired

(b) This is very similar to part (a), except that we now want the probability from 0 to t_{med} to equal 1/2. That is,

$$\frac{1}{2} = \int_0^{t_{\text{med}}} \frac{e^{-t/\tau}}{\tau}\,dt = -e^{-t/\tau}\Big|_0^{t_{\text{med}}} = -e^{-t_{\text{med}}/\tau} - (-e^{-0}). \tag{4.81}$$

This yields $e^{-t_{\text{med}}/\tau} = 1/2$. Taking the natural log of both sides then gives

$$-t_{\text{med}}/\tau = -\ln 2 \implies t_{\text{med}} = (\ln 2)\tau \approx (0.7)\tau. \tag{4.82}$$

So the median value of t is $(0.7)\tau$. In other words, $(0.7)\tau$ is the value of t for which the two shaded areas in Fig. 4.28 are equal.

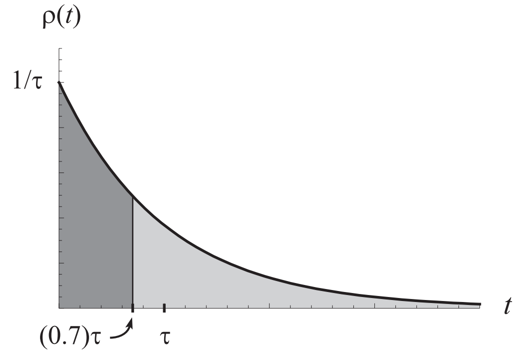


Figure 4.28: The areas on either side of the median are equal.

Note that the *median* value of t, namely $(0.7)\tau$, is *smaller* than the *mean* value (the expectation value) of t, namely τ. The reason for this is that the exponential distribution has a tail that extends to large values of t. These values of t drag the mean to the right, more so than the small values of t near zero drag it to the left (because the former are generally farther from t_{med} than the latter). Whenever you have an asymmetric distribution like this, the mean always lies on the "tail side" of the median.

(c) In the specific case of the exponential distribution, Eq. (4.55) in the solution to Problem 4.2 gives

$$\text{Expectation value} = \int_0^\infty t \cdot \frac{e^{-t/\tau}}{\tau}\,dt. \tag{4.83}$$

You can evaluate this integral by performing "integration by parts," or you can just look it up. It turns out to be

$$\begin{aligned}\text{Expectation value} &= -(t+\tau)e^{-t/\tau}\Big|_0^\infty \\ &= -(\infty+\tau)e^{-\infty} + (0+\tau)e^{-0} \\ &= 0+\tau, \end{aligned} \tag{4.84}$$

as desired. In the first term here, we have used the fact that the smallness of $e^{-\infty}$ wins out over the largeness of ∞. You can check this on your calculator by replacing ∞ with, say, 100.

(d) Let's use T to denote the random variable whose value is t. Since the mean of the exponential distribution is τ, Eq. (3.34) tells us that the variance is $E(T^2) - \tau^2$. So we need to find $E(T^2)$. Eq. (4.55) gives

$$E(T^2) = \int_0^\infty t^2 \cdot \frac{e^{-t/\tau}}{\tau}\,dt. \tag{4.85}$$

Evaluating this by integration by parts is rather messy, so let's just look up the integral. It turns out to be

$$\begin{aligned} E(T^2) &= -(t^2+2\tau t+2\tau^2)e^{-t/\tau}\Big|_0^\infty \\ &= -0 + (0+0+2\tau^2)e^{-0} \\ &= 2\tau^2. \end{aligned} \tag{4.86}$$

As in part (c), we have used the fact that the smallness of $e^{-\infty}$ makes the term associated with the upper limit of integration be zero. The variance is therefore

$$\mathrm{Var}(T) = E(T^2) - \tau^2 = 2\tau^2 - \tau^2 = \tau^2. \tag{4.87}$$

The standard deviation is the square root of the variance, so it is simply τ, which interestingly is the same as the mean. As with all other quantities associated with the exponential distribution, the variance and standard deviation depend only on τ, because that is the only parameter that appears in the distribution.

4.9. **Total probability**

The sum over k ranges from 0 to ∞. The upper limit is ∞ because with continuous time (or space, or whatever), theoretically an arbitrarily large number of events can occur in a given time interval (although if k is much larger than a, then $P(k)$ is negligibly small). We have (invoking Eq. (7.7) from Appendix B to obtain the third line)

$$\begin{aligned} \sum_{k=0}^{\infty} P(k) &= \sum_{k=0}^{\infty} \frac{a^k e^{-a}}{k!} \\ &= e^{-a}\sum_{k=0}^{\infty}\frac{a^k}{k!} \\ &= e^{-a}e^{a} \\ &= 1, \end{aligned} \tag{4.88}$$

as desired. You are encouraged to look at the derivation of Eq. (7.7) in Appendix B.

4.10. **Location of the maximum**

First solution: In this solution we'll use the fact that the expression for $P(k)$ in Eq. (4.40) is actually valid for all positive values of k, not just integers (even though we're really only concerned with integers). This is due to the fact that it is possible to extend the meaning of $k!$ to non-integers. We can therefore treat $P(k)$ as a continuous distribution. The maximum value of this distribution might not occur at an integer value of k, but we'll be able to extract the appropriate value of k that yields the maximum when k is restricted to integers.

A convenient way to narrow down the location of the maximum of $P(k)$ is to set $P(k) = P(k+1)$. (In calculus, this is equivalent to finding the maximum by setting the first derivative equal to zero.) This will tell us roughly where the maximum is, because this relation can hold only if k and $k+1$ are on opposite sides of the maximum. This is true because the relation $P(k) = P(k+1)$ can't be valid on the right side of the curve's peak, because the curve is decreasing there, so all those points have $P(k) > P(k+1)$. Similarly, all the points on the left side of the peak have $P(k) < P(k+1)$. The only remaining possibility is that k is on the left side and $k+1$ is on the right side. That is, they are on opposite sides of the maximum.

Setting $P(k) = P(k+1)$ gives (after canceling many common factors to obtain the second line)

$$\begin{aligned} P(k) = P(k+1) \quad &\Longrightarrow \quad \frac{a^k e^{-a}}{k!} = \frac{a^{k+1} e^{-a}}{(k+1)!} \\ &\Longrightarrow \quad \frac{1}{1} = \frac{a}{k+1} \\ &\Longrightarrow \quad k+1 = a \\ &\Longrightarrow \quad k = a-1. \end{aligned} \tag{4.89}$$

The two relevant points on either side of the maximum, namely k and $k+1$, are therefore $a-1$ and a. So the maximum of the $P(k)$ plot (extended to non-integers) lies between $a-1$ and a. Since we're actually concerned only with integer values of k, the maximum is located at the integer that lies between $a-1$ and a (or at both of these values if a is an integer). In situations where a is large (which is often the case), the distinction between $a-1$ and a (or somewhere in between) isn't all that important, so we generally just say that the maximum of the probability distribution occurs roughly at a.

Second solution: We can avoid any issues about extending the Poisson distribution to non-integer values of k, by simply finding the integer value of k for which both $P(k) \geq P(k+1)$ and $P(k) \geq P(k-1)$ hold. $P(k)$ is then the maximum, because it is at least as large as the two adjacent $P(k \pm 1)$ values.

By changing the "=" sign in Eq. (4.89) to a "≥" sign, we immediately see that $P(k) \geq P(k+1)$ implies $k \geq a-1$. And by slightly modifying Eq. (4.89), you can show that $P(k) \geq P(k-1)$ implies $a \geq k$. Combining these two results, we see that the integer value of k that yields the maximum $P(k)$ satisfies $a-1 \leq k \leq a$. The desired value of k is therefore the integer that lies between $a-1$ and a (or at both of these values if a is an integer), as we found in the first solution.

4.11. **Value of the maximum**

Since we know from the previous problem that the maximum of $P(k)$ occurs essentially at $k = a$, our goal is to find $P(a)$. Stirling's formula allows us to make a quick approximation to this value. Plugging $k = a$ into Eq. (4.40) and using Stirling's formula, $n! \approx n^n e^{-n}\sqrt{2\pi n}$, yields

$$P(a) = \frac{a^a e^{-a}}{a!} \approx \frac{a^a e^{-a}}{a^a e^{-a}\sqrt{2\pi a}} = \frac{1}{\sqrt{2\pi a}}. \qquad (4.90)$$

We see that the height is proportional to $1/\sqrt{a}$. So if a goes up by a factor of, say, 4, then the height goes down by a factor of 2.

It is easy to make quick estimates using this result. Consider the $a = 10$ plot in Fig. 4.21. The maximum is between 0.1 and 0.2, a little closer to 0.1. Let's say 0.13. And indeed, if $a = 10$ (for which Stirling's formula is quite accurate, from Table 2.6), Eq. (4.90) gives

$$P(10) \approx \frac{1}{\sqrt{2\pi(10)}} \approx 0.126. \qquad (4.91)$$

This is very close to the exact value of $P(10)$, which you can show is about 0.125. Since a is an integer here (namely, 10), Problem 4.10 tells us that $P(9)$ takes on this same value.

4.12. **Expectation of the Poisson distribution**

The expectation value is the sum of $k \cdot P(k)$, from $k = 0$ to $k = \infty$. However, the $k = 0$ term contributes nothing, so we can start the sum with the $k = 1$ term. Using Eq. (4.40), the expectation value is therefore

$$\begin{aligned}
\sum_{k=1}^{\infty} k \cdot P(k) &= \sum_{k=1}^{\infty} k \cdot \frac{a^k e^{-a}}{k!} \\
&= \sum_{k=1}^{\infty} \frac{a^k e^{-a}}{(k-1)!} \qquad \text{(canceling the } k\text{)} \\
&= a \cdot \sum_{k=1}^{\infty} \frac{a^{k-1} e^{-a}}{(k-1)!} \qquad \text{(factoring out an } a\text{)} \\
&= a \cdot \sum_{j=0}^{\infty} \frac{a^j e^{-a}}{j!} \qquad \text{(defining } j \equiv k-1\text{)} \\
&= a \cdot \sum_{j=0}^{\infty} P(j) \qquad \text{(using Eq. (4.40))} \\
&= a \cdot 1, \qquad \text{(total probability is 1)} \qquad (4.92)
\end{aligned}$$

as desired. In the fourth line, we used the fact that since $j \equiv k - 1$, the sum over j starts with the $j = 0$ term, because the sum over k started with the $k = 1$ term. If you want to show explicitly that the total probability is 1, that was the task of Problem 4.9.

4.13. **Variance of the Poisson distribution**

As suggested in the statement of the problem, let's find the expectation value of $k(k - 1)$. Since we've already done a calculation like this in Problem 4.12, we won't list out every step here as we did in Eq. (4.92). The $k = 0$ and $k - 1$ terms don't contribute anything to the expectation value of $k(k - 1)$, so we can start the sum with the $k = 2$

term. We have (with $j \equiv k-2$ in the 3rd line)

$$\begin{aligned}\sum_{k=2}^{\infty} k(k-1)\cdot P(k) &= \sum_{k=2}^{\infty} k(k-1)\cdot \frac{a^k e^{-a}}{k!}\\ &= a^2\cdot\sum_{k=2}^{\infty}\frac{a^{k-2}e^{-a}}{(k-2)!}\\ &= a^2\cdot\sum_{j=0}^{\infty}\frac{a^j e^{-a}}{j!}\\ &= a^2\cdot\sum_{j=0}^{\infty}P(j)\\ &= a^2\cdot 1. \end{aligned}\tag{4.93}$$

The expectation value of k^2 equals the expectation value of $k(k-1)$ plus the expectation value of k. The latter is just a, from Problem 4.12. So the expectation value of k^2 is a^2+a. To obtain the variance, Eq. (3.34) tells us that we need to subtract off $\mu^2 = a^2$ from this result. The variance is therefore

$$(a^2+a)-a^2 = a,\tag{4.94}$$

as desired. The standard deviation is then $\sqrt{a}$.

We will show in Section 5.3 that the standard deviation of the Poisson distribution equals $\sqrt{a}$ when a is large (when the Poisson looks like a Gaussian). But in this problem we demonstrated the stronger result that the standard deviation of the Poisson distribution equals $\sqrt{a}$ for *any* value of a, even a small one (when the Poisson *doesn't* look like a Gaussian).

Remark: We saw in Problem 4.11 that for large a, the height of the bump in the Poisson $P(k)$ plot is $1/\sqrt{2\pi a}$, which is proportional to $1/\sqrt{a}$. The present $\sigma = \sqrt{a}$ result is consistent with this, because we know that the total probability must be 1. For large a, the $P(k)$ plot is essentially a continuous curve, so we need the total area under the curve to equal 1. A rough measure of the width of the bump is 2σ. The area under the curve equals (roughly) this width times the height. The product of 2σ and the height must therefore be of order 1. And this is indeed the case, because $\sigma = \sqrt{a}$ implies that $(2\sigma)(1/\sqrt{2\pi a}) = \sqrt{2/\pi}$, which is of order 1. This order-of-magnitude argument doesn't tell us anything about specific numerical factors, but it does tell us that the height and standard deviation must have inverse dependences on a. ♣

4.14. **Poisson accuracy**

Replacing a with pn in the Poisson distribution in Eq. (4.40), and setting $k = pn$ as instructed, gives

$$P_{\text{P}}(pn) = \frac{(pn)^{pn}e^{-pn}}{(pn)!}.\tag{4.95}$$

Similarly, setting $k = pn$ in the exact binomial expression in Eq. (4.32) gives

$$\begin{aligned}P_{\text{B}}(pn) &= \binom{n}{pn}p^{pn}(1-p)^{n-pn}\\ &= \frac{n!}{(pn)!(n-pn)!}p^{pn}(1-p)^{n-pn}.\end{aligned}\tag{4.96}$$

The $(pn)!$ term here matches up with the $(pn)!$ term in $P_\text{P}(pn)$, so it will cancel in the ratio $P_\text{P}(pn)/P_\text{B}(pn)$. Let's apply Stirling's formula, $m! \approx m^m e^{-m}\sqrt{2\pi m}$, to the other two factorials in $P_\text{B}(pn)$. Since $n - pn = n(1-p)$, we obtain (we'll do the simplification gradually here)

$$\begin{aligned}
P_\text{B}(pn) &\approx \frac{n^n e^{-n}\sqrt{2\pi n}}{(pn)! \cdot \big(n(1-p)\big)^{n(1-p)} e^{-n(1-p)}\sqrt{2\pi n(1-p)}} \cdot p^{pn}(1-p)^{n(1-p)} \\
&= \frac{n^n e^{-n}}{(pn)! \cdot n^{n(1-p)} e^{-n(1-p)}\sqrt{1-p}} \cdot p^{pn} \\
&= \frac{1}{(pn)! \cdot n^{-pn} e^{pn}\sqrt{1-p}} \cdot p^{pn} \\
&= \frac{1}{\sqrt{1-p}} \cdot \frac{(pn)^{pn} e^{-pn}}{(pn)!}\,. \qquad (4.97)
\end{aligned}$$

This result fortuitously takes the same form as the $P_\text{P}(pn)$ expression in Eq. (4.95), except for the factor of $1/\sqrt{1-p}$ out front. The desired ratio is therefore simply

$$\frac{P_\text{P}(pn)}{P_\text{B}(pn)} = \sqrt{1-p}\,. \qquad (4.98)$$

This is the factor by which the peak of the Poisson plot is smaller than the peak of the (exact) binomial plot.

In the two plots in Fig. 4.20, the p values are $1/10$ and $1/100$, so the $\sqrt{1-p}$ ratios are $\sqrt{0.9} \approx 0.95$ and $\sqrt{0.99} \approx 0.995$. These correspond to percentage differences of 5% and 0.5%, or equivalently to fractional differences of $1/20$ and $1/200$. These are consistent with a visual inspection of the two plots; the 0.5% difference is too small to see in the second plot.

With the above $\sqrt{1-p}$ result, we can say that the Poisson approximation is a good one if $\sqrt{1-p}$ is close to 1, or equivalently if p is much smaller than 1. How much smaller? That depends on how good an approximation you want. If you want accuracy to 1%, then $p = 1/100$ works, but $p = 1/10$ doesn't.

Remarks:

1. A helpful mathematical relation that is valid for small p is $\sqrt{1-p} \approx 1 - p/2$. (You can check this by plugging a small number like $p = 0.01$ into your calculator. Or you can square both sides to obtain $1 - p \approx 1 - p + p^2/4$, which is correct up to the quadratic $p^2/4$ difference, which is very small if p is small.) With this relation, our $\sqrt{1-p}$ result becomes $1 - p/2$. The difference between this result and 1 is therefore $p/2$. This makes it clear why we ended up with the above ratios of 0.95 and 0.995 for $p = 1/10$ and $p = 1/100$.
2. Note that our "goodness" condition for the Poisson approximation involves only p. That is, it is independent of n. This isn't terribly obvious. For a given value of p (say, $p = 1/100$), we will obtain the same accuracy whether n is, say, 10^3 or 10^5. Of course, the $a = pn$ expected values in these two cases are different (10 and 1000). But the ratio of $P_\text{P}(pn)$ to $P_\text{B}(pn)$ is the same (at least in the Stirling approximation).
3. In the language of balls and boxes, since $p = 1/b$, the $p \ll 1$ condition is equivalent to saying that the number of boxes satisfies $b \gg 1$. So the more boxes there are, the better the approximation. This condition is independent of the number n of balls (as long as n is large).

4. The result in Eq. (4.98) is valid even if the expected number of events pn is small, for example, 1 or 2. The is true because the $(pn)!$ terms cancel in the ratio of Eqs. (4.95) and (4.96), so we don't need to worry about applying Stirling's formula to a small number. The other two factorials, $n!$ and $(n - pn)!$, are large because we are assuming that n is large. ♣

4.15. **Bump or no bump**

We saw in Section 4.7.2 that the Poisson distribution is obtained by taking the $n \to \infty$ and $p \to 0$ limits of the binomial distribution (p took the form of $\lambda\epsilon$ in the derivation in Section 4.7.2). But in the $n \to \infty$ limit, the $p = 1/(n+1)$ condition for $P(0) = P(1)$ in the binomial case becomes $p \approx 1/n$. So $pn \approx 1$. But pn is just the average number of events a in the Poisson distribution. So $a \approx 1$ is the condition for $P(0) = P(1)$ in the Poisson case, as desired.

4.16. **Typos**

First solution: Under the assumption that the typos occur randomly, the given setup calls for the Poisson distribution. If the expected number of typos in 50 pages is one, then the expected number of typos in a 350-page book is $a = 350/50 = 7$. So Eq. (4.40) gives the probability of zero typos in the book as

$$P(0) = \frac{a^0 e^{-a}}{0!} = e^{-a} = e^{-7} \approx 9 \cdot 10^{-4} \approx 0.1\%. \tag{4.99}$$

Second solution: (This is an approximate solution.) If there is one typo per 50 pages, then the expected number of typos per page is $1/50$. This implies that the probability that a given page has at least one typo is approximately 2%, which means that the probability that a given page has *zero* typos is approximately 98%. We are using the word "approximately" here, because the probability of zero typos on a given page must in fact be slightly larger than 98%. This is true because if it were exactly 98%, then in the 2% of the pages where a typo occurs, there might actually be two (or three, etc.) typos. Although these occurrences are rare in the present setup, they will nevertheless cause the expected number of typos per page to be (slightly) larger than $1/50$, in contradiction to the given assumption. The actual probability of having zero typos per page must therefore be slightly larger than 98%, so that slightly fewer than 2% of the pages have at least one typo.

However, if we work in the (reasonable) approximation where the probability of having zero typos per page equals 0.98, then the probability of having zero typos in 350 pages equals $(0.98)^{350} = 8.5 \cdot 10^{-4}$. This is close to the correct probability of $9 \cdot 10^{-4}$ in Eq. (4.99). Replacing 0.98 with a slightly larger number would yield the correct probability of $9 \cdot 10^{-4}$.

Remarks: What should the probability of 0.98 (for zero typos on a given page) be increased to, if we want to obtain the correct probability of $9 \cdot 10^{-4}$ (for zero typos in the book)? Since the expected number of typos per page is $1/50$, we simply need to plug $a = 1/50$ into the Poisson expression for $P(0)$. This gives the true probability of having zero typos on a given page as

$$P(0) = \frac{a^0 e^{-a}}{0!} = e^{-a} = e^{-1/50} = 0.9802. \tag{4.100}$$

As expected, this is only a tiny bit larger than the approximate value of 0.98 that we used above. If we use the new (and correct) value of 0.9802, the result of our second solution is modified to $(0.9802)^{350} = 9 \cdot 10^{-4}$, which agrees with the correct answer in Eq. (4.99).

The relation between the (approximate) second solution and the (correct) first solution can be seen by writing our approximate answer of $(0.98)^{350}$ as

$$(0.98)^{350} = \left(1 - \frac{1}{50}\right)^{350} = \left(\left(1 - \frac{1}{50}\right)^{50}\right)^{7} \approx (e^{-1})^7 = e^{-7}, \qquad (4.101)$$

which is the correct answer in Eq. (4.99). We have used the approximation in Eq. (7.4) to produce the e^{-1} term here. ♣

4.17. **Boxes with zero balls**

First solution: The given information that 20 out of the 1000 boxes contain zero balls (on average) tells us that the probability that a given box contains zero balls is $P(0) = 20/1000 = 0.02$. The process at hand is approximately a Poisson process, just as the balls-in-boxes setup in the example on page 213 was. We therefore simply need to find the value of a in Eq. (4.40) that makes $P(0) = 0.02$. That is,

$$\frac{a^0 e^{-a}}{0!} = 0.02 \implies e^a = 50 \implies a = \ln 50 = 3.912. \qquad (4.102)$$

This a is the average number of balls in each of the 1000 boxes. The total number of balls in each trial is therefore $n = (1000)a = 3912$.

Note that once we know what a is, we can determine the number of boxes that contain other numbers of balls. For example $P(3) \approx (3.9)^3 e^{-3.9}/3! \approx 0.20$. So about 200 of the 1000 boxes end up with three balls, on average. $P(4)$ is about the same (a hair smaller). About 4.5 boxes (on average) end up with 10 balls, as you can show.

Second solution: We can solve the problem from scratch, without using the Poisson distribution. With $k = 0$, Eq. (4.33) tells us that the probability of obtaining zero balls in a given box is $P(0) = (1 - 1/1000)^n$. Setting this equal to 0.02 and using the approximation in Eq. (7.14) gives

$$\begin{aligned}(1 - 1/1000)^n = 0.02 \quad &\implies \quad e^{-n/1000} = 0.02 \quad \implies \quad e^{n/1000} = 50 \\ &\implies \quad n/1000 = \ln 50 \quad \implies \quad n = 3912. \qquad (4.103)\end{aligned}$$

Alternatively, we can solve for n exactly, without using the approximation in Eq. (7.14). We want to find the n for which $(999/1000)^n = 0.02$. Taking the log of both sides gives

$$n \ln(0.999) = \ln(0.02) \implies n = \frac{-3.912}{-1.0005 \cdot 10^{-3}} = 3910. \qquad (4.104)$$

Our approximate answer of $n = 3912$ was therefore off by only 2, or equivalently 0.05%.

4.18. **Twice the events**

(a) This part of the problem is a repeat of Problem 4.11. The Poisson distribution is $P_a(k) = a^k e^{-a}/k!$, so the probability of obtaining a events is (using Stirling's formula for $a!$)

$$P_a(a) = \frac{a^a e^{-a}}{a!} \approx \frac{a^a e^{-a}}{a^a e^{-a}\sqrt{2\pi a}} = \frac{1}{\sqrt{2\pi a}}. \qquad (4.105)$$

(b) The average number of events during the time $2t$ is twice the average number during the time t. So we now have a Poisson process governed by an average of $2a$. The distribution is therefore $P_{2a}(k)$, and our goal is to calculate $P_{2a}(2a)$. In the same manner as in part (a), we find

$$P_{2a}(2a) = \frac{(2a)^{2a}e^{-2a}}{(2a)!} \approx \frac{(2a)^{2a}e^{-2a}}{(2a)^{2a}e^{-2a}\sqrt{2\pi(2a)}} = \frac{1}{\sqrt{4\pi a}}. \tag{4.106}$$

This is smaller than the result in part (a) by a factor of $1/\sqrt{2}$. In retrospect, we could have obtained the result of $1/\sqrt{4\pi a}$ by simply substituting $2a$ for a in the $1/\sqrt{2\pi a}$ result in part (a). The setup is the same here; we're still looking for the value of the distribution when k equals the average number of events. It's just that the average is now $2a$ instead of a.

(c) Since we're back to considering the original time t here, we're back to the Poisson distribution with an average of a. But since k is now $2a$, we want to calculate $P_a(2a)$. This equals

$$\begin{aligned} P_a(2a) &= \frac{a^{2a}e^{-a}}{(2a)!} \approx \frac{a^{2a}e^{-a}}{(2a)^{2a}e^{-2a}\sqrt{2\pi(2a)}} \\ &= \frac{1}{2^{2a}e^{-a}\sqrt{4\pi a}} = \left(\frac{e}{4}\right)^a \frac{1}{\sqrt{4\pi a}}. \end{aligned} \tag{4.107}$$

This is smaller than the result in part (a) by a factor of $(e/4)^a/\sqrt{2}$. The $(e/4)^a$ part of this factor is approximately $(0.68)^a$, which is very small for large a. For example, if $a = 10$, then $(e/4)^a \approx 0.02$. And if $a = 100$, then $(e/4)^a \approx 1.7 \cdot 10^{-17}$.

For $a = 10$, the above three results are summarized in Fig. 4.29. The three dots indicate (from highest to lowest) the answers to parts (a), (b), and (c). This figure makes it clear why the answer to part (c) is much smaller than the other two answers; the $P_{10}(20)$ dot is on the tail of a curve, whereas the other two dots are near a peak. Although we have drawn the Poisson distributions as continuous curves, remember that the distribution applies only to integer values of k. The two highest dots aren't right at the peak of the curve, because the peak of the continuous curve is located at a value of k between $a - 1$ and a; see Problem 4.10.

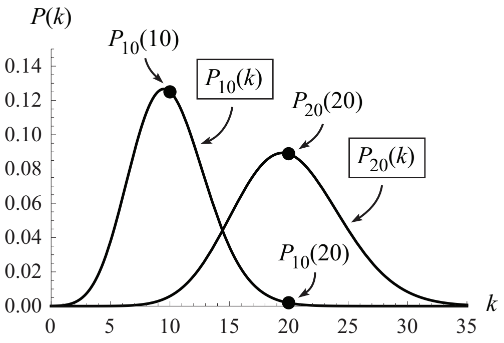


Figure 4.29: The Poisson curves for $a = 10$ and $a = 20$.

4.19. **P(0) the hard way**

The given interval (of time, space, or whatever) associated with the Poisson process has a expected events. Let's divide the interval into a very large number n of tiny intervals, each with a very small probability p of an event occurring. For simplicity, we are using p here instead of the $\lambda\epsilon$ we used at the beginning of Section 4.7.2. As in that section, we can ignore the distinction between the probability of an event in a tiny interval and the expected number of events in that interval, because we are assuming that $p \equiv \lambda\epsilon$ is very small; see Eq. (4.18).

The tasks of Problems 2.2 and 2.3 were to derive the "Or" rules for three and four events. Our goal here is basically to derive the "Or" rule for a large number n of independent events, each with a small probability p. These independent events are of course nonexclusive; we can certainly have more than one event occurring. Throughout this solution, you will want to have a picture like Fig. 2.17 in your mind. Although that picture applies to three events, it contains the idea for general n. Simple circles (each of which represents the probability that an event occurs in a given tiny interval) won't work for larger n, but it doesn't matter what the exact shapes are.

As in the solution to Problem 2.2(d), our goal is to determine the total area contained in the n partially overlapping regions (each with tiny area p) in the generalization of Fig. 2.17. The total area equals the probability of "Event 1 or Event 2 or . . . Event n," which is the desired probability that at least one event occurs in the original interval. As in the solution to Problem 2.2(d), we can proceed as follows.

- If we add up the individual areas of all n tiny regions, we obtain np. (Each region represents the probability p that an event occurs in that particular tiny interval, with no regard for what happens with any of the other $n - 1$ tiny intervals.) But np equals the total expected number of events a in the original interval, because p is the expected number of events in each of the n tiny intervals. The sum of the individual areas of all n tiny regions therefore equals a. This a is the first term in the parentheses in Eq. (4.53).
- However, in adding up the individual areas of all n tiny regions, we have double counted each of the overlap regions where two events occur. The number of these regions is $\binom{n}{2} = n(n-1)/2$, which essentially equals (in a multiplicative sense) $n^2/2$ for large n. The area (probability) of each double-overlap region is p^2, because that is the probability that two given events occur (with no regard for what else happens). The sum of the individual areas of the $n^2/2$ double-overlap regions is therefore $(n^2/2)p^2 = (np)^2/2 = a^2/2$. Since we have counted this area twice, and since we want to count it only once, we must correct for this by subtracting it off once. Hence the $-a^2/2!$ term in the parentheses in Eq. (4.53).
- We have now correctly determined the areas (probabilities) where *exactly one* or *exactly two* events occur. But what about the regions where three (or more) events occur? Each of these "triple" regions was counted $\binom{3}{1} = 3$ times when dealing with the "single" regions (because a triple region contains $\binom{3}{1}$ different single regions), but then uncounted $\binom{3}{2} = 3$ times when dealing with the "double" regions (because a triple region contains $\binom{3}{2}$ different double regions). We have therefore counted each triple region $\binom{3}{1} - \binom{3}{2} = 0$ times. There are $\binom{n}{3} = n(n-1)(n-2)/3! \approx n^3/3!$ of these regions. The area of each region is p^3, because that is the probability that three given events occur (with no regard for what else happens). The sum of the individual areas of the $n^3/3!$ triple regions is therefore $(n^3/3!)p^3 = (np)^3/3! = a^3/3!$. Since we have not counted

this area at all, and since we want to count it once, we must correct for this by adding it on once. Hence the $+a^3/3!$ term in the parentheses in Eq. (4.53).

- One more iteration for good measure: We have now correctly determined the areas (probabilities) where *exactly one* or *exactly two* or *exactly three* events occur. But what about the regions where four (or more) events occur? Each of these "quadruple" regions was counted $\binom{4}{1} = 4$ times when dealing with the single regions, then uncounted $\binom{4}{2} = 6$ times when dealing with the double regions, then counted $\binom{4}{3} = 4$ times when dealing with the triple regions. We have therefore counted each quadruple region $\binom{4}{1} - \binom{4}{2} + \binom{4}{3} = 2$ times. There are $\binom{n}{4} = n(n-1)(n-2)(n-3)/4! \approx n^4/4!$ of these regions. The area of each region is p^4, because that is the probability that four given events occur (with no regard for what else happens). The sum of the individual areas of the $n^4/4!$ quadruple regions is therefore $(n^4/4!)p^4 = (np)^4/4! = a^4/4!$. Since we have counted this area twice, and since we want to count it only once, we must correct for this by subtracting it off once. Hence the $-a^4/4!$ term in the parentheses in Eq. (4.53).

Continuing in this manner gives the entire area in Fig. 2.17, or rather, the entire area in the analogous figure for the case of n events instead of three. In the $n \to \infty$ limit, we will obtain an infinite number of terms inside the parentheses in Eq. (4.53). All of the multiple counting is removed, so each region is counted exactly once. The total area represents the probability that at least one event occurs. Subtracting this from 1 gives the probability $P(0)$ in Eq. (4.53) that zero events occur.

As mentioned in the remark in the solution to Problem Eq. (2.3), we have either overcounted or undercounted each region *once* at every stage. This is the *inclusion–exclusion principle,* and it follows from the binomial expansion of $0 = (1-1)^m$. Using the expansion in Eq. (1.21) with $a = 1$ and $b = -1$, we have

$$(1-1)^m = \binom{m}{0} - \binom{m}{1} + \binom{m}{2} - \binom{m}{3} + \cdots + \binom{m}{m-1}(-1)^{m-1} + \binom{m}{m}(-1)^m. \quad (4.108)$$

The lefthand side equals zero, and the $\binom{m}{0}$ and $\binom{m}{m}$ terms equal 1, so we obtain

$$\binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \binom{m}{m-1}(-1)^{m-1} = 1 + (-1)^m. \quad (4.109)$$

From the pattern of reasoning in the above bullet points, the lefthand side here is the number of times we have already counted each m-tuple region, in our handling of all of the 'lesser" regions – the single regions up through the $(m-1)$-tuple regions. (We are assuming inductively that we have overcounted or undercounted by 1 at each earlier stage.) The righthand side is either 2 or 0, depending on whether m is even or odd. We have therefore either overcounted or undercounted each m-tuple region by 1, which is consistent with the above results for $m = 2$, 3, and 4. There are $\binom{n}{m}$ of the m-tuple regions, each of which has an area of p^m. So at each stage, we need to either subtract or add an area (probability) of $\binom{n}{m}p^m \approx (n^m/m!)p^m = (np)^m/m! = a^m/m!$. These are the terms in parentheses in Eq. (4.53).

Remark: In the end, the solution to this problem consists of the reasoning in the remark in the solution to Problem Eq. (2.3), combined with the fact that if n is large, we can say that $\binom{n}{m}p^m \approx (n^m/m!)p^m$, which equals $a^m/m!$. Now, taking into account all of the above double (and triple, etc.) counting is of course a much more laborious way to find $P(0)$ than simply using Eq. (4.40). Equivalently, the double-counting solution is more laborious than using Eq. (4.32) with $k = 0$, which quickly gives $P(0) = (1 - p)^n \approx e^{-pn} = e^{-a}$, using Eq. (7.14). (Eq. (4.32) is equivalent to Eq. (4.34), which led to Eq. (4.40).) The reasoning behind Eq. (4.32) involved directly finding the probability that *zero* events occur, by multiplying together all of the probabilities $(1 - p)$ that each event doesn't occur. This is clearly a much quicker method than the double-counting method of finding the probability that *at least one* event occurs, and then subtracting that from 1. This double-counting method is exactly the *opposite* of the helpful "art of not" strategy we discussed in Section 2.3.1! ♣

4.20. **Probability of at least 1**

(a) In finding the probability that *at least one* ball ends up in the given box, our strategy (in both parts of this problem) will be to find the probability that *zero* balls end up in the given box, and then subtract this probability from 1. The process at hand is approximately a Poisson process, just as the balls-in-boxes setup in the example on page 213 was. So from the Poisson distribution in Eq. (4.40), the probability that zero balls end up in the given box is $P(0) = a^0e^{-a}/0! = e^{-a}$. The probability that at least one ball ends up in the given box is then $1 - e^{-a}$. This is an approximate result, because the process isn't exactly Poisson.

In the given setup, we have $n = 10^6$ balls and $b = 10^9$ boxes. So the average number of balls in a given box is $a = n/b = 1/1000$. Since this number is small, we can use the approximation in Eq. (7.9) (with $x \equiv -a$) to write $e^{-a} \approx 1 - a$. The desired probability that at least one ball ends up in the given box is therefore

$$1 - e^{-a} \approx 1 - (1 - a) = a = \frac{1}{1000}\,. \tag{4.110}$$

This makes sense. The expected number, a, of balls is small, which means that double (or triple, etc.) events are rare. The probability that at least one ball ends up in the given box is therefore essentially equal to $P(1)$. Additionally, since double (or triple, etc.) events are rare, we have $P(1) \approx a$, because the expected number of balls can be written as $a = P(1)\cdot 1 + \cancel{P(2)}\cdot 2 + \cdots \implies P(1) \approx a$. The two preceding sentences tell us that the probability that at least one ball ends up in the given box is approximately equal to a, as desired.

(b) The probability that a particular ball ends up in the given box is $1/b$, where $b = 10^9$ is the number of boxes. So the probability that the particular ball *doesn't* end up in the given box is $1 - 1/b$. This holds for all $n = 10^6$ of the balls, so the probability that *zero* balls end up in the given box is $(1 - 1/b)^n$. (This is just Eq. (4.33) with $k = 0$.) The probability that *at least one* ball ends up in the given box is therefore $1 - (1 - 1/b)^n$. This is the exact answer.

We can now use the $(1 + \alpha)^n \approx e^{n\alpha}$ approximation in Eq. (7.14) to simplify the answer. (We're using α in place of the a in Eq. (7.14), because we've already reserved the letter a for the average number of balls, n/b, here.) With $\alpha \equiv -1/b$, Eq. (7.14) turns the $1 - (1 - 1/b)^n$ probability into

$$1 - (1 - 1/b)^n \approx 1 - e^{-n/b} = 1 - e^{-a}. \tag{4.111}$$

The $e^{-a} \approx 1 - a$ approximation then turns this into a, as in part (a).

Remark: We have shown that for small $a = n/b$, the probability that at least one ball ends up in the given box is approximately a. This result of course doesn't hold for non-small a because, for example, if we consider the $a = 1$ case, there certainly isn't a probability of 1 that at least one ball ends up in the given box. And we would obtain a nonsensical probability larger than 1 if $a > 1$. From either Eq. (4.110) or Eq. (4.111), the correct probability (in the Poisson approximation) that at least one ball ends up in the given box is $1 - e^{-a}$. For non-small a, we can't use the $e^{-a} \approx 1 - a$ approximation to turn $1 - e^{-a}$ into a. ♣

4.21. **Comparing probabilities**

(a) The three events are independent. So with $p = 1/1000$, the desired probability is simply p^3, which equals 10^{-9}.

(b) The three trials of the process are independent, so the desired probability is again p^3, where $p = 1/1000$ is the probability that exactly one ball lands in the given box in a given trial of the process. So we again obtain an answer of 10^{-9}. This setup is basically the same as the setup in part (a).

(c) If we perform a single trial of throwing a million balls into a billion boxes, the probability that three *specific* balls end up in the given box is $(1/b)^3$ (where $b = 10^9$), because each ball has a $1/b$ chance of landing in the box.[6] There are $\binom{n}{3}$ ways to pick the three specific balls from the $n = 10^6$ balls, so the probability that exactly three balls end up in the box is $\binom{n}{3}/b^3$. We can simplify this result by making an approximation to the binomial coefficient. Using the fact that $n - 1$ and $n - 2$ are both essentially equal (multiplicatively) to n if n is large, we have

$$\begin{aligned}\binom{n}{3}\frac{1}{b^3} &= \frac{n(n-1)(n-2)}{3!}\frac{1}{b^3} \approx \frac{n^3}{3!}\frac{1}{b^3}\\ &= \frac{1}{3!}\left(\frac{n}{b}\right)^3 = \frac{(10^{-3})^3}{3!} = \frac{10^{-9}}{3!}\,. \end{aligned} \tag{4.112}$$

(d) The process in part (c) is approximately a Poisson process with $a = n/b = 1/000$. The probability that exactly three balls end up in the given box is therefore given by Eq. (4.40) as

$$P(3) = \frac{a^3 e^{-a}}{3!}\,. \tag{4.113}$$

Since $a = 1/000$ is small, the e^{-a} factor is essentially equal to 1, so we can ignore it. We therefore end up with

$$P(3) \approx \frac{a^3}{3!} = \frac{(10^{-3})^3}{3!} = \frac{10^{-9}}{3!}\,, \tag{4.114}$$

in agreement with the result in part (c).

In all of the parts to this problem, there is of course nothing special about the number 3 in the statement of the problem. If 3 is replaced by a general number k, then the results in parts (c) and (d) simply involve $k!$ instead of $3!$. (Well, technically k needs to be small compared with n, but that isn't much of a restriction in the present setup with $n = 10^6$.)

[6] There is technically a nonzero probability that other balls also land in the box. But this probability is negligible, so we don't have to worry about subtracting it off, even though we want *exactly* three balls in the box. Equivalently, the binomial distribution also involves a factor of $(1 - 1/b)^{n-3}$ (which ensures that the other $n - 3$ balls don't land in the box), but this factor is essentially equal to 1 in the present setup.

(e) The result in part (c) is smaller than the result in part (b) by a factor of $1/3! = 1/6$. Let's explain intuitively why this is the case.

In comparing the setups in parts (b) and (c), let's compare the respective probabilities (labeled $p_3^{(b)}$ and $p_3^{(c)}$) that three *specific* balls (labeled A, B, and C) end up in the given box. Although we solved part (b) in a quicker manner (by simply cubing p), we'll need to solve it here in the same way that we solved part (c), in order to compare the two setups. Note that in comparing the setups, it suffices to compare the probabilities for three specific balls, because both setups involve the same number of groups of three specific balls, namely $\binom{n}{3}$. So the total probabilities in each case are $\binom{n}{3}p_3^{(b)}$ and $\binom{n}{3}p_3^{(c)}$, with the $\binom{n}{3}$ factor being common to both.

Consider first the setup in part (c), with the single trial. There is only one way that all three of A, B, and C can end up in the box: If you successively throw down the n balls, then when you get to ball A, it must end up in the box (which happens with probability $1/b$); and then when you get to ball B, it must also end up in the box (which again happens with probability $1/b$); and finally when you get to ball C, it must also end up in the box (which again happens with probability $1/b$). The probability that all three balls end up in the box is therefore $p_3^{(c)} = (1/b)^3$. (This is just a repeat of the reasoning we used in part (c).)

Now consider the setup in part (b), with the three trials. There are now *six* ways that the three balls can end up in the box, because there are 3! permutations of the three balls. Ball A can end up in the box in the first of the three trials of n balls (which happens with probability $1/b$), and then B can end up in the box in the second trial (which again happens with probability $1/b$), and then C can end up in the box in the third trial (which again happens with probability $1/b$). We'll label this scenario as ABC. But the order in which the balls go into the boxes in the three successive trials can take five other permutations too, namely ACB, BAC, BCA, CAB, CBA. Each of the six possible permutations occurs with probability $(1/b)^3$, so the probability that all three balls (A, B, and C) end up in the box equals $p_3^{(b)} = 6(1/b)^3$. This explains why the answer to part (b) is six times the answer to part (b).

As mentioned above, if we want to determine the total probabilities in each setup, we just need to multiply each of $p_3^{(b)}$ and $p_3^{(c)}$ by the number $\binom{n}{3} \approx n^3/3!$ of groups of three balls. This was our strategy in part (c), and the result was $(n/b)^3/3!$. In part (b) this gives $(n^3/3!)(6/b^3) = (n/b)^3 = p^3$, in agreement with our original (quicker) solution. Note that it isn't an extra factor of 3! in the denominator that makes the answer to part (c) be smaller; parts (b) and (c) both have the 3! arising from the $\binom{n}{3}$ binomial coefficient. Rather, the answer to part (c) is smaller because it *doesn't* have the extra 3! in the numerator arising from the different permutations.

Remark: Alternatively, you can think in terms of probabilities instead of permutations. In part (c) the probability (as we noted above) that three specific balls end up in the box is $(1/b)(1/b)(1/b)$, because each of the three balls must end up in the box when you throw it down. In contrast, in part (b) the probability that three specific balls end up in the box is $(3/b)(2/b)(1/b)$, because in the first trial of n balls, any of the three specific balls can end up in the box. And then in the second trial, one of the two other balls must end up in the box. And finally in the third trial, the remaining one of the three balls must end up in the box. The probability in part (b) is therefore larger by a factor of $3! = 6$.

Intuitively, it makes sense that the probability in part (b) is larger, because in part (c) if ball A doesn't end up in the box when you throw it down, you are guaranteed failure (for the three specific balls A, B, and C). But in part (b) if ball A doesn't end up in the box in the first of the three trials of n balls, you still have two more chances (with balls B and C) in that trial to get a ball in the box. So you have three chances to put one of the balls in the box in the first trial. And likewise you have two chances in the second trial. ♣

4.22. **Area under a Gaussian curve**

Let I be the desired integral. Then following the hint, we have

$$\begin{aligned} I^2 &= \left(\sqrt{\frac{b}{\pi}}\int_{-\infty}^{\infty} e^{-bx^2}\,dx\right)\left(\sqrt{\frac{b}{\pi}}\int_{-\infty}^{\infty} e^{-by^2}\,dy\right) \\ &= \frac{b}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-b(x^2+y^2)}\,dx\,dy. \end{aligned} \tag{4.115}$$

If we convert from Cartesian to polar coordinates, then $x^2 + y^2$ becomes r^2 (by the Pythagorean theorem), and the area element $dx\,dy$ in the plane becomes $r\,dr\,d\theta$. This expression follows from the fact that we can imagine covering the plane with infinitesimal rectangles with sides of length dr in the radial direction and $r\,d\theta$ (the general form of an arclength) in the tangential direction.

The original double Cartesian integral runs over the entire x-y plane, so the new double polar integral must also run over the entire plane. The polar limits of integration are therefore 0 to ∞ for r, and 0 to 2π for θ. The above integral then becomes

$$I^2 = \frac{b}{\pi}\int_0^{2\pi}\int_0^{\infty} e^{-br^2}\,r\,dr\,d\theta. \tag{4.116}$$

The θ integral simply gives 2π. The indefinite r integral is $-e^{-br^2}/2b$, as you can verify by differentiating this. The factor of r in the area element is what makes this integral doable, unlike the original Cartesian integral. We therefore have

$$\begin{aligned} I^2 &= \frac{b}{\pi}\cdot 2\pi\cdot\left(-\frac{e^{-br^2}}{2b}\right)\bigg|_0^{\infty} \\ &= \frac{b}{\pi}\cdot 2\pi\cdot\frac{-1}{2b}\cdot\big(0-1\big) \\ &= 1. \end{aligned} \tag{4.117}$$

So $I = \sqrt{1} = 1$, as desired. Note that if we didn't have the factor of $\sqrt{b/\pi}$ in the distribution, we would have ended up with

$$\int_{-\infty}^{\infty} e^{-bx^2}\,dx = \sqrt{\frac{\pi}{b}}\,. \tag{4.118}$$

This is a useful general result.

The above change-of-coordinates trick works if we're integrating over a circular region centered at the origin. (An infinitely large circle covering the entire plane falls into this category.) If we want to calculate the area under a Gaussian curve with the limits of the x integral being arbitrary finite numbers a and b, then our only option is to evaluate the integral numerically. (The change-of-coordinates trick doesn't help with the rectangular region that arises in this case.) For example, if we want the limits to be $\pm\sigma = \pm 1/\sqrt{2b}$, then we must resort to numerics to show that the area is approximately 68% of the total area.

4.23. **Variance of the Gaussian distribution**

First solution: With $\mu = 0$, the variance of the second expression for $f(x)$ in Eq. (4.42) is

$$E(X^2) = \int x^2 f(x)\,dx = \sqrt{\frac{1}{2\pi\sigma^2}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2\sigma^2} dx. \tag{4.119}$$

We can evaluate this integral by using integration by parts. That is, $\int fg' = fg - \int f'g$. If we write the x^2 factor as $x \cdot x$, then with $f \equiv x$ and $g' \equiv xe^{-x^2/2\sigma^2}$, we can integrate g' to obtain $g = -\sigma^2 e^{-x^2/2\sigma^2}$. So we have

$$\begin{aligned}\int_{-\infty}^{\infty} x \cdot xe^{-x^2/2\sigma^2} dx &= x \cdot \Big(-\sigma^2 e^{-x^2/2\sigma^2}\Big)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} 1 \cdot \Big(-\sigma^2 e^{-x^2/2\sigma^2}\Big)\,dx \\ &= 0 + \sigma^2 \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx. \end{aligned} \tag{4.120}$$

The 0 comes from the fact that the smallness of $e^{-\infty^2}$ wins out over the largeness of the factor of ∞ out front. The remaining integral can be evaluated by invoking the general result in Eq. (4.118). With $b \equiv 1/2\sigma^2$ the integral is $\sqrt{2\pi\sigma^2}$. So Eq. (4.120) gives

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2\sigma^2} dx = \sigma^2 \sqrt{2\pi\sigma^2}. \tag{4.121}$$

Plugging this into Eq. (4.119) then gives

$$E(X^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \cdot \sigma^2 \sqrt{2\pi\sigma^2} = \sigma^2, \tag{4.122}$$

as desired.

Second solution: This solution involves a handy trick for calculating integrals of the form $\int_{-\infty}^{\infty} x^{2n} e^{-bx^2} dx$. Using the $\int_{-\infty}^{\infty} e^{-bx^2} dx = \sqrt{\pi} b^{-1/2}$ result from Eq. (4.118) and successively differentiating both sides with respect to b, we obtain

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-bx^2} dx &= \sqrt{\pi} b^{-1/2}, \\ \int_{-\infty}^{\infty} x^2 e^{-bx^2} dx &= \frac{1}{2}\sqrt{\pi} b^{-3/2}, \\ \int_{-\infty}^{\infty} x^4 e^{-bx^2} dx &= \frac{3}{4}\sqrt{\pi} b^{-5/2}, \end{aligned} \tag{4.123}$$

and so on. On the lefthand side, it is indeed legal to differentiate the integrand (the expression inside the integral) with respect to b. If you have your doubts about this, you can imagine writing the integral as a sum over, say, a million terms. It is then certainly legal to differentiate each of the million terms with respect to b. In short, the derivative of the sum is the sum of the derivatives.

The second line in Eq. (4.123) is exactly the integral we need when calculating the variance. With $b \equiv 1/2\sigma^2$, the second line gives

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2\sigma^2} dx = \frac{1}{2}\sqrt{\pi}\left(\frac{1}{2\sigma^2}\right)^{-3/2} = \sqrt{2\pi}\,\sigma^3, \tag{4.124}$$

in agreement with Eq. (4.121).