

INTRODUCTION TO STATISTICS

Mark E. Reeves, Dept. of Physics, GWU

Why do we study statistics? Clearly one answer is that this knowledge can help us to understand data that we measure in the lab. We calculate the average and the variance (miscalled the standard deviation) and use these to report the “true value” and our confidence in that true value. Is it that simple? Can we make five measurements and be confident that we know the true average and the values that bracket it? In other words, do we really know that the mass is $26.5 \pm .2$ g?

We will see in this discussion that we will need to be more careful in our statements about measuring the true value of a quantity. We will also see that statistics themselves can be a powerful theoretical tool for building models to understand nature. In this reading, we will look at some descriptive ideas and apply them to a few examples that arise from thinking about the outcome of throwing dice. In class, we have discussed the problems of random sampling from a population of people, 1/5 of whom have green eyes, and we have considered the outcome of flipping a coin, flipping 2 coins, or flipping 4, 8, 16, ..., 128 coins. We will not revisit these here, but can hopefully think about the results of the class discussion in terms of what is written here.

Why do we study statistics in a physics class? The obvious answer is given above that the proper use and understanding of statistics will help us to better evaluate data that we measure. Not so obvious is that statistics can help us to understand the kinematics and dynamics of biological and chemical systems. Think about the limits of how small a sample of protein you can measure. Is it micromoles? Nanomoles? Picomoles? Femtomoles? Attomoles? Even if you said attomoles, and this is really stretching it, we are still talking about a very large number of proteins, almost a million. Following a million proteins or writing an equation forces acting on a million proteins and predicting the resulting dynamics is simply not possible by simply solving kinematical equations. Nevertheless, quantities like velocity, momentum, and impulsive force will translate to pressure and temperature via careful use of statistical principals. Ultimately, learning to do this is the major goal of the rest of the course.

Question 1.1: How large are these quantities: Rank in order, from least to greatest pico-, atto- micro-, nano-, femto-. How many proteins are there in a one attomole sample?

Preliminaries: the dice problem.

Let's think then about rolling 1 dice vs. rolling 2 die. If we only roll one dice, then we know that the possible values we can get are 1,2,3,4,5,6, and that each of these

values are equally probable. We plot the expected distribution, which has the shape of a small rectangle below.

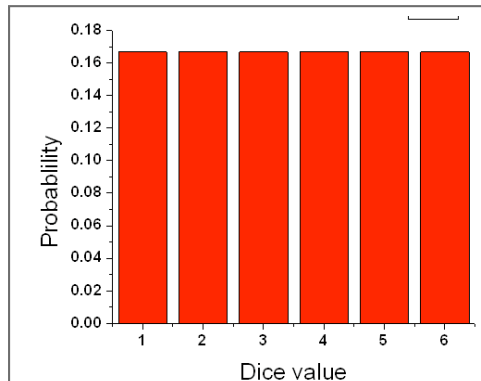


Figure 1: The probability density for rolling one dice. That is, for one dice, a plot of the dice value vs. the probability to roll that value. Adapted from <http://www.stat.sc.edu/~west/javahtml/CLT.html>

Notice several things about this:

1. For each value of the dice, the probability of getting that value is $1/6$. That is the number of ways to, for example, get a 5 (one way) divided by the total number of possible scores (six possibilities).
2. The sum of all probabilities is equal to 1. That is we have a 100% chance to roll a 1, 2, 3, 4, 5, or 6.
3. We could ask a different question, what is the probability to roll an even number, or to roll a number divisible by 3. More about this in the next section.

That was pretty easy. Now let's make it a little more challenging by asking what happens when we roll two die. You are familiar with this problem from your many hours playing

Backgammon or Monopoly. You know from these games that you can roll any value between 2 and 12. You also know that the chance of getting a 12 is smaller than for getting a six or a seven. We now know how to make this quantifiable and actually calculate the probability. First, we must figure out how many ways there are to roll each value. For example, there are 3 ways (1+3, 3+1, 2+2) to roll a 4. These are plotted below for all 11 possible values. Second, we add up all the possible rolls and find that there are 36 possibilities in all. Third, we divide the number of ways to roll each score by the total, 36, to get the probability. This is also plotted below, on the right.

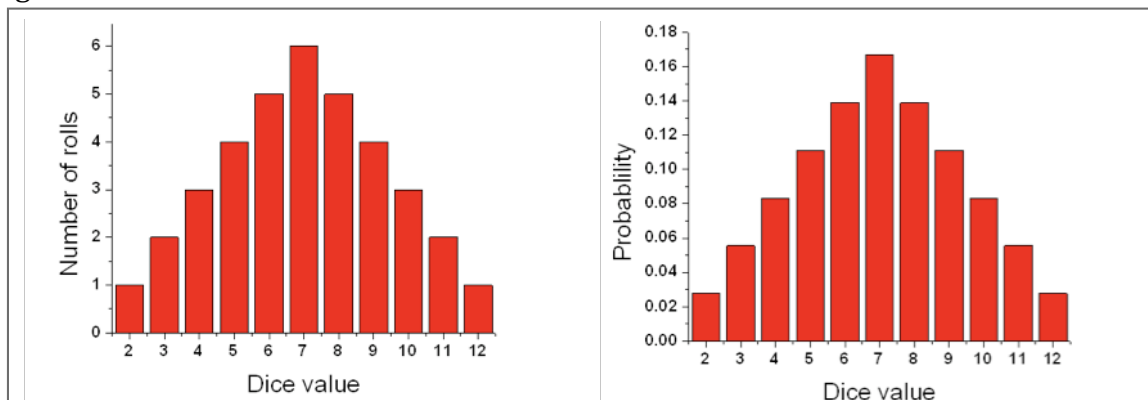


Figure 2: For two die, the number of ways to get a score vs. the total score (left), and the probability to get a score vs. the total score (right). The right-hand plot is the probability density for rolling two die. Adapted from <http://www.stat.sc.edu/~west/javahtml/CLT.html>

Questions 1.1: List the 6 ways to roll a 7. And show that the probability to roll a 7 is 16.667%. Go to the web site referenced in Fig. 2, and run the applet for larger numbers of die (set the number of rolls to 10,000 and roll several times – more on this later). Describe the shape of the distribution.

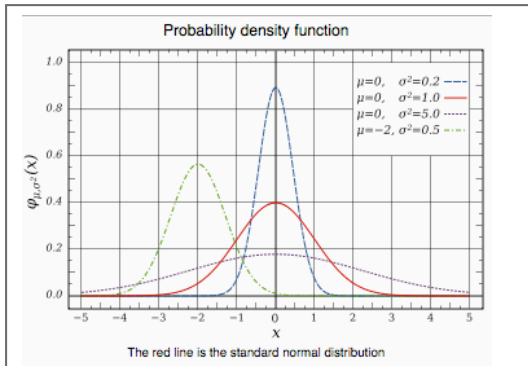


Figure 3: The probability density for rolling a large number of die. This is the normal distribution. Taken from Wikipedia article on the normal distribution.

We can see that the shape of the probability distribution changes as we roll one vs. two die. It can be expected to change as well, if we roll three, four, or five die. In fact, you can show (but we will not do so) that if we roll a large enough number of die, then the distribution will become a Gaussian

function, $p = \frac{1}{\sigma\sqrt{2\pi}} e^{-(V-\langle V \rangle)^2 / 2\sigma^2}$, which is

plotted below. To read more about the Gaussian function, also called the Bell curve or normal distribution, see the Wikipedia article on the normal distribution.

Characteristics of the distribution

We can go further and extract some

characteristic values from these distributions: the average and standard deviations. Looking at Figure 2, we can see that the average value is 7. We can calculate this average by adding $1 \times 2 + 2 \times 3 + 3 \times 4 + \dots$ (and so on for each value, V). We can write this in a compact form $\text{Sum} = \sum n_i V_i$ for

- i varying from 1 to 11 (number of possible values),
- n varying from 1 to 6 and
- V varying from 2 to 12, as in Figure 3.

If we divide this sum by the total number of theoretical rolls, 36 ... Total rolls $36 = \sum n_i$, then we get the average value of 7. We will give the average a special symbol, $\langle V \rangle$, and then write it as an equation:

$$\langle V \rangle = \text{Sum}/36 = (\sum n_i V_i)/36 = \sum (n_i V_i / 36).$$

In the last step we used the associative property to divide each term in the sum by 36.

If we are watching carefully, in the two-dice example, we notice that $n_i/36$ is just the probability P_i to measure the value V_i . Thus, we can rewrite the average in this way, $\langle V \rangle = \sum P_i V_i$. This is then the theoretical value of the average, also called the expectation value, and we can use the probabilities plotted in the right-hand Figure above to compute $\langle V \rangle$ in this way.

We can go even further and ask, “What is the value of the average deviation away from the average?” This quantity, the standard deviation, gives us some idea of how sharply peaked our probability distribution is. The standard deviation, also called

the root mean squared (RMS) deviation, is defined, just like its name implies, by the equation $\sigma = \sqrt{\langle (V - \langle V \rangle)^2 \rangle} = \sqrt{(1/N) \sum P_i (V_i - \mu)^2} = \sqrt{(1/N) \sum P_i (V_i)^2 - \mu^2}$, where in the last two steps we simply substitute the symbol μ for the average value ($\mu = \langle V \rangle$). Notice that the symbol, σ appears in the formula for the normal distribution. This is the same quantity, the standard deviation, and it is equal to the width of the Bell curve, at the half-way point of its height (Also called full-width at half max FWHM). So from a plot of the Bell curve, we can pick off both the average value (the peak value), and the standard deviation (the FWHM value).

Question 1.2: Show that $\frac{1}{N} \sum P_i (V_i - \mu)^2 = \frac{1}{N} \sum P_i (V_i)^2 - \mu^2$ (Hint, first simplify the quadratic term using the FOIL method).

How we count depends upon the rules we set

Now, in case you haven't noticed, we have been discussing statistics from an entirely theoretical perspective. The average value defined above is like no average we would measure in the lab, rather it is the theoretical value that we expect the average to take. Thus, $\langle V \rangle$ is often called the expectation value of V .

Further, we have set some arbitrary rules for the counting. In the examples from dice rolls, we used the rule (sensibly) that the score was the sum of the number of dots on the top face of the die. We could have also asked other questions, for example, what is the probability to roll an even number, or the probability to roll a 7, or the probability to roll a value divisible by three. These are examples where the answer has two values, did we roll a 7 or not, did we succeed or fail?, to which we can assign values 1 or 0. This two-possibility distribution has a special mathematical form, called the binomial distribution, which is given by

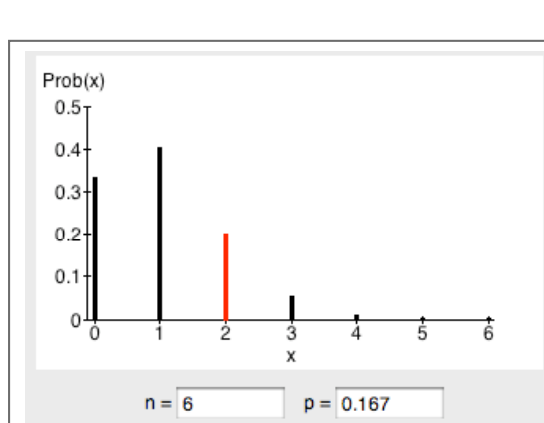


Figure 4: The probability density for rolling one dice. That is, for one dice, a plot of the dice value vs. the probability to roll that value. Adapted from <http://www.stat.tamu.edu/~west/applets/binomialdemo.html>

$$P = \binom{n}{r} p^r (1-p)^{n-r}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

and

$$n! = n(n-1)(n-2)\dots 3(2)1$$

In this equation, p is the probability to get a single yes (or 1 value), n is the number of attempts, r is the number of times to get the value of 1 (or yes). The first factor, $\binom{n}{r}$,

comes from considering the combinatorics or the number of ways to

arrange identical objects (see the discussion in the reading material from the Book of Numbers). How would you use this formula? For example, consider rolling one dice. This formula will give you the answer to the question, in 6 rolls, what would be the probability two of these will be fives? In this case, $n = 6$, $r = 2$, $p = 1/6$.

$P = \frac{6!}{2!4!} (1/6)^2 (5/6)^4 = 0.2014$. In another example, we could ask, what is the probability to roll 2 even values after 5 rolls of one dice? In this case $n = 5$, $r = 2$, $p = 1/2$.

Questions 1.3: Use the binomial calculator, referenced in Fig. 4, with several values of n . At what value of n does the probability distribution begin to look like the normal distribution? Once you find a large enough value of n , then vary i . For $i = 0.5$, does the average value (estimated from the distribution) make sense? How about for $i = 0.2$, $i = 0.75$? Other values?

More on the shape of the distribution

We could see from Question 1.3 that the binomial distribution does begin to resemble the normal distribution for large values of n . This is also the observation for rolling large numbers of die. These two cases are examples of what is called the central limit theorem. One very useful result that can be found by looking at the binomial distribution for progressively larger values of n is that the average value and the standard deviation may be calculated for any value of n , once they have been calculated for $n = 1$. This is expressed by the simple formulae below:

$$\mu_n = n\mu_1$$

and

$$\sigma_n^2 = n\sigma_1^2$$

What does this tell us about measurements?

In case you are still have not noticed, everything that we have discussed has been theoretical. How does it relate to what we would measure in the lab. For example, in the example with one dice, the average value would be 3.5. If we rolled a dice 6 times, would we get a 1, a 2, a 3, a 4, a 5, and a 6 to get an average value of 3.5? Probably not. How about for two die, after 36 rolls, would we roll a 7, exactly 6 times? Again, probably not. You can see that the actual number of rolls has to be very large in order to get a measurement that looks the same as the theoretical distribution.

The take-away message is that you should be careful how much significance you attach to average values measured from sample populations. Are these sample populations large enough to be representative of the theoretical, parent distributions.

Question 1.4. Try this yourself by invoking the applet at the website, <http://www.stat.sc.edu/~west/javahtml/CLT.html>. Go there and set the number of die equal to one. How many rolls does it take to recover the distributions in Figure 1? Answer the same for two die, how many rolls are required to recover Figure 2?

What does this tell us about nature and particularly about biology?

The short answer is yes, we can learn from statistics. In fact, there are a number of models for biological structures and processes that are described by statistical principles. We have already discussed one of them, the Brownian motion, which is modeled by a random walk, a process whereby a particle moves a short distance, collides with another object, generally a water molecule, and then moves off in a completely random direction. This process is governed by binomial statistics and ultimately by the normal distribution in the limit of a large number of steps. We covered this in class, and in the readings from Berg's textbook, which are on the class web site.

Furthermore, the very same statistical model links Brownian motion to a very fundamental way of moving materials in cells, that is diffusion. We have demonstrated that small molecules do diffuse in gels, in a manner that is quantitatively described by an increasingly spreading normal distribution. The very fact that there are so many molecules in the measurements that we made allows us to reach the large sampling limit where 1) the binomial distribution goes over to the normal distribution (large n = the number of steps = "number of coin flips") and 2) the number of molecules sampled is large enough that we can confidently say that our measurements are a representative sample of the parent distribution (which happens to be the normal distribution). This happens over and over in physics, chemistry, and particularly in biology, where the thermal energy at room temperature is large enough to jostle biological molecules out of place so that they sample a statistically large number of the possible configurations available to them. For the counter example, think of freezing a cell culture (or your gel) and asking whether the thermal energy is now sufficiently large to push the molecules from one place to another. Clearly, it is not, and so the die molecules will not diffuse at all.

Where will we go from here?

We can and will go well beyond the example of diffusion when using statistical models to describe biological processes. We will define a new quantity, entropy, not in terms of a popularized measure randomness and disorder, but rather in terms of a measure of the number of ways that a system can arrange itself. We can go back to the idea of rolling two die, and ask "For any given roll, what state are we likely to observe?" We have seen that safe money is on observing a 5, 6, 7, 8 or 9 as opposed to a 2, 3, 4, 10, 11, or 12. Why is this? Simply because there are more arrangements of the two die that will give a 5, 6, 7, 8 or 9 than will give a 2, 3, 4, 10, 11, or 12. We say that the state of "5, 6, 7, 8, or 9" has a higher entropy than the state of 2, 3, 4, 10, 11, or 12. Just as the dice find their way into the higher entropy (multiplicity) state, nature also finds the way to be in the most likely state. This only makes sense and thinking in this very sensible way will

allow us to spend some time over the rest of this course to quantitatively model a number of phenomena from membrane formation to protein folding.

Question 1.5 How many ways can we roll a 5, 6, 7, 8 or 9? How many ways can we roll a 2, 3, 4, 10, 11, or 12? What is the probability to roll a 5, 6, 7, 8 or 9? To roll a 2, 3, 4, 10, 11, or 12?