(Lane 3, 5)

Module 3 of 3 – Probability & Statistics

Probability review.

- Probabilities are degrees of certainty $0 \le p \le 1$.
- We can ~~[illegible]~~ calculate the probability of an event ~~[illegible]~~ as $Pr(\text{event}) = \frac{|\text{success outcomes}|}{|\text{all outcomes}|}$, if all outcomes are equally likely.
- A conditional probability $Pr(A|B)$ is the probability of A given ~~[illegible]~~ event B is true. Counting, this is $\frac{|\text{outcomes with A and B true}|}{|\text{outcomes with B true}|}$.

For example, Pr (6-sided die rolls 4) is 1/6, but Pr (6-sided die rolls 4 | roll is even) is 1/3 because only 3 outcomes satisfy event B true.

Also, when rolling 2 dice, not all sums are equally likely – we can't say all rolls 2–12 have the same likelihood of 1/11, because the outcomes aren't equally likely to start. But each (a, b) outcome of (1st roll, 2nd roll) is equally likely, and this way we can tell a 2 has 1/36 probability, while a 7 has 6/36 = 1/6 probability (1, 6), (2, 5), ... (6, 1). (This fact will lead to many natural distributions where not all outcomes are equally likely – especially the Gaussian distribution.)

We can also have random variables that take on different values instead of being true or false – like a die roll, which takes on the values 1 to 6 with equal likelihood. If a random variable has a uniform distribution, then all outcomes are equally likely. But other distributions, which assign probabilities to outcomes, are possible; if X = $ won from lottery, that's unlikely to be uniform.

With random variables, we can find expectations. E[X] is the expected value of X. It's what the average value of X would be if we averaged over a very large number of games.

The definition of $E[X]$, recall, is

$$\sum_{X} X \Pr(X)$$

or in other words, the sum over all outcomes of probability * outcome. We can think of this as a "probability-weighted average." If all outcomes are equally likely, it's just the average. But if they aren't, then we sum $p \cdot$ outcome instead of $\frac{1}{n} \cdot$ outcome to get the answer.

$$E[\text{20-sided die}] = \sum_{x=1}^{20} \frac{1}{20} \cdot X = \frac{1}{20} \sum_{x=1}^{20} X = \text{average roll} = 10.5$$

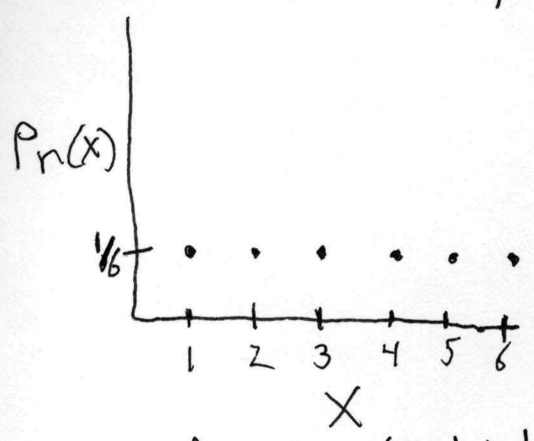
Raffle ticket costs \$5, $\frac{1}{100}$ chance of \$100

X = net winnings

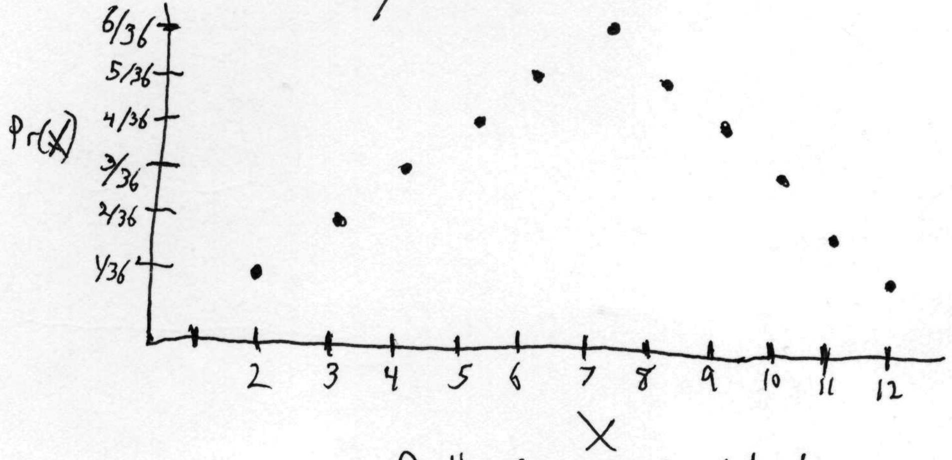
$$E[X] = \frac{99}{100} \cdot -5 + \frac{1}{100} \cdot 95 = -\frac{400}{100} = -\$4$$

We'll bring up other reminders as needed.

We can visualize distributions by graphing outcomes on the x-axis, and probabilities on the y-axis.



Roll of a 6-sided die



Roll of 2 6-sided dice

Something is clearly different here – on the left, there's no particular tendency to give us the expected value, except on average in the long run. But on the right, a 7 is more likely than an 8 even on just one trial. If we somehow "bet" on a non-extreme result, it's a safer bet on the right.

The number that characterizes how much of a tendency ~~[illegible]~~ to stray from the mean there is in the data is called the <u>variance</u> of the distribution.

$$Var(X) = E\left[(X - E[X])^2\right]$$

This is the expected value of the <u>squared</u> difference between the result and expectation. (As with least squares regression we assume we care more about big differences than little ones.) You may sometimes in statistical settings see $E(X)$ written μ if it's just the average value, which is true in some common distributions. Thus:

$$Var(X) = E\left[(X-\mu)^2\right].$$

The square root of the variance is called the "standard deviation," and sometimes the greek letter σ = sigma ("s") is used to denote the <u>standard deviation.</u>

So $\sigma^2 = E\left[(X - E[X])^2\right]$ or $\sigma = \sqrt{E\left[(X-E[X])^2\right]}$.

As with variance, a big σ indicates a spread-out distribution.

Let's take as examples the 1-die and two-die cases. These have $E[X] = 3.5$ and $E[X+Y] = 7$, respectively.

Variance for 1 die is $\frac{1}{6}(-2.5)^2 + \frac{1}{6}(-1.5)^2 + \frac{1}{6}(-0.5)^2 + \frac{1}{6}(0.5)^2 + \frac{1}{6}(1.5)^2 + \frac{1}{6}(2.5)^2$

$$= \frac{1}{6}(6.25 + 2.25 + 0.25) * 2 = 2.916$$

So the expected value of the square of the difference from the mean is 2.92. The standard deviation is thus 1.71 or so, and tells us roughly how far from the mean we can expect to be.

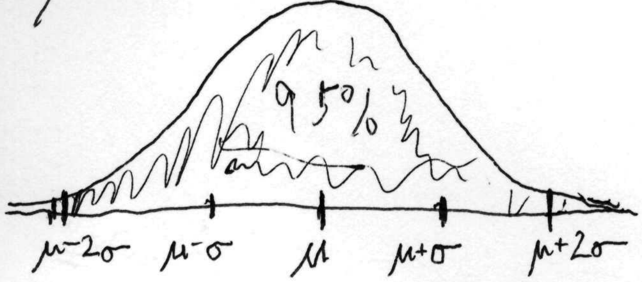
Variance for 2 dice is $\frac{1}{36}(-5)^2 + \frac{2}{36}(-4)^2 + \frac{3}{36}(-3)^2 + \dots + \frac{6}{36}(0)^2 + \frac{5}{36}(1)^2 + \dots + \frac{1}{36}(5)^2$

$$= 5\tfrac{5}{6} \text{ or } 5.83.$$

When we add the results of independent events, the variance sums as well. We can see this here: each die roll had variance 2.916, so the variance of the sum is $2.916 \cdot 2 = 5.83$.

Notice that the variance doesn't really describe the shape per se: the distribution for 2 dice has higher variance despite looking more concentrated (compared to 1 die). The variance just describes the size of differences, so distributions across wider ranges, like 2-12 instead of 1-6, will naturally tend toward higher numbers. If we kept the range the same, then lower variance would tend to imply more concentration toward the mean.

For the very common "normal" or "Gaussian" distribution, standard deviations have a useful rule of thumb associated with them: there is a 95% chance that a sample will fall within 2 standard deviations to either side of the mean. (1.96 to be more precise) Or, alternately, we can expect 95% of all samples to fall in this range. This is specifically for normal/Gaussian distributions, which we'll cover in more detail later, but it's the most common way to encounter standard deviations.



One last variance fact: $Var(aX) = a^2 Var(X)$. If the original variable is multiplied by a, that affects the variance by a factor of a^2. This will come up in some derivations later.

The Binomial Distribution

Some "named" distributions come about because they result naturally from doing some particular process or experiment repeatedly. The binomial distribution is a distribution on the number of coin flips that come up heads after flipping n of them. The coins could also be "biased" and have a probability of success p instead of being 0.5 for sure. (Flipping biased coins is often used as a model or metaphor for any kind of process with a fixed probability of success.)

You may recall from Discrete Structures that the number of ways to get k heads on n flips is $\binom{n}{k}$ since you're choosing which flips are heads. The probability of getting k successes is the sum over all these sequences of their probabilities. The probability of each specific sequence with the right number of heads is $p^k(1-p)^{n-k}$ — each head in the sequence had probability p, each tails had probability $1-p$, and if we multiply out these terms over all n symbols, we get the above.

Example: $p = 3/4$, $k=2$

Probability of HTTH: $3/4 \cdot 1/4 \cdot 1/4 \cdot 3/4$

Probability of TTHH: $1/4 \cdot 1/4 \cdot 3/4 \cdot 3/4$ same terms, just rearranged

When we ask "how many sequences have the right k," it's $\binom{n}{k}$, and they all have the same probability of $p^k(1-p)^{n-k}$, so the total probability of k successes is $\binom{n}{k}p^k(1-p)^k$.

What does this look like? If the coin isn't biased, and $p=(1-p)=0.5$, then this is $\binom{n}{k}\left(\frac{1}{2}\right)^n$. The overall shape of the distribution can be read off Pascal's triangle, which has the values for $\binom{n}{k}$ in row n.

row 0 → 1

1 1

1 2 1

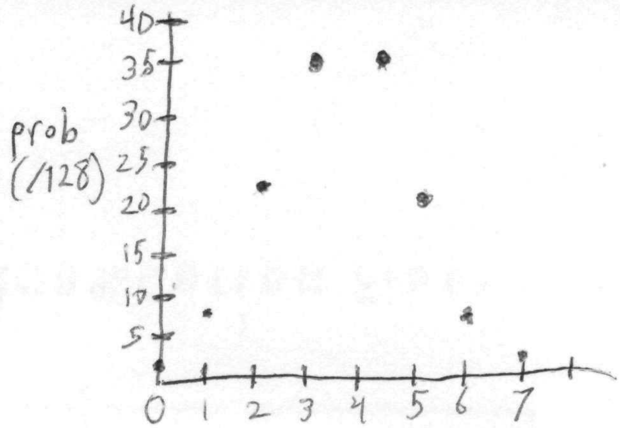
1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

row 7 → 1 7 21 35 35 21 7 1



This is starting to look a little like a Gaussian or normal, and that's no accident: as any independent random variables are added, the distribution of the sum of their outcomes will start to look more and more Gaussian. (Which also means plotting Pascal's triangle values starts to look more and more Gaussian.)

Two properties that we'll often want to be curious about for a distribution are its mean and its variance.

Mean: What is the expected number of successes for n flips with probability p? Intuitively, this should be np. For example, if $p = 0.6$ and $n = 10$, we'd expect 6 successes. This is correct, and one way to show it is with linearity of expectation: $E[X]$ for one flip is 0.6, so

$$E[X_1 + X_2 + \cdots + X_{10}] = E[X_1] + E[X_2] + \cdots + E[X_{10}] = 6.$$

To find the variance we can use a similar trick; recall $Var(X+Y) = Var(X) + Var(Y)$ if the two are independent. So we really just need to find the variance of one flip.

$$\underbrace{(1-p)}_{\text{fail prob}}\underbrace{(-p)^2}_{\text{fail is off by } p} + \underbrace{p}_{\text{success prob}}\underbrace{(1-p)^2}_{\text{success off by } (1-p)} = p^2 - p^3 + p^3 - 2p^2 + p = p - p^2 = p(1-p)$$

So by our variance rule, the overall variance must be $np(1-p)$.