Week 6

17 June 2021

26 Tail formula

Last time we talked about the expected value of a distribution:

We saw that, if we're just trying to figure out whether something succeeded or not, we can look at indicator functions instead of the probability itself. In this case, we let X_i denote the indicator function and we end up with:

Since this equation looks so nice, we're going to see whether we can turn our original definition of $E(X)$ into something nicer. It turns out, we can in the case where our x are coming from a set $\{0, 1, \ldots, n\}$. In this case, if we calculate $E(X)$ we get:

$$\begin{aligned} E(X) =& P(X = 1) \\ &+ P(X = 2) + P(X = 2) \\ &+ P(X = 3) + P(X = 3) + P(X = 3) \\ &+ P(X = 4) + P(X = 4) + P(X = 4) + P(X = 4) \\ &+ \cdots \end{aligned}$$

We can rewrite this into what's known as the *tail sum formula for expected value*.

Example 26.1 Suppose that we roll four fair six-sided dice and we let M be the minimum value of the four numbers rolled. What is the expected value of M?

27 Markov's Inequality

Let's go back to indicators again. For indicators, we took the sum of the indicators to get the expected value of success of our event X. How do we know if our sum of indicators is itself an indicator?

So let $X_1, X_2, \ldots, X_n$ be a bunch of events and let $I_1, I_2, \ldots, I_n$ be

their indicator functions. In other words, if X_i occurs then $I_i = 1$, if X_i does not occur then $I_i = 0$. If we think about it, an indicator only has values 0 and 1; that is the definition of an indicator function. So what about if we're asking for the sum of indicators?

$$I = I_1 + I_2 + \cdots + I_n$$

When is I an indicator? Well, I is an indicator if and only if I is equal to 0 or 1 (by definition). Since each of the I_i can only have values 0 or 1, that means *at most* one of the indicator functions can be 1 at any given calculation. But this is true if and only if the events X_i are mutually exclusive, *i.e.*, no two events can happen at the same time. In this case, we get the following equality:

If the events X_i are *not mutually exclusive*, then we get what's known as *Boole's inequality*.

Wikipedia: Boole's inequality

We can notice a couple things from this inequality by looking at the indicator functions. Let $I = I_1 + \cdots + I_n$ as before, then the right hand side is equal to $E(I)$. The left hand side, we can view it as the probability that at least one thing is true. In other words $P(I \geq 1)$. So we can rewrite the above as:

We can generalize this for arbitrary expected value to get what's called *Markov's inequality*.

Wikipedia: Markov's inequality

Example 27.1 Suppose we have a (non-negative) random variable X and we know its expected value is equal to 5. What is the largest probability of success that $P(X \geq 100)$ could be?

28 Multiplication Moments

We've talked about how we can make random variables functions of one another. The example we gave wasn't that great, *but* we'll be using functions to see how we can easily calculate the expected value from one random variable into another.

Let $Y = f(X)$. Then

$$E(Y) =$$

There are particular functions which mathematicians like to look at in particular. Let $f(x) = x^k$ for some strictly positive integer k. Then

$$E(X^k) =$$

Wikipedia: Moment

The distribution X^k is known as the *kth moment of X*.

If $k = 1$ then we have the *first moment* which is usually called the expected value (or mean). If $k = 2$ then it's called the *second moment* or, alternatively, the *mean square*.

Example 28.1 Let's calculate the mean square of the discrete uniform distribution of the set $\{0, 1, 2\}$. First, notice that

$$E(X) =$$

The mean square is then given by:

$$E(X^2) =$$

Notice how in general $E(f(X)) \neq f(E(X))$. In other words

Let's use functions in order to see why $E(X + Y) = E(X) + E(Y)$. Suppose we have two random variables X and Y and we want to calculate $E(X + Y)$. To do this, let's make a function f which takes in two random variables and outputs a third: $f(X, Y) = X + Y$. By the above we have

Another thing to note is that if we multiply a random variable by a constant, then we can just pull the constant out:

$E(cX) =$

What happens when we try and multiply two random variables?

But we can't actually separate here! The only time we can actually separate is if X and Y are independent. In that case:

In other words, if X and Y are independent, then

29 Variance

When we were working on the binomial distribution, we had defined two new concepts: the expected value and standard deviation. We already handled the expected value, so let's move onto the standard deviation. We had that the standard deviation was defined as $\sqrt{np(1-p)}$ where np was the expected value.

What was the standard deviation keeping track of? It was keeping track of how far away from the expected value we were. If we look at what's inside the square root and we expand it, we get $np - npp = E(X) - E(X)p$. In other words, we get the expected value and then a little correction factor.

If we let $\mu = E(X)$ be our correction factor, then

Since we'll eventually want to take the square root, we avoid having a negative number by squaring our equation. This gives us what is called the *variance* of a distribution

Wikipedia: Variance

Wikipedia: Standard Deviation

The *standard deviation* is then the square root of the variance.

$$\mathrm{SD}(X) =$$

Example 29.1 Suppose we have a fair eight sided die. Find $\mathrm{SD}(X)$ where X is the number on the die after one roll.

What if we only cared about success/failure of an event? In this case, we look at indicator functions and the formulas above become much easier. Remember that whenever we have an event A, we can create a random variable X_A that indicates whether or not A succeeded. In other words: X_A is either equal to 1 or 0 and $E(X_A) = P(A)$. Then

since X_A is equal to 1 or 0 we have:

$$X_A^2 : \ 1^2 = 1 \quad 0^2 = 1 \ \Rightarrow \ X_A^2 = X_A$$

So we have that

$$\text{Var}(X_A) = E(X_A^2) - E(X_A)^2 = E(X_A) - E(X_A)^2 = P(A) - P(A)^2 = P(A)\left(1 - P(A)\right)$$

30 Standardization

If we have constants a and b then $E(aX + b) =$ _____.
Plugging this into the standard deviation we get:

We can use these formulas to create a standardized version of the expected value and the standard deviation. Let $X^\star = \frac{X - E(X)}{\text{SD}(X)}$ Then:

$E(X^\star) =$

and

$\text{SD}(X^\star) =$

These are called the *standard units* for the distribution X.

This should look super familiar! Let $\mu = E(X)$ and $\sigma = \text{SD}(X)$, then the fractions we're looking at are _____.

Example 30.1 The average height for men in Canada is 178 cm and the standard deviation is roughly 7.5 cm. Approximately what percentage of Canadian men are taller than 166 cm?

For women, the average Canadian is 164 cm with a standard deviation of 7 cm. Unfortunately there is no data on non-binary or intersex individuals.

31 Chebychev's Inequality

Since the standard deviation tells us how far away things will stretch from the expected value, there must be some relation between the two. It turns out that given a random variable, the probability that it differs from its expected value by more than k standard deviations is at most $\frac{1}{k^2}$. This is known as *Chebychev's inequality*:

Wikipedia: Chebyshev's inequality

Example 31.1 Say we're working for a financial company who are looking at all the transactions of their customers. They notice that given one billion transactions, the average transaction is roughly \$20 and if we square the values of the transactions, the average becomes \$404. Find an upper bound on how many transactions are over \$50.

32 Central Limit Theorem

Before we get into some laws and theorems, we're quickly going to state what variance looks like under addition. If X_1, X_2, ..., X_n are all mutually independent, then

This doesn't work for dependent variables unfortunately! For example, if $X = Y$ (so they are dependent on one another) then we have

$$\operatorname{Var}(X+Y) = \operatorname{Var}(2X) = \operatorname{SD}(2X)^2 = (2\operatorname{SD}(X))^2 = 4\operatorname{Var}(X)$$

and

$$\operatorname{Var}(X) + \operatorname{Var}(Y) = \operatorname{Var}(X) + \operatorname{Var}(X) = 2\operatorname{Var}(X)$$

Remember how for the binomial distribution we had something called the square root law where we basically stated that as we increase the number of trials then most trials would be close to the expected value. This works with any distribution. We'll go through this slowly.

First let's suppose we have n independent random variables X_i each with the same distribution X. In other words _____ for all i. Since expectation and variance are determined by distributions, we also know $E(X_i) =$ _____ and $\operatorname{Var}(X_i) =$ _____. We

let S_n be the sum of all the random variables:

$$S_n = X_1 + X_2 + \cdots + X_n$$

Then:

$E(S_n) =$ $\quad$ $\mathrm{Var}(S_n) =$ $\quad$ $\mathrm{SD}(S_n) =$

This gives us the *square root law*:

Theorem 32.1 (Square root law) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. Let $\bar{X}_n = \frac{S_n}{n}$ be the average value. Then*

We continue our trek by looking at the law of large numbers in this context. As n increases, we see that $\mathrm{SD}(S_n)$ will grow while $\mathrm{SD}(\bar{X}_n)$ decreases. This simple idea gives us the law of averages.

Wikipedia: Law of averages

Theorem 32.2 (Law of averages) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. Let $\bar{X}_n = \frac{S_n}{n}$ be the average value. Then for every $\varepsilon > 0$*

Notice that we don't have an approximation for S_n in the theorem above. That's because there is no simple formula for the distribution of S_n. Instead, we can use normal approximation to find a simple approximation for S_n.

Wikipedia: Central limit theorem

Theorem 32.3 (Central limit theorem) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. For large n, the distribution of S_n is approximately normal, i.e., $E(S_n) = nE(X)$ and $\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X)$. In other words:*

Where Φ is the standard normal CDF.

Wikipedia: Random walk

Example 32.4 In this example, we'll consider what's known as a random walk. The problem is normally told from a physics perspective as

this is where the idea came from.

Suppose you have an infinite line of slots and you put a particle in the middle. Each second, the particle moves left with probability p_ℓ, right with probability p_r or stays where it is with probability p_s, *i.e.*, $P_\ell + p_r + p_s = 1$. Let's suppose that the particle is having a lazy day. It stays where it is half the time $p_s = \frac{1}{2}$ and it moves with equal probability, *i.e.*, $p_\ell = p_r = \frac{1}{4}$.

There are roughly $86,400$ seconds in a day, so let's say we run this experiment for a little longer than a day. After $88,200$ seconds, what are the chances that the particle is 300 slots away to the right from where it began.

33 Skewness

Just like with the binomial distribution, sometimes our normal approximation isn't very good. We need to add some sort of correcting factor. Since in the central limit theorem we look at $\frac{S_n - E(S_n)}{\text{SD}(S_n)}$ we let $X_\star = \frac{X - E(X)}{\text{SD}(X)}$ and look at our approximations from this perspective. In this case, the first moment is given by:

$E(X_\star) =$

and the second moment is given by:

$E(X_\star^2) =$

In order to find how much we're off by, we look at the third moment, and define skewness in that way.

$$\text{skw}(X) = E(X_\star^3) =$$

If $S_n = X_1 + X_2 + \ldots + X_n$ where the X_i are independent each with the same distribution X, then

$$\text{skw}(S_n) =$$

We won't try and show where these formulas come because they are difficult to show. If you do want to try and prove it you can do it by first showing:

$$E\left((S_n - E(S_n))^3\right) = nE\left((X - E(X))^3\right)$$

and

$$\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X)$$

Putting these together gives you the above formula.

As a quick example when $n = 2$ we have the following:

$$\begin{aligned} E\left((X_1 + X_2 - E(X_1 + X_2))^3\right) &= E((X_1 + X_2)^3) + 3E((X_1 + X_2)^2)E(X_1 + X_2) \\ &\quad + 3E(X_1 + X_2)E(X_1 + X_2)^2 + E(X_1 + X_2)^3 \\ &= E(X_1^3) + 3E(X_1^2)E(X_1) + 3E(X_1)E(X_1)^2 + E(X_1)^3 \\ &\quad + E(X_2^3) + 3E(X_2^2)E(X_2) + 3E(X_2)E(X_2)^2 + E(X_2)^3 \\ &= 2E((X - E(X))^3) \end{aligned}$$

and

$$\begin{aligned} \mathrm{SD}(X_1 + X_2) &= \sqrt{\mathrm{Var}(X_1 + X_2)} \\ &= \sqrt{\mathrm{Var}(X_1) + \mathrm{Var}(X_2)} \\ &= \sqrt{2}\sqrt{\mathrm{Var}(X)} \\ &= \sqrt{2}\,\mathrm{SD}\,X \end{aligned}$$