Week 4

3 June 2021

16 Confidence intervals

This idea of approximating the binomial distribution through the normal curve brings up an important topic that we'll discuss now. If we're doing a series of independent trials and we're recording how often we get successes, how do we know how close our rate is to the actual probability of success? For example, if I roll a die three times and I get 6 once, how do I know that 6 doesn't appear $\frac{1}{3}$ of the time? What if I told you it's not a fair die?

This is where we can take the ideas from the law of large numbers and use them to our advantage. Remember how the higher our n is, the more likely that our relative frequency was going towards our unknown probability. So if we had performed the die roll three trillion times and one trillion times we had a 6 show up, then we'd probably state that it was biased towards 6 at $\frac{1}{3}$ the time. Letting $\_\_$ denote our relative frequency, we know as n increases $\_\_\_\_\_$, but how can we measure it?

We can actually use the normal approximation. As we increased z we saw that $\Phi(-z, z)$ got closer and closer to 1. So for example, if $z = 4$ then $\_\_\_\_\_$ which means that if n is large enough, we can be 99.99% certain that the number of successes $n\hat{p}$ differs from np by less than $4\sqrt{np(1-p)}$. This means that the relative frequency $\hat{p}$ will differ from p by $\_\_\_\_\_$. Recall that $\sqrt{p(1-p)/n} \leq 1/2\sqrt{n}$ and so $4\sqrt{p(1-p)/n}$ is at most $2/\sqrt{n}$. In other words p is somewhere in the interval $\_\_\_\_\_$. This interval is called a 99.99% *confidence interval for p*.

Wikipedia: Confidence interval (binomial)

This confidence level is *specific* to the binomial distribution. You can define confidence intervals for other distributions, but we won't cover that now.

Wikipedia: Confidence interval (general)

Example 16.1 Let's look at a couple of examples. Say I redo my rolling of a die example and I roll the die roughly one million times (so

n is very large). We note that 6 appears roughly $180,000$ times. Find a 99.99% confidence interval for the probability that the die rolls a six.

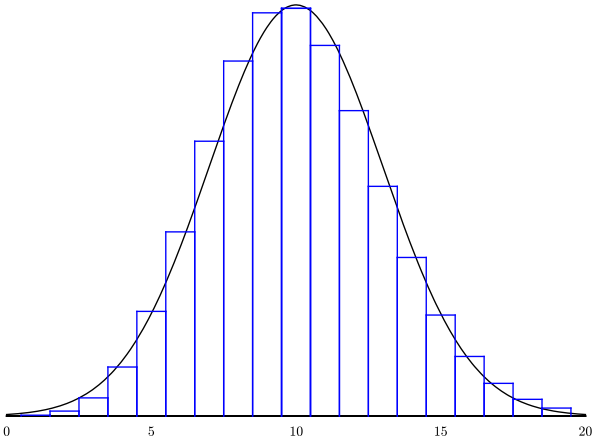
Notice that one of the key components to this approximation was the fact that $p(1-p) \leq \frac{1}{4}$. What this means is that the further from $\frac{1}{2}$ our probability p is, the less accurate this approximation becomes. Sure, we can just keep increasing n, but what happens if our probability is 1 in a bajillion? Then our approximation is not very good.

It turns out we can actually measure how good the normal approximation n is for a given n and p. Then let $N(a \text{ to } b)$ denote the normal approximation to a binomial probability $P(a \text{ to } b)$. Define $W(n,p)$ to be the *worst error* in the normal approximation to the binomial distribution, to be the biggest absolute difference between $P(a \text{ to } b)$ and $N(a \text{ to } b)$ over all integers a and b:

$$W(n,p) = \max_{0 \leq a \leq b \leq n} |P(a \text{ to } b) - N(a \text{ to } b)|$$

Running quick examples, you can check that for all $n \geq 10$ we know that $W\left(n, \frac{1}{2}\right) \leq 0.01$ and when $n \geq 20$ then $W\left(n, \frac{1}{2}\right) \leq 0.005$. This is why for $p = \frac{1}{2}$ this approximation is really good.

Now, suppose that $p \neq \frac{1}{2}$. Let's look at an example and see what happens and we might want to do to correct the problem. In the plot below, I set $n = 100$ and $p = \frac{1}{10}$.



What you'll notice is that the binomial distribution is slightly skewed to the left of the normal curve. This push to the left is known as the

skewness of the distribution. So what we will try and do is push the normal curve a little to the left in order to correct for this. We won't go to far into the details, but it turns out what we want to do is first take the third derivative of our function. Then

Wikipedia: Skewness

Wikipedia: Skew normal distribution

We then call the *skew-normal curve* the curve with the equation:

where $\text{skw}(n,p) = \frac{1-2p}{\sqrt{np(1-p)}} = \frac{1-2p}{\sigma}$.

Note that this skew function is *only* true for the binomial distribution. In general, it's a lot more complicated.

Notice that when $p = \frac{1}{2}$ then $\text{skw}(n,p) = 0$ giving us that $f(z) = \varphi(z)$. Depending on our choice of p then $\text{skw}(n,p)$ can either be positive or negative. It's positive if $p < \frac{1}{2}$ when the distribution is *skewed to the right*. It's negative if $p > \frac{1}{2}$ when the distribution is *skewed to the left*.

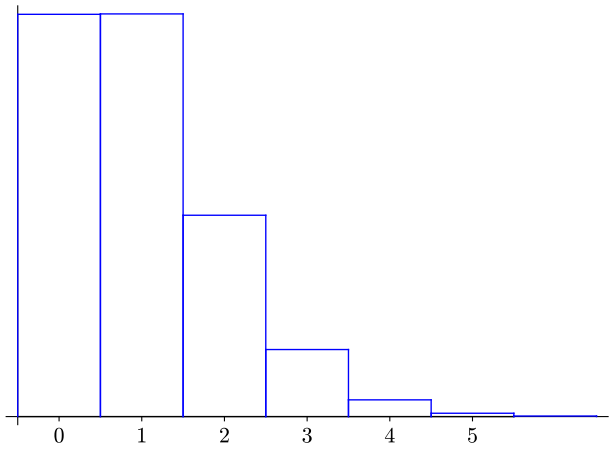
This gives us a second way to approximate the binomial distribution. For n independent trials where p is the probability of success, then

The term involving $\text{skw}(n,p)$ is known as the *skewness correction*.

Example 16.2 Let's use this approximation in an example. Suppose I want to calculate the distribution of 0s in 100 randomly selected digits. What is the probability of getting 4 or fewer 0s?

17 Poisson Approximation

Although normal approximation works nicely for probabilities near $p = \frac{1}{2}$, the further we get from 0.5, the worse the approximations will be. We tried to fix this by moving our approximation left and right, but even that has limits. The main issue is that the normal curve is symmetric (the left side and the right side look the same), whereas the binomial distribution is not a perfect curve. For example, we saw something like this earlier. If we let $n =$ _____ and $p =$ _____ we get the following distribution



It's so close to the 0 that there is no way we can make a nice bell curve that matches this. In this case we had let $p = \frac{1}{n}$ which means that the expected value is $\mu = np = \frac{n}{n} = 1$. If we follow this train of thought, as we do more and more trials, $n \to \infty$ and $p = \frac{1}{n} \to 0$, but μ stays at __. We can use this information to describe a curve which closely resembles the binomial distribution in this case. This "limit distribution" is called the Poisson distribution with parameter μ since we keep μ constant, but we take the limit as $n \to \infty$.

The *Poisson approximation* of the binomial distribution when n is large and p is small is given by

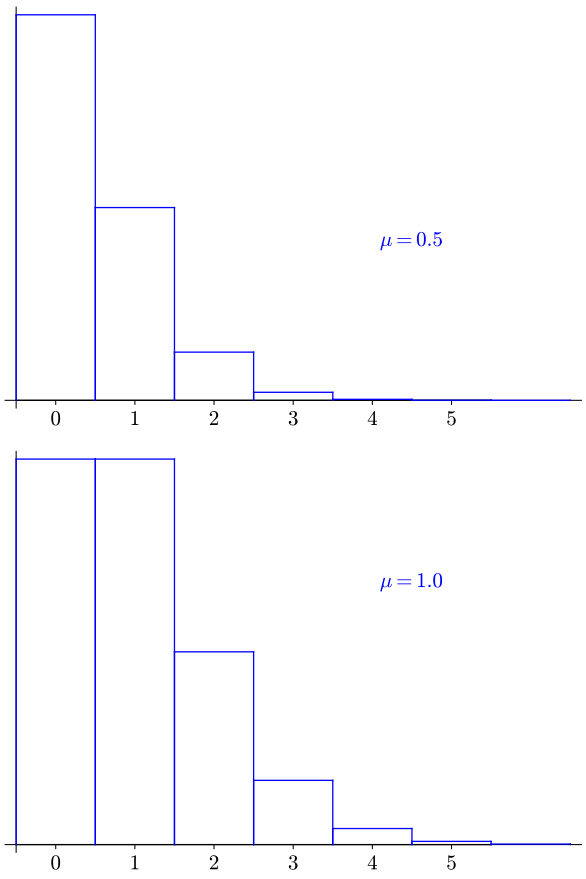
Just like with the normal approximation, the Poisson approximation is also itself a distribution. The *Poisson distribution* with parameter μ is the distribution given by:

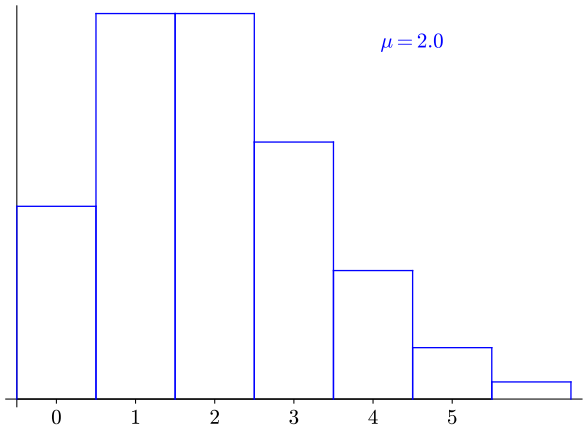
Wikipedia: Poisson distribution

Let's look at a couple examples.

Example 17.1 Every time you see a formula "$x - y$" there's a 1% chance you will copy it as "$x + y$". If you have to do 200 problems, what are the chances you will copy over 2 or more formulas incorrectly?

So what do these distributions look like as we change μ?





18 Random Sampling

Throughout the course so far we've seen random sampling various times without ever calling it random sampling. The idea of random sampling is to start off with a large population and to designate two categories of people. These categories might be something like "needs glasses" and "doesn't need glasses" or it might be "likes the colour red" and "dislikes the colour red". Basically the categories should be "opposite" of one another. We then take a small portion of the large population (a "random sample") and we look at the proportion of people that are in each category. This helps us figure out the proportion of people in each category for the population at large. Ideally, this would tell us exactly the proportion in the large population, but this is not always the case. Additionally, if we already know the proportion in the large population, we can ask what are the chances of a certain distribution happening in the smaller sample. We look at these questions through two main sampling methods

(1)

(2)

18.1 Sampling with replacement

Let N be the number of people in a large population and let n be the number of people in a sample which are drawn from the large population. These n individuals are drawn one at a time from the large population where each individual has the same chance of being chosen. After being drawn, they are put back into the large population (so a person *could* be chosen more than once). Therefore we would have a sequence of

length n of a set with N elements. This implies we have N^n different possible sequences which are all equally likely! Since we keep putting the individuals back into the large population, n *could* be larger than N.

An easy way to think about this is a bag full of N marbles. You take out a marble at a time and you record the colour and put the marble back into the bag. You do this n number of times. The question then becomes how well does the recording of the colours match with the actual distribution of colours in the bag of marbles?

Say R of the marbles are red and the rest B are not red (suppose they're blue just to make notation easier). Then _____. The distribution of red marbles is given by the probability $p = \frac{R}{N}$ which is the number we're trying to find. Notice here that we are working with a binomial distribution! Since p is static, the _____ approximation is a good approximation for the binomial distribution with parameters $\mu = np$ and $\sigma = \sqrt{npq}$. By the _____, if n is large enough then our sample of n marbles is highly likely to give the correct proportion of red to blue marbles. If you recall, from the section on confidence intervals, if we want a confidence interval of 99.99% we need to look at the interval: $\hat{p} \pm \frac{2}{\sqrt{n}}$ where $\hat{p}$ is the observed probability in our sample. If we only want a 95% accuracy, we can just look at the interval $\hat{p} \pm \frac{1}{\sqrt{n}}$.

Example 18.1 Say we have a bag with 20 marbles and we want to know how many are red. We pull out a marble, look at it's colour and put it back in. After doing this 35 times we notice that we have counted 25 red. What is the likely probability of a marble being red in the bag of 20 marbles? How about if we only pulled out 10 marbles and noticed 7 are red?

Recall that the other question we might want to ask is, if we already know the proportion in the large population, what is the probability of getting a certain distribution in the smaller sample? In this case, this should be fairly easy to calculate! If we already know that the probability of getting a red is $\frac{1}{p}$ then we have:

18.2 Sampling without replacement

Let's set things up the same as before, N individuals and we pull out n samples, but this time, we don't put the samples back in. Once we've taken someone out of the large population, we leave it out. This is just an ordering of n elements out of N individuals. So that means we'll have $(N)_n =$ _____ different options, which will be much less than our previous way of doing things.

In this case, we'll only look at the second question: if we already know the probability in the large population, what is the chances of getting the probability in n pulls?

So say that we have N total marbles and R of them are red and $B = N - R$ are blue. Then we know that the probability of red is given by $\frac{R}{N}$. Now we ask, say I pull out n marbles (without replacement), what is the probability that r of them are red?

First let's look at if the first r we pull are red and the rest $n - r$ are blue. Then we have:

for the chances of pulling a red out each time. Then, the rest of them are blue and so we need to multiply the above by:

Multiplying these two together we have:

But this is just for *one* option. We need to actually look at all combinations of r red appearing in n pulls. That means, we need to multiply the above by $\binom{n}{r}$ different combinations.

So in total we have:

19 Random Variables

We've been working on just one distribution so far: the binomial distribution. But if you were to look through Wikipedia, you'd notice that there are a ton of different distributions out there! The goal is to now generalize everything we have done into a more general setting. We start off with something that has been awkwardly missing in a math class: variables.

So far when we've talked about events, we've talked about them as subsets A of a sample space Ω. This works most of the time, but it can be hard to know what A is. For example, a lot of times we wrote $P(3)$ or something to mean "the probability of getting a 3". Instead of using the event "get a 3" which is represented by the subset $\{3\}$ we decided to just write the number 3. This is what we're going to try and generalize and we're going to create a whole new system for writing things out.

Variables have been used throughout mathematics as a placeholder for information, so it makes sense that we would use variables for our new system. We normally have these variables be large capital letters like X or Y, but they can be anything you prefer and they normally take the place of normal numbers. For example, if we are rolling a die, we might be interested in the event "the number 6 is rolled". To add a variable, we'd normally replace a number by a variable and go with that, *i.e.*, "the number X is rolled". Since we've been using capital letters for events, we'll let X be the variable we put into an event "the number X is rolled". In this case X is called a *random variable*.

Wikipedia: Random variable

As other examples, the random variable X might represent "the number X is rolled on a die" or "the side X of a coin is flipped" or "the

number X is drawn out of a hat", etc. Notice how in each case we don't actually say *which* number is rolled or *which* side is chosen, instead we let the random variable represent it.

How does this help? Well, let's look at the event from before represented by the random variable X: "the number X is rolled". We can let $X = 6$ mean _____. Notice how it gives us language to create new events. So if I were to say $X = 3$ then you'd know that we mean _____ in a much more compact way. We can then place these random variables into our probability function like _____ in order to mean _____.

At this point you might be asking, "why not just do something like _____" like before? Why are we complicating things? Because complicating things sometimes makes things less complicated! For example, what about if I want to ask for the event "the number rolled is less than or equal to 3"? Now, we can just write _____ and we know exactly what we're talking about! Even with more complicated events like "Let A be the subset $\{2, 4, 6\}$ of all even numbers and find $P(A)$", we can do this easier with our new random variables. We can represent the above as _____.

A random variable doesn't always have to represent a number! If we look at flipping a coin, I can ask for $P(X = H)$ where H represents getting a heads. These random variables can mean any particular outcome in our sample space!

Here's a quick table of everything from above to maybe help. (Suppose we are rolling a fair six-sided die)

English language	Random Var.	Subset	Probability
Number on the die is 6	$X = 6$	$\{6\}$	$\frac{1}{6}$
Number on the die is less than 3	$X < 3$	$\{1, 2\}$	$\frac{1}{3}$
Number on the die is x	$X = x$	$\{x\}$	$\frac{1}{6}$
Number on the die is less than x	$X < x$	$\{1, 2, \ldots, x-1\}$	$\frac{x-1}{6}$
Number on the die is in the subset B	$X \in B$	B	$\frac{\lvert B\rvert}{6}$

If we use X to help define an event A, then we say that the *event A is determined by X*. Although we *can* write $P(A)$ to let us know the probability, we will henceforth start writing $P(X \in A)$ to show that A has some variable in it. If we go over all possible subsets A, we (must) get a distribution which we call the *distribution of X*. In essence, the outcome of any particular outcome x is given by _____ and of any subset, via the addition rule, by _____

At this point, it might confusing and you might ask what the difference between x and X are. This is super confusing, so don't worry! A random variable is just a normal variable that represents certain things

inside of events. When we talk about the probability of an event then we must state what the random variable is equal to for the probability to make sense. We can have the random variable be equal to a standard variable (_____), but we can also have it mean anything else! One thing that never makes sense is _____ as this doesn't tell us anything.

The book uses the term "dummy variable" to mean a standard-/normal variable.

So say I have an event"I roll a five sided die and a I get an X" and an event "I pull the number Y out of a hat". Then if were to say _____ then what we mean is that the probability that we get u when rolling a five sided die is the same as the probability of pulling out v from out of a hat.

Aside: As an aside, if you look at the definition of a random variable in Wikipedia, it might be a little confusing since we're not defining it in exactly the same way. Since in mathematics we like to be precise, the exact definition of a random variable is given through a certain type of function. So X would be $X : \Omega \to E$ where E is some space (normally $\mathbb{R}$ for us). Then, more precisely we have $P(X \in A) = P(\{x \in \Omega \mid X(x) \in A\})$. Don't worry about this to much, but it's a good thing to know.

Example 19.1 Let's look at a quick example of what this looks like. Say I roll a fair six-sided die and I want to calculate some probability functions.

$$P(X = 2) =$$

$$P(X \leq 4) =$$

$$P\left(\frac{X}{2} \in \mathbb{Z}\right) =$$

20 Functions

Sometimes we want to look at random variables as a function of another random variable. Normally we see this in standard variables by: $y = f(x)$. Doing this for random variables we get: $Y = f(X)$. What does this mean though?

It means that if X has some value then Y has the value $f(X)$. The two are related to one another through the function. This implies that the distribution of Y can be derived from the distribution of X.

Example 20.1 Suppose we're rolling two dice and we want to calculate the sum of the two values. We've done this many times, but we're going to be using functions this time.

As before, we let (i, j) represent a roll of the two dice. If we let X be the sum of our dice, then we know that we get a distribution like the

following:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

These were calculated from the (i,j). Let Y represent the outcomes (i,j) and let f be the function $i+j$. In other words _____.

This is obviously a super easy pointless example. We're not really gaining any meaningful insights. What we *do* learn though is that we can use functions in conjunction with our variables. Addition doesn't do much, but there are limitless possibilities!

21 Joint Distributions

Let's take this one step further. What happens if an event has potentially more than one variable? For example, "What is the probability that I get the 5 of spades in a normal deck of cards?" I could technically represent this as one variable: "What is the probability that I get a X in a normal deck of cards?". But what if I want to ask for the probability of getting a 5, regardless of suit? I can no longer ask that! In this case it makes more sense to put in two variables: "What is the probability that I get a X of Y in a normal deck of cards?". Now we can ask actually ask for the probability! We would get _____ where S is the set of suits. How about our original example? This we can write as _____.

Wikipedia: Joint distribution

Whenever we have two or more random variables in a distribution we call it a *joint distribution*. Joint distributions can be a little weird to work with, but once you've done a few examples, they make more sense.

We'll look at two examples to try and make things more understandable.

Example 21.1 First we'll look at an example of pulling marbles out of two bags. Say that each bag has 3 red marbles and 1 blue marble. Our event is represented by "What is the probability of pulling out a X

out of bag 1 and a Y out of bag 2". Let's see what the chances are using a table.

	X =Red	X =Blue	Total
Y =Red	$\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{16}$	$\frac{3}{4}\cdot\frac{1}{4}=\frac{3}{16}$	$\frac{12}{16}=\frac{3}{4}$
Y =Blue	$\frac{1}{4}\cdot\frac{3}{4}=\frac{3}{16}$	$\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{16}$	$\frac{4}{16}=\frac{1}{4}$
Total	$\frac{3}{4}$	$\frac{1}{4}$	1

So we can read this table by looking at each column and row for each random variable. For example, $P(X=\text{Red}, Y=\text{Blue})$ gives us ____.

We can also technically look at any particular row and column to get the added probability. For example, if we want to know the probability that the first marble is red, then we're asking for X to be "Red". We can represent this as:

In particular, we can always use that equality whenever we have multiple random variables:

Example 21.2 Let's look at the previous example, but just look at one bag. So say we have one bag with 3 red marbles and 1 blue marbles and I pull out two marbles (one after another without replacement). Here I can represent this as the event "I pull out marble X then I pull out marble Y". So looking at the probability table we have:

	X =Red	X =Blue	Total
Y =Red			
Y =Blue			
Total			

We say that two random variables are *equal in distribution* if

The book calls this *same distribution*

If two random variables are equal in distribution, then we can change their variables whenever we want. For example _____ and we'll get the same thing!

We say that two random variables are *equal* if

Another way of saying this is:

Equal in distribution basically says the right column and the bottom row must be the same. Equal says the sum of the (main) diagonal must be 1

Notice how we can take this idea further:

$P(X < Y) =$

or even further:

$P(X+Y = z) =$

Example 21.3 Calculate the distribution of $X + Y$ if we roll two fair six-sided dice.

22 Conditional Distributions

Joint distributions might be confusing, but if you think about it, we had already been working with joint distributions for a while! Whenever we had a conditional probability statement $P(A \mid B)$, we were working with two different events/variables at the same time. The only issue was that it wasn't done in the framework of random variables. We now think about everything through the eyes of a random variable.

So let's now say we want to replace the event A by a random variable. We saw that $P(A)$ is written as _____ using random variables, so what would our conditional probability look like?

What if we now replace B with a random variable? We get:

Two random variables, easily put in! And just like before, we can switch these variables out to whatever we like. In particular, we'll define the *conditional distribution of X given $Y = y$* as

What happens when the events coming from X and Y are independent? What does independence mean in this case? We had this definition for independence $P(A \cap B) = P(A) \cdot P(B)$ but how do we move that over to joint distributions? We say that two random variables are *independent* if