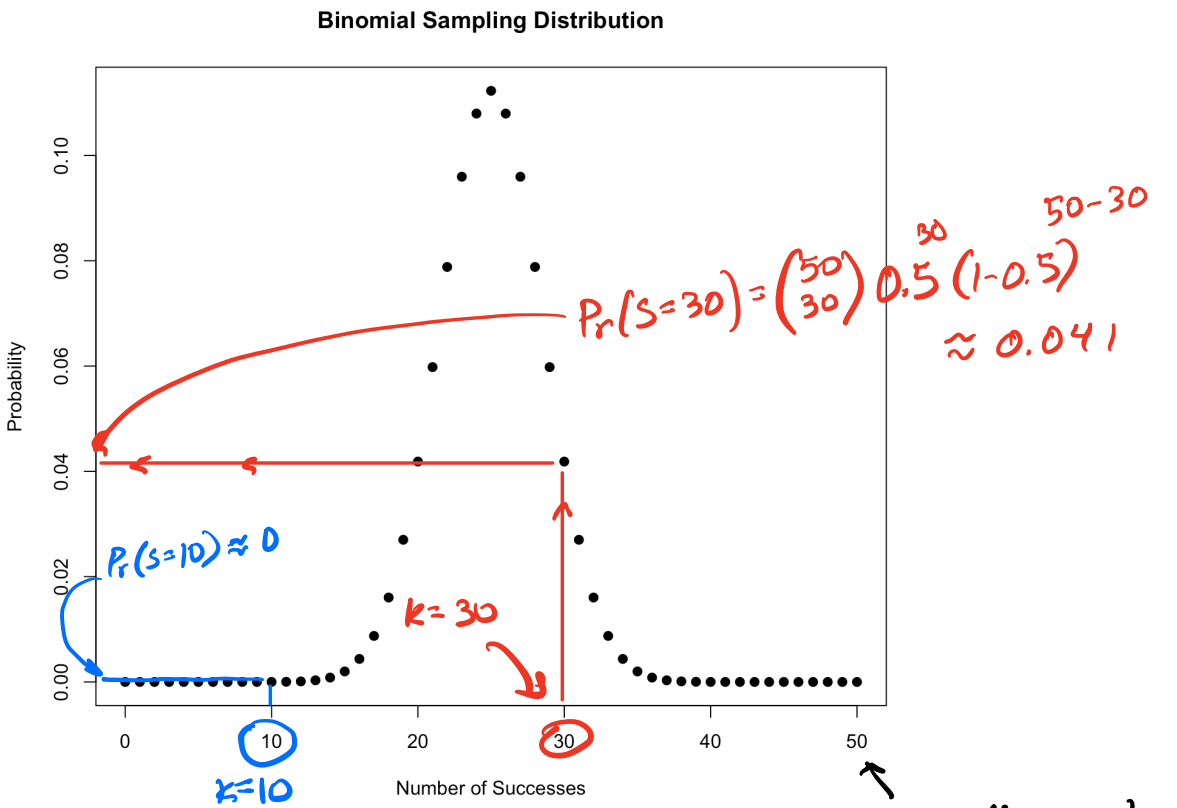
Sampling Distributions

We've been working to derive sampling distributions from probability distributions defined at the individual level. One of the most simple probability distributions at the individual level is the Bernoulli distribution, which models the probability of a categorical character with two categories. We then used probability trees to show that the sampling distribution for a binary character is given by a Binomial distribution. Here is an example with n = 50 individuals and Pr(success) = 0.5.



Let S be the number of successes in the sample.

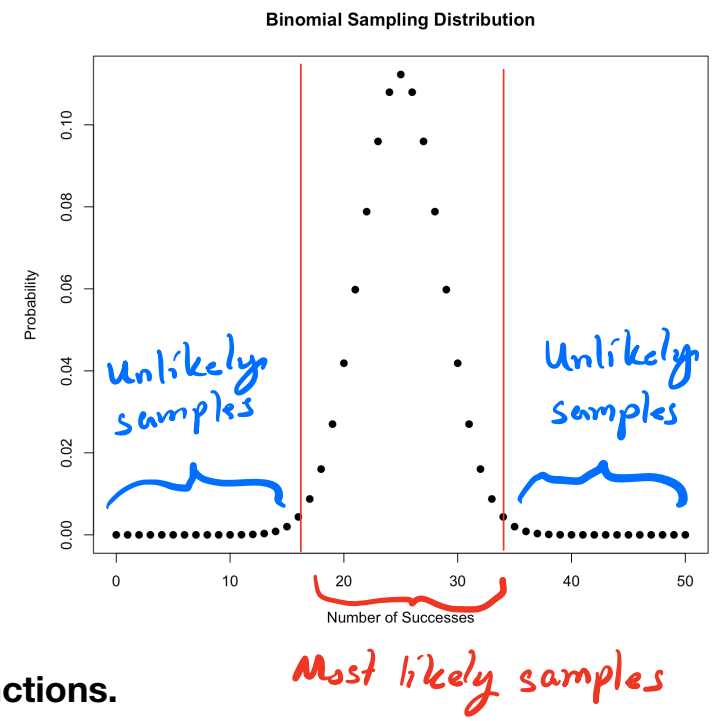
$S \sim Binomial(n=50, p=0.5)$

$Pr(S=k) = \binom{50}{k} 0.5^{k}(1-0.5)^{50-k}$

This distribution tells us how likely we are to see a certain set of sample properties, namely the number of successes in our sample of 50.

We can see that our most likely sample is one where 25 of the individuals are "successes". That should make some sense. We built this under the assumption that an individual has a 50:50 chance of being a "success".

We can use the graph of the sampling distribution to give us a sense of what is likely to happen and what is unlikely to happen based on the height of the curve.



Here, I have drawn that some are likely and the others are unlikely. Values near 25 are likely but values too small and too large are very unlikely. That should also make sense. This is just like saying, if I flip 50 coins, it's highly unlikely I only get a handful of heads (say 0 to 16) and its also highly unlikely they are almost all heads. But where do we make the cutoff? To do that, we need to use **cumulative density functions.**

Cumulative Density Functions

Cumulative density functions show how probability *accumulates* as you mean from small valued outcomes to large valued outcomes. Let's see how this works in practice before going any further.

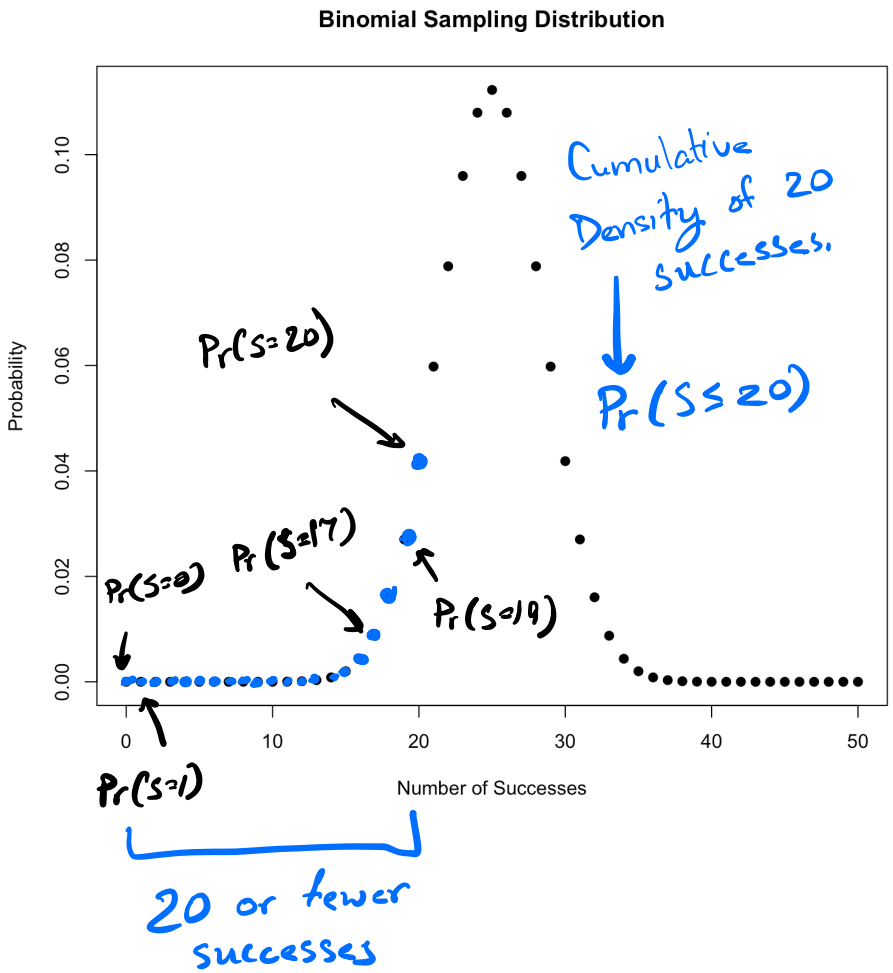
For the binomial, the cumulative density function for *k* successes defines the probability of seeing *k* successes or fewer. Mathematically, we write this as

$$Pr(S \leq k) = Pr(S=0) + Pr(S=1) + Pr(S=2) + \ldots + Pr(S=k-1) + Pr(S=k)$$

Cumulative density of k successes

Probability of 0 successes

Probability of K success



If $k = 20$, $Pr(S \leq 20)$ is the sum of the colored points.

Sum of the blue points is $Pr(S \leq 20)$.

<u>Cumulative distribution functions are helpful for answering how probable ranges of outcomes are.</u>

For example, we might to know what the chances are that, if the population fraction of successes is 50%, we would get a sample with between 40% and 60% successes. That is, if we sample 50 individuals, what is the probability that we get between 20 and 30 successes (20/50 = 0.4 and 30/50 = 0.6)?

Here is what that looks like in probability form.

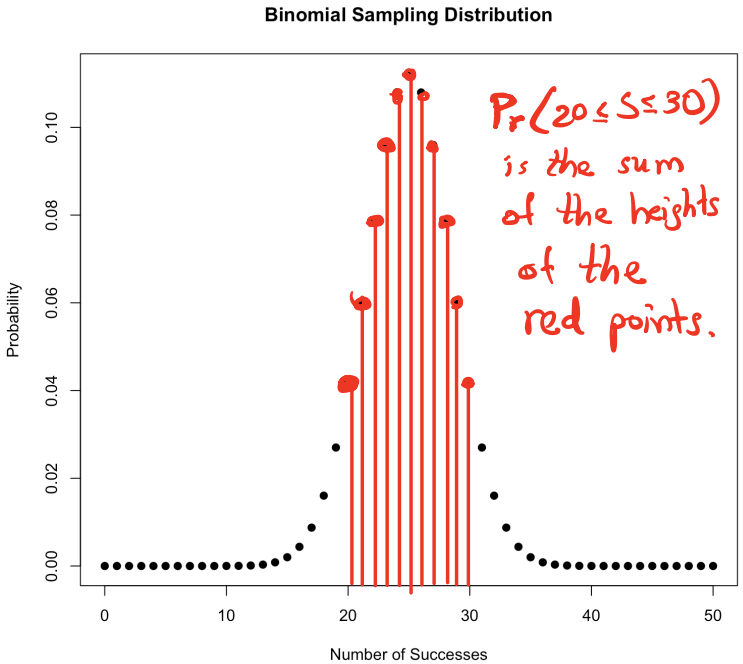
$$Pr(20 \leq S \leq 30) = Pr\left(0.4 \leq \frac{S}{n} \leq 0.6\right)$$

S = #successes

$\frac{S}{n}$ = $\frac{\#\text{Successes}}{\text{Sample size}}$ = fraction successes.

$$Pr(20 \le S \le 30) = Pr(S=20) + Pr(S=21) + Pr(S=22) + Pr(S=23) + \ldots + Pr(S=29) + Pr(S=30).$$

Here is what that looks like graphically.



We could find this probability by adding up the height of every red point.

But this would be tedious, especially when we get to very large numbers of possible outcomes.

Rather than add them all up, we can use a special property of cumulative distribution functions.

To see how, notice that we can find the red value as the difference between two cumulative probabilities. In math terms, we have

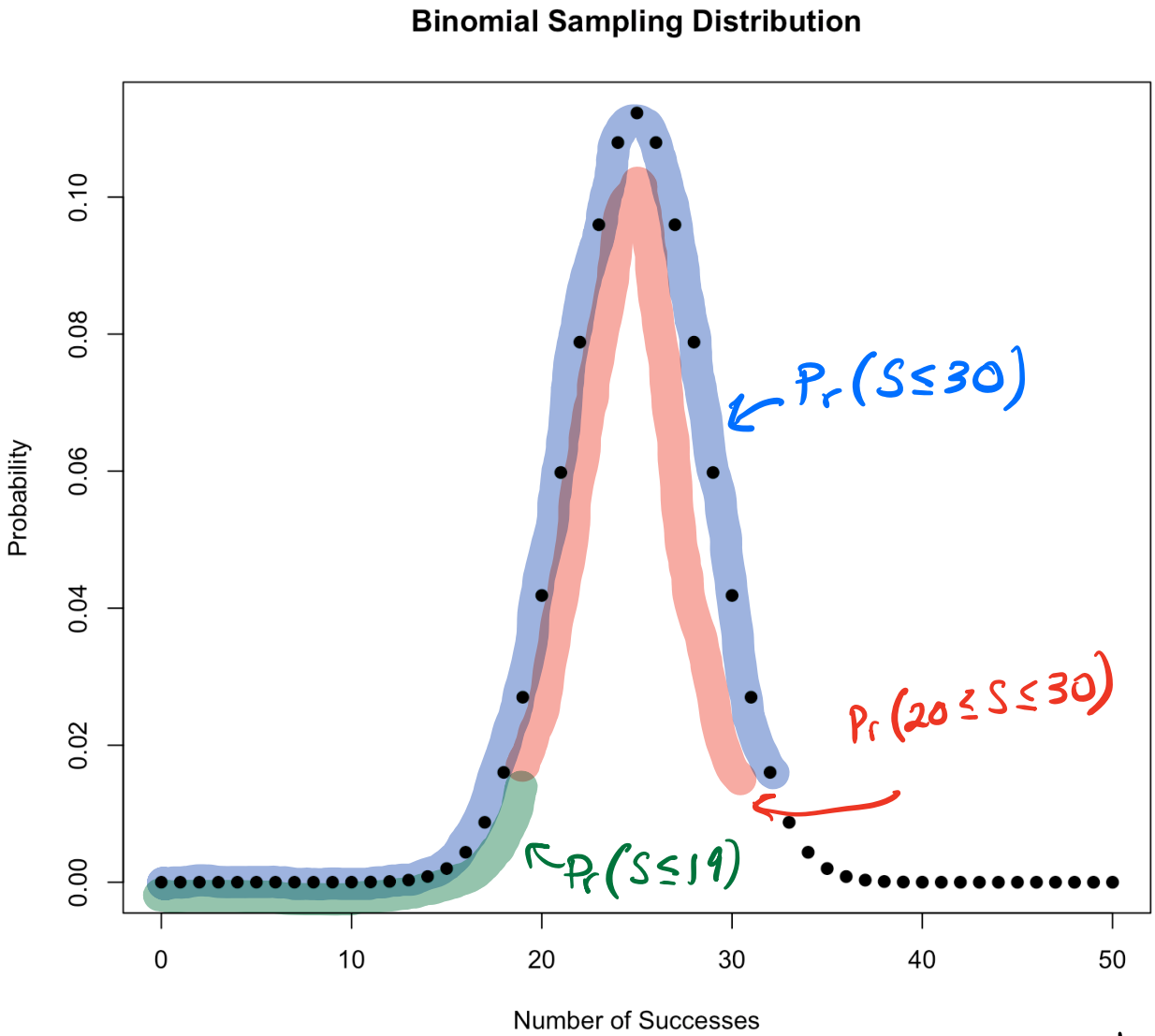
$$Pr(20 \le S \le 30) = Pr(S \le 30) - Pr(S < 20)$$

$$= Pr(S=0) + Pr(S=1) + \ldots + Pr(S=19) + Pr(S=20) + \ldots + Pr(S=30)$$

$$- Pr(S=0) + Pr(S=1) + \ldots + Pr(S=19)$$

$$= Pr(S=20) + \ldots + Pr(S=30)$$

Here is a graphical version of the exact same idea.



This is quite useful because cumulative densities are encoded in R. For example

$$Pr(20 \le S \le 30) = Pr(S \le 30) - Pr(S < 20)$$
$$= Pr(S \le 30) - Pr(S \le 19)$$

corresponding R code →

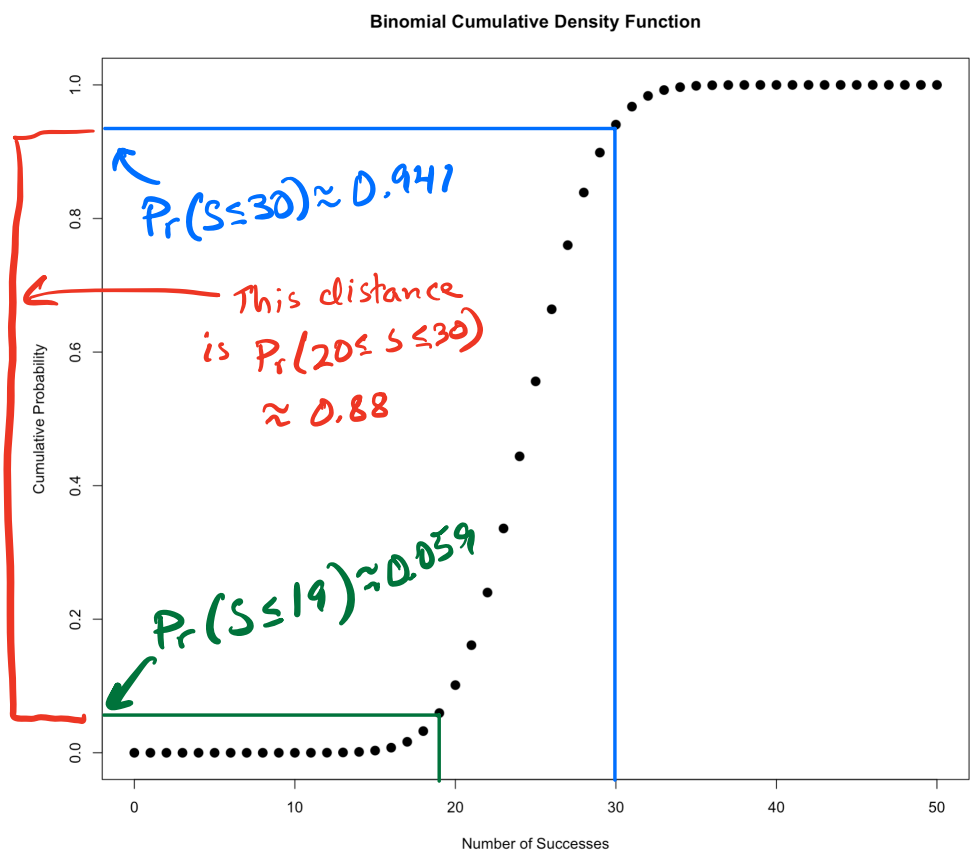
```
pbinom(30,50,0.5) - pbinom(19,50,0.5)
```

$$\approx 0.941 - 0.059$$
$$\approx 0.88$$

This structure tells us how likely a particular range of outcomes are that we pick and are interested in. When we specify the outcomes between 20 and 30, we see that the probability of seeing a sample in this range is about 88%. That represents a lot of outcomes.

In some circumstances, we want to specify the probability ahead of time and then ask which outcomes correspond to that probability. For example, what is the range of sample successes that we are likely to see 95% of the time, or 99% of the time. To do this, we invert the problem and look for **quantiles**.

To understand quantiles, we need the graph of the cumulative density function. It looks like this



On the x-axis is the number of successes in a sample. On the y-axis is the *probability of observing that many success or fewer.* This function only increases because the cumulative probability of any particular value contains the cumulative probability of every value smaller than it. This function is normally written with a capital *F* and a subscript indicating the random variable it applies to. Probability functions are often written with a lowercase *f* and a subscript that it applies to. For example, the probability density and cumulative density functions for this sampling distribution are

$S \sim \text{Binomial}(n = 50, p = 0.5)$

- S: random variable indicating the # of successes in a sample of 50.
- $n = 50$: sample size
- $p = 0.5$: individual probability of success.

Probability Density Function

$$f_S(k) = \Pr(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Cumulative Density Function

$$F_S(k) = \Pr(S \leq k) = \sum_{i=0}^{k} \Pr(S = i)$$

$$\Pr(S=0) + \Pr(S=1) + \cdots + \Pr(S=k-1) + \Pr(S=k)$$

When specify the outcomes and evaluate the probability, we are asking for the values of these cumulative density functions. This goes from the x-axis to the y-axis on the graph of the cumulative density function.

When we specify the probability and evaluate the outcomes, we are asking for the values that satisfy particular cumulative densities. This is akin to going from the y-axis to the corresponding x-axis on the graph of the cumulative density function.

Here is what that looks like mathematically

Specify k (outcome)

$Pr(S \leq k) = ?$

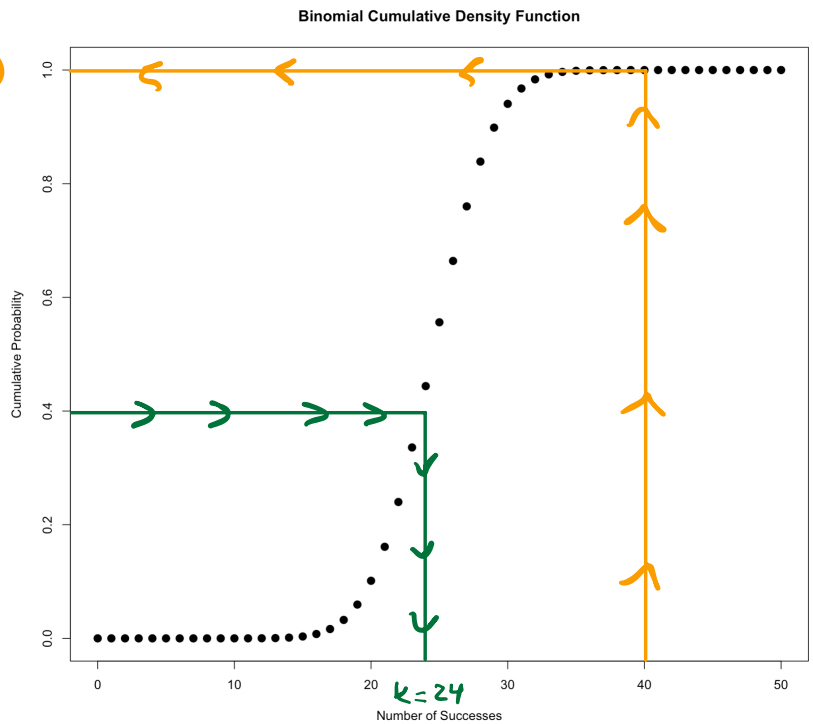
known — unknown

Specify P, the cumulative prob.

$Pr(S \leq ?) = P$

unknown — known

Here it is graphically



$Pr(S \leq k)$

Specify cumulative probability

$Pr(S \leq ?) = 0.4$

The value of *k* that corresponds to a particular cumulative probability is called a **quantile.**

Specify $k = 40$

To see why this is a quantile, let's start with a concrete example. Let's find the 50% quantile. The 50% quantile (i.e., the median) in **data** represents the value where half the data is small and half the data is larger. The 50% quantile in **probability** represents the value where half of all the observations in many many samples would be below this value and half would be above. That is to say, there is a 50% chance an outcome is smaller than this value (and a 50% chance any given outcome is larger). **This is exactly the definition of a cumulative density!**

Using quantiles to find the most likely outcomes

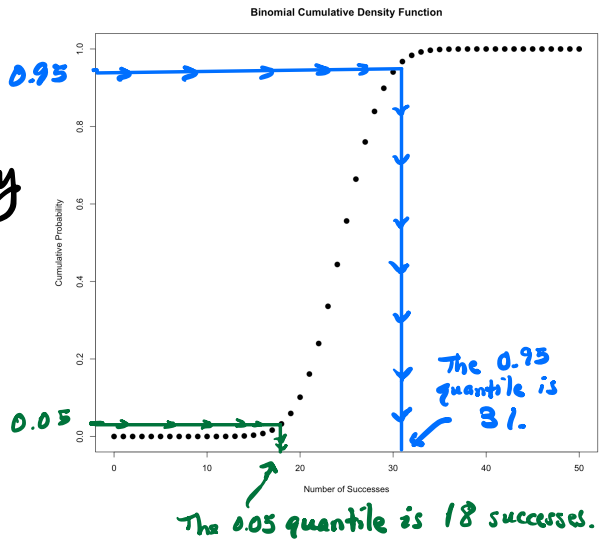
This way of finding quantiles is particular helpful for finding ranges of outcomes that are the most likely. To see this, let's return to our binomial distribution. The most unlikely outcomes are the very large and very small number of successes. Let's say we aren't interested in the 5% of least likely small numbers of successes. That means we want to find the value of *k* (the number of successes) that corresponds to a cumulative probability of 0.05.

Here that is in math.

$Pr(S \leq k) = 0.05$

We want to know the # of successes that makes this true.

We specify 5%.



We can also get rid of the 5% of least likely outcomes that are really large.
That means we want to find the value where Pr(*S* > *k*) = 0.05. Here it is in math.

$Pr(S > k) = 0.05 \rightarrow 1 - Pr(S \leq k) = 0.05$

We can do this because $Pr(A) = 1 - Pr(\text{Not } A)$.

If A is > k, then "Not A" is "Not larger than k", which is k or smaller ($\leq k$).

cumulative probability

$\rightarrow Pr(S \leq k) = 1 - 0.05 = 0.95$

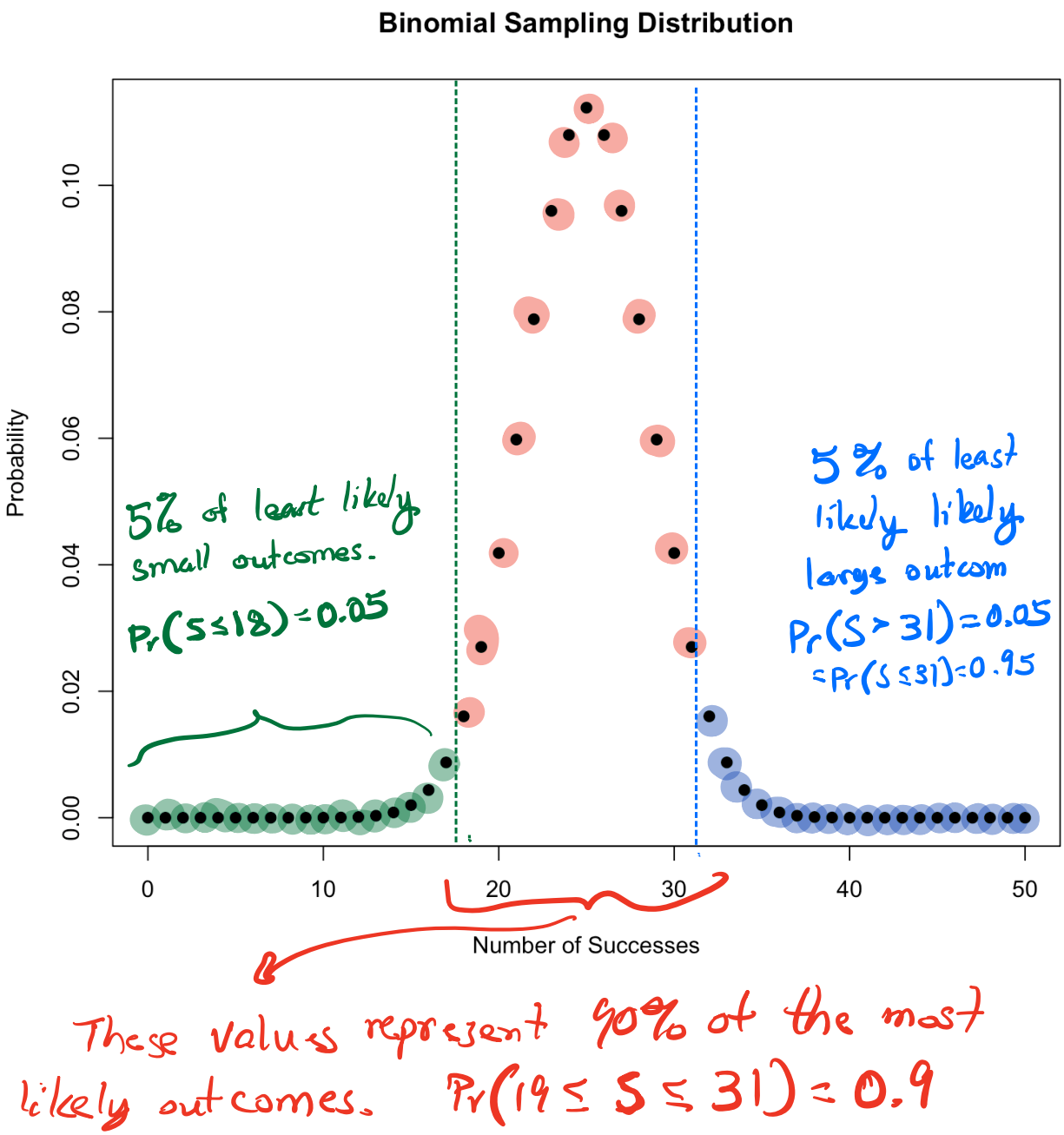
95% quantile

95% cumulative probability

The blue highlighted statements are the same. The k value that gives $Pr(S > k) = 0.05$ is the same k value that gives $Pr(S \leq k) = 0.95$

You can see from the cumulative density plot above that the *k* value that accumulates only 5% **above *k*** must be the *k* value that accumulates 95% probability **up to *k*.**

Now that we have found the 5% least likely small numbers of successes and the 5% least likely large numbers of successes, all that is left in the middle is 100% - 5% - 5% = 90% of the most likely values. Here that is mathematically and graphically using the probability density function.



In general, we can use the following approach to find the most likely outcomes from a probability distribution.

1. Define the probability you are interested in.
2. Look at the probability density function and see where the most likely and least likely outcomes are.
3. Break the probability density function into regions corresponding to the probabilities you care about.
4. Use the cumulative density function to find the quantiles corresponding to the break points to partition the outcomes into the regions of most and least likely.

Example

People are often interested in 95% of the most likely samples from a sampling distribution. The idea is, if we go out and collect a sample, we want to know a region of values we are likely to observe and values we are very unlikely to observe. For the binomial, this often looks like getting rid of the largest and smallest values, as we have done here (although not always). Say we have a sample of 200 beetles and we are interested in whether they are males or females. We have a strong suspicion that the sex ratio in this species is 50:50. As such, a suitable model for the individual characteristic (male/female) is a Bernoulli($p = 0.5$). If we let a 'success' be a female, then the number of females in our sample of 200 beetles follows a binomial distribution with $n = 200$ and $p = 0.5$. Formally, we write that S is the number of females in our sample and $S \sim \text{Binomial}(n = 200, p = 0.5)$.

Ahead of time, we could ask, what is a reasonable range of females we might see in this sample? Is only 2 females reasonable? Probably not. Let's find the 95% of most likely outcomes from our sample. We do the following calculation.

$$Pr(k_{small} \leq S \leq k_{large}) = 0.95$$ ← Statement of problem.

$$= Pr(S \leq k_{large}) - Pr(S < k_{small}) = 0.95$$

$Pr(S \leq k_{large}) = 0.975$

$\rightarrow Pr(S > k_{large}) = 0.025$

Least likely large outcomes

$Pr(S < k_{small}) = Pr(S \leq k_{small} - 1) = 0.025$

Least likely small outcomes.

We find these using the quantile function in R.

$Pr(S \leq K_{large}) = 0.975$

qbinom(0.975, size = 200, p = 0.5)

$K_{large} = 114$

$Pr(S \leq K_{small} - 1) = 0.025$

qbinom(0.025, size = 200, prob = 0.5)

$K_{small} - 1 = 86$

$\rightarrow K_{small} = 86 + 1 = 87$

There we go. 95% of the time we sample this population, we see between 87 and 114 females in our 200 beetles

In terms of the fraction female, we simply divide these outcomes by 200 to get

0.43 and 0.57.

Now we have an expectation. **If the true population fraction female is 50%**, then we expect the vast majority of our samples to estimate prevalence between 43% and 57%.

This kind of exercise sets up the potential for calculating both confidence intervals and hypothesis testing.