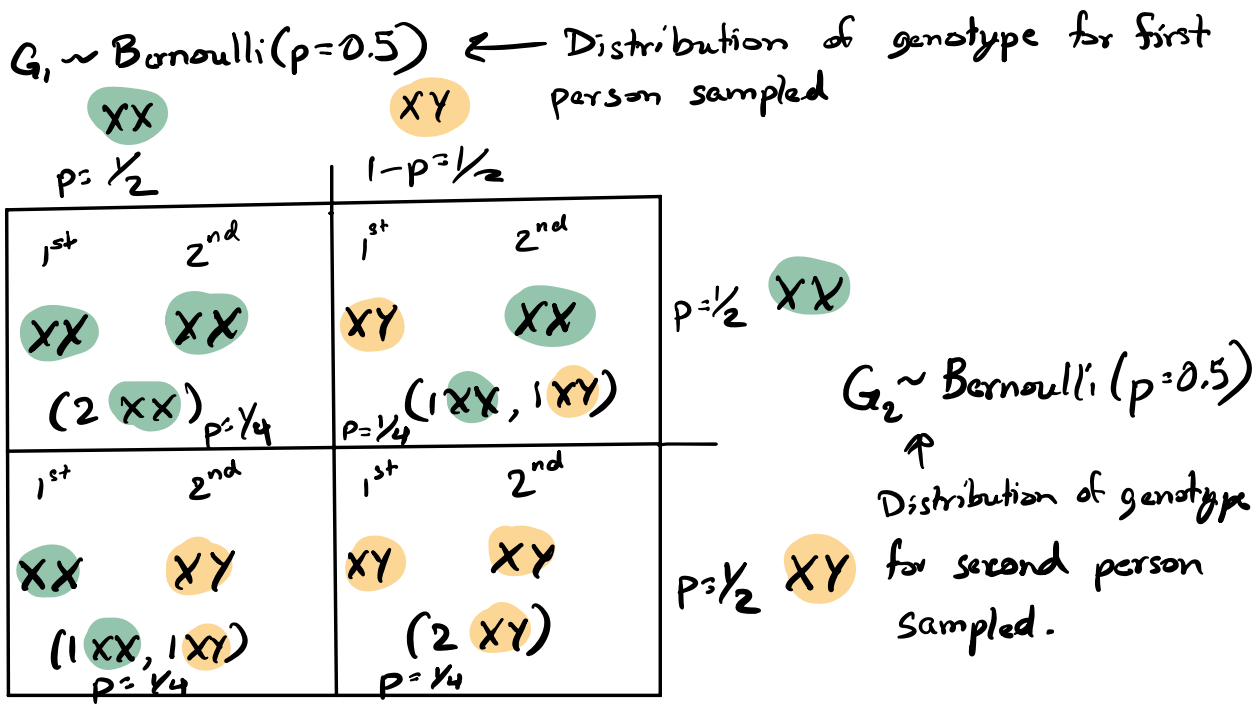
Probability of Sampling Distributions

We've now seen some examples of how probabilities are formulated and how they can be used to model, based on some assumptions, biological processes that lead to variable characters across individuals. The example we used before was for the probability of either XX or XY sex chromosome genotype. We saw that under typical assumptions you learn during an introductory biology class that the random variable giving sex chromosome genotype can be described by a Bernoulli distribution with parameter $p = 0.5$.

But this is a random variable of only a single individual. This probability suggests that, in the world population, humans have a 50:50 sex ratio at birth. But samples have only a few individuals in them, not the entire population. So how do we use probability to describe the process of sampling?

One way to do it is to think about the possible individuals we could sample when we look at just two people in a sample. Here's an example.



What this shows is that there are 4 mutually exclusive outcomes of a sample of 2 individuals. They are

1. {XX, XX} with probability 1/4
2. {XY, XX} with probability 1/4
3. {XX, XY} with probability 1/4
4. {XY, XY} with probability 1/4.

This is a <u>sampling distribution</u> because it shows all possible samples and their probabilities of occurrence.

Let's formalize this with the random variable S2.
S2 is the sample of XX-XY individuals in a sample of 2 individuals.

Let's see what S3 looks like.

← S2, the distribution of samples with 2 individuals.

1/4 {XX, XX}	1/4 {XX, XY}	1/4 {XY, XX}	1/4 {XY, XY}
1st XX 2nd XX 3rd XX	1st XX 2nd XY 3rd XX	1st XY 2nd XX 3rd XX	1st XY 2nd XY 3rd XX
1st XX 2nd XX 3rd XY	1st XX 2nd XY 3rd XY	1st XY 2nd XX 3rd XY	1st XY 2nd XY 3rd XY

p=1/2 XX

p=1/2 XY

$G_3 \sim \text{Bernoulli}(p=0.5)$

↑ Distribution of genotype for 3rd individual.

Each box is the same size, 1/8 the total area.

Each box also represents each possible sample when we look at 3 individuals who each have the same probability of being XX or XY. This is the S3 sampling distribution.

Often in statistics, we don't care about the exact order of the samples. We just care about how many have a particular characteristic.

For example, if the question is about sex ratio, do the samples

$\{XX, XY, XX\}$, $\{XY, XX, XX\}$, or $\{XX, XY, XY\}$

differ? Not really. For questions related to sex ratio, only the number of one genotype matters. And these three samples, while different, all have the same number of XX (=2) and XY (=1) individuals.

From this perspective, let's reframe S3 by the number of XX individuals in the sample.

Let Y be the # of XX individuals in the random variable S3. Here is a picture of Y.

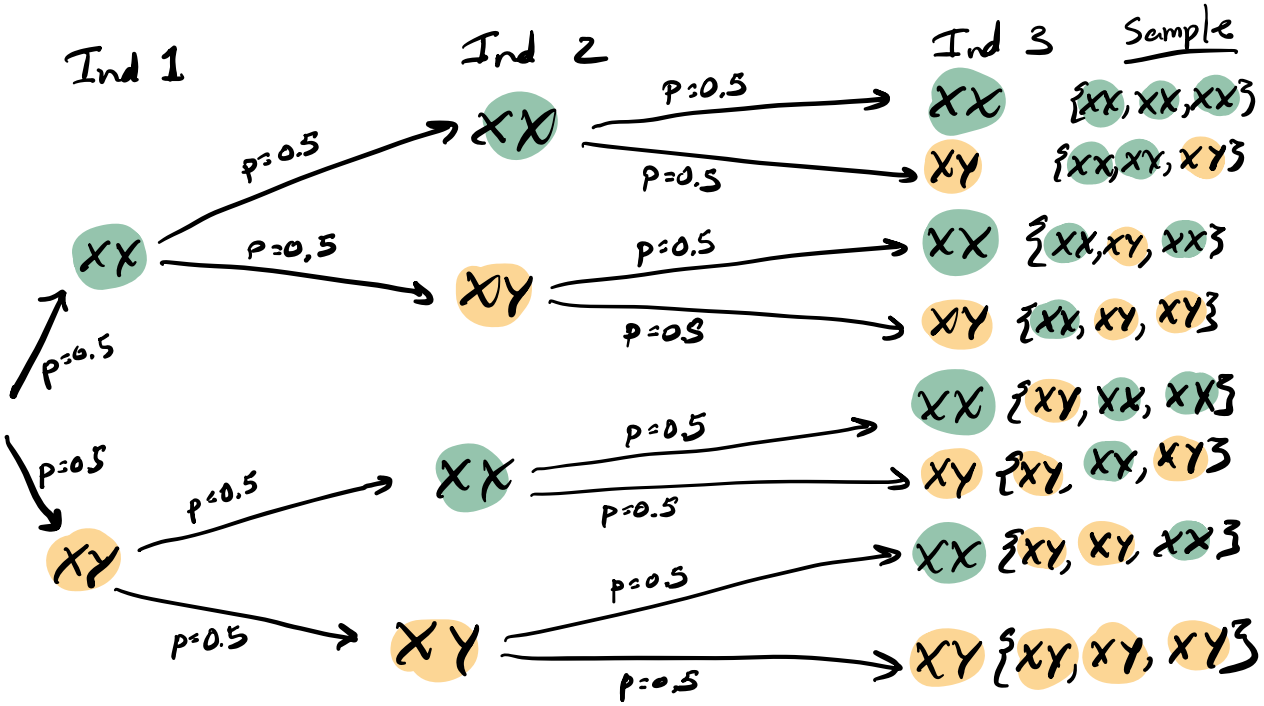
XY XX XX → 3, $p=1/8$	XX XY XX → 2, $p=1/8$	XY XX XX → 2, $p=1/8$	XY XY XX → 1, $p=1/8$
XX XX XY → 2, $p=1/8$	XX XY XY → 1, $p=1/8$	XY XX XY → 1, $p=1/8$	XY XY XY → 0, $p=1/8$

Outcomes (#XX)	Probability (total area)
0	$1 \cdot 1/8 = 1/8$
1	$3 \cdot 1/8 = 3/8$
2	$3 \cdot 1/8 = 3/8$
3	$1 \cdot 1/8 = 1/8$

The distribution of Y is called a **Binomial Distribution.**

Binomial Distributions model the number of successes in a repeated number of Bernoulli random variables.

Another way to do this is with a probability tree.

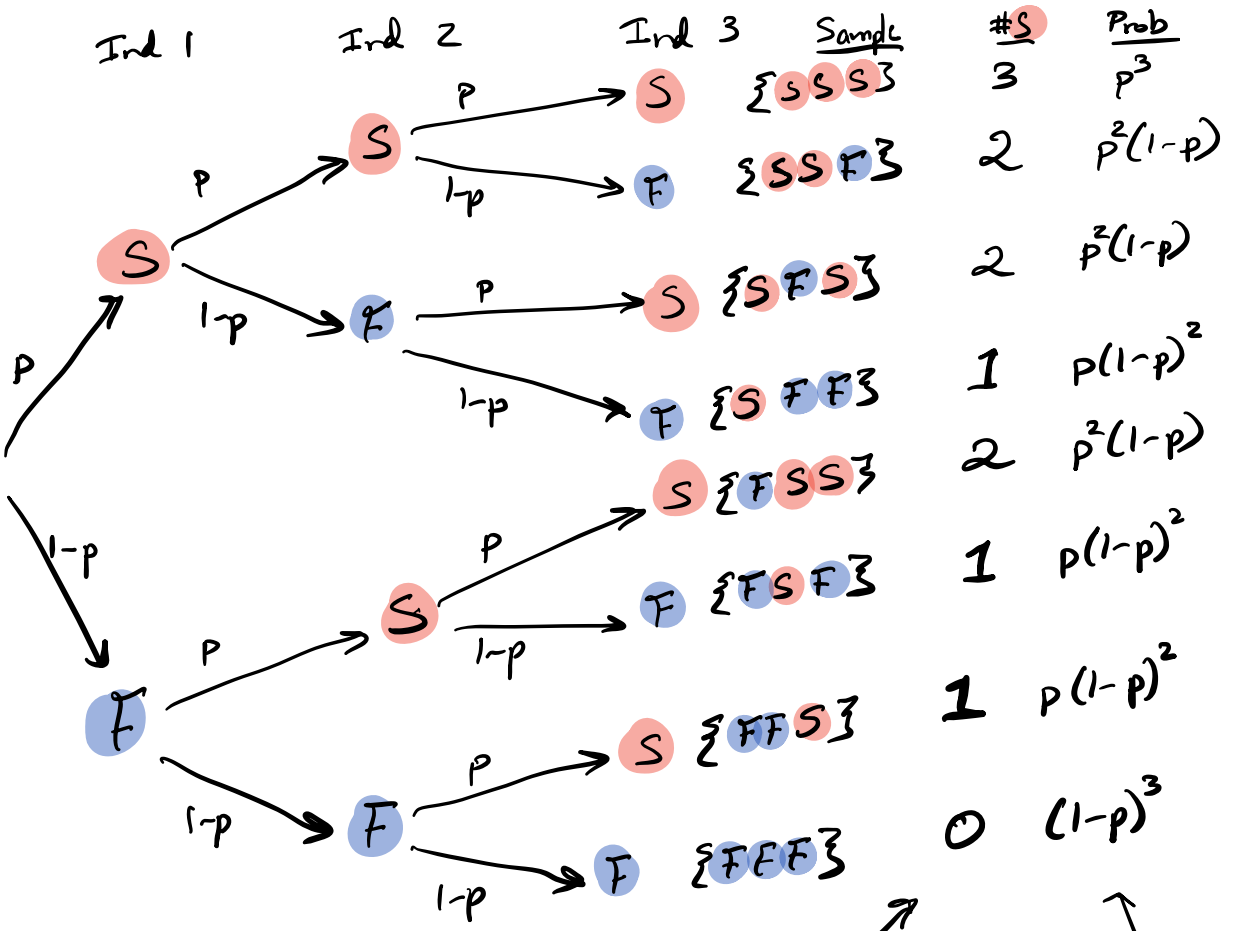


To find the sequence, follow the tree from left to right.

To find the probability of the sequence, multiply the probabilities along the the "branch" from left to right.

We can do this with any Bernoulli distribution.

Let's try with an arbitrary "Success" shown by "S" with probability p. "Failure" has probability $1-p$.



All possible outcomes in terms of # successes.

0, 1, 2, ..., number of individuals

Distribution of # Successes

0	$1 \times p^0 (1-p)^3$
1	$3 \times p(1-p)^2$
2	$3 \times p^2(1-p)$
3	$1 \times p^3(1-p)^0$

in front represents the # of ways to get a particular # successes.

Probabilities have a particular form

$$p^{\#\text{successes}} \times (1-p)^{\#\text{failures}}$$

failures = # individuals − # successes.

Note: $x^0 = 1$

These features all all present in the Binomial distribution.

If $Y \sim \text{Binomial}(n, p)$, then $\Pr(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$

Parameters:
- $n \rightarrow$ # individuals in sample
- $p \rightarrow$ probability of success for a single individual.

$k \rightarrow$ # successes you are interested in. k can be any integer from 0 to n.

Above, we figured out that the probability of 2 successes in 3 individuals is $3p^2(1-p)$.

Let's use the Binomial equation to confirm.

$n=3$, $k=2$, $p=p$

$$\Pr(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Pr(2 \text{ successes}) = \binom{3}{2} p^2 (1-p)^{3-2}$$

$$= \frac{3 \times 2 \times 1}{2 \times 1 \times 1} p^2 (1-p)^1$$

$$= 3p^2(1-p) \checkmark$$

This equation is nice because probability trees get very large for even small sample sizes.

The relationship is $\#\text{possible samples} = 2^{\#\text{individuals}}$

Note: $\binom{n}{k}$ reads as "n choose k" and is called "The Binomial Coefficient".

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and gives the number of ways to arrange k items in a collection of n.

Example:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!\,2!}$$

$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (2 \times 1)}$$

$$= \frac{5 \times 4}{2 \times 1} = 10.$$

This means there are 10 ways to arrange 3 people in 5 seats, for example.

# individuals	# possible samples
1	$2^1 = 2$
2	$2^2 = 4$
3	$2^3 = 8$
4	$2^4 = 16$
5	$2^5 = 32$
6	$2^6 = 64$
7	$2^7 = 128$
8	$2^8 = 256$
9	$2^9 = 512$
10	$2^{10} = 1024$
⋮	
20	$2^{20} = 1,048,576$ ($\approx$ 1 million tree tips)
⋮	
30	$2^{30} = 1,073,741,824$ ($\approx$ 1 billion tree tips)
⋮	
40	$2^{40} = 1,099,511,627,776$ ($\approx$ 1 trillion tree tips)
⋮	
50	$2^{50} =$ A LOT!

Binomial distributions help us easily calculate the number of successes in even very large samples without having to build probability trees or probability diagrams. For example, imagine you are interested in the prevalence of a disease. Based on prior knowledge, you hypothesize that the prevalence is 70%. This implies an individual's infection status can be modeled by a Bernoulli with parameter $p = 0.7$ (setting "infected" to be a "success").

Let's assume we sample 100 individuals (n = 100). Here are the probabilities of seeing different numbers of infected individuals in our sample of 100.

Number Infected (K)	Probability
0	$\binom{100}{0} 0.3^{100} = 5.15 \times 10^{-53}$
20	$\binom{100}{20} 0.7^{20} 0.3^{80} = 6.32 \times 10^{-25}$
40	$\binom{100}{40} 0.7^{40} 0.3^{60} = 3.71 \times 10^{-10}$
60	$\binom{100}{60} 0.7^{60} 0.3^{40} = 0.00849$
80	$\binom{100}{80} 0.7^{80} 0.3^{20} = 0.00757$
100	$\binom{100}{100} 0.7^{100} = 3.23 \times 10^{-16}$

If you want to see what this distribution looks like, pull up R and copy in the following commands

```
n <- 100
outcomes <- 0:100
p <- 0.7
plot(outcomes, dbinom(outcomes, n, p))
```

Assumptions of the Binomial

Binomial distributions rely on a number of assumptions.

1. Individual outcomes follow a Bernoulli distribution
   Outcomes of interest are categorical, with two categories
2. Individual outcomes are **identical** across individuals
   In essence, every individual has the same probability distribution
3. Probabilities are **independent** across individuals.
   This means that the outcome of one individual that is sampled does not change the probability of the outcome for any other individuals.

These three are summarized as saying the Binomial distribution models, "*independent and identical Bernoulli trials"*.

The independent and identical assumptions are most important. When this is the case, we have a random sample. When our sampling methodology does not match the assumptions of identical and independent, then we run into problems making inferences based on expectations from Binomial distributions.

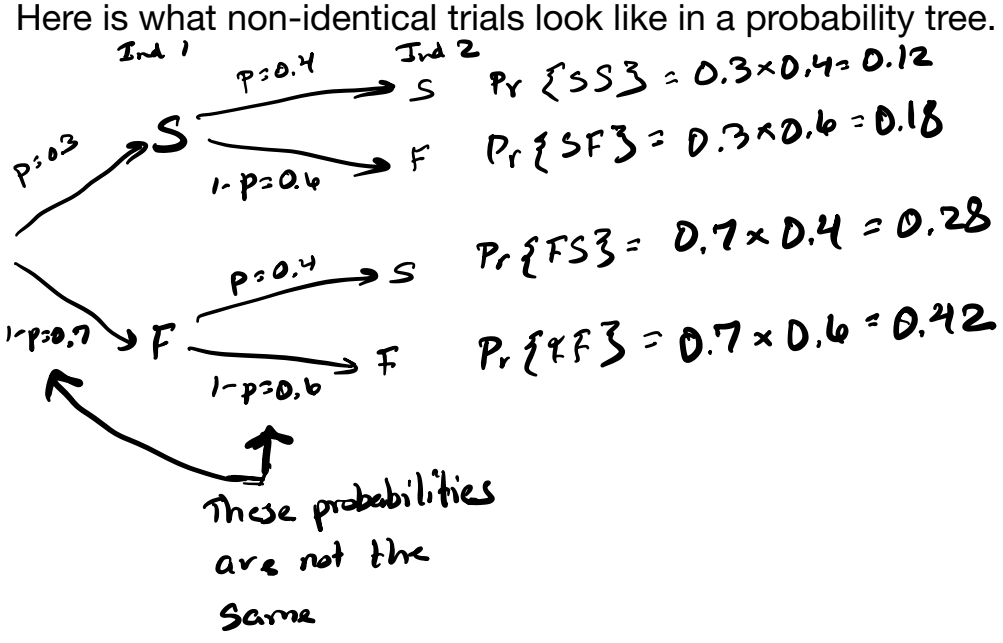
Here are a couple cases where the assumption of identical trails is broken.

- You are studying prevalence of a disease and prevalence is highly structured in space (i.e., some places have much higher prevalence than others).
- You are studying a particular species and are trying to find how abundant it is. You go out and record whether it is present in a trap. But the species' abundance changes drastically over time. This means the probability of catching it changes from time to time (here, individuals are times).

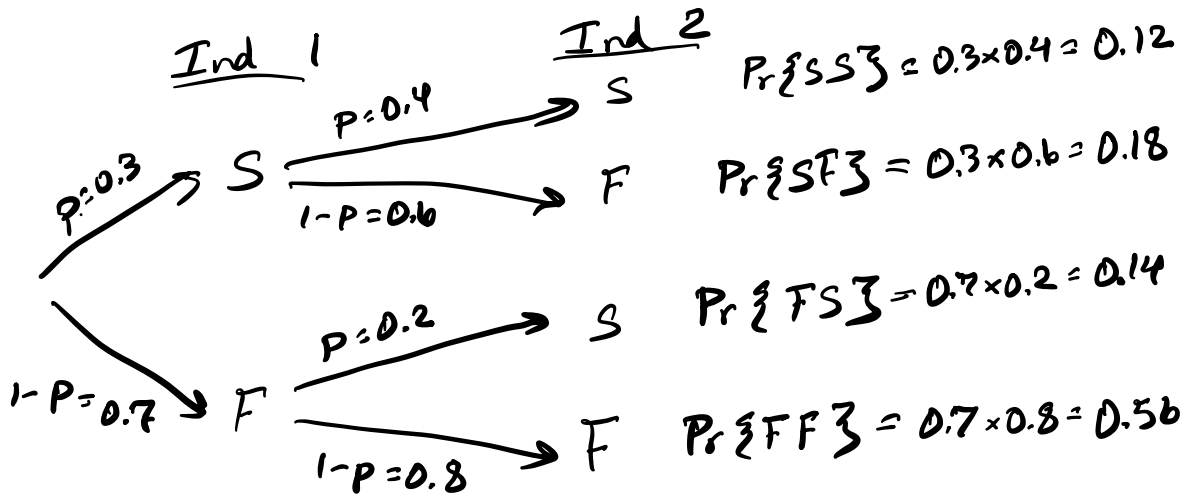
Here are a couple of cases where the assumption of independent trials is broken.

- You are studying the prevalence of a disease that is infectious and so transmits between individuals with high contact. You find an individual who is infected and go visit their house to ask them questions. While there, you take the opportunity to go ahead and sample members of their family. But family members typically have high contact, so you should expect that the probability that the family members are infected to be higher if the focal individual is also infected.
- You are studying a species and trying to find out how abundant it is. At first, you are bad at trapping and catching the species (and so you get a bunch of "failures"). After you catch a few, you understand better how to catch them. As such, the probability of catching individuals at a later time goes up.

Here is what non-identical trials look like in a probability tree.



This is what non-independent trials look like on a probability tree.



In this case, you can see that the probability of success and failure in the second individual changes based on whether the first individual was a success or a failure. Here, success in the first individual leads to a higher probability of success in the second. Likewise, failure in the first individual leads to a higher chance of failure in the second.