Week 3 - Probability

The mathematical tools of probability help us make inferences about how we are to interpret our data in the context of particular assumptions about the world. For example, imagine that you are interested in the color variation of a butterfly species. You collect some butterflies from an area and find two color morphs among 20 sampled individuals. Probability allows us to quantify how likely this outcome is given a presumption that there are two color morphs, three color morphs, or more.

Probability theory is the formal mathematical study of randomness. Random in this sense simply means that something cannot be predicted with exact certainty. This is certainly the purview of statistics. We cannot know exactly how any biological sample might look before we take it. If we could, why would we take a sample in the first place? Most importantly, the tools of probability theory allows us to quantify uncertainty in our estimation of population parameters, evaluation of hypotheses, and in our predictions (the main objectives of statistics).

Most of probability theory is concerned with **random variables**. As their name implies, they represent quantities that vary and that have uncertainty. Random variables represent an as yet unmeasured outcome of an "experiment." Experiment is a probability term, not a scientific term, and it means any activity for which you are interested in the outcome but do not yet know it. Flipping a coin is an experiment; catching a frog and identifying its species is an experiment; walking across campus and counting the number of students is an experiment. In each case, you may have some idea what the outcome will be of performing said "experiment", but you do not know with exact certainty. Moreover, these experiments might lead to a range of possible outcomes. Hence, their as yet undetermined outcome is represented by a random variable. Here is an example.

X_i

random variable

index i that means it is the random variable for experiment i.

If the experiment is recording the hair color of my brother, $X_{brother}$ represents the as yet undetermined hair color of my brother. Without knowing, we simply substitute a set of plausible guesses and some measure of how likely they are to be the correct answer.

To characterize a random variable, we need two items:

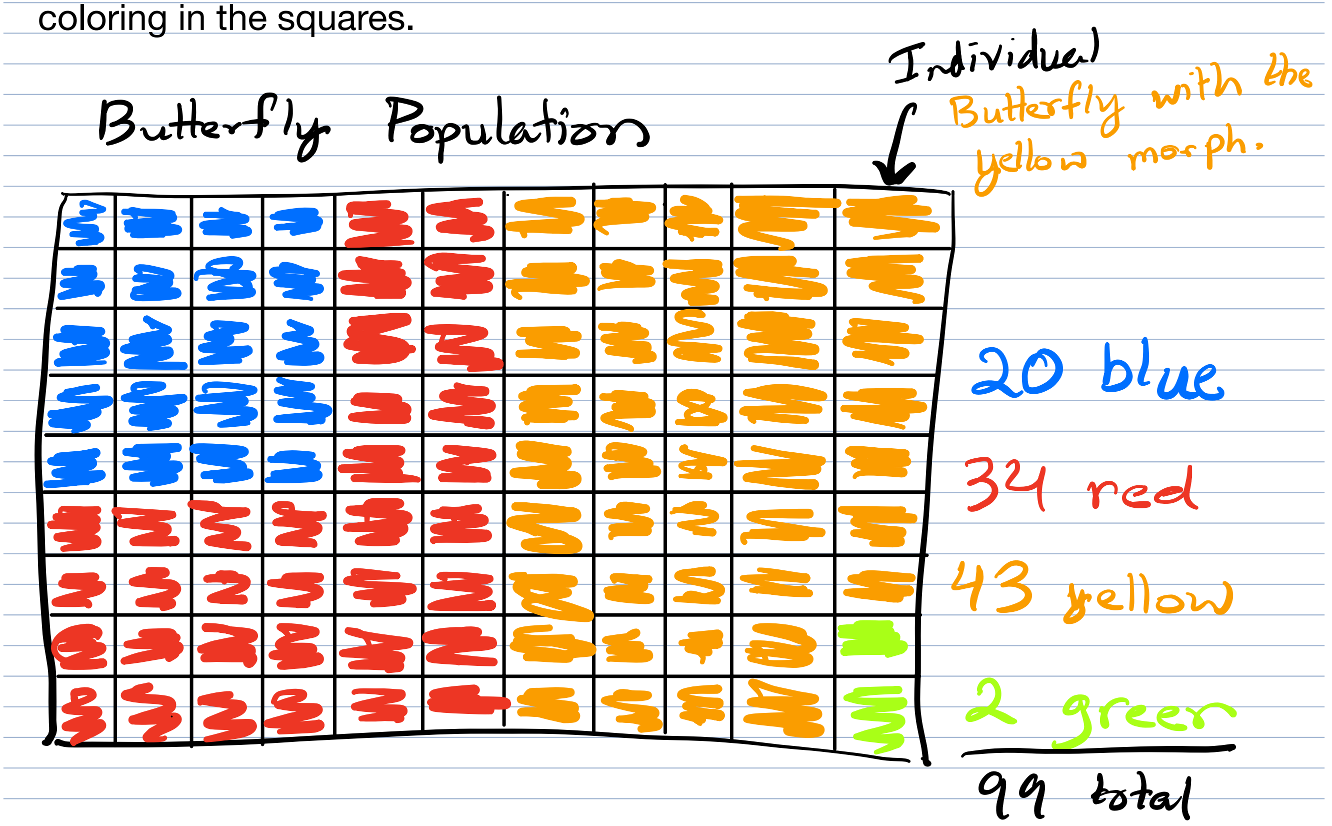
(1) a list of all possible outcomes of an experiment, and

(2) probabilities associated with each outcome.

To list all possible outcomes, we need the idea of a *mutual exclusivity*. Mutual exclusivity means that two outcomes cannot both occur simultaneously in a single experiment. For example, when you role a single die, the die cannot come up both 1 and 6. The outcomes 1 and 6 are *mutually exclusive.* A list of all possible outcomes in an experiment is a list of all mutually exclusive outcomes. Mutual exclusivity is not a difficult concept. Simply ask if two events can be measured in a single sample individual.

Probabilities measure how likely a particular outcome is to occur. Probabilities have the property that they are always between 0 and 1 (including 0 and 1). A probability of zero means that outcome will not occur with exact certainty. A probability of 1 means that the outcome will occur with exact certainty.

One way to measure probabilities is to think about the population of individuals (i.e., statistical individuals) and to label the individuals according to their characteristics. Here's an example with color morphs of butterflies. Each square in a butterfly in the population and we can indicate their color by coloring in the squares.



The probability of selecting a butterfly of any particular color is related to how common it is in the population.

Green seems unlikely to be caught. Yellow seems much more likely.

We can quantify this thinking by taking the probability as the fraction of the population taken up by an outcome.

$Pr(\text{blue}) = 20/99$

$Pr(\text{red}) = 34/99$

$Pr(\text{yellow}) = 43/99$

$Pr(\text{green}) = 2/99$

Note that all of these are positive and less than 1. Moreover, the sum of all probabilities is 1.

- A general feature of random variables is that the sum of the probabilities of all possible outcomes is 1.

$Pr(\text{blue or red or yllw or green})$

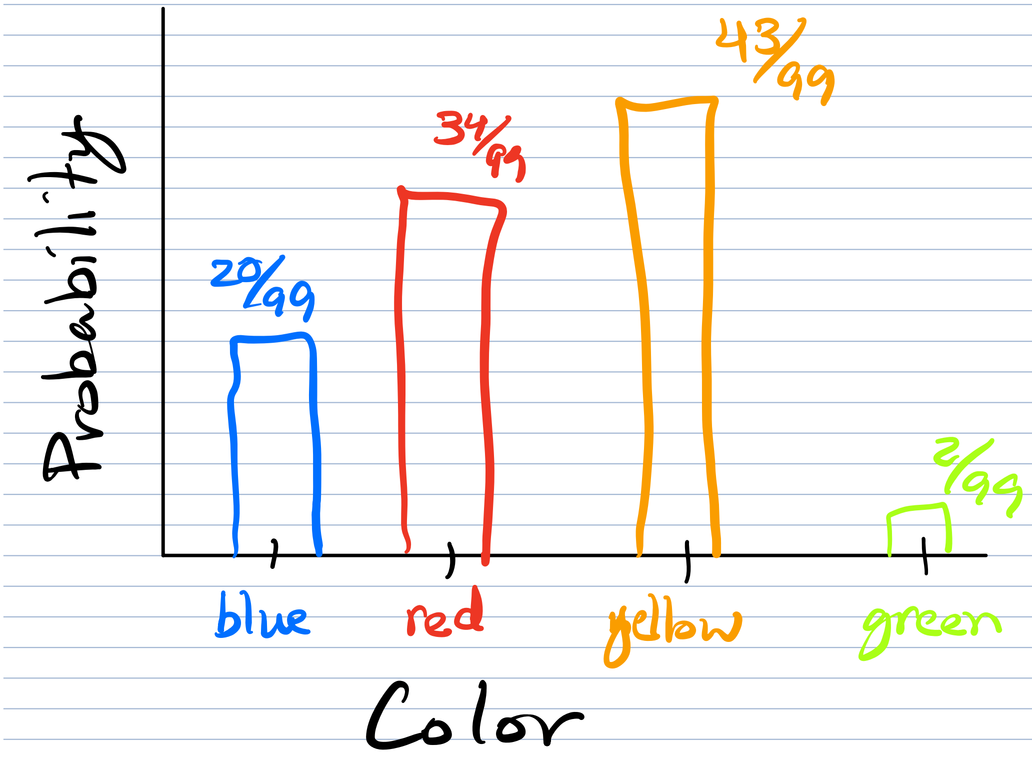
$$= \frac{20 + 34 + 43 + 2}{99}$$

← individuals that are any of these four colors.

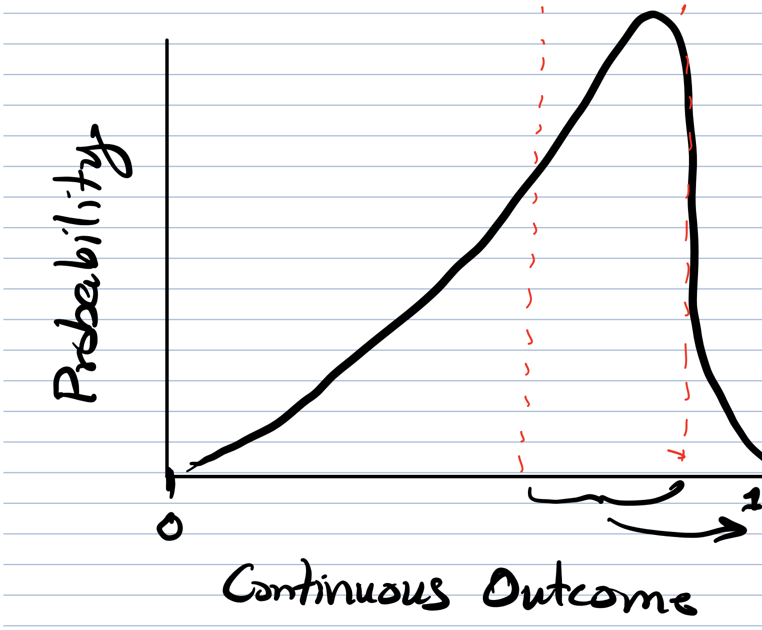
← total individuals

$$= 1.$$

Another way to visualize a random variable is by plotting possible outcomes on the x-axis and their associated probabilities on the y-axis. Since probabilities are always positive, they can be thought of as amounts and can be plotted with barcharts.



These bar charts work very well for outcomes that are categorical or numerical and discrete. The work less well for outcomes that are numerical and continuous. Such continuous probability distributions are graphically represented by density plots.



Region of most likely outcomes.
Probability is highest here.

When populations are very large, a population diagram with boxes for individuals has very small boxes, like pixels in a tv. With enough of these small boxes, we can treat the distribution of possible characteristics as continuous. We can then use Venn diagrams to evaluate and derive probability rules. These are detailed in the book chapter. We won't use many of them, but you should be aware of them. Here is an example.

Population with
4 individuals

A	A
A	B

2 outcomes = {A, B}

Pr(A) = 3/4

Pr(B) = 1/4

Population with
20 individuals

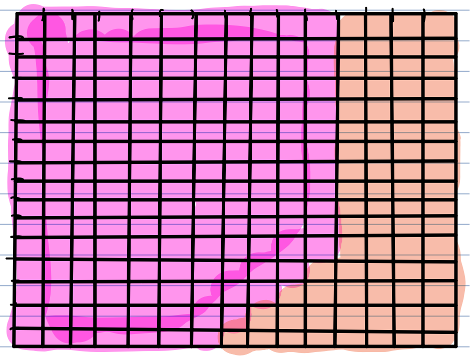
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	B
B	B	B	B

2 outcomes = {A, B}

Pr(A) = 15/20 = 3/4

Pr(B) = 5/20 = 1/4

Population with
240 individuals

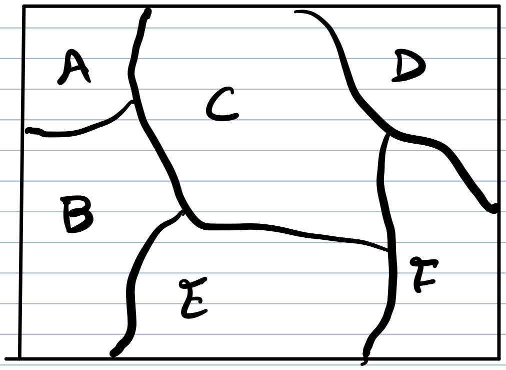


2 outcomes = {🟣, 🟠}

Pr(🟣) = 166/240

Pr(🟠) = 74/240

Very Large Population



Six Mutually Exclusive Outcomes = {A, B, C, D, E, F}

Area A + Area B + Area C + ... Area F = 1

Pr(A) = Area A

Based on diagram, C seems to be the most likely outcome and A least likely.

You can see that there are many similarities between the way to think about frequencies of different outcomes in data and to think about probabilities. But there is an important distinction to be made. **Probabilities in statistics represent assumptions** about how the world works, both in terms of assumptions

(1) about the biology determining individual characteristics and

(2) about the sampling process.

To see how, let's consider the sex-ratio of a diploid organism such as humans. Most humans have XX or XY sex chromosome pairs. (There are certainly others that are important for human biology, but we will simplify things a bit for the purpose of illustration.)

To make assumptions about the biology, let's think about the possible processes that generate a particular chromosome pair.

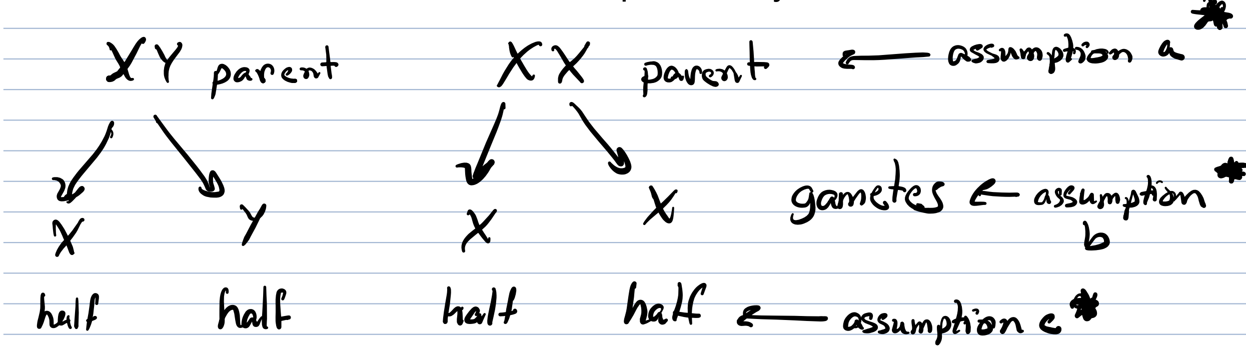
a. An XX individual must mate with an XY individual.

b. During mating each individual makes gametes that are diploid and so only has a single chromosome from the parental genome.

c. The chromosome that makes it into any given gamete is random.

d. Fertilization pairs gametes to make the offspring chromosome combination. Fertilization is random with respect to gametes.

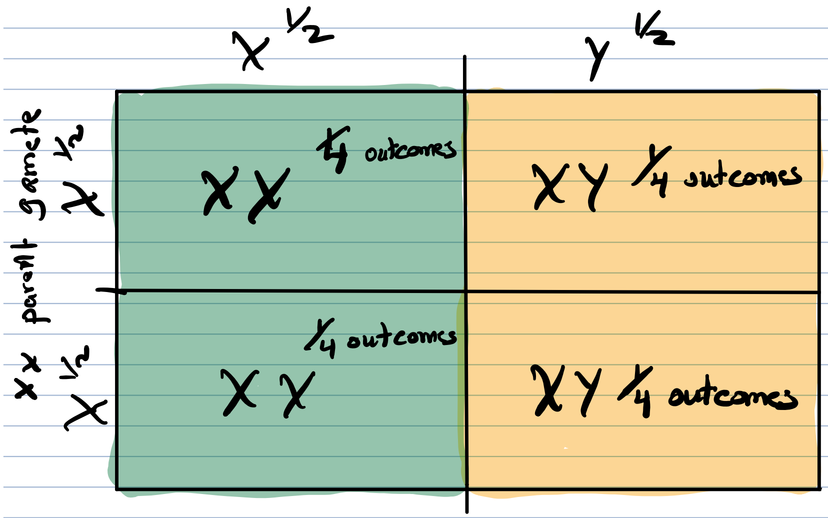
Let's formalize these with math and probability.



These combine during fertilization.

XY parent gamete

assumption d*



XX parent gamete	$X^{1/2}$	$Y^{1/2}$
$X^{1/2}$	XX $\frac{1}{4}$ outcomes	XY $\frac{1}{4}$ outcomes
$X^{1/2}$	XX $\frac{1}{4}$ outcomes	XY $\frac{1}{4}$ outcomes

all possible fertilization events

All possible offspring outcomes.

XX individuals make up $\frac{1}{2}$ total area. $Pr(XX) = \frac{1}{2}$

XY individuals make up $\frac{1}{2}$ total area. $Pr(XY) = \frac{1}{2}$

Let G_i be the sex chromosome haplotype for individual i. Based on the arguments above, a probability distribution for this random variable is

$$G_i = \begin{cases} XX & \text{with probability } \frac{1}{2} \\ XY & \text{with probability } \frac{1}{2} \end{cases}$$

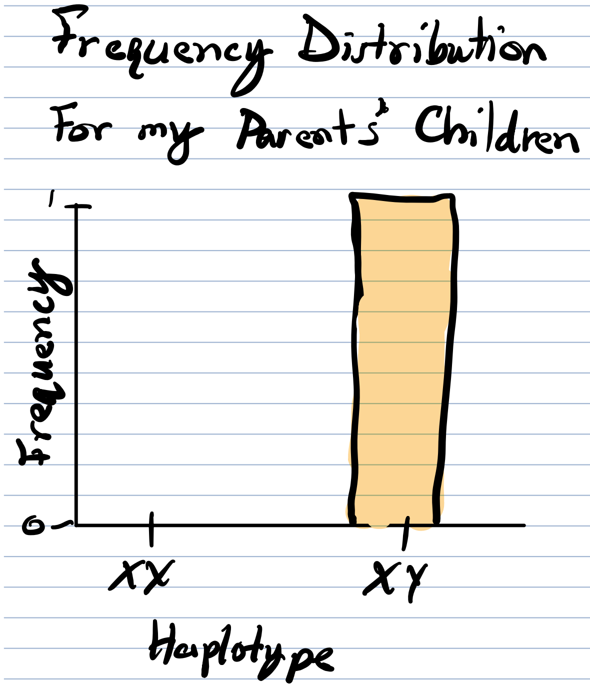
↑ all possible, mutually exclusive outcomes

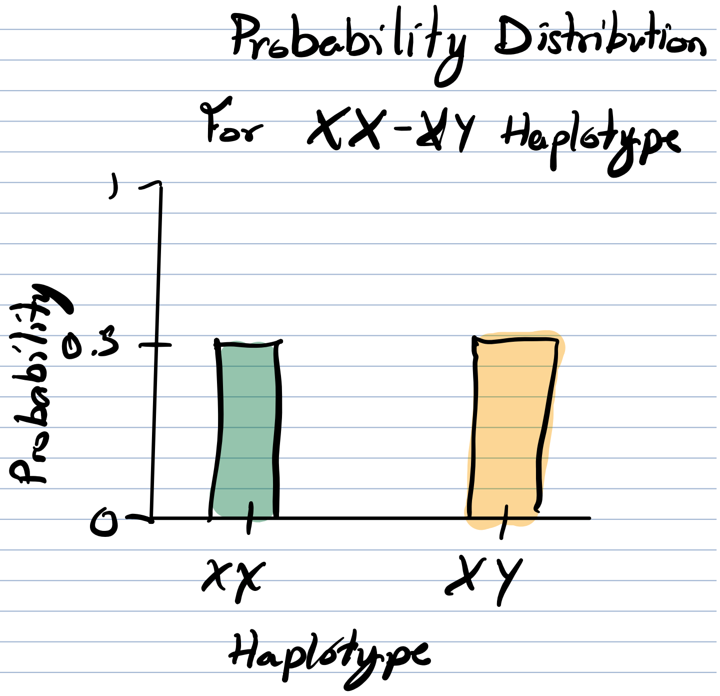
↑ associated probabilities.

We now have a probability model for how an individual's characteristics are acquired. The assumptions here are the typical ones you learn in basic biology of sex determination. They give rise to the prediction that in large enough numbers of people, sex ratios should be evenly split between XX and XY individuals.

Of course, this doesn't always happen. When there are a small number of events, observed frequencies differ from probabilistic expectations. My family is one example. My brother and I are the only children of our parents and we are both XY individuals. As such, the frequency of XY = 100% among my parents children and the frequency of XX individuals is 0%.

Data are described using **frequency distributions**, which show the proportion of the data that fall among the multiple possible outcomes. Probabilities are described using **probability distributions,** which show expectations from many many instantiations of a random variable.





This biologically-based probability distribution is an example of what is called a **Bernoulli distribution**.

> Bernoulli distributions are probability distributions with two categorical outcomes, arbitrarily named “success” and “failure”. It has a single parameter *p* which describes the probability of a “success”.

If we define a “success” as the XX haplotype, then we can write

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

The random variable G for individual i

is distributed as

a Bernoulli random variable with probability of success equal to 0.5.

Another way to write this is

$$G_i = \begin{cases} XX \ (\text{“success”}); & Pr(XX) = 0.5 \\ XY \ (\text{“failure”}); & Pr(XY) = 0.5. \end{cases}$$