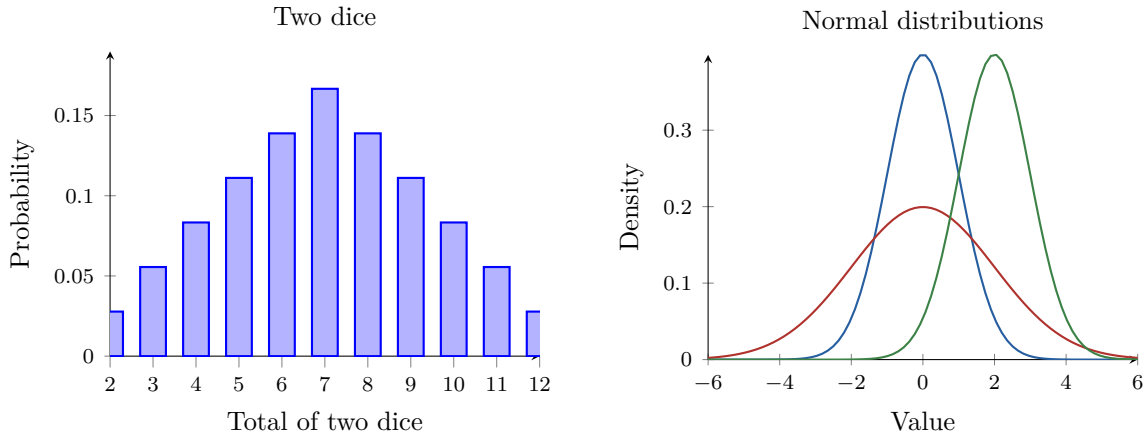
**Data Analytics Workshop, JHU Pre-College Program**
**Day 3: Distributions and Outliers (optional notes)**

**TL;DR** Data is random, but the randomness has structure. If we roll dice many times and tally the results, the tally takes on a predictable shape called a **distribution**. The more samples we take, the closer the tally gets to the true distribution. A few distributions come up often (the names matter less than the idea that data spreads itself out in different ways):

- **Uniform**: every value is equally likely, so the graph is flat (one die, one card from a deck).
- **Bernoulli**: exactly two outcomes, one with probability $p$ and one with probability $1 - p$ (heads or tails, hit or miss).
- **Binomial**: the number of successes in repeated independent Bernoulli trials (hits in 10 at bats).
- **Normal** (Gaussian, the "bell curve"): set by a mean $\mu$, where it is centered, and a standard deviation $\sigma$, how wide it is (heights of people).

An **outlier** is a data point that does not fit with the rest. Some outliers come from mistakes or broken experiments and should be removed (a stopwatch someone forgot to stop). Others are real data that just happened to be unusual, and should be kept. An outlier raises a question; it does not answer one.




Left: exact probabilities for the total of two dice. Right: normal curves with $(\mu, \sigma) = (0, 1)$, $(0, 2)$, and $(2, 1)$.

# 1 Distributions of Data

So far we have said that data is random. The nice thing is that this randomness is not just a jumbled mess we cannot learn anything from. Let's start very simple. Roll a single die, and keep a bar graph with the die value on the horizontal axis and the number of times we rolled it on the vertical axis. Each roll adds one block to one bar. Since every value is equally likely (if it weren't, a lot of games would be unfair), after a thousand rolls the bars should all be about the same height.

Now roll two dice and add them. The totals go from 2 to 12, but they are not equally likely, and if you have played Monopoly you already know this. The only way to get a 2 is a 1 on each die, while a 7 can come from

$$1+6, \quad 2+5, \quad 3+4, \quad 4+3, \quad 5+2, \quad 6+1.$$

There are $6 \cdot 6 = 36$ equally likely ways for two dice to land, so a total of 2 has probability $\frac{1}{36}$ and a total of 7 has probability $\frac{6}{36} = \frac{1}{6}$. After many rolls, the bars form the triangle shape on the first page, with many more 6s, 7s, and 8s than 2s or 12s.

**Definition 1.1** (Distribution)

*The **distribution** of a random variable describes which values it can take and how likely each one is. The distribution of a dataset is how often each value (or range of values) shows up in the data.*

Other data has a shape of its own. If we ask many people their height, most answers cluster around the average height of the group, and far fewer are near seven feet. Notably, the more samples we take, the closer the tally gets to the true distribution. With three rolls of a single die we could get three sixes, which looks nothing like the flat distribution we expect.

# 2 Common Distributions

The idea is that if we can model a situation with a distribution we already understand, we can draw conclusions from it. I want to emphasize that the names below are less important than the idea that data can distribute itself in different ways. For each one, I invite you to play around with the distributions Desmos graph from Discussion 3.

**Definition 2.1** (Uniform distribution)

*In a **uniform distribution**, every value is equally likely, so its graph is flat.*

Rolling one die and drawing one card from a full deck are uniform. Data from the real world is rarely uniform though.

**Definition 2.2** (Bernoulli distribution)

*A **Bernoulli distribution** has exactly two outcomes, one with probability $p$ and the other with probability $1 - p$.*

The two outcomes could be heads and tails, win and lose, or true and false. Exactly one of them must happen (you can't have neither, and you can't have both), so the probabilities add to 1.

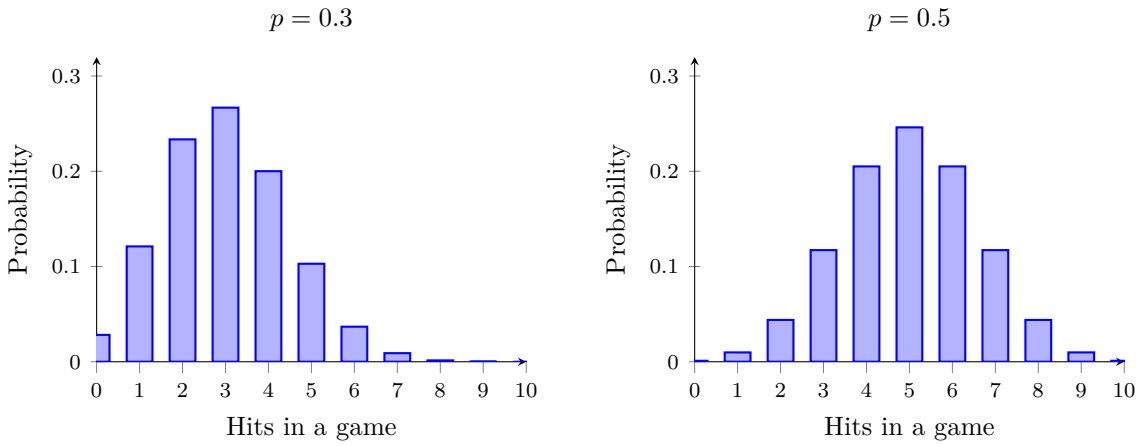
A fair coin has $p = \frac{1}{2}$, while an unfair coin could land heads with probability 0.75 and tails with probability 0.25. Rolling a total of four or higher with two dice is also Bernoulli: either it happens or it does not.

**Definition 2.3** (Binomial distribution)

*A **binomial distribution** counts the number of successes in $n$ repeated, independent Bernoulli trials that all have the same probability $p$ of success.*

**Example 2.1**

*Suppose a batter hits the ball 30% of the time. A single at bat is Bernoulli with $p = 0.3$. If the batter has 10 at bats every game, the number of hits in a game, anywhere from 0 to 10, follows a binomial distribution with $n = 10$ and $p = 0.3$. After many games, the tally peaks at 3 hits and falls off on either side (left plot below). A better hitter has a larger $p$, which moves the whole shape to the right (right plot).*




Binomial distributions with $n = 10$ at bats

**Definition 2.4** (Normal distribution)

*A **normal** (or **Gaussian**) distribution has the shape of a bell curve. It has two parameters: the mean $\mu$, which is where the curve is centered, and the standard deviation $\sigma$, which is how wide it is.*

The normal distribution is the one we will see the most, which is why it is called "normal." Notice the difference in parameters: Bernoulli and binomial distributions are set by how likely a success is, while a normal distribution needs both a center $\mu$ and a spread $\sigma$. You can see on the first page how changing $\mu$ slides the curve left or right, and increasing $\sigma$ makes it shorter and wider.

Asking many people their height gives a roughly normal distribution centered at the average height of the group: being close to the average is likely, and being far from it is unlikely. The height of the curve can be read either as how many people fall in that range or as the probability that a randomly chosen person lands in it.

**Remark 1**

*Look back at the binomial plots. They already look a lot like bell curves! This is no accident, and the same thing happens when we add up more and more dice. Adding up many independent random things tends to produce a normal distribution, which is a big reason the normal distribution shows up everywhere. We come back to this on Day 7 with the central limit theorem. (The Galton board in Discussion 3 is exactly this idea: each ball is bounced left or right many times.)*

## 3 Outliers

One task in the project is cleaning your data, and there is an Excel video (and a section of the Excel notes) on how to actually do it. Here, let's think about *what* we should remove. We want data to be organized, but more importantly, we want it to represent what we are actually trying to measure.

Suppose a class times each student running 100 meters. Most times fall between 13 and 18 seconds, with one recorded time of 41 seconds. It turns out the timer got distracted and stopped the watch late. That time does not measure what we were trying to measure, the student's speed, so it is an **outlier** that breaks the control of the experiment, and we remove it.

Not every outlier should be removed though. Suppose we roll two dice many times and see far more 10s than we would expect. We should be careful not to assume we know the distribution in advance; maybe the dice are weighted, and something strange is going on that is worth investigating. Or maybe we just got lucky. Because data is random, an unusual value can be completely valid, and that is the tricky part of data analysis. More samples reduce the chance of this happening.

**Example 3.1**

*Five titrations in a chemistry lab give concentrations of*

$$0.201, \quad 0.237, \quad 0.198, \quad 0.503, \quad 0.221.$$

*The value* 0.503 *stands out. Perhaps the pipette was opened too far, or the solution was not mixed properly. Unlike a statistic, deciding whether something is an outlier is not deterministic: an outlier raises a question. Once we have the mean and standard deviation (Day 4), we will have a better way to decide. If a value falls outside a range we are comfortable with around the mean, it may be an outlier we want to remove.*

## 4 Practice

**Problem 4.1** *1) What is the probability that two dice total* 10 *or more?*

*2) Which distribution (uniform, Bernoulli, binomial, or roughly normal) best describes each? (a) whether a free throw goes in, (b) the number of free throws made out of* 20*, (c) the weights of* 1000 *apples from one orchard, (d) a random digit from* 0 *to* 9*.*

*Solution:*

1) A total of 10, 11, or 12 can happen in $3 + 2 + 1 = 6$ ways out of 36, so the probability is $\frac{6}{36} = \frac{1}{6}$.
2) (a) Bernoulli, (b) binomial (if every shot has the same chance and they are independent), (c) roughly normal, (d) uniform.