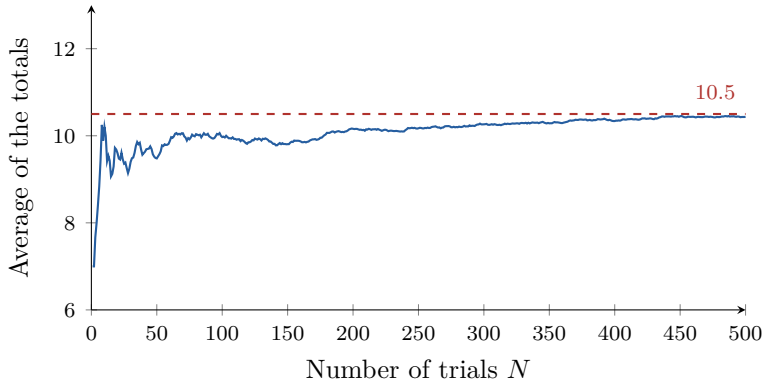
Data Analytics Workshop, JHU Pre-College Program
Day 2: Data, Sampling, and Statistics (optional notes)

TL;DR A **dataset** is a collection of data points, and each data point carries information stored in **variables**. Person i in a survey might be the list $p_i = (\text{height}, \text{weight}, \text{age})$. Data almost always has some randomness to it, whether from nature (rolling dice) or from measurement error (timing a race with a stopwatch), so we think of each data point as coming from a **random variable**. Because of that randomness, one observation tells us very little. We **sample** many times, and the more samples we have, the more confident we can be in our conclusions.

(Statistic)

*A **statistic** is a deterministic function of the data: a formula that takes in the data and returns a value. Anyone given the same data computes the same statistic, with no surprises.*

- Examples: the mean (average) $\frac{1}{N}\sum_{k=1}^{N} x_k$, the maximum, the minimum, the mode.
- Before we collect the data, a statistic is random (it depends on random data). Once the data is collected, it is just a number.
- Averages settle down as we collect more data. The plot below shows the average of the total of three dice after more and more rolls. It jumps around at first, then settles near 10.5.



Rolling three dice N times and averaging the totals (a simulation)

1 Data

Like a set, a dataset is a collection of points. The difference is that each data point carries information, and that information is stored in **variables**. If we collect the height (inches), weight (pounds), and age (years) of several people, we can record each person as a list, such as

$$p_1 = (65, 150, 27), \qquad p_2 = (68, 175, 36).$$

As long as we agree on the order of the entries, we can read off that person 2 is 68 inches tall, weighs 175 pounds, and is 36 years old. The subscript tells us which data point we are looking at, and the position in the list tells us which variable. In a spreadsheet, this is exactly one row per person and one column per variable.

	Height (in)	Weight (lb)	Age (yr)
Person 1	65	150	27
Person 2	68	175	36
⋮	⋮	⋮	⋮

Data usually comes with some randomness. This may come from nature, as with flipping coins or rolling dice, or from how we measure, as with timing a race with a stopwatch (no two people will press stop at exactly the same moment). Just as a formula has a variable we plug values into, the data we collect can be represented by a **random variable**: a variable whose value is determined by some random process. Notably, because of this randomness, a single observation tells us very little. To reach a confident conclusion about a whole population, we **sample** it many times.

2 Sampling

Let's make this concrete. Roll three dice and add them up. Each time we do this is a **trial**, and say we do N trials. If $d_{i,k}$ is the value showing on die i in trial k, then the result of trial k is

$$x_k = d_{1,k} + d_{2,k} + d_{3,k} = \sum_{i=1}^{3} d_{i,k}.$$

With seven dice, the only change is the ending point of the sum: $\sum_{i=1}^{7} d_{i,k}$. This is exactly why sigma notation is worth learning.

Each x_k is a data point, and because rolling dice is random, so is each data point. With only two trials we could get very lucky and see $x_1 = x_2 = 18$. That requires six sixes in a row, which has probability $\left(\frac{1}{6}\right)^6 = \frac{1}{46656}$. Very unlikely, but possible! If we concluded from those two trials that three dice usually total 17 or 18, we would be badly wrong. If we instead do a million trials and compute the average

$$\frac{1}{N} \sum_{m=1}^{N} x_m,$$

we get something much more realistic, very close to 10.5. You can see this in the plot on the first page: early on the average is all over the place, and as N grows it settles down. (Notice the index here is m instead of k. The letter does not matter, as long as we are consistent.)

Remark 1

Where does 10.5 come from? One die averages $\frac{1+2+3+4+5+6}{6} = 3.5$ *in the long run, and three dice should average three times that. We make this precise on Day 8 with expected value.*

Data comes from sampling, whether by running a trial or by asking people a question. Asking one specific person their height is not very random, but asking a *randomly chosen* person on the street is. This leads to two words worth knowing.

Definition 2.1 (Population and sample)

The ***population*** *is the entire group we want to learn about. A* ***sample*** *is the part of the population we actually collect data on.*

If we want to know the average height of high school students in Maryland, the population is every high school student in Maryland, and our sample might be 200 students we survey. We almost never get to measure the whole population, so we use the sample to make conclusions about it. For this to work, the sample should look like the population. Surveying only the basketball team would give a very misleading average height!

3 Statistics

Now say we have gone out and collected our data. What do we actually do with it? This is where statistics come in. You hear people say "here's a statistic" on the news all the time, but what do we really mean?

Definition 3.1 (Statistic)

A ***statistic*** *is a deterministic function of the data. It is calculated from the data with a formula, so anyone given the same data computes the same value.*

By *deterministic* I mean there are no surprises. If you give two people the exact same dataset and ask for the same statistic, they should always get the same answer. Some examples are the mean (average), the maximum, the minimum, and the mode (the most common value). We can also make up our own: "the maximum plus the mean" is not something people look at, but it is still a statistic, since it comes from a formula applied to the data.

In the dice example, the average

$$T(x_1, \ldots, x_N) = \frac{1}{N} \sum_{m=1}^{N} x_m$$

is a statistic. It is literally a function: the data goes in, a number comes out. Here is the subtle part. The data is random, so *before* we collect it, the value of a statistic is random too (roll the dice again and you will get a different average). Once the data is collected, the statistic is just a fixed number.

Remark 1

Statisticians usually want the statistic itself to tell us something about the population. The

average of our three dice samples is meant to estimate the "true" long run average of 10.5. *Over the next few days we will see how close we can expect such an estimate to be, and why more data makes it closer.*

4 Practice

Problem 4.1 *1) Which of these are statistics of a list of heights* $h_1, \ldots, h_N$*? (a) the median, (b) the range,* $\max - \min$*, (c) the height of the person I think looks the tallest, (d)* $h_1 + h_N$*.*

2) You want to know how many hours per week students at your school spend on homework, so you survey everyone in your AP Calculus class. What is the population and what is the sample? Is this a good sample?

Solution:

1) (a), (b), and (d) are statistics, since each is computed from the data by a fixed rule (even if (d) is a strange one). (c) is not: two people could disagree about who *looks* tallest, so it is not deterministic.
2) The population is all students at the school, and the sample is the AP Calculus class. It is probably not a good sample, since students in an advanced class may spend more (or less) time on homework than a typical student.