Data Analytics Workshop, JHU Pre-College Program
Day 1: Sets, Functions, and Sigma Notation (optional notes)

TL;DR Today is mostly about the language we will use for the rest of the course. A **set** is a collection of objects, written in curly brackets, like $\{1, 3, 7, \text{dog}\}$. A **function** f takes in an element of one set and returns an element of another according to some rule, such as $f(x) = mx + b$. Functions let us **model** data: a line of best fit through a plot of age versus height is a function that predicts a height for any age, even ages we never asked about.

Data analysis involves a lot of addition, so we use **sigma notation** to write long sums in a short way.

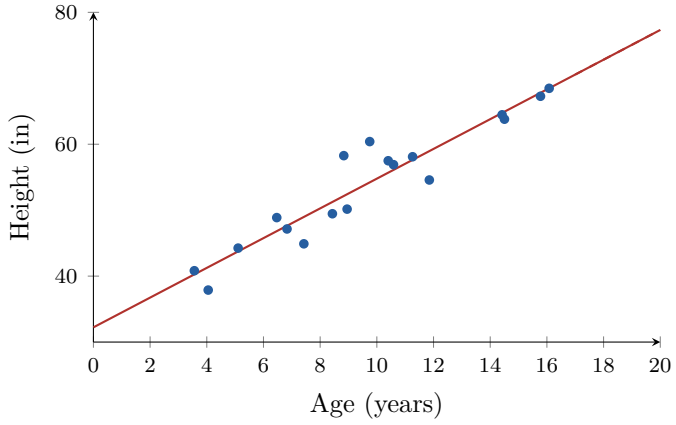
(Sigma notation)

$$\sum_{k=a}^{b} f(k) = f(a) + f(a+1) + \cdots + f(b).$$

Here k is the index, a is the starting point, b is the ending point, and $f(k)$ is the rule we follow at each step.

- When we collect data on N people, we label each person with an index: $h_1, h_2, \ldots, h_N$ for their heights. The total height is then $\sum_{i=1}^{N} h_i$.
- Constants can be pulled out of a sum, and sums can be split up: $\sum c\, a_i = c \sum a_i$ and $\sum (a_i + b_i) = \sum a_i + \sum b_i$.
- But be careful: $\left(\sum a_i\right)^2$ is **not** the same as $\sum a_i^2$ in general.

Finally, most of what we do comes down to **algebra**: moving variables around to solve for the one we care about. Whatever we do to one side of an equation, we do to the other.



Made up data of age versus height, with a line of best fit $f(x) \approx 2.26x + 32.2$

1 Sets and Functions

There is an entire course you can take on set theory, but for this course we really just want to think of a **set** as a collection of objects. I like picturing a set as a little box that we can put anything into: numbers, letters, words, even a (crude) picture of an apple. We list everything in the box inside curly brackets, such as

$$\{1, 3, 7, A, C, \text{dog}\}.$$

Some sets come up often enough to get their own symbols. The **integers** are all the positive and negative whole numbers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, and the **real numbers**, every number on the number line, are written $\mathbb{R}$. We can also build new sets out of old ones by describing what is in them. The set of even numbers is

$$E = \{x \in \mathbb{Z} : x \text{ is even}\}.$$

The symbol $\in$ means "is in," so $x \in \mathbb{Z}$ says x is an integer, and the colon is read "such that." So the whole line reads "E is the set of integers x such that x is even." None of this notation is scary; it is just a short way of saying what is in the box.

Remark 1

A set does not care about order or repeats: $\{1, 2, 3\}$ and $\{3, 1, 2, 2\}$ are the same set. Data usually does care about both (the order tells us whose height is whose, and two people can have the same height), so when we start collecting data we will write it as a list instead, with round brackets like $(65, 150, 27)$.

A big theme in math is taking two things and finding a way to relate them. If sets are our collections of objects, then **functions** are how we relate two sets.

Definition 1.1 (Function)

*A **function** f takes in an element x of one set and returns an element $f(x)$ of another set, following some rule.*

You have probably seen functions before as the line $y = mx + b$, with slope m (rise over run) and y-intercept b, or as the quadratic $y = ax^2 + bx + c$. Rather than thinking of these as a relation between x and y, I want you to think of a function as a little machine: some x goes in, and out comes $f(x) = mx + b$. The formula tells us everything about the machine. It is this idea of an input, an output, and a rule connecting them that makes functions so useful for modeling.

2 Modeling Data with Functions

Suppose we went around and asked a bunch of people their age and their height, and plotted each person as a point with age on the horizontal axis and height on the vertical axis. You can see some made up data like this on the first page. The plot alone already tells us something (older kids tend to be taller), but it is often more useful to model the data with a function we already understand, like a line or a parabola, since we know a lot about how those behave.

A common choice is the **line of best fit**: the line that passes through the data in the best possible way. We write it as

$$f(x) = mx + b,$$

where the slope m and intercept b are calculated from the data (Excel will do this for us, and we will see how on Day 7). For the data on the first page, $f(x) \approx 2.26x + 32.2$, which says that each extra year of age goes with about 2.26 more inches of height.

Recall that a function takes a value from one set and returns a value in another. Here the first set is ages and the second is heights. For a person in our data, the output is their actual height. For an age that is *not* in our data, the output is a *prediction* of what that height might be. Using a model to predict values we never collected is called **extrapolation**.

Remark 1

Extrapolation should be done carefully. Our line says a 10 year old should be about 55 inches tall, which is reasonable. It also says a 40 year old should be about 122 inches tall, which is over ten feet! The line only describes the data well in the range we collected (ages 3 to 17), and people stop growing. Always ask whether a prediction is inside the range of your data.

3 Sigma Notation and Indices

When we start calculating averages and other statistics, we will be doing a lot of addition. With only a few values, this is not too bad. Say the tortoise and the hare race in three legs.

	Leg 1 (s)	Leg 2 (s)	Leg 3 (s)	Total (s)
Hare	85	33	400	518
Tortoise	150	100	150	400

The hare is much faster on the first two legs, but takes its famous nap on the third, so the tortoise wins by 118 seconds. Adding three numbers is easy to write out. Adding the first 100 even numbers, $2 + 4 + 6 + \cdots + 200$, is not, and when we work with data we often do not even know ahead of time how many numbers we will be adding. So we want a short way to write a long sum.

Definition 3.1 (Sigma notation)

For integers $a \leq b$ and a rule $f(k)$,

$$\sum_{k=a}^{b} f(k) = f(a) + f(a+1) + \cdots + f(b).$$

The variable k is the index, a is the starting point, b is the ending point, and $f(k)$ is the rule applied at each step.

The big Σ is the Greek letter sigma, which is our S, for "sum." To reiterate, there are three parts: where we start, where we stop, and the rule we follow along the way. We plug in $k = a$, then $k = a + 1$, and so on up to $k = b$, and add up everything we get.

Example 3.1

Since $2k$ is even for every integer k, the sum of the first 10 even numbers is

$$\sum_{k=1}^{10} 2k = 2 + 4 + 6 + \cdots + 20 = 110.$$

If we do not know ahead of time how many values there are, we replace the ending point with a variable. The sum of the first N even numbers is $\sum_{k=1}^{N} 2k$.

Notice that sigma notation does not make the arithmetic any easier. We still have to add all those numbers up (or have Excel do it). It is *notation*: a way of writing a long sum in a short and precise way.

Indices

When we collect data on many people, we label each person with an **index** rather than inventing a new letter for everyone. If we ask 100 people for their height, weight, and age, we write

$$h_1, h_2, \ldots, h_{100}, \qquad w_1, w_2, \ldots, w_{100}, \qquad a_1, a_2, \ldots, a_{100},$$

so that h_i, w_i, and a_i all belong to person i. The total height is then $\sum_{i=1}^{100} h_i$. The letter we use for the index does not matter: $\sum_{i=1}^{100} h_i$ and $\sum_{k=1}^{100} h_k$ mean exactly the same thing.

Rules for Sums

Two rules will save us a lot of work later. Both are really just the usual rules of addition, written in sigma notation.

Proposition 3.2

For any numbers $a_1, \ldots, a_N$, $b_1, \ldots, b_N$, and any constant c,

$$\sum_{i=1}^{N} c\, a_i = c \sum_{i=1}^{N} a_i \qquad \text{and} \qquad \sum_{i=1}^{N} (a_i + b_i) = \sum_{i=1}^{N} a_i + \sum_{i=1}^{N} b_i.$$

Also, adding the same constant N times gives $\sum_{i=1}^{N} c = Nc$.

The first one is just factoring: $ca_1 + ca_2 + \cdots + ca_N = c(a_1 + a_2 + \cdots + a_N)$. The second one is just regrouping the terms. Not everything works this nicely though.

Remark 1

In general,

$$\left(\sum_{i=1}^{N} a_i \right)^2 \neq \sum_{i=1}^{N} a_i^2.$$

For example, take $a_1 = 1$, $a_2 = 3$, $a_3 = 7$. Then

$$\left(\sum_{i=1}^{3} a_i \right)^2 = (1 + 3 + 7)^2 = 121, \qquad \sum_{i=1}^{3} a_i^2 = 1 + 9 + 49 = 59.$$

One example where two formulas disagree is all it takes to show they are not equal in general. This is called a counterexample.

Remark 2

There is a famous story that a young Carl Friedrich Gauss was asked to add the numbers 1 *through* 100 *and did it in seconds. He paired the first and last numbers* $(1 + 100)$*, the second and second to last* $(2 + 99)$*, and so on. Each pair adds to* 101 *and there are* 50 *pairs, so* $\sum_{k=1}^{100} k = 50 \cdot 101 = 5050$*. Can you use this (and the rules above) to find the sum of the first* 100 *even numbers?*

4 Algebra: Solving for a Variable

There is a hidden notion of algebra built into all of mathematics. If you go on to take calculus, you will find out pretty quickly that most of it comes down to how good your algebra chops are! This course is similar: we will be moving a lot of variables around, so let's warm up.

Suppose your grade G in a class depends on your homework H, midterm M, final F, and participation P through

$$G = 0.25H + 0.25M + 0.35F + 0.15P.$$

This is a function of four variables. Near the end of the semester you know $H = 90$, $M = 80$, and $P = 100$, but you have not taken the final yet. Plugging in what we know,

$$G = 0.25(90) + 0.25(80) + 0.35F + 0.15(100) = 22.5 + 20 + 0.35F + 15 = 57.5 + 0.35F.$$

Now the grade only depends on F. To find the final exam score needed for $G = 90$, we undo the operations on F one at a time, doing the same thing to both sides:

$$57.5 + 0.35F = 90 \quad \Longrightarrow \quad 0.35F = 32.5 \quad \Longrightarrow \quad F = \frac{32.5}{0.35} \approx 92.86.$$

Since a higher F always gives a higher G, this is the *lowest* final exam score that gives a 90 in the course. As a sanity check, this should not be too surprising: the homework was right at 90, the midterm was lower, and participation boosts us up a bit, so we need a little more than 90 on the final.

We could also have solved for F before plugging in any numbers. If the grade we want is K, then subtracting the other terms and dividing by 0.35 gives

$$F = \frac{K - 0.25H - 0.25M - 0.15P}{0.35}.$$

This one formula answers the question for any target grade and any homework, midterm, and participation scores. That is the real power of algebra: do the work once with letters, then plug in numbers as often as you like.

5 Practice

Problem 5.1 *1) Compute* $\sum_{k=1}^{5}(k^2-1)$.

2) Write $3+6+9+\cdots+30$ *in sigma notation.*

3) Using the grade formula above, what homework score H *would you have needed for a* 90 *in the course if* $M=80$, $F=85$, *and* $P=100$*?*

Solution:

1) $(1-1)+(4-1)+(9-1)+(16-1)+(25-1)=0+3+8+15+24=50$. Alternatively, by the rules for sums, $\sum k^2-\sum 1=55-5=50$.

2) Each term is 3 times a counting number, from $3\cdot 1$ up to $3\cdot 10$, so it is $\sum_{k=1}^{10}3k$.

3) $90=0.25H+20+29.75+15=0.25H+64.75$, so $0.25H=25.25$ and $H=101$. Unless there was extra credit, a 90 was out of reach!