JOHNS HOPKINS UNIVERSITY
Pre-College Program
AS.110.100 Data Analytics Workshop

Lecture Notes

Aaron Zoll

How to use these notes. These notes are optional. Everything you need for the quizzes, guided practice, discussions, and project is in the lecture videos, and these notes cover the same material, so you can use them in place of a video or alongside one. Each day starts with a TL;DR: a one page summary with the main definitions. If you just want a refresher, that page is enough. The rest of each day goes through the ideas in more detail, with pictures, examples, and a few practice problems (solutions in red right below them). Anything marked as a remark is extra: interesting, but not something you will be tested on.

Contents

Day 1: Sets, Functions, and Sigma Notation **3**
1 Sets and Functions . . . 4
2 Modeling Data with Functions . . . 4
3 Sigma Notation and Indices . . . 5
4 Algebra: Solving for a Variable . . . 7
5 Practice . . . 7

Day 2: Data, Sampling, and Statistics **9**
1 Data . . . 10
2 Sampling . . . 10
3 Statistics . . . 11
4 Practice . . . 12

Day 3: Distributions and Outliers **13**
1 Distributions of Data . . . 14
2 Common Distributions . . . 14
3 Outliers . . . 16
4 Practice . . . 16

Day 4: Center, Spread, and Standard Error **18**
1 Mean, Median, and Mode . . . 19
2 Variance and Standard Deviation . . . 20
3 Standard Error of the Mean . . . 21
4 Practice . . . 22

Day 5: Questions, Hypotheses, and Correlation **24**
1 Meaningful Questions . . . 25
2 Hypothesis Testing and p-Values . . . 25
3 Correlation versus Causation . . . 26
4 Practice . . . 28

Day 6: Probability and Conditional Probability **29**
1 Probability and Independence . . . 30
2 Conditional Probability and Bayes' Theorem . . . 31
3 Practice . . . 33

Day 7: Visualizing Results **34**
1 The Normal Distribution and Error Bars . . . 35
2 Histograms . . . 37
3 Scatter Plots and Lines of Best Fit . . . 37
4 Practice . . . 38

Day 8: Expected Value **40**
1 Expected Value . . . 41
2 Expected Values Add . . . 41
3 The Law of Large Numbers . . . 42
4 Is It a Fair Game? . . . 42
5 Practice . . . 43

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.

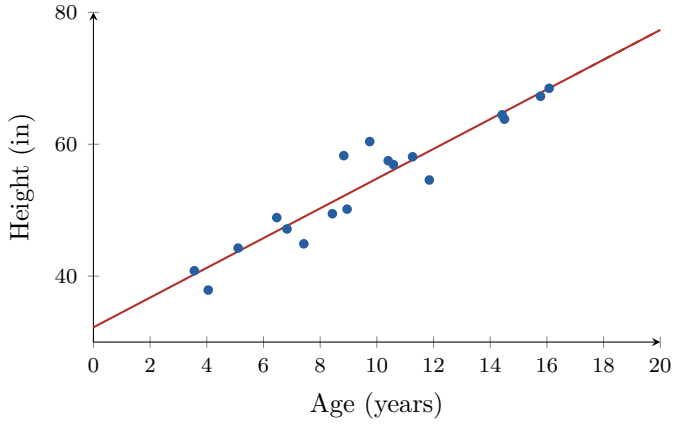
(Sigma notation)

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

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!

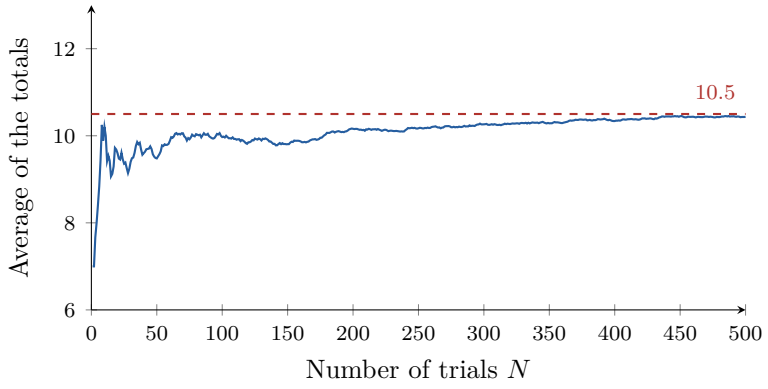
Data Analytics Workshop, JHU Pre-College Program
Day 2: Data, Sampling, and Statistics (optional notes)

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:

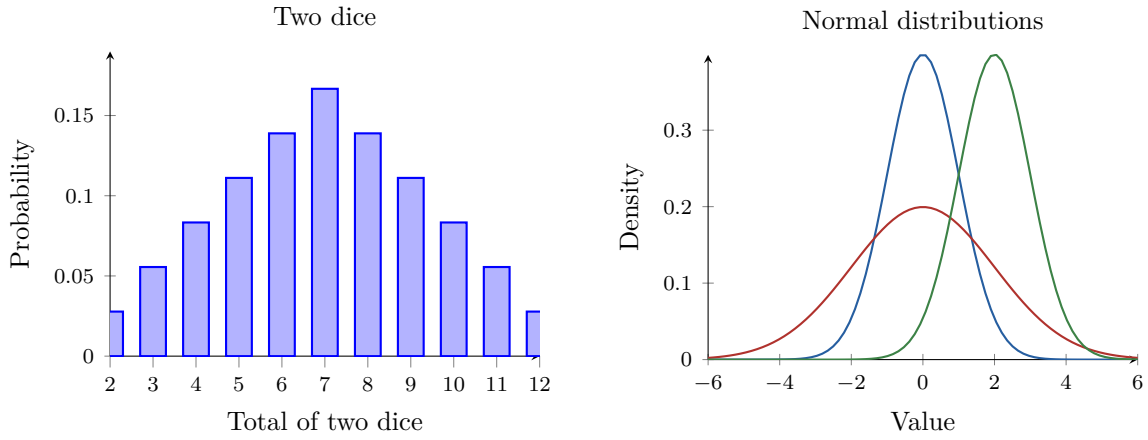
1) (a), (b), and (d) are statistics, since each is computed from the data by a fixed rule (even if (d) is a strange one). (c) is not: two people could disagree about who *looks* tallest, so it is not deterministic.
2) The population is all students at the school, and the sample is the AP Calculus class. 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.

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 μ, where it is centered, and a standard deviation σ, 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.

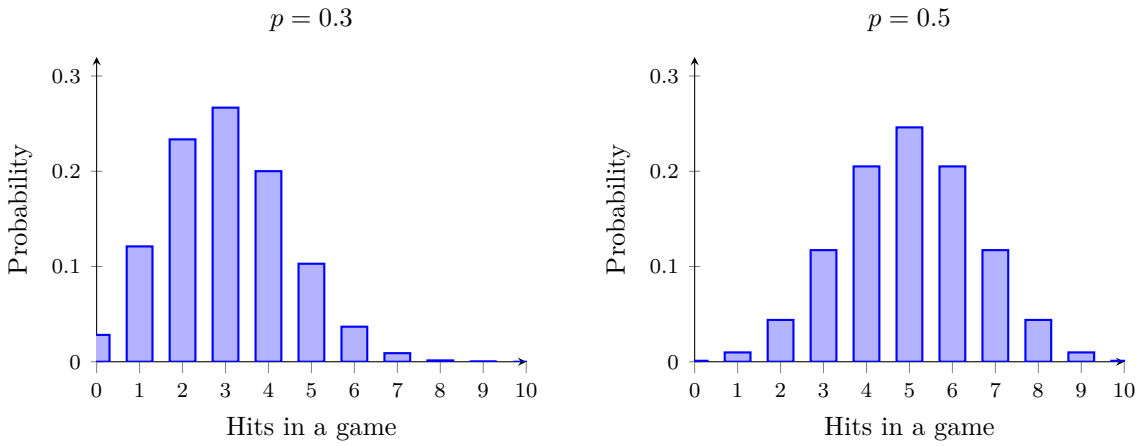
A fair coin has $p = \frac{1}{2}$, while an unfair coin could land heads with probability 0.75 and tails with probability 0.25. Rolling a total of four or higher with two dice is also Bernoulli: either it happens or it does not.

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 μ, which is where the curve is centered, and the standard deviation σ, 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 μ and a spread σ. You can see on the first page how changing μ slides the curve left or right, and increasing σ 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.

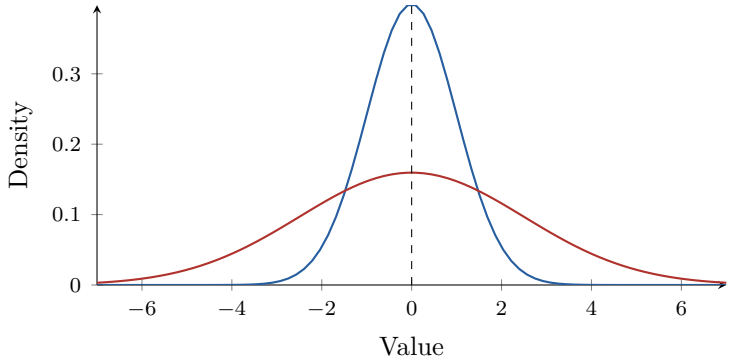
Data Analytics Workshop, JHU Pre-College Program
Day 4: Center, Spread, and Standard Error (optional notes)

TL;DR A distribution can hold a lot of information. Today we summarize it with just a few numbers: where the data is **centered** and how **spread out** it is. For data $x_1, \ldots, x_N$:

(Mean, variance, standard deviation, SEM)

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k, \qquad \sigma^2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \bar{x})^2, \qquad \sigma = \sqrt{\sigma^2}, \qquad \text{SEM} = \frac{\sigma}{\sqrt{N}}.$$

- The **mean** $\bar{x}$ is the balance point (center of mass) of the data. The **median** is the middle value, with half the data on each side. The **mode** is the most common value.
- The median ignores how big the values are, while the mean takes every value into account. For $1, 2, 4, 80, 100$ the median is 4 but the mean is 37.4.
- The **variance** σ^2 is the average squared distance from the mean. The **standard deviation** σ undoes the square, so it has the same units as the data.
- The **standard error of the mean** (SEM) measures how much the *mean* of our data would wobble if we redid the whole experiment. It shrinks as the sample size N grows, which is exactly why large samples give more confident results.



Same mean, different spread: $\sigma = 1$ versus $\sigma = 2.5$

1 Mean, Median, and Mode

A distribution can contain a lot of information, and the goal today is to tell nearly the same story with only a few numbers. The first question is: where is the data centered? There are two natural ways to answer this. We can look for the point with *half of the data* on each side, or we can look for the point that balances *the weight* of the data. These sound similar, but they are different.

For the data $1, 2, 4, 8, 10$, written in increasing order, the middle value 4 has half of the data below it and half above it. This is the **median**. With an even number of points, such as $1, 2, 4, 8$, no single value splits the data in half, so we take the midpoint of the two middle values, here 3.

Now notice that the median ignores how large the values are. If we change the data to $1, 2, 4, 80, 100$, the median is still 4. If these were the weights in pounds of five gerbils, reporting that a typical gerbil weighs 4 pounds would not give the full picture when two of them weigh 80 and 100 (those are some big gerbils). This is where the mean comes in.

Definition 1.1 (Mean)

*The **mean** (or average) of data $x_1, x_2, \ldots, x_N$ is*

$$\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_k.$$

You may have seen a *weighted* average when computing a final grade, like on Day 1. The mean is a weighted average with the *same* weight $\frac{1}{N}$ on every data point, so large values count just as much as small ones. For the gerbils,

$$\bar{x} = \frac{1 + 2 + 4 + 80 + 100}{5} = \frac{187}{5} = 37.4.$$

The mean also works for data that are not single numbers. For points in the plane, we average each coordinate separately. The mean of the five points below is

$$\left(\frac{1 + 3 + 2 - 1 - 2}{5}, \ \frac{4 + 2 - 3 - 3 + 2}{5} \right) = (0.6, 0.4).$$

If these points were the corners of a thin plate, the mean is where we could balance the plate on a finger. In this sense the mean is the *center of mass* of the data. A median for points in the plane is much harder to even define (which point is "in the middle"?), which is one reason the mean is used so much more often.

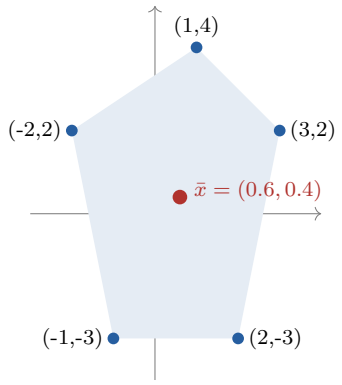
For completeness, here are the formal definitions of the other two.

Definition 1.2 (Median)

*Reorder the data $x_1, \ldots, x_N$ in increasing order as $x_{k_1} \le x_{k_2} \le \cdots \le x_{k_N}$. The **median** $\hat{x}$ is*

$$\hat{x} = \begin{cases} x_{k_{(N+1)/2}}, & N \text{ odd}, \\ \frac{1}{2}\left(x_{k_{N/2}} + x_{k_{N/2+1}} \right), & N \text{ even}. \end{cases}$$

The indices $k_1, \ldots, k_N$ "index the indices": they put the data in increasing order without renaming it. It looks complicated, but it just says "the middle one, or the average of the two middle ones."



The mean of five points is their balance point

Definition 1.3 (Mode)
*The **mode** $\tilde{x}$ of data $x_1, \ldots, x_N$ is the value that occurs most often.*

In $1, 2, 4, 8, 10$ every value occurs once, so every value is a mode. In $1, 2, 4, 8, 8, 10$ the mode is 8. Of the three, the mean is the one we will use the most.

2 Variance and Standard Deviation

Knowing where data is centered is not enough. Look at the two curves on the first page: they have the same center but clearly look very different. To tell them apart, we need to measure how spread out the data is.

Definition 2.1 (Variance)
*The **variance** of data $x_1, \ldots, x_N$ is*

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})^2 .$$

Let's break this down. Each term $x_k - \bar{x}$ measures how far a data point is from the mean. We square it so that every term is positive, since a point far to the left should not cancel out a point far to the right. (An absolute value or a fourth power would also work, but squaring makes later calculations much nicer, and if you take calculus you will see why.) Dividing by N makes the variance the *average squared distance from the mean*. So if most of the data is close to the mean, the variance is small, and if much of the data is far from the mean, the variance is large.

Definition 2.2 (Standard deviation)
*The **standard deviation** is $\sigma = \sqrt{\sigma^2}$.*

This may seem silly at first: why have a name for a number and a separate name for its square root? It is best to think of σ^2 as its own symbol rather than "the standard deviation squared." The real reason we use both is *units*. If the data is measured in seconds, the mean is in seconds, but the

variance is in seconds squared, which is hard to picture. Taking the square root brings us back to seconds, and the standard deviation usually gives a much better sense of how far the data actually spreads. The variance is still great for comparing: a bigger variance always means a bigger spread.

Example 2.1

Let's compute everything for $1, 2, 4, 8, 10$. *The mean is* $\bar{x} = \frac{25}{5} = 5$. *Then*

x_k	1	2	4	8	10
$x_k - \bar{x}$	-4	-3	-1	3	5
$(x_k - \bar{x})^2$	16	9	1	9	25

The squared distances add to 60, *so* $\sigma^2 = \frac{60}{5} = 12$ *and* $\sigma = \sqrt{12} \approx 3.46$. *As a sanity check, notice that the distances* $x_k - \bar{x}$ *add up to* 0. *This always happens (why?), and it is exactly why we need to square them.*

This is the same table you will build in Excel with a column of differences and a column of squares; see the Excel notes for how.

3 Standard Error of the Mean

I have said many times that we want a large sample size. Here is where the math explains why.

Definition 3.1 (Standard error of the mean)

For data $x_1, \ldots, x_N$ *with standard deviation* σ, *the* ***standard error of the mean*** *is*

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}.$$

The variance σ^2 describes the distribution our data comes from. It tells us how far a single *new* data point should be from the mean. If we measured one more plant, σ says roughly how far from the average its height might land. The SEM instead describes the *mean of the data we collected*. If we redid the entire experiment with N new data points and took the mean again, the SEM says roughly how much that mean would change.

Where does the $\sqrt{N}$ come from? The idea is this (we will not prove it). If we add up N independent random variables that each have variance σ^2 and divide by N to take their average, the average has variance $\frac{\sigma^2}{N}$. Its standard deviation is then

$$\sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}},$$

which is exactly the SEM. Averaging many values cancels out a lot of the randomness, since some land above the mean and some below.

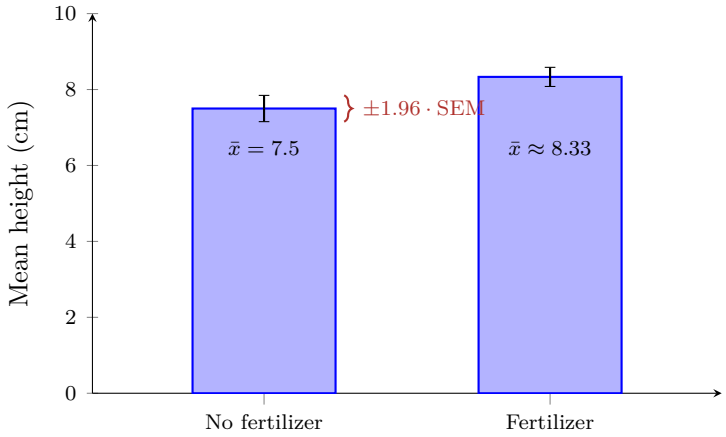
This is why a large sample size matters. If we keep drawing from the same distribution, σ stays about the same while $\sqrt{N}$ grows, so the SEM shrinks. Notably, to cut the SEM in half we need *four times* as much data, since $\sqrt{4N} = 2\sqrt{N}$.

Example 3.1

Suppose we grow plants with and without a fertilizer, measure their heights after two weeks, and plot the two mean heights as bars. The means alone do not say whether the difference is meaningful, because they say nothing about how the data is spread. The error bars on each bar were

$$\bar{x} \pm 1.96 \cdot \text{SEM}.$$

Each error bar is centered at the top of its bar (the mean) and reaches $1.96 \cdot \text{SEM}$ *above and below it, as in the picture below. Whether or not the error bars overlap tells us whether the difference is statistically significant. With a larger sample, the error bars shrink. We return to this (and where* 1.96 *comes from) on Day 7.*



Made up plant heights (6 plants each). The error bars do not overlap, so the difference looks significant.

Remark 1

Excel has two variance functions, `VAR.P` *(divide by* N*, our formula) and* `VAR.S` *(divide by* $N-1$*). In this course we always use* `VAR.P` *and* `STDEV.P`*. The Excel notes say a bit more about the difference.*

4 Practice

Problem 4.1 *1) For the data* $3, 5, 5, 6, 11$*, find the mean, median, mode, variance, and standard deviation.*

2) A dataset has $\sigma = 5$ *and* $N = 100$ *points. What is the SEM? How many points would we need for an SEM of* 0.25*?*

Solution:

1) Mean $\frac{30}{5} = 6$, median 5, mode 5. The distances from the mean are $-3, -1, -1, 0, 5$, with squares $9, 1, 1, 0, 25$ adding to 36. So $\sigma^2 = \frac{36}{5} = 7.2$ and $\sigma \approx 2.68$. Notice the mean is pulled above the median by the 11.
2) SEM $= \frac{5}{\sqrt{100}} = 0.5$. To halve it to 0.25 we need four times the data, $N = 400$. Check: $\frac{5}{\sqrt{400}} = \frac{5}{20} = 0.25$.

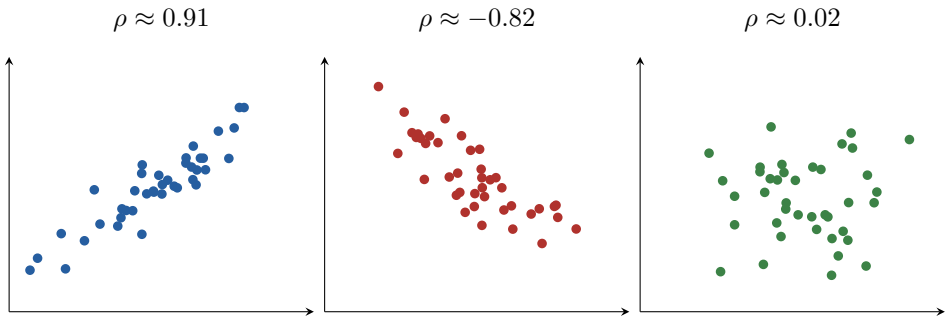
Data Analytics Workshop, JHU Pre-College Program
Day 5: Questions, Hypotheses, and Correlation (optional notes)

TL;DR Today is about the project: how to ask good questions of a dataset and how to tell whether the answer is real or just luck.

- A meaningful question is **observable** (our data can see it), **measurable** (with numbers or categories), and **repeatable** (someone could redo the study). Data is **numerical** (salary, age) or **categorical** (college attended, favorite flavor).
- The **null hypothesis** H_0 says any effect we see is due to chance. The **alternative hypothesis** H_1 says there is a real relationship. We collect data and ask whether we have enough evidence to **reject the null**.
- The ***p*-value** is the probability, *assuming the null hypothesis is true*, of seeing data at least as extreme as ours. If it is below a threshold (often 0.05), the result is **statistically significant**.
- Data analysis does not show *causation*. It measures **correlation**: a number ρ between -1 and 1 describing how one variable tends to change as another changes.

(Covariance and correlation)

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}), \qquad \rho_{x,y} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}.$$



Strong positive, strong negative, and no correlation

1 Meaningful Questions

By now you may have a large dataset for your project. Suppose ours is a sample of 1000 employees at a company, with each person's salary, years at the company, age, race, and gender. Before testing anything, we need questions worth asking. A meaningful question should be

1. **Observable**: we can actually observe it with the data we have. We cannot ask about favorite ice cream flavors if the dataset does not record them.
2. **Measurable**: we can measure it, though not necessarily with numbers.
3. **Repeatable**: the study could be done again, for example at a different company.

Definition 1.1 (Numerical and categorical data)

Numerical *data can be assigned a number, such as salary, years worked, or age.* ***Categorical*** *data sorts each data point into a group, such as race, favorite ice cream flavor, or college attended.*

Categorical data is still measurable. To compare salary across races, we can group people by race, take the average salary of each group, and plot the averages as a bar chart. The kind of data also tells you the kind of visual to make: categories go on a bar chart, while two numerical variables go on a scatter plot (more on this on Day 7).

Meaning is subjective (ask ten philosophers and you'll get ten thousand answers), but a question that is observable, measurable, and repeatable is on the right track. Examples are whether age relates to salary, or whether years at the company relate to salary. Notice the word *relate*. Data analysis rarely tells us that one thing *causes* another, so we try not to phrase questions that way.

2 Hypothesis Testing and p-Values

Even among people who have worked at a company the exact same number of years, salaries vary. They come from some distribution with some mean and some variance. Because of this randomness, we might see that people who have worked longer earn more in our sample, even if no real relationship exists in the company as a whole. Hypothesis testing is how we guard against being fooled by luck.

Definition 2.1 (Null and alternative hypotheses)

The ***null hypothesis*** H_0 *states that any effect we see in the data is due to chance. The* ***alternative hypothesis*** H_1 *states that there is a relationship that is not due to chance alone.*

A null hypothesis can also compare two groups. For two coins that may not be fair, H_0 could state that both coins land heads with the same probability, and H_1 that coin 1 lands heads more often.

Example 2.1

We flip each coin 100 *times.*

	Heads	*Tails*
Coin 1	*89*	*11*
Coin 2	*78*	*22*

Coin 1 came up heads more often, but this could be luck. Coin 2 could secretly be the more unfair one and we just got unlucky with it. The question is whether the data gives us enough evidence to ***reject the null hypothesis****. Notably, with* 1000 *flips each, a gap this large would be much harder to blame on luck.*

So the process is:

1. Ask a meaningful question.
2. State a null hypothesis (the effect is due to chance) and an alternative hypothesis.
3. Collect data.
4. Use a measure of statistical significance, such as error bars or a p-value, to decide whether we can reject the null hypothesis.

Definition 2.2 (p-value)

The ***p*-value** *is the probability, assuming the null hypothesis is true, of observing data at least as extreme as the data we observed.*

A p-value is always between 0 and 1. Since it is computed from the observed data, it depends on the sample size. Here is the key idea: we *pretend* the null hypothesis is true, and ask how surprising our data would be. A small p-value means our data would be very unlikely if only chance were at play. If the p-value is below a chosen threshold, often 0.05, we reject the null hypothesis and call the result **statistically significant**. For example, a p-value of 0.003 for the two coins would give strong evidence that coin 1 really does land heads more often than coin 2.

Remark 1

A very common mistake (even in published papers!) is to say the p-value is "the probability that the result is due to chance," or "the probability the null hypothesis is true." It is neither. It is computed assuming the null hypothesis holds, so it only measures how surprising the data would be in that world. After Day 6 you will have the language for this: the p-value is about $\mathbb{P}(\text{data} \mid H_0)$*, not* $\mathbb{P}(H_0 \mid \text{data})$*, and those can be very different. In Discussion 4 you read an article on why p-values, though one of the most used tests, are not always the best one.*

3 Correlation versus Causation

You may have heard that "correlation does not imply causation." In data analysis we try to avoid the word causation altogether, because it is just too strong. Causation says one thing definitely causes another. Even dropping your phone does not always crack the screen, and we have all had the opposite happen too.

Correlation, on the other hand, is exactly the kind of thing we like: it is observable, repeatable, and most importantly measurable. It is a number describing how one variable tends to change as another changes. If we plot age against salary, we might see salary increase with age, decrease with age, or show no pattern at all, like the three plots on the first page. In the first two we could easily draw a line through the data. In the third, we could not even say which direction the line should go.

Example 3.1

Take $x = (1, 2, 3, 4, 5)$ and $y = (1, 2, 3, 4, 5)$, where the lists are paired by position. As x increases, y increases by exactly the same amount. This is perfect correlation, with correlation 1. If instead $z = (5, 4, 3, 2, 1)$, then z decreases by the same amount each time x increases, and the correlation of x and z is -1.

Definition 3.1 (Covariance and correlation)

For paired data $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$, the ***covariance*** *is*

$$\operatorname{cov}(x, y) = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})\,(y_k - \bar{y})\,,$$

and the ***correlation*** *is*

$$\rho_{x,y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}.$$

The correlation always satisfies $-1 \le \rho_{x,y} \le 1$.

A correlation near 1 means that increasing one variable tends to increase the other. A correlation near -1 means that increasing one variable tends to decrease the other. A correlation near 0 means there is no clear linear relationship.

Why does the sign come out right? Suppose x and y are positively correlated. When x_k is above $\bar{x}$, y_k tends to be above $\bar{y}$, so both factors in the sum are positive. When x_k is below $\bar{x}$, y_k tends to be below $\bar{y}$, so both factors are negative, and their product is positive again! Either way we mostly add positive numbers. When the variables are negatively correlated, the factors tend to have opposite signs and the sum is negative. Dividing by $\sigma_x \sigma_y$ is what keeps the answer between -1 and 1 (the reason why takes more advanced math).

Example 3.2

Let $x = (1, 2, 3, 4, 5)$ and $y = (2, 1, 4, 3, 5)$. Both have mean 3.

$x_k - \bar{x}$	-2	-1	0	1	2
$y_k - \bar{y}$	-1	-2	1	0	2
product	2	2	0	0	4

The products add to 8, so $\operatorname{cov}(x, y) = \frac{8}{5} = 1.6$. Both variances are $\frac{4+1+0+1+4}{5} = 2$, so $\sigma_x = \sigma_y = \sqrt{2}$ and

$$\rho_{x,y} = \frac{1.6}{\sqrt{2}\sqrt{2}} = \frac{1.6}{2} = 0.8.$$

That is, y mostly goes up with x, but not perfectly. I invite you to make up some data of your own, compute its correlation by hand, and then check it with `CORREL` *in Excel.*

Remark 1

The covariance of x with itself is the variance:

$$\operatorname{cov}(x,x) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2 = \sigma_x^2, \qquad \text{so} \qquad \rho_{x,x} = \frac{\sigma_x^2}{\sigma_x \sigma_x} = 1.$$

Any data is perfectly correlated with itself, which is exactly what we should expect.

Two warnings to keep in mind for your project.

Remark 2

Correlation only sees lines. *Take $x = (-2,-1,0,1,2)$ and $y = x^2 = (4,1,0,1,4)$. Here y is completely determined by x, but you can check that $\rho_{x,y} = 0$! As x increases, y first goes down and then goes up, and these cancel. Always plot your data, don't just trust one number.*

Remark 3

Hidden variables. *Across a year, ice cream sales and sunburns are strongly positively correlated. Ice cream does not cause sunburns (I hope). Both go up in the summer, when it is hot and sunny. A third variable driving both is called a confounding variable, and it is one of the main reasons correlation does not imply causation.*

4 Practice

Problem 4.1 *1) Is each question observable, measurable, and repeatable with a dataset of students' GPA, hours of sleep, and favorite subject? (a) Do students who sleep more have higher GPAs? (b) Does sleep cause better grades? (c) Do students who like math more have higher GPAs?*

2) Compute $\rho_{x,z}$ for $x = (1,2,3,4,5)$ and $z = (5,4,3,2,1)$ using the formula.

Solution:

1) (a) Yes: both are numerical and recorded. (b) Not as stated: data can show a relationship, but not causation. Rephrase it as (a). (c) Not quite: we know each student's favorite subject, not how much they like math. Comparing the average GPA of students whose favorite subject is math against the others would work.
2) Both means are 3. The deviations are $(-2,-1,0,1,2)$ and $(2,1,0,-1,-2)$, with products $-4,-1,0,-1,-4$ adding to -10. So cov $= -2$, both variances are 2, and $\rho_{x,z} = \frac{-2}{2} = -1$.

Data Analytics Workshop, JHU Pre-College Program
Day 6: Probability and Conditional Probability (optional notes)

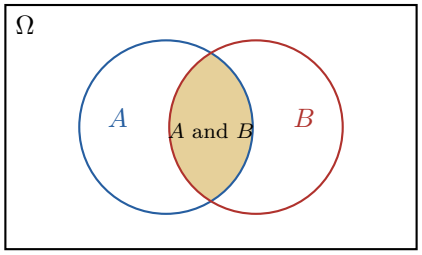
TL;DR Today is a more theoretical day, but it is what makes sense of everything else we do with data. Everything in this course has finitely many outcomes, so probability comes down to counting.

(Probability)

*The **sample space** Ω is the set of all possible outcomes, and an **event** A is a subset of Ω. If every outcome is equally likely,*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\textit{number of outcomes in } A}{\textit{number of outcomes in } \Omega}.$$

- Events A and B are **independent** if $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\,\mathbb{P}(B)$. Intuitively, one has no effect on the other (two coin flips). This **product rule** only works for independent events!
- The **conditional probability** of A given B is $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}$. Knowing B happened shrinks the sample space down to B.
- **Bayes' theorem** flips a conditional probability around: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$.
- In general $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$. Mixing these up has caused real damage, including in court cases.



Given B, only the red circle matters, and the event becomes the overlap

1 Probability and Independence

We have used the word probability a lot without saying what it is. Everything in this course involves only finitely many outcomes, so we can define probability by counting. I am going to introduce some new symbols, but bear with me; they are just symbols, nothing too scary.

Definition 1.1 (Sample space, event, probability)

*The **sample space** Ω is the set of all possible outcomes. An **event** A is a subset of Ω. If every outcome is equally likely, the **probability** of A is*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|},$$

the number of outcomes in A divided by the number of outcomes in Ω.

For a coin, $\Omega = \{H, T\}$. For a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. For a deck of cards, Ω is all 52 cards, and events could be "the card is a heart" or "the card is a three." Two notable probabilities are

$$\mathbb{P}(\Omega) = \frac{|\Omega|}{|\Omega|} = 1 \qquad \text{and} \qquad \mathbb{P}(\emptyset) = \frac{0}{|\Omega|} = 0,$$

where $\emptyset$ is the empty set (nothing happens). Since every event is somewhere between nothing and everything, every probability is between 0 and 1.

Definition 1.2 (Independence)

*Two events A and B are **independent** if*

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Intuitively, independence means one event has no effect on the other. Flipping a coin today has no effect on flipping it again in thirty seconds, or in thirty years. The same is true at a slot machine: losing 100 times in a row does *not* make the next pull more likely to win, no matter how much it feels like it should. Let's check the definition against our intuition.

Example 1.1

Flip a coin twice. Let A be the event that the first flip is tails and B the event that the second flip is tails. The sample space is

$$\Omega = \{HH, HT, TH, TT\},$$

and only TT is in both A and B, so $\mathbb{P}(A \text{ and } B) = \frac{1}{4}$. Since $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$,

$$\mathbb{P}(A)\,\mathbb{P}(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = \mathbb{P}(A \text{ and } B),$$

and the events are independent, just as we expected.

Example 1.2

Flip a coin once. Let A be the event that the flip is tails and B the event that the same flip is heads. Then $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$, but no outcome is both heads and tails, so

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(\emptyset) = 0 \neq \frac{1}{4}.$$

These events are not independent. In fact they are as dependent as it gets: if A occurs, B cannot.

Example 1.3

Shuffle a deck, draw a card, and let A be the event that it is a club. Put the card back, shuffle again, draw, and let B be the event that this card is a club. There are 13 clubs, so $\mathbb{P}(A) = \mathbb{P}(B) = \frac{13}{52} = \frac{1}{4}$. There are 52^2 ways to draw two cards and 13^2 ways to draw two clubs, so

$$\mathbb{P}(A \text{ and } B) = \frac{13^2}{52^2} = \frac{1}{16} = \mathbb{P}(A)\,\mathbb{P}(B).$$

Drawing with replacement gives independent events, just like flipping a coin.

Now suppose we do not put the first card back. If the first card was a club, there are fewer clubs left, so the second draw is less likely to be a club. If it was not a club, the second draw is more likely to be one. So the first draw affects the second, and indeed

$$\mathbb{P}(A \text{ and } B) = \frac{13}{52} \cdot \frac{12}{51} = \frac{1}{17} \neq \frac{1}{16}.$$

Without replacement the events are not independent.

The **product rule** for independent events makes many probabilities easy to compute, since we can find each probability on its own and multiply.

Example 1.4

Suppose you pick an outfit from 3 shirts, 4 pairs of pants, and 2 pairs of shoes, each chosen independently at random. There are $3 \cdot 4 \cdot 2 = 24$ outfits. Rather than listing all 24, the probability of choosing shirt 1, pants 2 or 3, and shoe 1 is just

$$\frac{1}{3} \cdot \frac{2}{4} \cdot \frac{1}{2} = \frac{1}{12}.$$

The product rule only works for independent events. In Discussion 4 you read about real trials where an expert multiplied probabilities of events that were *not* independent, and got a number that was wildly wrong, with serious consequences.

2 Conditional Probability and Bayes' Theorem

Probability always depends on the conditions at play. Even independence is a condition. If I roll a die, hide it, and tell you the result is odd, the sample space shrinks from $\{1, 2, 3, 4, 5, 6\}$ to $\{1, 3, 5\}$,

and the probability that the result is 3 becomes $\frac{1}{3}$ instead of $\frac{1}{6}$. The information changed the probability.

Definition 2.1 (Conditional probability)

The probability of A given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}.$$

Picture A and B as two overlapping circles inside Ω, like the first page, and think of probability as area. $\mathbb{P}(A)$ is the area of A compared to all of Ω. Knowing B is true means we only look inside B, so B becomes our new sample space and the only part of A we can land in is the overlap. That is exactly the formula: the overlap, compared to B.

Being in the overlap is the same as having A occur, and then having B occur given that A did. So $\mathbb{P}(A \text{ and } B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$. Substituting this into the top of the definition gives the following.

Theorem 2.2 (Bayes' theorem)

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

This can look like it says nothing new, but often we know three of these four quantities and not the fourth, and Bayes' theorem hands us the last one. It also suggests a way to think about drawing conclusions from data: to make an inference, we must take into account prior information.

Example 2.1

Alice and Bob have two children, and each child is equally likely to be a boy or a girl. Given that one of them is a boy, what is the probability that the other is a boy?

This is not the same as asking about the next child after a first boy. That probability is $\frac{1}{2}$, since the first child has no effect on the second. Here both children are already born. The possibilities are

$$\{BB, BG, GB, GG\}.$$

Knowing one child is a boy removes GG, leaving $\{BB, BG, GB\}$. Only BB has the other child also a boy, so the answer is $\frac{1}{3}$, not $\frac{1}{2}$ as most people guess! With the formula: $\mathbb{P}(\text{both boys} \mid \text{at least one boy}) = \frac{1/4}{3/4} = \frac{1}{3}$.

Remark 1

In general, $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ are very different. The probability that someone speaks English, given that they are President of the United States, is essentially 1. The probability that someone is President of the United States, given that they speak English, is about one in a

billion. Mixing these two up in court is called the prosecutor's fallacy, and it shows up in the Discussion 4 readings. It is also exactly the p-value mistake from Day 5.

3 Practice

Problem 3.1

Roll two dice, one red and one blue. Let A be the event that the red die shows a 3.

1) *Let B be the event that the total is 7. Are A and B independent?*
2) *Let C be the event that the total is 8. Are A and C independent?*
3) *What is $\mathbb{P}(C \mid A)$?*

Solution:

There are 36 equally likely outcomes, and $\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}$.

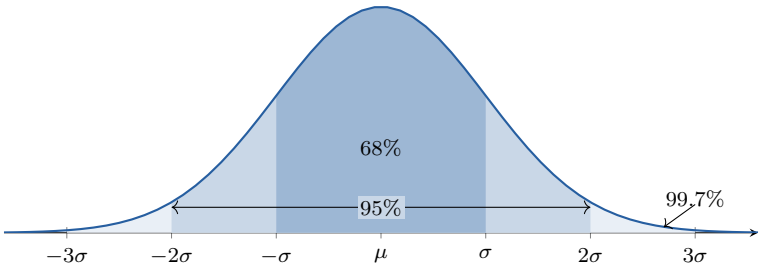
1) $\mathbb{P}(B) = \frac{6}{36} = \frac{1}{6}$, and the only way to get both is red 3, blue 4, so $\mathbb{P}(A \text{ and } B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$. Independent! Whatever the red die shows, there is exactly one blue value that makes 7.
2) $\mathbb{P}(C) = \frac{5}{36}$, and $\mathbb{P}(A \text{ and } C) = \frac{1}{36}$ (red 3, blue 5). But $\frac{1}{6} \cdot \frac{5}{36} = \frac{5}{216} \neq \frac{1}{36}$, so these are *not* independent. (If the red die showed a 1, a total of 8 would be impossible.)
3) $\mathbb{P}(C \mid A) = \frac{1/36}{1/6} = \frac{1}{6}$, which is bigger than $\mathbb{P}(C) = \frac{5}{36}$. Knowing the red die is a 3 makes an 8 slightly more likely.

Data Analytics Workshop, JHU Pre-College Program
Day 7: Visualizing Results (optional notes)

TL;DR Today gives you the last tools for your project: how to show your answers in a picture, and how to tell from the picture whether the answer is significant.

(68-95-99.7 rule)
For normally distributed data with mean μ and standard deviation σ, about 68% of the data lies within σ of the mean, 95% within 2σ, and 99.7% within 3σ.

- The **central limit theorem**: the average of many independent samples is approximately normally distributed, with standard deviation equal to the SEM, $\frac{\sigma}{\sqrt{N}}$.
- So about 95% of the time, the mean of our data lands within $2 \cdot$SEM (more precisely $1.96 \cdot$SEM) of the true mean. This is why we draw **error bars** at $\bar{x} \pm 2 \cdot \text{SEM}$.
- To compare groups (categorical data), use a **bar chart with error bars**. If the error bars of two groups do not overlap, we have evidence the groups truly differ.
- To see how one numerical variable is spread out, use a **histogram**. To compare two numerical variables, use a **scatter plot**, and add a **line of best fit**. Its R^2 value (between 0 and 1) measures how well the line fits.



The 68-95-99.7 rule

1 The Normal Distribution and Error Bars

For your project you will make visuals that answer the questions you asked on Day 5. For comparing groups, such as average salary by race, or plant height with and without a fertilizer, a bar chart with error bars works really well. Each bar is the mean of a group, and each error bar extends a multiple of the standard error of the mean above and below it. The bars on their own are not enough: with enough variability and randomness, two means can differ just by luck. The error bars are what let us judge significance. Let's see why.

Theorem 1.1 (68-95-99.7 rule)

For normally distributed data with mean μ and standard deviation σ, about

- 68% *of the data lies within* σ *of* μ,
- 95% *lies within* 2σ *of* μ,
- 99.7% *lies within* 3σ *of* μ.

Equivalently, if the total area under the bell curve is 100%, then 68% of the area lies between $\mu - \sigma$ and $\mu + \sigma$, as on the first page. Notice how quickly the tails die off: three standard deviations account for nearly everything.

But our data is not always normal. Heights might be, but dice, coin flips, and salaries usually are not. Here is the amazing part.

Theorem 1.2 (Central limit theorem, informally)

If we take the average of N independent samples from the same distribution, then for large N that average is approximately normally distributed, no matter what the original distribution looked like.

You can see this below. On the left are 2000 random numbers between 0 and 1, every value equally likely, so the histogram is flat (uniform). On the right, each data point is instead the *average* of 10 such random numbers. Same ingredients, but now a bell curve! This is the same thing you saw with the Galton board in Discussion 3 and with binomial distributions on Day 3.

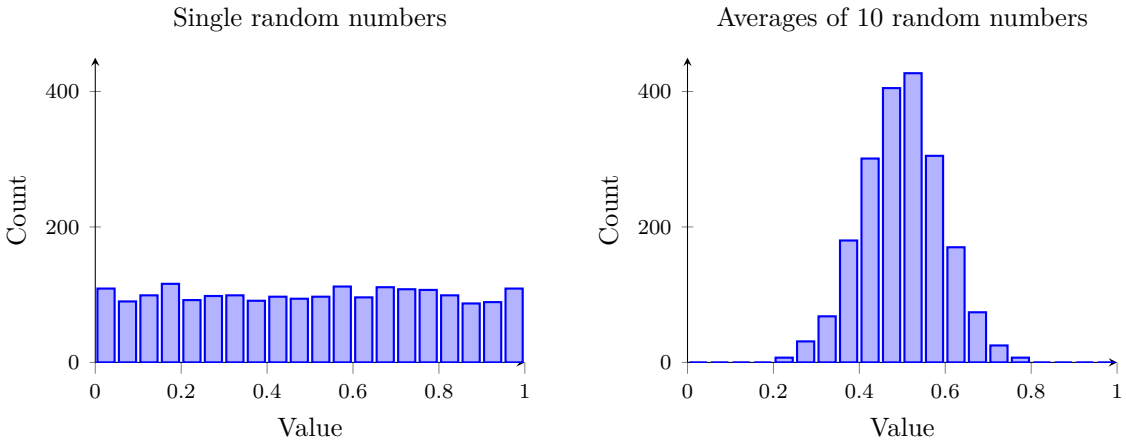
Now put the pieces together. On Day 4 we saw that the average of N samples has standard deviation $\frac{\sigma}{\sqrt{N}}$, which is the SEM. The central limit theorem says that average is roughly normal. So applying the 68-95-99.7 rule to the *average*, about 95% of the time the mean of our data falls within $2 \cdot \text{SEM}$ of the true mean. (In the simulation above, the averages have standard deviation about 0.091, which matches $\frac{\sigma}{\sqrt{10}} \approx 0.091$ for a uniform distribution. As a sanity check, they are also centered at 0.5.)

This is why we draw error bars at

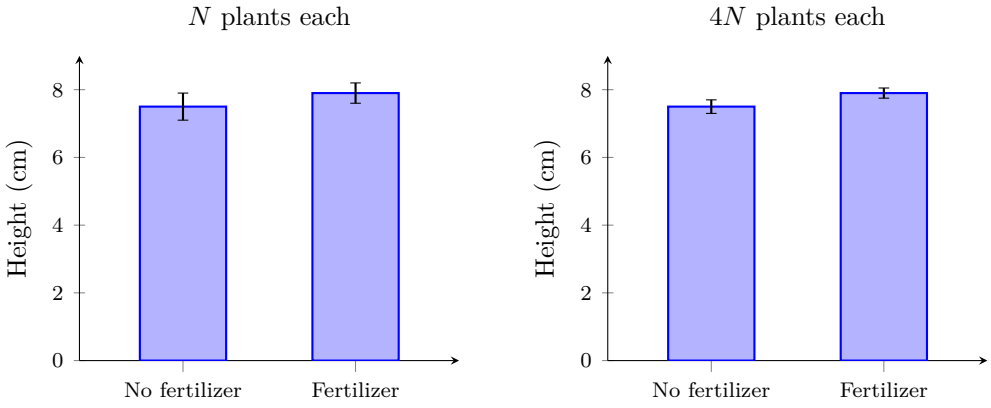
$$\bar{x} \pm 2 \cdot \text{SEM}.$$

The more precise multiple for exactly 95% is the 1.96 we used on Day 4. If the error bars of two groups do not overlap, we have evidence that the groups truly differ, and not just by luck. A bar chart with error bars gives both an answer to the question and a quick visual of it.

In the left chart the error bars overlap (7.1 to 7.9 against 7.6 to 8.2), so we cannot conclude the fertilizer does anything; the difference could be luck. In the right chart, with four times as many



The central limit theorem in action (a simulation with 2000 data points each)



Made up plant heights with error bars $\bar{x} \pm 2 \cdot \text{SEM}$. Same means, but four times the data halves the error bars.

plants, the error bars are half as long and no longer overlap (7.3 to 7.7 against 7.75 to 8.05). Same averages, but now the difference is significant. This is the whole story of why sample size matters, in one picture.

Remark 1

When you make a bar chart, the vertical axis should start at 0. Starting it at, say, 7 makes a small difference look enormous. The Excel notes have an example of how misleading this can be.

2 Histograms

A **histogram** is a bar chart for numerical data. We split the range of values into intervals called **bins** and count how many data points fall in each one. The two plots above showing the central limit theorem are histograms. Histograms do not need error bars to be useful: they show how the data is distributed, and we can still read off roughly where it is centered and how spread out it is.

A common use is with images, which is what Discussion 2 is about. A video (like the lectures) is a sequence of images in quick succession, and each pixel of an image has a brightness, a color, and a saturation. In a black and white image only brightness matters, and tallying how many pixels have each brightness gives a histogram. There is a whole field of image analysis built on this. For example, increasing the brightness of every pixel by the same amount shifts the whole histogram to the right. You can try this with your own photos in the Pixel Data Explorer.

Remark 1

Choosing the bin width matters. Bins that are too wide hide the shape of the data (everything lands in two or three bars), and bins that are too narrow make it look spiky and random. Try a few widths and pick the one that shows the shape most honestly.

3 Scatter Plots and Lines of Best Fit

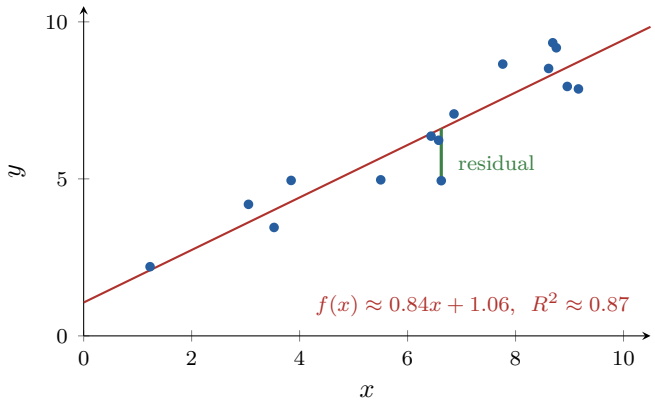
To compare two numerical variables, like age and salary, we use a **scatter plot**: one point per data point, with one variable on each axis. On Day 1 we drew a line of best fit through such a plot, and on Day 5 we measured how "line-like" it is with the correlation ρ. Here is how those connect.

Definition 3.1 (Line of best fit and residuals)

*For a line $f(x) = mx + b$ and a data point (x_k, y_k), the **residual** is $y_k - f(x_k)$, the vertical distance from the point to the line. The **line of best fit** (or **linear regression** line) is the line that makes the sum of the squared residuals, $\sum_{k=1}^{N} (y_k - f(x_k))^2$, as small as possible.*

This should feel familiar: just like variance, we square the distances so that points above and below the line do not cancel out. Excel computes this line for you as a **trendline** (see the Excel notes).

Excel also reports an R^2 value for the trendline. It is between 0 and 1, and it measures how much of the up and down in y the line accounts for. An R^2 near 1 means the points hug the line, and an R^2 near 0 means the line barely helps. For a line of best fit, R^2 is exactly the correlation squared, $R^2 = \rho_{x,y}^2$, which ties everything together. In the plot above, $\rho \approx 0.935$ and $R^2 \approx 0.87$.



A line of best fit, with one residual drawn in green

Remark 1

The slope tells you how much y changes when x increases by 1, and R^2 tells you how reliable that is. A steep line with a low R^2 is a weak result, and a shallow line with a high R^2 can be a strong one. Report both, and remember from Day 1 to be careful predicting outside the range of your data.

4 Practice

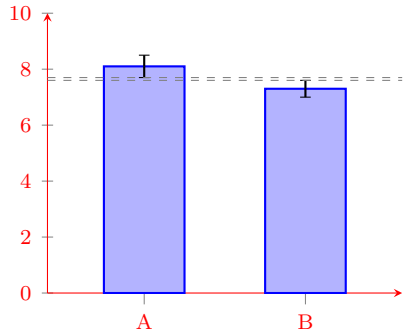
Problem 4.1 *1) Group A has mean 8.1 and SEM 0.2. Group B has mean 7.3 and SEM 0.15. Using error bars of $\pm 2 \cdot$ SEM, is the difference significant?*

2) IQ scores are designed to be normal with mean 100 and standard deviation 15. About what percent of people have an IQ between 85 and 115? Above 130?

3) A trendline has $R^2 = 0.64$ and a positive slope. What is the correlation?

Solution:

1) Group A's error bar runs from $8.1 - 0.4 = 7.7$ to 8.5, and group B's from $7.3 - 0.3 = 7.0$ to 7.6. They do not overlap, so yes. It helps to draw it:



The dashed lines at 7.7 (bottom of A) and 7.6 (top of B) show the small gap between them.

2) 85 to 115 is within one standard deviation, so about 68%. Above 130 is more than 2σ above the mean. 95% is within 2σ, leaving 5% outside, split evenly between the two tails, so about 2.5%.

3) $\rho^2 = 0.64$, so $\rho = \pm 0.8$, and since the slope is positive, $\rho = 0.8$.

Data Analytics Workshop, JHU Pre-College Program
Day 8: Expected Value (optional notes)

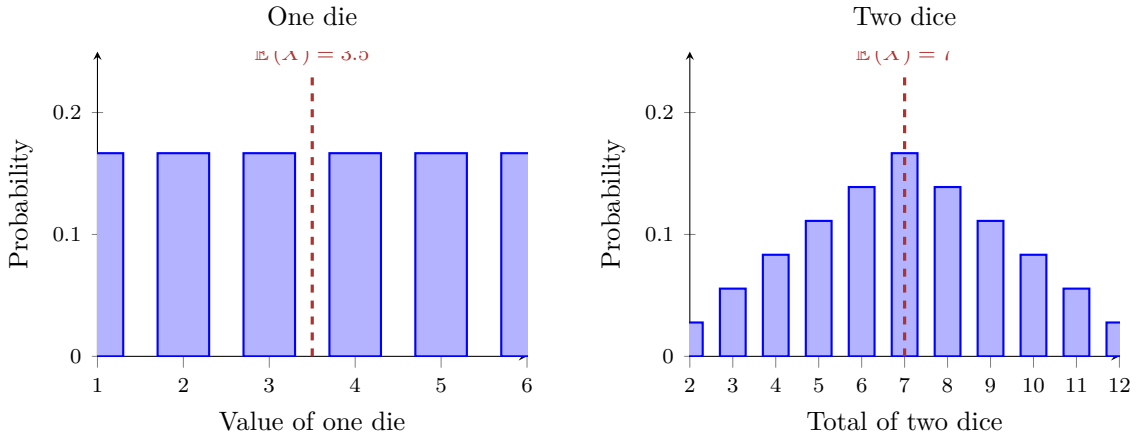
TL;DR The mean is an average of data we already collected. The **expected value** is the average we *expect* in the long run, computed from the distribution itself, before any data is collected.

(Expected value)

*If a random variable X can take the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, its **expected value** is*

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

- It is a weighted average of the possible values, weighted by how likely each one is. One die has $\mathbb{E}(X) = 3.5$, even though you can never roll a 3.5.
- Expected values add: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. So three dice have expected total $3 \cdot 3.5 = 10.5$, exactly where the average settled on Day 2.
- The **law of large numbers**: as we collect more and more data, the mean of the data gets closer and closer to the expected value. The SEM from Day 4 says how close.
- Expected value tells us whether a game or bet is fair: a fair game has expected winnings of 0.



The expected value is the balance point of the distribution

1 Expected Value

On Day 4 we computed the mean of data we had already collected. But often we want to know what to expect *before* collecting anything. If I hand you a die, what should you expect to roll, on average? We do not need to roll it a thousand times to answer this, since we already know the distribution: each face comes up with probability $\frac{1}{6}$.

Definition 1.1 (Expected value)

If a random variable X can take the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, its ***expected value*** *is*

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \, p_i.$$

That is, multiply each value by how likely it is, and add them up. Notably, this is a weighted average, just like the grade formula on Day 1, except now the weights are probabilities.

Example 1.1

For one die,

$$\mathbb{E}(X) = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = \frac{21}{6} = 3.5.$$

You will never actually roll a 3.5! The expected value does not have to be a possible outcome. It is the long run average, and the balance point of the distribution, like the mean was the balance point of the data.

Example 1.2

For a Bernoulli random variable that equals 1 with probability p (success) and 0 with probability $1 - p$ (failure),

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

So the expected number of hits in one at bat, for our batter from Day 3 who hits 30% of the time, is 0.3.

2 Expected Values Add

The most useful fact about expected value is this.

Proposition 2.1 (Linearity of expectation)

For any random variables X and Y and any constant c,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \qquad \text{and} \qquad \mathbb{E}(cX) = c\,\mathbb{E}(X).$$

This looks just like the rules for sums on Day 1, and that is no coincidence: an expected value *is* a sum. What is surprising is that this works even when X and Y are not independent. Let's see how much work it saves.

Example 2.1

The expected total of two dice is $3.5 + 3.5 = 7$*, which matches the peak of the distribution on the first page. The expected total of three dice is* $3 \cdot 3.5 = 10.5$*. This is exactly the number the running average settled on in the Day 2 simulation! Computing it directly would mean listing all* $6^3 = 216$ *outcomes.*

Example 2.2

Our batter gets 10 *at bats per game and hits each with probability* 0.3*. The number of hits is a sum of* 10 *Bernoulli random variables, so the expected number of hits per game is* $10 \cdot 0.3 = 3$*. That is exactly where the binomial distribution peaked on Day 3. In general, a binomial distribution with* n *trials and success probability* p *has expected value* np*.*

3 The Law of Large Numbers

How does expected value connect back to data? The mean of our data is a statistic, and it is our best guess at the expected value.

Theorem 3.1 (Law of large numbers, informally)

If we collect more and more independent samples of a random variable X*, the mean of the samples gets closer and closer to* $\mathbb{E}(X)$*.*

This is what you saw on Day 2: the average of three dice wandered around at first and then settled near 10.5. The SEM from Day 4 tells us how close to expect: with N samples, the mean is usually within about $2 \cdot \frac{\sigma}{\sqrt{N}}$ of the expected value. So more data does not change the expected value; it changes how well our data can find it.

4 Is It a Fair Game?

Expected value is exactly how casinos (and insurance companies) make money. A game is **fair** if your expected winnings are 0.

Example 4.1

You pay \$1 *to roll a die. If you roll a* 6*, you get* \$5 *back; otherwise you get nothing. Your profit is* \$4 *with probability* $\frac{1}{6}$ *and* $-$\$1 *with probability* $\frac{5}{6}$*, so*

$$\mathbb{E}(\textit{profit}) = 4 \cdot \tfrac{1}{6} + (-1) \cdot \tfrac{5}{6} = -\tfrac{1}{6} \approx -\$0.17.$$

On average you lose about 17 *cents per game. You might win the first few, but by the law of large numbers, play long enough and you will lose about* 17 *cents a game. For the game to be fair, a* 6 *would need to pay back* \$6 *(check this!).*

Remark 1

Expected value can be surprising, and that is the point of Discussion 5. Straws weighing anywhere from 0 to 20 pounds are piled on a camel until the total goes over 100 pounds. A single straw weighs 10 pounds on average, so you might expect the final straw to weigh 10 pounds on average too. Does it? Run the camel straw simulator a few hundred times and see what the data says before you decide.

5 Practice

Problem 5.1 *1) A spinner lands on 1 with probability $\frac{1}{2}$, on 2 with probability $\frac{1}{3}$, and on 6 with probability $\frac{1}{6}$. What is its expected value?*

2) What is the expected total of 20 dice?

3) A raffle sells 500 tickets for \$2 each, and one ticket wins \$600. What is the expected profit of buying one ticket? Is it fair?

Solution:

1) $1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{3} + 6 \cdot \frac{1}{6} = \frac{1}{2} + \frac{2}{3} + 1 = \frac{13}{6} \approx 2.17$.
2) By linearity, $20 \cdot 3.5 = 70$.
3) You win \$598 with probability $\frac{1}{500}$ and lose \$2 with probability $\frac{499}{500}$, so the expected profit is $\frac{598-998}{500} = -\$0.80$. Not fair: the raffle collects \$1000 and pays out \$600, so on average each ticket loses $\frac{400}{500} = \$0.80$.