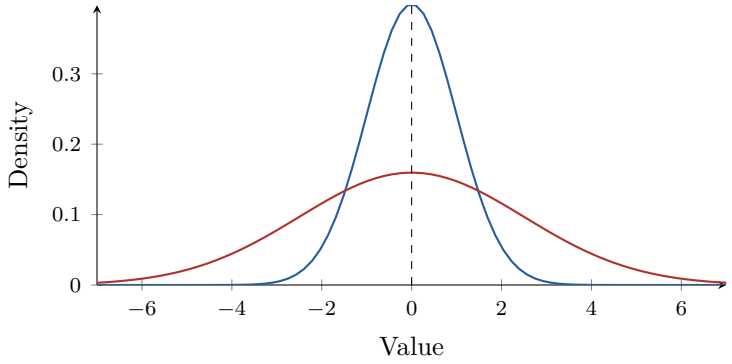
Data Analytics Workshop, JHU Pre-College Program
Day 4: Center, Spread, and Standard Error (optional notes)

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.

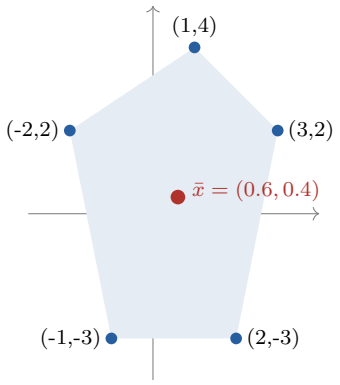
For completeness, here are the formal definitions of the other two.

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.

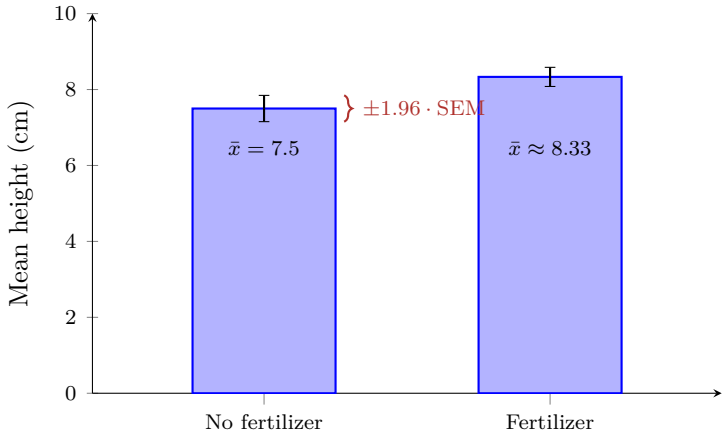
This is why a large sample size matters. If we keep drawing from the same distribution, σ stays about the same while $\sqrt{N}$ grows, so the SEM shrinks. Notably, to cut the SEM in half we need *four times* as much data, since $\sqrt{4N} = 2\sqrt{N}$.

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: $\frac{5}{\sqrt{400}} = \frac{5}{20} = 0.25$.