LESSON 7: MEASURES OF CENTER
Where is the Data?

PART A: FOUR MEASURES OF CENTER

Here is a list of 100 test scores. How would you find a **central (or representative)** score?

70	40	60	93	86	87	13	86	66	80
71	65	95	75	69	47	72	100	39	99
90	63	81	88	82	51	89	58	46	95
50	15	81	52	47	76	90	70	89	44
46	53	46	52	37	38	54	64	70	40
94	60	63	56	92	92	63	86	56	87
56	51	83	79	91	87	53	97	84	49
91	81	69	82	77	91	71	90	51	92
68	88	38	84	54	47	50	14	48	29
61	97	79	36	52	93	46	92	28	38

Let's make things easier. Let's go from 100 scores to five.

Example 1 (Five Test Scores)

The five students in a class take a test. Their scores in points are as follows:

80 76 100 83 100

How can we find a single number that tells us how well the class did? *§*

Let's look at four possibilities for measuring the center of a quantitative data set.

1) [Arithmetic] Mean or Average

There are other measures called means, but the arithmetic mean (or simply "the mean") is by far the most common.

In Example 1

$$\begin{aligned}\text{Mean} &= \frac{\text{Sum of all data values}}{\text{Number of values}} \\ &= \frac{80+76+100+83+100}{5} \\ &= \frac{439}{5} \\ &= 87.8 \text{ points}\end{aligned}$$

WARNING 1: Group the numerator. You must group or compute ("process") the numerator before dividing by 5. You can do this by placing grouping symbols like parentheses around the numerator … or by pressing "ENTER" (for example) on a calculator before dividing.

Think About It: What would be wrong with entering the following on your calculator: $80+76+100+83+100 \div 5 = ?$

Remember to write units, such as "points", in your final answer where appropriate.

Notes on Rounding:

• For now, we will typically round off our **final** answers to **one more decimal place** than the number of decimal places provided in the given data. In Example 1, because the given data values are integers (rounded off to zero decimal places), we round off our final answers to one decimal place.

• **Avoid rounding intermediate results.** Calculator memory can help.

• Always read instructions on exams. They take precedence.

We will discuss trimmed means in Part D.

2) Median

Assume that we have a population data set of size N.

If N is **odd**, then the **median** is the data value in the **middle** position **after sorting** the data in increasing or decreasing order.

• Since ties are permitted, we should technically say nondecreasing or nonincreasing order.

In Example 1, the test scores (in points) were:

80 76 100 83 100

We must first **sort** the five data values. **WARNING 2: Sort first!**

76 80 **83** 100 100
↑

The median is 83 points.

Observe that there are as many data values **below** the median as above. Two values are below; two are above.

• If the median is tied with other scores, then this might not be the case.

If N is **even**, then the **median** is the average of (the midpoint between) the two data values in the two middle positions after sorting the data.

Example 2 (Test Scores; N is Even)

Let's say there were only four test scores:

80 76 100 83

We must first sort them.

76 80 83 100
↑ ↑

The two middle values are 80 and 83, so we take their **average**, $\frac{80+83}{2} = 81.5$. The median is 81.5 points. §

Median Position Number

The position number of the median in a **sorted** list of N quantitative data values is given by: $\frac{N+1}{2}$

• If $N = 5$, as in Example 1, the **median position number** is:

$$\frac{N+1}{2} = \frac{5+1}{2} = 3$$

Therefore, the median is both the 3rd-lowest and the 3rd-highest number in the list.

Position #:	1	2	**3**	4	5
	76	80	**83**	100	100
			↑		

Median = 83 points

• If $N = 4$, as in Example 2, the **median position number** is:

$$\frac{N+1}{2} = \frac{4+1}{2} = 2.5$$

Therefore, the median is the **average** of the 2nd- and 3rd-lowest numbers in the list (or the 2nd- and 3rd-highest).

Position #:	1	**2**	**3**	4
	76	80	83	100
		↑	↑	

Median = 81.5 points

• If $N = 99$, the **median position number** is:

$$\frac{N+1}{2} = \frac{99+1}{2} = 50$$

Therefore, the median is both the 50th-lowest and the 50th-highest number in the list.

Position #:	1	2	3	…	49	**50**	51	…	97	98	99
	<– –49 positions– –>						<– –49 positions– –>				

3) Mode

The **mode** is the **most frequent** data value, if any, in the data set.

- We are using the term "mode" a bit differently here compared to Lesson 6, when we referred to "humps" in histograms.

In Example 1, the test scores (in points) were:

80 76 100 83 100

The **mode** is 100 points.

- How wonderful! But is it really? See **WARNING 3** below.
- We do not have to go one decimal point further for modes.

WARNING 3: Pros and cons of modes. Although it is often easy to find **modes** (and they can be used for **qualitative data** – see below), they might be **questionable as measures of centrality** for quantitative data. The mode may actually be an **extreme value**, as in Example 1, or it may just be the result of simple **coincidence**.

A data set could have no mode, one mode, or more than one mode.

- The data set $\{2.41, 3.62, 7.25\}$ has **no mode**.
- The data set $\{50, 50, 70, 90, 90\}$ has **two modes**, 50 points and 90 points. The data set is bimodal. (We are using the term "bimodal" a bit differently from Lesson 6.)

The mode can also be used for **qualitative data**, such as the presidential party data in Lesson 4, Example 3.

Party	Frequency
Democratic (**D**)	16
Republican (**R**)	19
Whig (**W**)	4
Other (**O**)	6

The mode here would be "Republican."

4) Midrange

The **midrange** is the **average** of (or midpoint between) the **lowest** and **highest** values in the data set.

$$\text{Midrange} = \frac{\text{Min} + \text{Max}}{2}, \text{ where}$$

- Min = the **lowest** value in the data set, and
- Max = the **highest** value.

• As with computations for the mean, make sure to **group the numerator** before dividing by 2.

In Example 1, the test scores (in points) were:

80 76 100 83 100

$$\begin{aligned}\text{Midrange} &= \frac{\text{Min} + \text{Max}}{2} \\ &= \frac{76 + 100}{2} \\ &= 88 \text{ points}\end{aligned}$$

PART B: NOTATION

Let N = the **population size** for a population data set.

Let n = the **sample size** for a sample data set.

We can label data values x_i, where $1 \le i \le N$ for population data, or $1 \le i \le n$ for sample data. The "i" is called a subscript.

In Example 1, we had a population data set of test scores (in points).
Its population size was: $N = 5$.

x_1	x_2	x_3	x_4	x_5
80	76	100	83	100

Summation Notation

The Greek uppercase sigma, $\sum$, is a summation operator.
$\sum x$ denotes the **sum** of the given data values.

More precisely:

$\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$ denotes the sum of the values in a **population** data set, and

$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$ denotes the sum in a **sample** data set.

In Example 1: $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$.

Notations for Means

The notation we use depends on whether we're dealing with a **population** data set or a **sample** data set.

Population (Size N)

↓

Sample (Size n)

Population Mean

• This is denoted by the Greek letter mu, μ.

• We typically use Greek letters to denote population parameters, values that describe a **population**.

$$\mu = \frac{\sum x}{N}$$

Sample Mean

• This is denoted by x-bar, $\bar{x}$.

• We typically use Roman / English letters to denote sample statistics, values that describe a **sample**.

$$\bar{x} = \frac{\sum x}{n}$$

Many calculators have an $\bar{x}$ button that allows you to compute the mean of inputed data.

PART C: OUTLIERS

Example 3 (The Effect of Outliers)

The five students in a class take a test. Their test scores in points are:

70 80 90 90 100

Median = 90 points
Mean = 86.0 points
Midrange = 85.0 points

Let's say the person who received the "70" actually cheated, and we drop that score down to 0 points. What effect does that have on the median and the mean?

0 80 90 90 100

Median = 90 points
Mean = 72.0 points
Midrange = 50.0 points

Outliers

The "0" is an outlier, because it is **extremely low** relative to the other data values. An outlier can also be **extremely high**.

The **mean** is sensitive to outliers; that is, the mean can be greatly affected by outliers. In this Example, it drops dramatically when the "70" is changed to a "0."

The **median** is **not** sensitive to outliers, and it remains unchanged in this Example.

The **midrange** is **very** sensitive to outliers; it plunges from 85.0 points down to 50.0 points in this Example.

§

WARNING 4: Omitting outliers. Although it is tempting to simply omit outliers, we're generally not supposed to. If you do omit an outlier, you better **say so** … and give a **good reason** why! You don't want to be accused of **tampering** with your data. Sometimes, we give results both **before and after** outliers are removed.

PART D: TRIMMED MEANS

Trimmed means can be used to **reduce the effect of outliers**.

• For example, a 10% trimmed mean is obtained by taking the average of the data set **after** the bottom 10% and the top 10% of the sorted data values have been deleted.

• For a 20% trimmed mean, the bottom 20% and the top 20% are deleted.

Example 4 (Trimmed Means)

To obtain a **20% trimmed mean** for both data sets in Example 3, we delete the lowest and highest scores: "70" and "100" in the first data set … and "0" and "100" in the second data set.

In both cases, we are left with these test scores (in points), which we now average:

80 90 90

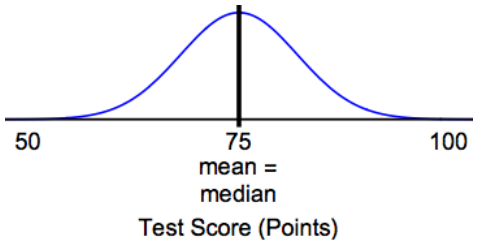
The **20% trimmed mean** in both cases is about 86.7 points.

§

PART E: COMPARING MEAN vs. MEDIAN: SYMMETRY, SKEWNESS, and OUTLIERS

Example 5 (Symmetric Distributions)

Here is a smoothed-out histogram of test scores near the beginning of a term:

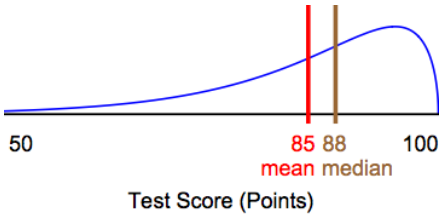


The distribution is **perfectly symmetric**, so the mean is **equal to** the median.

- If a distribution is **approximately symmetric**, and if there are **no outliers**, then the mean should be **close to** the median.
- **Outliers could change that!** §

Example 6 (The Effect of Skewness)

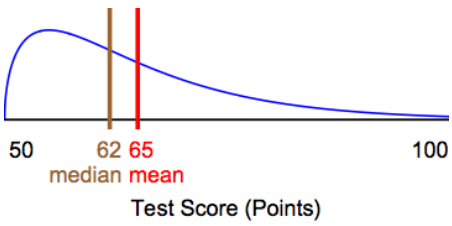
Here is a smoothed-out histogram of test scores near the beginning of a term:



- The **left-skewness** of the distribution is pulling the mean **lower** than the median. The very **low** values pull the **mean** to the **left**.

  Think About It: What typically happens to exam means and medians as the term progresses and students drop out of the class?

- For a **right-skewed** distribution, the mean would be pulled **higher** than the median. The very **high** values pull the **mean** to the **right**. See below.



§

• The **mean** may be better than the median for roughly **symmetric** distributions, especially when there are **few (if any) outliers**. The mean is more often used in **formulas** than the median, mode, and midrange are.

• The **median** may be more appropriate than the mean as a measure of center for **skewed** distributions. The presence of **outliers** would also tend to make us favor the median.

FOOTNOTES (OPTIONAL)

#1) Notation. The median is sometimes denoted by $\tilde{x}$. The mode is sometimes denoted by $\hat{x}$.