THE OHIO STATE UNIVERSITY

COLLEGE OF ENGINEERING

Data Analysis and Excel

Class 7

Today's Learning Objectives

After this class, students will be able to:

- Define accuracy, precision, and resolution.
- Define systematic variation and random variation.
- Define the terms mean, median, mode, central tendency, and standard deviation.
- Analyze data using mean, median, and standard deviation.

Review of Important Terms

- "Clicker activity"

- When measuring the length of an object, do accuracy and precision mean the same thing?
  1. Yes
  2. No

Review of Important Terms

- What is accuracy?
  - The degree of closeness between a measured value and the correct value. (implies there is a correct value!)

- Precision is based on repeatability and resolution

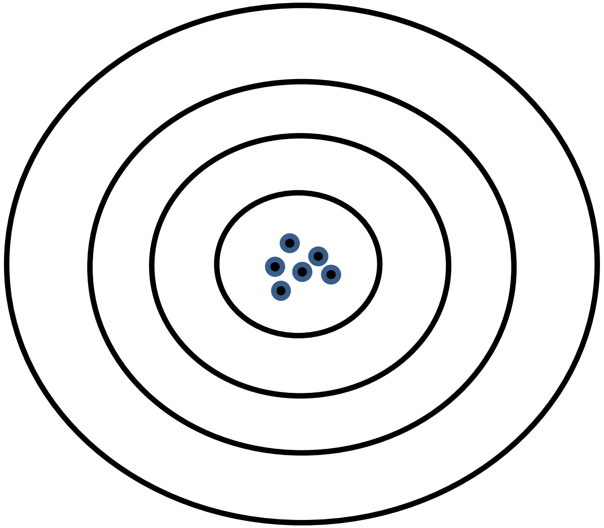
Review of Important Terms

- What is repeatability?
  - The degree of closeness between repeated measurements of the same feature.
- What is resolution?
  - The smallest measurable increment on a measurement device.

Review of Important Terms

- “Clicker activity”

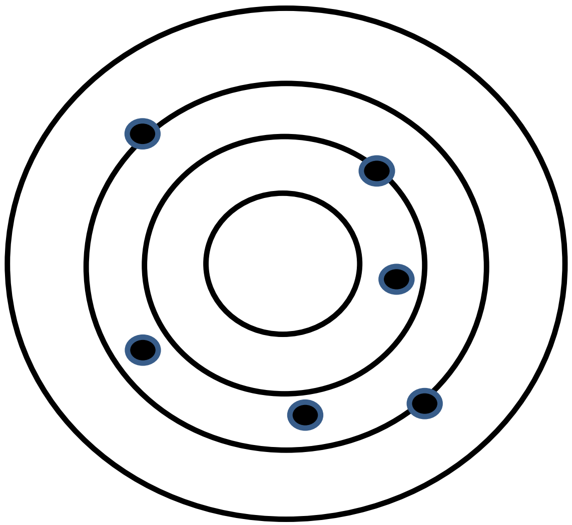
- Is this:
  1. Repeatable and accurate?
  2. Accurate, but not repeatable?
  3. Repeatable, but not accurate
  4. Neither accurate nor repeatable



Review of Important Terms

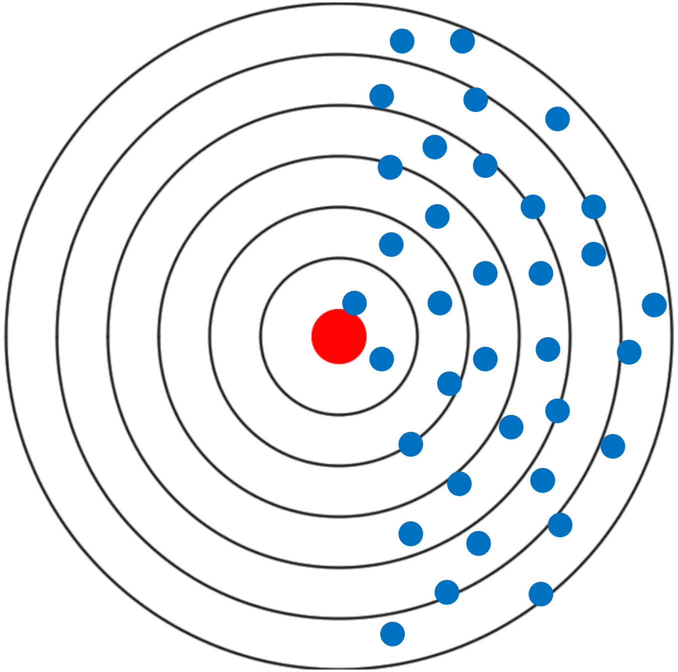
- "Clicker activity"

- Is this:
  1. Repeatable and accurate?
  2. Accurate, but not repeatable?
  3. Repeatable, but not accurate
  4. Neither accurate nor repeatable



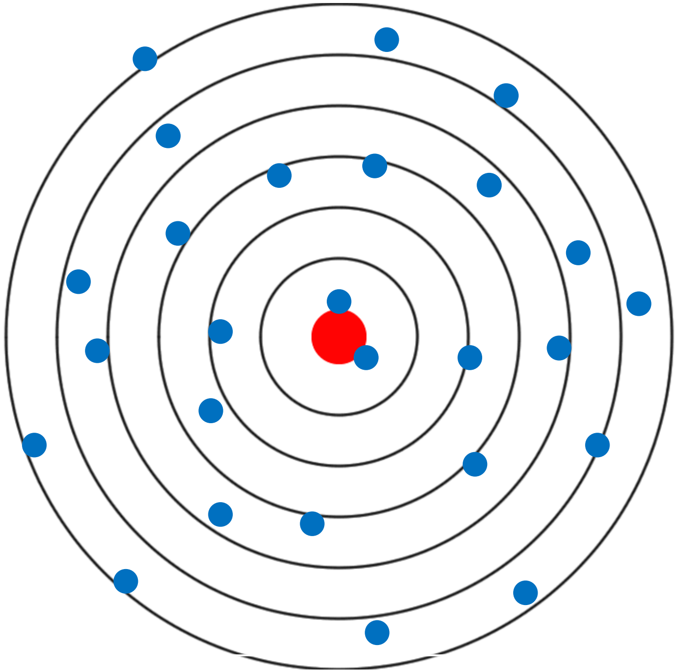
Review of Important Terms

- What is an example of a Systematic Error?



Review of Important Terms

- What is an example of a Random Error?



Example: Collecting and Analyzing Data

- The next few slides will walk through an example regarding collecting data and then analyzing the data.

Sample Data

- 8 sample measurements were collected using two different instruments
  - Scale
  - Caliper

Sample Data

- 8 sample measurements were collected using two different instruments
- The expected value is 1.020 inches

- Which instrument has a higher resolution?
- Which measurement offer good accuracy?
- Which measurements offer good repeatability?
  - 1. Scale
  - 2. Caliper

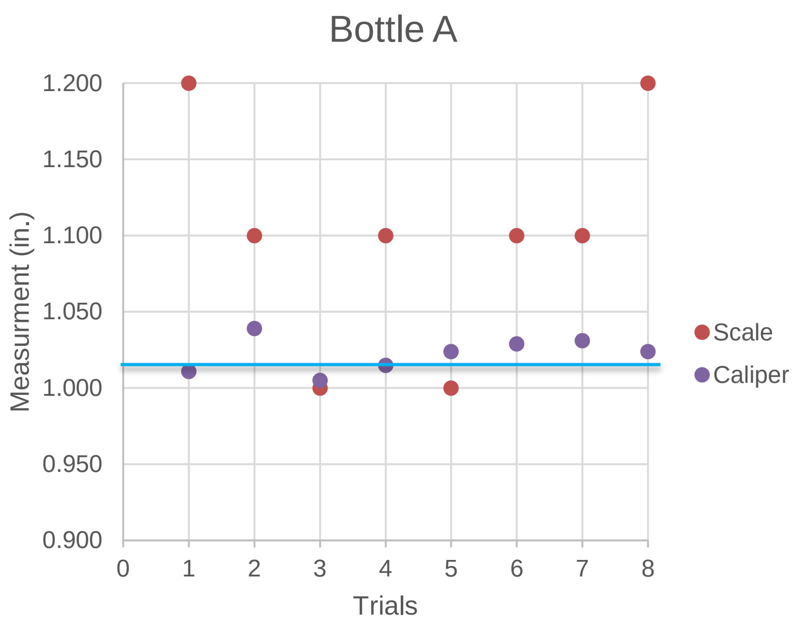
Sample Diameter Measurements, Shampoo Bottle A

Trial	Scale	Caliper
1	1.2	1.011
2	1.1	1.039
3	1.0	1.005
4	1.1	1.015
5	1.0	1.024
6	1.1	1.029
7	1.1	1.031
8	1.2	1.024

Sample Data

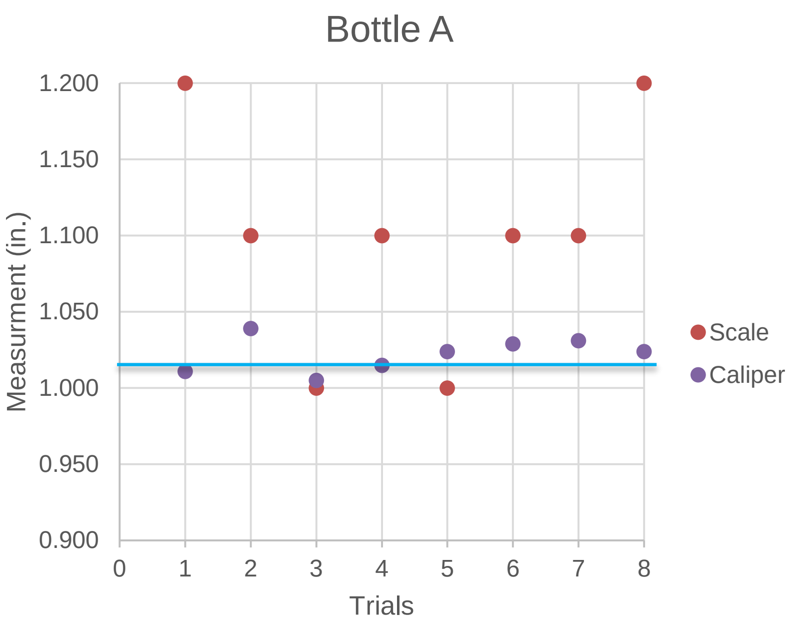
- Notice how the measurements fall above and below the specified value of 1.02”

- Which measurement tool provided more accurate values?
  - 1. Scale
  - 2. Caliper



Outliers

- Outliers are values that differ greatly from the specified value.
- Do we have any outliers here?
- Outliers can occur from systematic or random variation, or they can be a misrepresented value.
  - The important thing is to figure out why the values are so different.



Propagation of Error

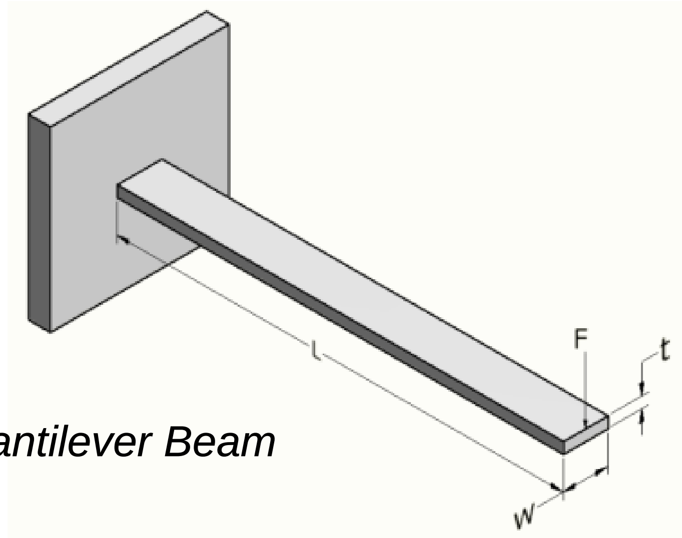
- The errors from inaccurate measurements can be magnified when the erroneous value is used in an equation.
- What seems like a small error in the initial measurement can propagate to a large overall error in the calculation of an equation.

Example: Propagation of Error

- To calculate the moment of inertia of a cantilever beam, we must measure thickness and width of the beam.

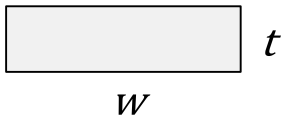
- Moment of inertia is calculated by:
  $I = \frac{1}{12}wt^3$

- What if there is an error measuring w?
- What about an error measuring t?

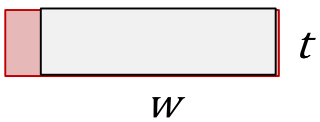


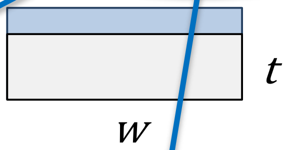
Cantilever Beam

Example



Base Case: $I=1/12wt^3$ w=3.421 in t=2.129 in $I_1=\underline{2.751\ in^4}$





Assume 9.5% error in *w*

w=3.096 in

$I=1/12(3.096)(2.129)^3$

$I_2=\underline{2.490\ in^4}$

(difference of 0.261 in^4)

➔ 9.5% error

Assume 9.5% error in *t*

t=1.927 in

$I=1/12(3.421)(1.927)^3$

$I_3=\underline{2.040\ in^4}$

(difference of 0.711 in^4)

➔ 25.8% error

Analyzing Measurement Data

- When collecting data, there will always be variation. We can use statistical tools to help us determine:
  - Is the variation (or error) systematic or random?
  - What is the cause of the variation?
  - Is the variation in an acceptable range?
  - What is an acceptable range of variation for this data?

Example: Slingshot Experiment!

- An engineer is performing a data collection experiment using a slingshot and a softball.
  - She predicted that if the slingshot is pulled back by 1 meter before launching the ball, the softball would land 17 meters downrange.
- Data is collected from 20 trials. Let's analyze the data and see how the experiment went…

Example: The Data

- Most of this data falls in the range of 14-20 meters.
- Do you see any data that appears much outside this range?
  - These rogue data points are called outliers.

Table 1. Results of 20 trial launches with slingshot spring pulled back 1 m

Trials 1-5:	17.5	45.2	16.4	19.3	16.6
Trials 6-10:	16.0	17.4	16.7	18.1	17.5
Trials 11-15:	15.1	14.2	17.4	15.7	17.8
Trials 16-20:	19.3	18.5	15.7	17.9	17.0

Outliers: Deal With It.

- Outliers will happen even in good data sets. Engineers know how to deal with them!

- Engineers must determine whether an outlier is a valid data point, or if it is an error and thus invalid.

- Invalid data points can be the result of measurement errors or of incorrectly recording the data.

Characterizing The Data

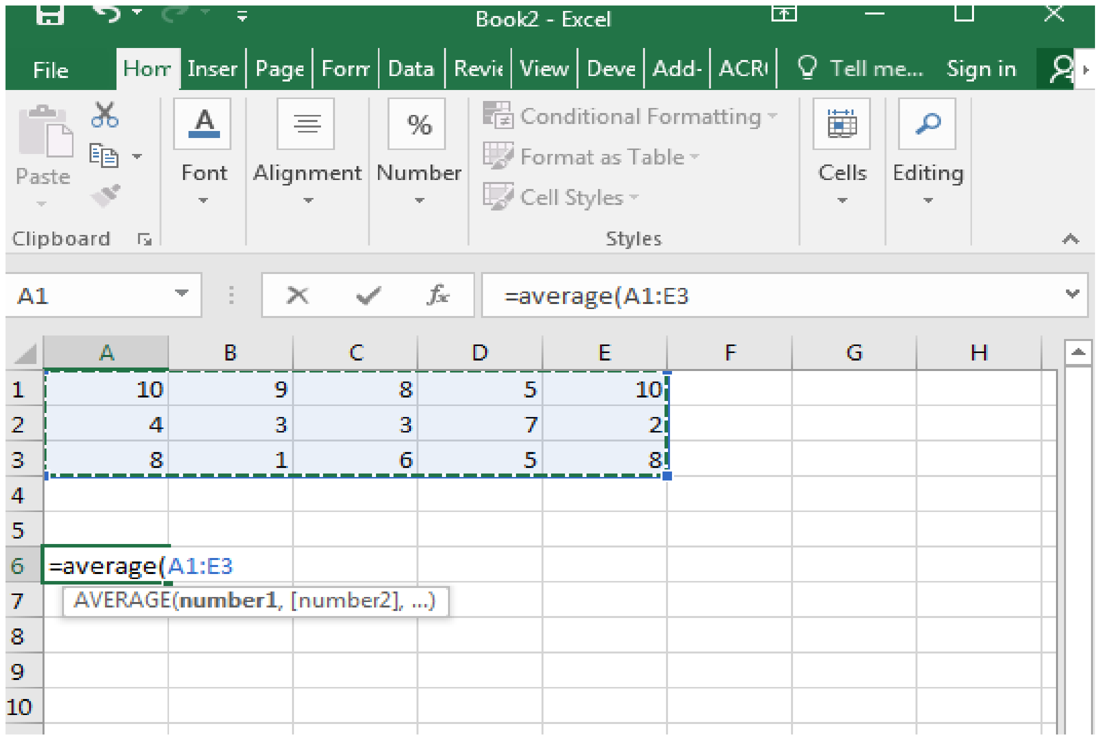
- Statistics allows us to characterize the data numerically as well as graphically.

- We characterize data numerically in two ways:
  - Central Tendency
  - Variation

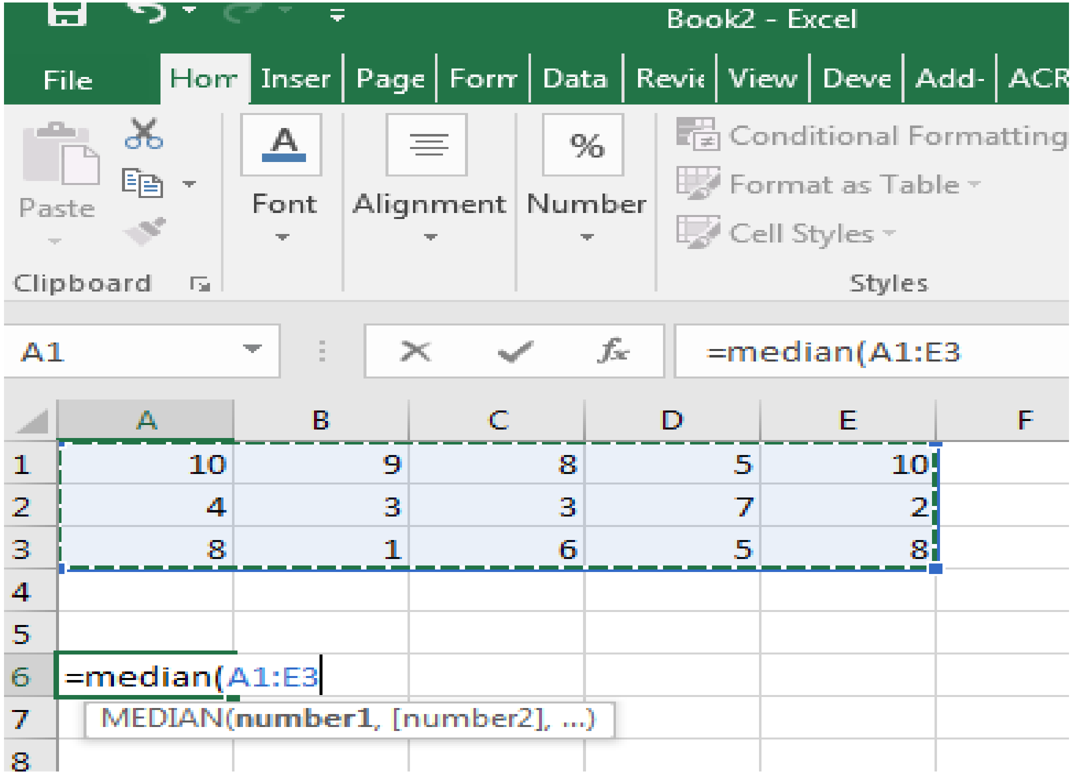
Central Tendency

- This is a single value that best represents the data. This value could be determined by:
  - Mean
  - Median
  - Mode
- For many engineering applications, the mean and median are most relevant.

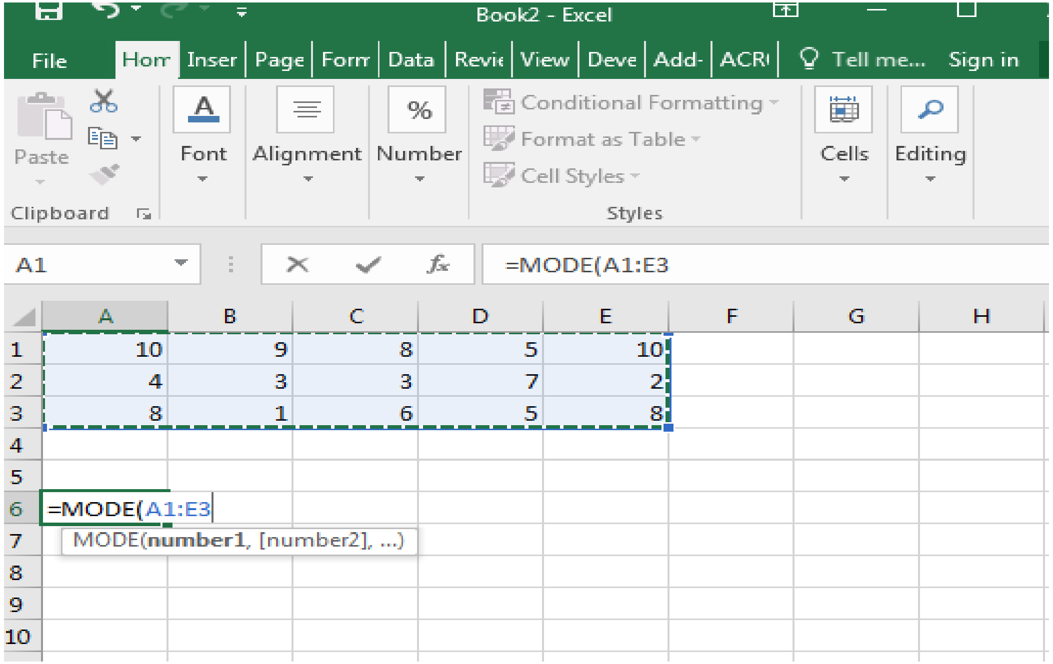
Calculate Mean in Excel



Calculate Median in Excel

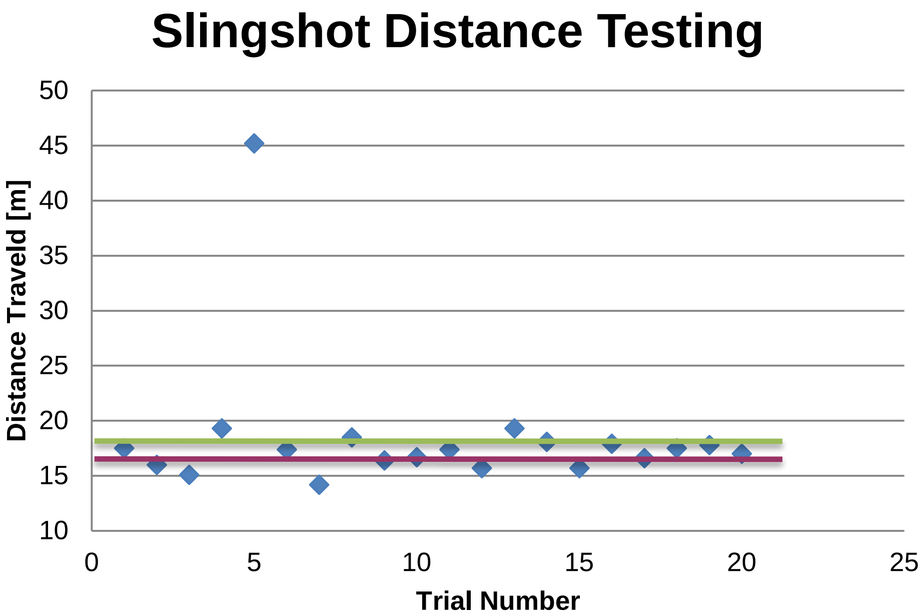


Calculate Mode in Excel



Central Tendency: Comparison

- Which value is a better representation of the slingshot data?
  - Mean = 18.47 m
  - Median = 17.4 m



Characterizing the Data

- We can select a value of central tendency to represent the data, but is just one number enough?

- No! It is also important to know how much variation is present in the data.

- Variation describes how the data is distributed around the central tendency value.

Representing Variation

- As with central tendency, there are multiple ways to represent variation in a set of data:
  - ± (“Plus/Minus”) gives the range of values
  - Standard Deviation provides a more sophisticated look at how the data is distributed around the central value.

Standard Variation

- Definition:
  How closely the values cluster around the mean; how much variation there is in the data.

- Equation:

Mean = Average in Excel

$$\sigma = \sqrt{\sum \frac{(each\ value - mean)^2}{(number\ of\ values - 1)}} = \sqrt{\sum \frac{(x - x)^2}{(n-1)}}$$

OR

$$= \sqrt{\frac{\sum (x - \bar{x})^2}{(N - 1)}}$$

Calculating Standard Deviation

x (each value)	x - mean	(x-mean)^2/(n-1)
14.2	-4.3	0.97
15.1	-3.4	0.61
15.7	-2.8	0.41
15.7	-2.8	0.41
16.0	-2.5	0.33
16.4	-2.1	0.23
16.6	-1.9	0.19
16.7	-1.8	0.17
17.0	-1.5	0.12
17.4	-1.1	0.06
17.4	-1.1	0.06
17.5	-1.0	0.05
17.5	-1.0	0.05
17.8	-0.7	0.03
17.9	-0.6	0.02
18.1	-0.4	0.01
18.5	0.0	0.00
19.3	0.8	0.03
19.3	0.8	0.03
45.2	26.7	37.52
18.5		**41.32**

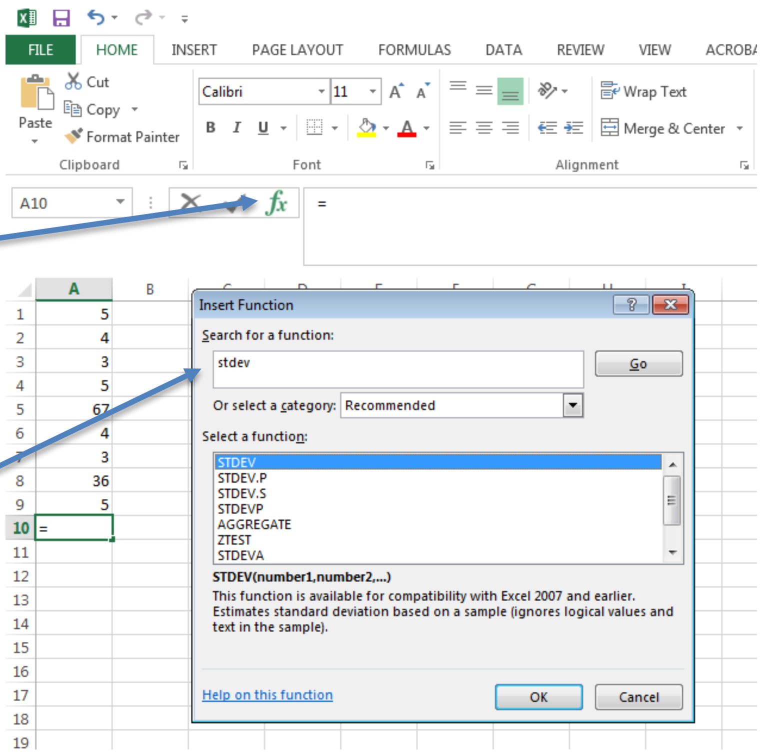
$$\sigma = \sqrt{\sum \frac{(each\ value - mean)^2}{(number\ of\ values - 1)}}$$

$$\sigma = \sqrt{41.32}$$

$$\sigma = 6.4281$$

Calculating Standard Deviation in Excel

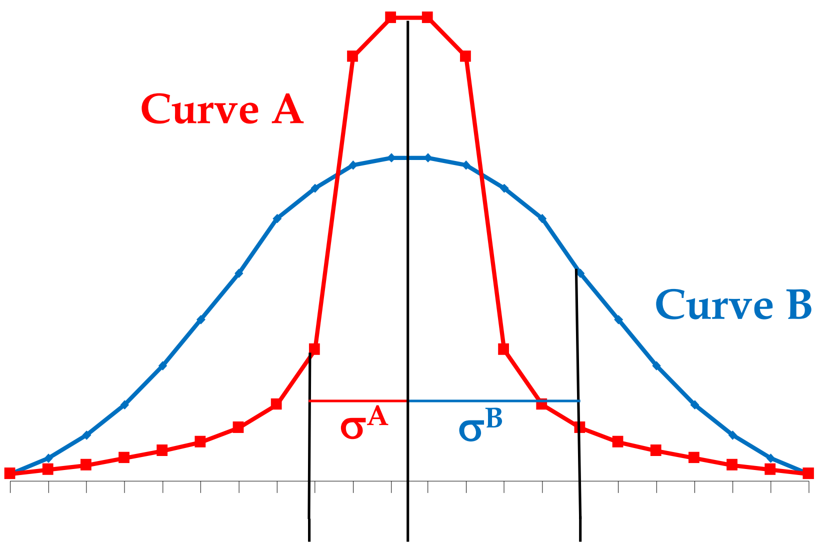
- Put the cursor where you want the standard deviation
- Click on the Function Wizard (fx) button
- Type in the function name (use stdev or stdev.s in this class)



Interpreting Standard Deviation

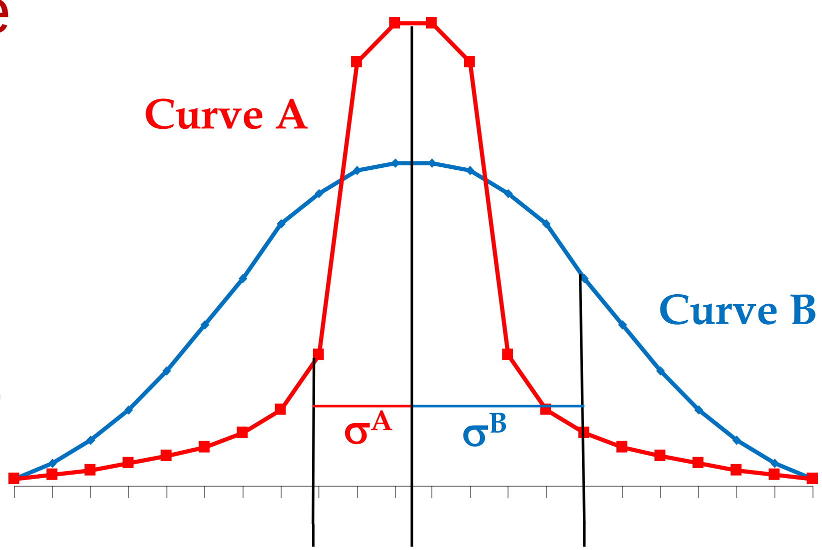
- Curve A has a small σ. Data points are clustered close to the mean.

- Curve B has a large σ. Data points are far from the mean.



What do YOU think?

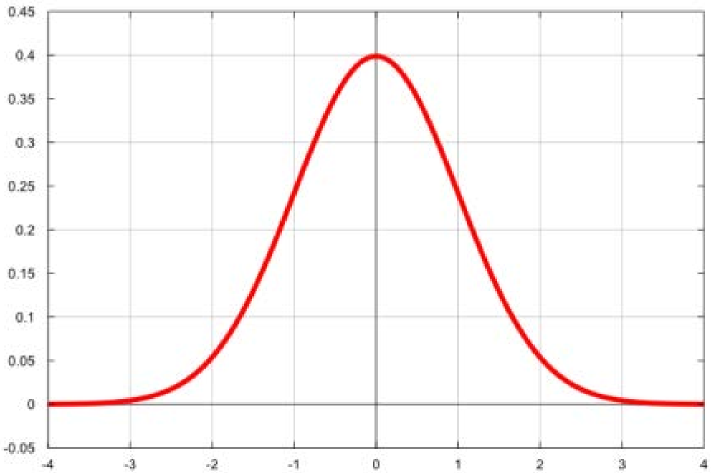
- Say these curves describe the distribution of grades from an exam, with an average score of 83%...
- What would be the benefit if you were in class with Curve A?
- What about Curve B?



Normal Distribution

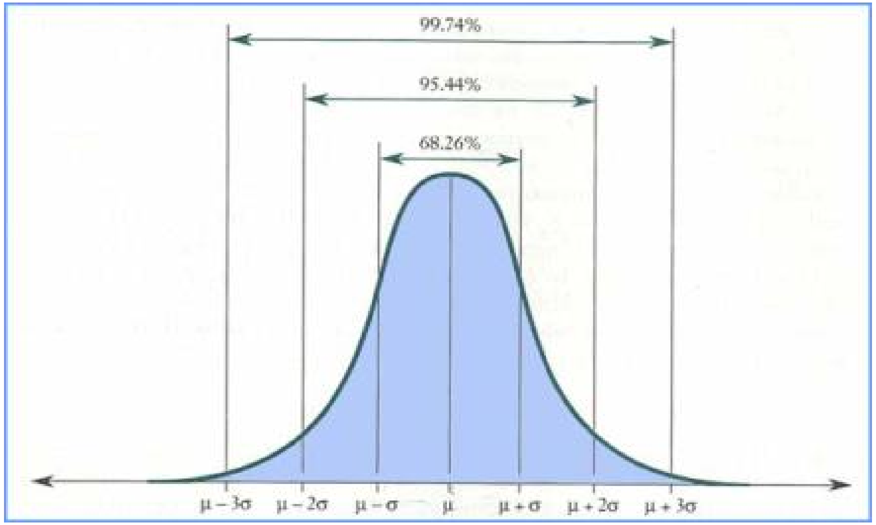
- Data that is normally distributed occurs with greatest frequency around the mean.

- Normal distributions are also known as Gaussian Distributions or Bell Curves

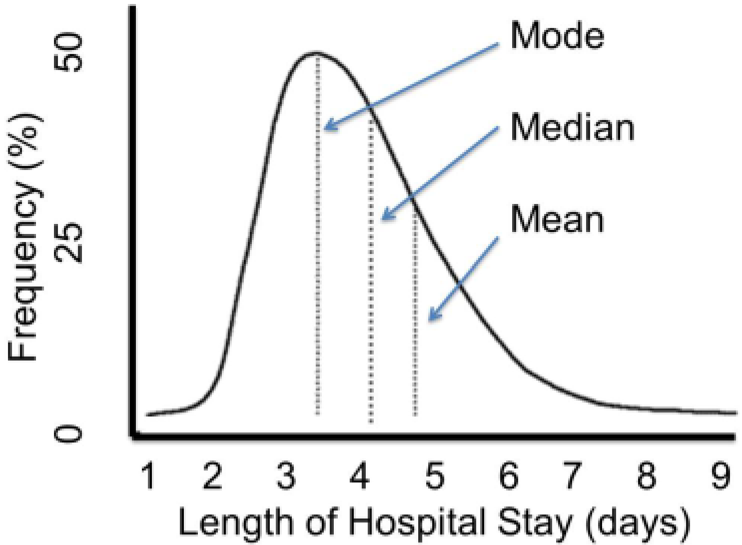


Normal Distribution

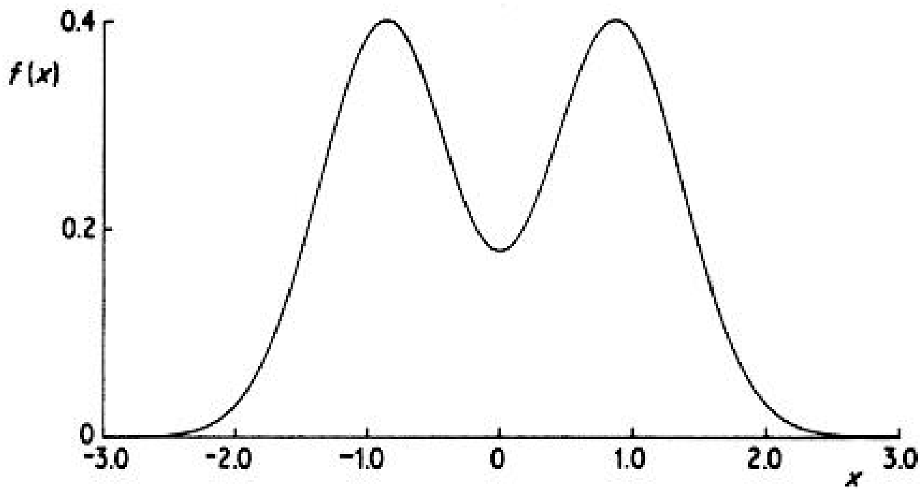
- Mean = Median = Mode
  - 68% of values within 1 σ
  - 95% of values within 2 σ
  - 99.99966% of values are within 6 σ



Other Distributions



Skewed Distribution

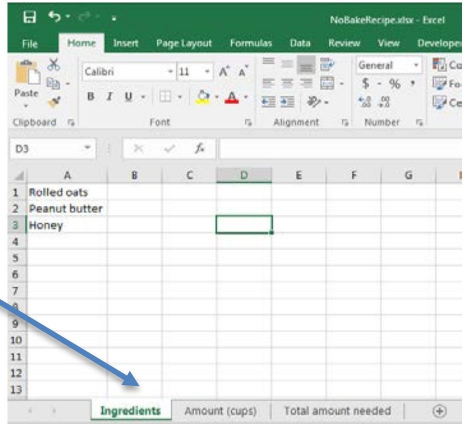


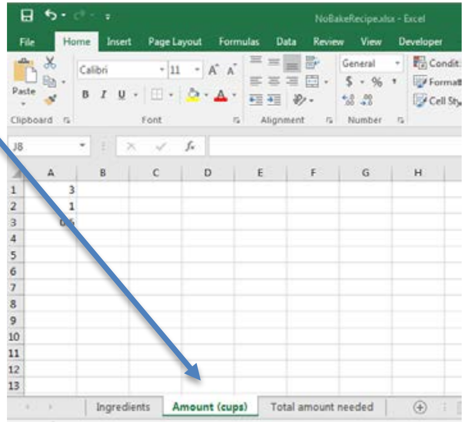
Bimodal Distribution

Referencing Other Worksheets

- For a no-bake peanut butter oat squares recipe, ingredients are stored in one worksheet, titled “Ingredients”. Amounts of each ingredient needed are listed in a different worksheet, titled “Amount (cups)”.

- How could we calculate the total volume of ingredients needed in a third worksheet?





Referencing Other Worksheets

- ='Amount (cups)'!A1 + 'Amount (cups)'!A2 + 'Amount (cups)'!A3