Lesson 2
Part 1
Summarizing Numerical Data with Summary Statistics

Course Material Organization

- The concepts covered in this course are organized by Measurement Scale of the data
- This first section of the course covers methods of summarizing data
  - Lesson 2 – summarizing numerical data
    - Part 1: Summary statistics
    - Part 2: Tables and Graphs
  - Lesson 3 – summarizing categorical data (ordinal and nominal)

Lesson 2 Overview

- The first step in statistical analysis is to summarize the data with summary statistics, tables and graphs
- Lesson 2 Part 1 covers summary statistics for numerical (quantitative) data
- *Descriptive Statistics or Summary Statistics: statistics used to organize and describe collected data*
  - *Pocket Dictionary of Statistics*

Lesson 2 Part 1 Outline

Summary statistics covered in Lesson 2 Part 1:

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
  - Geometric Mean
- Measures of Variability
  - Range and Interquartile Range
  - Variance and Standard Deviation
  - Coefficient of Variation

Some Mathematical Notation

- Datasets will be identified as numbers inside { }. For example here is a dataset with 4 observations: {8,2,3,5}
- The numbers in a dataset are denoted as x_i's

  $x_1=8$ $x_2=2$ $x_3=3$ $x_4=5$

- The Greek letter Σ (sigma) is the summation sign
  - Sum= $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 8+2+3+5=18$
- The sample size of a dataset is denoted by n or N. For the dataset {8,2,3,5}, n = 4

.

Measures of Central Tendency

- Measures of Central Tendency identify the center of the distribution of observations.
- The most commonly used measure of central tendency for numerical data is the Mean.
- The mean is also called the sample average or the arithmetic mean
- Mathematical notation for the mean: $\overline{X}$

Sample Mean: $\overline{X}$

- **Formula**

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- The mean is the sum of all the observations divided by n (the number of observations).
- For the dataset {8, 2, 3, 5}

  n = 4, $\sum$= 18, mean = 18/4 = 4.5

Outliers and the mean

- We asked two groups of 5th graders 'How many hours a week do you play video or computer games?'

  Sample 1: { 4, 1, 3, 8, 10, 9, 7 }. What is the mean?

  Mean = 6

  Sample 2: { 4, 1, 3, 8, 10, 37, 7 } What is the mean?

  Mean = 10

- Sample 2 has an extreme value of 37. Extreme values in a dataset are called 'outliers'.
- Outliers can have a large influence on the mean, especially if the dataset is small.

Outliers

Strategies for dealing with outliers

- Is the outlier an error in the dataset? If so, correct the error
  - There might be a measurement error
  - There might be a recording error
- If the outlier isn't the result of measurement or recording error, it is a valid observation and must be included in the data
- When the mean is unduly* influenced by an outlier, the median might be a better measure of central tendency

*The mean is unduly influenced if it no longer represents the center of most of the data – look at the mean for sample 2 in the previous slide. The mean of 10 doesn't represent the center of the data because of the influence of the outlier.

Median

- The Median is the middle observation that divides the dataset into equal halves.
- Finding the median
  - Data need to be arranged in order
  - If N is odd, the median is the middle value
  - If N is even, the median is the average of the 2 middle values
- The median is an appropriate measure of Central Tendency for data with outliers because it is not influenced by a small number of outliers.

Finding the Median

- We asked two groups of 5th graders 'How many hours a week do you play video or computer games?'

  Sample 1: { 4, 1, 3, 8, 10, 9, 7 }

  Sample 1 ordered: {1, 3, 4, 7, 8, 9, 10}

  The median is the middle value = 7

  Sample 2: { 4, 1, 3, 8, 10, 37, 7 }

  Sample 2 ordered: {1, 3, 4, 7, 8, 10, 37}

  The median is the middle value = 7

- The middle value is not changed by the outlier.

Another use of the Median

- Sometimes ordinal data are coded with numbers
- For example, responses to a survey question might be coded as
  - ‘never’ = 0
  - ‘sometimes’ = 1
  - ‘always’ = 2
- These arbitrarily assigned numerical codes represent the order of the responses but do not represent true numerical data.
- The median is an appropriate summary statistic for numerically coded ordinal data.
  - Mathematical functions such as addition, subtraction, multiplication and division are not appropriate for ordinal data.

Mode

Another measure of central tendency is the mode

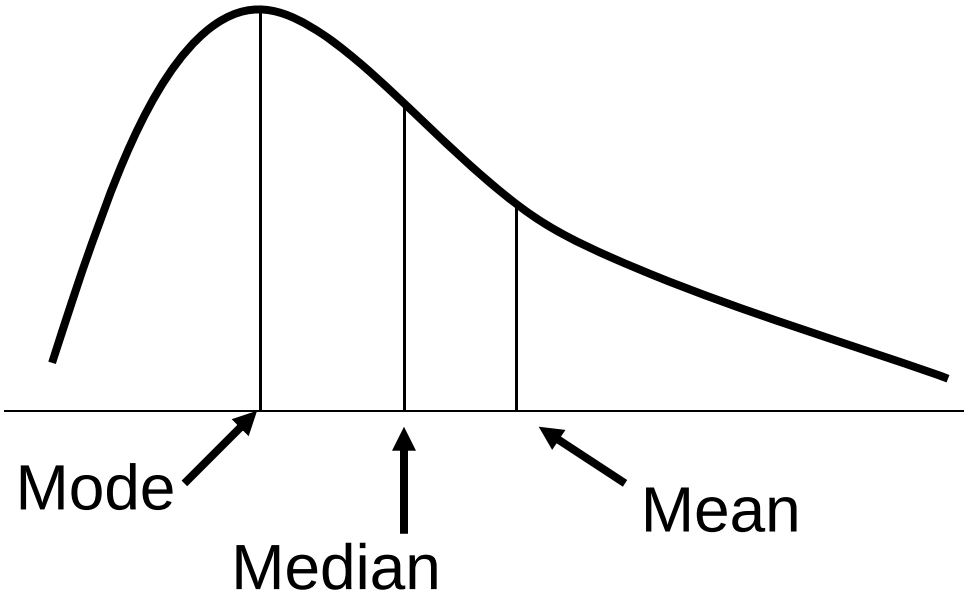
- The Mode is defined as the most commonly occurring value in a dataset
- Unimodal distribution: one mode in the dataset
  - {3, 5, 7, 7, 7, 10, 12, 14} - the mode is 7
- Bimodal distribution: two modes in the dataset
  - {1, 2, 3, 3, 7, 10, 10} - both 3 and 10 are modes.
- Not all datasets have a mode
  - {3, 5, 6, 8, 9, 12, 15} - no mode.

Mean, Median, and Mode

- The mean, median and mode identify unique aspects of the distribution, or ‘shape’ of the observations
- The relationship between these measures of central tendency can be used to identify whether the data are symmetric or skewed
- If the data have a few outliers in one direction, the distribution of the data is ‘skewed’
  - A few large extreme values: positively skewed
  - A few small extreme values: negatively skewed

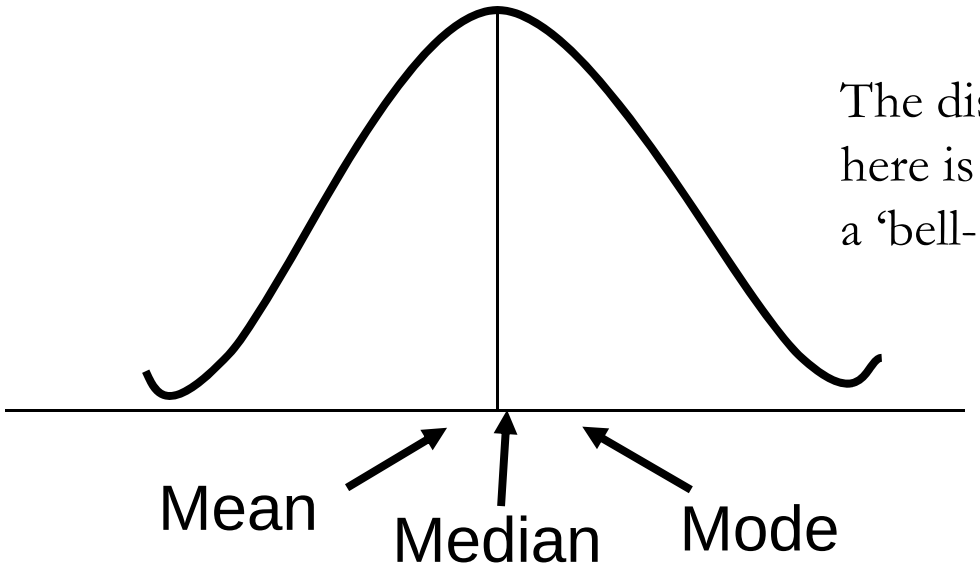
Distribution Characteristics

- Mode: Identifies the Peak(s) of the data
- Median: Identifies the equal areas point
- Mean: Identifies the balancing point



Symmetric Distribution

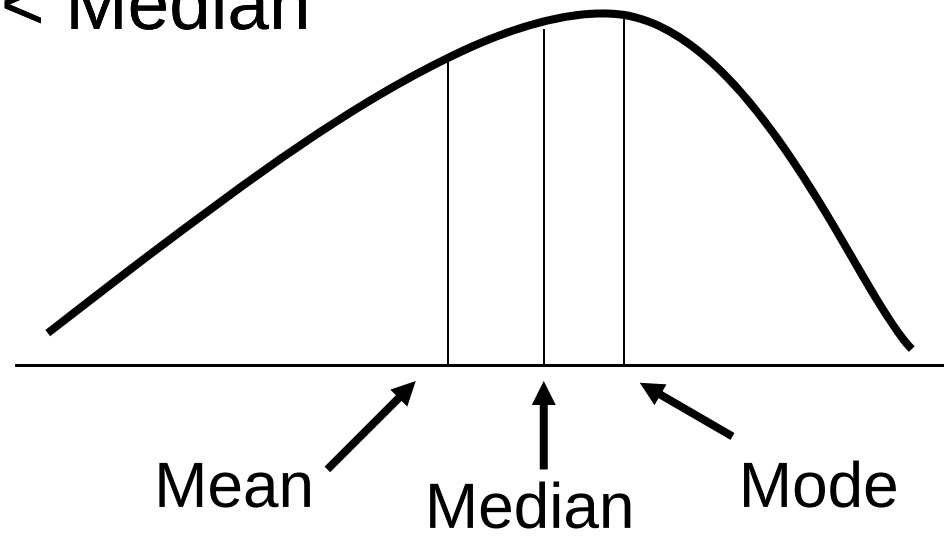
- Right and left sides are mirror images in a symmetric distribution
  - Left tail and right tail are identical
  - Mean = Median = Mode



The distribution illustrated here is sometimes called a 'bell-shaped' distribution

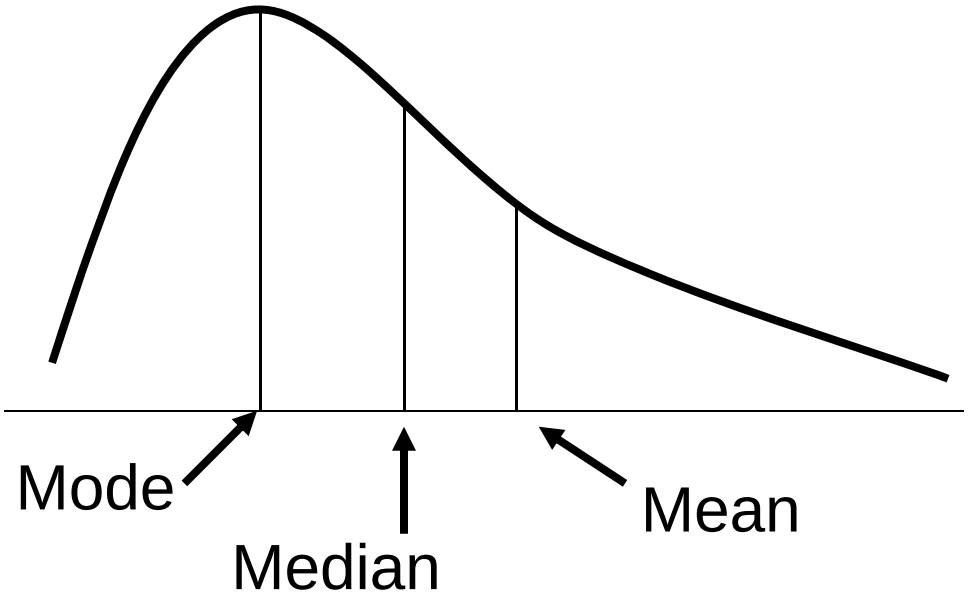
Negatively skewed Distribution

- Negatively skewed is also called ‘left skewed’.
  - Long left tail
  - Outlier(s) are small values relative to the data
  - Mean < Median



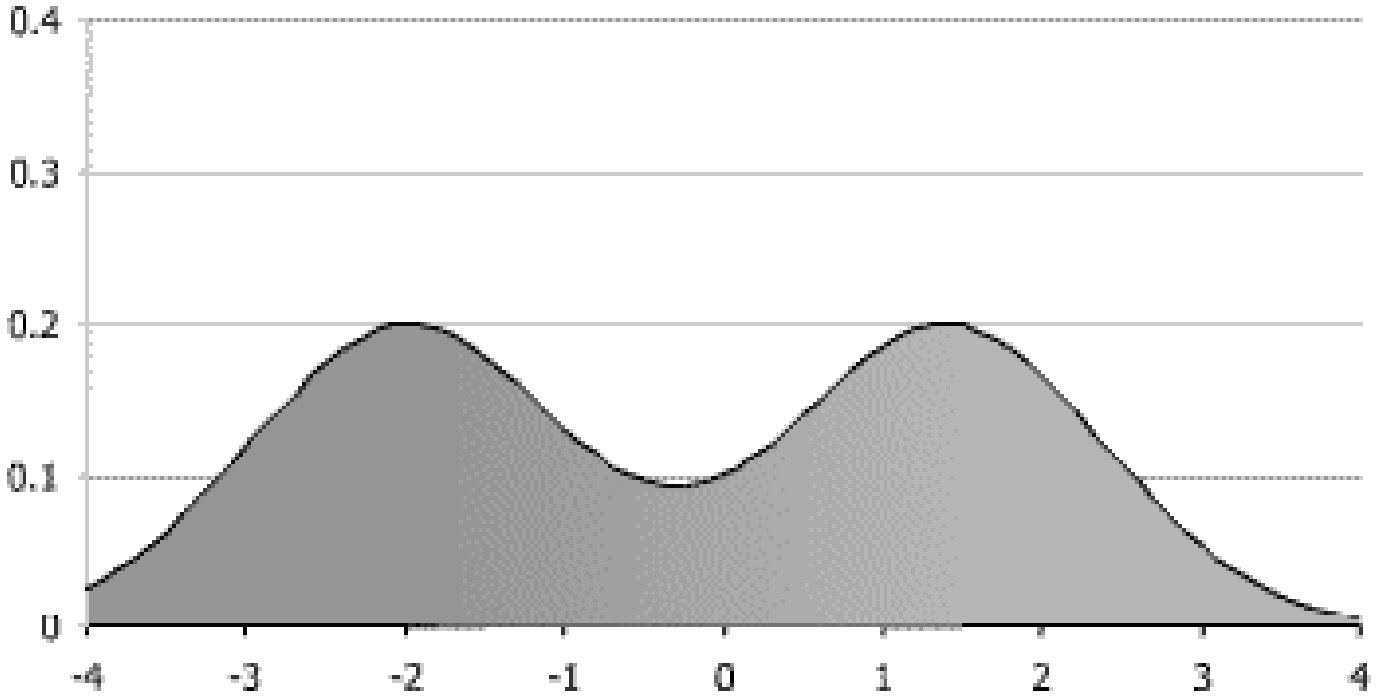
Positively skewed Distribution

- Positively skewed is also called right skewed
  - Long right tail
  - Outlier(s) are large values relative to the data
  - Mean > Median



Bimodal Distribution

there are two peaks in a bimodal distribution



- Source: Wikipedia, bimodal distribution

Summary Statistics for Skewed data

- Skewed data have outliers so the median is a better measure of the center than the mean for skewed data
- Another approach for positively skewed data is to summarize the data on the log scale
  - The log transformation minimizes the effect of extreme observations.
  - If the natural log transformation is used, the center of the data can be summarized with the geometric mean

Geometric Mean

- Method of obtaining the geometric mean:
  - Take the natural log (base e) of each data value
    - The notation for natural log is ln
  - Calculate the mean on the ln scale
  - Take the antilog of the mean to return to the original scale of measurement.
  - This is called the “GEOMETRIC” Mean.

Geometric Mean Example

x	ln(x)
8	2.08
5	1.61
4	1.39
12	2.48
15	2.71
7	1.95
28	3.33
79	15.55

The mean using the raw data is :

$$\bar{x} = \frac{79}{7} = 11.3$$

While on the log scale :

$$\frac{\sum \ln x}{n} = \frac{15.5}{7} = 2.22$$

Now take the antilog of 2.22 :

$e^{2.22} = 9.22$

The geometric mean $= 9.22$

Geometric Mean Applications

- The Geometric mean is an appropriate measure of central tendency for positively skewed data. Some examples are:
  - Exponential data such as bacterial concentrations
  - Dilution data such as antibody concentrations
- Limitations to using Geometric mean
  - Cannot be calculated if any data are negative
  - Calculating the geometric mean is problematic if there are many ‘0’ values

Which Measure of Central Tendency should you report?

- Mean
  - For symmetric (not skewed) numerical data
- Median
  - For skewed numerical data
  - For ordinal data that have been coded numerically
- Mode
  - Report the modes if the data are bimodal
- Geometric Mean
  - For positive data measured on a logarithmic scale
  - For positively skewed data

Measures of Variability

- Measures of variability are used along with measures of central tendency to more completely describe the data.
- Measures of variability (also called measures of dispersion) describe the spread of the data
- Measures of Variability:
  - Range
  - Quartiles and Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Range

- Range = largest value (maximum) minus the smallest value (minimum)
- Pros: Easy to calculate
- Cons: The value of the range is only determined by two values and provides no information about values between the Minimum and the Maximum

  Sample 1: {20,28,30,30,31,38,40,42,48,50,51,100}

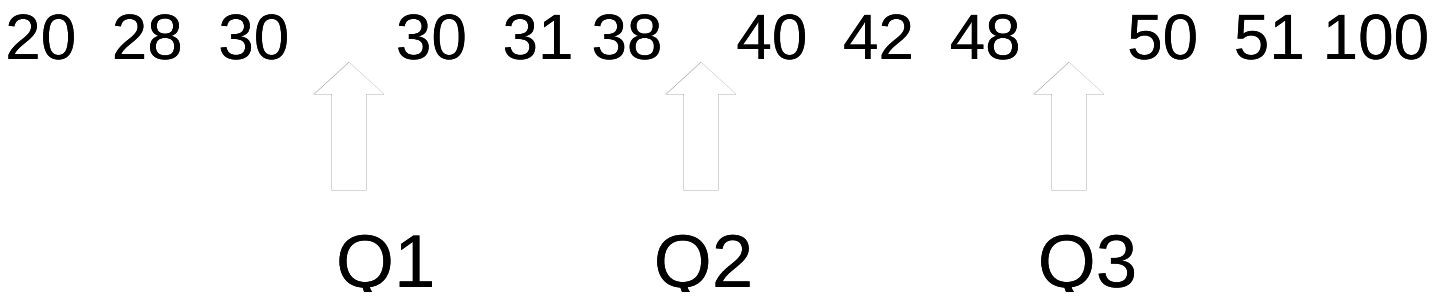
  Sample 2: {20,32,41,53,63,68,71,82,89,94, 100 }

  Range for both data sets = 100 – 20 = 80

  However, the data distributions are not very similar.
  Sample 1 has an outlier and Sample 2 is more evenly distributed across the range of the data

Quartiles

Quartiles divide the data into 4 equal parts



Q1 = ½ (30 + 30) = 30

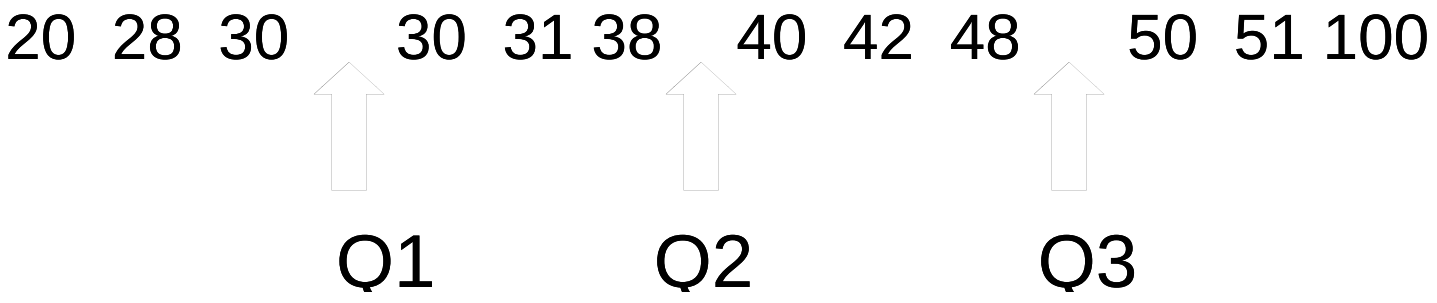
Q2 = ½ (38 + 40) = 39

Q3 = ½ (48 + 50) = 49

Each quartile is the average of the values on either side of the division when the data divide equally into 4 parts

Interquartile Range (IQR)

The Interquartile Range = Q3 – Q1



Q1 = ½ (30 + 30) = 30

Q3 = ½ (48 + 50) = 49

IQR = 49 – 30 = 19

The IQR is the range of the middle 50% of the data

Variance: A measure of Dispersion

- Each observation has a deviation from the mean which is the difference between the observation and the mean.
- The variance (s^2) is obtained by squaring these deviations and dividing the sum of squared deviations by n-1 where n = the sample size.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

The variance is measured in the square of the units in which the x's were measured.

Calculating the Variance

- Consider the dataset {8,5,4,12,15,5,7}
  - Calculate the mean
  - Subtract the mean from each x
  - Sum the squared deviations and divide by n-1

$$\bar{x} = 8$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 100$$

$$s^2 = 100/6 = 16.67$$

x	$x_i - \bar{x}$ deviations	$(x_i - \bar{x})^2$ Squared deviations
8	0	0
5	-3	9
4	-4	16
12	4	16
15	7	49
5	-3	9
7	-1	1

Deviations from the Mean

- The difference between each observation and the mean is called the deviation from the mean.
- The sum of the deviations from the mean for any dataset always = 0.
- This is why the squared deviations are summed as a measure of variability.
- Because squared deviations are summed, the variance is always a positive value

Why do we use n-1 instead of n for the denominator?

- The denominator for the variance calculation is n-1 because the variance has n-1 degrees of freedom.
- Degrees of freedom for a statistic are the number of observations that are “free” to vary
- Since the deviations from the mean sum to 0, only n-1 of the deviations are free to vary. The last deviation can be determined if the other n-1 deviations are known.

N-1 degrees of freedom

- A dataset {10, 6, 7, 12, 15, 6, 7} and n-1 deviations from the mean are given. What is the missing deviation from the mean?

x	$x_i - \bar{x}$
10	1
6	-3
7	-2
12	3
15	6
6	-3
7	-2

The Standard Deviation

The standard deviation (SD or *s)* is the positive square root of the variance.

The SD can be thought of as the average deviation of the individual observations from the mean

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Calculating the Standard Deviation:

- Consider the dataset {1,3,6,8,10}
  - Use the data from the table to calculate s

$$\bar{x} = 5.6$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 53.2$$

$$s^2 = 53.2 / 4 = 13.3$$

$$s = \sqrt{13.3} = 3.65$$

x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-4.6	21.16
3	-2.6	6.67
6	0.4	0.16
8	2.4	5.76
10	4.4	19.36

Properties of the Standard Deviation:

- Unlike the variance, the standard deviation is in the same units as the original data
- The Standard deviation as a measure of variability is appropriate when the mean is used as the measure of central tendency.
- A Standard deviation = 0 indicates “no variation” in the data. All of the data have the same value if $s = 0$.
- Like the mean, the Standard deviation is affected by outliers

Applications of Standard Deviation

- The standard deviation will be used in confidence intervals and statistical tests of the mean.
- For any distribution of data (skewed or symmetric), *at least* 75% of the data are between the mean minus 2 SD and the mean plus 2 SD.
  - This is a result of Chebyshev's Theorem.

Which Measure of Variability Should you Report?

- If you are reporting the Mean, report the standard deviation as a measure of variability.
- If you are reporting the Median, report the Range or (Min, Max) as a measure of variability.
- The Interquartile Range is used to describe the Central 50% of the data
- If you are reporting the Geometric mean, there isn't a corresponding measure of variability calculated on the log scale.

Given 15 numbers…

39	45	59	46	47	59	47	69
37	67	47	53	79	89	99	

- How do we calculate the mean?

$$mean = \bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$= \frac{39 + 45 + \cdots + 99}{15} = 58.8$$

Given 15 numbers…

39	45	59	46	47	59	47	69
37	67	47	53	79	89	99	

- How do we calculate the median?
- Since *n* is odd, we arrange the numbers and select the middle number.

37 39 45 46 47 47 47 53

59 59 67 69 79 89 99

Median =

Given 15 numbers…

39	45	59	46	47	59	47	69
37	67	47	53	79	89	99	

- How do we calculate the mode?
- Observing the data we see...

37	39	45	46	47	47	47	53
59	59	67	69	79	89	99	

Mode =

Given 15 numbers…

39	45	59	46	47	59	47	69
37	67	47	53	79	89	99	

- We can also calculate the variance and standard deviation.

$$S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x}) = 341.5$$

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})} = 18.5$$

which measure of central tendency is best?

1. Mean
2. Median
3. Mode

What happens when we add ten to every number?

49	55	69	56	57	69	57	79
47	77	57	63	89	99	109	

What Changes?

1. The Mean
2. The Standard Deviation
3. Both
4. Neither

What stays the same?

1. The Median
2. The Variance
3. Both
4. Neither

Let’s take a look.

49	55	69	56	57	69	57	79
47	77	57	63	89	99	109	

mean=68.8 median=53.0

std. dev.=18.5 variance=341.5

Coefficient of Variation

- The coefficient of variation (CV) is a measure of the relative spread of the data
- CV = the standard deviation as a percentage of the mean:

$$CV = \frac{s}{\bar{x}} * 100$$

- CV is used to compare variability between distributions that are measured on different scales.

Comparing Variability between measures

- Data on the following slide provide the mean and SD for three different laboratory blood measures.
- Notice that the units of measurement are not the same for the different laboratory measures
- The differences in units of measurements means that direct comparisons of variability using the SD may not be valid

Data for three laboratory Measures

Variable	Mean and Standard Deviation		Coefficient of Variation (%)
Erythrocyte count (adult males, millions/mm3)	**Mean**	**5.4**	**16.67**
	SD	**0.9**	
Total serum – iron binding capacity (ug/dL)	**Mean**	**273.7**	**21.78**
	SD	**59.6**	
Hematocrit (neonates, 1-13 days, mL/dL)	**Mean**	**54.0**	**18.52**
	SD	**10.0**	

Comparing Variability using CV

- In the previous slide, the SD of total serum-iron binding capacity is 66 times greater than the SD of erythrocyte count.
- After calculating the coefficients of variation, serum iron-binding capacity has a relative variation (CV) that is only 1.3 times that of erythrocyte count.
- Comparisons of the coefficients of variation are more valid than comparisons of SD when the units of measurement are not the same

Corrections to Text

- Pg 32 Coefficient of Variation
- “The mean and the standard deviation of shock index in the total sample are 0.69 and 0.20, respectively; for systolic blood pressure, they are 138 and 26 [not 0.26], respectively”
- The formula for CV should have $\bar{X}$ [not X] in the denominator.

Formulas vs. Excel Function

- Calculating summary statistics using the formulas is time-consuming for large data sets.
- Fortunately, the summary statistics described in this lesson can be quickly calculated using Excel functions if the data are recorded on an Excel spreadsheet
- It's still important to understand the formulas and to work through the calculations at least once on the Practice Exercises. However, for most applications and for the homework assignments, use the Excel functions.

Excel: Mean

- Use the SUM function to add up the individual observations and then divide by the number of observations

  or

- Use the AVERAGE function
  - If data are in cells A1:A10, the AVERAGE function will return the mean of the data

    = AVERAGE(A1:A10)

Median and Mode in Excel

- MEDIAN function:

  If the data are in cells A1: A10, the MEDIAN function will return the Median value of the data

  = MEDIAN(A1:A10)

- MODE function

  If the data are in cells A1: A10, the MODE function will return the Mode of the data

  = MODE(A1:A10)

Problems with MODE in EXCEL

Problems with MODE function in Excel:

	Case 1	Case 2	Case 3	Case 4
				1
	1	2	1	1
	1	2	2	2
	1	1	3	2
Mode	1	2	#N/A	1

The "mode" is supposed to be the most frequent number.

•Cases 1 and 2 are straight forward.

• In Case 3 there is no mode, so Excel chooses "not applicable."

•In case 4 there is a tie and Excel chooses the smallest value. This is Excels' way of selecting, and is not standard practice.

Geometric Mean in Excel

If the data are in cells A1:A10

- Calculate the natural log of each observation in a separate column using the LN function
  - = LN(A1) will return the natural log of the value in cell A1
- Find the mean of the values on the log scale using the AVERAGE function
- Use the EXP function to find the antilog
  - If the mean on the log scale is in cell B11, =EXP(B11) will return the geometric mean

Or use the GEOMEAN function on the original data: = GEOMEAN(A1:A10) will return the geometric mean

Finding the Range and Interquartile Range in Excel

- There is no range function in Excel. Use MAX and MIN functions to find the Maximum and the Minimum
  - If the data are in A1:A10, the range is:

    =MAX(A1:A10) – MIN(A1:A10)
- For the Interquartile range, use the QUARTILE function to find Q1 and Q3.
  - If the data are in A1:A10, the IQR is:

=QUARTILE(A1:A10, 3) – QUARTILE(A1:A10, 1)

Variance and SD in Excel

- Use the VAR and STDEV functions in Excel to find the variance and standard deviation of the data
- If your data are in cells A1:A10:

  =VAR(A1:A10) will return the Variance

  =STDEV(A1:A10) will return the SD

Coefficient of Variation

- There isn't a specific function for the coefficient of variation
- Calculate the mean using the AVERAGE function
- Calculate the SD using the STDEV function
- Divide the SD by the mean and multiply by 100 to calculate the coefficient of variation.

Readings and Assignments

- Reading: Chapter 3 pgs. 27-32
- Excel Module 1
- Lesson 2 Part 1 Practice Exercises
- Homework 1: Problem 3 (3.2a and 3.2c)
- Look up any new terms in the Pocket Dictionary of Statistics