Things To Look For in Data

- Shape of data
  - Histogram, Symmetry, Peaks, Skewness
- Center
  - Mean, Median, Mode, Midrange
- Variation
  - Standard deviation

Measures of Center

Mean/Average

Sum of data values divided by number of values

WE WILL NOT CALCULATE THIS BY HAND (USUALLY)

1, 1, 2, 3, 3, 3, 4, 4, 4, 5

$$\frac{1+1+2+3+3+3+4+4+4+5}{10} = 3$$

Notation

- Σ (Greek Sigma) denotes a SUM
- x - variable to denote individual data values
- Σx - sum of all values of x
- n - number of data values in a SAMPLE
- N - number of data values in a POPULATION
- $\bar{x}$ - the mean or average of a SAMPLE
- μ - the mean or average of a POPULATION

mu

Notation

$$\textbf{Sample Mean: } \bar{x} = \frac{\sum x}{n}$$

$$\textbf{Population Mean: } \mu = \frac{\sum x}{N}$$

Median

Middle value when data is sorted

~~2~~, ~~3~~, ~~3~~, 4, ~~5~~, ~~6~~, ~~45~~

~~2~~, ~~3~~, ~~5~~, ~~5~~, 6, 7, ~~7~~, ~~9~~, ~~12~~, ~~15~~

6.5

Mode

The MODE is the most common data value

2, 3, 3, 4, 4, 5, 5, 5, 6, 7

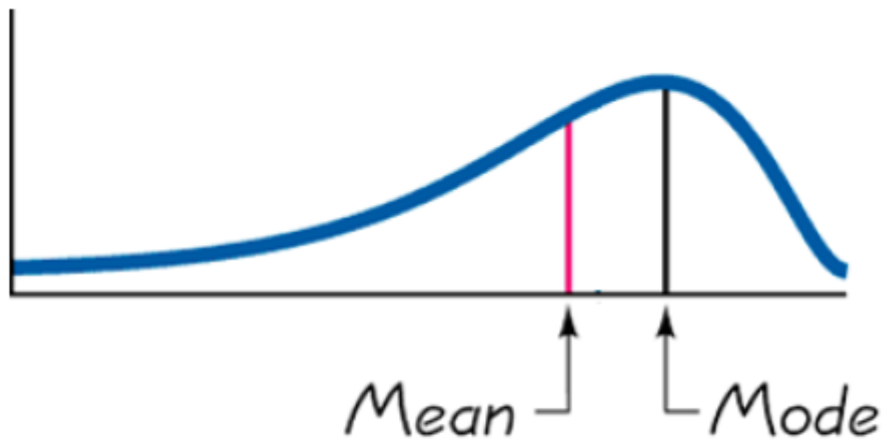
a, a, a, b, b, b, b, c, c, c

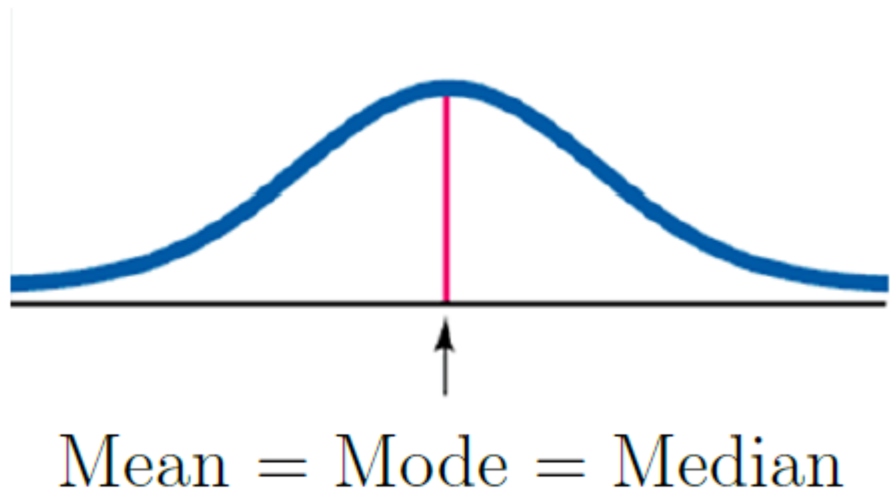
can use w/ categorical data

Bimodal: 1, 1, 1, 2, 2, 2, 3, 4, 5

Multimodal: 1, 1, 2, 2, 3, 3, 4, 5

Skewness





Midrange

Average of the maximum and minimum values

2, 3, 4, 4, 5, 5, 6, 7, 9, 10

$$\frac{10+2}{2} = 6$$

Comparison of Middle Values

	Mean	Median	Mode	Midrange
Reliable (sample values similar)	Yes	No	No	No
Resistant to outliers	No	Yes	Yes	No
Can be used with categorical data	No	No	Yes	No

Variation

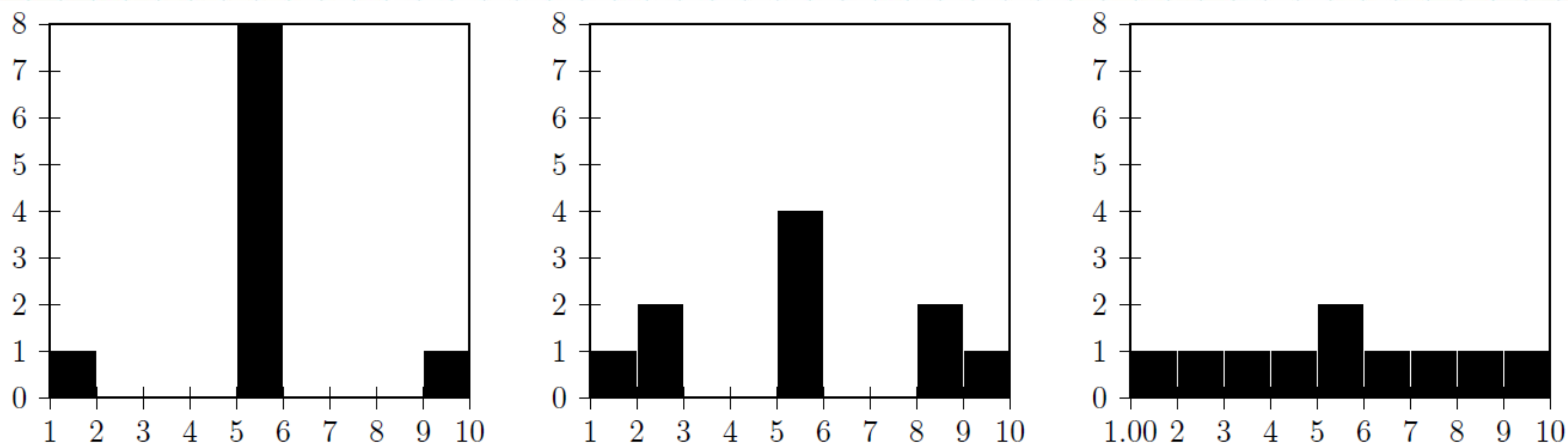
Math 122

Find the mean, median, mode, midrange of these three sets of data:

- 1, 5, 5, 5, 5, 5, 5, 5, 5, 9
- 1, 2, 2, 5, 5, 5, 5, 8, 8, 9
- 1, 2, 3, 4, 5, 5, 6, 7, 8, 9

mean = mode = median = midrange = 5

Measures of center do not distinguish between these sets of data, but we can see in their histograms that the data is distributed differently.



Measures of Variation

Range

- **Range** = maximum – minimum
- 1, 5, 5, 5, 5, 5, 5, 5, 5, 9 Range $= 9 - 1 = 8$
- 1, 2, 2, 5, 5, 5, 5, 8, 8, 9
- 1, 2, 3, 4, 5, 5, 4, 3, 2, 9

Average Distance From the Mean

- A better measure of variation might be the average distance from the mean.
  - Calculate the mean $\bar{x}$
  - For each data value x calculate $|x - \bar{x}|$
  - Average these values
- But absolute values are difficult to work with

Standard Deviation

- Sample:

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

- Population:

Sigma

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

- Think: Average Distance From the Mean

We will calculate these with technology the same way we find means

Variance

- The **variance** is the square of the standard deviation.
- Notation
  - Sample: s^2
  - Population: σ^2
- Find the variance of this data:
  - 1, 2, 3, 4, 5, 5, 4, 3, 2, 9

Mean = 3.8
St. Dev. = 2.25
Variance = 5.07

Some values from an NHANES data set of 1000 college age males and females

	Female		Male	
	Mean	St. Dev.	Mean	St. Dev.
Height	63.8 in	2.7 in	69.6 in	3.2 in
Weight	154.7 lb	43.0 lb	179.7 lb	47.7 lb
Waist	35.2 in	6.7 in	35.2 in	6.3 in
Pulse	79.1	13.2	70.5	10.7

	Mean	Standard Deviation
Annual Snowfall in Lincoln	26.7 in	11.1 in
Annual Mean Temperature in Lincoln	51.5 degrees	1.6 degrees
IQ	100	15
Infant Birth Weight	7.5lb	1.1lb
SAT Area Test	500	100
ACT	18	6

Range Rule of Thumb

- For many types of data (whose histograms are bell-shaped) about 95% of data values are within two standard deviations of the mean.
- **Maximum Usual Value** = $\mu + 2\sigma$
  - Any value larger than this is unusually large.
- **Minimum Usual Value** = $\mu - 2\sigma$
  - Any value smaller than this is unusually small.

Snowfall

$\mu = 26.7$ and $\sigma = 11.1$

- Find the maximum usual annual snowfall for Lincoln.

  $max = \mu + 2\sigma = 26.7 + 2 \times 11.1 = 48.9$

- Find the minimum usual annual snowfall for Lincoln.

  $min = \mu - 2\sigma = 26.7 - 2 \times 11.1 = 4.5$

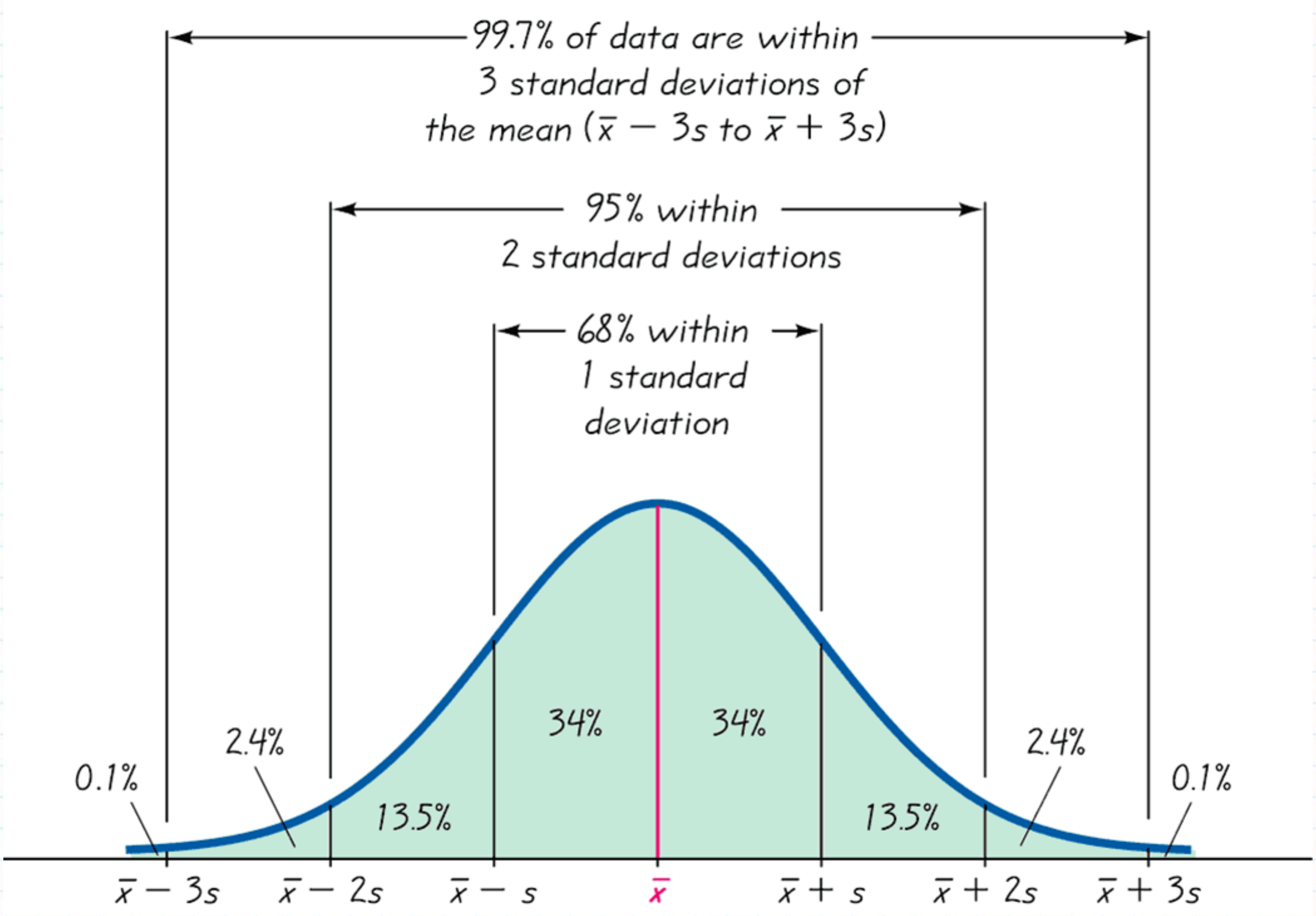
- Would 40 inches in one year be unusual?

  NO!

Empirical Rule

- For many types of data (whose histograms are bell-shaped)
  - About 68% of the data values are within one standard deviation of the mean.
  - About 95% of the data values are within two standard deviations of the mean.
  - About 99.7% of the data values are within three standard deviations of the mean.

Empirical Rule



Infant Birth Weights

$\mu = 7.5$ and $\sigma = 1.1$

- Use the Empirical Rule to find an interval which contains 68% of infant birth weights.

$\mu - \sigma = 7.5 - 1.1 = 6.4$ lb

$\mu + \sigma = 7.5 + 1.1 = 8.6$ lb

- Find an interval which contains 95% of infant birth weights.

$\mu - 2\sigma = 7.5 - 2 \times 1.1 = 5.3$ lb

$\mu + 2\sigma = 7.5 + 2 \times 1.1 = 9.7$ lb

- Find an interval which contains 99.7% of infant birth weights.

$\mu - 3\sigma = 7.5 - 3 \times 1.1 = 4.2$ lb

$\mu + 3\sigma = 7.5 + 3 \times 1.1 = 10.8$ lb

Summary

- Range = max – min
- Standard Deviation
  - Think average distance from the mean
- Range Rule of Thumb
  - Find max and min usual values
  - Identify unusual values
- Empirical Rule
  - Find intervals for 68%, 95%, 99.7%