Dispersion

- Dispersion: Variance, Standard Deviation

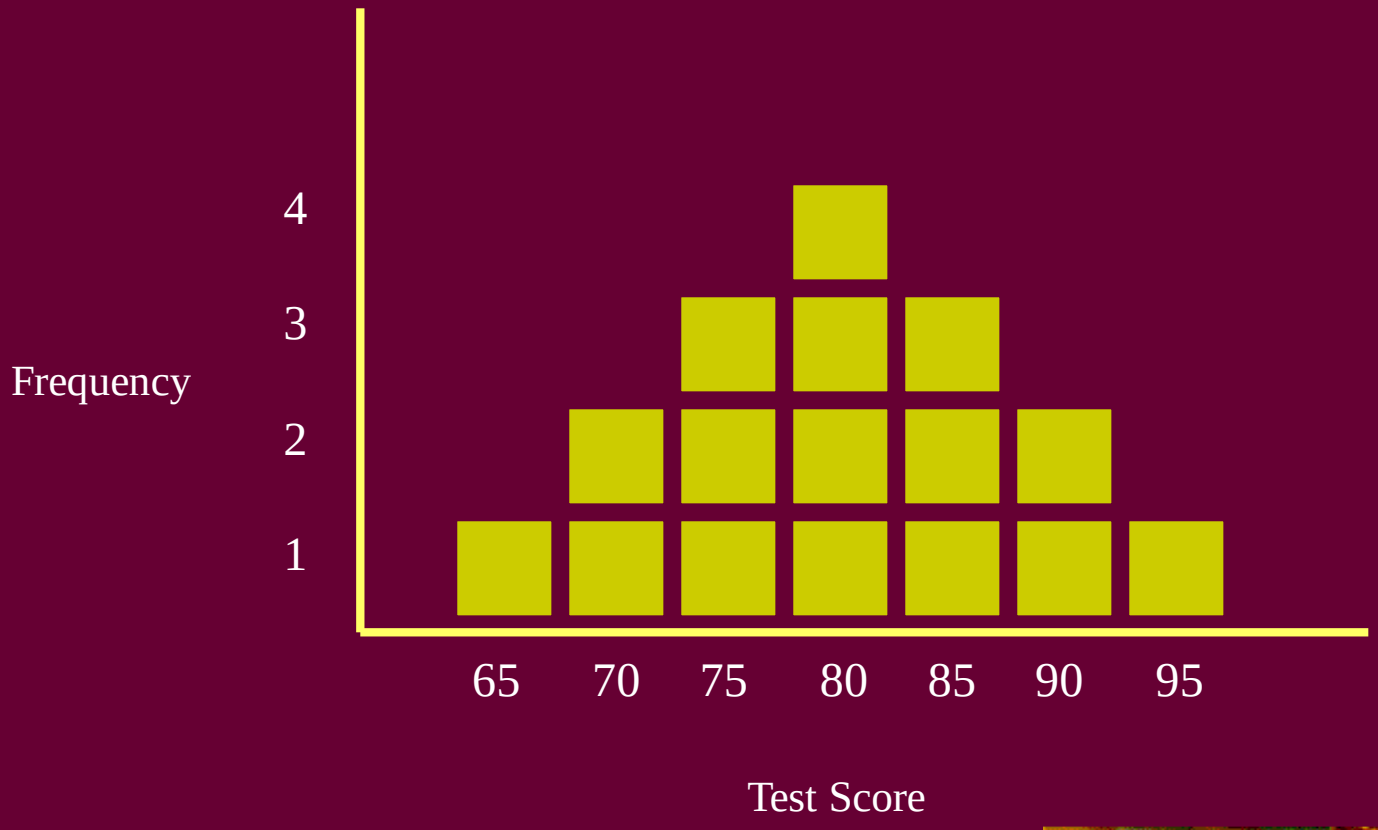
Frequency Distributions

- Simple depiction of all the data
- Graphic — easy to understand
- Problems
  - Not always precisely measured
  - Not summarized in one number or datum

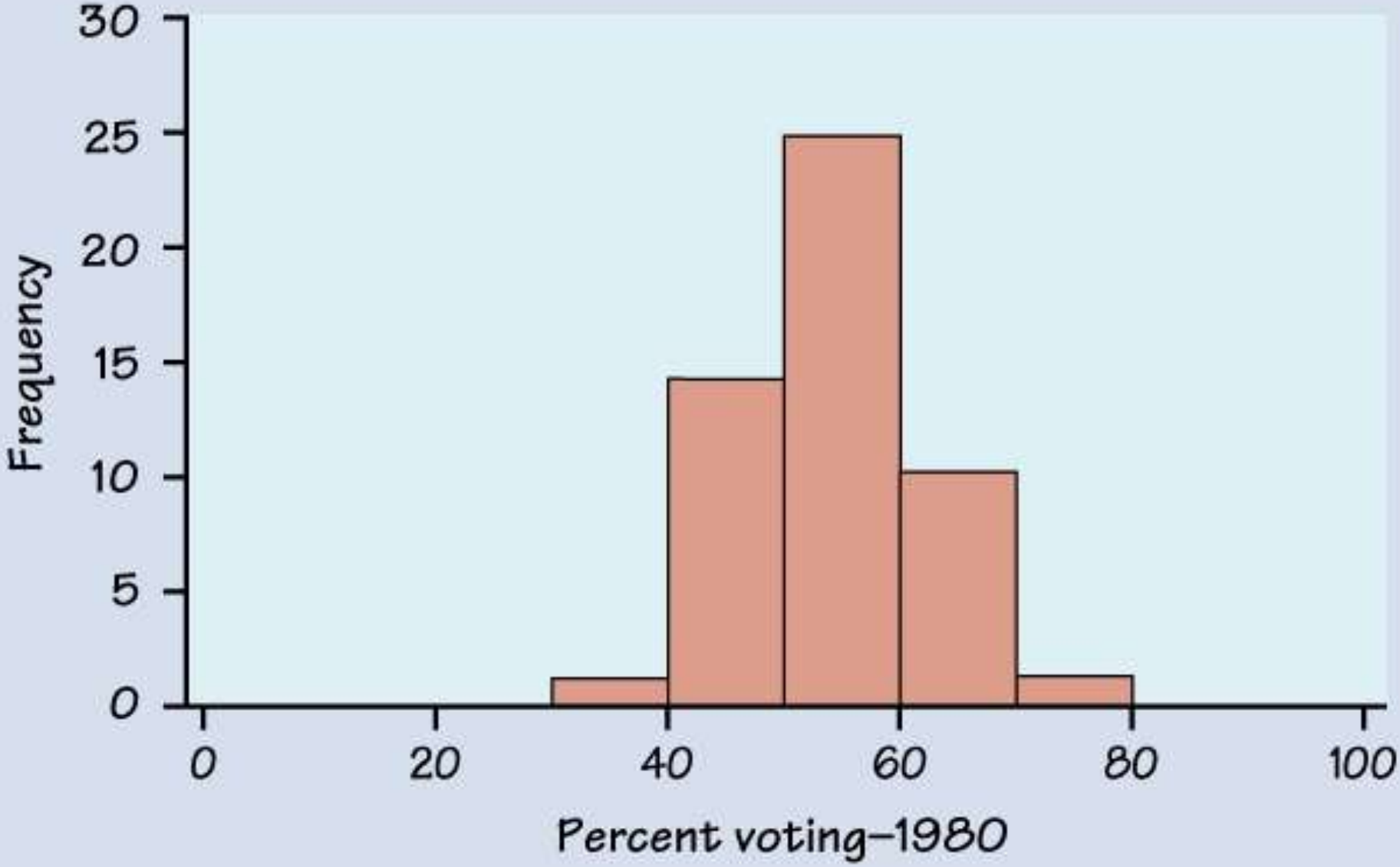
Frequency Table
Test Scores

Observation	Frequency
65	1
70	2
75	3
80	4
85	3
90	2
95	1

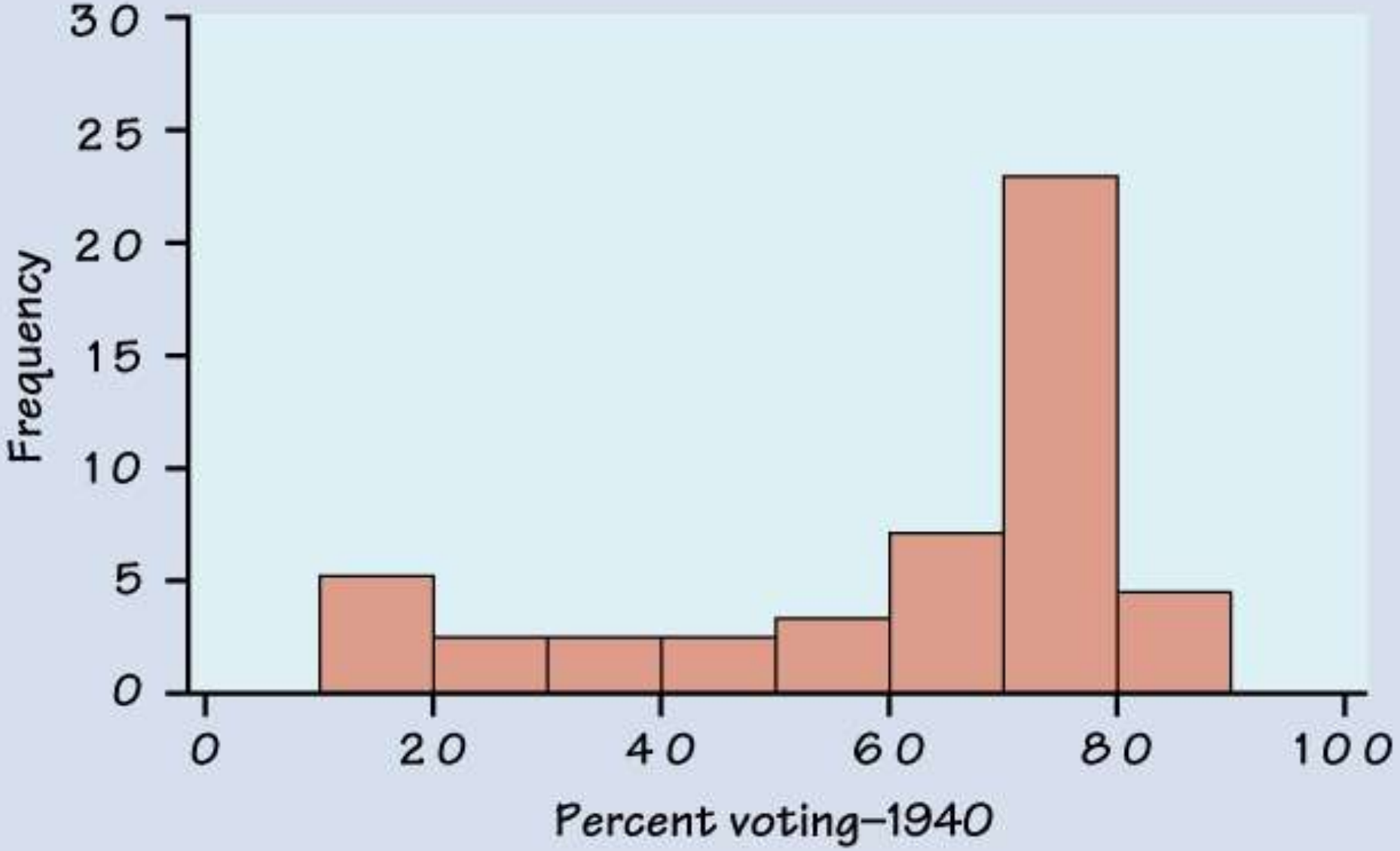
Frequency Distributions



Voter Turnout in 50 States - 1980



Voter Turnout in 50 States - 1940



Normally Distributed Curve

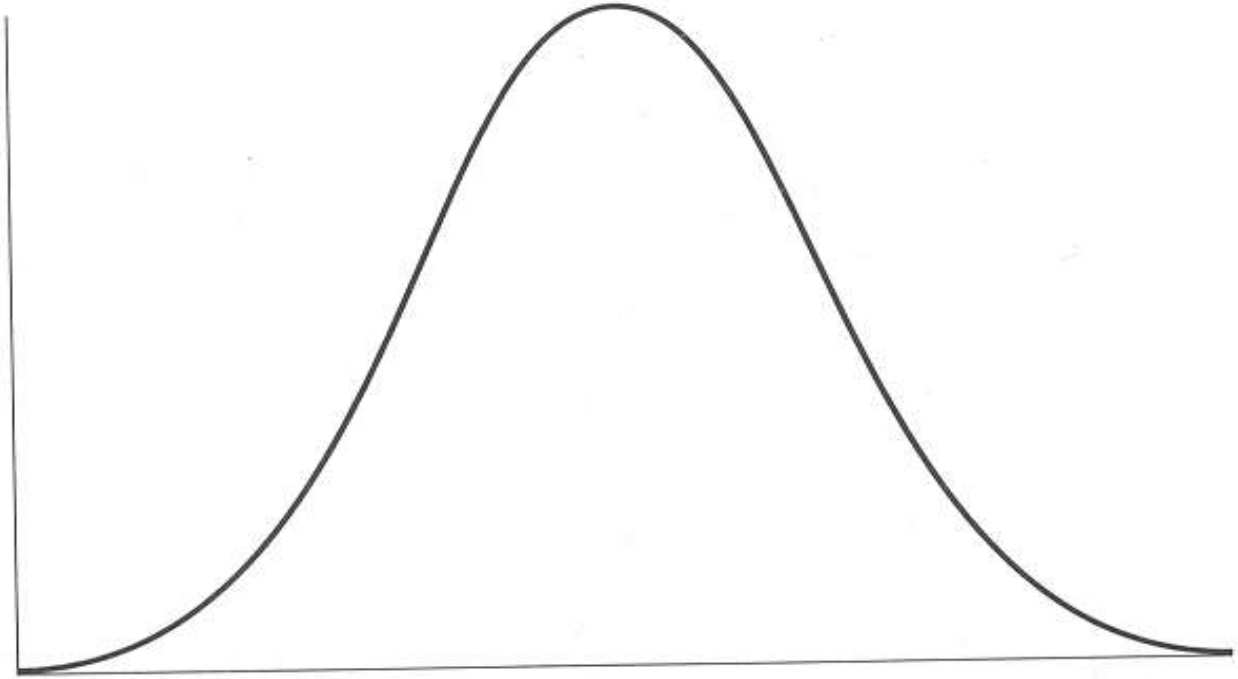
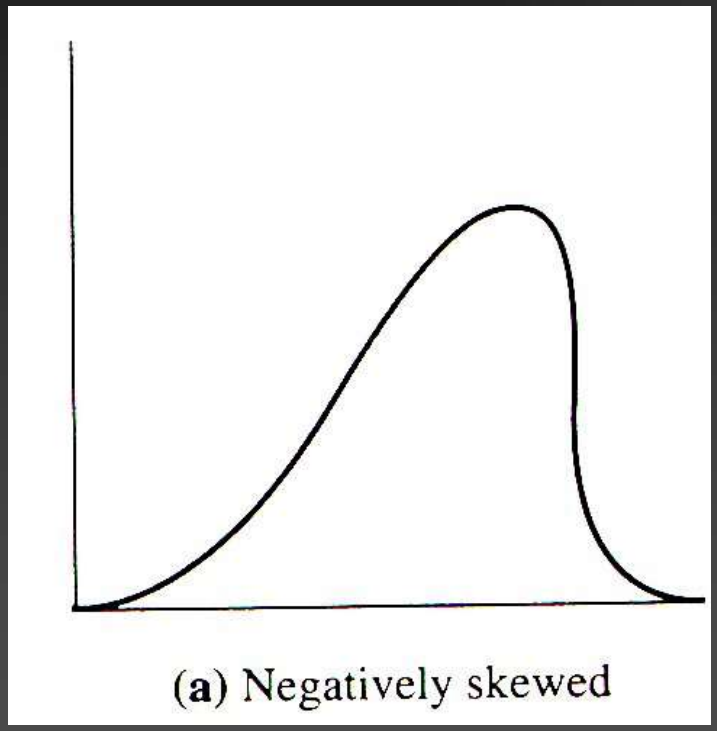
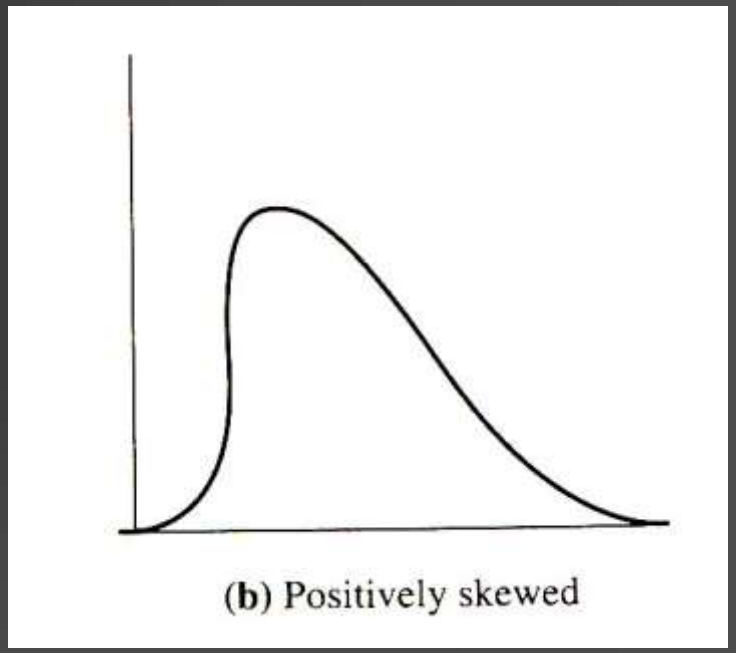


FIGURE 5.4 The Shape of the Normal Curve

Skewed Distributions



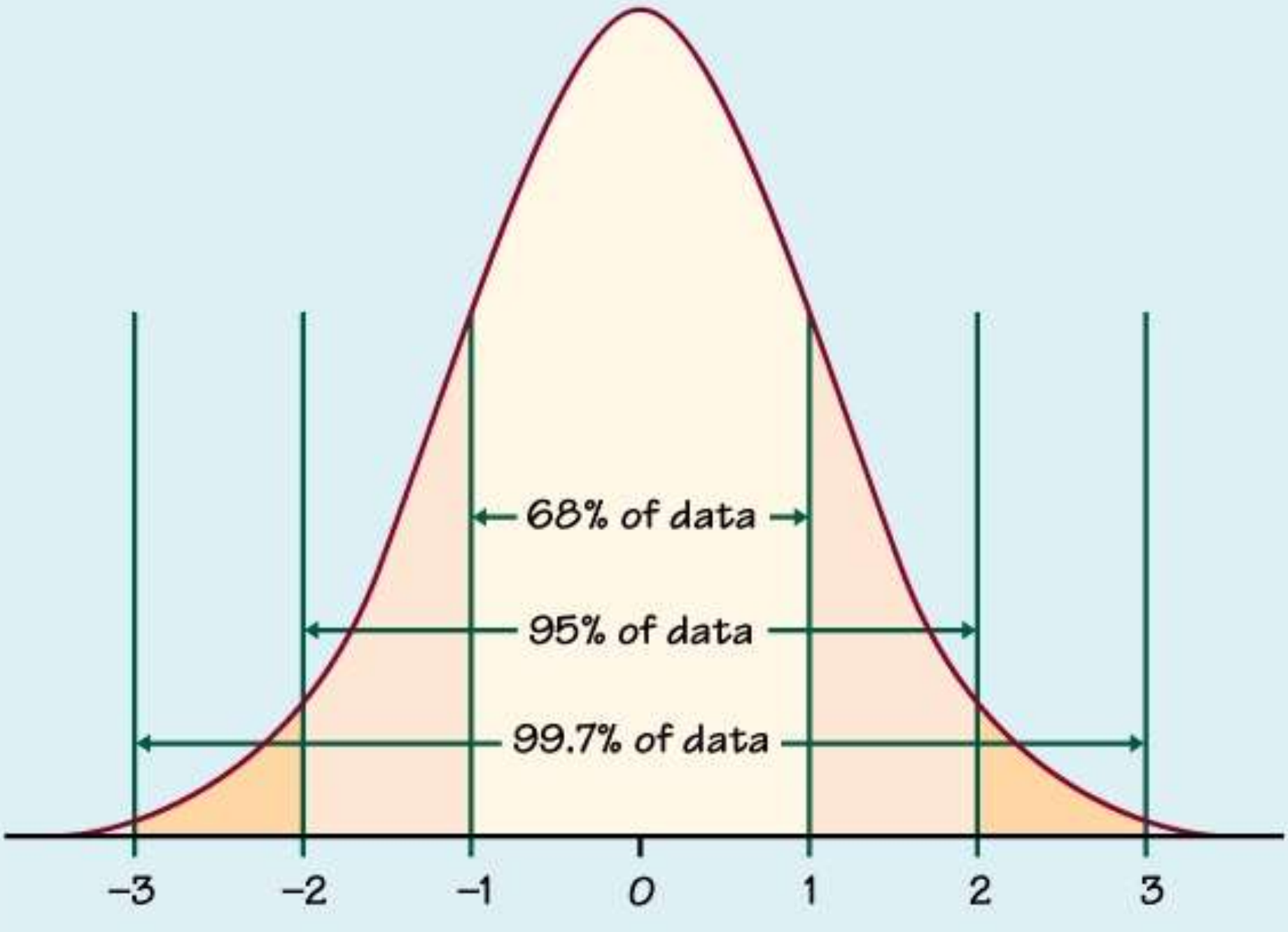
(a) Negatively skewed



(b) Positively skewed

Characteristics of the Normal Distribution

- It is symmetrical -- Half the cases are to one side of the center; the other half is on the other side.
- The distribution is single peaked, not bimodal or multi-modal
- Most of the cases will fall in the center portion of the curve and as values of the variable become more extreme they become less frequent, with "outliers" at each of the "tails" of the distribution few in number.
- It is only one of many frequency distributions but the one we will focus on for most of this course.
- The Mean, Median, and Mode are the same.
- Percentage of cases in any range of the curve can be calculated.


68% of data
95% of data
99.7% of data
−3
−2
−1
0
1
2
3

Family of Normal Curves

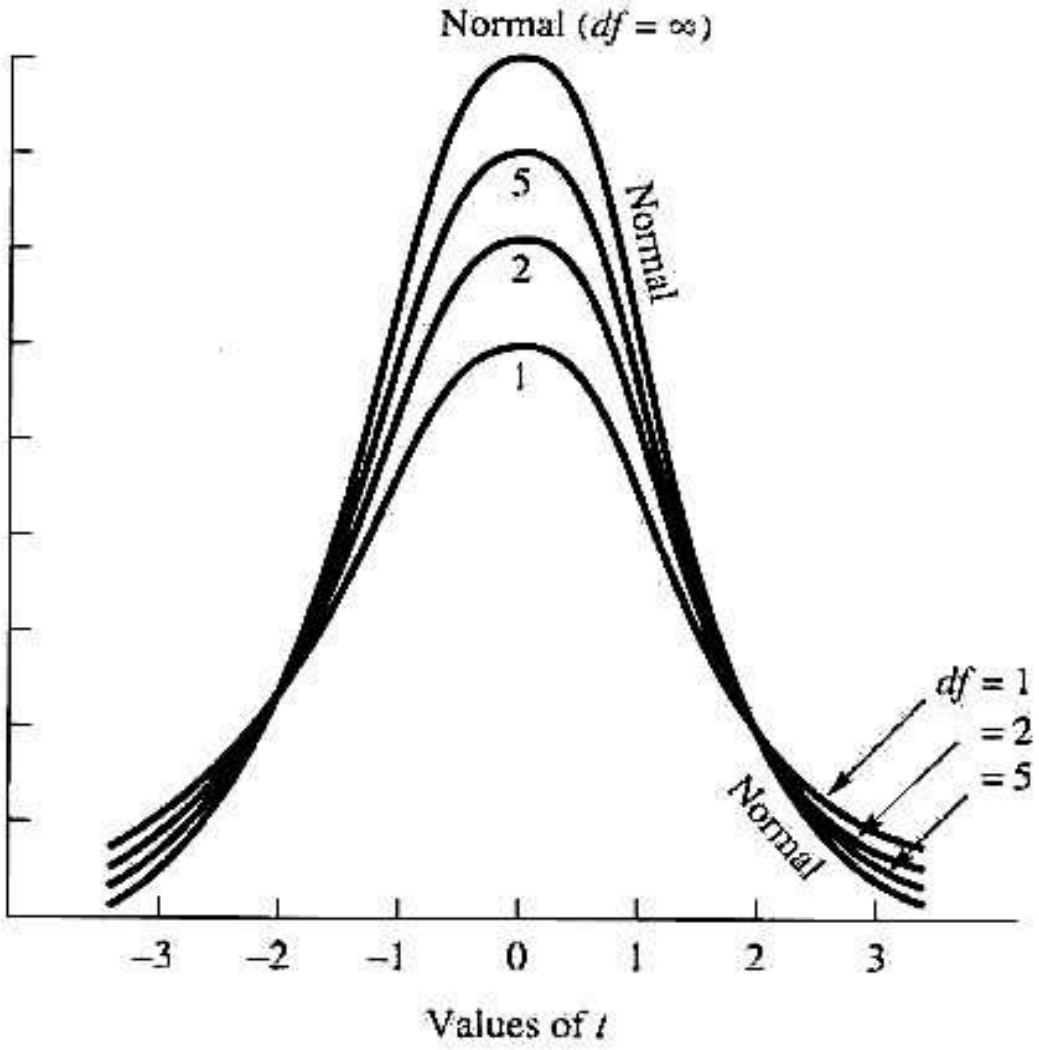


FIGURE 6.8 The Family of *t* Distributions

Summarizing Distributions

Two key characteristics of a frequency distribution are especially important when summarizing data or when making a prediction from one set of results to another:

- Central Tendency
  - What is in the “Middle”?
  - What is most common?
  - What would we use to predict?
- Dispersion
  - How Spread out is the distribution?
  - What Shape is it?

Three measures of central tendency are commonly used in statistical analysis - the mode, the median, and the mean

Each measure is designed to represent a typical score

The choice of which measure to use depends on:

- the shape of the distribution (whether normal or skewed), and
- the variable's "level of measurement" (data are nominal, ordinal or interval).

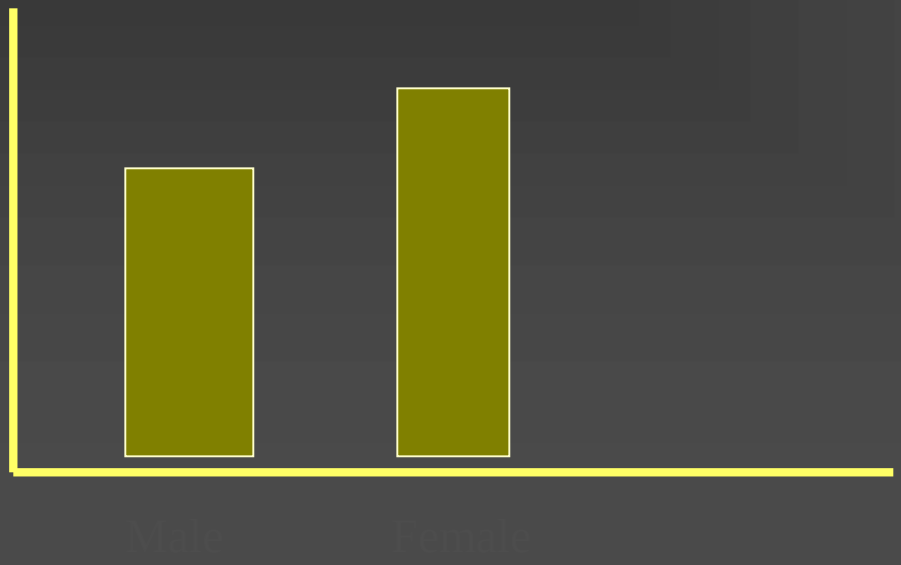
Appropriate Measures of Central Tendency

- Nominal variables → Mode
- Ordinal variables → Median
- Interval level variables → Mean

  - If the distribution is normal (median is better with skewed distribution)

Mode

- Most Common Outcome



Median

- Middle-most Value
- 50% of observations are above the Median, 50% are below it
- The difference in magnitude between the observations does not matter
- Therefore, it is not sensitive to outliers
- Formula Median = n + 1 / 2

To compute the median

- first you rank order the values of X from low to high: → 85, 94, 94, 96, 96, 96, 96, 97, 97, 98
- then count number of observations = 10.
- add 1 = 11.
- divide by 2 to get the middle score → the 5 ½ score

  here 96 is the middle score score

Median

- Find the Median

  4 5 6 6 7 8 9 10 12

- Find the Median

  5 6 6 7 8 9 10 12

- Find the Median

  5 6 6 7 8 9 10 100,000

Mean - Average

- Most common measure of central tendency
- Best for making predictions
- Applicable under two conditions:

1. scores are measured at the interval level, and
2. distribution is more or less normal [symmetrical].

- Symbolized as: $\overline{X}$
  - for the mean of a sample
  - μ for the mean of a population

Finding the Mean

- $\bar{X} = (\Sigma X) / N$
- If $X = \{3, 5, 10, 4, 3\}$

$$\bar{X} = (3 + 5 + 10 + 4 + 3) / 5$$
$$= 25 / 5$$
$$= 5$$

Find the Mean

Q: 4, 5, 8, 7

A: 6

Median: 6

Q: 4, 5, 8, 1000

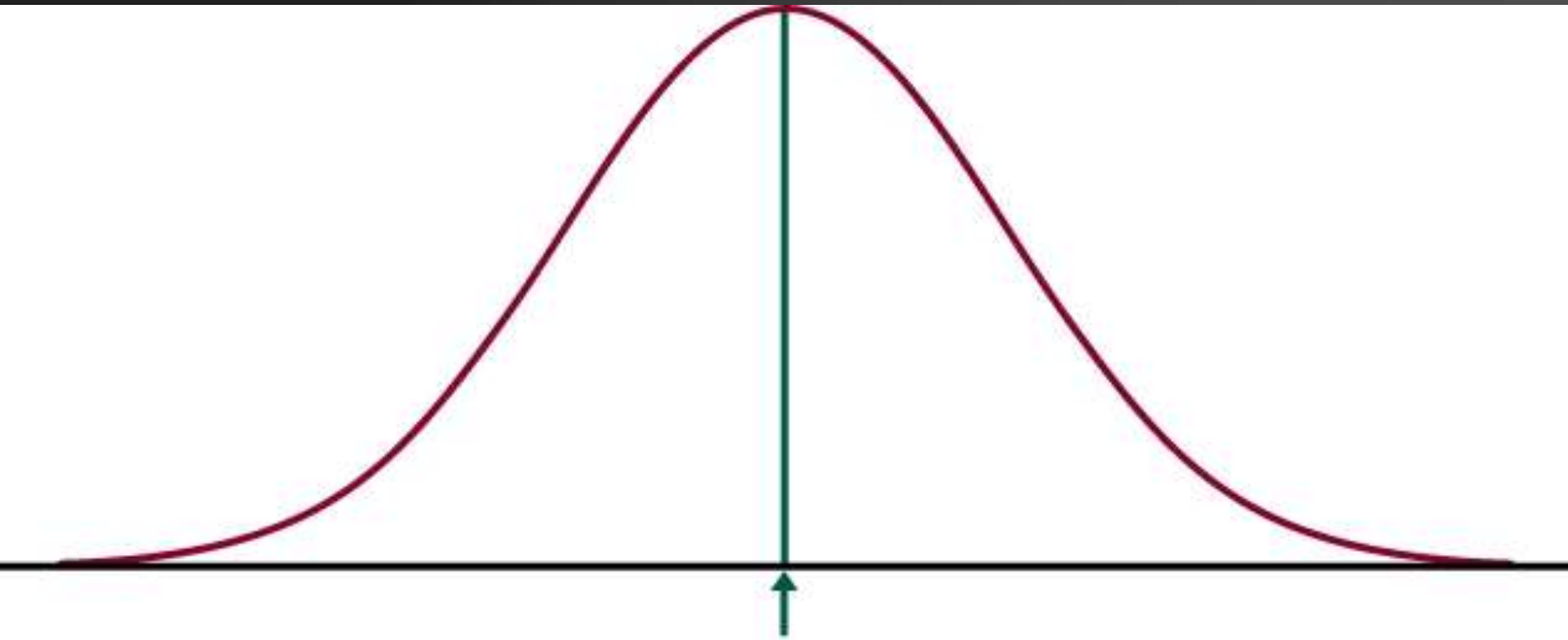
A: 254.25

Median: 6.5

IF THE DISTRIBUTION IS NORMAL

Mean is the best measure of central tendency

- Most scores “bunched up” in middle
- Extreme scores less frequent ➔ don’t move mean around.



Median and mean

Measures of Variability

Central Tendency doesn't tell us everything

Dispersion/Deviation/Spread tells us a lot about how a variable is distributed.

We are most interested in Standard Deviations (σ) and Variance (σ^2)

Why can’t the mean tell us everything?

- Mean describes Central Tendency, what the average outcome is.
- We also want to know something about how accurate the mean is when making predictions.
- The question becomes how good a representation of the distribution is the mean? How good is the mean as a description of central tendency -- or how good is the mean as a predictor?
- Answer -- it depends on the shape of the distribution. Is the distribution normal or skewed?

Family of Normal Distribution Curves

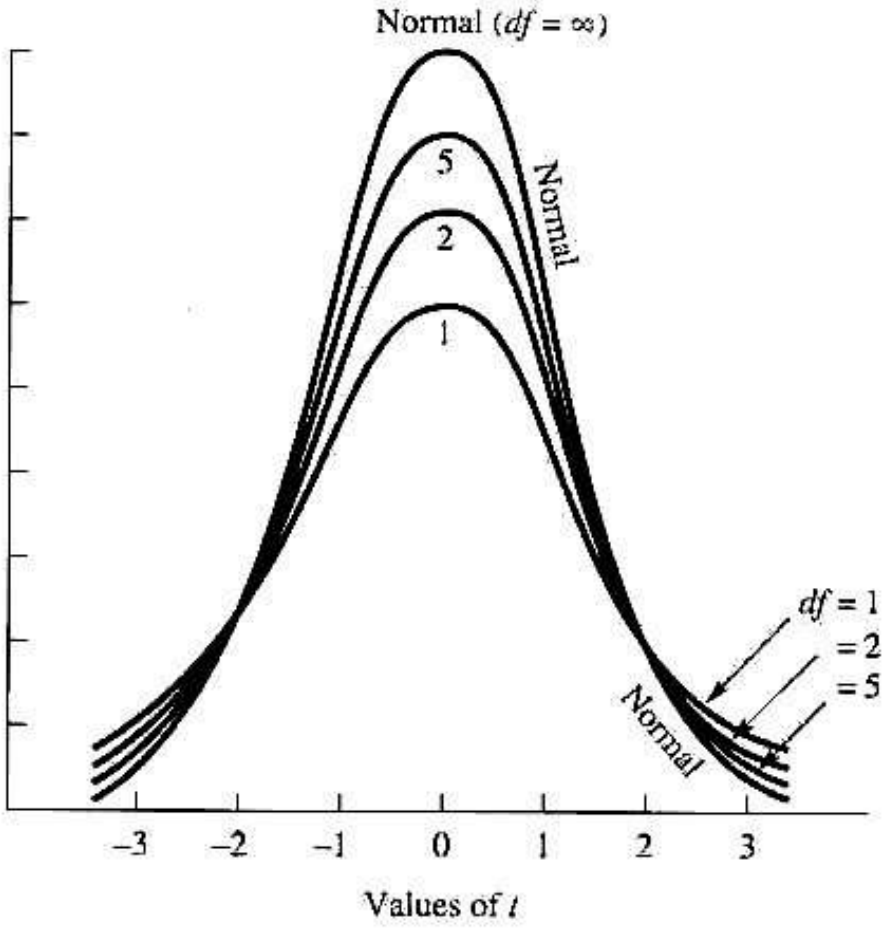


FIGURE 6.8 The Family of *t* Distributions

Dispersion

Once you determine that the variable of interest is normally distributed, ideally by producing a histogram of the scores, the next question to be asked about the NDC is its dispersion: **how spread out are the scores around the mean.**

Dispersion is a key concept in statistical thinking.

The basic question being asked is how much do the scores deviate around the Mean? The more "bunched up" around the mean the better your ability to make accurate predictions.

Means

- Consider these means for weekly candy bar consumption.

$X = \{7, 8, 6, 7, 7, 6, 8, 7\}$

$\overline{X} = (7+8+6+7+7+6+8+7)/8$

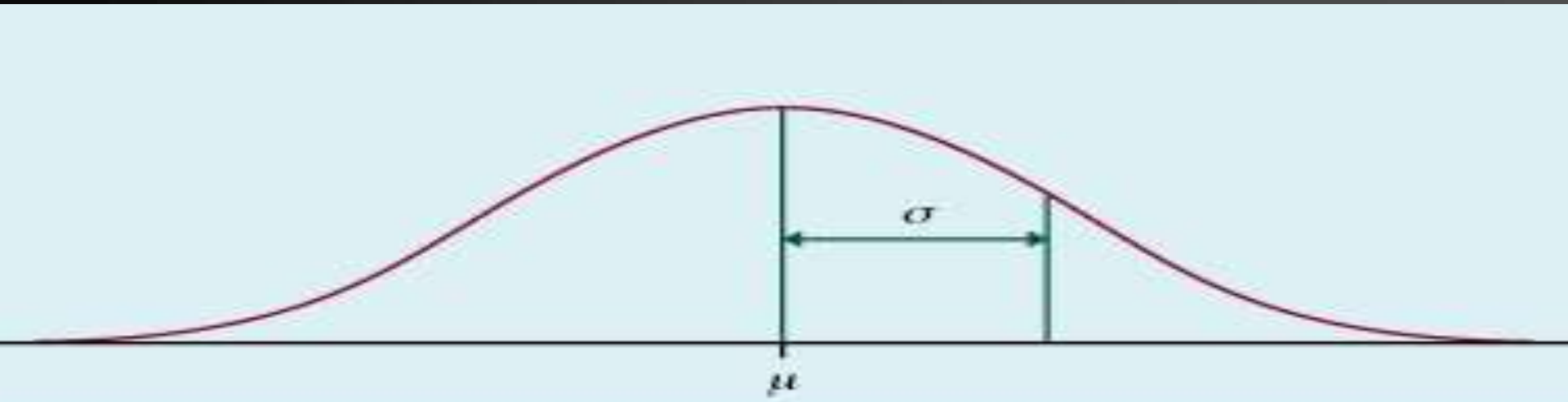
$\overline{X} = 7$

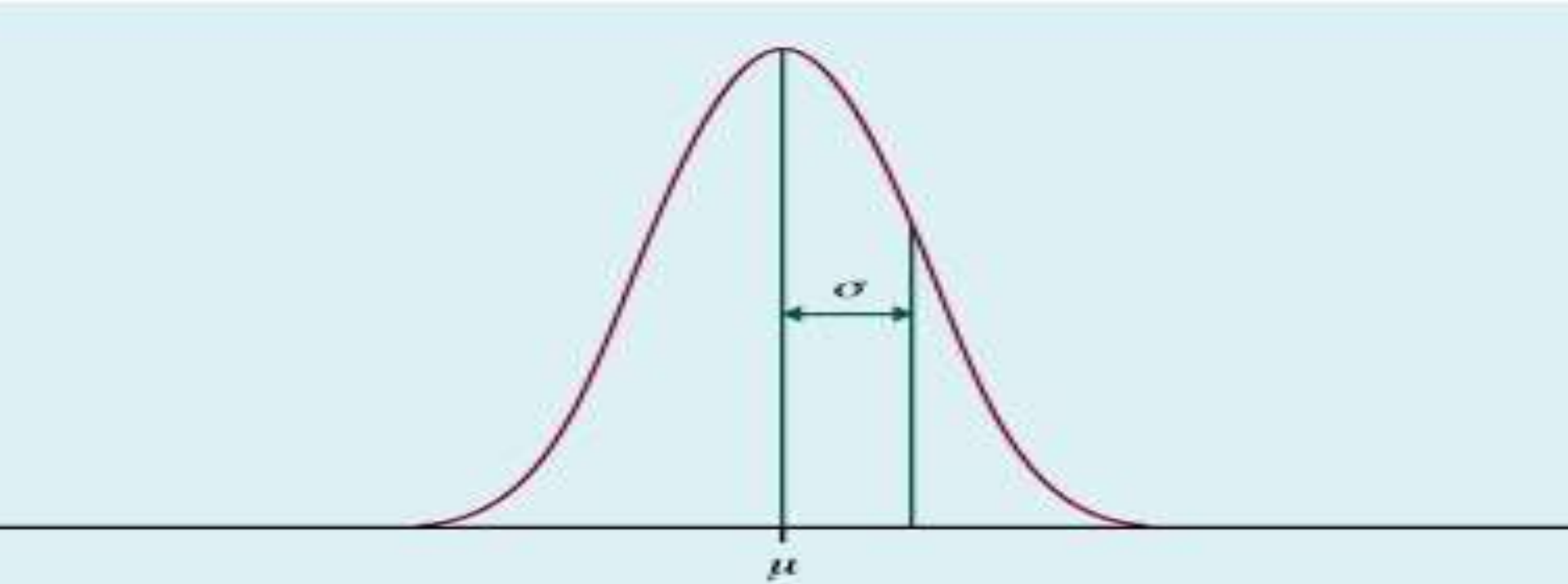
$X = \{12, 2, 0, 14, 10, 9, 5, 4\}$

$\overline{X} = (12+2+0+14+10+9+5+4)/8$

$\overline{X} = 7$

What is the difference?


σ
μ


σ
μ

How well does the mean represent the scores in a distribution? The logic here is to determine **how much spread** is in the scores. How much do the scores "deviate" from the mean? Think of the mean as the **true score** or as your **best guess**. If every X were very close to the Mean, the mean would be a very good predictor.

If the distribution is very sharply peaked then the mean is a good measure of central tendency and if you were to use the mean to make predictions you would be right or close much of the time.

What if scores are widely distributed?

The mean is still your best measure and your best predictor, but your predictive power would be less.

How do we describe this?

- Measures of variability
  - Mean Deviation
  - Variance
  - Standard Deviation

Mean Deviation

The key concept for describing normal distributions and making predictions from them is called **deviation from the mean**.

We could just calculate the average distance between each observation and the mean.

- We must take the absolute value of the distance, otherwise they would just cancel out to zero!

Formula:

$$\sum \frac{|\overline{X} - X_i|}{n}$$

Mean Deviation: An Example

Data: $X = \{6, 10, 5, 4, 9, 8\}$ $\quad \bar{X} = 42 / 6 = 7$

$\bar{X} - X_i$	Abs. Dev.
7 – 6	1
7 – 10	3
7 – 5	2
7 – 4	3
7 – 9	2
7 – 8	1
Total:	12

1. Compute X (Average)
2. Compute X – X and take the Absolute Value to get Absolute Deviations
3. Sum the Absolute Deviations
4. Divide the sum of the absolute deviations by N

12 / 6 = 2

What Does it Mean?

- On Average, each observation is two units away from the mean.

Is it Really that Easy?

- No!
- Absolute values are difficult to manipulate algebraically
- Absolute values cause enormous problems for calculus (Discontinuity)
- We need something else…

Variance and Standard Deviation

- Instead of taking the absolute value, we square the deviations from the mean. This yields a positive value.
- This will result in measures we call the Variance and the Standard Deviation

Sample-	Population-
s: Standard Deviation	σ: Standard Deviation
s^2: Variance	σ^2: Variance

Calculating the Variance and/or Standard Deviation

Formulae:

Variance:

$$s^2 = \frac{\sum(\overline{X} - X_i)^2}{N}$$

Standard Deviation:

$$s = \sqrt{\frac{\sum(\overline{X} - X_i)^2}{N}}$$

Examples Follow . . .

Example:

Data: X = {6, 10, 5, 4, 9, 8}; N = 6

X	$X - \bar{X}$	$(X - \bar{X})^2$
6	-1	1
10	3	9
5	-2	4
4	-3	9
9	2	4
8	1	1
Total: 42		Total: 28

Mean:

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

Variance:

$$s^2 = \frac{\sum (\bar{X} - X)^2}{N} = \frac{28}{6} = 4.67$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$