Lecture 5

Summaries for Symmetric Distributions

§3.1

1. Summaries for Symmetric Distributions

When a numerical distribution is approximately symmetric, the **mean** provides a useful measure of the center of the data.

For symmetric distributions, we will also use the **standard deviation** to describe the spread.

For an approximately symmetric distribution:

$$\boxed{\text{Center} = \text{Mean} \qquad\qquad \text{Spread} = \text{Standard Deviation}}$$

The mean describes a typical value, while the standard deviation describes how far observations tend to be from the mean.

2. The Mean

Definition.
Given a collection of data

$$\{x_1, x_2, \ldots, x_n\},$$

the **mean** of the data is the arithmetic mean:

$$\bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Here,

- n is the number of observations;
- x_i represents the i-th observation;
- $\bar{x}$ represents the sample mean.

In words, the mean is obtained by

$$\boxed{\text{adding all observations and dividing by the number of observations.}}$$

2.1 Example: Computing a Mean

Suppose the data are

$$4, \quad 6, \quad 7, \quad 8, \quad 10.$$

There are $n = 5$ observations, so

$$\bar{x} = \frac{4 + 6 + 7 + 8 + 10}{5}.$$

Thus,

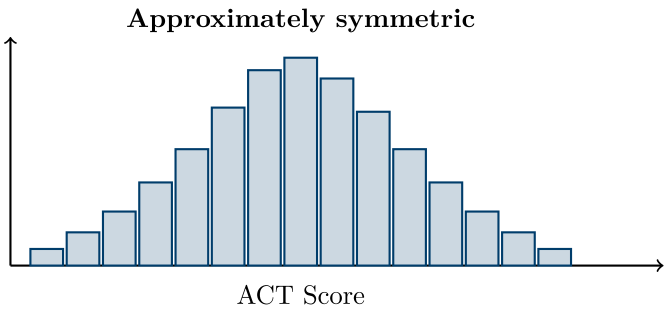
$$\bar{x} = \frac{35}{5} = \boxed{7}.$$

Therefore, the mean of the data is 7.

2.2 Example: ACT Scores

An instructor at Peoria Junior College in Illinois collected data from two classes, including the students' ACT scores.

Below is the distribution of self-reported ACT scores for one statistics class.



The distribution is approximately symmetric and mound-shaped.

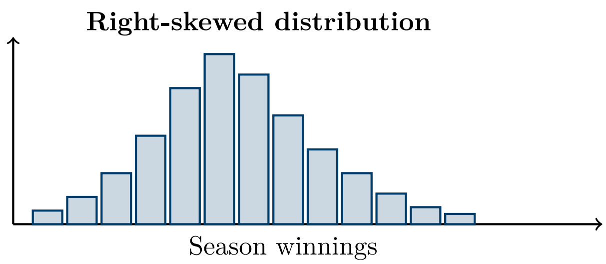
Therefore, the mean is an appropriate measure of center.

When a distribution is approximately symmetric, the mean represents a typical value in the data.

$$\boxed{\text{Symmetric distribution} \quad \Rightarrow \quad \text{mean is a good measure of center}}$$

2.3 Example: Professional Tennis Winnings

The winnings of the top-ranked professional tennis players in the 2018 season are represented by the graph below.



The distribution of professional tennis winnings is strongly right-skewed. A small number of players have extremely large winnings, creating a long right tail.

Important:
For a skewed distribution, the mean may not represent a typical observation well because extreme observations can pull the mean toward the long tail.

$$\boxed{\text{Skewed distribution} \quad \Rightarrow \quad \text{mean may not be a good measure of a typical value}}$$

3. When Is the Mean a Good Measure of Center?

The mean works especially well when the distribution is approximately symmetric.

Note:

- When the distribution is roughly symmetric, the mean represents a typical value in the data.
- The mean is not a good estimate of a typical value of a skewed distribution.

Why?

The mean uses every observation in its calculation. Consequently, unusually large or small observations can have a substantial effect on the mean.

$$\boxed{\text{Extreme values} \quad \longrightarrow \quad \text{can pull the mean toward them}}$$

3.1 Example: Gas Prices in Austin

According to GasBuddy.com, the prices of one gallon of regular gas at 12 service stations near the downtown area of Austin, Texas, were reported one winter day in 2018.

The prices were

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

We want to find the mean price.

3.2 Step 1: Add the observations

$$\begin{aligned}\sum x_i &= 2.19 + 2.19 + 2.39 + 2.19 \\ &\quad + 2.24 + 2.39 + 2.27 + 2.29 \\ &\quad + 2.17 + 2.29 + 2.30 + 2.29 \\ &= 27.40.\end{aligned}$$

There are $n = 12$ service stations.

3.3 Step 2: Divide by the number of observations

$$\bar{x} = \frac{\sum x_i}{n} = \frac{27.40}{12}.$$

Therefore,

$$\boxed{\bar{x} \approx \$2.2833}$$

and, rounded to the nearest cent,

$$\boxed{\bar{x} \approx \$2.28}.$$

3.4 Interpretation

The mean price of a gallon of regular gas at these 12 service stations was approximately

$$\boxed{\$2.28 \text{ per gallon.}}$$

In context, this means that the average reported price across the 12 service stations was about \$2.28 per gallon.

The data are fairly concentrated around \$2.28, with no extremely large or small observations. Thus, the mean provides a reasonable description of a typical gas price for these stations.

Answer:

$$\boxed{\text{Mean gas price} = \$2.28 \text{ per gallon}}$$

Interpretation: The average price of regular gasoline among the 12 sampled service stations was approximately \$2.28 per gallon.

4. The Standard Deviation

The mean describes the center of a distribution, but it does not tell us how spread out the observations are.

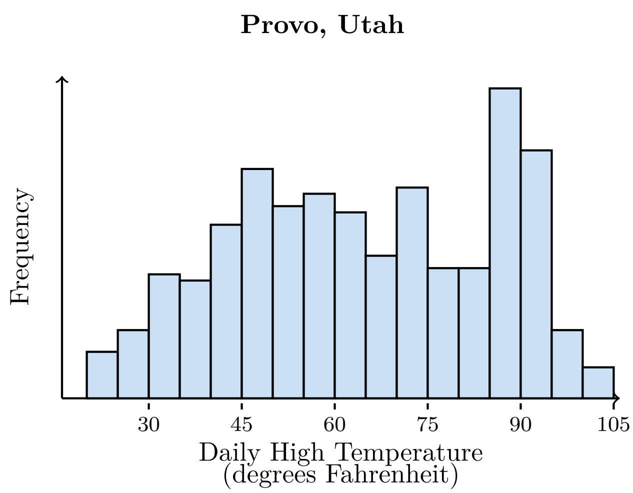
For symmetric distributions, we use the **standard deviation** to measure spread.

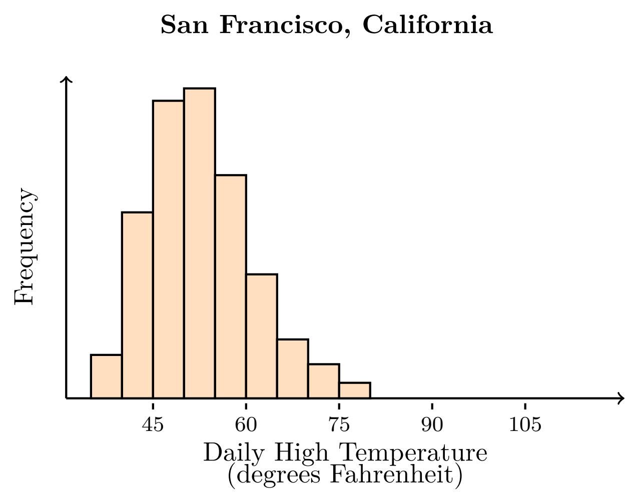
Definition.
The **standard deviation** is a number that measures how far the typical observation is from the mean.
For symmetric, unimodal distributions, a majority of the data is within one standard deviation of the mean.
The standard deviation is given by

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

4.1 Example: Daily High Temperatures

The histograms below show the daily high temperatures, in degrees Fahrenheit, recorded over one year in Provo, Utah, and San Francisco, California.





Which city do we expect to have a higher standard deviation?

Standard deviation measures the typical distance of the observations from their mean. Therefore, when comparing two distributions, we look at how spread out the observations are around the center of the distribution.

The daily high temperatures in Provo are spread across a much wider range of temperatures, while San Francisco's temperatures are concentrated in a relatively narrow range. This indicates that Provo's temperatures tend to be farther from their mean.

Therefore,

$$\boxed{s_{\text{Provo}} > s_{\text{San Francisco}}}$$

Thus, we expect **Provo** to have the higher standard deviation.

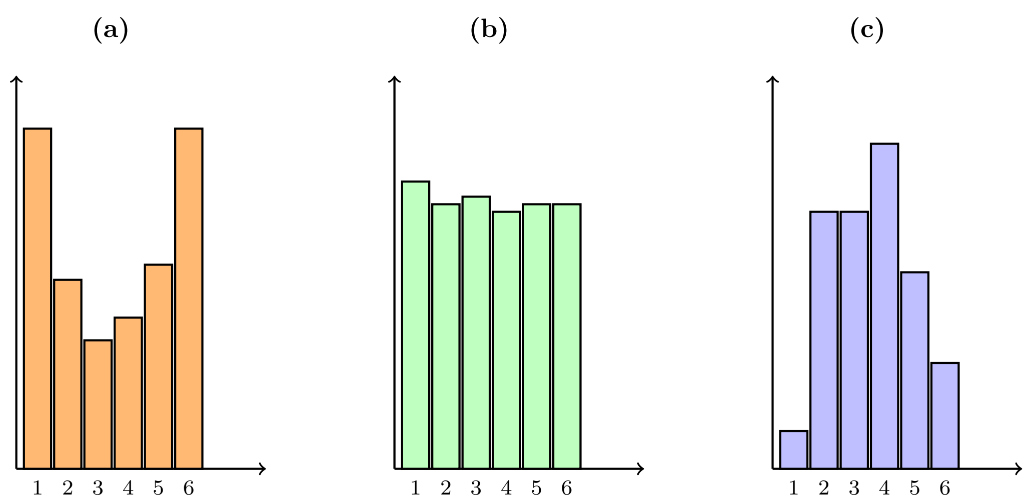
$$\boxed{\text{Greater typical distance from the mean} \implies \text{larger standard deviation}}$$

When comparing histograms, focus on how spread out the observations are around their center. A distribution with observations that are typically farther from its mean will have the larger standard deviation.

4.2 Example: Comparing Standard Deviations Graphically

Below are three histograms representing distributions with the same mean.

Which distribution has the largest standard deviation? Which has the smallest?



All three distributions have the same mean. However, the observations are distributed differently around that mean.

- Distribution (a) has many observations at the extremes, far from the mean.
- Distribution (b) spreads observations relatively evenly across the entire range.
- Distribution (c) has most of its observations concentrated closer to the center.

Therefore,

$$\boxed{s_{(a)} > s_{(b)} > s_{(c)}.}$$

Thus,

Largest standard deviation: (a)

Smallest standard deviation: (c)

The shape of a distribution (symmetric, skewed, or uniform) does not by itself determine the standard deviation.
Instead, ask:

How far are the observations from the mean?

More observations far from the mean produce a larger standard deviation. More observations close to the mean produce a smaller standard deviation.

4.3 Example: Interpreting Standard Deviation

Suppose two symmetric distributions have the same mean.

$$\bar{x} = 50.$$

Suppose distribution A has

$$s_A = 3$$

while distribution B has

$$s_B = 12.$$

Both distributions are centered at 50, but the observations in distribution B are much more spread out.

$$\boxed{\text{Larger standard deviation} \Rightarrow \text{more variability}}$$

$$\boxed{\text{Smaller standard deviation} \Rightarrow \text{less variability}}$$

The standard deviation is therefore a measure of the typical distance between an observation and the mean.

4.4 Example: Gas Prices and Standard Deviation

Recall the data set of gas prices from earlier:

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

We can use **StatCrunch** to compute the standard deviation of this data set and interpret the result.

Using StatCrunch

1. Click "Open StatCrunch."
2. Enter the data into the spreadsheet.
3. Under the "Stat" menu, select "Summary Stats," then select "Columns" or "Rows."

StatCrunch will report the sample standard deviation along with other summary statistics.

For these 12 gas prices, the sample standard deviation is approximately

$$\boxed{s \approx \$0.07}.$$

Interpretation: The gas prices typically differ from the mean gas price of approximately \$2.28 by about \$0.07.

4.5 Variance

Definition.
The **variance** is the standard deviation squared.
For a sample, the variance is

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

where

- s^2 is the sample variance;
- x_i is the i-th observation;
- $\bar{x}$ is the sample mean;
- n is the number of observations.

Since

$$s^2 = s \cdot s,$$

we can recover the standard deviation from the variance by taking the square root:

$$\boxed{s = \sqrt{s^2}}.$$

For the gas-price data,

$$s^2 \approx 0.0052,$$

so

$$s = \sqrt{0.0052} \approx \boxed{0.07}.$$

The variance measures spread using squared units, while the standard deviation converts the measure of spread back to the original units.
For gas prices:

$$\boxed{s \approx \$0.07 \text{ per gallon}}$$

This tells us that the gas prices typically vary from their mean by about seven cents per gallon.