Lecture 9

§3.3 Summaries for Skewed Distributions and §3.4 Comparing Measures of Center

1. §3.3 Summaries for Skewed Distributions

When a distribution is skewed, the mean and standard deviation can be strongly affected by unusually large or small observations.

For skewed distributions, we typically use the **median** to describe the center and the **interquartile range (IQR)** to describe the spread.

1.1 The Median

Definition.

The **median** of a sample of data is the middle value when the data are sorted from smallest to largest.

- If the data contain an **odd number** of observations, the median is the middle observed value.
- If the data contain an **even number** of observations, the median is the average of the two middle observed values.

The median is the preferred measure of center when the data are skewed because approximately 50% of the observations lie below the median and approximately 50% lie above the median.

For skewed data, use the

$$\boxed{\text{Median}}$$

to describe the center because the median is resistant to extreme observations and outliers.

Example 1: Gas Prices

The prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, Texas, were as follows one winter day in 2018:

2.19	2.19	2.39	2.19
2.24	2.39	2.27	2.29
2.17	2.29	2.30	2.29

Find the median price for a gallon of gas and interpret the value.

Step 1: Sort the Data

Arrange the prices from smallest to largest:

$$2.17, 2.19, 2.19, 2.19, 2.24, 2.27, 2.29, 2.29, 2.29, 2.30, 2.39, 2.39.$$

There are 12 observations, so there is an even number of values.

Step 2: Identify the Two Middle Values

For 12 observations, the two middle positions are

$$\frac{12}{2} = 6$$

and

$$6 + 1 = 7.$$

The sixth and seventh values are

$$2.27 \qquad \text{and} \qquad 2.29.$$

Therefore,

$$\text{Median} = \frac{2.27 + 2.29}{2}.$$

Thus,

$$\text{Median} = \frac{4.56}{2} = \boxed{2.28}.$$

Example 2: Fat Percentage in Sliced Turkey

Below are the percentages of fat for some brands of sliced turkey:

$$14,\ 10,\ 20,\ 20,\ 40,\ 20,\ 10,\ 10,\ 20,\ 50,\ 10.$$

Find the median percentage of fat and interpret the value.

Step 1: Sort the Data

Arrange the observations from smallest to largest:

$$10,\ 10,\ 10,\ 10,\ 14,\ 20,\ 20,\ 20,\ 20,\ 40,\ 50.$$

There are 11 observations.
Because 11 is odd, there is one middle value.

Step 2: Find the Middle Observation

The middle position is

$$\frac{11 + 1}{2} = 6.$$

The sixth observation is

$$20.$$

Therefore,

$$\boxed{\text{Median} = 20\%}.$$

1.2 Range and Quartiles

Definition.
The **range** is the difference between the maximum and minimum values:

$$\boxed{\text{Range} = \text{Maximum} - \text{Minimum}}.$$

The **quartiles** divide the ordered data into approximately four equal parts.

The three quartiles are:

- Q_1: first quartile
- Q_2: second quartile, which is the median
- Q_3: third quartile

The **interquartile range (IQR)** measures approximately how much space the middle 50% of the data occupy.

$$\boxed{\text{IQR} = Q_3 - Q_1}$$

Quartile Percentage Rules

$$\boxed{\begin{array}{rl} Q_1 : & 25\% \text{ below} \\ \text{Median } (Q_2) : & 50\% \text{ below} \\ Q_3 : & 75\% \text{ below} \end{array}}$$

Therefore,

$$\boxed{\begin{array}{r} \text{Above } Q_3 = 25\% \\ \text{Between } Q_1 \text{ and } Q_3 = 50\% \\ \text{Below } Q_1 = 25\% \end{array}}$$

Example 3: Computing Quartiles by Hand

A group of eight children have the following heights, in inches:

$$48.0,\ 48.0,\ 53.0,\ 53.5,\ 54.0,\ 60.0,\ 62.0,\ 71.0.$$

Find:

- the median, also called Q_2;
- the first quartile Q_1;
- the third quartile Q_3;
- the IQR.

The data are already sorted from smallest to largest.

Step 1: Find the Median Q_2

There are 8 observations, so the median is the average of the fourth and fifth observations.
The two middle values are

$$53.5 \qquad \text{and} \qquad 54.0.$$

Therefore,

$$Q_2 = \frac{53.5 + 54.0}{2} = \frac{107.5}{2} = \boxed{53.75}.$$

Step 2: Find Q_1

The lower half of the data is

$$48.0,\ 48.0,\ 53.0,\ 53.5.$$

The median of this lower half is the average of 48.0 and 53.0:

$$Q_1 = \frac{48.0 + 53.0}{2}.$$

Therefore,

$$\boxed{Q_1 = 50.5}.$$

Step 3: Find Q_3

The upper half of the data is

$$54.0,\ 60.0,\ 62.0,\ 71.0.$$

The median of this upper half is the average of 60.0 and 62.0:

$$Q_3 = \frac{60.0 + 62.0}{2}.$$

Therefore,

$$\boxed{Q_3 = 61.0}.$$

Step 4: Find the IQR

Use

$$\text{IQR} = Q_3 - Q_1.$$

Thus,

$$\text{IQR} = 61.0 - 50.5.$$

Therefore,

$$\boxed{\text{IQR} = 10.5 \text{ inches}}.$$

Example 4: Total Energy Consumption per Capita

Data were collected on the total energy consumption per capita (in million BTUs) for some states. A summary of the data is shown below.

	Minimum	Q_1	Median	Q_3	Maximum
Total BTU	187.4	215.8	302.4	391.3	913.2

Complete parts (a) through (d).

Part (a): Percentage Above Q_3

What percentage of the states consumed more than 391.3 million BTUs per capita?
Recall that

$$Q_3$$

is the third quartile. By definition, approximately 75% of the observations are at or below Q_3, while approximately 25% of the observations are above Q_3.
Since

$$Q_3 = 391.3,$$

approximately 25% of the states consumed more than 391.3 million BTUs per capita.

Part (b): Percentage Below the Median

What percentage of the states consumed less than 302.4 million BTUs per capita?
Since

$$\text{Median} = 302.4,$$

approximately 50% of the observations lie below the median.
Therefore,

$$\boxed{50\%}.$$

Part (c): Percentage Below Q_1

What percentage of the states consumed less than 215.8 million BTUs per capita?
Since

$$Q_1 = 215.8,$$

approximately 25% of the observations lie below Q_1.
Therefore,

$$\boxed{25\%}.$$

Part (d): Percentage Between Q_1 and Q_3

What percentage of the states consumed between 215.8 and 391.3 million BTUs per capita?
The interval from Q_1 to Q_3 contains the middle 50% of the observations.
Therefore,

$$\boxed{50\%}.$$

2. §3.4 Comparing Measures of Center

The appropriate measures of center and spread depend on the shape of the distribution.

Shape	Measure of Center	Measure of Spread
Symmetric	Mean	Standard Deviation
Skewed	Median	IQR

2.1 Why Does Shape Matter?

Skewed data and outliers can have a large effect on the mean and standard deviation.

The median is **resistant** to outliers.

This means that changing the size of an extreme observation generally does not substantially change the median.

By contrast, the mean is pulled toward extreme observations.

2.2 Mean Versus Median

The relationship between the mean and median provides useful information about the shape of a distribution.

Shape	Relationship	Direction
Skewed left	Mean < Median	Mean pulled left
Symmetric	Mean = Median	No substantial pull
Skewed right	Mean > Median	Mean pulled right

Quick Memory Rule

Skewed left:	Mean < Median
Symmetric:	Mean ≈ Median
Skewed right:	Mean > Median

The mean is pulled in the direction of the tail.

2.3 Example 5: Fast-Food Restaurant Salaries

A very small fast-food restaurant has five employees, all of whom work full-time for approximately \$7 per hour.

Each employee's annual income is about

$$\$16,000.$$

The owner, on the other hand, makes

$$\$100,000$$

per year.

Find both the mean and the median.

Which would you use to represent the typical income at this business?

Step 1: Write the Data

The six annual incomes are approximately

$$16,000,\ 16,000,\ 16,000,\ 16,000,\ 16,000,\ 100,000.$$

Step 2: Find the Mean

The mean is

$$\bar{x} = \frac{\sum x}{n}.$$

The total income is

$$5(16,000) + 100,000.$$

Thus,

$$\sum x = 80,000 + 100,000 = 180,000.$$

There are 6 employees, so

$$\bar{x} = \frac{180,000}{6} = \boxed{30,000}.$$

Step 3: Find the Median

The data are already sorted:

$$16,000,\ 16,000,\ 16,000,\ 16,000,\ 16,000,\ 100,000.$$

There are 6 observations, so the median is the average of the third and fourth observations.
Both are $16,000$:

$$\text{Median} = \frac{16,000 + 16,000}{2} = \boxed{16,000}.$$

Step 4: Choose the Better Measure

The owner's income of $\$100,000$ is an extreme value compared with the other incomes.
It pulls the mean upward:

$$\$30,000 > \$16,000.$$

The median is resistant to the owner's unusually high income. Therefore, the median is the better measure of the typical income.

2.4 Comparing Two Distributions

When comparing two distributions, always use the same measures of center and spread for both distributions.

For example, it would not be appropriate to compare:

Mean of Distribution A

with

Median of Distribution B.

Likewise, we should not compare the standard deviation of one distribution with the IQR of another.

Comparison Rule
When comparing distributions:

1. Use the same measure of center for both distributions.
2. Use the same measure of spread for both distributions.
3. If either distribution is skewed, use the **median and IQR** for both distributions.