

Describing Distributions with Numbers

Diana Mindrila, Ph.D.

Phoebe Balentyne, M.Ed.

Based on Chapter 2 of *The Basic Practice of Statistics* (6th ed.)

Concepts:

- Measuring Center: Mean and Median
- Measuring Spread: Quartiles
- Five-Number Summary and Boxplots
- Spotting Suspected Outliers
- Measuring Spread: Standard Deviation
- Choosing Measures of Center and Spread

Objectives:

- Calculate and interpret mean, median, and mode
- Compare mean and median
- Calculate and interpret quartiles
- Construct and interpret the five-number summary and boxplots
- Determine suspected outliers
- Calculate and interpret standard deviation
- Choose appropriate measures of center and spread
- Organize a statistical problem

References:

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Measuring Center: The Mean

To find the **mean** \bar{x} (pronounced “x-bar”) of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, x_3, \dots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or, in more compact notation,

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

The most common measure of center is the arithmetic average, or **mean**.

- ❖ The **mean** of a data set is the average of all its values.
- ❖ Calculate the mean by finding the sum of all the values, then dividing by the number of values in the data set.
- ❖ The mean is a measure of the center, because it is an indicator of where most values are located.
- ❖ Notation for mean: \bar{x} (spoken: x bar)
- ❖ The mean is not resistant to outliers.

Example of Calculating the Mean:

Student ID	Test Score
1	91
2	89
3	89
4	88
5	95
6	100
7	94
8	30
9	92
10	93

Entire Sample:

Sum = 861

Number of Students: $N = 10$

Mean = Sum/ N

Mean = $861/10 \rightarrow$ Mean = 86.1

Sample with outlier removed:

Sum = 831

$N = 9$

Mean = 92.33

- Calculate the mean:
 - Find the sum: $91+89+89+88+95+100+94+30+92+93 = 861$
 - Divide by the number of values: $861/10$
 - Mean = 86.1
- Looking at the data after calculating the mean shows that all but one student (student 8) scored above the mean. Even without a graph, it is evident that this score does not follow the general pattern; it is an outlier.
- Remove the outlier and recalculate the mean:
 - Find the sum: $91+89+89+88+95+100+94+92+93 = 861$
 - Divide by the number of values: $831/9$
 - Mean = 92.33
- There is a considerable difference between the values of the mean with and without the outlier.
- As this example portrays, the mean is considered a nonresistant measure, because outliers can have a strong impact on its value.
- This is why it is important to use additional descriptive statistics to inform our decisions, not just the mean.

Measuring Center: The Median

Because the mean cannot resist the influence of extreme observations, it is not a **resistant measure** of center. Another common measure of center is the **median**.

The **median** M is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1. Arrange all observations from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list.
3. If the number of observations n is even, the median M is the average of the two center observations in the ordered list.

Middle Point

- ❖ Odd # of Observations: $(N+1)/2$ [Use only one value]
Example: $(9+1)/2 = 5$ > the median is the value of the 5th observation
- ❖ Even # of Observations: $(N+1)/2$ [Use two consecutive values]
Example: $(10+1)/2 = 5.5$ > find the midpoint (or average) of the 5th and 6th observations
- The **median** is a summary statistic that indicates the midpoint of a distribution.
- The following general rule can be helpful:
 - If the data set contains an odd number of observations, like 9, add one to this number and get 10. Then, divide 10 by 2, and obtain five, which means that the fifth observation is the median. In this case, only this value is used and an average does not need to be computed.
 - If the data set contains an even number of observations, like 10, add 1 to this number and get 11. Then, divide 11 by 2, and obtain 5.5. This means that the fifth and sixth observations should be used to compute the median.
- Statistical software computes the median, so it is not necessary to do it by hand, but it is important to understand how it is obtained to be able to understand what it means and how to interpret its values.

Example of Finding the Median:

Student ID	Test Score
8	30
4	88
2	89
3	89
1	91
9	92
10	93
7	94
5	95
6	100

Stem	Leaf
3	0
4	
5	
6	
7	
8	899
9	12345
10	0

$$\text{Median: } (91+92)/2 = 91.5$$

- The median can be found using a table of values or a stem-and-leaf plot. In a stem-and-leaf plot, the values are already presented in ascending order.
- In the above example, the middle values have been highlighted. Since there are two middle values, they were averaged to find the median.

Example with Outlier Removed:

Student ID	Test Score
4	88
2	89
3	89
1	91
9	92
10	93
7	94
5	95
6	100

Stem	Leaf
8	899
9	12345
10	0

$$\text{Median} = 92$$

- When there is only one middle value, it is the median. No calculation is necessary.
- It is important to note that the median value without the outlier is very close to the median with the outlier. Outliers do not have a strong impact on the median.
- The median is not as sensitive to outliers as the mean, so it is important to always report both the mean and the median, especially when data are skewed.

Measuring Center: The Mode

The **mode** is the value that occurs most often in a data set.

Example:

Student ID	Test Score
8	30
4	88
2	89
3	89
1	91
9	92
10	93
7	94
5	95
6	100

Mode = 89

- In this example, 89 is the only value that occurs more than once. It occurs twice. It has been highlighted in red. This value is the mode.

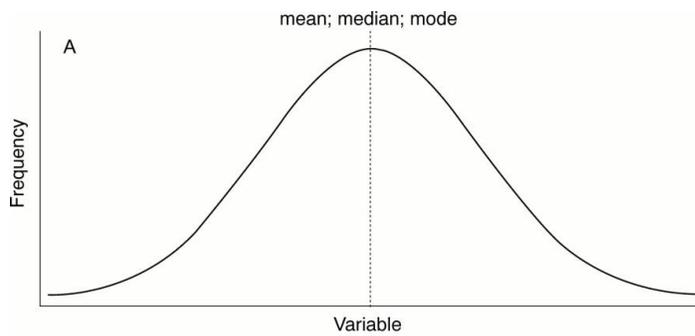
Comparing the Mean and Median

The mean and median measure center in different ways, and both are useful.

The mean and median of a roughly **symmetric** distribution are close together.

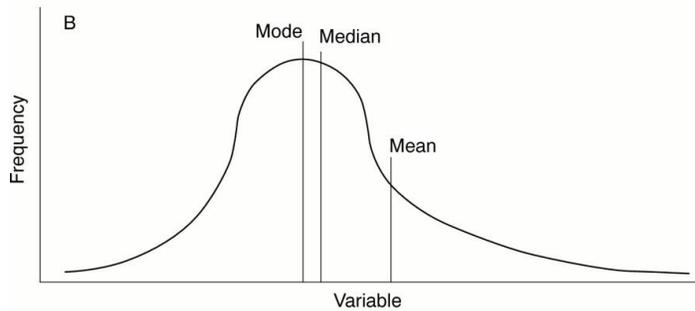
If the distribution is exactly **symmetric**, the mean and median are exactly the same.

In a **skewed** distribution, the mean is usually farther out in the long tail than the median.



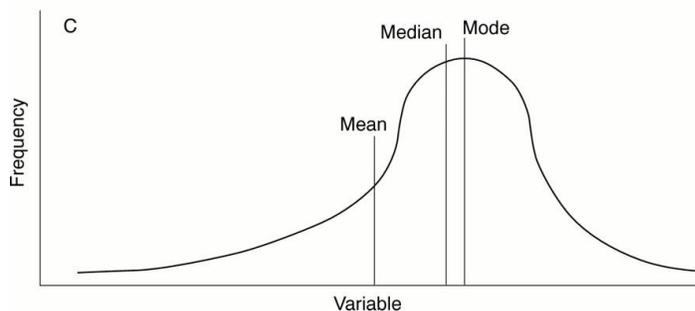
Symmetric Distribution

Mean, Median, Mode:
approximately equal



Positively Skewed Distribution

Mean larger than median



Negatively Skewed Distribution

Mean smaller than median

- Figure A displays a **symmetric** distribution. The mean, median, and mode are all approximately equal. With real data, these will not have they exact same value, but they will be very close.
- Outliers cause a **skewed** distribution resulting in a larger difference between the mean and median. When there are extreme observations, the mean is pulled in that direction.
- The **mode** is the value that occurs most often, so it is always located at the highest point in the curve.
- If the outliers in a data set are larger values, then the distribution will be positively skewed, as shown in Figure B. In this case, the mean has been pulled to the right with the extremely large values. The mean is larger than the median in a **positively skewed distribution**.
- If the outliers in a data set are smaller values, then the distribution will be negatively skewed, as shown in Figure C. In this case, the mean has been pulled to the left with the extremely small values. The mean is smaller than the median in a **negatively skewed distribution**.

Measuring Spread: Range

The **range** of a data set is the difference between the minimum and maximum recorded values.

Example of Calculating the Range:

Student ID	Test Score
1	91
2	89
3	89
4	88
5	95
6	100
7	94
8	30
9	92
10	93

Min = 30 Max = 100

Range = Max - Min

Range = 100 - 30

Range = 70

- Calculate the range: **Range = Maximum Value - Minimum Value**

Measuring Spread: Quartiles

A measure of center alone can be misleading. A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles**:

- 1) Arrange the observations in increasing order and locate the median M .
- 2) The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.
- 3) The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.

The **interquartile range (IQR)** is defined as:

$$IQR = Q_3 - Q_1$$

- Mean, median, mode, and range are helpful measures, but they are not enough.
- To provide even more information, the quartiles and interquartile range should be identified.

Example:

Student ID	Test Score	
8	30	Min
4	88	
2	89	Q1=89
3	89	
1	91	Median=91.5
9	92	(Q2)
10	93	
7	94	Q3=94
5	95	
6	100	Max(Q4)

Diagram illustrating the calculation of the Interquartile Range (IQR) from a data set. The data set is sorted, and the first quartile (Q1) is 89, the median is 91.5, and the third quartile (Q3) is 94. The IQR is calculated as Q3 - Q1 = 94 - 89 = 5. The IQR represents the spread of the middle 50% of values.

- Quartiles get their name because they each represent a quarter, or 25%, of the values in the data set.
- The first quartile, Q_1 , is found by finding the median of the values between the minimum and the median. In this case, the values in the first half of the data set are: 30, 88, 89, 89, and 91. The median of these values: $Q_1 = 89$
- The median would be the second quartile, Q_2 , but should be simply referred to as the median.
- The third quartile, Q_3 , is found by finding the median of the values between the median and the maximum. In this case, the values in the second half of the data set are: 92, 93, 94, 95, and 100. The median of these values: $Q_3 = 94$
- Each quartile contains 25% of the values in the data set, so 50% of the values are between Q_1 and Q_3 . The **range** represents the spread of the entire data set.
- The **interquartile range** represents the spread of the middle 50% of values. This value is found with the following equation: $IQR = Q_3 - Q_1$
- In this example, $IQR = Q_3 - Q_1 = 94 - 89 = 5$
- The value of Q_3 in this example is very close to the maximum. However, the value of Q_1 is very far from the minimum. This indicates that the distribution is skewed to the left.

Five-Number Summary

- The maximum and minimum values alone do not give enough information about the distribution as a whole.
- Likewise, the median and the quartiles do not give information about the tails of the distribution.
- The **five-number summary** combines all of these values to get a quick summary of both center and spread. It includes the minimum, the first quartile (Q_1), the median (Q_2), the third quartile (Q_3), and the maximum (Q_4).

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

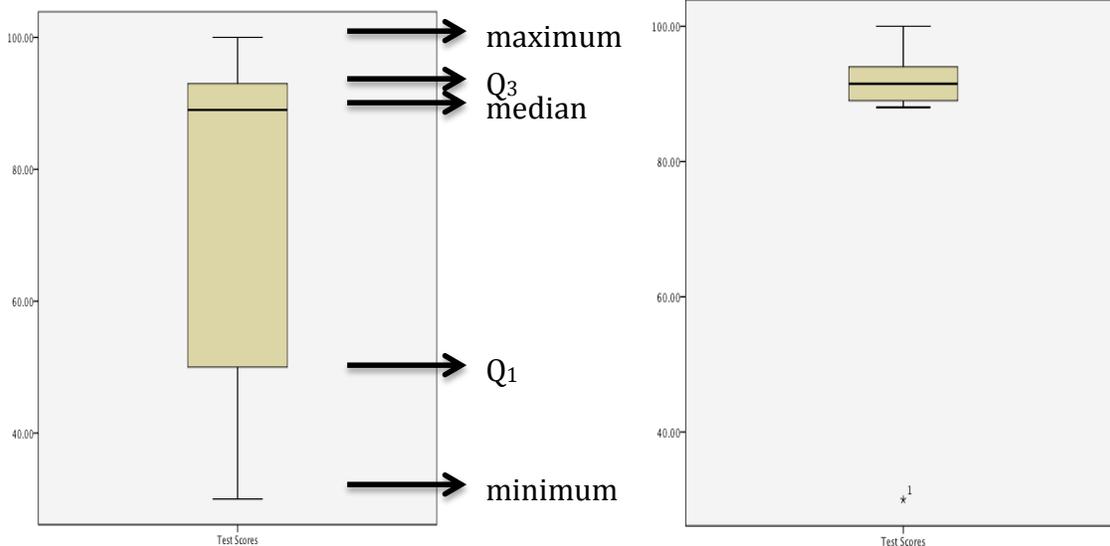
Minimum Q_1 *Median* Q_3 *Maximum*

Example:

Min	Q1	Median (Q2)	Q3	Max (Q4)
30	89	91.5	94	100

Boxplots

- The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data: the **boxplot**.



- The **boxplot**, or **box-and-whisker plot**, is a graphical representation of the five-number summary.
- The above images are examples of boxplots obtained using SPSS software. The variable of interest is on the horizontal axis and the values that variable takes are on the vertical axis.
- The box represents the interquartile range, or the values that are located between the first and third quartiles, so the box includes 50% of the sample. The line across the box displays the location of the median value. The whiskers extend from the box to the minimum and the maximum values. The whiskers help determine whether or not the data are skewed. If they have approximately the same length, the variable has a symmetric distribution. In this example, the lower whisker is much longer, so the data are negatively skewed.
- Boxplots sometimes represent the outliers as dots or little stars outside of the whiskers. In this way, extreme observations are identified, but are not considered as being the actual minimum or maximum values. The example on the right represents the data set from the previous example, but the individual with a score of 30 is considered an outlier, and the whisker starts from the next lowest value, which is 88.
- In the boxplot on the right, where the outlier has been removed, each section of the box is approximately the same size and the whiskers are

approximately the same length. This represents a more symmetric distribution.

Suspected Outliers: The 1.5 X IQR Rule

- In addition to serving as a measure of spread, the interquartile range (IQR) is also used as part of a rule for identifying outliers.

The 1.5 × IQR Rule for Outliers

An observation is considered an outlier if it falls more than $1.5 \times \text{IQR}$ above the third quartile or below the first quartile.

- A researcher may need to make a decision as to whether or not a data point is an outlier.
- The 1.5 X IQR Rule can be used as a rule of thumb to test suspected outliers.
- To test outliers:
 - Multiply the IQR by 1.5
 - Check for small outliers by subtracting $(\text{IQR} \times 1.5)$ from Q_1 . Any values below that number would be considered outliers.
 - Check for large outliers by adding $(\text{IQR} \times 1.5)$ to Q_3 . Any values above that number would be considered outliers.
- This rule helps identifying outliers to be less arbitrary.

Example:

Stem	Leaf
3	0
4	
5	
6	
7	
8	899
9	12345
10	0

$$Q_1 = 89 \quad Q_3 = 94 \quad \text{IQR} = 5$$

$$1.5 \times \text{IQR} = 1.5 \times 5 = 7.5$$

$$Q_1 - 1.5 \times \text{IQR} = 89 - 7.5 = \mathbf{81.5}$$

$$Q_3 + 1.5 \times \text{IQR} = 94 + 7.5 = \mathbf{101.5}$$

**Any score lower than 81.5 or higher than 101.5 is considered an outlier.

- In this example:
 - Values lower than 81.5 would be considered outliers.
 - Values greater than 101.5 would be considered outliers.
- Using this rule, the value of 30 can be considered an outlier.

Measuring Spread: Standard Deviation

- The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation**.

The **standard deviation** s_x measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances, and then taking the square root. This average squared distance is called the **variance**.

$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n - 1} \sum (x_i - \bar{x})^2}$$

- To find the standard deviation:
 - Compute the mean of all values.
 - Find the difference between each value in the data set and the mean. (Find out how far each value is from the mean).
 - Square each of these differences. (This will make all the values positive).
 - Calculate the sum of these values (the squares).
 - Divide the sum by $(N - 1)$ where N is the number of values in the data set. ($N - 1$ is also called the **degrees of freedom** for the data set).
 - The value obtained is the **variance**, which shows how much spread there is in the data.
 - To obtain the **standard deviation** take the square root of the variance.
- The variance and standard deviation are rarely computed by hand. This task can become extremely cumbersome with large data sets.
- Variance and standard deviation are easily obtained using statistical software, but it is important to understand their meaning.
- The standard deviation is simply the average distance from the mean.
 - A large standard deviation means the spread is large.
 - A small standard deviation means the values in the data set are clustered around the mean.
- Like the mean, the standard deviation is very sensitive to outliers. If only one difference is very large, the overall average distance from the mean increases significantly.

Calculating the Standard Deviation

Example:

Student ID	Test Score	Distance from the mean	Squared Distance
1	91	-4.9	24.01
2	89	-2.9	8.41
3	89	-2.9	8.41
4	88	-1.9	3.61
5	95	-8.9	79.21
6	100	-13.9	193.21
7	94	-7.9	62.41
8	30	56.1	3147.21
9	92	-5.9	34.81
10	93	-6.9	47.61
Mean	86.1	Sum	3608.9
		Variance (Average of Squared Distances)	400.99
		Standard Deviation (Square Root of the Variance)	20.02

- The mean in the above example is 86.1.
- First, the distance from the mean was computed for each test score by subtracting each value from the mean. These differences are displayed in the third column. Some of these differences are negative, because all of the values except 30 (the outlier) are larger than the mean.
- Second, each of the distances was squared. These values are displayed in the fourth column.
- Next, the variance was computed by finding the sum of the squares (3608.9) and dividing it by $N - 1$. Since there are 10 values in this data set: $N - 1 = 10 - 1 = 9$. Therefore, the variance = $3608.9/9 = 400.99$
- Finally, the standard deviation was obtained by taking the square root of the variance. (20.02)

Choosing Measures of Center and Spread

- Two choices for descriptions of center and spread have been discussed thus far:
 - Mean and Standard Deviation
 - Median and Interquartile Range

Choosing Measures of Center and Spread

- The median and *IQR* are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.
 - Use mean and standard deviation only for reasonably symmetric distributions that don't have outliers.
 - **NOTE: Numerical summaries do not fully describe the shape of a distribution. ALWAYS PLOT THE DATA!**
-
- Depending on the purpose of a study and the nature of the data, some measures of center and spread are more appropriate or more informative than others.
 - The mean and standard deviation are sensitive to outliers. Therefore, especially when the data are skewed, it is important to also report the median and the interquartile range and to explain their meaning.
 - Usually, it is helpful to report all of these numbers and explain the significance of each one.

Organizing a Statistical Problem

- There is a four-step process that can be used to solve statistical problems as they become more complex.

How to Organize a Statistical Problem: A Four-Step Process

State: What's the practical question, in the context of the real-world setting?

Plan: What specific statistical operations does this problem call for?

Do: Make graphs and carry out calculations needed for the problem.

Conclude: Give a practical conclusion in the setting of the real-world problem.

- The purpose of the work of a researcher is to respond to questions based on statistical evidence. In order to respond to questions and to provide evidence, researchers must conduct some statistical analyses. This is a complex process, and, in real life, it is often difficult to determine how to begin.
- The first thing that the researcher should do is to formulate a question. The question relies on his or her knowledge and experience in the field and must also take into account the constraints of reality.
- The second step is to identify the statistical procedures that will help the researcher respond to the question. Regardless of the statistical analysis that is employed, the researcher must first screen the data, which means to examine potential missing values, data entry errors, etc. Then, the researcher must always examine the data using descriptive analyses. Descriptive analyses can be presented using both graphs and numbers, depending on the nature of the data. It is up to the researcher to choose the descriptive analysis that is the most relevant to the study.
- A researcher should always examine the data using descriptive analyses prior to employing other statistical procedures and interpreting the results.