# Describing Distributions with Numbers (Page 40-43, Chapter 2)

**TODAY YOU WILL BE ABLE TO…**

* Calculate and interpret mean and median
* Compare mean and median

**NOTATION**

denotes the *addition* of a set of values (pronounced “summation”)

x the *variable* usually used to represent the individual data values

n represents the *number of values* in a *sample*

represents the *mean* of a set of *sample* values

M represents the *median* of a set of *sample* values

**MEASURE OF CENTER**

A measure of center is a value at the center or middle of a data set.

* Mean or “average”
* Median

**MEAN**

To find the **mean**, , of a set of observations, add their values and divide by the number of observations. If the *n* observations are , , ,…, , their mean is:

or in more compact notation:

***Example 1:*** Listed below are measured amounts of lead (in micrograms per cubic meter, μg/m3) in the air. The Environmental Protection Agency has established an air quality standard for lead: 15 μg/m3. The following measurements were recorded at building 5 of the World Trade Center site on different days immediately following the September 11 terrorist attacks because of concern for the quality of air. Find the mean for this sample of measured levels of lead in the air.

5.4 1.1 0.42 0.73 0.48 1.1

Because the mean involves every data value in the set, it is sensitive to extreme values. It cannot resist the influence of extreme values. Another measure of center, the median, is not calculated from the values and so is not influenced by extreme values.

**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 calculate the median…

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.

To find the location of the median, use the formula

**Location** of M =

***Example 2:*** What is the median travel time for 15 North Carolina workers? Here are the data arranged in order:

5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

***Example 3:*** What is the median travel time for 20 New York workers? Here are the data arranged in order:

5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85

The median may be a more appropriate measure of center for skewed distributions.

***Example 3:*** The measures of lead from ***example 1*** include what appears to be an extreme value. Calculate the median of the data set and it compare to the mean:

5.4 1.1 0.42 0.73 0.48 1.1

***Ordered:***

Mean:

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 is the median.

# Describing Distributions with Numbers (Page 43-49, Chapter 2)

**TODAY YOU WILL BE ABLE TO…**

* Calculate and Interpret Quartiles
* Construct and Interpret the Five-Number Summary and Boxplots
* Determine Suspected Outliers

**QUARTILES AND INTERQUARTILE RANGE**

A measure of center alone does not fully describe a distribution. Two distributions can have similar measures of center but otherwise look very different. A useful numerical description of a distribution requires both a measure of center and a measure of spread.

A **measure of spread** tells us how much a data set is spread out or scattered. We can use the range and the interquartile range (*IQR*) to measure the spread of a sample.

To calculate the quartiles:

Arrange the observations in increasing order and locate the median M.

* The **first quartile** (Q1) is the median of the observations located to the left of the overall median, M, in the ordered list.
* The **third quartile** (Q3) is the median of the observations located to the right of the overall median in the ordered list.
* The **interquartile range** (IQR) is defined as: **IQR = Q3 – Q1**

***Example 4:*** What is the spread of travel time for 15 North Carolina workers? Find the median, Q1, Q3, and IQR:

***Note:*** When there is an odd number of observations, leave out the overall median when you locate the quartiles in the ordered list.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 5 | 10 | 10 | 10 | 10 | 12 | 15 | 20 | 20 | 25 | 30 | 30 | 40 | 40 | 60 |
| Q1 = 10 | | | | | | |  | Q3 = 30 | | | | | | |
| M = 20 | | | | | | | | | | | | | | |

**IQR: 30-10=**

***Example 5:*** What is the spread of travel time for 20 New York workers? Find the median, Q1, Q3, and IQR:

***Note:*** When there is an even number of observations, include all the observations when you locate the quartiles in the ordered list.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 5 | 10 | 10 | 15 | 15 | 15 | 15 | 20 | 20 | 20 | 25 | 30 | 30 | 40 | 40 | 45 | 60 | 60 | 65 | 85 |
| Q1=(15+15)/2=15 | | | | | | | | | | Q3=(40+45)/2=42.5 | | | | | | | | | |
| M=(20+25)/2=22.5 | | | | | | | | | | | | | | | | | | | |

**IQR: 42.5-15=**

The minimum and maximum values in the data set tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of the distribution. To get a quick summary of both center and spread, combine all five numbers. The IQR is not included in the five number summary but is used in another way to evaluate the distribution (i.e., in detecting outliers).

**Minimum Q1 M Q3 Maximum**

These five number summaries offer a reasonable complete description of center and spread.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q1** | **M** | **Q3** | **Max** |
| **North Carolina** | 5 | 10 | 20 | 30 | 60 |
| **New York** | 5 | 15 | 22.5 | 42.5 | 85 |

**OUTLIERS**

An **outlier** is a data point that is extremely different from other data points. The interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

***General Rule:*** A data point is a suspected outlier if it falls more than 1.5 × IQR above the third quartile (Q3) or below the first quartile (Q1).

***Example 6:***

|  |  |
| --- | --- |
| **North Carolina Travel Times** IQR = 20 | **New York Travel Times** IQR = 27.5 |
| **30**  Q1 – ( = 10 – **30** = -20  Q3 – ) = 30 + **30** = 60  Any points not falling between -20 and 60 are suspected outliers. | **41.25**  Q1 – ( = 15 – **41.25** = -26.25  Q3 – ) = 42.5 + **41.25** = 83.75  Any points not falling between -26.25 and 83.75 are suspected outliers. |

**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 often includes the outliers.



Sample Boxplot showing spread of race times for the Flying Monkey 5K.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q1** | **M** | **Q3** | **Max** |
| **Flying Monkey** | 950 | 1325.3 | 1559.6 | 1865.73 | 3290.3 |
| **Labor Day 5K Classic** | 1015.5 | 1504.45 | 1848.7 | 2274.35 | 3994 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q1** | **M** | **Q3** | **Max** |
| **North Carolina** | 5 | 10 | 20 | 30 | 60 |
| **New York** | 5 | 15 | 22.5 | 42.5 | 85 |

North Carolina IQR = 20 New York IQR = 27.5

1.5 × 20 = 1.5 × 27.5 =

Q1 - (1.5 × IQR) = Q1 - (1.5 × IQR) =

Q3 + (1.5 × IQR) = Q3 + (1.5 × IQR) =

***Example 7:***

1. Draw and label a number line that includes the range of the distribution.
2. Draw a central box from *Q*1to *Q*3.
3. Divide the box with a line at the median *M*.
4. Extend lines (whiskers) from the box out to the minimum and maximum values that are not outliers: Q1 - (1.5 × IQR) and Q3 + (1.5 × IQR). Your whiskers cannot extend beyond the ranges of the data so if you have no outliers in one or both directions, use the min, max, or both.

***Note:*** The placement of the whiskers can vary – the text instructs you to ignore the outliers and use the min and max values, but we will use the interquartile range.

|  |
| --- |
|  |

Which state appears to have longer travel times, overall?

# Describing Distributions with Numbers (Page 49-52, Chapter 2)

**TODAY YOU WILL BE ABLE TO…**

* Calculate and Interpret Standard Deviation
* Choose Appropriate Measures of Center and Spread

**NOTATION**

denotes the *addition* of a set of values (pronounced “summation”)

x the *variable* usually used to represent the individual data values

n represents the *number of values* in a *sample*

represents the *mean* of a set of *sample* values

sx represents the *standard deviation* of the set of *sample* values represented by x

represents the *variance* of the set of *sample* values represented by x (variance is the standard deviation squared)

**STANDARD DEVIATION**

The most common measure of spread, the **standard deviation**, looks at how far each observation is from the mean. The standard deviation (sx)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**, .

***Example 8:*** The mean, measure of lead in the air from ***example 1*** is 1.5438.

|  |  |  |
| --- | --- | --- |
| **OBSERVATIONS** (μg/m3) | **DIFFERENCE FROM MEAN** (μg/m3) | **SQUARE THE DIFFERENCE** (μg/m3)2 |
| =0.42 | (0.42 - 1.5438) = -1.1183 | (-1.1183) 2 = 1.2507 |
| =0.48 | (0.48 - 1.5438) = -1.0583 | (-1.0583) 2 = 1.1201 |
| =0.73 | (0.73 - 1.5438) = -0.8083 | (-0.8083) 2 = 0.6534 |
| =1.1 | (1.1 - 1.5438) = -0.4383 | (-0.4383) 2 = 0.1921 |
| =1.1 | (1.1 - 1.5438) = -0.4383 | (-0.4383) 2 = 0.1921 |
| =5.4 | (5.4 - 1.5438) = 3.8617 | (3.8617) 2 = 14.9125 |
| **SUM** | **0.0000** | **18.3209** |

The **variance**, *average squared difference*, is computed in a way that is similar to calculating the mean of the data set. Sum the squared differences and divide the sum by (n-1).

From example 8, the variance is 18.3209 ÷ (6-1) = 3.66 (μg/m3)2

The variance is in squared units, which is not an intuitive measure for understanding the spread of the data. To change the unit measure to match that of the data, compute the square root of the variance. The result is the standard deviation.

**Variance**

or in more compact notation…

**Standard Deviation**

***Example 9 (data from Ex 2.7 p50):***

|  |  |  |
| --- | --- | --- |
| **OBSERVATIONS** (SAT units) | **DIFFERENCE FROM MEAN** (SAT units) | **SQUARE THE DIFFERENCE** (SAT units)2 |
| = 650 | ( - ) = | ( - ) 2 = |
| = 490 | ( - ) = | ( - ) 2 = |
| = 580 | ( - ) = | ( - ) 2 = |
| = 450 | ( - ) = | ( - ) 2 = |
| = 570 | ( - ) = | ( - ) 2 = |

1. Calculate the mean (): \_\_\_\_\_\_\_\_\_\_\_\_
2. Calculate the difference each point is from the mean and record in the table.
3. Square each difference and record in the table.
4. Sum the squared differences, :\_\_\_\_\_\_\_\_\_\_\_\_
5. Compute the variance, (sum of squared differences): \_\_\_\_\_\_\_\_\_\_\_\_
6. Compute the standard deviation : \_\_\_\_\_\_\_\_\_\_\_\_

**PROPERTIES OF THE STANDARD DEVIATION**

* Sx measures spread about the mean
* Sx is always greater than or equal to zero
* Sx has the same units of measurement as the original observations
* Like the mean, Sx is sensitive to outliers. A few extremely large values can make Sx large.

**CHOOSING MEASURES OF CENTER AND SPREAD**

You now have a choice between two descriptions of the center and spread of a distribution:

* Median and Five Number Summary with Interquartile Range (IQR)
* Mean and Standard Deviation

**Median and Five Number Summary/IQR**

The median and *IQR* are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.

**Mean and Standard Deviation**

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 GRAPH YOUR DATA!*