8-1 Measures of Central Tendency and Variation

Extension: Data Distributions

Essential question: *How can you use shape, center, and spread to characterize a data distribution?*

COMMON CORE **Standards for Mathematical Content**

CC.9-12.S.ID.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).*

CC.9-12.S.ID.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).*

Recall that the *mean*, *median*, and *mode* are measures of central tendency—values that describe the center of a data set.

The *mean* is the sum of the values in the set divided by the number of values. It is often represented as x.

The *median* is the middle value or the mean of the two middle values when the set is ordered numerically.

The mode is the value or values that occur most often. A data set may have one mode, no mode, or several modes.

Find the mean, median, and mode of the data set.

{6, 9, 3, 8}

Mean: $\frac{6+9+3+8}{4} = \frac{26}{4} = 6.5$

Median: 3 6 8 9 $\quad \frac{6+8}{2} = 7$

Mode: None

For numerical data, the weighted average of all of those outcomes is called the **expected value** for that experiment.

The **probability distribution** for an experiment is the function that pairs each outcome with its probability.

The probability distribution of successful free throws for a practice set is given below. Find the expected number of successes for one set.

Number of Good Free Throws, n	0	1	2	3
Prob. of n Good Free Throws	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{2}$

$\text{expected value} = 0\left(\frac{3}{20}\right) + 1\left(\frac{3}{20}\right) + 2\left(\frac{1}{5}\right) + 3\left(\frac{1}{2}\right)$ *Use the weighted average.*

$= 0 + \frac{3}{20} + \frac{2}{5} + \frac{3}{2}$ *Simplify.*

$= 0 + \frac{3}{20} + \frac{8}{20} + \frac{30}{20} = \frac{41}{20} = 2.05$

The expected number of successful free throws is 2.05.

The probability distribution of the number of accidents in a week at an intersection, based on past data, is given below. Find the expected number of accidents for one week.

Number of accidents n	0	1	2	3
Probability of n accidents	0.75	0.15	0.08	0.02

Use the weighted average.

expected value = 0(0.75) + 1(0.15) + 2(0.08) + 3(0.02)

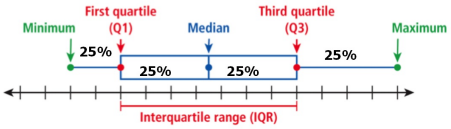
= 0.37 *Simplify.*

The expected number of accidents is 0.37.

The data sets **{19, 20, 21}** and **{0, 20, 40}** have the same mean and median, but the sets are very different. The way that data are spread out from the mean or median is important in the study of statistics.

A *measure of variation* is a value that describes the spread of a data set. The most commonly used measures of variation are the *range*, the interquartile range, the *variance*, and the *standard deviation*.

A *box-and-whisker plot* shows the spread of a data set. It displays 5 key points: the **minimum** and **maximum** values, the **median**, and the **first** and **third quartiles**.

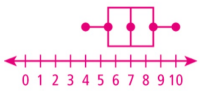


The quartiles are the medians of the lower and upper halves of the data set. If there are an odd number of data values, do not include the median in either half.

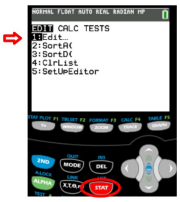
The *interquartile range*, or IQR, is the difference between the 1st and 3rd quartiles, or Q3 – Q1. It represents the middle 50% of the data.

Make a box-and-whisker plot of the data. Find the interquartile range.

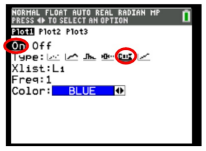
{6, 8, 7, 5, 10, 6, 9, 8, 4}



The interquartile range is 3, the length of the box in the diagram.



STAT: Edit
enter your data on L1

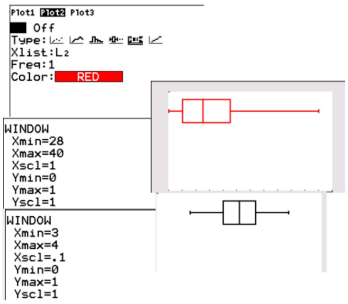


Turn on STAT Plots

Baby	Birth month	Birth weight (kg)	Mother's age
1	5	3.3	28
2	7	3.6	31
3	11	3.5	33
4	2	3.4	35
5	10	3.7	39
6	3	3.4	30
7	1	3.5	29
8	4	3.2	30
9	7	3.6	31
10	6	3.4	32
11	9	3.6	33
12	10	3.5	29
13	11	3.4	31
14	1	3.7	29
15	6	3.5	34
16	5	3.8	30
17	8	3.5	32
18	9	3.6	30
19	12	3.3	29
20	2	3.5	28

L2 L1

Make a box and whisker plot on your calculator and describe the shape of the graph. Describe the shape of Mother's age and of the birth weight.



Suppose one of the mothers' ages is chosen at random. Based on the box plot and not the original set of data, what can you say is the approximate probability that the age falls between the median, 30.5, and the third quartile, 32.5? Explain your reasoning.

25% or 0.25 because Q_1, the median, and Q_3 divide the data into four almost-equal parts.

What do you notice about the mean and median for the symmetric distribution (baby weights) as compared with the mean and median for the skewed distribution (mothers' ages)? Explain why this happens.

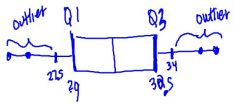
Mean and median for a symmetric distribution are equal, but mean and median for a skewed distribution are not. This happens because the mean is pulled toward the data values in the longer tail, but the median is not.

1-Var Stats
x̄=31.15
Σx=623
Σx²=19543
Sx=2.680828623
σx=2.612948526
n=20
minX=28
↓Q1=29
↑Sx=2.680828623
σx=2.612948526
n=20
minX=28
Q1=29
Med=30.5
Q3=32.5
maxX=39

1-Var Stats
x̄=3.5
Σx=70
Σx²=245.42
Sx=.1486783883
σx=.1449137675
n=20
minX=3.2
↓Q1=3.4
↑Sx=.1486783883
σx=.1449137675
n=20
minX=3.2
Q1=3.4
Med=3.5
Q3=3.6
maxX=3.8

For a data set with a first quartile of Q1 and a third quartile of Q3, a value less than Q1 – 1.5(IQR) or greater than Q3 + 1.5(IQR) may be considered to be an outlier. Use this rule to identify any outliers in each data set

Baby	Birth month	Birth weight (kg)	Mother's age
1	5	3.3	28
2	7	3.6	31
3	11	3.5	33
4	2	3.4	35
5	10	3.7	39
6	3	3.4	30
7	1	3.5	29
8	4	3.2	30
9	7	3.6	31
10	6	3.4	32
11	9	3.6	33
12	10	3.5	29
13	11	3.4	31
14	1	3.7	29
15	6	3.5	34
16	5	3.8	30
17	8	3.5	32
18	9	3.6	30
19	12	3.3	29
20	2	3.5	28



Mother's age: 39
Birth weight: none
Birth month: none

The **variance**, denoted by σ^2, is the average of the squared differences from the mean. **Standard deviation**, denoted by σ, is the square root of the variance and is one of the most common and useful measures of variation.

$$\sigma^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

$$\sigma = \sqrt{\sigma^2}$$

Low standard deviations indicate data that are clustered near the measures of central tendency, whereas high standard deviations indicate data that are spread out from the center.

Finding Variance and Standard Deviation
Step 1. Find the mean of the data, $\bar{x}$.
Step 2. Find the difference between the mean and each data value, and square it.
Step 3. Find the variance, σ^2, by adding the squares of all of the differences from the mean and dividing by the number of data values.
Step 4. Find the standard deviation, σ, by taking the square root of the variance.

Find the mean and standard deviation for the data set of the number of people getting on and off a bus for several stops.

{6, 8, 7, 5, 10, 6, 9, 8, 4}

Step 1 - Find avg/mean $\bar{x} = \frac{63}{9} = 7$

Step 2 - Find differences + square them + divide by number of values (sum)

Variance $\sigma^2 = \frac{(6-7)^2+(8-7)^2+(7-7)^2+(5-7)^2+(10-7)^2+(6-7)^2+(9-7)^2+(8-7)^2+(4-7)^2}{9}$

Standard Deviation $\sigma = \sqrt{3.3} = 1.83$

An **outlier** is an extreme value that is much less than or much greater than the other data values. Outliers have a strong effect on the mean and standard deviation. If an outlier is the result of measurement error or represents data from the wrong population, it is usually removed. There are different ways to determine whether a value is an outlier. One is to look for data values that are more than 3 standard deviations from the mean.

Find the mean and the standard deviation for the heights of 15 cans. Identify any outliers, and describe how they affect the mean and the standard deviation.

Can Heights (mm)		
92.8	92.8	92.9
92.9	92.9	92.8
92.7	92.9	92.1
92.7	92.8	92.9
92.9	92.7	92.8

$\bar{x} = 92.77$

$\sigma = (.195) \times 3 = .585$

92.77 + .585 = 92.355

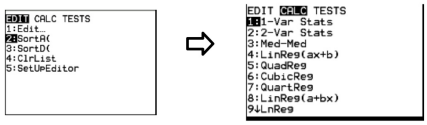
92.770 − .585 = 92.185

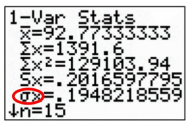
92.185 — 92.355

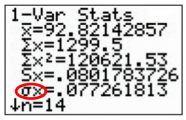
92.77

All Data Without outlier

Sort your data and remove the outlier







The outlier in the data set causes the mean to decrease from 92.82 to 92.77 and the standard deviation to increase from ≈ 0.077 to ≈ 0.195.

In the 2003-2004 American League Championship Series, the New York Yankees scored the following numbers of runs against the Boston Red Sox: 2, 6, 4, 2, 4, 6, 6, 10, 3, 19, 4, 4, 2, 3. Identify the outlier, and describe how it affects the mean and standard deviation.

The mean is about 5.4, and the standard deviation is about 4.3.

Three standard deviations is about 3(4.3) = 12.9.

Values less than −7.5 and greater than 18.3 are outliers, so 19 is an outlier.

The outlier in the data set causes the mean to increase from ≈ 4.3 to ≈ 5.4, and the standard deviation increases from ≈ 2.2 to ≈ 4.3.

Which measures of center and spread would you report for the symmetric distribution? For the skewed distribution? Explain your reasoning.

Report either mean and standard deviation or median and IQR for symmetric, but use only median and IQR for skewed because mean and standard deviation are too sensitive to the data values in the long tail.