Week Two: Z-scores and the Empirical Rule

Week Two Goals

- Learn about Z-scores.
- Learn about Percentiles.
- Learn about The Empirical Rule.
- Compare data sets.
- Learn about Stacked and Unstacked Data.
- Instructional videos using StatHelper and Minitab 19.

Z-Score

A z-score describes the position of a data point in terms of its distance from the mean, when measured in standard deviation units.

Positive Z-score

If a data point has a positive z-score, it means that the data point is larger than the mean. A z-score of +1.35 indicates that the data point you are studying is 1.35 standard deviations above the mean.

Negative Z-score

A negative z-score means that the data point is smaller than the mean. A z-score of -2.55 indicates that the data point you are studying is 2.55 standard deviations below the mean.

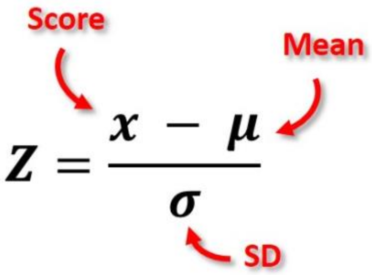
Z-score Formula

If we are working with sample data, then we can calculate the sample z-score for any data point using this formula:

$$z = \frac{x - \bar{x}}{s}$$

If we are working with population data, then we can calculate the z-score for any data point using the same formula but with different symbols to represent population values:

$$z = \frac{x - \mu}{\sigma}$$



Z-scores are always rounded to the nearest hundredth.

Unusual Data Point versus Outlier

Data points that are more than 2 standard deviations above or below the mean are considered UNUSUAL. Therefore, z-scores that are greater than +2.00 or less than -2.00 indicate that the data point is unusual. If we get a z-score that is less than 2, then we can consider the data point not unusual for the sample. A very small Z-score would indicate the data point is typical for the data set.

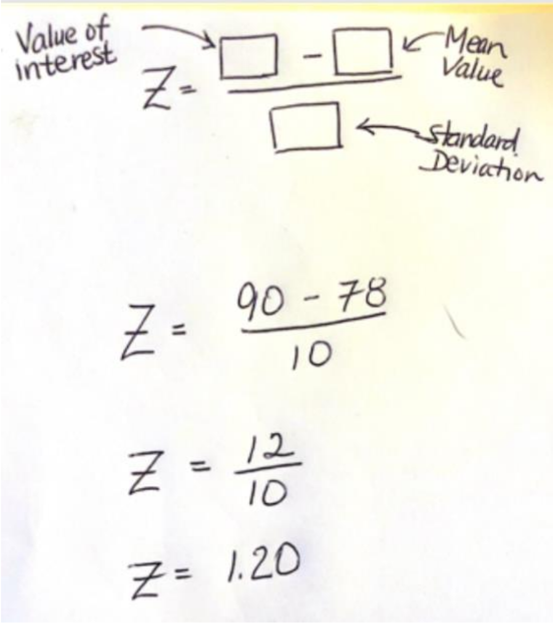
Data points that are more than 3 standard deviations above or below the mean can be called OUTLIERS.

Z-Score Examples

Example 1: A random sample of Applied Calculus final exam scores is taken and the mean determined to be 78 with a standard deviation of 10. If you received a 90, what is the z-score for your final exam score? Interpret the z-score in a complete sentence. Is your score unusual? Explain.

ANSWER:

The z-score would calculate to be +1.20 (see handwritten work below). This means that your exam score is 1.2 standard deviations above the mean final exam score of 78. Your exam score of 90 is **not** considered unusual for this sample since your score is not more than 2 z-scores from the mean final exam score. In other words, your exam score is not more than 2 standard deviations from the mean final exam score.

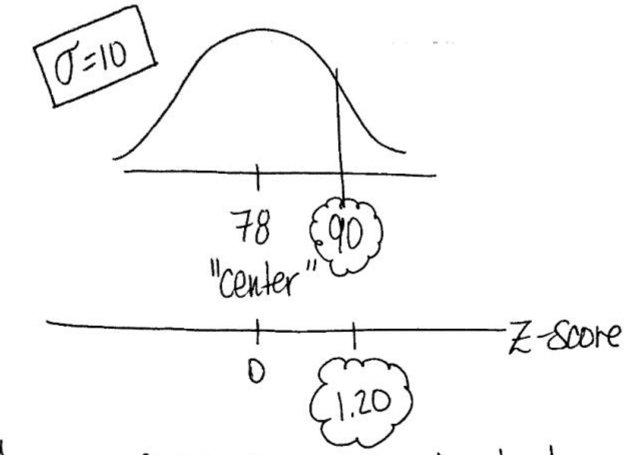


$$Z = \frac{\square - \square}{\square}$$

$$Z = \frac{90 - 78}{10}$$

$$Z = \frac{12}{10}$$

$$Z = 1.20$$



A score of "90" is 1.2 standard deviations above (to the right of) the mean score of 78.

Video Explaining Z-Scores from Prof. Coffey

Here is a short video discussing Z-scores: https://youtu.be/xnwe6wFJtCQ

Here is the same video with no sound and INTERPRETING: https://youtu.be/jOQRKMBS1pc

Example 2: The results of a cognitive abilities test for patients with Alzheimer's disease have been shown to be normally distributed with a mean of 52 and a standard deviation of 5. A patient recently diagnosed with Alzheimer's disease takes a cognitive abilities test and scores a 44.5. Calculate the patient's Z-score. Interpret the z-score in a complete sentence. Is your score unusual? Explain.

ANSWER:

z = (44.5 – 52) / 5 = -1.50

The recently diagnosed patient's cognitive abilities test score is 1.5 standard deviations below the mean score. This score is not unusual since the z-score is less than 2.

Example 3:

Over 6-months in 2018, Rochester, NY gas prices had a mean of $2.73 per gallon and a standard deviation of $0.16 per gallon. Consider the gas price I saw near my house: $2.67/gallon

What is the z-score for $2.67 per gallon? Interpret the meaning of this z-score. Is this price per gallon *unusual*?

ANSWER:

$$z = \frac{2.67 - 2.73}{0.16} = -0.38$$

(round to nearest hundredth)

A price per gallon of $2.67 is .38 standard deviations BELOW the mean price per gallon of $2.73.

This price per gallon is NOT unusual since the z-score is not past +/- 2. In fact, we would consider this $2.67 gas price to be typical for the sample.

A friend was visiting from Akron, OH and stated that she is used to paying $2.36 per gallon of gas in Akron. Is this price per gallon unusual for Rochester's average?

ANSWER: Z-score for the friend from Akron, OH:

$$z = \frac{2.36 - 2.73}{0.16} = -2.31$$

A price per gallon of $2.36 is 2.31 standard deviations BELOW the mean Rochester price per gallon of $2.73.

This price per gallon is unusual since the z-score is past +/- 2.

Finding a Value given a Z-Score

If you are provided a z-score, along with the mean and standard deviation, we can use Algebra to solve the specific data value that corresponds to that z-score.

$$z - score = \frac{x - \mu}{\sigma}$$

Example: The monthly utility bills in a city are normally distributed with a mean of $70 and a standard deviation of $8. Find the monthly utility bill that corresponds to a z-score of -0.75. View this YouTube video for worked out solution.

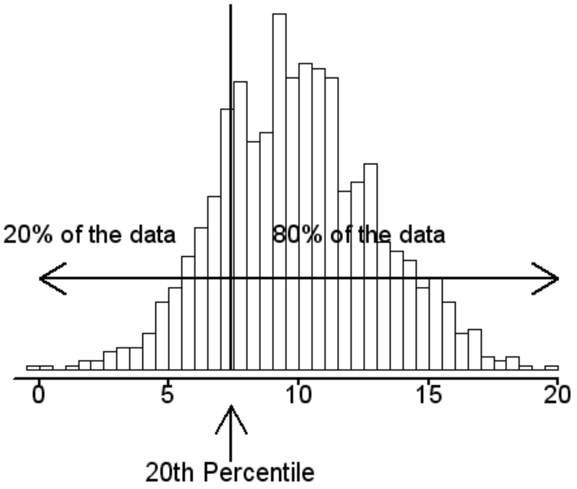
ANSWER: A z-score of -0.75 corresponds with a $64 monthly utility bill.

$$-0.75 = \frac{x - 70}{8}$$
$$(-0.75)(8) = x - 70$$
$$(-0.75)(8) + 70 = x$$
$$-6 + 70 = x$$
$$64 = x$$

Percentile

A percentile is a measure used in statistics indicating the value ***below which*** a given percentage of observations in a group of observations falls.

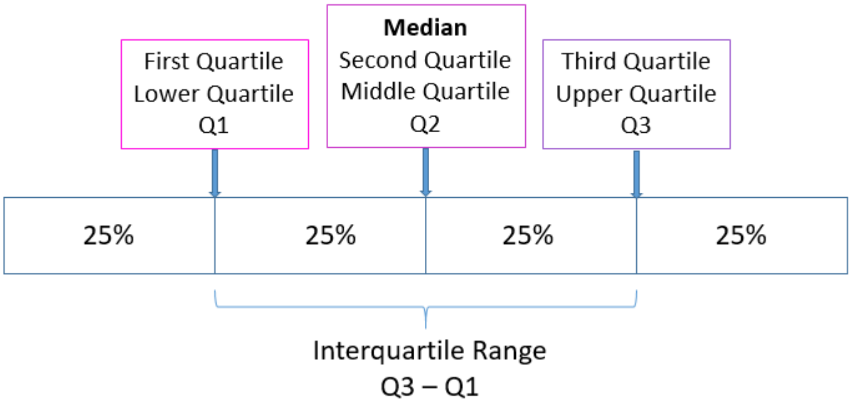
For example, the 20th percentile is the value below which 20% of the observations may be found. The 20th percentile can also be stated as the value ***above which*** 80% of the observations may be found.



If you know that your SAT score is in the 90th percentile, that means **you** scored better than 90% of people who took the SAT (and **you** scored worse than 10% of people who took the test). A percentile helps you locate where the data point lies relative to the sample/group.

Recall, we used quartiles in building a modified box plot.
The first quartile (Q1) is the 25th percentile.
The second quartile (Q2) is the 50th percentile; we call it the MEDIAN.
The third quartile (Q3) is the 75th percentile.

Median and Quartiles



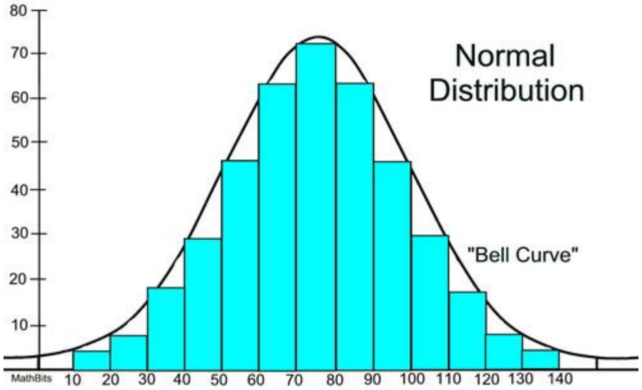
Video Reviewing Percentiles from Prof. Coffey

Here is a short video reviewing percentiles: https://youtu.be/C-Mh_nwn0Sk

Here is the same video with no sound and INTERPRETING: https://youtu.be/_6hqH7EOqXQ

The Normal Curve

When the distribution of data appears symmetric and bell-shaped, we will refer to it as a Normal Distribution or a Normal Curve. If a curve were to be drawn on a symmetric histogram, it might look like this:



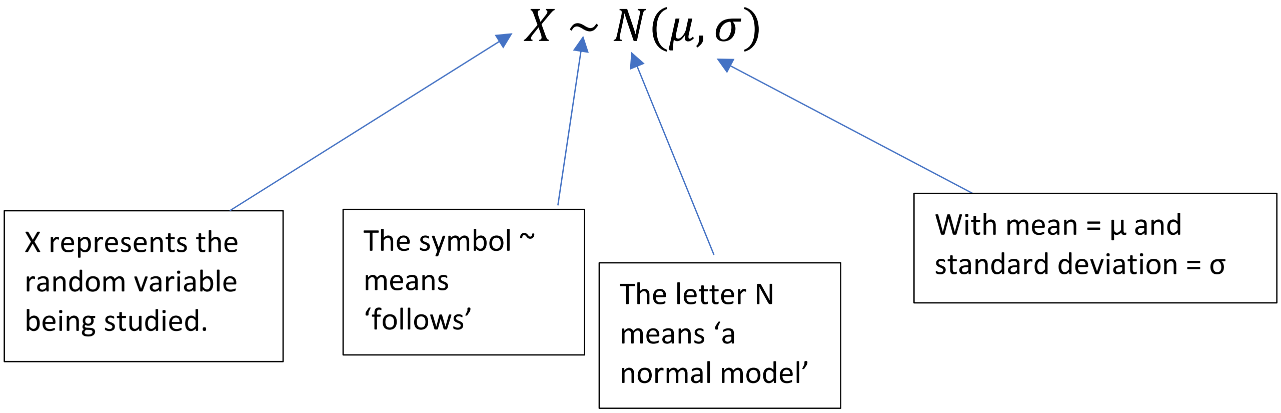
Calculus can be used to study the area under a bell-shaped, symmetric curve and, as a result, we are able to determine the approximate areas of different regions under the normal curve. These areas are provided to us as percentages and are considered the Empirical Rule intervals. These percentages are approximate and should only be used when we are told or can determine that the data set with which we are working can be assumed "Normal".

When we work with Z-scores, we are no longer working in the units of the problem; we are working with a standardized value. In this case, we call the curve a **Standard Normal Curve**.

When we work with the normal or standard normal distribution, you will need to know two numerical descriptive measures: the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing:

$$X \sim N(\mu, \sigma).$$

If you see such notation, it is a short-hand way of saying: We are studying this variable called x and it follows a Normal Distribution with a mean value, μ, and standard deviation, σ.



$$X \sim N(\mu, \sigma)$$

X represents the random variable being studied.

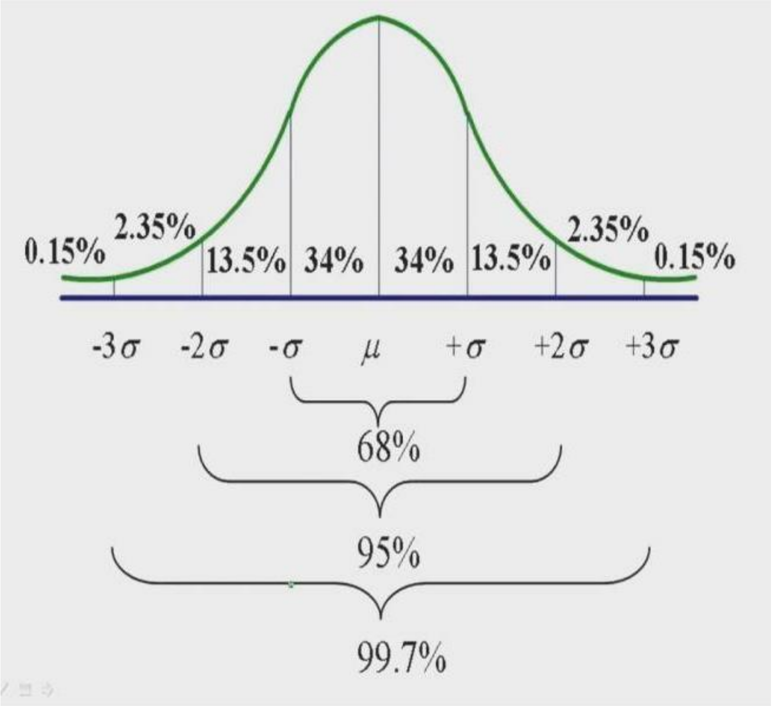
The symbol ~ means 'follows'

The letter N means 'a normal model'

With mean = μ and standard deviation = σ

The Empirical Rule Intervals

The Empirical Rule states that if you place the mean in the center of a somewhat symmetric distribution and add/subtract one standard deviation value from the mean, you will have approximately 68% of your data.



If you add/subtract two standard deviation values from the mean, you will have approximately 95% of your data. A data point that falls beyond the middle 95% is considered ***unusual*** for the data set.

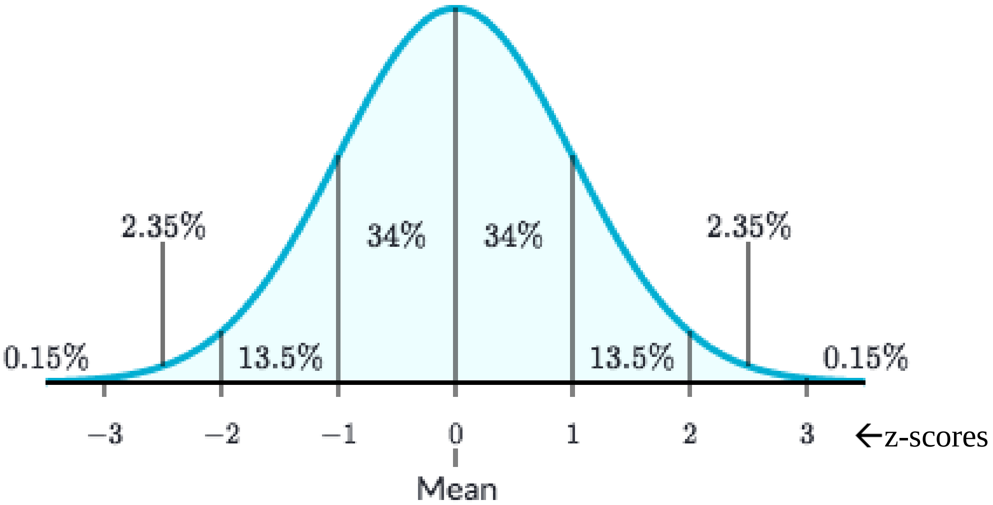
If you add/subtract three standard deviation values from the mean, you will have approximately 99.7% of your data.

The remaining 0.3% of the data (0.15% in the left tail and 0.15% in the right tail) is considered more than 3 standard deviations away. A data point that falls this far away from the mean is considered an ***outlier***.

Unusual versus Outlier

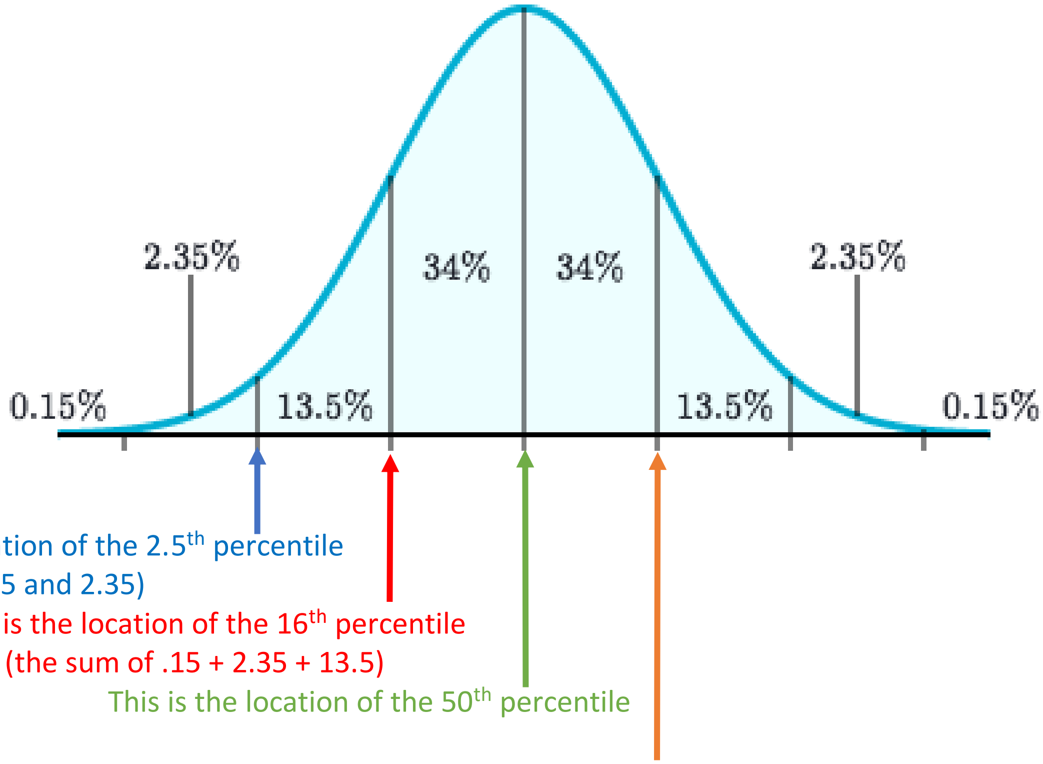
When we know where a data point lies on the empirical rule intervals, we can use the term unusual when the data point is outside of the middle 95% region; in other words, more than 2 standard deviations from the mean. We can use the term outlier when the data point is outside the middle 99.7% region; in other words, more than 3 standard deviations from the mean.

The breakdown of the empirical rule percentages can also be seen compared to z-scores, where a z-score of ZERO lies in the center of the distribution.
One standard deviation above the center corresponds to a z-score of +1. One standard deviation below the center corresponds to a z-score of -1.
Two standard deviations above the center corresponds to a z-score of +2. Two standard deviations below the center corresponds to a z-score of -2, etc.



When you are studying the distribution of data with empirical rule intervals and percentages, the middle 95% of your data represents data points that are **NOT** unusual for the data set.

The data points beyond 2 standard deviations from the mean are considered UNUSUAL.
When you add up the percentages of the empirical rule intervals beginning from the left, you can find the corresponding PERCENTILES.



This is the location of the 2.5th percentile
(the sum of .15 and 2.35)

This is the location of the 16th percentile
(the sum of .15 + 2.35 + 13.5)

This is the location of the 50th percentile

This is the location of the 84th percentile, etc.

Video Explaining the Empirical Rule from Prof. Coffey

Here is a short video discussing the Empirical Rule: https://youtu.be/ODt4yiGCEvs

Here is the same video with no sound and INTERPRETING: https://youtu.be/MVyeXyKySiM

Empirical Rule and Percentile Examples

Expenditures Example

The National Retail Federation reported that first-year college students spend more on back-to-school items than any other college group (USA Today 8/4/2006). Back-to-school expenditures ($) for a sample of twelve first-year students are provided below:

1479	983	1133	1286	1409	1259	1579	1113	945	1354	1255	1121

A. The Empirical Rule is more accurate when the data is "bell-shaped". Produce a histogram using the data above and determine if the back-to-school expenditures data appear to be somewhat "bell-shaped"?

B. Draw a bell-shaped curve and label the center with the mean expenditure from earlier ($1243) and a standard deviation of $193.3. In short-hand, we could tell you this same information by stating $X \sim N(1243, 193.3)$. Label the empirical rule intervals and percentages.

C. Fill in the blanks: About 68% of the back-to-school expenditures in this sample are between $____ and $____.

D. What percentage of the back-to-school expenditures in this sample are between $856.4 and $1629.6?

E. Find the back-to-school expenditure that represents the 84th percentile.

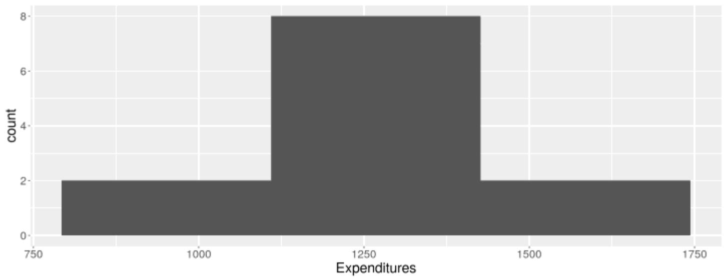
F. Fill in the blank: $1049.70 represents the _______ percentile.

G. Are any of the back-to-school expenditure data points considered unusual?

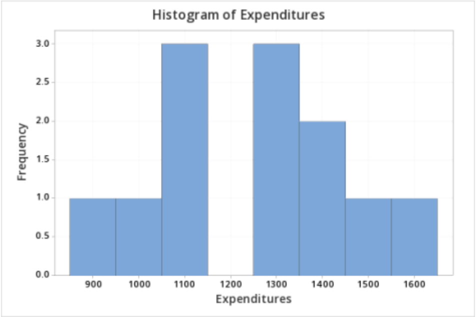
ANSWER:

Use technology to build the histogram. Depending on the width of the histogram bars--which affects the number of bins-- each of you may have a different histogram. That is ok. Here are two that I produced. I would say that the data appear somewhat bell-shaped. Yes, we can use the empirical rule. Do you agree? Does your histogram look somewhat bell-shaped? If so, let's assume a bell-shaped distribution and answer the following questions with the empirical rule intervals.

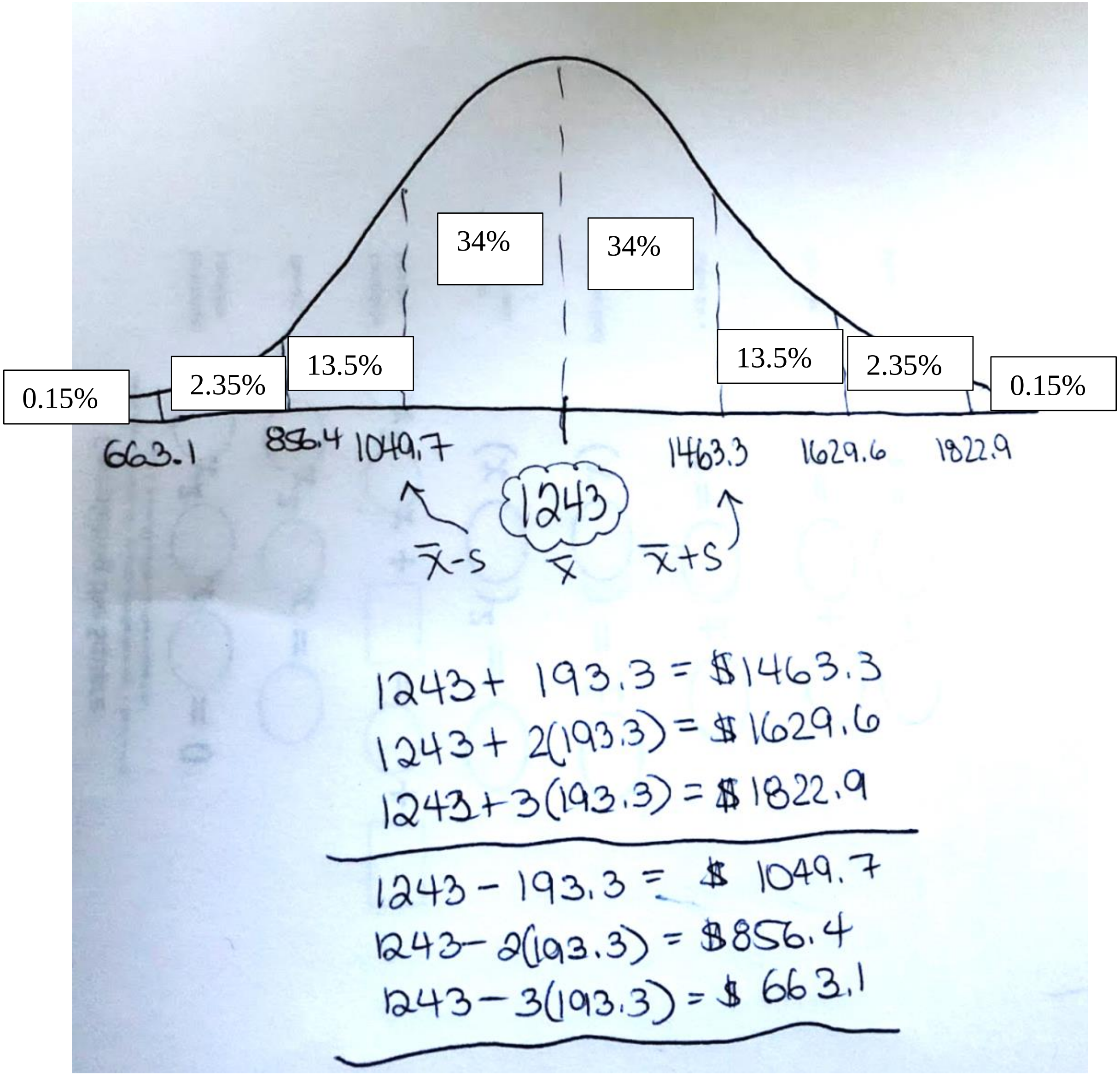
This is the histogram that StatHelper built:



This is the histogram that Minitab 19 built:



A. Draw a bell-shaped curve and label the center with the mean expenditure from earlier ($1243) and a standard deviation of $193.3. Label the empirical rule intervals and percentages.



$$1243 + 193.3 = \$1463.3$$
$$1243 + 2(193.3) = \$1629.6$$
$$1243 + 3(193.3) = \$1822.9$$
$$1243 - 193.3 = \$\ 1049.7$$
$$1243 - 2(193.3) = \$856.4$$
$$1243 - 3(193.3) = \$\ 663.1$$

B. Fill in the blanks: About 68% of the back-to-school expenditures in this sample are between $____ and $____.

ANSWER: $1049.7 and $1463.3

C. What percentage of the back-to-school expenditures in this sample are between $856.4 and $1629.6?

ANSWER: 95%

D. Find the back-to-school expenditure that represents the 84th percentile.

ANSWER: $1463.3

E. Fill in the blank: $1049.70 represents the _______ percentile.

ANSWER: 16th percentile

F. Are any of the back-to-school expenditure data points considered unusual?

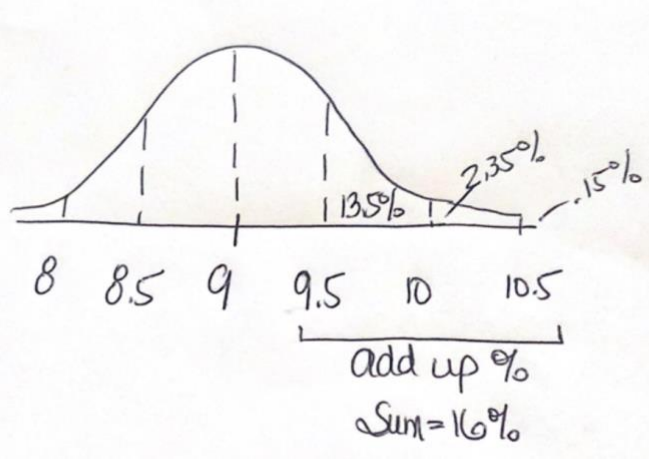
ANSWER: No, since all of the data points in the sample are within 2 standard deviations of the mean amount of $1243.

Basketball Diameter Example

If the diameter of a basketball is normally distributed, with a mean (μ) of 9″, and a standard deviation (σ) of 0.5″, what is the probability that a randomly chosen basketball will have a diameter between 9.5″ and 10.5″.

NOTE: In short-hand we could say the same question like this: if X = the diameter of a basketball and $X \sim N(9, 0.5)$, find P (9.5 ≤ X ≤ 10.5).

ANSWER:
The percentage that a randomly chosen basketball will have a diameter between 9.5 and 10.5 inches is 16%. That makes the probability 0.16.



Comparing Data

When we compare data sets, we can compare shape, center and spread. If we are able to assume a somewhat symmetric, bell-shaped distribution, then we can use the Empirical Rule intervals. Z-scores can be especially helpful when comparing since you are not working with the raw data, but rather the standardized z-score values.

If you want to compare two data sets and you have the raw data, first look over the data to determine if you have unstacked data or stacked data.

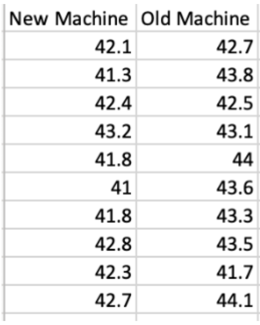
Learn About Stacked and Unstacked Data

Unstacked Data is when you have more than one column of the variable you are studying.

Stacked Data is when the variable you are studying is all in one column and a second variable can be used to separate your data into separate groups.

Here is an example to illustrate: In a packing plant, a machine is used to pack cartons with jars. A new machine has been introduced and we are interested in comparing the packing times with the old machine. The times (in seconds) that it takes each machine to pack ten cartons are recorded below.

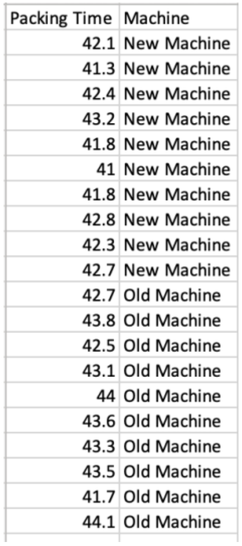
The variable we are studying is numerical and it represents the time to pack ten cartons (in seconds). When the numerical data is provided in two separate columns, it is considered **Unstacked Data**.



New Machine	Old Machine
42.1	42.7
41.3	43.8
42.4	42.5
43.2	43.1
41.8	44
41	43.6
41.8	43.3
42.8	43.5
42.3	41.7
42.7	44.1

NOTE: This is the exact same data. If your data is unstacked, then you would run the summary statistics for each column. You can build two separate box plots. If your data is stacked, you can build a comparative box plot.

When the numerical data is provided in one column and a second column has a different variable, a variable that can separate the data into two columns, then we have **Stacked Data**.



Packing Time	Machine
42.1	New Machine
41.3	New Machine
42.4	New Machine
43.2	New Machine
41.8	New Machine
41	New Machine
41.8	New Machine
42.8	New Machine
42.3	New Machine
42.7	New Machine
42.7	Old Machine
43.8	Old Machine
42.5	Old Machine
43.1	Old Machine
44	Old Machine
43.6	Old Machine
43.3	Old Machine
43.5	Old Machine
41.7	Old Machine
44.1	Old Machine

Statistical Software and Comparing Data Sets

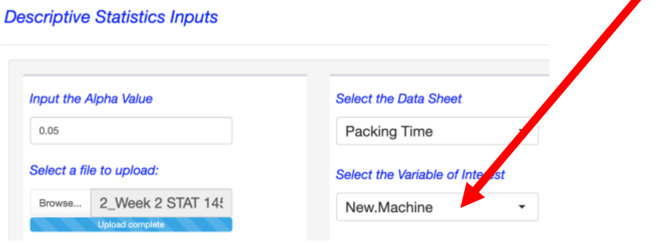
StatHelper Instructions

Video Explaining how StatHelper can work with Stacked and Unstacked Data from Prof. Coffey

Here is a short video discussing StatHelper: https://youtu.be/M0yJ8cNP1GU

The unstacked data are in the 2_Week 2 STAT 145 Data.xlsx, sheet labeled **Packing Time**, there one variable labeled **New Machine** and another variable labeled **Old Machine**.

If we want to get summary statistics for both the new machine and the old machine packing times, we would run them twice.



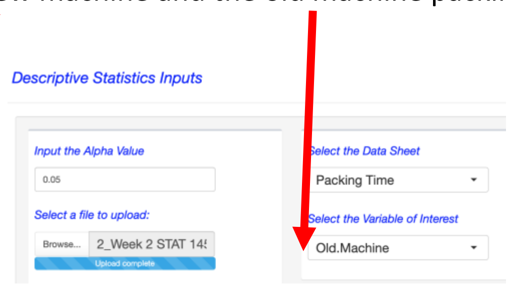
Descriptive Statistics Inputs

Input the Alpha Value: 0.05

Select a file to upload: Browse... 2_Week 2 STAT 14

Select the Data Sheet: Packing Time

Select the Variable of Interest: New.Machine



Descriptive Statistics Inputs

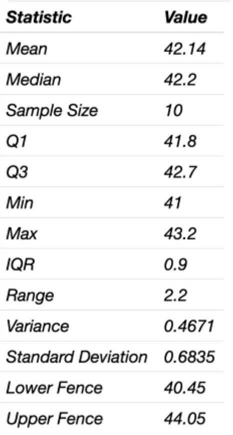
Input the Alpha Value: 0.05

Select a file to upload: Browse... 2_Week 2 STAT 14

Select the Data Sheet: Packing Time

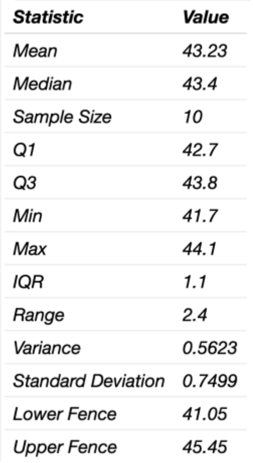
Select the Variable of Interest: Old.Machine

New Machine

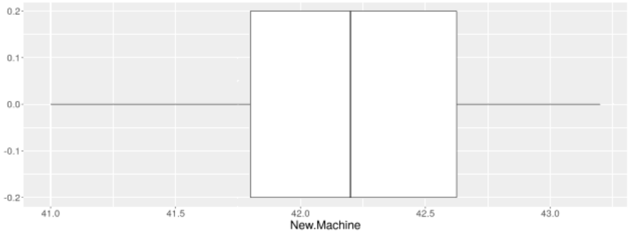


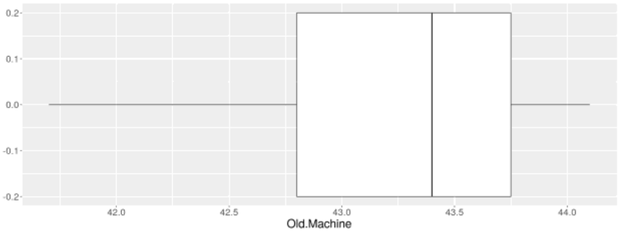
Statistic	Value
Mean	42.14
Median	42.2
Sample Size	10
Q1	41.8
Q3	42.7
Min	41
Max	43.2
IQR	0.9
Range	2.2
Variance	0.4671
Standard Deviation	0.6835
Lower Fence	40.45
Upper Fence	44.05

Old Machine



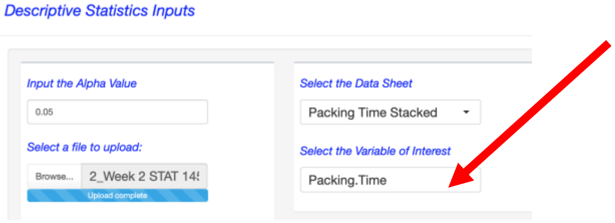
Statistic	Value
Mean	43.23
Median	43.4
Sample Size	10
Q1	42.7
Q3	43.8
Min	41.7
Max	44.1
IQR	1.1
Range	2.4
Variance	0.5623
Standard Deviation	0.7499
Lower Fence	41.05
Upper Fence	45.45





Comparative Box Plots are box plots that are built on the same set of axes. If you are comparing data sets, it is much better to work with comparative box plots. You can compare the median line easily, compare the size of the IQR rectangle, etc.

To build comparative box plots with StatHelper, the data must first be **Stacked**. Then, the two box plots will be built on the same axes. The stacked data are in the 2_Week 2 STAT 145 Data.xlsx, sheet labeled **Packing Time Stacked**, the numerical variable is **Packing Time** and the categorical variable is **Machine**. The variable of interest is the packing time.


Descriptive Statistics Inputs

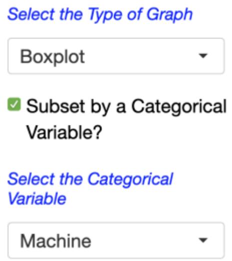
Input the Alpha Value: 0.05

Select a file to upload: Browse... 2_Week 2 STAT 14!
Upload complete

Select the Data Sheet: Packing Time Stacked

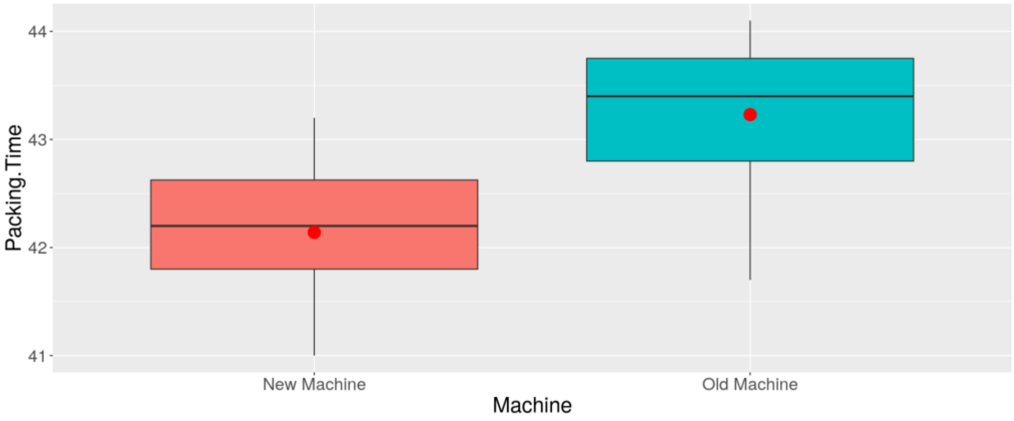
Select the Variable of Interest: Packing.Time

Once you click **RUN**, then go to **Graphs→Select type of graph→ Box Plot** and Subset by a Categorical Variable **→Machine**


Select the Type of Graph: Boxplot

☑ Subset by a Categorical Variable?

Select the Categorical Variable: Machine



<u>**Once created, we are able to describe and compare the new and old machines:**</u>

Shape: The shape of the new machine data appears somewhat symmetric (since the mean is approximately equal to the median). The old machine data appears slightly skewed negatively/left (since the mean is less than the median)

Center: The packaging time for the new machine appears lower than the old machine (since the center--mean and median—are lower for the new machine.

Spread: The spread of the data appears slightly greater for the old machine (since the size of the rectangle is slightly larger—meaning the IQR is slightly greater for the old machine.

Minitab 19 Instructions

Video Explaining how Minitab can work with Stacked and Unstacked Data from Prof. Coffey

Here is a short video discussing Minitab 19: https://youtu.be/P9DxA5Ud92I

Unstacked Data with Minitab

The unstacked data are in the 2_Week 2 STAT 145 Data.xlsx, sheet labeled **Packing Time**, there one variable labeled **New Machine** and another variable labeled **Old Machine**. You can copy and paste them into Minitab 19 as needed. If you are working with unstacked data, then use the following instructions to get the summary statistics you may need:
Stat→Basic Statistics→Display Descriptive Statistics; Bring in both variables (new and old) in the box labeled **Variables**.

Stacked Data with Minitab

The stacked data are in the 2_Week 2 STAT 145 Data.xlsx, sheet labeled **Packing Time Stacked**, the numerical variable is **Packing Time** and the categorical variable is **Machine**. The variable of interest is the packing time.

If you are working with stacked data, then use the following instructions to get the summary statistics you may need.
Stat→Basic Statistics→Display Descriptive Statistics; Where it says **Variable**: bring in the numerical variable we are studying and where it says **By Variable**: bring in the categorical variable that separates into groups.

Comparative Box Plots with Minitab

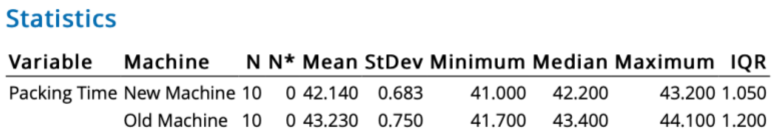
Comparative Box Plots are box plots that are built on the same set of axes. If you are comparing data sets, it is much better to work with comparative box plots. You can compare the median line easily, compare the size of the IQR rectangle, etc.
To build comparative box plots with Minitab, the data must first be **Stacked**. Then, the two box plots will be built on the same axes.

Examples Comparing Data Sets

Example 1:

In a packing plant, a machine is used to pack cartons with jars. A new machine has been introduced and we are interested in comparing the packing times with the old machine. The times (in seconds) that it takes each machine to pack ten cartons are recorded below. The packing times for two different machines can be found in the file: 2_Week 2 STAT 145 Data.xlsx, sheet labeled **Packing Time**, there one variable labeled **New Machine** and another variable labeled **Old Machine**. It can also be found in the sheet labeled **Packing Time Stacked**. (see information above on stacked and unstacked data). Describe and compare the two machines.

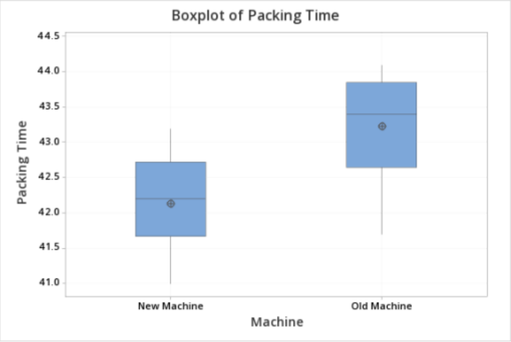
Minitab Summary Statistics:



Statistics

Variable	Machine	N	N*	Mean	StDev	Minimum	Median	Maximum	IQR
Packing Time	New Machine	10	0	42.140	0.683	41.000	42.200	43.200	1.050
	Old Machine	10	0	43.230	0.750	41.700	43.400	44.100	1.200

Minitab Comparative Boxplots:



ANSWER:

Once created, we are able to describe and compare the new and old machines:

Shape: The shape of the new machine data appears somewhat symmetric (since the mean is approximately equal to the median). The old machine data appears slightly skewed negatively/left (since the mean is less than the median).

Center: The packaging time for the new machine appears lower than the old machine (graphically and numerically this is visible: since the mean and median for the new machine are lower than for the old machine).

Spread: The spread of the data appears slightly greater for the old machine (graphically: since the size of the rectangle is slightly larger—meaning the IQR is slightly greater for the old machine. Numerically: this is also seen by comparing the IQR or standard deviation).

Example 2:

In international swimming, the mean time for the men's 100 m freestyle is 50.46 seconds (sec) with a standard deviation of 0.6 sec. For the 200 m freestyle, the mean time is 110.4 sec with a standard deviation of 1.4 sec. Jeff's best time for the 100 m is 48.76 sec and for the 200 m, his best time is 108.43 sec.

If he can only enter one of these events in the competition, which one should he enter?

ANSWER: Calculate the Z-score for Jeff's 100 m and 200 m best swim times. The Z-score that is further from zero will represent the event that Jeff is better at, relative to the mean times.

100 m

$$\begin{aligned} z &= \frac{x-\bar{x}}{s} \\ &= \frac{48.76-50.46}{0.6} \\ &= -2.83 \end{aligned}$$

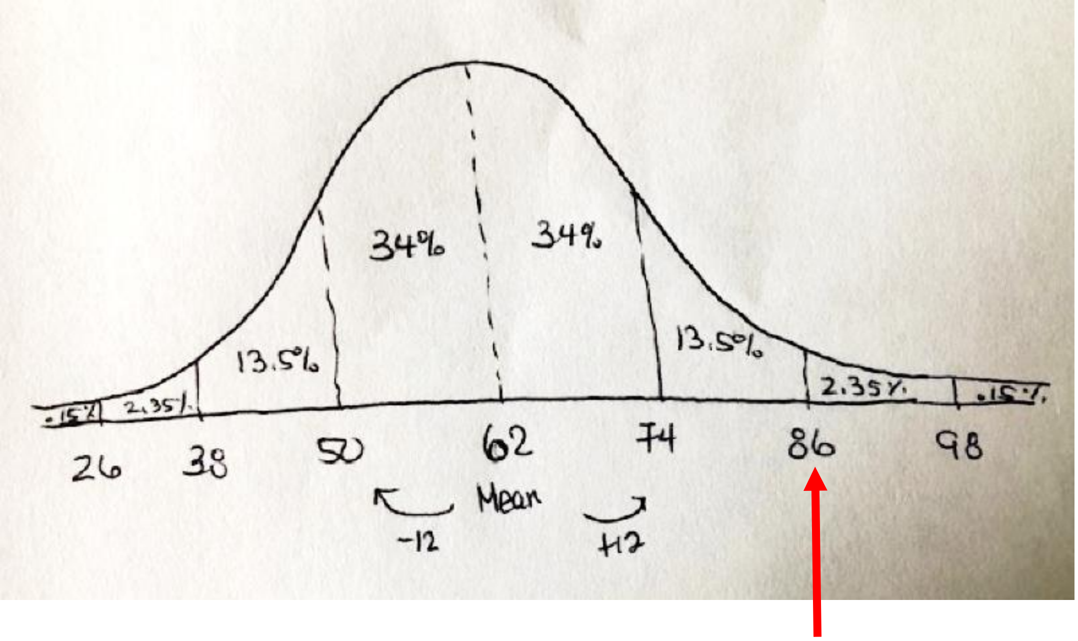
200 m

$$\begin{aligned} z &= \frac{x-\bar{x}}{s} \\ &= \frac{108.43-110.4}{1.4} \\ &= -1.41 \end{aligned}$$

Based on the z-score results, Jeff's z-score for the 100 m is lower, indicating that his time is ***further below the mean for this event*** than for the 200 m event. **Jeff should enter the 100 m event!**

Example 2: On an end-of-semester math test out of 100 points, the mean result was 62 points and the standard deviation was 12 points. What percentage of the results would lie above a test score of 86 points?

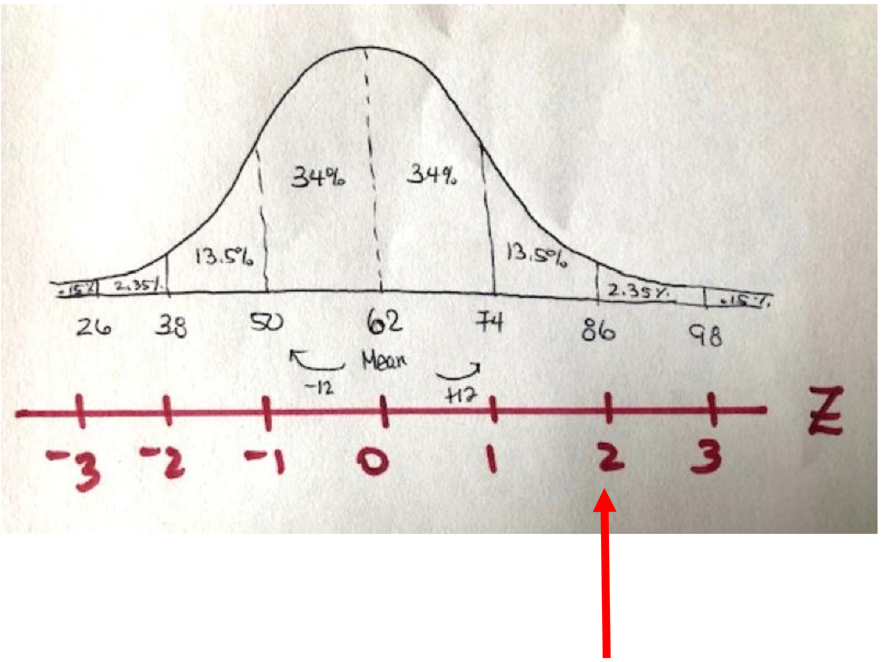
ANSWER: This can be solved by working in the raw values or working with z-scores. If you work with the raw values, begin by sketching the empirical rule intervals for this scenario. From the image below, you can see that a score of 86 points results in 2.35% and .15% greater than it. This sum is 2.5%. The percentage of results that would lie above a score of 86 points is 2.5%.



If you were to tackle this problem with z-scores, then you would convert the exam scores to z-scores and look at the empirical rule with z-scores.

$$
\begin{aligned}
z &= \frac{x-\bar{x}}{s} \\
&= \frac{86-62}{12} \\
&= \frac{24}{12} \\
&= 2
\end{aligned}
$$

A score of 86 results in a z-score = +2. Looking at the empirical rule intervals, the percentage of values above a z-score of 2 is 2.5%.



Instead of drawing the curves by hand, you may use a template such as this one and you can fill in the percentages. You can place the mean value in the center—an actual number—and then add/subract the standard deviations and record those values.