Chapter 3:
Statistics for Describing, Exploring, & Comparing Data

3.1 Review and Preview

Important Characteristics when Describing, Exploring, & Comparing Data (CVDOT)

1. Center
2. Variation
3. Distribution
4. Outliers
5. Changing Characteristics of Data overTime

Methods of Descriptive Statistics: the objective is to summarize or describe the important characteristics of a set of data

Methods of Inferential Statistics: Uses sample data to make inferences (or generalizations) about a population

3.2 Measures of Center

Measure of Center: a value at the center or middle of a data set

→ Several ways to determine the center:

- *Mean (also called the arithmetic mean or the average):* the measure of center found by adding the values & dividing by the total number of values

  --Notation:

Σ	sum of a set of values
x	represents the individual data values
n	the number of values in a sample
N	the number of values in a population
$\bar{x} = \frac{\sum x}{n}$	the mean of a set of sample values ($\bar{x}$ is called "x-bar")
$\mu = \frac{\sum x}{N}$	the mean of all values in a population (μ is called "mu")

--Example #1: Intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park

98, 92, 95, 87, 96, 90, 65, 92, 95, 93, 98, 94

--The mean uses every data value but this can be problematic if there's an outlier because it will affect the mean dramatically

*<u>*Outlier:*</u> a value that is located very far away from almost all of the other values

→ An extreme value that falls well outside the general pattern of almost all of the data

→ May reveal important information, may strongly affect the value of the mean & standard deviation, may distort the scale of a histogram

→ May be an actual recorded value or may be a typo

--Calculator: Press Stat→1:Edit

Enter the data into L1

Press Stat→Calc→1:1-Var Stats and then Enter

--Example #2: After the collapse of the two World Trade Center buildings, the following samples were obtained to measure the levels of lead in the air (in $\mu g / m^3$)

5.40 1.10 0.42 0.73 0.48 1.10

a. Find the mean of all samples.

b. What is the outlier?

c. Remove the outlier and then find the mean with the remaining data values.

d. Was the mean strongly affected by the outlier?

- *Median:* The measure of center that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude.

  --Notation: $\tilde{x}$ (pronounced "*x-tilde*")

  --To find the median, first sort the data, then…

  a. If the number of values is odd, the median is the number located in the exact middle of the list

  b. If the number of values is even, the median is found by computing the mean of the two middle numbers

  --Example #1: Use the Old Faithful Eruption Times (from above)
  65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

  -What if we took out the outlier 65?
  87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

  --The median is not strongly influenced by outliers because it does not really use every data value (like the mean does)

  --Calculator: Med represents the median in the calculator on the same 1-Var Stats screen but you have to scroll down to see it

  --Example #2: Use the following sample data describing the lead levels in the air,
  5.40 1.10 0.42 0.73 0.48 1.10

  a. Find the median.

  b. Remove the outlier and find the median using the remaining data values.

  c. Was the median strongly affected by the outlier?

- *Mode:* the value that occurs most frequently

--A data set is ***bimodal*** when two values occur with the same greatest frequency & each one is a mode.

--A data set is ***multimodal*** when more than two values occur with the same greatest frequency & each is a mode

--There is ***no mode*** when every data value occurs the same number of times

--Example #1: Use the Old Faithful Eruption Times
65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

--Mode is not often used with numerical data but can be used for nominal measurement (names, labels, categories)

- Ex: most common pets

--Calculator: The mode is not listed on the Calculator under 1-Var Stats
- You could sort the list and then see which occurs the most
-Enter the data into L1 then press Stat 2:Sort, type L1, and press Enter

--Example #2: Use the samples describing the amount of lead in the air to find the mode.

5.40 1.10 0.42 0.73 0.48 1.10

- *Midrange:* the measure of center that is the value midway between the maximum and minimum values in the original data set.

$$\text{Midrange} = \frac{\text{maximum value} + \text{minimum value}}{2}$$

--Example #1: Use the Old Faithful Eruption Times

65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

--Example #2: Use the samples describing the amount of lead in the air to find the midrange.

5.40 1.10 0.42 0.73 0.48 1.10

--Midrange is rarely used because it is too sensitive to the maximum & minimum extremes

--3 Benefits to the Midrange:

1. easy to compute
2. helps to reinforce that there are several ways to define the center of a data set
3. sometimes incorrectly used for the median so it helps to show how the midrange is different from the median

-- *Calculator: The midrange is not listed on the Calculator under 1-Var Stats - but you can use 1-Var Stats to find the max and min & use those to find the midrange*

Rule for Rounding: Carry one more decimal place than is present in the original set of values.

→ Round only the final answer (not intermediate values that occur during calculations)

→ Do not round the mode

Interpreting the Measures of Center:

- → No single best measure of center: it depends on the data set
- → Mean will be used often, median occasionally, the mode & midrange will rarely be used
- → The mean is relatively reliable: When samples are selected from the same population, the sample means tend to be more consistent than other measures of center

- → The mean takes every value into account, yet can be dramatically affected by a few extreme vales (this can be overcome by using a trimmed mean)

- → It doesn't make sense to do numerical calculations (mean & median) with data at the nominal level of measurement

  -Examples of Data at the Nominal Level of Measurement:

  1. Zip codes

  2. Ranks of stress levels from different jobs

  3. Surveyed respondents are coded 1 for Democrat, 2 for Republican, 3 for Independent

- → The mean of a population is not necessarily equal to the mean of the means found from different subsets of the population

  -Example of Data at the Ratio Level of Measurement: For each of the 50 states, a researcher obtains the mean salary of secondary school teachers to be $42,210 (data from the National Education Association). Is this the mean salary of all secondary school teachers?

Mean from a Frequency Distribution:

1. Find the class midpoint of each interval
2. Multiply each class midpoint by its frequency
3. Add these products to find the total of all sample values
4. Divide this sum by the number of sample values
   (# of sample values = $\sum f$ = sum of the frequencies)

$$\bar{x} = \frac{\sum(f \bullet x)}{\sum f}$$

*gives an approximation of $\bar{x}$ because it is not based on the exact original list of sample values

Example #1: Use the following frequency distribution for Boston's Sunday rainfall amounts (over 52 weeks) to determine the mean.

Rainfall (in.)	**Frequency (days) *f***	**Class Midpoint *x***	***f* • *x***
0.00—0.19	44		
0.20—0.39	6		
0.40—0.59	1		
0.60—0.79	0		
0.80—0.99	0		
1.00—1.19	0		
1.20—1.39	1		
	$\sum f$		$\sum(f \bullet x)$
$\bar{x} = \frac{\sum(f \bullet x)}{\sum f} =$			

Example #2: The following sample data was collected from a statistics class to see how many cars were registered in each household.

Number of cars	Frequency f	Class Midpoint x	$f \cdot x$
0-1	10		
2-3	12		
4-5	3		
6-7	1		
	Σf		$\Sigma(f \cdot x)$
$\bar{x} = \frac{\Sigma(f \bullet x)}{\Sigma f}$			

Weighted Mean: a mean computed with different values assigned different weights denoted by *w* because their values vary in their degree of importance

$$\bar{x} = \frac{\Sigma(w \bullet x)}{\Sigma w}$$

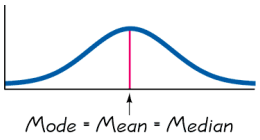
-Example #1: A final grade is weighted 30% tests, 30% quizzes, and 40% final exam. A student had an 86 test average, a 92 quiz average, and earned an 85 on the final. Calculate the weighted mean.

-Example #2: A particular college's grading policy assigns quality points to letter grades as follows:

A = 4 B = 3 C = 2 D = 1 F = 0

In a student's first semester at college, he took 5 courses. His final grades were A (3 credits), A (3 credits), B (4 credits), C (1 credit) and a D (3 credits). Compute his grade point average by finding the weighted mean.

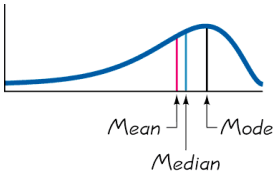
→ A distribution of data is **symmetric** if the left half of its histogram is roughly a mirror image of its right half



(b) Symmetric

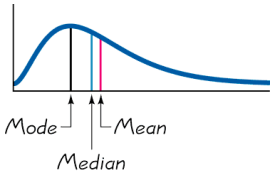
→ A distribution of data is **skewed** if it is not symmetric & extends more to one side than the other

-*Skewed to the left (negatively skewed)*: longer left tail, the mean & median are to the left of the mode, mean is usually to the left of the median



(a) Skewed to the Left (Negatively)

-*Skewed to the right (positively skewed)*: longer right tail, the mean & median are to the right of the mode, mean is usually to the right of the median



(c) Skewed to the Right (Positively)

3.3 Measures of Variation

One of the most important sections in the book

Basic Concepts of Variation:

→ *Range:* the difference between the maximum value and minimum value

Range = Max – Min

-range is easy to compute but isn't as useful as the other measures of variation that use every value

→ *Standard Deviation* of a set of sample values is a measure of variation of values about the mean.

$$s = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}} \qquad \text{OR} \qquad s = \sqrt{\frac{n\sum(x^2)-\left(\sum x\right)^2}{n(n-1)}}$$

- Standard deviation is the measure of variation that is generally the most important & useful
- Standard deviation is a measure of how much the data values deviate away from the mean
- The value of the standard deviation (s) is usually positive. It is zero only when all of the data values are the same number. (It is never negative).
- Larger values of s indicate greater amounts of variation
- The value of the s can increase dramatically with the inclusion of one or more outliers.
- The units of s (such as minutes, feet, …) are the SAME as the units of the original data values

- Procedure for Finding the Standard Deviation
  1. Compute the mean ($\bar{x}$)
  2. Subtract the mean from each individual value to get a list of deviations of the form $(x - \bar{x})$
  3. Square each of the differences obtained from step 2 which will produce numbers of the form $(x - \bar{x})^2$
  4. Add all of the squares obtained from step 3 which gives the value of $\Sigma(x - \bar{x})^2$
  5. Divide the total from step 4 by the number (n – 1) which is 1 less than the total number of values present
  6. Find the square root of the result of step 5

- Example #1: Intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park
  98, 92, 95, 87, 96, 90, 65, 92, 95, 93, 98, 94

  - What is the range?
  - What is the mean? $\bar{x}$

  - Find the standard deviation using the first formula:

x	$x - \bar{x}$	$(x - \bar{x})^2$
98		
92		
95		
87		
96		
90		
65		
92		
95		
93		
98		
94		
		$\Sigma(x - \bar{x})^2$

$$s = \sqrt{\frac{\sum\left(x - \bar{x}\right)^2}{n-1}} =$$

- Find the standard deviation using the 2nd formula:

x	x^2
98	
92	
95	
87	
96	
90	
65	
92	
95	
93	
98	
94	
Σx	$\Sigma(x^2)$

$$s = \sqrt{\frac{n\sum(x^2) - \left(\sum x\right)^2}{n(n-1)}}$$

- Standard Deviation of a **<u>Population</u>**:

  - Notation: σ (lower case "sigma"—looks like a p turned on its side)
  - Formula: $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$

  -The population standard deviation divides by N (the number of data values in the entire population)

  -The sample standard deviation divides by n-1 independent data values (otherwise you will underestimate the value of the population standard deviation/variance)

- *Advantages of the 1st formula given for standard deviation*:

$$s = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}$$

-reinforces the concept that standard deviation is a type of average deviation

- *Advantages of the 2nd formula given for standard deviation*:

$$s = \sqrt{\frac{n\sum(x^2) - \left(\sum x\right)^2}{n(n-1)}}$$

-easier to use when you must calculate standard deviations on your own

-eliminates the intermediate rounding errors introduced when the exact value of the mean is not used in the 1st formula

-used in calculators & programs because it requires less memory locations (only n, $\sum x$, & $\sum x^2$ rather than one for every value in the set of data)

→ *Variance:* A measure of variation equal to the square of the standard deviation

- Notation for sample variance: s^2
- Notation for population variance: σ^2

- The sample variance (s^2) is said to be an *unbiased estimator* of the population variance (σ^2)

  --Values of s^2 tend to target the value of σ^2 rather than over or underestimating

- Example #1: Find the sample variance (s^2) for the Intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park

- One serious disadvantage of variance: the units of variance are different than the units of the original data set (the units of variance are squared)

➔ *Calculator: Press Stat→ 1: Edit*

- *Enter the data into L1*
- *Press Stat→Calc→1:1-Var Stats*
- *Sample standard deviation is Sx*
- *Population standard deviation is σx*
- *To find variance, copy down the ENTIRE decimal for the appropriate standard deviation (sample or population) then square it*

- Example #2: Use the samples describing the amount of lead in the air to find the sample standard deviation and variance.

  5.40 1.10 0.42 0.73 0.48 1.10

Interpreting & Understanding Standard Deviation:

♦ Standard deviation measures the variation among values:

values close together = small standard deviation

values farther apart = larger standard deviation

♦ *Range Rule of Thumb:* based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean

--To interpret a known value of the standard deviation:

Minimum "usual" value = $\bar{x}$ – 2s

Maximum "usual" value = $\bar{x}$ + 2s

--If the standard deviation is unknown, you can estimate it by dividing the range by 4: $s \approx \frac{range}{4}$

--Example #1: Estimate the minimum and maximum usual values for the Intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park

98, 92, 95, 87, 96, 90, 65, 92, 95, 93, 98, 94

-- Example #2: Use the samples describing the amount of lead in the air to find the minimum and maximum usual values.

5.40 1.10 0.42 0.73 0.48 1.10

-Would the air quality test that showed 5.40 $\mu g / m^3$ be considered unusual? Why?

- *Mean Absolute Deviation:* the mean distance of the data from the mean

Mean Absolute Deviation = $\frac{\sum|x-\bar{x}|}{n}$

--Example #1: Use the following sample of intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park to find the mean absolute deviation.

x	$x - \bar{x}$	$\lvert x-\bar{x}\rvert$
98		
92		
95		
87		
96		
90		
65		
92		
95		
93		
98		
94		
		$\sum\lvert x-\bar{x}\rvert =$

$\frac{\sum|x-\bar{x}|}{n}$

--Why not use the mean absolute deviation?

-Because it requires that we use absolute values, it uses an operation that is not algebraic (the algebraic operations include addition, multiplication, extracting roots, and raising to powers that are integers or fractions)

-The mean absolute deviation lacks the additive property of variance (if you have 2 independent populations & you randomly select one value from each population & add them, the sum should have a variance equal to the sum of the variances of the 2 populations)

-The mean absolute value is biased (when you find mean absolute values of samples, you do not tend to target the mean absolute value of the population)

♦ *Empirical (or 68-95-99.7) Rule for Data with a Bell-Shaped Distribution:*

--about 68% of all values fall within 1 standard deviation of the mean

-between ($\bar{x}$ - s) and ($\bar{x}$ + s)

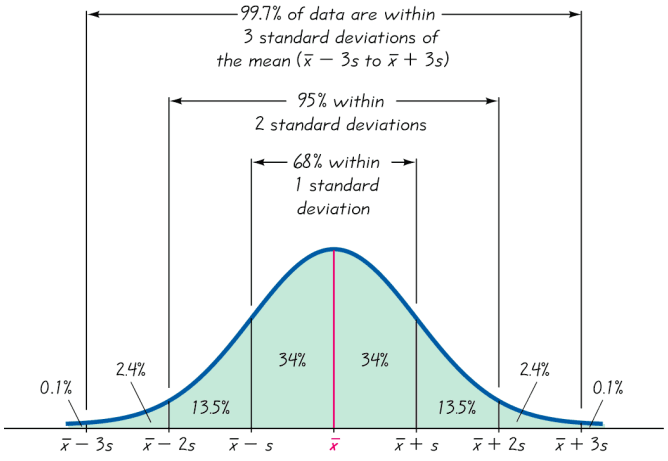
--about 95% of all values fall within 2 standard deviations of the mean

-between ($\bar{x}$ -2s) and ($\bar{x}$ +2s)

--about 99.7% of all values fall within 3 standard deviations of the mean

-between ($\bar{x}$ -3s) and ($\bar{x}$ +3s)

--Example:



- *Chebyshev's Theorem:* the proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $1 - 1/K^2$, where K is any positive number greater than 1. For K = 2 and K = 3, we get…

  --At least $\frac{3}{4}$ or (75%) of all values lie within 2 standard deviations of the mean

  --At least 8/9 or (89%) of all values lie within 3 standard deviations of the mean.

  --Ex:

  --The Empirical Rule applies only to data sets with bell-shaped distributions, whereas, Chebyshev's Theorem applies to ANY DATA SET

  --Results are approximations because the results are lower limits

- *Coefficient of Variation (CV)* for a set of nonnegative sample or population data describes the standard deviation relative to the mean, is expressed as a percent, & is given by the following formula:

  --Sample: $CV = \frac{s}{\bar{x}} \bullet 100\%$ --Population: $CV = \frac{\sigma}{\mu} \bullet 100\%$

  --When comparing variation in 2 data sets, the standard deviations should only be compared if the 2 data sets use the same scale and have a similar mean. If not, we can use the coefficient of variation to compare variation among data sets.

  *The higher the coefficient of variation, the more variation among the data values.

  --Example #1: Find the coefficient of variation for the intervals (in minutes) between eruptions of the Old Faithful geyser in Yellowstone National Park if s = 8.9 min & $\bar{x}$ = 91.25

--Example #2: Find the coefficient of variation for the following samples describing the amount of lead in the air if s = 1.914 and $\bar{x}$ = 1.538

--Which data set varies more—the data set from Example #1 or #2? Explain.

- ***Biased Estimator:***
  The sample standard deviation (s) is a biased estimator of the population standard deviation (σ) because the values of s do not target but rather tend to underestimate the value of σ

- ***Unbiased Estimator:***
  The sample variance (s^2) is an unbiased estimator of the population variance (σ^2) because the value of s^2 tend to target and not underestimate or overestimate the value of σ^2

3.4 Measures of Relative Standing

z score (or standardized value): the number of standard deviations that a given value x is above or below the mean

sample z score: $z = \frac{x - \bar{x}}{s}$ population z score: $z = \frac{x - \mu}{\sigma}$

→ used to compare values from different data sets

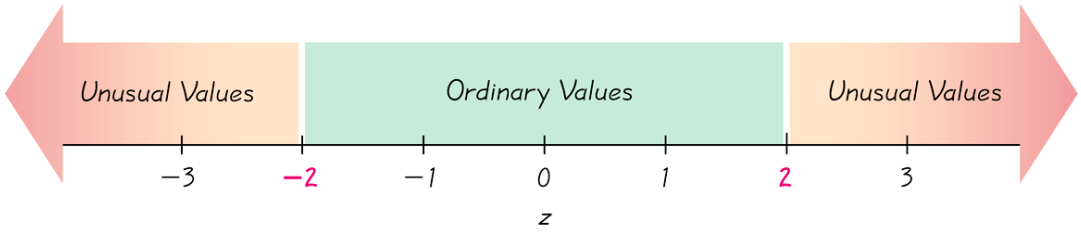
→ whenever a value is less than the mean, its z score is negative

→ z scores describe the location of a value (in terms of standard deviations) relative to the mean

→ ordinary or usual values have z scores between -2 and +2

→ unusual values have z scores less than -2 or greater than +2

→ We round z-scores to two decimal places & are expressed with no units of measurement



→ Example #1: With a height of 67 inches, William McKinley was the shortest president of the past century. The presidents of the past century have a mean height of 71.5 in. and a standard deviation of 2.1 in.

- What is the difference between McKinley's height and the mean height of presidents from the past century?
- How many standard deviations is that (the difference in his height from the mean)?
- Convert McKinley's height to a z score. $z = \frac{x - \bar{x}}{s}$

- If we consider "usual" heights to be those that convert z scores between -2 and 2, is McKinley's height usual or unusual?

→ Example #2: The Weschler Adult Intelligence Scale (better known as the IQ test) has a mean of 100 and a standard deviation of 15. A psychologist tests a woman's IQ and finds that it is 121.

- Find the woman's IQ as a z-score: $z = \frac{x - \bar{x}}{s}$

- Would this woman's IQ be considered unusual?

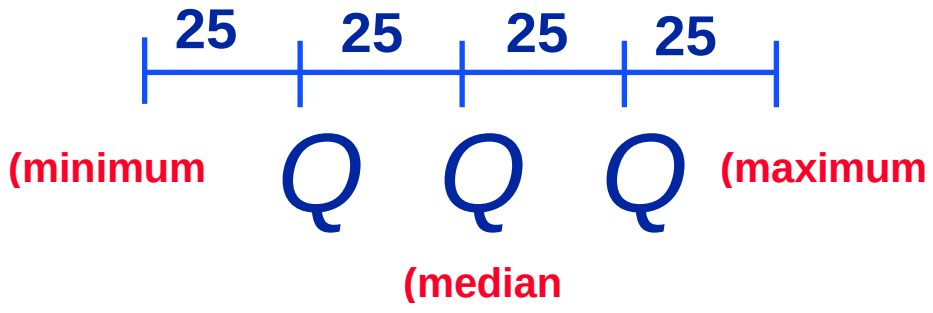
Quantiles or Fractiles: partition data into groups with roughly the same number of values

Median: the middle value that divides the data into 2 equal parts

→ 50% of the values are equal to or less than the median & 50% of the values are greater than or equal to the median

Quartiles (Q_1, Q_2, and Q_3): divide the sorted data values into four equal parts

→ Q_1: at least 25% of the sorted values are less than or equal to Q1 and at least 75% of the values are greater than or equal to Q1
→ Q_2: same as the median—separates the bottom 50% of the sorted values from the top 50%
→ Q_3: separates the bottom 75% of the sorted values from the top 25%



→ There is not a universal agreement on calculating quartiles & different programs often yield different results

→ *Calculator: Min, Q1, Med, Q3, Max can all be found on the 1-Var Stats screen (just don't forget to enter the data into L1 and scroll down after finding the calculations)*

--Example #1: Use the following sample of ages from 11 wild bears to find the Minimum, 1st Quartile, Median, 3rd Quartile, and the Maximum.

19 55 81 115 104 100 56 51 57 53 68

--Example #2: Use the following sample of weights (in lbs) from 12 wild bears to find the Minimum, 1st Quartile, Median, 3rd Quartile, and the Maximum.

80 344 416 348 166 220 262 360 204 144 332 34

Percentiles (P_1, P_2, …, P_{99}): divide the sorted data values into 100 groups with about 1% of the values in each group

$$\text{Percentile value of } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

→ round the percentile to the nearest whole number

→ percentiles tell you the percent of data that is at or below a given value (x)

→ Make sure the data is arranged in order from *smallest to largest* before finding a percentile value

→ Example #1: Find the percentile of 90 min. intervals between eruptions of the Old Faithful geyser in Yellowstone National Park

65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

→ Example #2: Find the percentile of an age of 81 for a bear based on the following sample data

19, 51, 53, 55, 56, 57, 68, 81, 100, 104, 115

→ How to convert from a given percentile to the corresponding value in the data set:

1. Sort the data (lowest to highest)

2. Compute $L = \left(\frac{k}{100}\right)n$ *where k = the percentile & n = # of values*

3. Is L a whole number?

--if yes, find P_k by adding the L^{th} value & the next value & dividing by 2

--if no, change L by rounding it up to the next larger whole # & the value of P_k is the L^{th} value, counting from the lowest

→ Example #1: Find the quartiles of intervals between eruptions of the Old Faithful geyser in Yellowstone National Park

65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

Q_1

Q_2

Q_3

→ Example #2: Find the 90^{th} percentile if we take out the outlier of 65 min.
87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

→ Example #3: Find the 10^{th} percentile if we take out the outlier of 65 min.
87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

Interquartile range: $Q_3 - Q_1$

→ Example #1:

Semi-interquartile range: $\frac{Q_3 - Q_1}{2}$

→ Example #1:

Midquartile: $\frac{Q_3 + Q_1}{2}$

→ Example #1:

10-90 Percentile Range: $P_{90} - P_{10}$

→ Example #1:

To Determine if a Data Value is an Outlier:

1. Find Q_1 and Q_3
2. Find the interquartile range (IQR): $Q_3 - Q_1$
3. Multiply the IQR by 1.5
4. The data value is an outlier if it is above Q_3 by an amount greater than 1.5 x IQR
5. The data value is an outlier if it is below Q_1 by an amount greater than 1.5 x IQR

→ Example #1: Use the following sample of weights (in lbs) from 12 wild bears to determine if a weight of 416 lbs would be considered an outlier.

80 344 416 348 166 220 262 360 204 144 332 34

5-number summary: consists of the minimum value (min), the first quartile (Q1), the median or second quartile (Q2), the third quartile (Q3), and the maximum value (max)

Boxplot (or box-and-whisker diagram): a graph of a data set that consists of a line extending from the min to the max, & a box with lines drawn at Q1, Q2, and Q3.

→ Also referred to as skeletal or regular boxplots

→ To construct a boxplot:

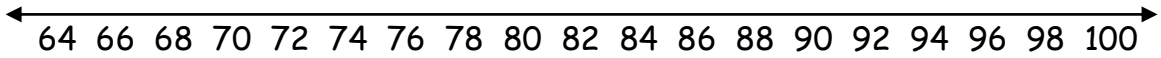
1. Find the 5-number summary (min, Q1, Q2, Q3, max)
2. Construct a scale with values that include the min & max
3. Construct a box from Q1 to Q3 and draw a line in the box at the median
4. Draw lines extending outward from the box to the min & max

→ Useful for revealing the center of the data, the spread of the data, the distribution of the data, & the presence of outliers (when comparing two or more data sets)

→ Not the best choice when you have a single data set because they don't show as much detailed information as histograms or stem-and-leaf plots

→ It's very important to use the same scale when comparing data sets

→ Ex: Let's make a boxplot of the Old Faithful eruption data:
65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98



Modified boxplots: a boxplot constructed with these modifications

1. A special symbol (such as an asterisk) is used to identify outliers as defined by the interquartile range (IQR)

2. The solid horizontal line extends only as far as the minimum data value that is not an outlier & the maximum data value that is not an outlier

→ Ex: Let's make a modified boxplot of the Old Faithful eruption data:
65, 87, 90, 92, 92, 93, 94, 95, 95, 96, 98, 98

→ *Calculator: Enter the data into L1*
- *Press Stat plot (2^{nd} y=)*
- *Turn the Plot ON*
- *Regular Box and Whisker Plot is the middle example on the bottom row under Type*
- *Modified Box and Whisker Plot is the first example on the bottom row under Type*
- *Then press Zoom 9:Stat*
- *If you press trace and hit the arrow keys, the cursor will jump to show you the 5 number summary and/or outlier values*