Chapter 3: Describing, Exploring, and Comparing Data

Section 3.1: Measures of Center

Def **Measure of Center** - a value at the center or middle of a data set.

The three most widely-used measures of center are the mean, median, and mode.

The (**arithmetic) mean** of a data set is computed by _____ all of the values of the variable in the data set and _____ by the number of observations.	
The **population** arithmetic mean, μ, is computed using _____ of the individuals in a population. The population mean is a parameter.	The **sample** arithmetic mean, $\bar{x}$, is computed by using some of the individuals in a population. The sample mean is a _____.
$= \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x}{}$	$= \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{}$

Ex: Of the 42 students enrolled in an Introductory Statistics course, the data below are the first 10 exam scores. Treat the 10 students as a _____ of the population, which means you use _____ to find the mean.

Student	**Score**
Michelle	82
Ryanne	77
Bilal	90
Pam	71
Jennifer	62
Dave	68
Joel	74
Sam	84
Justine	94
Juan	88

The **median** of a data set is the value that lies in the _____ of the data when arranged in ascending order. We use M to represent the median.	
ODD number of data **1st:** Arrange the data in _____ order **2nd:** The median will be the middle number *Ex:* 11, 14, 16, 19, 28	**EVEN** number of data **1st:** Arrange the data in _____ order **2nd:** The median will be the _____ of the middle numbers *Ex:* 14, 18, 20, 26, 31, 39

The **midrange** of a data set is the value midway between the minimum and maximum values.

$$Midrange = \frac{min\ value + max\ value}{2}$$

Ex: Use the data from the Introductory Statistics example from above to find the median and midrange.

Round-Off Rule: Carry one more decimal place than is present in the original set of values.

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.
*If no observation occurs more than once, we say that the data have _____.
*If the data set has more than one observation that repeat the same number of time, then it is considered _____.

Ex: Find the mode for each example below.

a) The following data represent the number of O-ring failures on the shuttle *Columbia* for its 17 flights prior to its fatal flight: $0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3$	b) The data of the test scores from above: $82, 77, 90, 71, 62, 68, 74, 84, 94, 88$	c) Hair color of ten people in line: Brown, Blonde, Red, Brown, Brown, Blonde, Brown, Blonde, Blonde, Red

Mean from a Frequency Distribution

Formula: $$\bar{x} = \frac{\sum (f \cdot x)}{n}$$

Ex: The following table gives the weights of a sample of 100 babies born at a local hospital.

Weight (in lbs)	*Freq* (f)	*class midpt* (x)	$f \cdot x$
3-4.9	5		
5-6.9	32		
7-8.9	40		
9-10.9	18		
11-12.9	5		
	$n = \sum f =$		$\sum (f \cdot x) =$

Find the sample mean.

Resistance Statistics

A numerical summary of data is said to be _____ if extreme values (very large or small) relative to the data do not affect its value substantially.

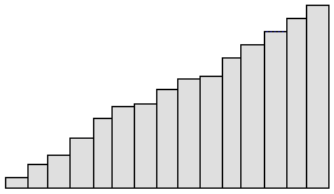
Ex: The following are wait times (in minutes) at a dentist office: $1, 1, 2, 2, 3, 5$. (a) Find the mean and median.

b) Note the value of 102 minutes added to this data. Find the mean and median. Which measure is resistant to the added value?

$$1, 1, 2, 2, 3, 5, 102$$

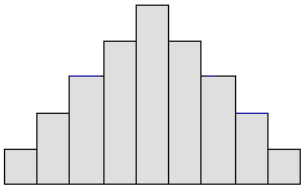
When data are skewed, there are extreme values in the tail, which tend to pull the _____ in the direction of the tail.

SKEWED LEFT



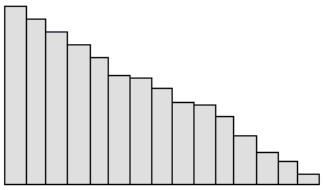
_____ < _____

SYMMETRIC



_____ = _____

SKEWED RIGHT



_____ > _____

General rule: If the data are symmetric use the _____ as the best measure of center.

If the data are skewed use the _____ as the best measure of center.

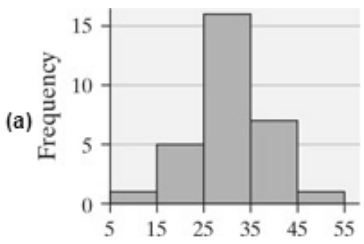
Ex: FICO scores range in value from 300 to 850, with a higher score indicating a more creditworthy individual. The distribution of FICO scores is skewed left with a median score of 723.

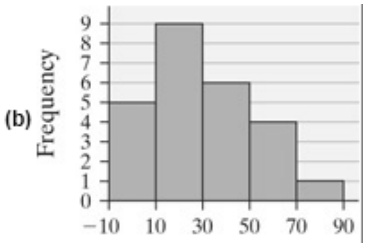
(a) Do you think the mean FICO score is greater than, less than, or equal to 723? Justify your response.

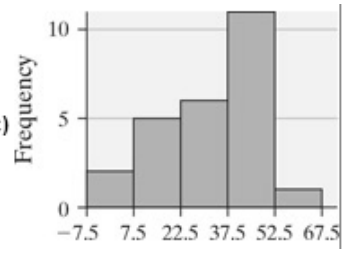
(b) What proportion of individuals have a FICO score above 723?

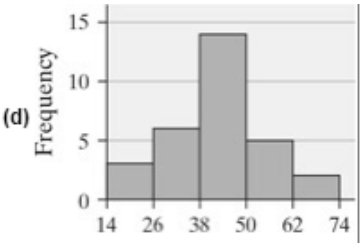
Ex: Match the histograms shown to the appropriate summary statistics by writing the appropriate number under each histogram.

	Mean	Median
1	42	42
2	31	36
3	31	26
4	31	32









3.2 Measures of Variation

Importance of Variation

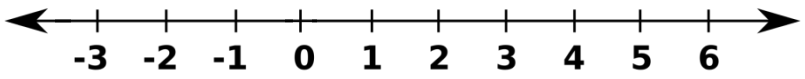
Ex: Advil and Motrin IB produce the same headache relief medication with the active ingredient ibuprofen. Each pill should contain 200 mg of ibuprofen. A health agency obtains a sample of ten tablets from both manufacturers and measures how much ibuprofen each pill actually contains.

	Number of milligrams measured									
Advil	199.25	198.50	200.10	200.75	201.00	198.00	200.10	199.00	201.10	202.20
Motrin IB	205.00	195.80	195.20	203.20	205.80	194.40	204.60	194.60	207.20	194.20

Each sample has a mean value of 200 mg. However, based on the given sample values, which company would you prefer to buy from?

Ex: The following are temperatures (in degrees) on four consecutive days in Mongolia in January: $-3, -1, 2, 6$

(a) Find the mean.



(b) How far away is each number from the mean?

Measures of variation

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

$$range = maximum\ value - minimum\ value$$

Standard Deviation of a Sample

Def The **standard deviation** (denoted by s) of a set of sample values is a measure of variation of values about the mean. It is a type of average deviation of values from the mean.

Formula: $$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$ *Shortcut:* $$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$$

Note: *Each form can be tricky, but the alternative form tends to be easier.*

Standard Deviation of a Population

Def The **standard deviation** (denoted by σ) of a complete set of values is a measure of variation of values about the mean. It is a type of average deviation of values from the mean.

Formula: $$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

Note: It's rare to compute a population standard deviation. Therefore, when using technology, be sure to use the sample standard deviation unless otherwise noted.

Variance

Def The variance (denoted by s^2 or σ^2) of a set of values is a measure of variation equal to the square of the standard deviation.

Ex: Find the range, standard deviation, and variance for the following sample of the number of chips in nine randomly sampled fun-sized bags of Doritos.

25 31 28 19 24 26 29 32 20

x	$x - \bar{x}$	$(x - \bar{x})^2$
19		
20		
24		
25		
26		
28		
29		
31		
32		
$\sum x =$		$\sum (x - \bar{x})^2 =$

Use the alternative form to find the standard deviation.

x	x^2
19	
20	
24	
25	
26	
28	
29	
31	
32	
$\sum x =$	$\sum x^2 =$

Question: If you bought a bag of chips everyday, would you prefer to have a small or large standard deviation between bags?

Standard Deviation from a Frequency Distribution

Formula: $$s = \sqrt{\frac{n\left[\Sigma\left(f \cdot x^2\right)\right] - \left[\Sigma\left(f \cdot x\right)\right]^2}{n\left(n-1\right)}}$$

Ex: Recall: The following table gives the weights of a sample of 100 babies born at a local hospital.

Weight (in lbs)	*Freq* (f)	*class midpt* (x)	$f \cdot x$	$f \cdot x^2$
3-4.9	5			
5-6.9	32			
7-8.9	40			
9-10.9	18			
11-12.9	5			
	$n = \Sigma f =$		$\Sigma(f \cdot x) =$	$\Sigma(f \cdot x^2) =$

Find the sample standard deviation and variance.

Range Rule of Thumb

Data are significantly LOW if the value is $\mu - 2\sigma$ or *lower.*	***Data are not significant*** if the value is between $\mu - 2\sigma$ and $\mu + 2\sigma$.	***Data are significantly HIGH*** if the value is $\mu + 2\sigma$ or *higher.*

Ex: The data below are free download wifi speeds (in Mbps) from ten of the busiest international airports.

AIRPORT CODE	WIFI SPEED	AIRPORT CODE	WIFI SPEED
DEN	78.2	YYC	41.8
YVR	55.1	BOS	32.2
PHL	48.4	DFW	32.0
SFO	45.3	MEX	27.7
SEA	43.7	DTW	22.9

https://www.speedtest.net/insights/blog/fastest-airports-north-america-2017/

If an international airport began to provide new 60.8 Mbps free wifi, and claimed that their speed is "miles above" many others, would you agree with their claim?

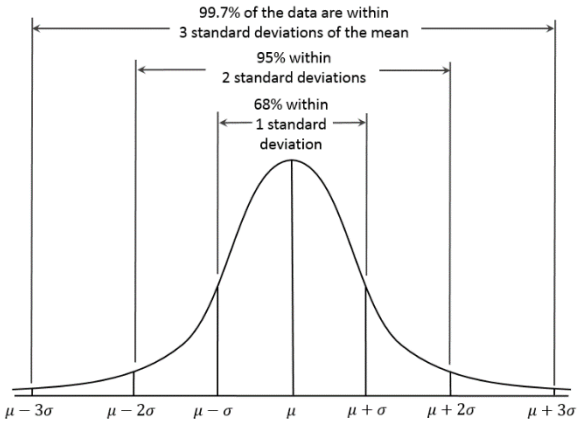
Empirical Rule

Empirical Rule says that a _____ distribution has…

○ Approximately _____% of the data will lie within **1** standard deviation of the mean.

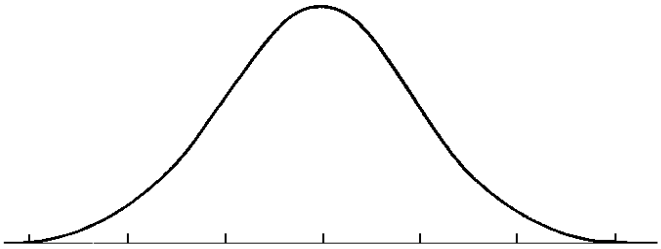
○ Approximately _____% of the data will lie within **2** standard deviation of the mean.

○ Approximately _____% of the data will lie within **3** standard deviation of the mean.



Ex: The following data represent the ages of all of the forty-two female patients of a family doctor. We are told that the data has a bell-shaped distribution and that the population mean, μ, is 57.2 years and the population standard deviation, σ, is 12.3 years.

41	48	43	38	35	37	44
62	75	77	58	82	39	85
67	69	69	70	65	72	74
60	60	60	61	62	63	64
54	54	55	56	56	56	57
45	47	47	48	48	50	52



(a) Determine the percentage of all patients whose age is within 3 standard deviations of the mean.

(b) Between what two values will this percentage lie?

Finding summary statistics for a data set

Use Graphing Calculator (TI-84 Plus)

Instructions: (a) STAT ⇒ 1: Edit… ⇒ Enter data into a list

(b) STAT ⇒ CALC ⇒ 1: 1-Var Stats ⇒ List: (Choose list) ⇒ Calculate

3.3 Measures of Relative Standing and Boxplots

z Scores

Def A ***z* score** is the number of standard deviations that a given value x is above or below the mean.

Formula: Sample : $z = \dfrac{x - \bar{x}}{s}$ Population: $z = \dfrac{x - \mu}{\sigma}$

Round-Off Rule: Round z scores to two decimal places.

The z score is a standardized value that describes a data value's relative standing.

1. A negative z score corresponds to a data value below the mean.
2. Unusual data values are more than two standard deviations from the mean.

   Ordinary values: $-2 \le z\ score \le 2$

   Unusual data values: $z < -2$ or $z > 2$

3. The z score allows us to compare data values drawn from different samples or populations.

Ex: Two college roommates are taking different physics courses at a university. They agree to a wager regarding their midterm scores, whereby the loser must do the dishes for a month. After scoring an 82, Jacob insists that Michael lost since he earned a 70 on his exam. However, Michael argues that he performed better relative to the rest of his class than did Jacob. Use the given class results to determine who won the bet?

Jacob's class: $\bar{x} = 78,\ s = 6$ Michael's class: $\bar{x} = 55,\ s = 12$

Percentiles

Def Percentiles (denoted P_k) are measures of location in relation to all the other data values.

Notation

Symbol	*Represents*
L	locator that gives the rank of a value
P_k	k^{th} percentile

Finding a Percentile Associated to a Given Score

Formula: $percentile\ of\ score = \dfrac{\#\ of\ scores\ less\ than\ given\ score}{total\ number\ of\ scores}$

Finding the Score Associated to a Given Percentile

Formula: $$L = \frac{k}{100} \cdot n$$

Note: *If L is a decimal, then always round up to find the score with that specific rank.*

If L is a whole number, then average the k^{th} score and the next higher score.

Ex: The following data set represents the selling price (in thousands) of 38 randomly selected homes.

128	135	138	145	149	152	155	158	159	163	163	165	167
168	170	170	172	173	176	177	180	180	185	188	191	193
199	205	210	212	215	229	233	250	325	450	500	525	

(a) Find the percentile corresponding to a selling price of \$188,000.

(b) Find the home price corresponding to the 85th percentile.

Quartiles

Def **Quartiles** (denoted Q_1, Q_2, and Q_3) are measures of location which divide a data set into four groups with about 25% of the values in each group.

Note: $Q_1 = P_{25}$, $Q_2 =$ *the median, and* $Q_3 = P_{75}$

Boxplot

Def A **boxplot** is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, the median, and the third quartile.

Note: *A five-number summary refers to the five values used to draw the boxplot.*

Ex: Find the five-number summary for the previous data regarding home prices.

(a) Find the minimum and maximum values in the data set.

(b) Find Q_2 (the median).

(c) Find $Q_1 = P_{25}$.

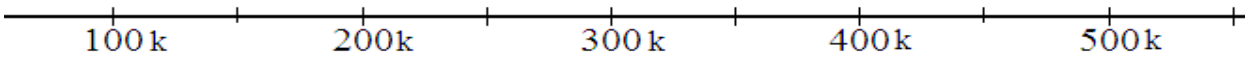
(d) Find $Q_3 = P_{75}$.

(e) List the five-number summary.

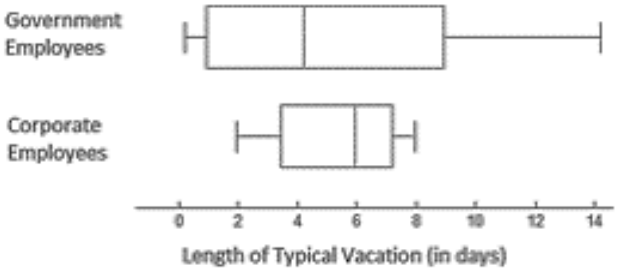
Ex: Construct a boxplot for the previous data set regarding home prices.

Graphing a boxplot:

(a) Place the five number summary for a data set on a number line, and draw a straight line connecting them.
(b) Draw vertical lines at each of the five number summary values.
(c) Draw a rectangle connecting Q_1 to Q_3.



Ex: Below, boxplots are shown for the length of a typical vacation for California residents who work for the government in some capacity and for those who work for a private company.



Based on these graphs, would you prefer a government job or corporate job based solely on the length of their vacations? There is no one right answer, but please consider **center**, **spread**, and any other relevant statistics or values from the boxplots shown to support your data.