Compiled by: Mr. Loren L. Spencer

Displaying & Describing Data Vocabulary

Describing Data-C.U.S.S.

Center

- When a constant is added to each data value, the center increases by that constant.
- When each data value is multiplied by a constant, the center is increased by the multiple of that constant.

Mean-(μ, $\bar{x}$)-the sum of all the data divided by the number of data-**the expected value.**

Significantly impacted by outliers.

Median-the middle most data value in a line of data arranged from least to greatest. 50% of the data are greater and 50% of the data are smaller.

Not significantly impacted by outliers.

Note: Place data values in **STAT EDIT**
Press **STAT CALC 1-Var STATS**

Unusual Features

Gaps-Places in a distribution with **no data values.**

Clusters-groups of data points-give the value of the centers of the groups and the ranges.

Outliers-data points that are significantly larger or smaller than the remaining data values.
1.5 x IQR or more than **3 standard deviations** from the mean.

Shape

Unimodal-one mound-the most common occurring data value-highest probability of occurring.

Bimodal-two mounds or high points-the two most common-the two values most likely to occur.

Multimodal-more than two mounds

Uniform-all values are **equally likely** to occur. The bars of a histogram are of equal height.

Mound Shaped- The mean equals the median and mode. Use this description for bell curves, t-distributions and normal curves.

Skewed Right-the mean is to the **right/greater** than the median. The tail is to the right.
All Chi-squared distributions are skewed right

Skewed Left-the mean is to the **left/smaller** than the median. The tail is to the left.

Symmetric-The mean equals the median.
The mode may different-bimodal or inverse bell.

Describing Data-C.U.S.S.-cont'd.

Spread

- When a constant is added to each data value, the spread does not change.
- When each data value, is multiplied by a constant, the measure of spread is increased by that multiple of that constant. Exception: variance is increased by the constant's square.

Range-the largest data value minus the smallest.

Significantly impacted by outliers.

Interquartile Range:$Q_3 - Q_1$. The data value at the 1st quartile subtracted from the data value at the 3rd quartile.

The range of the Middle 50% of the data

Not significantly impacted by outliers.

Standard Deviation-(σ, s)-The square root of the average of the squared data differences from the mean. The square root of the variance.

Z-scores are a measure of standard deviations from the mean and are related to percentile ranks.

Variance- The standard deviation squared. The average of the squared differences from the mean.

Univariate Data Displays

- Label the graph's axes
- Title the graph

Categorical Displays-Displays that have one axis which is comprised of a list of **qualitative** values rather than quantitative values and measure frequency or relative frequency *(percent).*

Note: There is no description of spread, shape, or center for categorical data. The **Mode** measures frequency & is a valid measure for categorical data.

DotPlots: Measure the frequency of categorical variables and provide the exact counts of the data.

Barcharts/Bargraphs-Provide frequencies or relative frequencies as *(percents)* and represent the categorical data counts as areas.

Note: Bars should not touch because the data categories are not sequential/numerical in nature.

PieCharts/Graphs- Represent categorical data as areas & relative frequencies as *(percents)* .

Compiled by: Mr. Loren L. Spencer

Displaying & Describing Data Vocabulary

Quantitative Displays- Displays in which one axis is a listing of sequential quantitative values.

- Label the graph's axes
- Title the graph

Histogram: a frequency distribution whose class/bar widths have a height that is proportional to the frequency of the values in that class.

- Useful for large data sets
- Provides shape and an idea of spread
- Individual data values are usually lost
- The area of a bar is proportional
- Vertical axis can be either frequencies or relative frequencies/percents.

Note: Bars should touch because the data is sequential/numerical in nature.

Stemplot- a graphical displaying that separates the ones digits from the remaining digits.

- Provides shape
- Maintains all data values so exact summary statistics can be calculated
- Unwieldy for large data sets.

Note: must provide a legend

Cumulative Relative Frequency Plot (ogive)- Plots cumulative frequencies for data from left to right such that the largest data point furthest to the right will be at 100%.

- Skewed right data will increase rapidly at first then more slowly later. **Begins convex**
- Skewed left data will increase slowly at first and more rapidly later. **Begins concave**

Box Plot (box and whisker plot)- a visual representation of **the five number summary**. The smallest value, the 1st quartile value, the median, the 3rd quartile value and the largest value.

Note: Each section contains 25% of the data

Remember: The mean is not shown & the variance cannot be calculated. The range can be computed.

Calculator:

- Enter Data into STAT Edit
- Press 2nd STAT PLOT
- Select the Box plot with outliers & Press graph Press Zoom 9

Do Not Forget to label your axis

Comparing Distributions

- Be Specific
- Compare and contrast the centers
- Compare and contrast the spread
- Compare and contrast the shape

Bivariate Data

Scatterplot: Both axes represent variables either the response variable (output) or the explanatory variable (input).

Note: Neither axis represents a frequency

Empirical Rule

- Applies to symmetric mound/bell shaped curves.
- **68%** of the data is within **± 1** standard deviations of the mean
- **95%** of the data is within **± 2** standard deviations of the mean
- **99.7%** of the data is within **± 3** standard deviations of the mean

Quick Hints:

Percentile ranks are the areas under the distributions and should be read left to right.

Z-scores- measure the distance in number of standard deviations a value is from the mean. The related area under the curve tells us the chance that an event will happen.

Remember: we are always calculating the likelihood that an event occurs is either **greater than or less than** a specific value. In essence we are calculating the chance of a range of values.

We do not calculate the probability of a specific data value when working with continuous data.

Standard Normal-the distribution which is created **After** a z-score has been computed.

$\mu = 0$ and $\sigma = 1$

Before the z-score is computed the mean is the mean of the population and the standard deviation is the standard deviation of the population.

Compiled by: Mr. Loren L. Spencer

Notes: C.U.S.S. & B.S.

As we work through this section on quantitative data, we will learn to create data displays and we will develop a common language to describe the distributions of the data. We will begin with developing the common language which follows the acronym **C.U.S.S.**

C: Measures of the Middle (mean and median)

U: Unusual Features (gaps, clusters and outliers)

S: Shape (skewed left, skewed right, mound, symmetric, uniform, bi-modal and multi-modal)

S: Spread (variation, range, interquartile range, variance and standard deviation)

When describing or comparing distributions you will need to address all four of the parts of **C.U.S.S.** and you will need to B.S. those in your descriptions, that is to say you must **B**e **S**pecific.

When it comes to being able to C.U.S.S. appropriately, all parts are important and should be addressed. However, I will give you a rank ordering of importance.

1. **Center and Spread:** These must always be addressed. Failure to do so will result in no points being awarded. (... and may God have mercy on your soul)— Billy Madison
2. **Shape:** If you want to receive full credit, you should be discussing the shape of the distribution
3. **Unusual Features:** If there exist unusual features they need to be addressed. If they are not addressed, you can expect point deductions.

C: Measures of the Middle

- **Mean:** The average of all of the values (the mean is impacted by outliers and non-symmetric distributions). The mean is the balance point on a scale or see-saw.
  Calculation: Sum all of the values and divide by **n** the number of values summed.
- **Median:** the middle-most number (median is not impacted by outliers and non-symmetric distributions). 50% of the values are larger and 50% are smaller than the median.
  Calculation: Arrange the numbers smallest to greatest and find the middle number.

U: Unusual Features

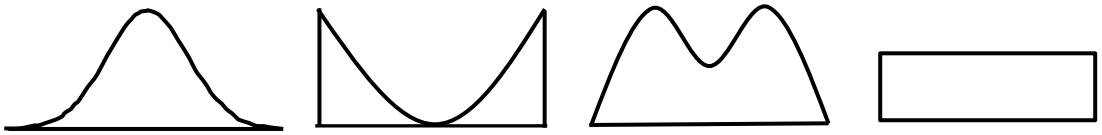
- **Gap:** a section in the distribution without a data point
- **Clusters:** Clusters are distinct groups of data as in two mounds. (**Note:** if clusters exist the centers spread and shape of the clusters must be given)
- **Outliers-**data points that are either too large or too small. In this course, we identify outliers utilizing 2 methods:
  1. More than 2-3 standard deviations above or below the mean. (We will learn about standard deviations, a measure of variation, in the near future.)
  2. More than 1.5 Interquartile ranges below the 1^{st} quartile or more than 1.5 interquartile ranges above the 3^{rd} quartile. (We will discuss the interquartile range and the 1^{st} and 3^{rd} quartile when we discuss box plots)

Compiled by: Mr. Loren L. Spencer

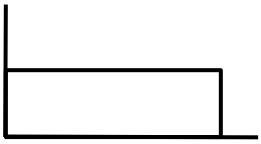
Notes: C.U.S.S. & B.S.

S: Shape

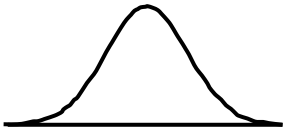
- **Symmetric:** The vertical line can divide the data into two matching mirror images. If the data is symmetric the mean and median will be equal.

  Examples: 

- **Uniform:** A graph that is approximately the same height.

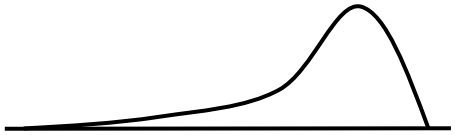
  Example: 

- **Mound Shape:** The graph is symmetric with most of the data in the center of the graph. The data density diminishes as you move towards each tail. The mean, median and mode are all equal.

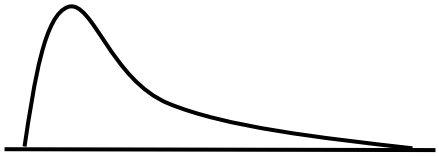
  Example: 

Note: Because data is rarely if ever perfect, we need to qualify the above shapes as approximately symmetric or reasonably symmetric or reasonably mound shaped or approximately uniform.

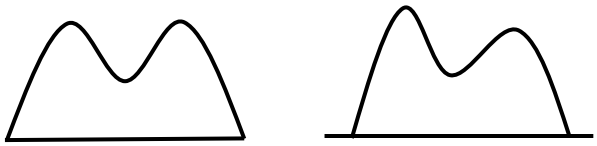
- **Skewed Left:** The tail is to the left and the mean is less than the median. The mean is to the left of the median. (imagine pulling on a sticky wad of gum towards the left)

  Example: 

- **Skewed Right:** The tail is to the right and the mean is greater than the median. The mean is to the right of the median. (imagine pulling on a sticky wad of gum towards the right)

  Example: 

- **Bi-modal:** The distribution has two distinct peaks (modes). The peaks may not be equal in height.

  Examples: 

Compiled by: Mr. Loren L. Spencer

Notes: C.U.S.S. & B.S.

S: Spread

- **Range:** The range provides an idea as to how spread out the data is by subtracting the smallest value from the largest value. The range is a singular value and is never negative. Because the range uses the largest and smallest value it is greatly impacted by outliers.

- **Interquartile Range:** Then interquartile range or IQR gives an idea as to how spread out the data is by focusing on the middle 50% of the data and is computed by subtracting the value of the 1st quartile from that of the 3rd quartile. We express it in this manner IQR= $Q_3 - Q_1$. Because the interquartile range focuses on the middle 50% of the data it is not impacted by outliers. IQR is very useful for skewed distributions.

- **Variance:** The variance takes an average of the data distances from the mean. Because some values are below the mean the distances are negative and obviously some of those distances are greater than the mean and are positive. As a consequence, the distances are squared and then summed. $\boldsymbol{\sigma^2}$ is the symbol for the true population variance.
  Remember: we use means as a measure of center when the data tends to be mound shaped and symmetric. Because variance uses the mean in the calculation we tend to reserve the use of the variance for distributions that are symmetric and mound shaped.

  The formula for variance of the entire population is $\boldsymbol{\sigma^2 = \sum \frac{(x_i - \mu)^2}{n}}$ where $\boldsymbol{\mu}$ is the true population mean.

  Unfortunately we rarely know the true population mean and have to rely on the sample mean which has the symbol $\boldsymbol{\bar{x}}$. The symbol for variance of the sample is $\boldsymbol{s^2}$ and the formula is as follows: $\boldsymbol{s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}}$ You may notice that the sample differences are divided by n-1 this is done because we are using the sample mean, $\boldsymbol{\bar{x}}$ an estimate of the true population mean, in our calculation.

- **Standard Deviation:** The standard deviation is just the square root of the variance. Because of advances in technology we are able to easily take square roots and as consequence we tend to talk about standard deviations more than we do variances. The symbol for the standard deviation of the population is $\boldsymbol{\sigma}$ and the formula is: $\boldsymbol{\sigma = \sqrt{\sum \frac{(x_i - \mu)^2}{n}}}$. As expected, the symbol for the standard deviation of the sample is just $\boldsymbol{s}$ and the formula is: $\boldsymbol{s = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}}$. We will typically say something is so many standard deviations from the mean. Because we are relying on the mean, we usually reserve standard deviation for data that is mound shaped. The smaller the standard deviation the closer the data is to the mean.

- Z-scores: A Z-score is a ratio that provides a measure as to how far a value is from the mean and takes into account both the center and the dispersion of the data. Z-scores act as a ruler and can be used to compare different shaped distributions the basic z score formula is $\boldsymbol{z = \frac{x - \mu}{\sigma}}$ where **z is the number of standard deviations a value lies from the mean.**