Sampling Terms

Population: The set of all elements of interest in a particular study

Sample: A subset of the population

Inferential Statistics: The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics/attribute of a population

If the measures are computed for data from a sample, they are called ***sample statistics*** .

If the measures are computed for data from a population, they are called ***population parameters*** .

In Statistical Inference, a ***sample statistic*** is referred to as the ***point estimator*** of the corresponding ***population parameter*** .

Since it is usually impossible to get all the numbers for a population, we have to use **sample statistics** to **reasonably estimate** the **population parameter**!

Example: calculate an sample average of number of salmon in Puget Sound and use it as the estimate of the population average of number of salmon in Puget Sound - "We can't count all the fish in the sea!!"

More about Point estimation in chapter 7.

Measures of Location:

Location

Location locates, or positions, a data point against the full data set such as:

The **Mean** calculation which calculates a central value, a typical value, a value near the middle,

The **Percentile** calculation that can say things like "for that particular number, 75% of the other values are below that value and 25% of the other values are below that value".

The **Z-Score**, which tells you the relative position of a number in units of Standard Deviation.

Average = Typical Value = Measure of central location

"Typical Values" calculated so that we have one value that can represent all the data points.

Mean

Arithmetic Mean: Add them up and divide by the count. Use these math formulas:

Population Mean = $\mu = \text{Mew} = \frac{\sum x_i}{N}$

Sample Mean = $\overline{x} = \text{Xbar} = \frac{\sum x_i}{n}$

Good for quantitative data when there are not extreme values - extreme values can make the mean look too big or too small (Median more representative of a typical value in that case)

The sample mean is a point estimator of the population mean

In Excel use: AVERAGE function

Median

Sort, then take the one in the middle. If count odd, take one in middle, if even, average middle two.

Marks the point in the sorted list (an actual number) where approximately 50% of the numbers are above and 50% of the numbers are below

Good for quantitative data when there are extreme values (like house prices and salaries)

Median can be a better measure of central location, than mean, when there are extreme values

In Excel use: MEDIAN function

Mode

One that occurs most frequently (can be bimodal, multimodal)

Bimodal means two numbers are tied with highest frequency. Multimodal means there are multiple modes.

Good for Categorical Data (Nominal and Ordinal)

Mode is the only way to calculate an "average" for categorical data

In Excel use: MODE.SNGL or MODE.MULT functions for quantitative data and PivotTables for Categorical or quantitative data.

MODE.SNGL will only show 1 mode if the data set is bi-modal or multi-modal. MODE.MULT can be used for multiple modes.

MODE.MULT can be used for multiple modes and will spill the results from the top cell where the formula lives.

Weighted Mean

Fast way to calculate Mean when you have a Frequency Table or Relative Frequency Table. Use these math formulas:

$$\overline{x} = \text{Weighted Mean} = \frac{\sum w_i x_i}{\sum w_i}$$

$\overline{x}$ = Weighted Mean = $\sum PF_i * x_i$

Use Excel Formula: SUMPRODUCT(Values, Weights)/SUM(Weights)

Geometric Mean

Used to calculate the average compounding rate per period for % change or % growth numbers. The average compounding rate per period is the rate that can be used in multiplicative growth formulas to calculate an end amount from a begin amount across equal size time periods. In finance it is used to calculate the "Average Compounding Rate Per Period" for an investment. Use one of two formulas:

Geometric Mean Formula 1:

Use when you are given all the "Growth Rates" or "Rates of Change" or "% Change" amounts across all equal sized periods

$$Geometric\ Mean\ 1 = \sqrt[n]{(1+GR_1)} * (1+GR_2) * \cdots (1+GRn)\text{, then subtract 1}$$

$$= [(1+GR_1)*(1+GR_2)*...(1+GRn)]\text{^}(1/n)\text{-}1$$

"Growth Rates" or "Rates of Change" or "% Change" = % change from one equal size period to the next

Growth Factor = 1 + Growth Rate = Factor you can multiply by Begin Value to get End Value. Growth Factor ALWAYS >= 0

In Excel use:

GEOMEAN(RangeOfGrowthRates+1) = 1 + Geometric Mean = 1 + Average Compounding Rate per Period

GEOMEAN(RangeOfGrowthRates+1)-1 = Geometric Mean = Average Compounding Rate per Period

Geometric Mean Formula 2:

Use when you are given the Begin Value, End Value and the number of periods

$$Geometric\ Mean\ 2 = \left(\frac{EndValue}{BeginValue}\right)\text{^}(1/n)\text{-}1$$

In Excel use:

RRI function. Function arguments: RRI(n,PV,FV) = RRI(NumberOfPeriods,BegValue, EndValue) = Geometric Mean = Average Compounding Rate per Period

Formula: =(EndValue/BegValue)^(1/NumberOfPeriods)-1 = Geometric Mean = Ave. Compounding Rate per Period

MUST use Geometric Mean (not arithmetic mean) if you want the true "average" compounding rate per period

Arithmetic Mean overestimates

Arithmetic Mean is for additive processes. Geomean is for multiplicative processes

In finance, when calculating End Values, use Geometric Mean; however, arithmetic mean is used in some situations like for Standard Deviation, Correlation, and other calculations that do not require true "average" compounding rate per period.

Percentiles/Quartiles

Marker that divides the sorted data set and says what % are above and what % are below

Percentiles

Provides information about how the data are spread over an interval from smallest to largest value (ascending)

A number that divides the ascending sorted data set, or marks the point in the ascending sorted list where approximately X% of the numbers are below and approximately 1-X% of the numbers are above the marking point. For example, on a recent exam a score of 78.75 marks the point in the ascending sorted data set, where 75% of the scores were below that score ands 25% of the scores were above that score.

Use Percentiles for quantitative data without a lot of duplicates

Note: There is NOT just one way to calculate percentiles. Percentile calculations are estimations, this is why we use the word "approximate" when we talk about percentiles

In Excel use:

PERCENTILE.EXC function. Function arguments: PERCENTILE.EXC(array,k) = PERCENTILE.EXC(ArrayOfNumbers, %PercentileValue) = Number that divides the ascending sorted list. The function uses formula: **P%*(n+1)** to get the position. If the k argument contains a 0% or 100% (min or max values), the function yields an error. THE "EXC" in the function name indicates that the function excludes the min and max values. This is the function that the textbook uses.

PERCENTILE.INC function. Function arguments: PERCENTILE.INC(array,k) = PERCENTILE.INC(ArrayOfNumbers, %PercentileValue) = Number that divides the ascending sorted list. The function uses formula: **p*n+(1-p)** to get the position. This function allows the k argument to contain a 0% or 100% to deliver the min or max values. THE "INC" in the function name indicates that the function includes the min and max values.

For big data sets, these two function deliver similar results.

Quartiles

Provides the three values that will divides the ascending sorted data set into four parts: first 25%, second 25%, third 25% and fourth 25%

In Excel use:

QUARTILE.EXC function. Function arguments: QUARTILE.EXC(array,quart) = QUARTILE.EXC(ArrayOfNumbers, 1or2or3). The function uses formula: **P%*(n+1)** to get the position. THE "EXC" in the function name indicates that the function excludes the min and max values. This is the function that the textbook uses.

QUARTILE.INC function. Function arguments: QUARTILE.INC(array,quart) = QUARTILE.INC(ArrayOfNumbers, 0or1or2or3,or4). The function uses formula: **p*n+(1-p)** to get the position. THE "INC" in the function name indicates that the function includes the min and max values. To deliver the Five Number Summary, use the formula: =QUARTILE.INC(ArrayOfNumbers, {0;1;2;3;4}) to display the numbers vertically, or =QUARTILE.INC(ArrayOfNumbers, {0,1,2,3,4}) to display the numbers horizontally.

Percentile Rank

PERCENTRANK.EXC and PERCENTRANK.INC opposite of PERCENTILE.EXC and PERCENTILE.INC. Whereas PERCENTILE functions calculates the marker value when you give it a % Percentile Value, PERCENTRANK functions calculates % Percentile Value when you give it a value.

Rank

Ranks numbers, 1st, 2nd, 3rd and so on for "Biggest to Smallest" or "Smallest to Biggest"

In Excel use:

RANK.AVG function. This provides an average of ranks when there is a tie. Function arguments: RANK.AVG function(number,ref,order) = RANK.AVG function(NumberToRank,ArrayOfNumbers, AscendingOrDescending)

RANK.EQ function. This provides equivalent ranks when there is a tie. Function arguments: RANK.EQ function(number,ref,order) = RANK.EQ function(NumberToRank,ArrayOfNumbers, AscendingOrDescending)

Variability

Variability answers the questions:

What is the dispersion in the data?

How spread out is the data?

Range

Range = Max value - Min value

The bigger the range, the more the variability or spread in the data

It is simple to calculate, but sensitive to extreme values

Interquartile Range

Interquartile Range = Quartile 3 - Quartile 1

The bigger the range, the more the variability or spread in the data

The Interquartile Range tells you the range for the middle 50% of the data. It overcomes the sensitivity to extreme values that range has

Deviation

Deviation = Particular X value - Mean

It tells you how far one datum is from the mean. "How far above or below the mean is the particular value?"

For any data set, the sum of the deviations is always zero!!!! This is why mathematically, when we calculate variance or standard deviation, we either square the values before we add.

Variance

A Numerical Measure that says how much variability there is in the data points

Variance uses all the data points, not just some like Range and Interquartile Range

Variance has squared units, which makes interpreting it difficult. Standard Deviation undoes the squared units and is thus easier to interpret and is more commonly used

$$Population\ Variance = \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$Sample\ Variance = s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

In Excel use: VAR.P function for population data and VAR.S for sample data.

Standard Deviation = SD

Standard Deviation uses all the data points, not just some like Range and Interquartile Range

Standard Deviation does not have squared units (like Variance) and is thus easier to interpret; the standard deviation has the same units as the data.

Population Standard Deviation = "little sigma" = σ ; Sample Standard Deviation = s

The sample standard deviation is a point estimator of the population standard deviation

Interpretation of Standard Deviation:

- A numerical measure that says how much variability/dispersion there is in the data in relation to the mean
- Standard Deviation is like an average of the deviations
- Standard Deviation tells us how fairly the mean represents its data points
- Standard Deviation tells us how clustered the data points are around the mean

$$Population\ Standard\ Deviation = \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$Sample\ Standard\ Deviation = s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

In Excel use: STDEV.P function for population data and STDEV.S for sample data.

Coefficient of Variation

In Excel create math formula = (Standard Deviation)/Mean = Coefficient of Variation

Coefficient of Variation converts Standard Deviation to Standard Deviation per unit of Mean so you can compare:

1) Data in different units.

or

2) Data in the same units, but the means are far apart.

Coefficient of Variation answers the question: "For every one unit of mean, what is the Standard Deviation?"

In Finance you see the inverse formula = (Average Return)/(Standard Deviation) = Return for 1 unit of risk

Shape of Distribution, Relative Location, Outliers

Shape of Distribution

Histograms

Histograms show the shape of the distribution visually.

A few short column heights to the left means skew negative of left

A few tall column heights to the right means skew positive or right

Bell shape or Symmetrical Shape indicates no skew

Skew

Skew measures the shape of the data distribution

Skew tells to which way the distribution is tilting

Skew = tilt of Histogram

In Excel use: The SKEW function to calculate the Skew. SKEW function uses the formula listed below =>

$$Skewness = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^3$$

When we calculate skew, the result can be negative (skew left), positive (skew right) or zero (no skew).

Left or Negative Skew

The frequency for small numbers is low and the frequency for large numbers is high

A few small values will tend to pull the Mean down away from Median

Mean usually less than Median

Right or Positive Skew

The frequency for large numbers is low and the frequency for small numbers is high

A few big values will tend to pull the Mean up away from Median

Mean usually greater than Median

No Skew = Bell Shape or Symmetrical Shape

Mean and Median equal, and Mean, Median and Mode tend to all be right near the middle of the data set.

Relative Location

The goal of relative location is to determine, in a data set, how far a particular x value is from the mean.

Z-Score = Standardized Value

Z-score measures the relative location of a particular x in the data set (as compared to the mean), in units of standard deviation.

Z-Score tells you "**How Many Standard Deviations The Particular Value Is Away From The Mean**". 1? -1? 1.5? 0?

$z_i <$ 0, value below mean

$z_i >$ 0, value above mean

z_i = 0, value is equal to mean

Observations in 2 different data sets that have the same z-score are said to have the same relative location or the same number of standard deviations away from the mean.

In Excel use the formulas:

Z-Score = (Particular Value - Mean)/StandardDeviation = z_i = (x_i-Xbar)/s

Use STANDARDIZE function to calculate z-score. Function arguments: STANDARDIZE(a,mean, standard_dev) = STANDARDIZE(ParticularX$_i$Value,Mean,StandardDeviation

Chebyshev's Theorem

Chebyshev's Theorem can determine the proportion of data values that lie between +/- a given number of standard deviations (z-score) for any shape distribution

Russian Mathematician, P.L. Chebyshev (1821-1894)

Chebyshev's Theorem:

Allows us to make a statement about the proportion of data values that must be within a specified number of standard deviations of the mean.
For example:
75% of the students must have a test score between 34 and 86

Rule:
At least **(1 - 1/z^2)** of the values in any data set will be
within z standard deviations of the mean, where z is
any value greater than 1.

Real power of this Theorem:
Applies to any data set regardless of the shape of the distribution of the data!

At least 0.75, or 75.00%, of the data values must be within z = 2 standard deviations of the mean

At least 0.89, or 89.00%, of the data values must be within z = 3 standard deviations of the mean

At least 0.94, or 94.00%, of the data values must be within z = 4 standard deviations of the mean

Empirical Rule:

The Empirical Rule can determine the proportion of data values that lie between +/- a given number of standard deviations (z-score) for a bell-shape or normal distribution (symmetric distribution)

Empirical Rule:

Approximately 68% of the data values will be within 1 Standard Deviation of the Mean

Approximately 95% of the data values will be within 2 Standard Deviations of the Mean

Approximately 99% of the data values (almost all the data) will be within 3 Standard Deviations of Mean

* MUCH more about Normal Distribution / Bell Shaped Distribution and the Empirical Rule later in the class.

Detecting Outliers

Outlier

An outlier is an unusually small or unusually large value in a data set.

+/- 3 Z Scores Outlier Rule

A data value with a z-score less than -3 or greater than +3 might be considered an outlier.

An Outlier can be:

1)an incorrectly recorded data value

2)a data value that was incorrectly included in the data set

3) a correctly recorded data value that belongs in the data set

In Excel, the ABS function determines the absolute value of a number (distance away from zero)

5 Number Summary

5 Number Summary Provides information like:

Minimum Value

Maximum Value

Median

Range (Max - Min)

Quartiles

Interquartile Range (Quart 3 - Quart 1 = range for middle 50%)

Basis for building Box Plot

Box Plot (Box and Whisker) Chart and Outliers Rule of 1.5

A Box Plot is a visual way to show the spread in the data. It visualizes the 5 Number Summary and shows outliers

In Excel use the: "Box and Whisker" Chart.

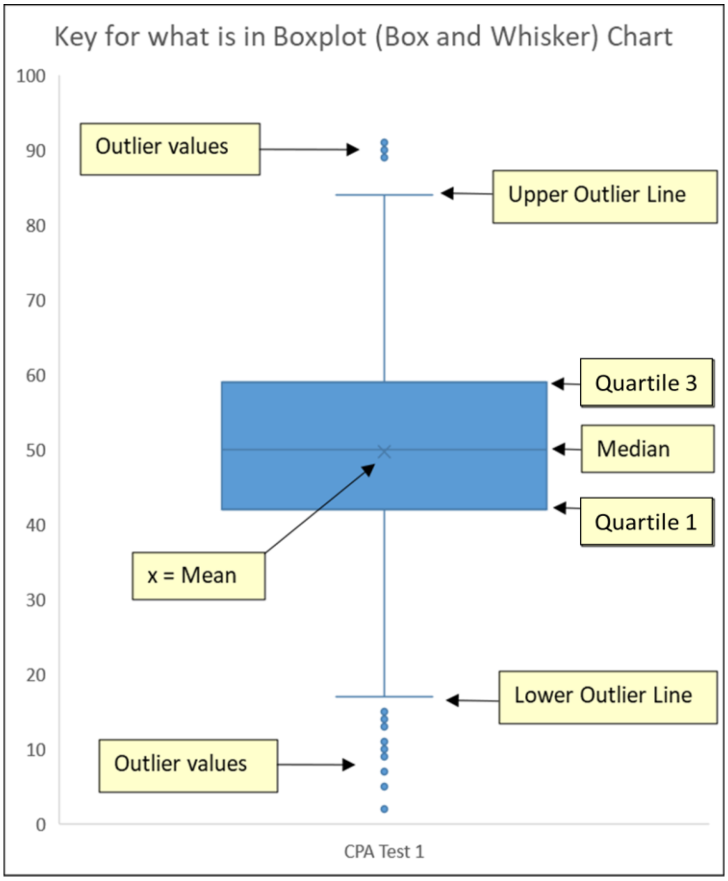
Horizontal lines for Lower and Upper Limits that indicate that anything past is an Outlier.

Lower Limit = **Quartile 1 - 1.5*Interquartile Range**, or 0.

Upper Limit = **Quartile 3 + 1.5*Interquartile Range**

**Although for this class we will use 1.5, other multiples can be used

Example of Excel Box and Whisker Chart:



Excel Notes:

Dynamic spilled array formulas details:

For dynamic spilled array formulas, the formula lives in the top left cell of the spilled range and must be edited in that cell. Spilled results below top cell are greyed out in the formula bar.

The spilled values can be dynamically referred to with the cell address of the top left cell in the spilled range and the spilled range operator, #. For example, if the spilled array formula resides in cell E3, the spilled range reference is #E3.

If there is data in the way of a spilled array, a #SPILL! error is displayed in the top left cell.

Dynamic spilled array formulas are dynamic because when the source data changes and the spilled array expands or contracts, the values emanating from the top left cell expand and contract accordingly.

Advantages of Dynamic Spilled Array Formulas vs. "Old School" formulas:

Don't have to lock cell reference

Don't have to copy formula

Editing is done in top cell and edited formula automatically spills down

Dynamic Spilled Array Worksheet Functions used in Video:

The **MODE.MULT array function** which delivers one or more modes.

The **UNIQUE array function** can deliver unique set of items from a table, a column, or a row. A unique set of items is when you select and list only items that occur exactly one time in the data set.

The **SORT array function** can sort a row, a column or a table in ascending or descending order. The default sort order is ascending (smallest to biggest).

The **SEQUENCE** array function generates a sequence of numbers in a row, a column or a table, based on the formula inputs for the number of rows in the final sequence, the number of columns in the final sequence, a start value and an increment value.

The **FILTER** array function allows you to filter a set of values to show only that values that meet a logical test. The **array** argument contains the values that you want to filter. The **include** argument requires an array of TRUE and FALSE values (same dimension as array argument values) to indicate which values to keep (TRUE) and with ones to filter out (FALSE).

Other Worksheet Functions used in Video:

The **COUNTIFS function** makes a conditional count calculation based on one or more logical tests. The **criteria_range1** argument contains the full range with all the conditional items. The **criteria1** argument contains one or more conditions for counting items from the criteria_range1 argument. You can place one or more conditions into the criteria1 argument: when you place one condition into the criteria1 argument, the COUNTIFS function delivers a single answer, but when you place more than one condition into the criteria1 argument, the COUNTIFS function spills an array of answers into the worksheet, one for each condition. If you need to use a comparative operator with the condition, you must join the comparative operator to the cell with the condition, like: "<"&J27. You can have up to 127 pairs of criteria_rangeN criteriaN arguments that will run an AND Logical Test to make the conditional count calculation.

Comparative Operator Note:

* When using comparative operators in functions like COUNTIFS, SUMIFS, AVERAGEIFS, MINIFS and MAXIFS, you must join the comparative operator to the cell with the condition, like: ">"&J28.

* But when you use a comparative operator in a formula that makes a direct logical test formula calculation, you do not use quotes or an ampersand (join operator), like: CPAScoreTable[CPA Test 1]<J27. Example of this note is in video #15.

Unpivot Power Query feature to convert an Improper Data Set to a Proper Data Set:

The Improper Data Set has a unique list of elements from a single variable column show as column headers, with data points below each column header. The Improper Data Set is more difficult to deal with when performing data analysis. The Proper Data Set lists only the variable fields and does not show elements from a variable field as column headers. A Proper Data Set shows only variable fields, in the CPA example: CPATest and Score. The feature that you use to convert the Improper Data Set to a Proper Data Set is the Unpivot feature.

Variability

Variability answers the questions:

What is the dispersion in the data?

How spread out is the data?

1) Range

Range = Max value - Min value

The bigger the range, the more the variability or spread in the data

It is simple to calculate, but sensitive to extreme values

2) Interquartile Range

Interquartile Range = Quartile 3 - Quartile 1

The bigger the range, the more the variability or spread in the data

The Interquartile Range tells you the range for the middle 50% of the data. It overcomes the sensitivity to extreme values that range has

3) Deviation

Deviation = Particular X value - Mean

It tells you how far one datum is from the mean. "How far above or below the mean is the particular value?"

For any data set, the sum of the deviations is always zero!!!!

This is why mathematically, we either square (Variance or Standard Deviation) or take the absolute value (Mean Absolute Value)

4) Variance

A Numerical Measure that says how much variability there is in the data points

Variance uses all the data points, not just some like Range and Interquartile Range

Variance has squared units, which makes interpreting it difficult. Standard Deviation undoes the squared units and is thus easier to interpret and is more commonly used

Math Formulas below ↓

In Excel use: VAR.P function for population data and VAR.S for sample data

5) Standard Deviation

Standard Deviation uses all the data points, not just some like Range and Interquartile Range

Standard Deviation does not have squared units (like Variance) and is thus easier to interpret; the standard deviation has the same units as the data.

Population Standard Deviation = "little sigma" = s ; Sample Standard Deviation = s

The sample standard deviation is a point estimator of the population standard deviation

Interpretation of Standard Deviation:

A numerical measure that says how much variability there is in the data points

Standard Deviation is like an average of the deviations

* Standard Deviation tells us how fairly the mean represents its data points (key concept is statistics)

* Standard Deviation tells us how clustered the data points are around the mean (key concept is statistics)

For financial assets standard deviation is a measure of risk or fluctuation in asset value

Math Formulas below ↓

In Excel use: STDEV.P function for population data and STDEV.S for sample data

6) Coefficient of Variation

Math Formula = (Standard Deviation)/Mean = Coefficient of Variation

Coefficient of Variation converts Standard Deviation to Standard Deviation per unit of Mean so you can compare:

1) Data in different units.

or

2) Data in the same units, but the means are far apart.

Coefficient of Variation answers the question: "For every one unit of mean, what is the Standard Deviation?"

In Finance you see the inverse formula = (Average Return)/(Standard Deviation) = Return for 1 unit of risk

$$Population\ Variance = \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

$$Sample\ Variance = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$Population\ Standard\ Deviation = \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

$$Sample\ Standard\ Deviation = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

These two not required for this class:

Alternative for variability measure called Mean Absolute Error

$$MAE = \frac{\sum |x_i - \bar{x}|}{n}$$

Alternative to calculate Sample SD

$$Sample\ Standard\ Deviation = s\sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$

Xbar	=	"Xbar"	Sample Mean
μ	=	"Mew"	Population Mean
s	=	"s"	Sample Standard Deviation
σ	=	"sigma"	Sample Standard Deviation
xi	=	"x sub i"	Particular Value
∑	=	"Sigma"	Greek Letter used for "adding"
n	=	"n"	Count of sample items
N	=	"n"	Count of population items

	A	B	C	D	E	F
1						
2		**Sample Standard Deviation = s**	$3.53		Cell C2: =STDEV.S(B8:B30)	Use STDEV.S for calculating Sample Standard Deviation = s
3		**Population Standard Deviation = σ**	$3.65		Cell C3: =STDEV.P(D$8:D$111)	Use STDEV.P for calculating Population Standard Deviation = σ
4						
5		Wage Data				
6						
7		Sample1		Population 1		
8		$22.35		$15.35		
9		$14.01		$32.16		
10		$15.20		$12.55		
11		$18.30		$13.74		
12		$13.63		$22.35		
13		$14.08		$15.40		
14		$15.44		$14.30		
15		$10.75		$14.79		
16		$14.67		$20.94		
17		$11.29		$11.91		
18		$15.02		$12.40		
19		$17.97		$12.05		
20		$22.12		$18.11		
21		$22.12		$14.83		
22		$20.68		$13.93		
23		$13.21		$11.66		
24		$18.74		$12.69		
25		$13.66		$14.01		
26		$12.94		$12.52		
27		$12.94		$14.35		
28		$13.34		$15.20		
29		$13.68		$12.68		
30		$12.36		$12.32		
31				$15.65		
105				$12.32		
106				$13.34		
107				$14.52		
108				$23.70		
109				$13.21		
110				$13.68		
111				$12.36		
112						

$$Sample\ Standard\ Deviation = s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$Population\ Standard\ Deviation = \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

	G	H	I	J	K
5		Xbar	=	"Xbar"	Sample Mean
6					
7		µ	=	"Mew"	Population Mean
8		s	=	"s"	Sample Standard Deviation
9		σ	=	"sigma"	Sample Standard Deviation
10		xi	=	"x sub i"	Particular Value
11		∑	=	"Sigma"	Greek Letter used for "adding"
12		n	=	"n"	Count of sample items
13		N	=	"n"	Count of population items

Relative Location

The goal of relative location is to determine, in a data set, how far a particular x value is from the mean.

Z-Score = Standardized Value

Z-score measures the relative location of a particular x in the data set (as compared to the mean), in units of standard deviation.

Z-Score tells you "How Many Standard Deviations The Particular Value Is Away From The Mean". 1? -1? 1.5? 0?

zi < 0, value below meanzi > 0, value above meanzi = 0, value is equal to mean

Observations in 2 different data sets that have the same z-score are said to have the same relative location or the same number of standard deviations away from the mean.

In Excel use the formulas:

Z-Score = (Particular Value - Mean)/StandardDeviation = zi = (xi-Xbar)/s

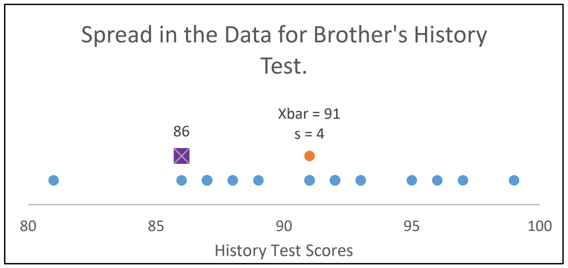
Use STANDARDIZE function to calculate z-score. Function arguments: STANDARDIZE(a,mean, standard_dev) = STANDARDIZE(ParticularXiValue,Mean,StandardDeviation

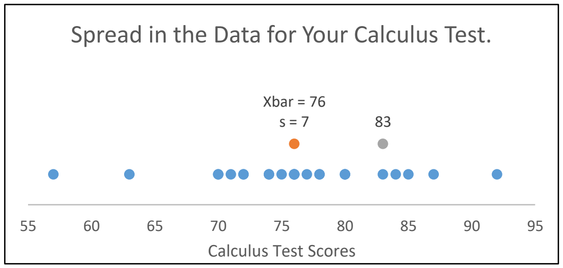
Xbar	=	Sample Mean	"Xbar"
xi	=	Particular Value	"x sub i"
s	=	Sample Standard Deviation	"s"
z	=	Number of Standard Deviations Away From Xbar	"z"

Who did better? You or your Brother?	
Your Calculus Test score, Xi =	83
Mean for Calculus Test, Xbar =	76
Standard Deviation for Calculus Test, s =	7
z = (x - Xbar)/s =	
STANDARDIZE function	
Brother's History Test score, Xi =	86
Mean for History Test, Xbar =	91
Standard Deviation for History Test, s =	4
z = (x - Xbar)/s =	
STANDARDIZE function	
Joe's Calculus Test score, Xi =	76
z score	

$$z = \frac{Xi - Xbar}{s} =$$

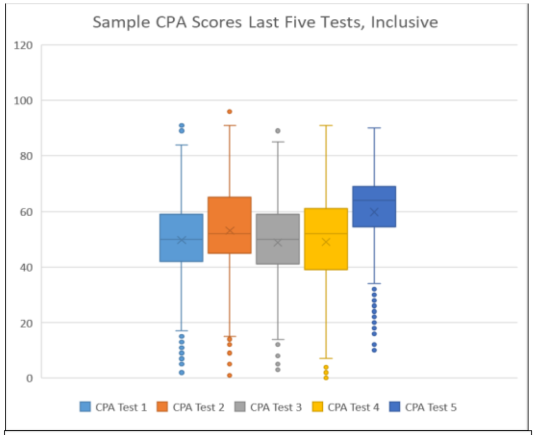
Number of Standard Deviations Away From Xbar

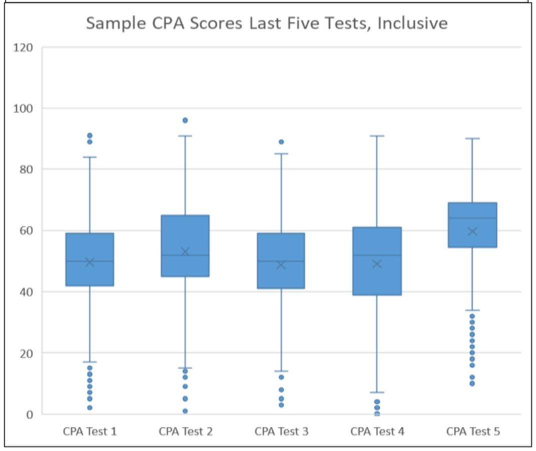


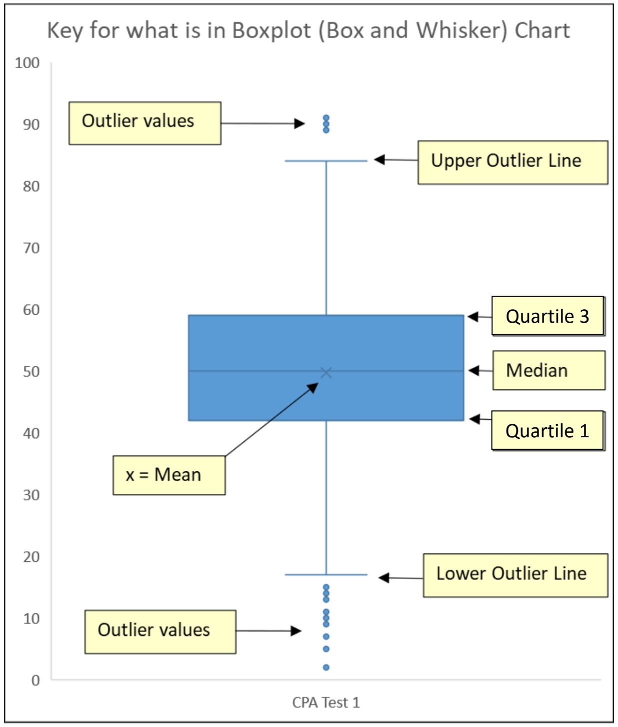


Z-score measures the relative location of a particular x in the data set (as compared to the mean), in units of standard deviation.

	CPA Test 1	CPA Test 2	CPA Test 3	CPA Test 4	CPA Test 5
Min	2	1	3	0	10
Quartile 1	42	45	41	39	54.5
Median	50	52	50	52	64
Quartile 3	59	65	59	61	69
Max	91	97	89	91	90
Interquartile Range	17	20	18	22	14.5
Lower Outlier Hurdle	16.5	15	14	6	32.75
Upper Outlier Hurdle	84.5	95	86	94	90.75
Count Lower Outliers	11	14	11	5	36
Count Upper Outliers	3	2	1	0	0
Outlier Multiple	1.5				
Mean	49.67885117	53.12793734	48.79895561	49.04960836	59.76501305







Unpivot Power Query feature to convert an Improper Data Set to a Proper Data Set:

The Improper Data Set has a unique list of elements from a single variable column show as column headers, with data points below each column header. The Improper Data Set is more difficult to deal with when performing data analysis. The Proper Data Set lists only the variable fields and does not show elements from a variable field as column headers. A Proper Data Set shows only variable fields, in the CPA example: CPATest and Score. The feature that you use to convert the Improper Data Set to a Proper Data Set is the Unpivot feature.

Improper Data Set with a Pivoted Column (Field)

CPA Test 1	CPA Test 2	CPA Test 3	CPA Test 4	CPA Test 5
22	28	43	35	55
49	14	32	45	53
70	5	52	61	30
13	65	5	37	68
52	80	51	52	58
44	67	62	55	31
18	81	56	45	61
52	47	31	7	57
49	51	47	44	87
37	67	12	57	60
78	66	64	58	66
75	81	61	50	66
70	63	49	53	64
51	15	67	39	32
50	54	34	34	62
70	50	89	60	17

Proper Data Set with only Variable Fields

CPATest	Score
CPA Test 1	22
CPA Test 2	28
CPA Test 3	43
CPA Test 4	35
CPA Test 5	55
CPA Test 1	49
CPA Test 2	14
CPA Test 3	32
CPA Test 4	45
CPA Test 5	53
CPA Test 1	70
CPA Test 2	5
CPA Test 3	52
CPA Test 4	61
CPA Test 5	30
CPA Test 1	13
CPA Test 2	50
CPA Test 3	13
CPA Test 4	50
CPA Test 5	61