DR. ALVIN'S PUBLICATIONS

DESCRIPTIVE STATISTICAL MEASURES

DR. ALVIN ANG

CONTENTS

PART I

DRAWING SIMPLE VISUALS

A. HOW TO DRAW SIMPLE BAR CHART

T-Shirt Size	Frequency
XS	68
S	136
M	170
L	272
XL	34

Figure 1: T-Shirt Sizes

Step 1: Highlight the Data
Step 2: Click on the Insert Tab
Step 3: Click on the Down Arrow showing 2-D Column
Step 4: Edit and Adjust Accordingly

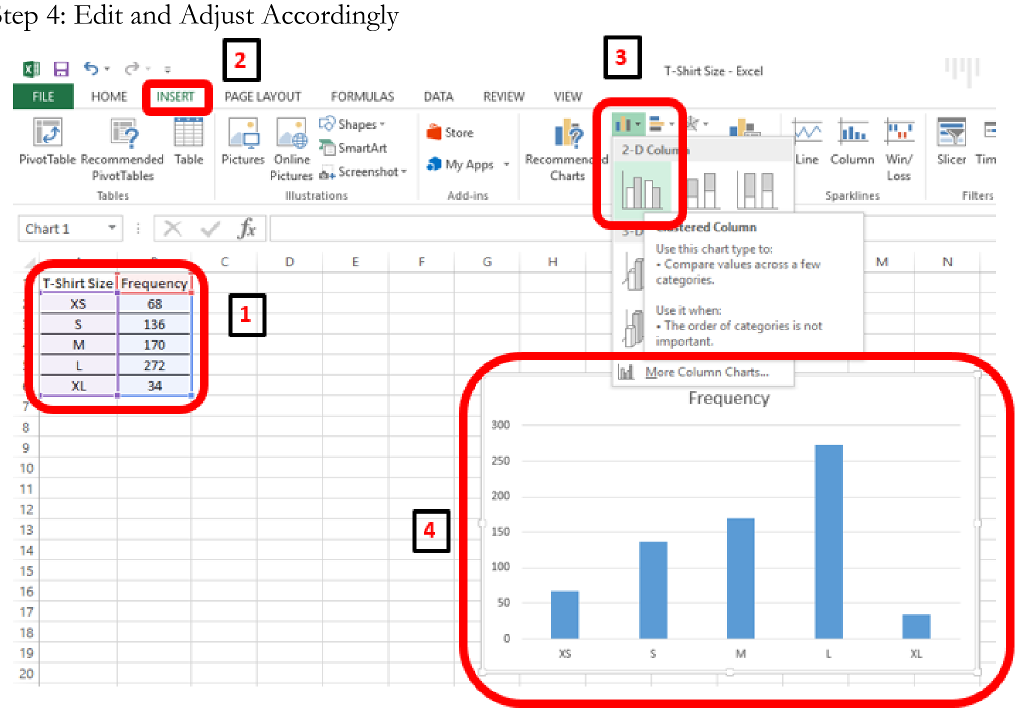


Figure 2: How to Draw Bar Chart

B. HOW TO DRAW SIMPLE PIE CHART

Step 1: Highlight the Data
Step 2: Click on the Insert Tab
Step 3: Click on the Down Arrow showing 2-D Column
Step 4: Edit and Adjust Accordingly

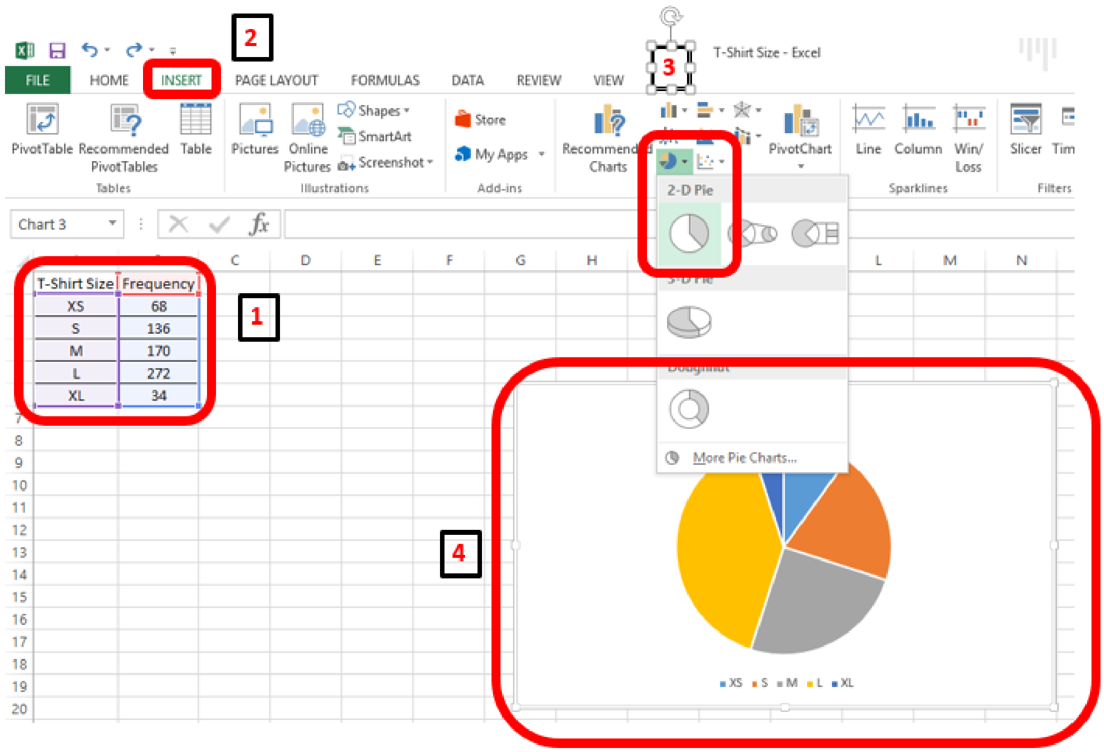


Figure 3: How to Draw Pie Chart

C. HOW TO DRAW SIMPLE HISTOGRAM

Step 1: This requires installing an additional add-on within Excel, the "Data Analysis Tool pak". Refer to Figure 24: Installing the Excel Analysis Tool Pak.

Step 2a: Given a set of data

Step 2b: Click on the Data Tab

Step 2c: Click on Data Analysis

Step 2d: Select the Histogram Option and click OK

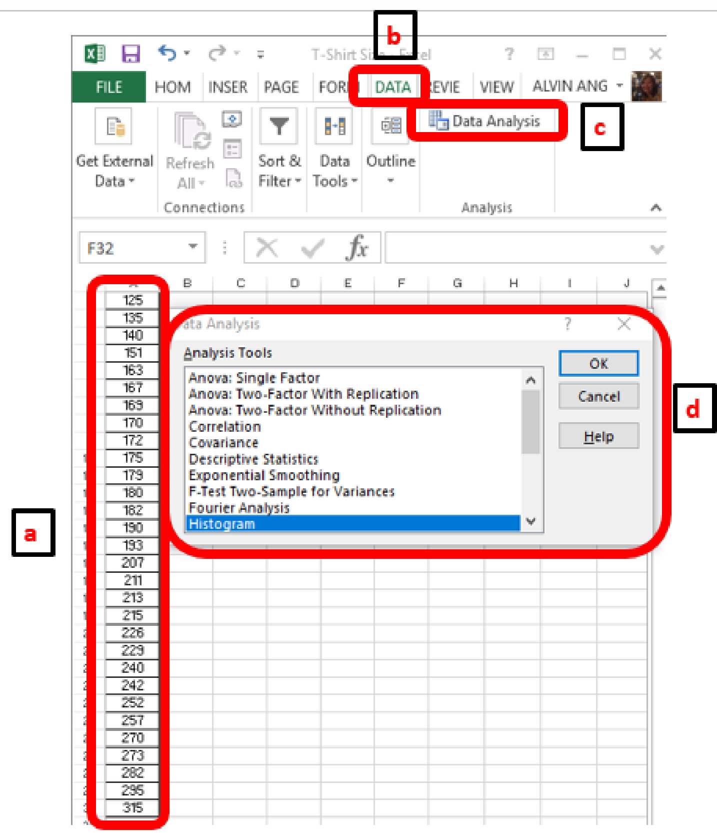


Figure 4: Navigating to the Histogram Option

Step 3a: For the Input Range, select all the data.

Step 3b: For the Output Range, select any empty cell on the sheet. We select C1 for now since its empty. It will appear as the top left hand corner for the output.

Step 3c: Select "Chart Output" and then click OK

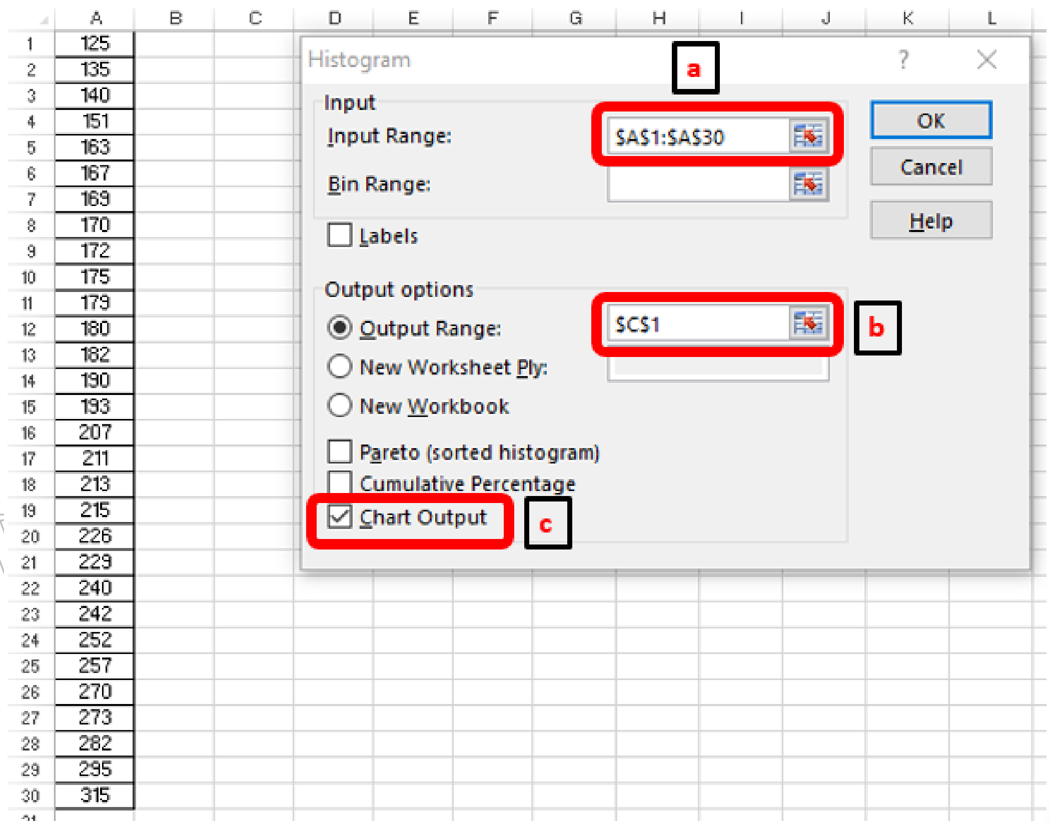


Figure 5: Setting the Histogram Parameters

Step 4: The output as shown

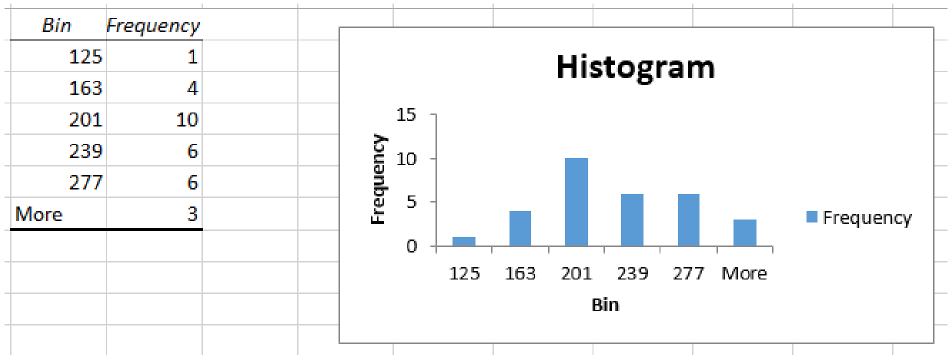


Figure 6: Histogram

Step 5: We are not satisfied with the Histogram drawn in Figure 6, thus we will do more editing.

Step 6a: We repeat the steps in Step 2, only this time, we create the Bins as shown in Figure 7.

Step 6b: We input the Bin Range.

Step 6c: We change the Output Range to cell E1. Anywhere on the spreadsheet, as long as it is an empty cell, is ok to select. Then click OK.

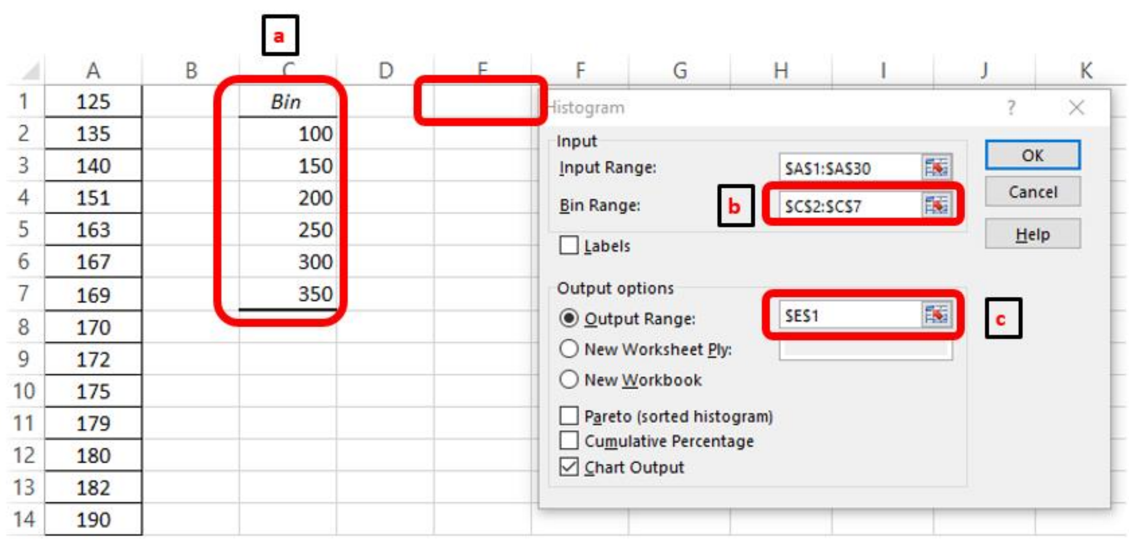


Figure 7: Setting up the Bins for Advanced Histogram

Step 7: New Output as shown.

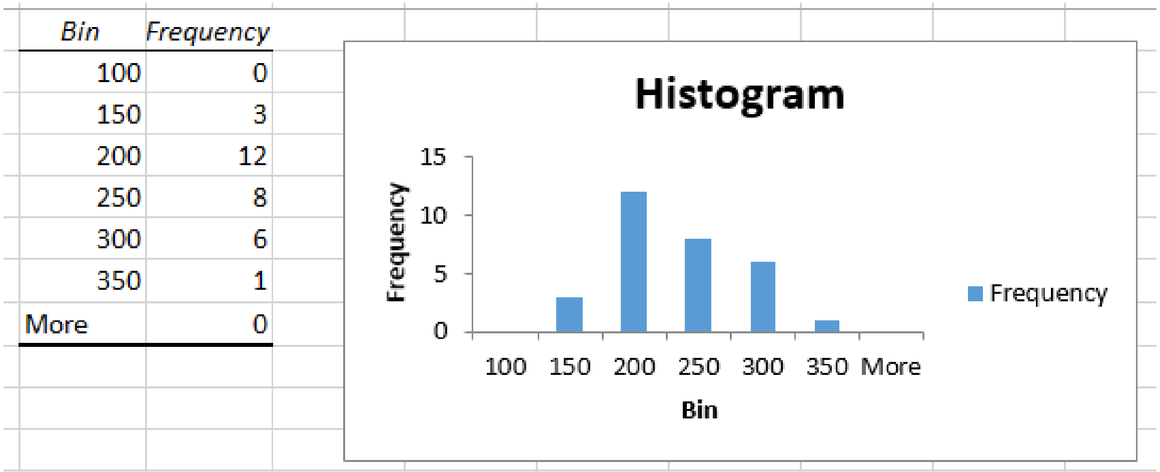


Figure 8: New Histogram with New Bins

Step 8a: Right click on any blue area within the rectangle.

Step 8b: Click on Format Data Series. A new side bar will appear.

Step 8c: Change the Gap Width to Zero.

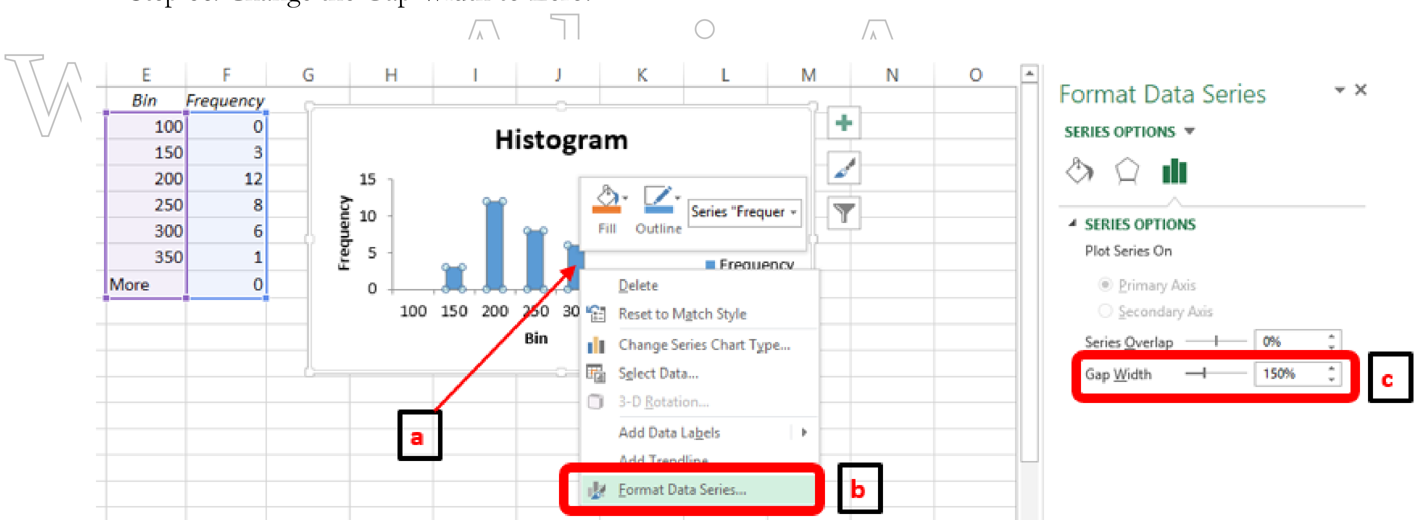


Figure 9: Adjusting the Bin Width

Bin	Frequency
100	0
150	3
200	12
250	8
300	6
350	1
More	0

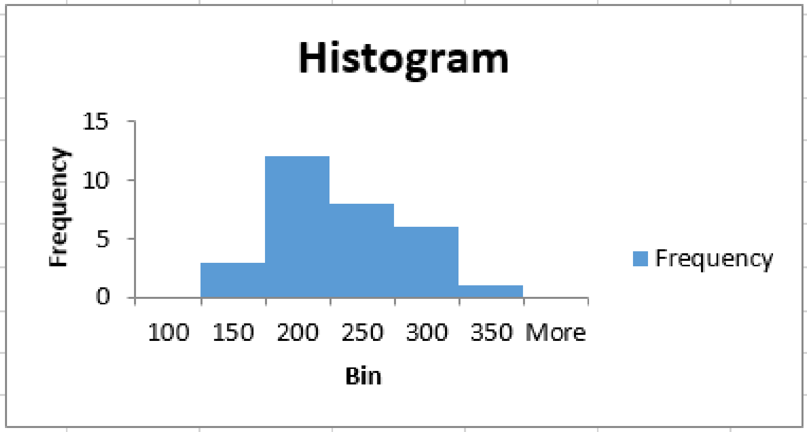


Figure 10: Final Histogram

Figure 10 shows how the final histogram looks like. Further editing to the main title, side title and legend can be done accordingly.

PART II

MEASURES OF LOCATION

D. MEAN

I. ARITHMETIC MEAN

For Population: $\mu = \frac{\sum X}{N}$

Where:

- μ: Population Mean
- X: Any Value
- N: Number of Items in Population

For Sample: $\overline{X} = \frac{\sum X}{n}$

Where:

- $\overline{X}$: Sample Mean
- X: Any Value
- n: Number of Items in Sample

II. WEIGHTED MEAN

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

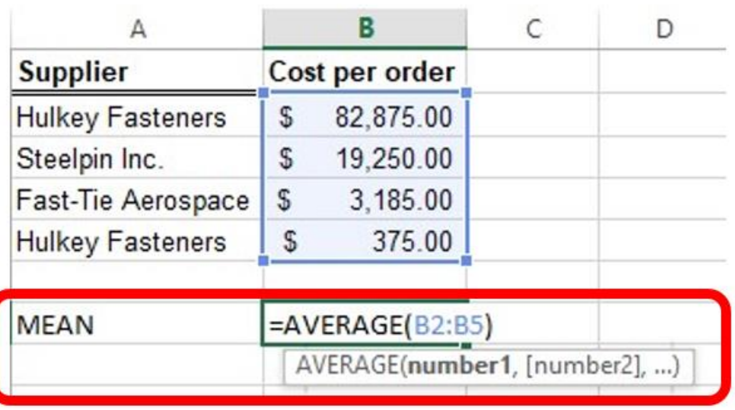
Where:

- $\overline{X}_w$: Weighted Mean
- w_i: Weight for that particular 'I'
- X_i: Value Associated for that particular 'i'

Weighted Mean Example:

Given:
- A hospital has 10 nurses
- 2 nurses earn \$14 per hour
- 3 nurses earn \$18 per hour
- 5 nurses earn \$28 per hour
- Weighted Mean = $\frac{(2\times\$14)+(2\times\$18)+(5\times\$28)}{2+3+5}$ = \$22.20 (ANS.)

How to use Excel to obtain Mean:
Use the “Average” Function



	A	B	C	D
	Supplier	**Cost per order**		
	Hulkey Fasteners	\$ 82,875.00		
	Steelpin Inc.	\$ 19,250.00		
	Fast-Tie Aerospace	\$ 3,185.00		
	Hulkey Fasteners	\$ 375.00		
	MEAN	=AVERAGE(B2:B5)		
		AVERAGE(**number1**, [number2], ...)		

Figure 11: Using the "Average" Function to obtain Mean

E. MEDIAN

HOW TO OBTAIN THE MEDIAN

1) Arrange all Data Values from Smallest to Largest
2) Select the Middle Value (this is the Median)
3) Half the observations are above the median and half are below it
4) If we have an even number of values, the Median is the average of the two middle numbers.

Median is a useful measure when we encounter data with an extreme value.
Example:

$115,000 $118,000 $126,000 $135,000 $350,000

↑ *median*

HOW TO USE EXCEL TO OBTAIN MEDIAN: USE THE "MEDIAN" FUNCTION

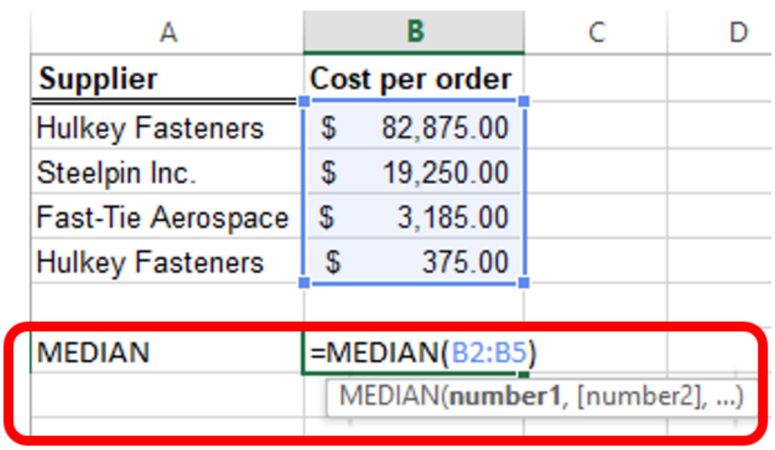
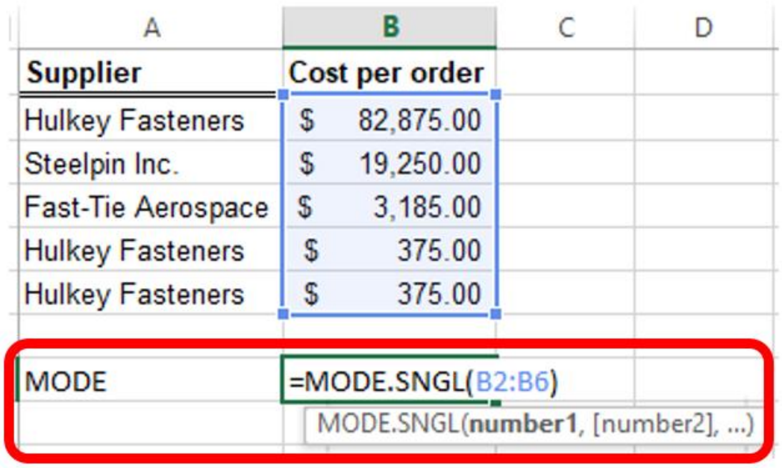


Figure 12: Using the "Median" Function to obtain Median

F. MODE

The mode is the observation that occurs most frequently. You can easily identify the mode from a frequency distribution by identifying the value having the largest frequency or from a histogram by identifying the highest bar. You may also use the Excel function MODE.SNGL(data range).



A	B	C	D
Supplier	Cost per order		
Hulkey Fasteners	$ 82,875.00		
Steelpin Inc.	$ 19,250.00		
Fast-Tie Aerospace	$ 3,185.00		
Hulkey Fasteners	$ 375.00		
Hulkey Fasteners	$ 375.00		
MODE	=MODE.SNGL(B2:B6)		
	MODE.SNGL(**number1**, [number2], ...)		

Figure 13: Using the "MODE.SNGL" to obtain the Mode

G. COMPARISON OF MEAN VS MEDIAN VS MODE

MEAN	***MEDIAN***	***MODE***
Advantages: • ***Most widely used measure of location*** • ***Easiest to understand and apply*** • ***All data values are included in the calculation*** • ***The mean is unique – there is only one mean for a set of data.*** • ***The sum of deviations of each value from the mean will always be zero i.e.*** $\sum\left(X-\overline{X}\right)=0$	Advantages: • Useful if we encounter data with extreme value/s. • Not affected by extremely large or small values. • The median is unique – there is only one median for a set of data.	Advantages: • Not affected by extremely large or small values. • Very easy to use. • Quite popular. • All data values are used in the calculation. • Most useful for data sets that contain small number of unique values
Disadvantages: • ***Mean gets affected by extreme value/s in dataset.***	Disadvantages: • Not all data values are included in the calculation. • The sum of deviations of each value from the median is not zero. • Not popular.	Disadvantages: • For data sets that have few repeating values, the mode does not provide much practical value

PART III

MEASURES OF DISPERSION

Dispersion is a measure of the spread of data. A small value for a measure of dispersion indicates that the data are clustered closely, say, around the arithmetic mean. Thus the mean is considered representative of the data, that is, it is reliable. Conversely, a large measure of dispersion indicates that the mean is not reliable and is not representative of the data.

A. RANGE

Range = Largest Value – Smallest Value

Advantages of using Range	*Disadvantages of using Range*
✓ ***It is easy to compute and understand.*** ✓ ***Only two values are used in the calculation***	✓ It is influenced by extreme values.

HOW TO OBTAIN RANGE USING EXCEL: USE "MAX" & "MIN" FUNCTIONS.

	A	B	C
	Supplier	**Cost per order**	
	Hulkey Fasteners	$ 82,875.00	
	Steelpin Inc.	$ 19,250.00	
	Fast-Tie Aerospace	$ 3,185.00	
	Hulkey Fasteners	$ 375.00	
	RANGE	=MAX(B2:B5) - MIN(B2:B5)	

Figure 14: Using "MAX" & "MIN" to obtain Range

B. VARIANCE AND STANDARD DEVIATION

POPULATION VARIANCE AND POPULATION STANDARD DEVIATION

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

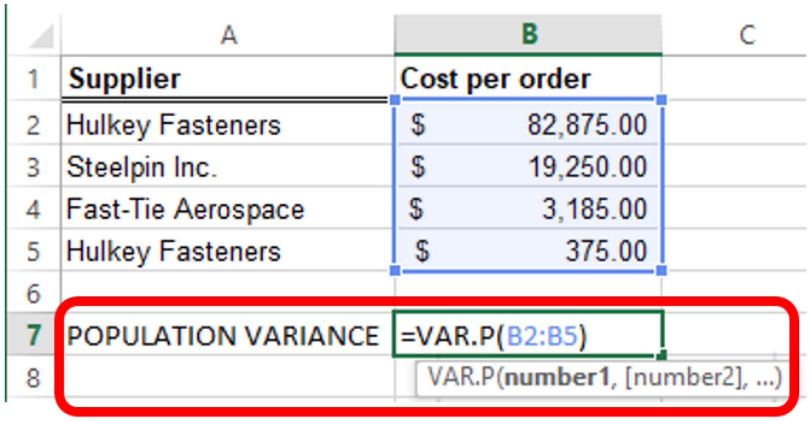
Where:

- σ^2 : Population Variance
- X : Observed Value in the Population
- μ : Mean of the Population
- N : Total number of Observations in the Population
-

The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations.

Advantages of using Variance	*Disadvantages of using Variance*
✓ ***Not distorted by extreme observations.*** ✓ ***All observations are used in the calculations.*** ✓ ***The squaring of the difference between X and μ helps by:*** i. ***Removing any negative differences*** ii. ***Any difference that is <1 becomes much smaller, and ignored. Any difference >1 is amplified and taken into account largely.***	✓ Units are difficult to work with because they are "Units Squared" – for e.g. Dollars² – which does not make any sense.

HOW TO OBTAIN POPULATION VARIANCE USING EXCEL: USE "VAR.P" FUNCTION



	A	B	C
1	Supplier	Cost per order	
2	Hulkey Fasteners	$ 82,875.00	
3	Steelpin Inc.	$ 19,250.00	
4	Fast-Tie Aerospace	$ 3,185.00	
5	Hulkey Fasteners	$ 375.00	
6			
7	POPULATION VARIANCE	=VAR.P(B2:B5)	
8		VAR.P(number1, [number2], ...)	

Figure 15: Using "VAR.P" to obtain Population Variance

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

Where:

- σ : Population Variance
- X : Observed Value in the Population
- μ : Mean of the Population
- N : Total number of Observations in the Population

A small standard deviation indicates that the data are clustered close to the mean, thus the mean is representative of the data. A large standard deviation indicates that the data are spread out from the mean and the mean is not as representative of the data.

Advantages of using Standard Deviation

- ✓ ***Easier to interpret than the variance because it uses the original units of measurement (e.g. Dollars, not Dollars²)***
- ✓ ***It is the positive square root of the Variance.***
- ✓ ***Easier to relate to the Mean. More widely used than Variance.***

HOW TO OBTAIN POPULATION STD. DEV. USING EXCEL: USE "STDEV.P" FUNCTION

	A	B
1	**Supplier**	**Cost per order**
2	Hulkey Fasteners	$ 82,875.00
3	Steelpin Inc.	$ 19,250.00
4	Fast-Tie Aerospace	$ 3,185.00
5	Hulkey Fasteners	$ 375.00
6		
7	POPULATION STD. DEV.	=STDEV.P(B2:B5)

Figure 16: Using "STDEV.P" to obtain Population Standard Deviation

SAMPLE VARIANCE AND SAMPLE DEVIATION

Sample Variance: $$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1}$$

Where:

- s^2 : Sample Variance
- X : Observed Value in the Sample
- $\overline{X}$: Mean of the Sample
- n : Total number of Observations in the Sample

Why is the denominator changed to (n – 1)? This is because statisticians have shown that this provides a more accurate representation of the true population variance. The use of (n –1) in the denominator provides an appropriate correction factor since "n" tends to underestimate the population variance.

HOW TO OBTAIN SAMPLE VARIANCE USING EXCEL: USE “VAR.S” FUNCTION

	A	B
1	**Supplier**	**Cost per order**
2	Hulkey Fasteners	$ 82,875.00
3	Steelpin Inc.	$ 19,250.00
4	Fast-Tie Aerospace	$ 3,185.00
5	Hulkey Fasteners	$ 375.00
6		
7	Sample Variance	=VAR.S(B2:B5)

Figure 17: Using "VAR.S" to obtain Sample Variance

Sample Std. Dev. $s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n-1}}$

Where:

- s: Sample Std. Dev.
- X : Observed Value in the Sample
- $\overline{X}$: Mean of the Sample
- n : Total number of Observations in the Sample

HOW TO OBTAIN SAMPLE STD. DEV. USING EXCEL: USE “STDEV.S” FUNCTION

	A	B
1	**Supplier**	**Cost per order**
2	Hulkey Fasteners	$ 82,875.00
3	Steelpin Inc.	$ 19,250.00
4	Fast-Tie Aerospace	$ 3,185.00
5	Hulkey Fasteners	$ 375.00
6		
7	Sample Variance	=STDEV.S(B2:B5)

Figure 18: Using "STDEV.S" to obtain Sample Standard Deviation

C. BOX PLOT, INTERQUARTILE RANGE (IQR), PERCENTILE

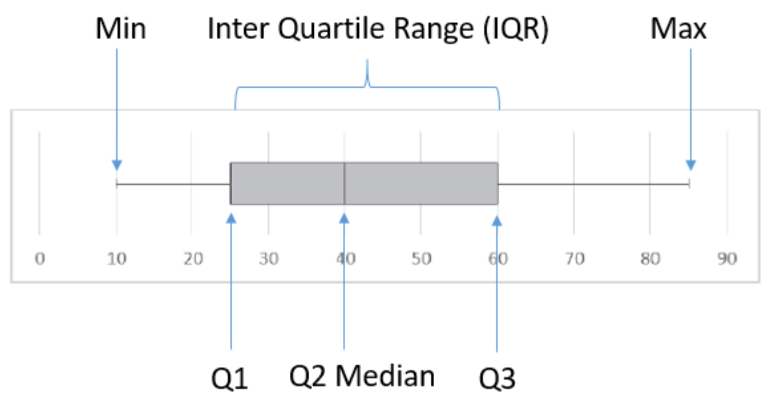


Figure 19: A Box Plot

Figure 19 shows a Box Plot. It shows:

1) The Minimum Value = 10
2) The Maximum Value = 85
3) The 1st Quartile (25%) = 25 (also called the 25th Percentile)
4) The 2nd Quartile (50%) = 40 (also called the Median, or the 50th Percentile)
5) The 3rd Quartile (75%) = 60 (also called the 75th Percentile)
6) The IQR = Q3 – Q1 = 60-25 = 35

We can calculate the "Outlier Zones" by:

a. ZONE 1 = (Q1 – 1.5*IQR) = (25 – 1.5*35) = -27.5
   - ✓ If Value < Zone 1 → Outlier

b. ZONE 2 = (Q3 + 1.5*IQR) = (60 + 1.5*35) = 112.5
   - ✓ If Value > Zone 2 → Outlier

*Excel 2013 does not support Box Plot Charts.

*Go here to see how to build Box Plots: https://www.contextures.com/excelboxplotchart.html

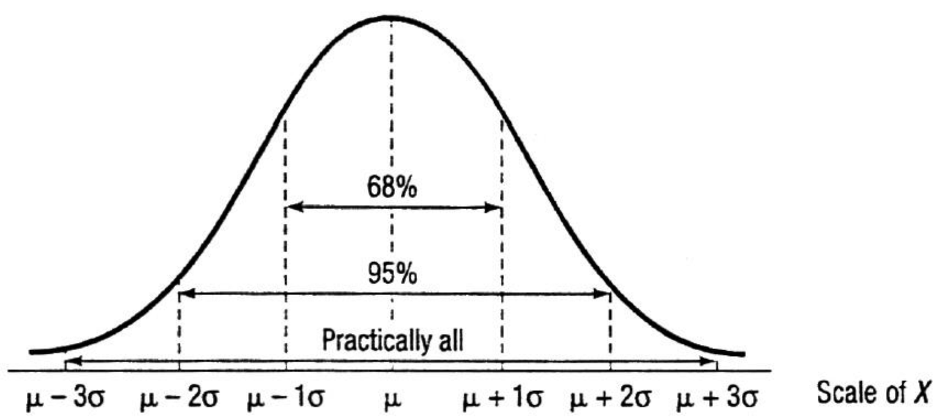
D. CHEBYSHEV'S THEOREM AND EMPIRICAL RULE

Chebyshev's Theorem:

- ✓ Let k be the number of Std. Deviations (k>1)
- ✓ CHEBYSHEV THEOREM: The Proportion of Values (or Percentage of Observations) that lie within k is $\leq 1-\frac{1}{k^2}$
- ✓ This theorem holds for all types of observations, regardless of the shape of its distribution.
- ✓ E.g. for k = 2 Std. Dev. → ≥75% of the data lie within the 2 Std. Dev.
- ✓ E.g. for k = 3 Std. Dev. → ≥89% of the data lie within the 3 Std. Dev.

Empirical Rule:

- ✓ Empirical Rule is derived from Chebyshev Theorem.
- ✓ Chebyshev Theorem holds for ***ALL*** distribution types, but Empirical Rule holds only for Normal or Approximately Normal Distribution.
- ✓ Figure 20 below shows the Empirical Rule:



✓ *Figure 20: Empirical Rule holds for Normal Distribution Curve*

- ✓ EMPIRICAL RULE:

  1) Within One Std. Dev. of the mean ($\mu \pm 1\sigma$) → ≈ 68% of all observations will lie within the area under the normal curve.

  2) Within Two Std. Dev. of the mean ($\mu \pm 2\sigma$) → ≈ 95% of all observations will lie within the area under the normal curve.

  3) Within Three Std. Dev. of the mean ($\mu \pm 3\sigma$) → ≈ 99.7% of all observations will lie within the area under the normal curve.

  4) Actual % may be higher or lower, depending on the shape of the distribution.

- ✓ To describe variability of practical data → 2~3 Std. Dev. around the mean are commonly used.

- ✓ E.g. suppose an order is delivered at an average of 8 days with a Std. Dev. of 1 day. Using the 2nd Empirical Rule, you can tell a customer with 95% Confidence that their package should arrive within 6 to 10 days.

EXAMPLE 1 FOR CHEBYSHEV'S THEOREM AND EMPIRICAL RULE

Given:
- Sample Mean Income, μ = $72,000
- Sample Std. Dev., s = $4,000

Find:
- % who earn $64,000 < X < $80,000

Answer:
- Step 1: Chebyshev Theorem:
- $k = \frac{X - \overline{X}}{s} = \frac{\$64,000 - \$72,000}{\$4,000} = -2$
- $k = \frac{X - \overline{X}}{s} = \frac{\$80,000 - \$72,000}{\$4,000} = 2$
- Formula: $1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75$
- 75% earns $64,000 < X < $80,000

- Step 2: Empirical Rule:
- If the distribution is normal, Rule 2 states: ($\mu \pm 2\sigma$) → ≈ 95% percent of all observations will lie within the area under the normal curve.
- Therefore 95% earns $\overline{X} \pm 2s$ = $72,000 ± 2($4,000) → $64,000 < X < $80,000

EXAMPLE 2 FOR EMPIRICAL RULE: THE PROCESS CAPABILITY INDEX (C_P)

The Process Capability Index (C_p) is a practical application of the Empirical Rule. C_p is used by manufacturers to evaluate the quality of their products. A C_p value less than 1.0 is not good; it means that the variation in the process is wider than the specification limits, signifying that some of the parts will not meet the specifications. In practice, many manufacturers want to have C_p values of at least 1.5. Figure 21 demonstrates how C_p can be implemented in Excel.

$$C_p = \frac{\text{Upper Specification - Lower Specification}}{\text{Total Variaion}}$$

Step 1: Data Set is given

Step 2: Mean and Std. Dev. Calculated using Excel (together with 3rd Empirical Rule, which is μ ± 3σ)

Manufacturing Measurements

5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11
5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11
4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08
5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03
5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94
5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88
4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91
5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86
4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89
5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91
5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87
5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93
4.85	5.04	5.12	4.97	5.02	4.97	5.02	5.14
4.90	5.09	4.89	5.07	4.99	5.04	5.03	4.87
4.97	5.07	4.91	5.03	5.02	4.94	5.18	4.98
5.09	4.99	4.97	4.81	5.03	4.98	5.08	4.88
4.89	5.01	4.98	4.95	5.02	5.03	5.14	4.88
4.87	4.88	5.01	4.89	5.07	5.05	4.92	5.01
5.01	4.93	5.01	5.08	4.95	4.91	4.97	4.93
4.97	5.10	5.09	4.93	4.95	5.09	4.92	4.93
4.76	4.94	4.93	4.99	4.94	5.21	5.14	4.99
4.94	4.88	5.04	4.94	5.12	4.87	4.92	4.91
4.92	4.89	5.11	5.13	5.08	5.02	5.03	4.96
4.91	4.89	5.07	5.02	4.91	4.81	4.98	4.78
4.96	5.02	5.13	5.13	4.92	4.98	4.89	4.88

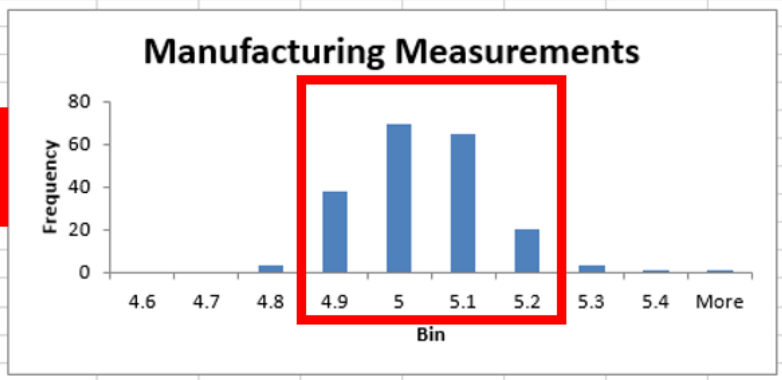
Mean	4.99
Standard deviation	0.117
Mean - 3*Stdev	4.640
Mean + 3*Stdev	5.340
Total variaton	0.700
Lower Specification	4.8
Upper Specification	5.2
Specification range	0.4
Cp	0.57

Step 3: Calculate Total Variation (which is 5.340 – 4.640)

Step 4: These Specifications are given by the Management

Step 5: Use the Total Variation and the Specification Range to calculate the Cp.

Bin	*Frequency*
4.6	0
4.7	0
4.8	3
4.9	38
5	69
5.1	65
5.2	20
5.3	3
5.4	1
More	1



Step 6: Draw histogram with different Bins.
Step 7: Anything outside the Specification Range is rejected (total 8/200 measurements are rejected = 4% defective and 96% were acceptable).

Conclusion:

- This shows that the 3rd Empirical Rule (μ ± 3σ) holds (≈ 99.7% coverage).
- Although this doesn't meet the empirical rule exactly (since its 96%), we are dealing with sample data.
- Other samples from the same process would have different characteristics.
- The Empirical Rule provides a good estimate of the total variation in the data that we can expect from any sample.

Figure 21: Steps to Implement C_p

.

PART IV

MEASURES OF SHAPE

A. SKEWNESS AND COEFFICIENT OF SKEWNESS (CS)

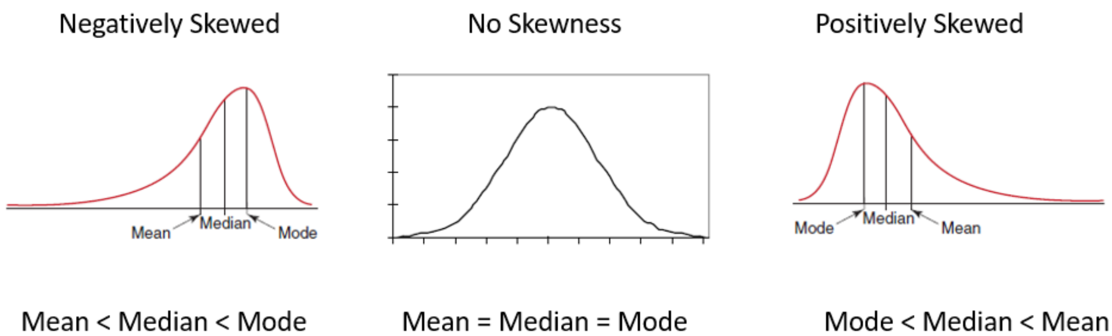


Figure 22: Skewness

$$CS = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^3}{\sigma^3}$$

Where:

- ✓ CS: Coefficient of Skewness
- ✓ Skewness = Lack of Symmetry of Data
- ✓ N: Population Size
- ✓ x_i : Individual Value of each of the Population
- ✓ μ : Population Mean
- ✓ σ : Population Std. Dev.
- ✓ If CS > 1 → Highly Positively Skewed
- ✓ If CS < -1 → Highly Negatively Skewed
- ✓ If CS = 0 → No Skewness

- ✓ If -0.5 < CS < 0.5 → Almost no Skewness
- ✓ If -0.5 < CS < -1 → Moderate Negative Skewness
- ✓ If 0.5 < CS < 1 → Moderate Positive Skewness
- ✓ If using Sample Data (rather than Population) → Replace the μ and σ (in the equation) with $\bar{x}$ (Sample Mean) and s (Sample Std. Dev.) respectively.
- ✓ To find Skewness using EXCEL Function → SKEW (data range)

B. KURTOSIS AND COEFFICIENT OF KURTOSIS (CK)

$$CK = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^4}{\sigma^4}$$

Where:

- ✓ CK: Coefficient of Kurtosis
- ✓ Kurtosis = Peakedness (High / Narrow) or Flatness (Short / Flat Topped) of the Histogram
- ✓ If CK > 1 → Highly Peaked
- ✓ If CK < -1 → Very Flat
- ✓ If -0.5 < CS < 0.5 → Relatively Normal Distribution
- ✓ If -0.5 < CS < -1 → Moderate Flat
- ✓ If 0.5 < CS < 1 → Moderate Peaked
- ✓ If using Sample Data (rather than Population) → Replace the μ and σ (in the equation) with $\bar{x}$ (Sample Mean) and s (Sample Std. Dev.) respectively.
- ✓ To find Kurtosis using EXCEL Function → KURT (data range)

EXCEL DESCRIPTIVE STATISTICS TOOL

Example: Given these 12 data:

	A
1	COST
2	$ 241.00
3	$ 262.00
4	$ 226.00
5	$ 179.00
6	$ 156.00
7	$ 142.00
8	$ 158.00
9	$ 158.00
10	$ 153.00
11	$ 151.00
12	$ 225.00
13	$ 244.00

Figure 23: Sample Data

Find: All Descriptive Statistics of this 12 data i.e.

- Mean
- Median
- Mode
- Variance
- Std. Dev.
- Range
- Skewness
- Kurtosis
- All Quartiles: Q_1, Q_2, Q_3

STEP 1

INSTALL EXCEL ANALYSIS TOOLPAK

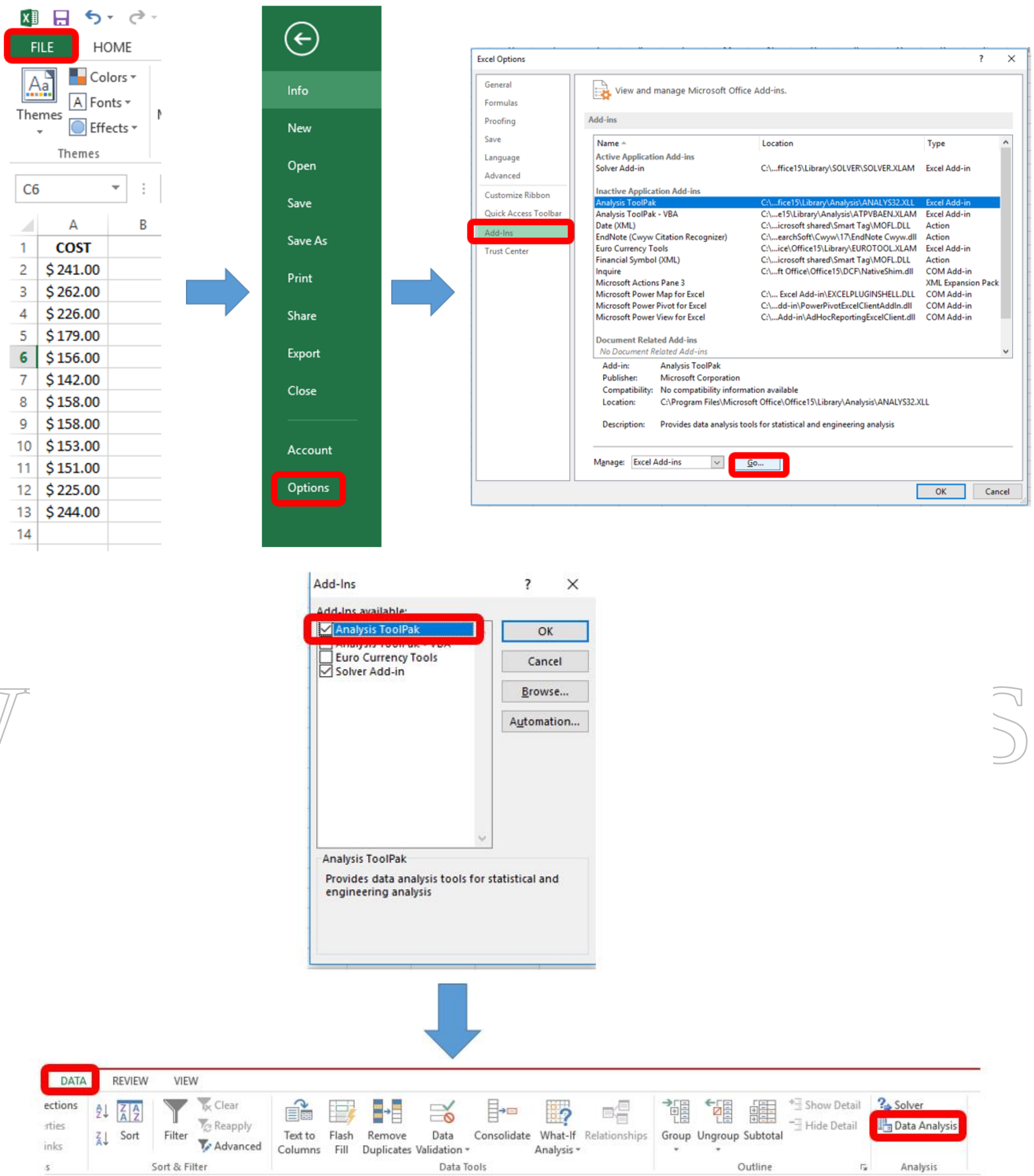


Figure 24: Installing the Excel Analysis Tool Pak

- ✓ Click File → Options → Add – Ins → Select "Excel Add-Ins" → Go…

- ✓ Select "Analysis Tool Pak" → OK → Data → Data Analysis should appear

STEP 2

RUNNING THE DESCRIPTIVE STATISTICS

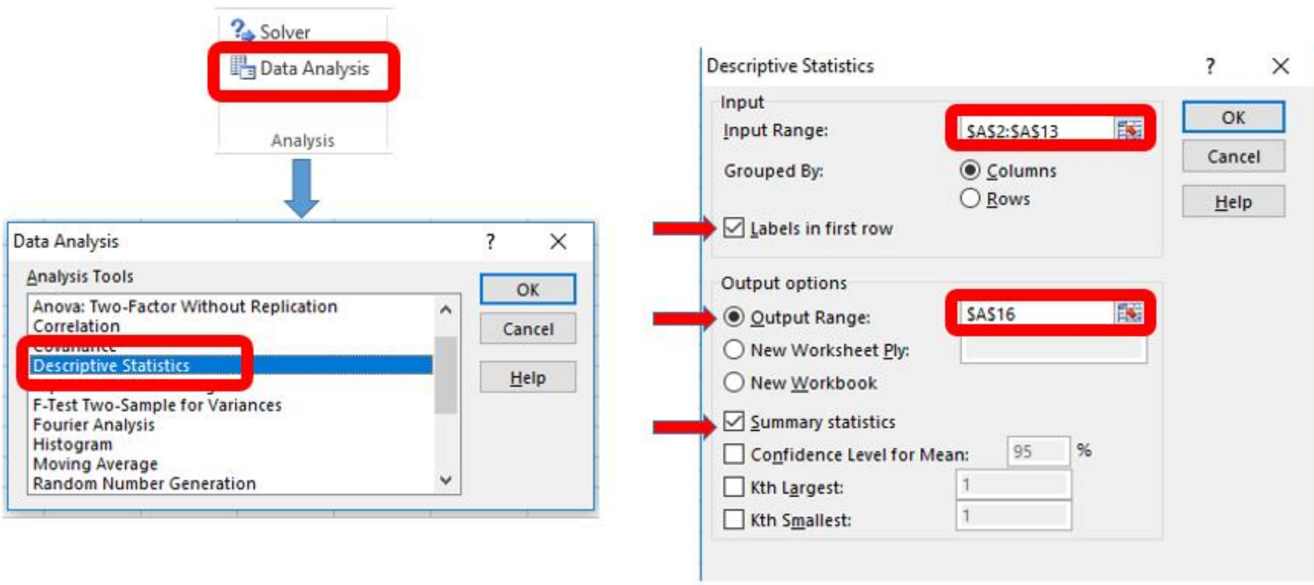


Figure 25: Running the Descriptive Statistics

- ✓ Click Data Analysis → Select Descriptive Statistics → OK
- ✓ Select "Input Range" of Data → Select "Labels in first row"
- ✓ Select "Output Range" for Display of Descriptive Statistics → Select "Summary Statistics → OK
- ✓ All Descriptive Statistics will appear.

COST	
Mean	191.25
Standard Error	12.82819
Median	168.5
Mode	158
Standard Deviation	44.43816
Sample Variance	1974.75
Kurtosis	-1.75678
Skewness	0.428482
Range	120
Minimum	142
Maximum	262
Sum	2295
Count	12

Figure 26: All Descriptive Statistics

STEP 3

FINDING THE QUARTILES

- ✓ Use the formula: =QUARTILE(A2:A13,1) to find Q1
- ✓ Use the formula: =MEDIAN(A2:A13) to find Q2
- ✓ Use the formula: =QUARTILE(A2:A13,3) to find Q3

=QUARTILE(A1:A13,3)

C	D	E
Q1	155.25	
Q2	$ 168.50	
Q3	229.75	

Figure 27: All Quartiles

PART VI

MEASURES OF ASSOCIATION

A. COVARIANCE

$$\mathrm{cov}_P(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Where:

- X: 1st Variable
- Y: 2nd Variable
- cov_P (X,Y): Population Covariance
- N: Population Size
- x_i : Random Variable x, where i = 1, 2, 3, …, N
- y_i : Random Variable y, where i = 1, 2, 3, …, N
- μ_x : Population Mean of X
- μ_y : Population Mean of Y

1. cov_P (X,Y) is a measure of the linear association between two variables, X and Y.
2. The larger the cov_P (X,Y) → the higher the degree of ***linear*** association between X and Y.
3. Positive cov_P (X,Y) → direct relationship (i.e., one variable increases as the other increases)
4. Negative cov_P (X,Y) → inverse relationship (i.e., one variable increases while the other decreases, or vice versa).
5. Scatter Diagram shows the strength of linear association between two variables and the sign of the covariance. (Figure 28)
6. Population Covariance Excel function = COVARIANCE.P (array1, array2).

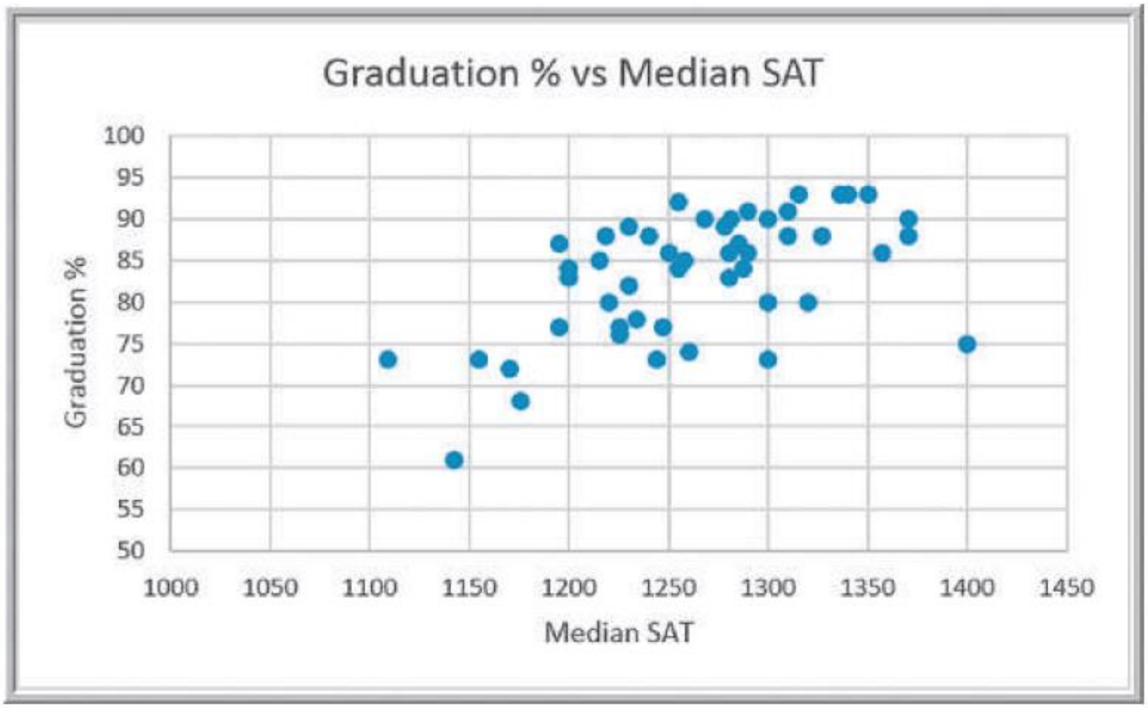


Figure 28: Scatter Diagram showing Positive Covariance (Evans, 2014)

$$\text{cov}_s(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Where:

- cov_s (X,Y): Sample Covariance
- n: Sample Size
- $\bar{x}$: Sample Mean of X
- $\bar{y}$: Sample Mean of Y

HOW TO OBTAIN SAMPLE COVARIANCE USING EXCEL

1. Sample Covariance Excel function = COVARIANCE.S(array1, array2)
2. Figure 29 shows how to obtain the Sample Covariance; which is reflected in Figure 28.

B52 | =COVARIANCE.S(A2:A50,B2:B50)

	A	B	C	D	E	F
1	Graduation % (X)	Median SAT (Y)				
2	93	1315				
3	80	1220				
4	88	1240				
5	68	1176				
6	90	1300				
7	90	1281				
8	84	1255				
9	75	1400				
10	80	1300				
46	78	1234				
47	86	1250				
48	91	1290				
49	93	1336				
50	93	1350				
51						
52	COVARIANCE.S	263.3703231				

Figure 29: Obtaining the Sample Covariance using Excel

B. CORRELATION

1. Correlation is a "better" version of Covariance.
2. Correlation is more widely used over Covariance.
3. Why? Because
    - Reason 1: Covariance has no definite measure of "**strength**" of relationship. Correlation has a definite measure.
        - The scale of Correlation is between -1 and +1
            - +1 = very strong positive ***linear*** correlation
            - -1 = very strong negative ***linear*** correlation
            - 0 = no ***linear*** relationship
        - The scale of Covariance is between -∞ and +∞
            - +∞ = very strong positive ***linear*** correlation
            - -∞ = very strong negative ***linear*** correlation
            - 0 = no ***linear*** relationship
        - Since Covariance is between -∞ and +∞, there is no actual/relative way to represent "strong" and "weak". (i.e. how strong is strong? How weak is weak? We can't tell by the numbers!)
        - Since Correlation is between -1 and +1, there is a definite way to represent "strong" and "weak".
    - Reason 2: Covariance has "units" but Correlation has no "units".
        - If you look at the Covariance equation i.e. cov (X,Y), you will realize that there are units tied to it. E.g. if X is in "cm" and Y is in "cm", then cov (X,Y) will end up in "cm^2" (which does not make any sense).
        - But Correlation has no units. Its value [-1, +1] solely represents "strength" of relationship.
    - Reason 3: Covariance is affected by scale. Correlation is not.
        - For example cov (X,Y) → initially we let X: cm and Y: cm, for standardization.
        - But later, we change X: m and Y: cm.

- This means that for all X, we have to divide by 100 to change it to m.
- cov (X: cm, Y: cm) will then change and be different from cov (X: m, Y: cm).
- This makes it difficult to assess the strength.
- But since Correlation is unit-less, its strength is still measured between [-1, +1]. There is no change.

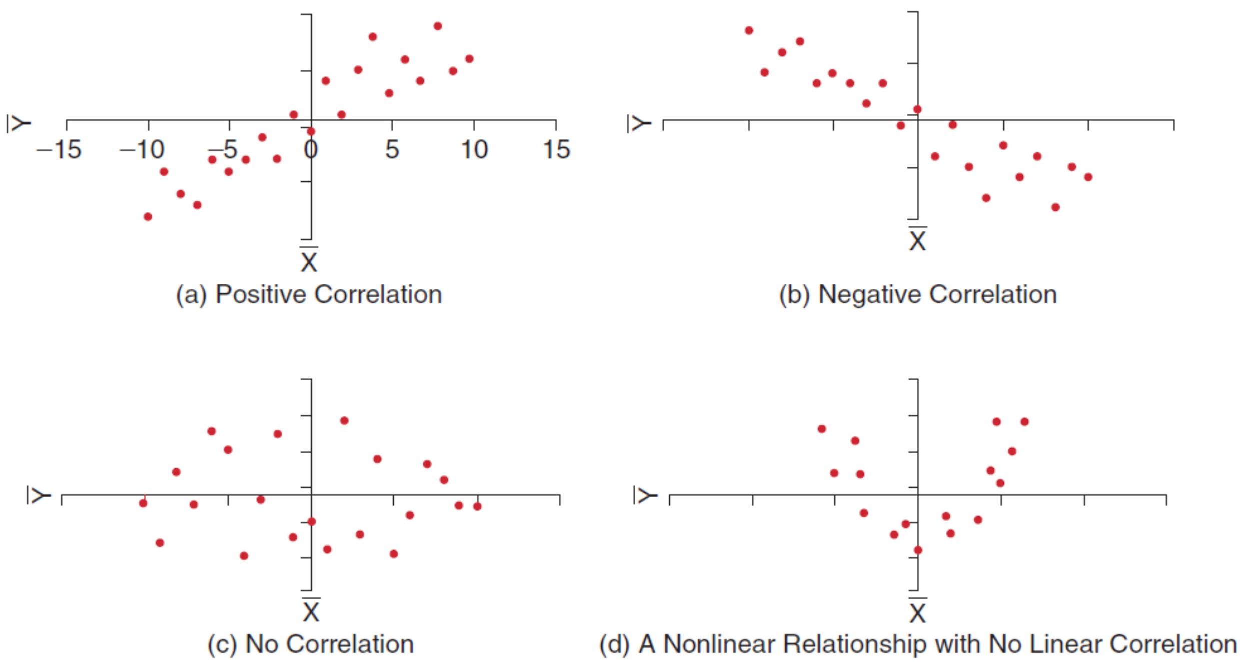


Figure 30: Scatter Diagrams showing Correlation (Evans, 2014)

4. For Figure 30(d), the relationship is not linear and the correlation is zero.
5. In real life, do a scatter plot to observe the relationship between two variables first. Do this before obtaining the Correlation Coefficient.

$$\rho_{xy} = \frac{\text{cov}_{\text{p}}(X,Y)}{\sigma_x \sigma_y}$$

Where:

- ρ_{xy} : Population Correlation
  - aka Pearson Product Moment Correlation Coefficient
  - aka Correlation Coefficient
- cov_{p} (X,Y): Population Covariance
- σ_x : Population Std. Dev. Of X
- σ_y : Population Std. Dev. Of Y
- By dividing the covariance by the product of the standard deviations, we are essentially scaling the numerical value of the covariance to a number between -1 and 1.

$$r_{xy} = \frac{\text{cov}_{\text{s}}(X,Y)}{s_x s_y}$$

Where:

- r_{xy} : Sample Correlation
- cov_{s} (X,Y): Sample Covariance
- s_x : Sample Std. Dev. Of X
- s_y : Sample Std. Dev. Of Y

HOW TO OBTAIN SAMPLE CORRELATION USING EXCEL

1. Sample Correlation Excel function = CORREL(array1, array2).
2. Figure 29 shows how to obtain the Correlation.
3. Note: in Excel, the CORREL function outputs only one Correlation (Population Correlation = Sample Correlation)

B52 | =CORREL(A2:A50,B2:B50)

	A	B	C	D	E
1	**Graduation % (X)**	**Median SAT (Y)**			
2	93	1315			
3	80	1220			
4	88	1240			
5	68	1176			
40	86	1280			
41	88	1218			
42	61	1142			
43	73	1109			
44	84	1287			
45	76	1225			
46	78	1234			
47	86	1250			
48	91	1290			
49	93	1336			
50	93	1350			
51					
52	**CORREL Function**	0.564146827			

Figure 31: Obtaining the Sample Correlation using Excel

C. EXCEL CORRELATION TOOL

HOW TO OBTAIN CORRELATIONS BETWEEN MULTIPLE VARIABLES USING EXCEL

1. Follow Step 1: Install Excel Analysis Toolpak (Page 29) to install the Analysis Toolpak
2. Click on Data Analysis → Correlation → OK

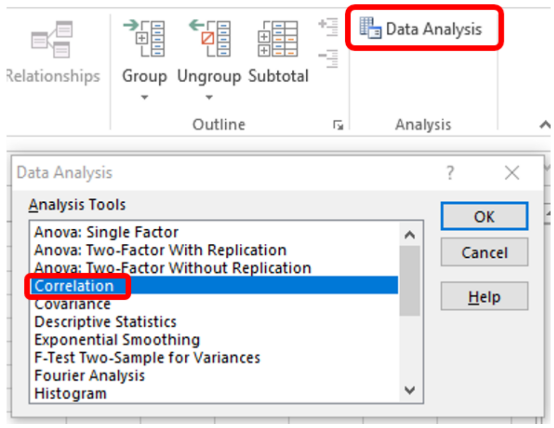


Figure 32: Go to "Data Analysis" to find "Correlation"

3. For the Input Range, select all the relevant data.
4. Tick the box "Labels in First Row".
5. Select the Output Range anywhere on the **SAME** sheet
   - *Note 1: All the Data Columns must be *contiguous* to each other (next to each other).
   - *Note 2: The output range must be on the **SAME** sheet, or else an error will pop up.

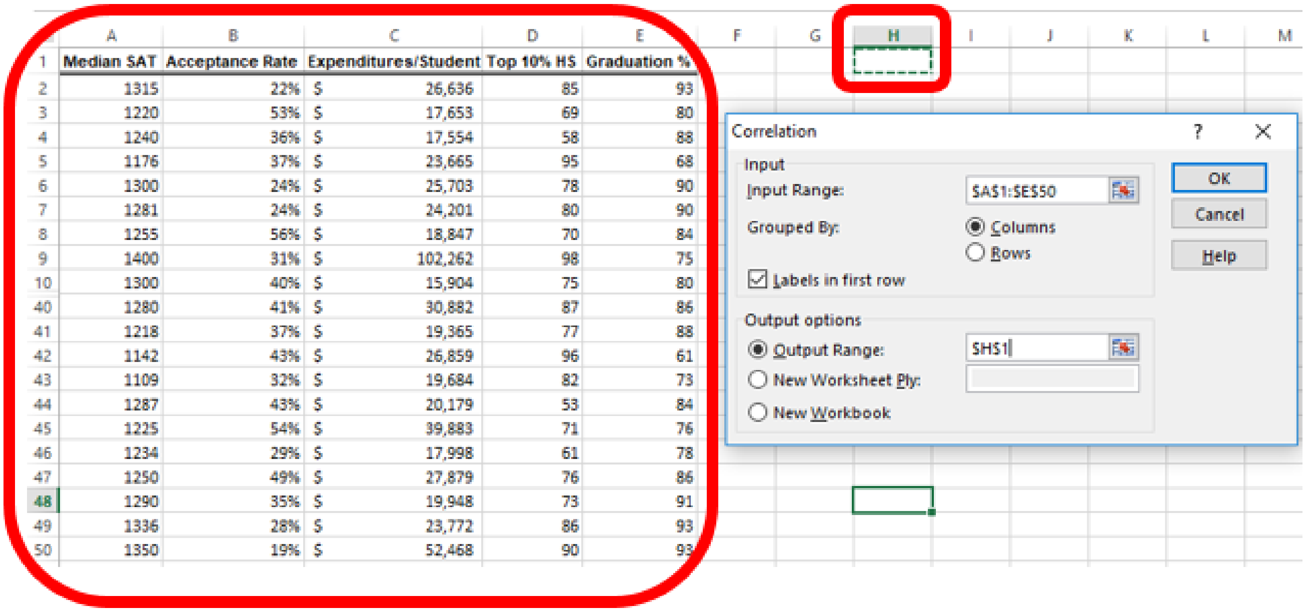


Figure 33: Select the relevant cells for Input and Output

6. Figure 34 shows the final output of the Excel Correlation Tool
   - The diagonal "1"s represent that variables are perfectly correlated with themselves.

	Median SAT	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
Median SAT	1				
Acceptance Rate	-0.601901959	1			
Expenditures/Student	0.572741729	-0.284254415	1		
Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

Figure 34: Output of Excel Correlation Tool

PART VII

OTHERS

OUTLIERS

- ✓ There is no standard definition of what constitutes an outlier.
- ✓ It is just an unusual observation as compared with the rest.
- ✓ Sometimes, individual variables might not exhibit outliers, but combinations of them might.

HOW TO DETERMINE OUTLIERS?

1. Method 1: Visual Inspection
    - Check the data for errors:
        - Misplaced decimal point?
        - Typo error?
        - Use histograms to identify outliers visually.

2. Method 2: Empirical Rule
    - Anything > 3σ or <3σ = Outliers

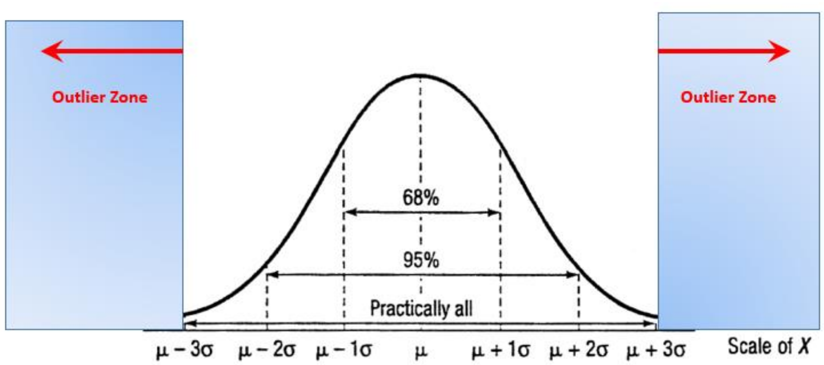


Figure 35: Using Empirical Rule to Determine Outliers

3. Method 3: Inter Quartile Range (IQR)

   - On Page 21 (Box Plot, Interquartile Range (IQR), Percentile), we determined one way of finding outlier – by zones.

   - Here, we refine the zones (Figure 36: IQR outliers)

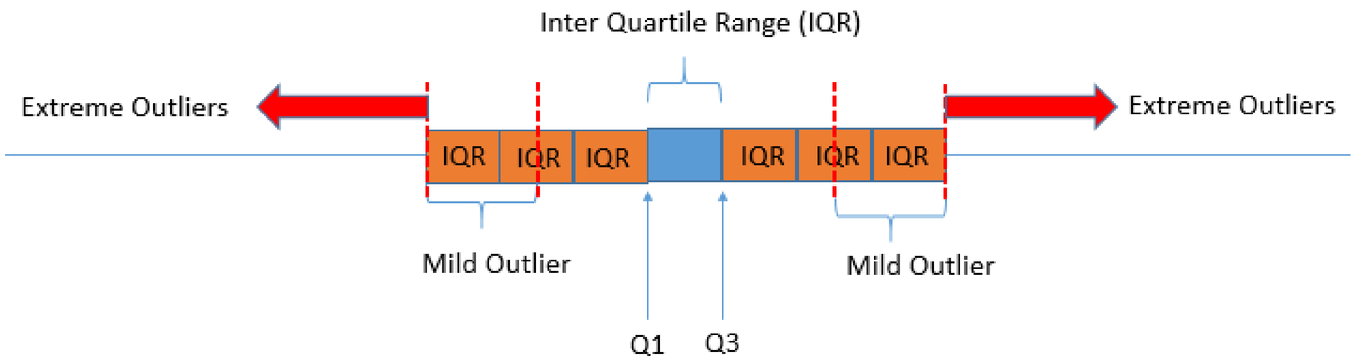


Figure 36: IQR outliers

WHAT SHOULD WE DO WITH THE OUTLIERS?

- ✓ Do NOT blindly eliminate outliers.
- ✓ Eliminate only if it does not make common sense.
- ✓ First, run the experiment with Outliers.
- ✓ Then, run the experiment without Outliers.
- ✓ Lastly, compared both results critically.

REFERENCES

EVANS, J. R. 2014. *Business analytics*, Harlow Pearson, [2014] Pearson new international edition.

ABOUT PROFESSOR JAMES EVANS

James R. Evans is a professor in the Department of Operations, Business Analytics, and Information Systems in the College of Business at the University of Cincinnati. He holds BSIE and MSIE degrees from Purdue and a PhD in Industrial and Systems Engineering from Georgia Tech. He has also served on numerous journal editorial boards and is a past-president and Fellow of the Decision Sciences Institute. A recognized international expert on quality management, he served on the Board of Examiners and the Panel of Judges for the Malcolm Baldrige National Quality Award. Much of his current research focuses on organizational performance excellence and measurement practices.

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.