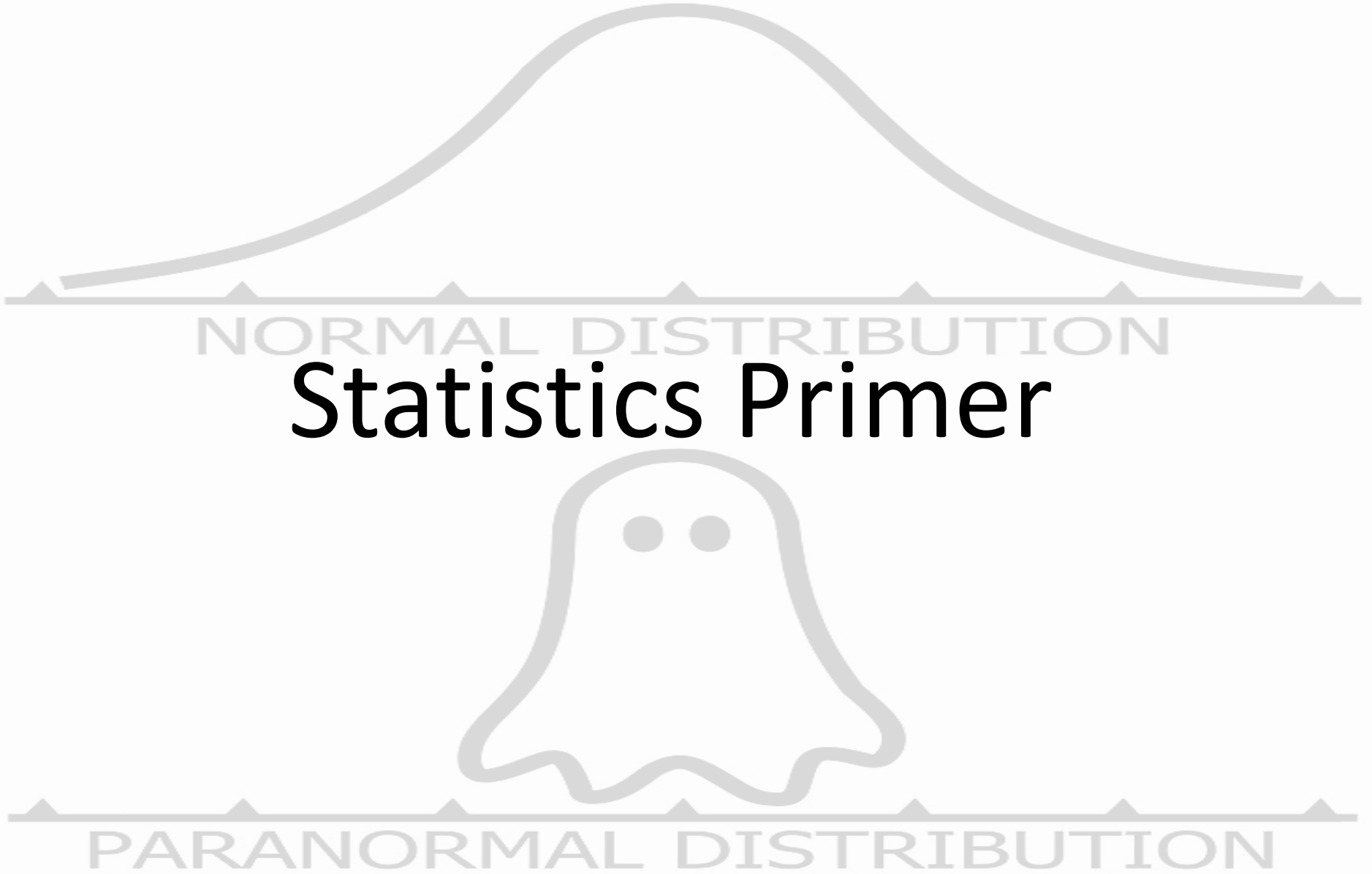


Statistics Primer

What is the objective of this course?
What is the assignment attached to this course ?

- The objective of this course is to learn fundamental statistical concepts which can be applied in day to day analytics

- We have a MCQ assignment attached to the course

Course Content

- Types of Variables
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Association
- Types of error metrics
- Standardization
- Outlier detection

Scales of Measurement

Nominal/Categorical

Values can be put into Categories (2 or more categories)

Values have no intrinsic order hence cannot be compared

E.g. City, Zip code

Ordinal

Values are categories with pre-defined order

The gap between categories may vary

E.g. Customer Tier (Platinum, Gold, Silver)

Here Platinum is better than Gold but the gap is not quantifiable

Continuous

Values are measurable quantities with equal gaps between values

They can be further divided into Interval and Ratio

Interval scale has 0 as one of the values in the data and values can exist on either sides (E.g. Temperature, Sea Level)

Ratio scale has an absolute 0 value and values below 0 are not possible (E.g. # Visits, Sales)

Scales of Measurement

	Indications Difference	Indications Direction of Differences	Indicates Amount of Differences	Absolute Zero
Nominal	X			
Ordinal	X	X		
Interval	X	X		
Ratio	X	X	X	X

Measures of Central Tendency

Mean

Quantiles

Mode

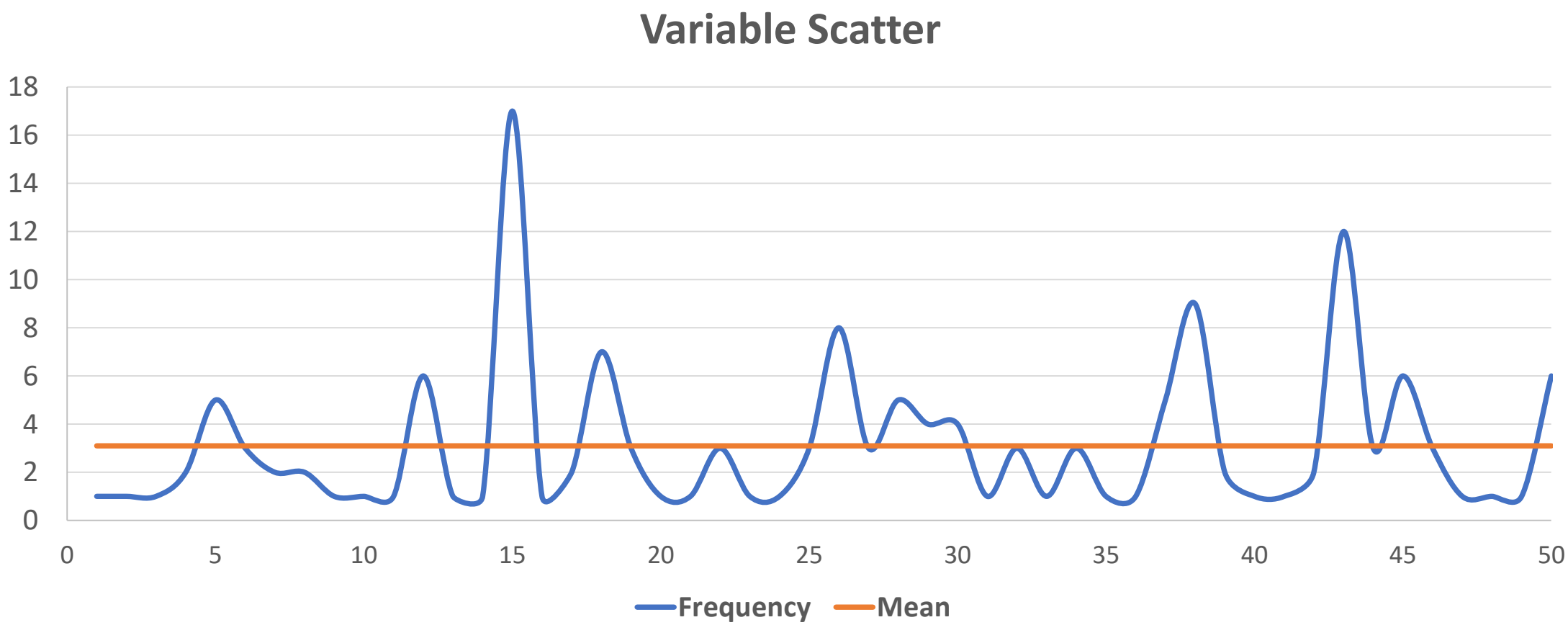
Mean

Mean is the central value of the data

Calculation : **sum of all values/total number of observations**

Pros	Cons
• Ease of calculation • Least affected by sample fluctuations • All values are accounted for	• Highly affected by presence of outliers • In absence of single term, value is inaccurate • Cannot be determined by inspection

Mean – Visual representation



Median

Median is the value which divides the data into 2 equal parts

Calculation :

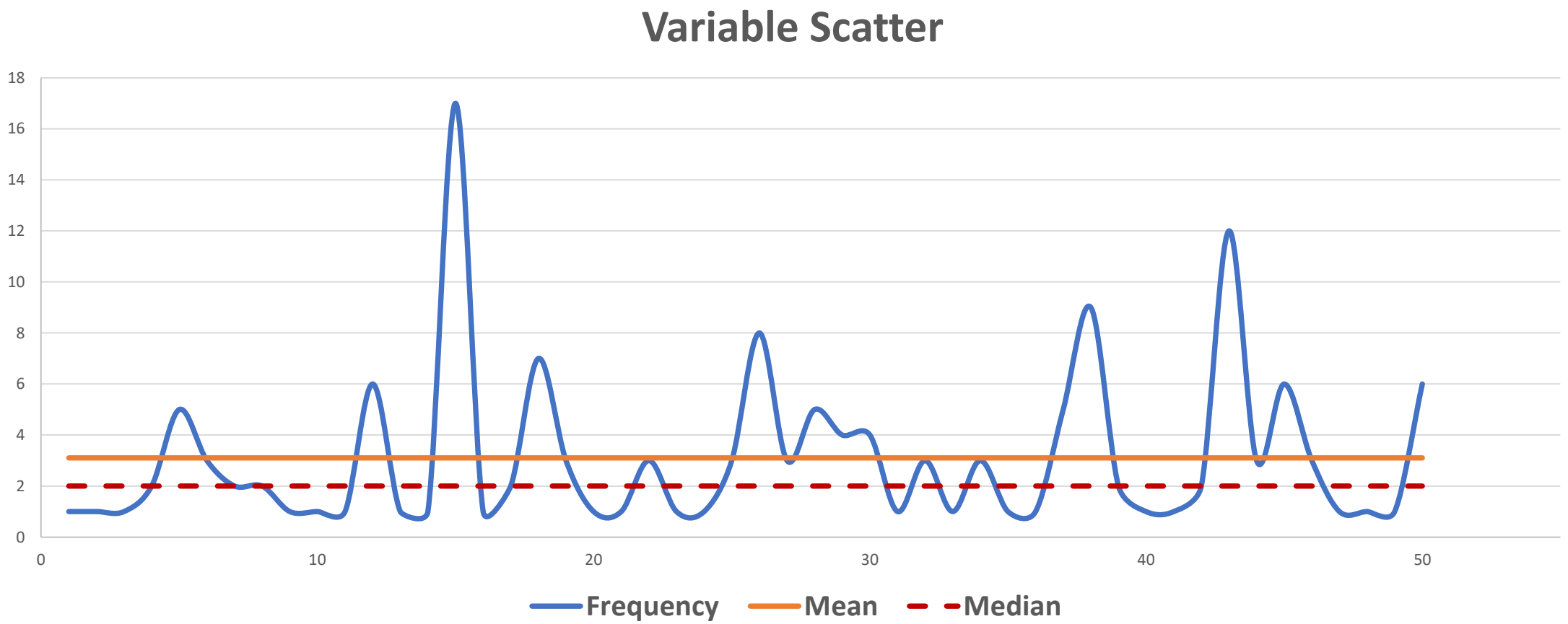
1. Arrange all the values in ascending order
2. Depending on whether N (# of observations) is odd or even it's calculated

For **odd N** - ((N+1)/2)th observation

For **even N** - mean of (N/2)th and ((N/2)+1)th observation

Pros	Cons
• Less affected by outliers • Less affected by skewed data	• Cannot be calculated from nominal data

Median – Visual representation



Quantiles

- Median is a part of larger set of metrics called quantiles
- These are used to divide the data into **n** parts

n	quantile name
4	quartile
10	decile
20	demi-deciles
100	percentiles

Mode

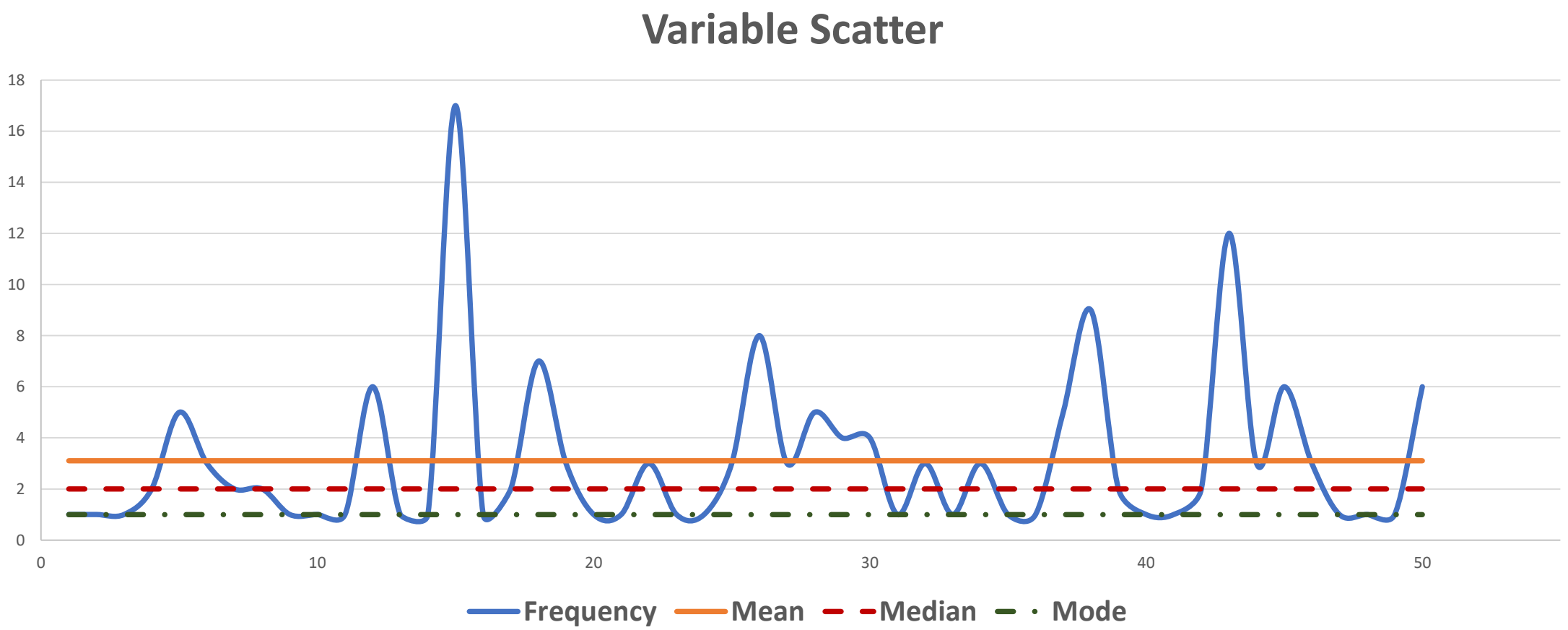
Mode is the most frequently occurring value

Calculation : **Compute frequency distribution of the data**

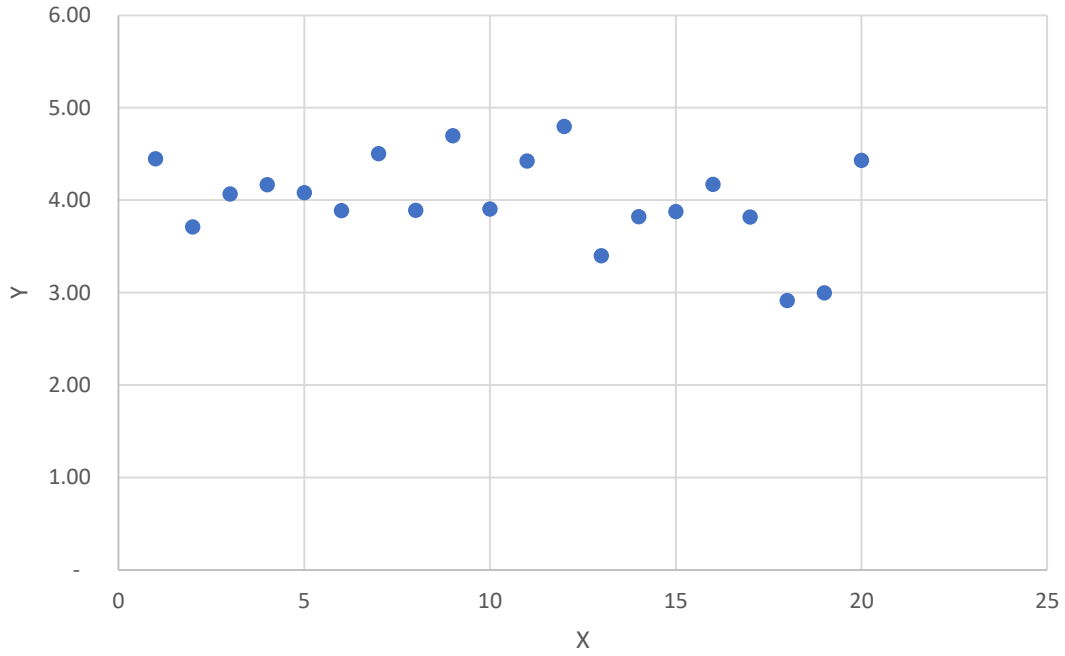
Value with highest frequency is the mode

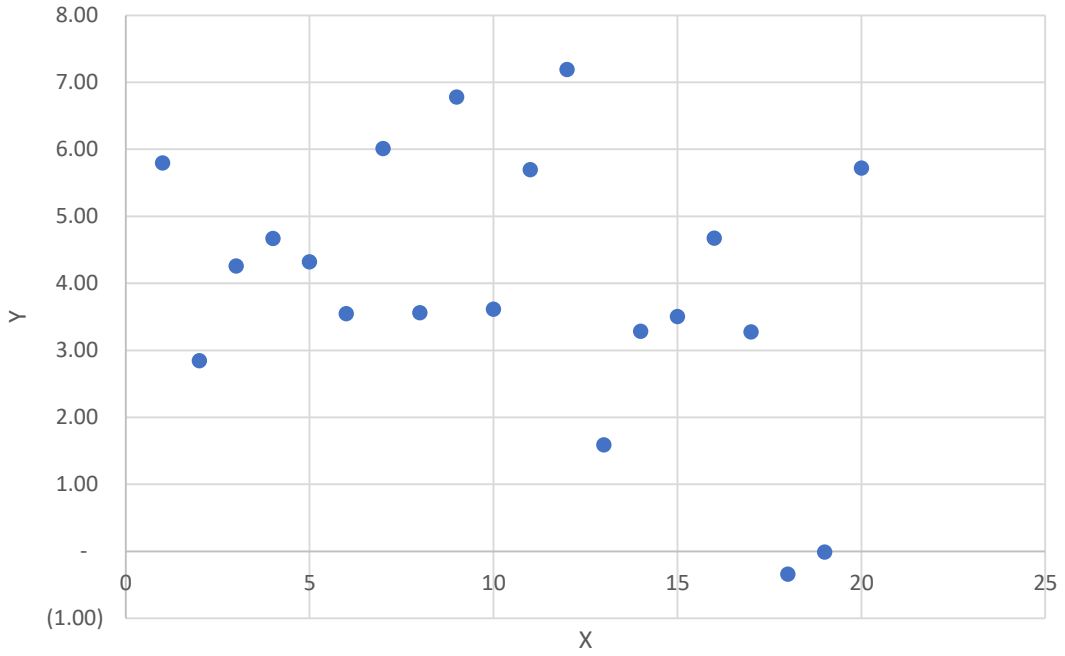
Pros	Cons
• Can be calculated for categorical as well as continuous data	• May not refer to the central value always • Data can be bi-modal (2 modal values) • No modal value is possible for continuous data

Mode – Visual representation



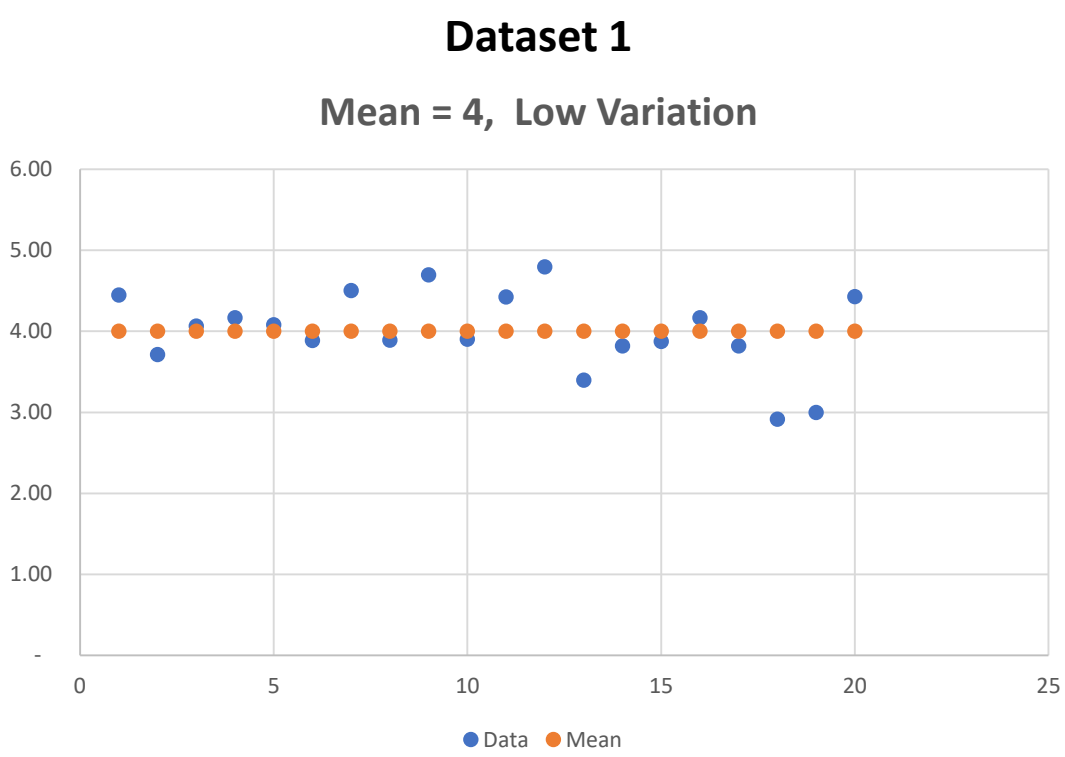
Understanding Dispersion

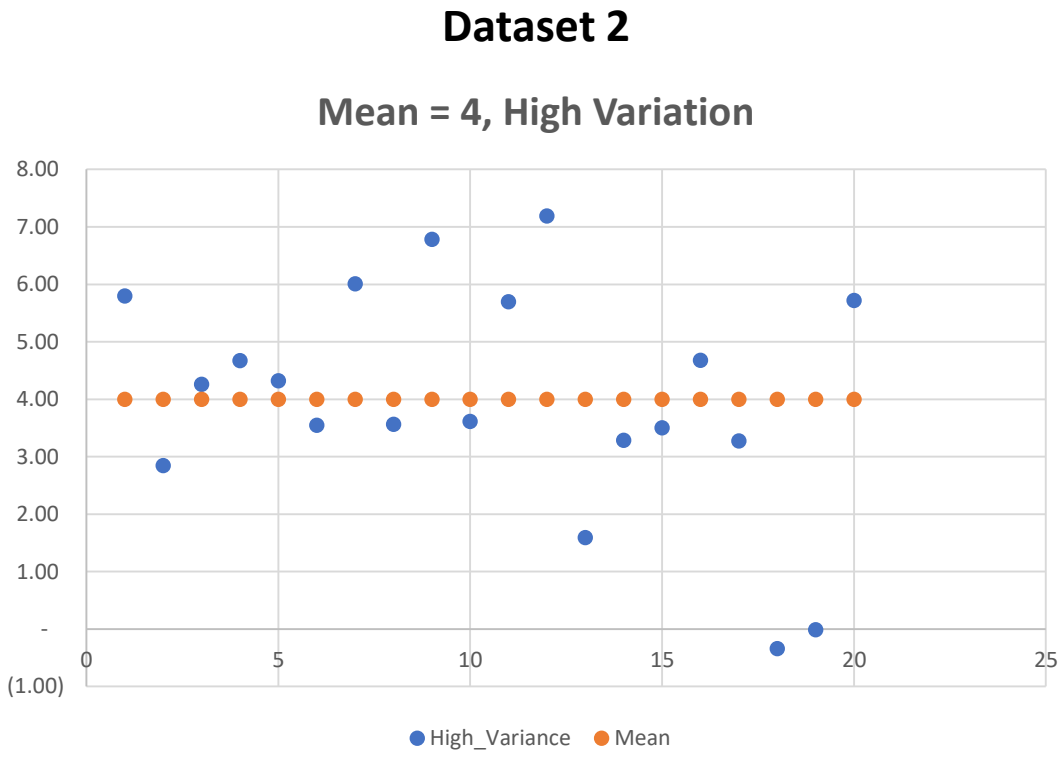




The data shown above looks very different at the first glance

Understanding Dispersion





Same mean, **Different** variance

Measures of Dispersion

Absolute Measures

Range

Inter-Quartile Range

Standard Deviation

Mean Absolute Deviation

Relative Measures

Coefficient of Variation

Range based statistics

The simplest measure for variation is **Range**

Calculation – **Max Value** – **Min Value**

Dataset 1 : Range **=** (4.7 – 2.9) = **1.8**

Dataset 2 : Range **=** (7.19 – (-0.33)) = **7.5**

Range based statistics

We could also use **Inter-Quartile Range**

Calculation – **3rd Quantile – 1st Quantile**

Dataset 1 : IQR(abbrv.) = (4.42 – 3.82) = **0.6**

Dataset 2 : IQR **=** (5.70 – 3.28) = **2.42**

Inter-Quartile Range will also be used for **Outlier detection and treatment**

Deviation based statistics

- A very well known measure is **Standard deviation**
- Calculation **–**
  1. Compute deviations (differences) of each observation from the mean
  2. Square the deviations and take average across all which is known as variance
  3. Take squared root of variance

Dataset 1 : SD**(abbrv.)** = 0.49

Dataset 2 : SD = 1.95

Deviation based statistics

- Another measure is **Mean Absolute Deviation**
- Calculation **–**
    1. Compute deviations (differences) of each observation from the **mean**
    2. Take absolutes of deviations and take average across all

Dataset 1 : MAD**(abbrv.)** = 0.31

Dataset 2 : MAD = 1.23

This statistic can be calculated with **Median** as well

Deviation based statistics - Relative

Many a times we need to compare variation across different variables.

But 2 variables can be of different scales e.g. Sales and Frequency

Since the absolute measures are scale dependent we need to use relative measures to compare across variables

Coefficient of Variation = **SD/Mean**

Variable	Mean	Standard Deviation	Coefficient of Variation
ASP	809.14	484.29	0.599
Frequency	3.1	3.13	1.00

ASP has higher absolute SD but **lower COV**

Measures of Association

So far we have been looking at variables in isolation

But in order to analyze relationships across variables we need measures of association.

Variable 1	Variable 2	Measures of association
Continuous	Continuous	Pearson's correlation coefficient
Ordinal	Ordinal	Spearman's rank correlation coefficient
Ordinal	Categorical	Spearman's rank correlation coefficient
Categorical	Categorical	Chi-Squared test of independence

Continuous vs Continuous

Pearson's Correlation Coefficient

The extent of **linear relationship** between 2 variables

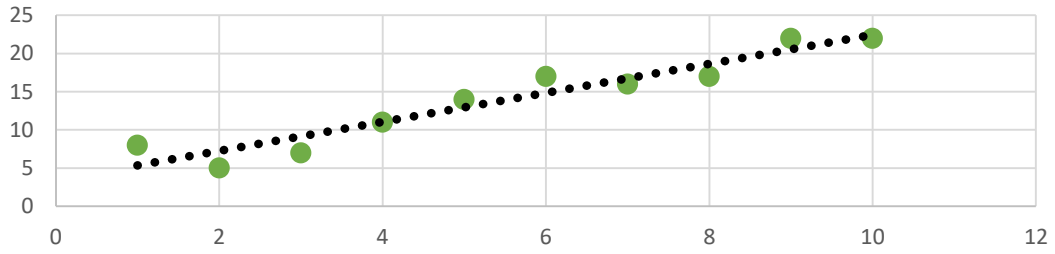
Value for Pearson's Correlation Coefficient (**r**) **lies between -1 and 1**

- **$0 < r \leq 1$:** Positive Correlation
- **$-1 < r < 0$:** Negative Correlation
- **$r \approx 0$:** No Correlation

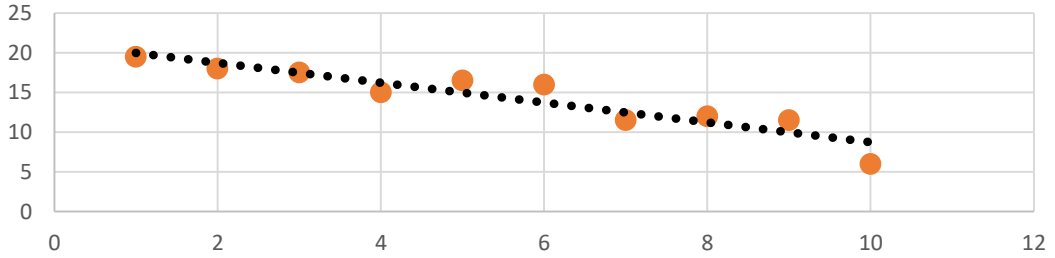
Continuous vs Continuous

Pearson's Correlation Coefficient

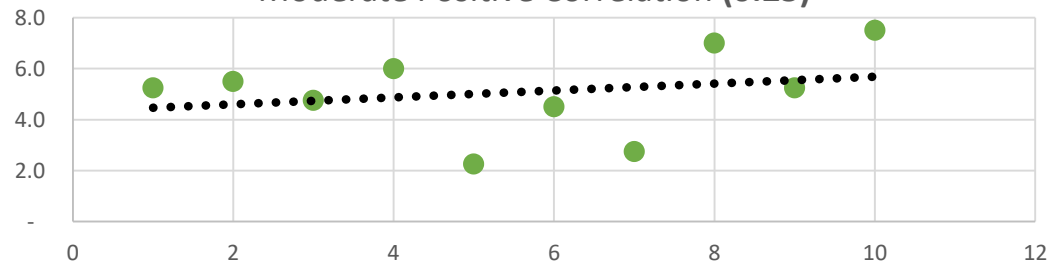
High Positive Correlation **(0.95)**



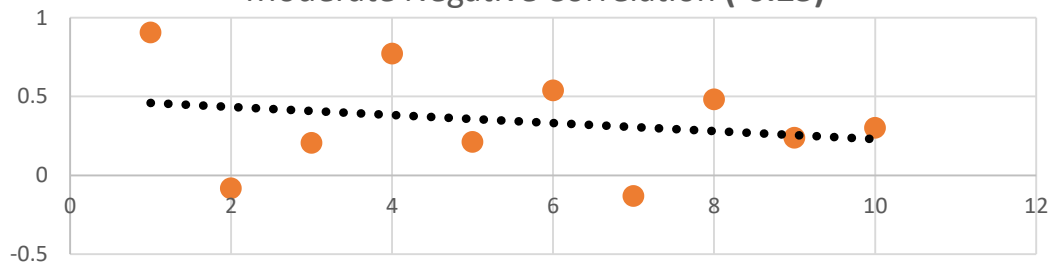
High Negative Correlation **(-0.92)**



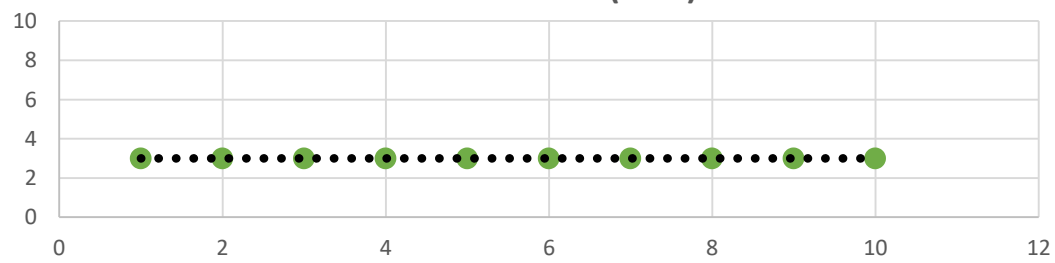
Moderate Positive Correlation **(0.25)**



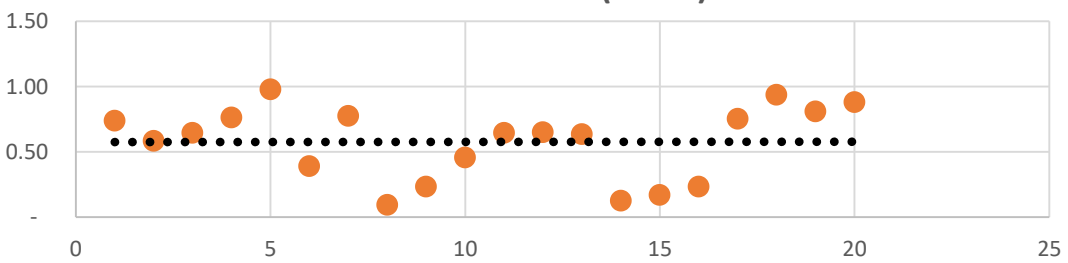
Moderate Negative Correlation **(-0.23)**



No Correlation **(0.00)**



No Correlation **(0.003)**



Ordinal vs Ordinal/Continuous

Spearman’s Rank Correlation Coefficient

Spearman's correlation assesses monotonic relationships (whether linear or not)

Intuitively

Spearman correlation is **high** when **observations have a similar rank**

Spearman correlation is **low** when **observations have a dissimilar rank**

Value for Spearman’s Correlation Coefficient (**ρ**) **lies between -1 and 1**

Appropriate for both continuous and discrete ordinal variables

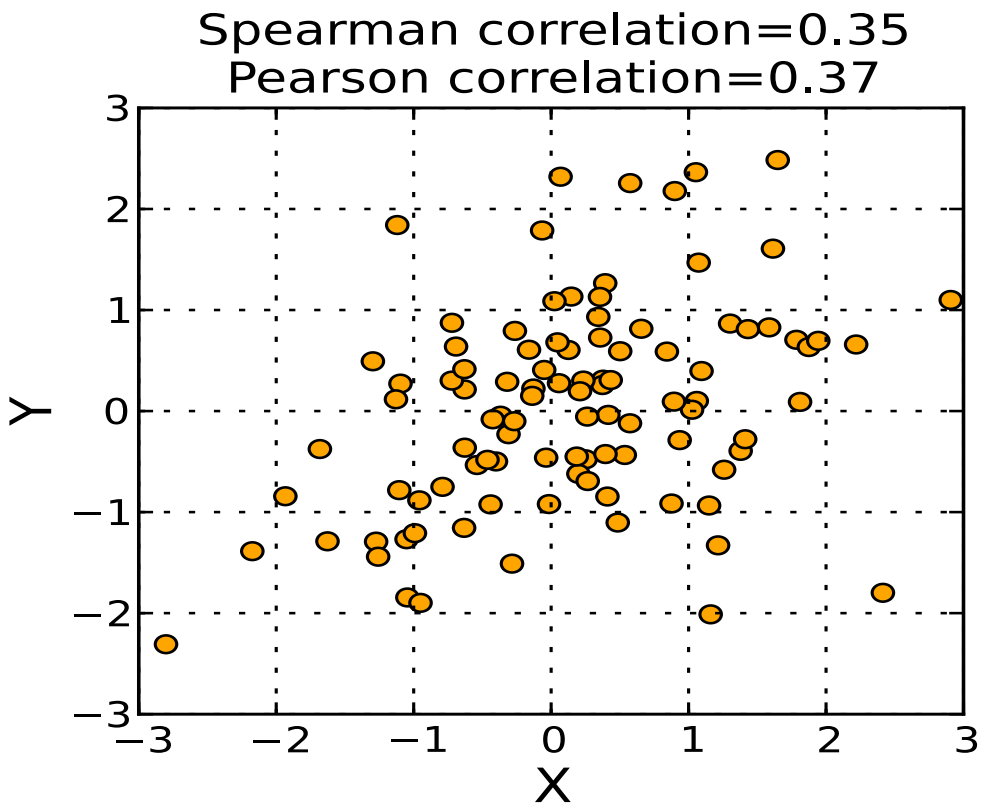
$0 < \rho \leq 1$: Positive Correlation

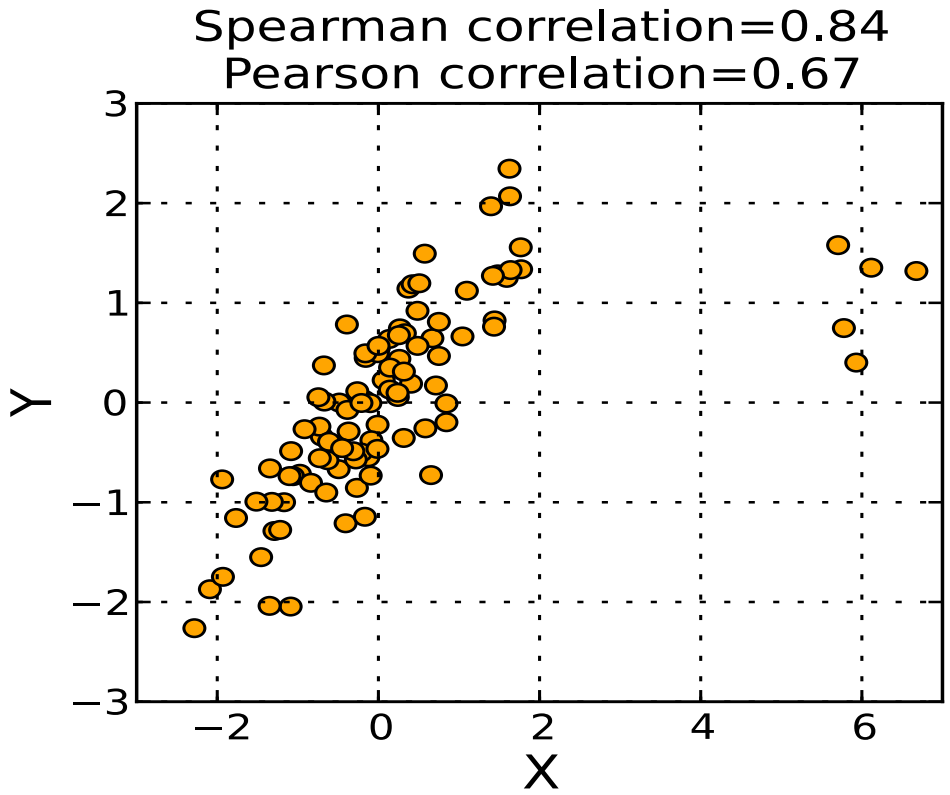
$-1 < \rho < 0$: Negative Correlation

$\rho \approx 0$: No Correlation

Ordinal vs Ordinal/Continuous

Spearman's Rank Correlation Coefficient





Spearman's correlation is robust to outliers

Source : **Wikipedia**

Categorical vs Categorical

Chi-Square test of Independence

The chi-squared test is used to determine whether there is a **significant difference** between the **expected frequencies** and the **observed frequencies** in one or more categories

Null hypothesis : Variable X and Y are **independent**

Alternative hypothesis : Variables X and Y are **dependent**

$$\mathbf{X^2} = \sum_{i=1}^{mn} \frac{(\boldsymbol{O_i} - \boldsymbol{Ei})^2}{\boldsymbol{E_i}} \sim \mathbf{X^2}_{(m\_1)(n\_1)}\ degrees\ of\ freedom$$

m – number of rows

n – number of columns

O_i : Observed Frequency

E_i : Expected Frequency

Categorical vs Categorical

Chi-Square test of Independence

	Avg Spend Bands				
Gender	**< 1K**	**1K - 5K**	**5K - 10K**	**10K+**	**Total**
Male	100	200	500	300	1,100
Female	20	50	70	80	220
Total	120	250	570	380	1,320

$\mathbf{X^2}_{cal} = 15$
p-value = 0.001687

Reject Null hypothesis
Variables are **dependent** on each other

	Avg Spend Bands				
Gender	**< 1K**	**1K - 5K**	**5K - 10K**	**10K+**	**Total**
Male	100	200	500	300	1,100
Female	110	190	520	270	1,090
Total	210	390	1,020	570	2,190

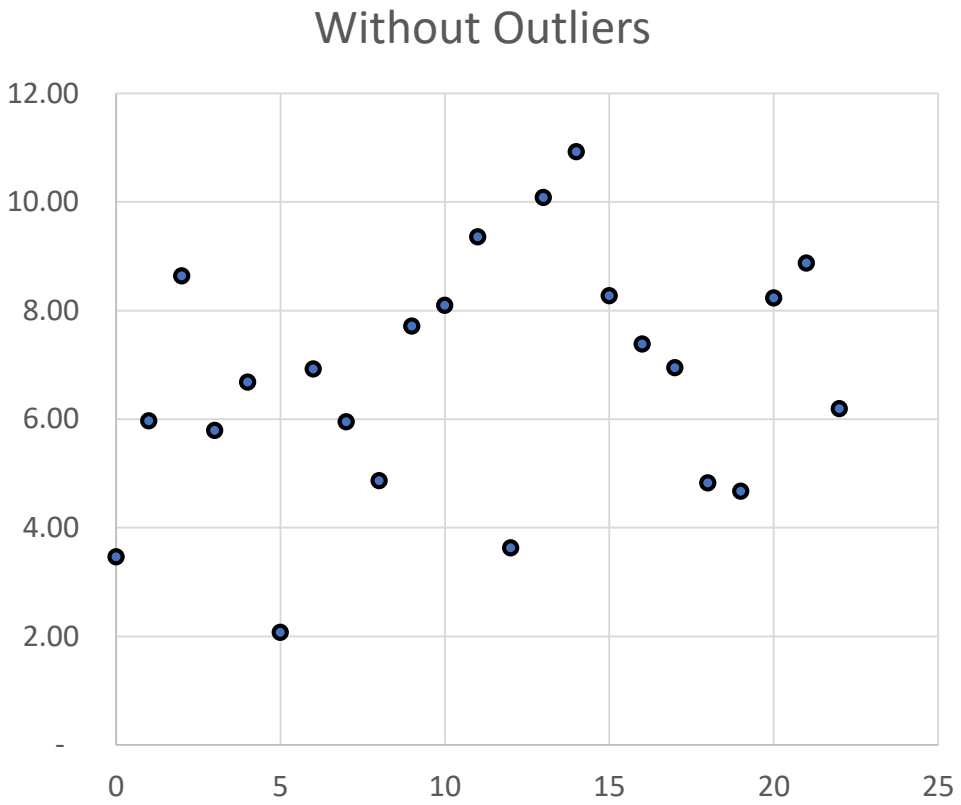
$\mathbf{X^2}_{cal} = 3$
p-value = 0.447395

Accept Null hypothesis
Variables are **independent** off each other

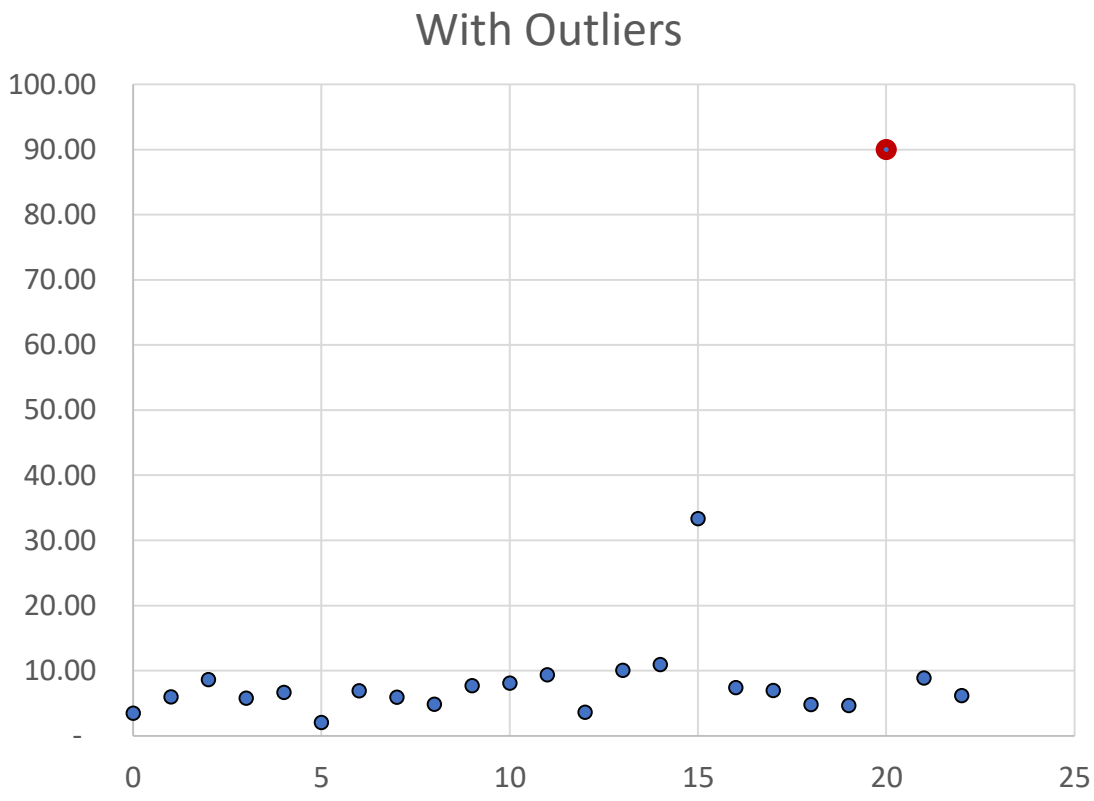
Outlier Detection & Treatment

What is a Outlier?

Scatter Plot



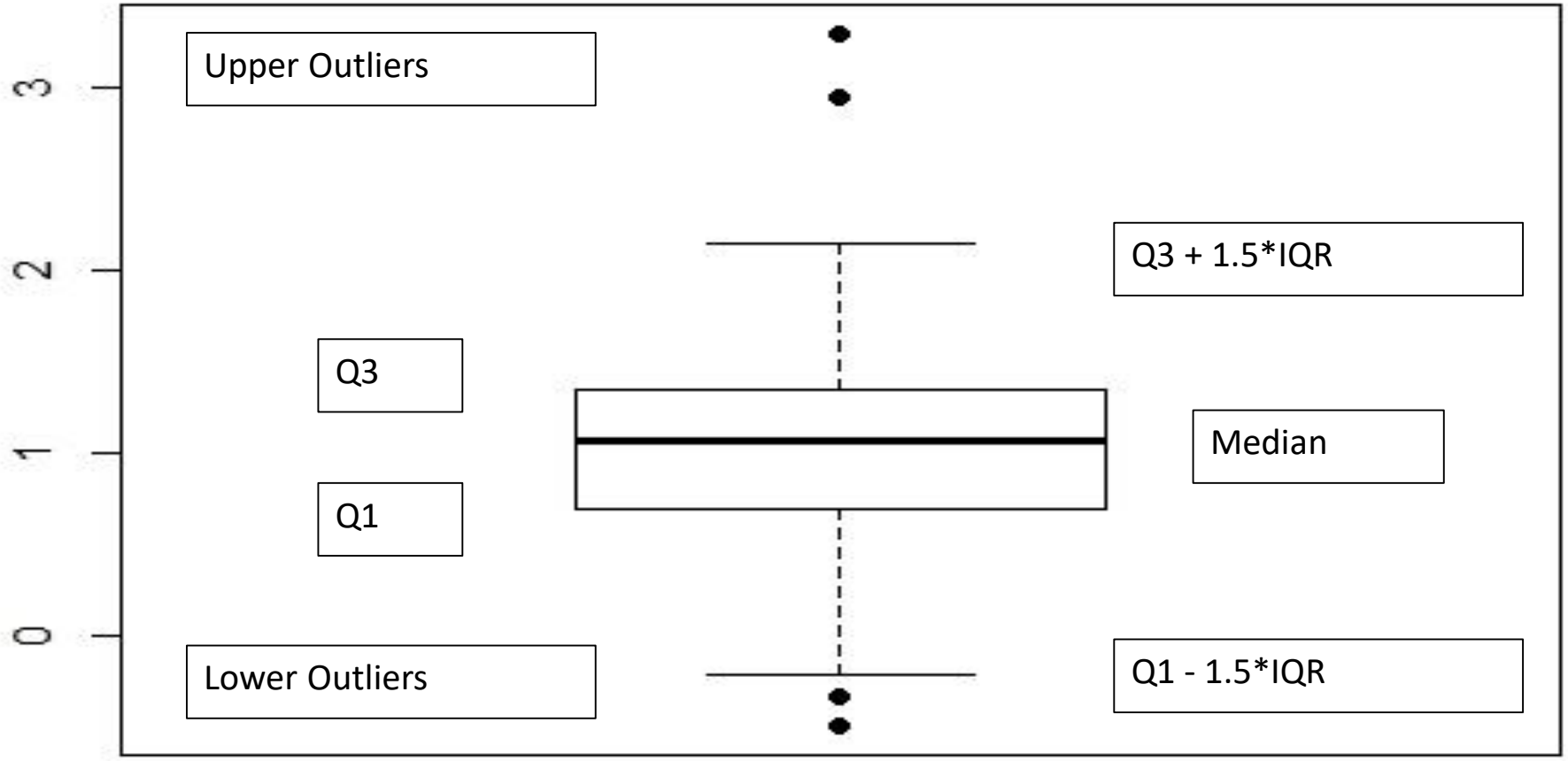
Mean : **6.76**



Mean : **20.54**

What is an Outlier

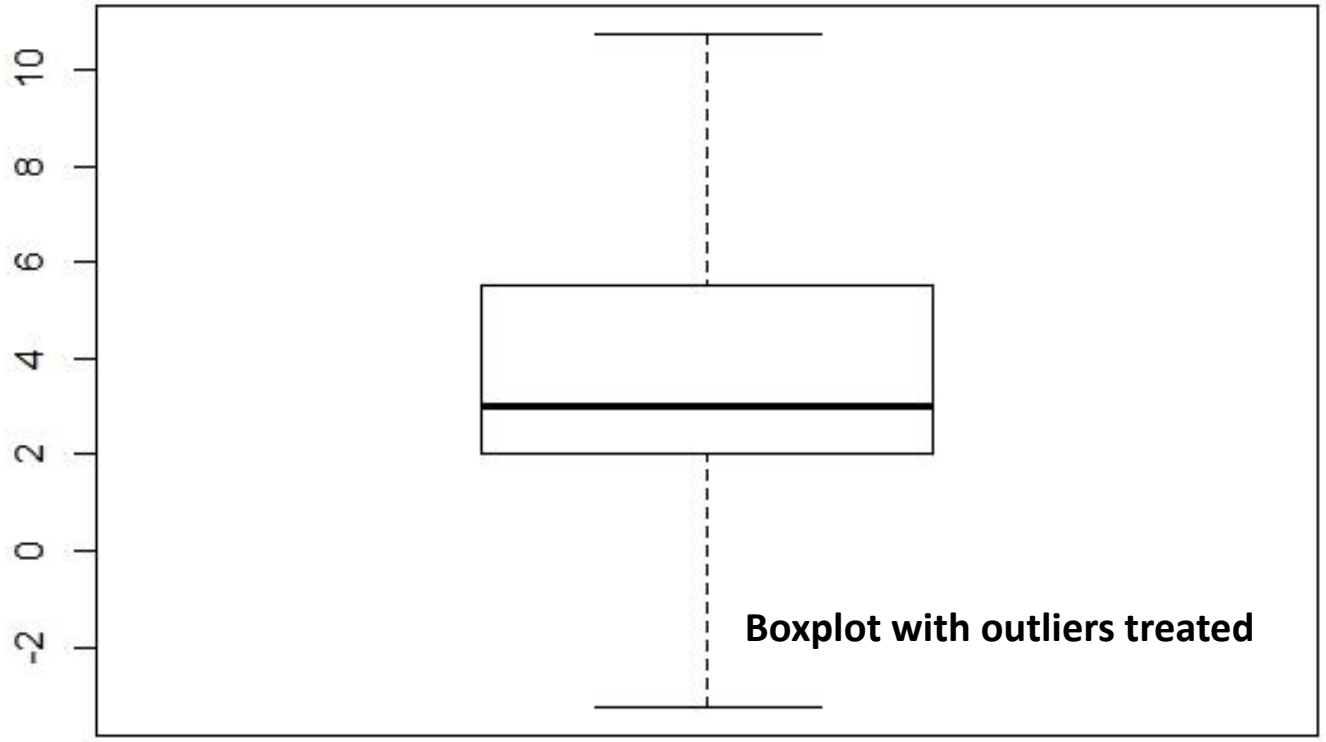
Boxplot



Treating Outliers

Post Outlier Treatment

Quartile based : Any value beyond the range of ± 1.5 x IQR of either quartiles should be capped



Thank You