APPLIED STATISTICS

Prashant Sahu

Data Science Trainer

Prashant is a highly motivated and passionate Corporate Trainer with experience in a variety of corporate trainings.

LinkedIn

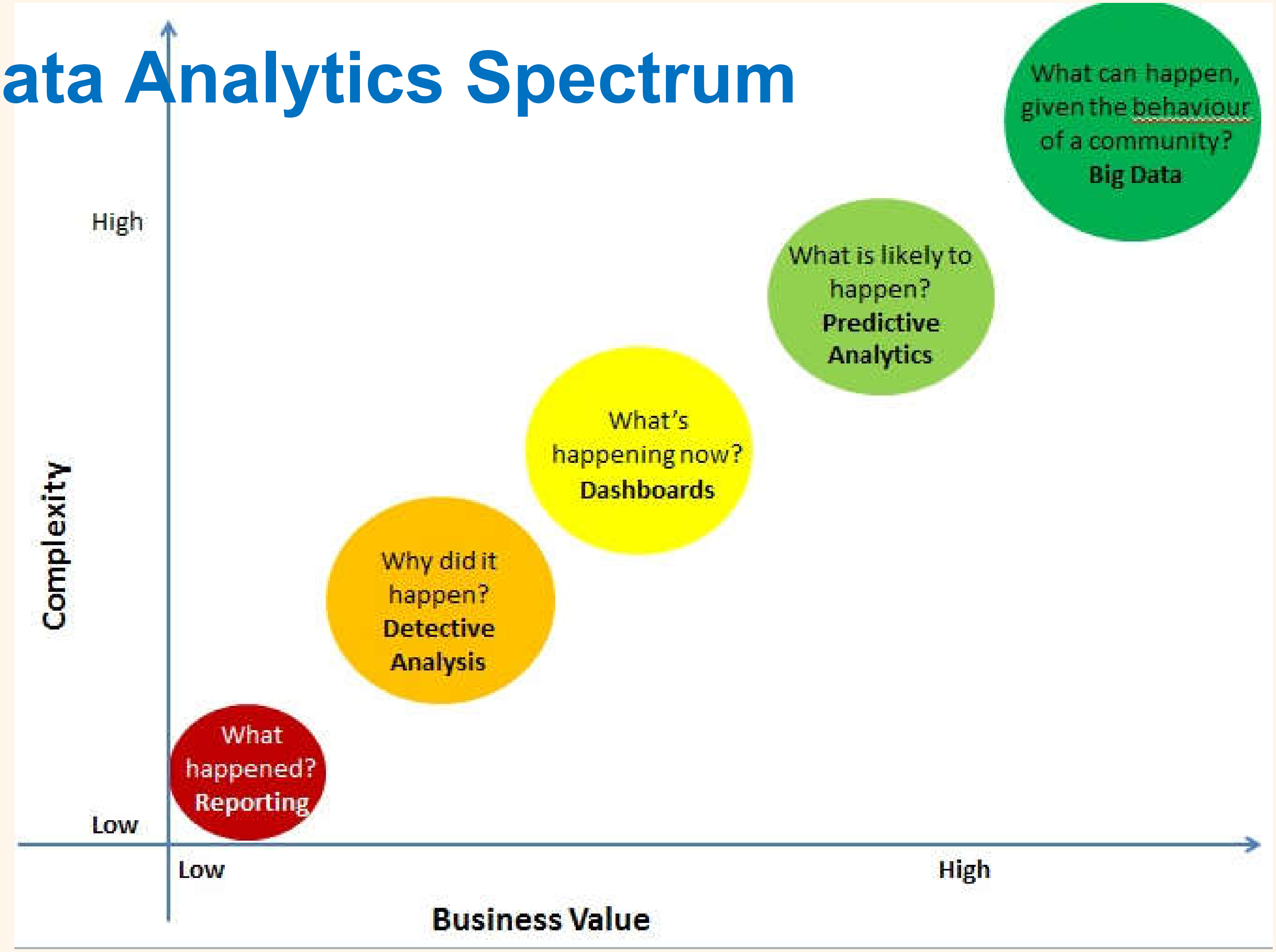
- Has overall 17 years of corporate experience
- Conducted many corporate training programs in AI, Machine Learning & Deep Learning using Python for engineers from Nomura, Bank of America, JPMorgan Chase, DBS Bank, HSBC India & HSBC HongKong, Reliance Industries Ltd., CRIF High Mark, Collins Aerospace Syntel, Qualys, Tata Communications, Tata Elxsi, Capgemini, Cybage, Cognizant Technology Solutions etc., both as lateral training and for freshers (induction programs) onsite.

What is Statistics?

- The art and science of learning from data
- Translating data into knowledge

It is a four-step process:
(1) Formulate a statistical question,
(2) Collect data,
(3) Analyze data, and
(4) Interpret results.

Data Analytics Spectrum



Descriptive Statistics?

- Descriptive analytics is about finding “**what has happened**” by summarizing the data using innovative methods and analyzing the past data using simple queries.

- Primary objective of descriptive analytics is simple comprehension of data using data summarization, basic statistical measures and visualization.

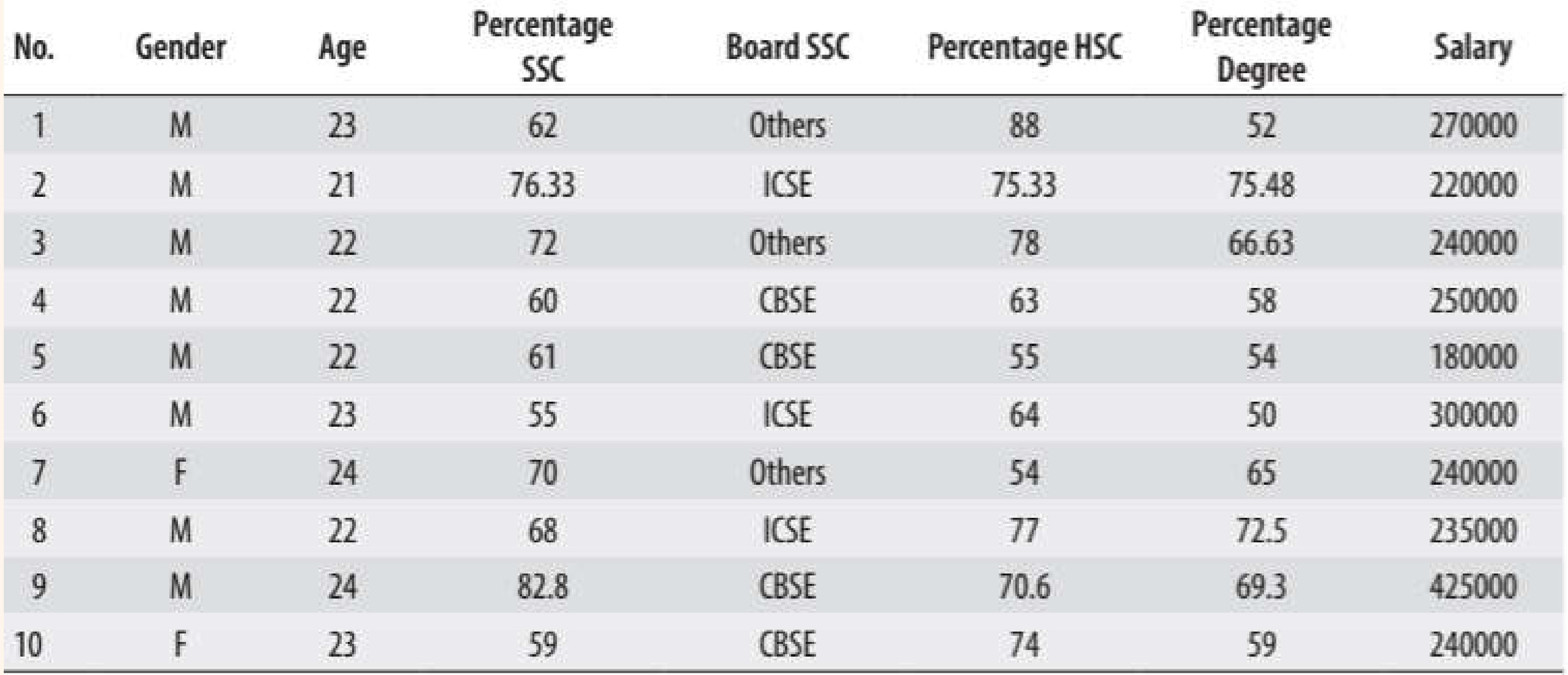
- Various tools and techniques are used in describing the data.

What is Data?

- The collected observations we have about something.
- Data is classified into different categories based on data structure and scale of measurement of the variables.
  - Structured and Unstructured Data
  - Cross-sectional, Time Series, and Panel Data

Structured Data

- Structured data means that the data is described in a matrix form with labelled rows and columns.



No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

Unstructured Data

- Data that is not originally in the matrix form with rows and columns is an unstructured data.
- For example,
    - E-mails, click streams, textual data, images (photos and images generated by medical devices), log data, and videos.
    - Machine generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG) and thermography are few examples of unstructured data.

Cross-sectional, Time Series, and Panel Data

- **Cross-sectional data** :-A data collected on many variables of interest at the same time or duration of time is called cross-sectional data.
  - For example, consider data on movies such as budget, box-office collection, actors, directors, genre of the movie during year 2017.
- **Time Series Data** :- A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.
- **Panel Data**:- Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).
  - Example of a panel data is data collected on variables such as gross domestic product (GDP), Gini index, and unemployment rate for several countries over several years.

Types of Data Measurement Scales

1. **Nominal Scale (Qualitative Data) :-** Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables.

   - For example, variables such as marital status (single, married, divorced) and industry type (manufacturing, healthcare, banking and finance)

2. **Ordinal Scale :-** Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.

   - For example, assume that a feedback is collected on a training program using 5-point Likert scale in which 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent.

3. **Interval Scale :-** Interval scale corresponds to a variable in which the value is chosen from an interval set.

   - Variable such as temperature measured in centigrade (°C) or intelligence quotient (IQ) score are examples of interval scale.

4. **Ratio Scale :-** Any variable for which the ratios can be computed and are meaningful is called ratio scale.

   - Ms. Rose's salary is 40,000 per month and Ms. June's salary is 90,000 per month then we can interpret that June earns 2.25 times the salary of Rose.

Population and Sample

- **Population** is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases.
  - For example, in 2014, close to 834.08 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the voters in 2014 was 834.08 million which included all eligible voters.
- **Sample** is the subset taken from a population.
  - It is very difficult (also practically impossible) to collect data from all 834.08 million eligible voters about their choice of candidate, so the opinion polls are based on opinion expressed by a subset of voters

Measures of Central Tendency

- Measures of central tendency are the measures that are used for describing the data using a single value.

- **Mean**, **median** and **mode** are the three measures of central tendency and are frequently used to compare different data sets.

- Measures of central tendency help users to summarize and comprehend the data.

Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency.

$$\text{Mean} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

If the entire population is available and if we calculate mean based on the entire population, then we get the population mean which is denoted by **μ** .

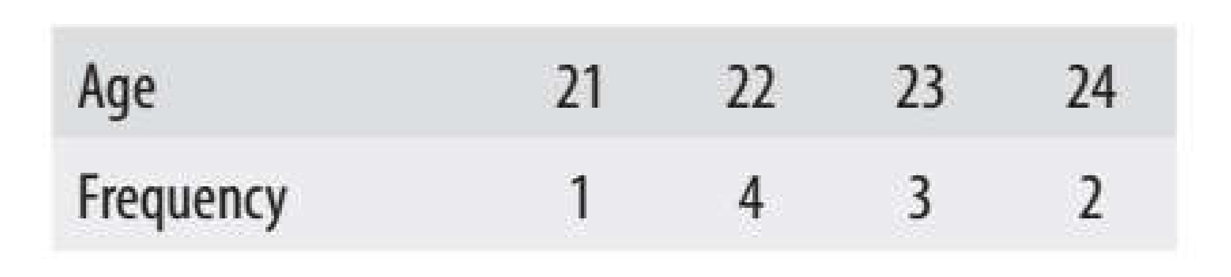
NOTE:- One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that, “if someone’s head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive”.

Making decisions solely based on mean value is not advisable.

- An important property of mean is that the summation of deviation of observations from the mean is zero, that is,

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

- If the data is captured in frequencies, then it can be used for calculating the average:

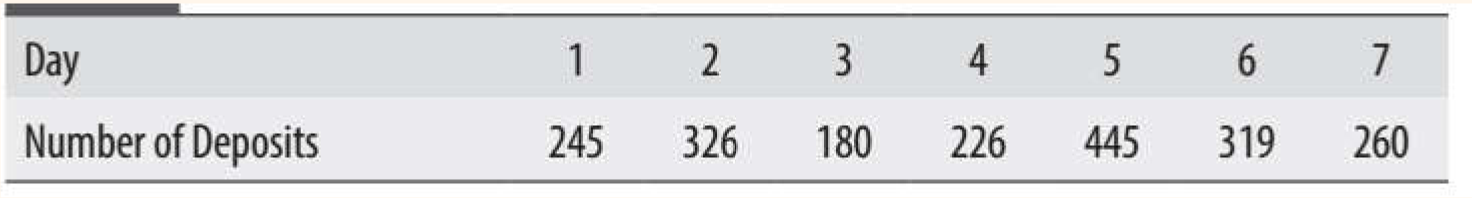


Age	21	22	23	24
Frequency	1	4	3	2

$$\bar{X} = \frac{1\times 21 + 4\times 22 + 3\times 23 + 2\times 24}{1+4+3+2} = 22.6$$

Median (or Mid) Value

- Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%.
- Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position (n + 1)/2 when n is odd.
- When n is even, the median is the average value of (n/2)th and (n + 2)/2th observation after arranging the data in the increasing order.

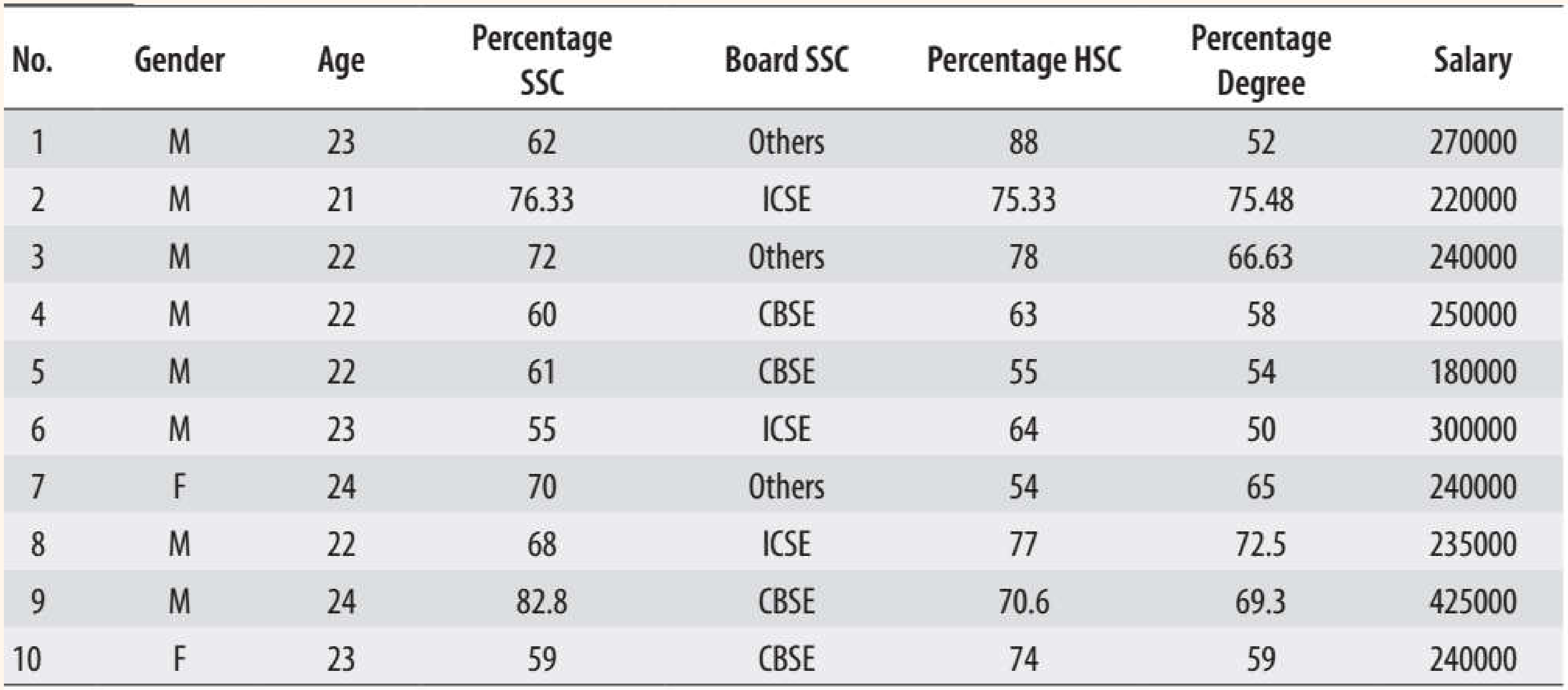


Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

The ascending order of the data:- 180, 226, 245, 260, 319, 326 and 445
Now (n + 1)/2 = (8/2) = 4. Thus the median is the 4th value in the data i.e. 260.

Mode

- Mode is the most frequently occurring value in the data set. The value 240000 is appearing three times and is the mode.



No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

Percentile, Decile and Quartile

- They are used to identify the position of the observation in the data set.
- **Percentile** score is frequently used in education to identify the position of a student in the group.
- Percentile, denoted as P_x ,is the value of the data at which x percentage of the data lie below that value. For example, P10 denotes the value below which 10 percentage of the data lies.

$$\text{Position corresponding to } P_x \approx \frac{x(n+1)}{100}$$

- **Decile** corresponds to special values of percentile that divide the data into 10 equal parts.
- First decile contains first 10% of the data and second decile contains first 20% of the data and so on.
- **Quartile** divides the data into 4 equal parts.
- The first quartile (Q1) contains first 25% of the data, Q2 contains 50% of the data and is also the median.

Measures of Variation

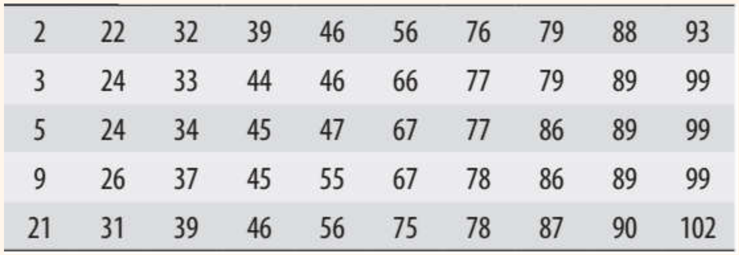
- One of the primary objectives of analytics is to understand the variability in the data.
- Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X).

Variability in the data is measured using the following measures:

1. Range
2. Inter-Quartile Distance (IQD)
3. Variance
4. Standard Deviation

Range

- Range is the difference between maximum and minimum value of the data. It captures the data spread.
- Time between failures of wire-cut (in hours) -



2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

- the range = 102 – 2 = 100.

Inter-Quartile Distance (IQD)

- Inter-quartile distance (IQD), also called inter-quartile range (IQR), is a measure of the distance between Quartile 1 (Q1) and Quartile 3 (Q3).
- IQD is a useful measure for identifying outliers in the data.
- **Outlier** is an observation which is far away (on either side) from the mean value of the data.
- Values of data below Q1 – 1.5 IQD and above Q3 + 1.5 IQD are classified as outliers.

Exercise

- Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in the Table. The function of the wire-cut is to cut the dough into cookies of desired size.

Variance and Standard Deviation

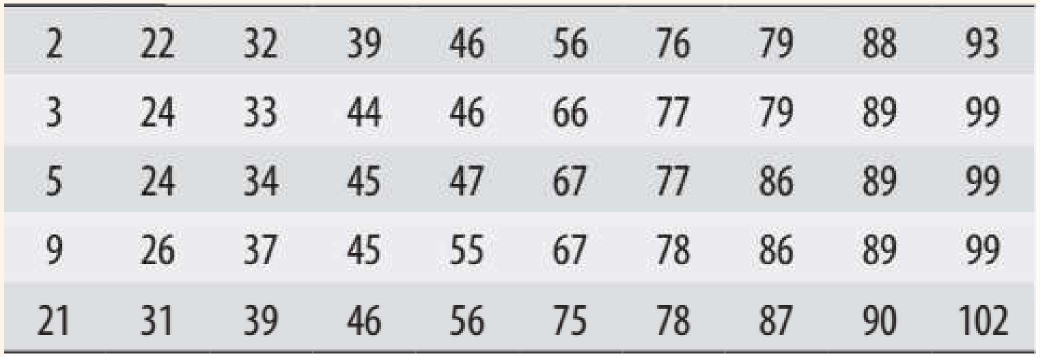
- **Variance is a measure of variability in the data from the mean value.**
- Variance for a population and sample is calculated using:

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \qquad S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

- Deviation from mean is squared since sum of deviations from mean will always add up to zero.
- The Population and Sample standard deviation s are given by:-

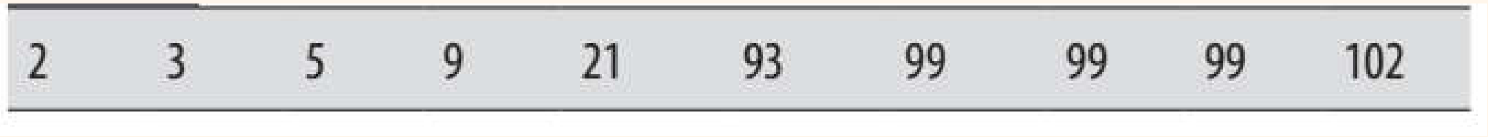
$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \qquad S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$$

- Population



2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

- Sample



2	3	5	9	21	93	99	99	99	102

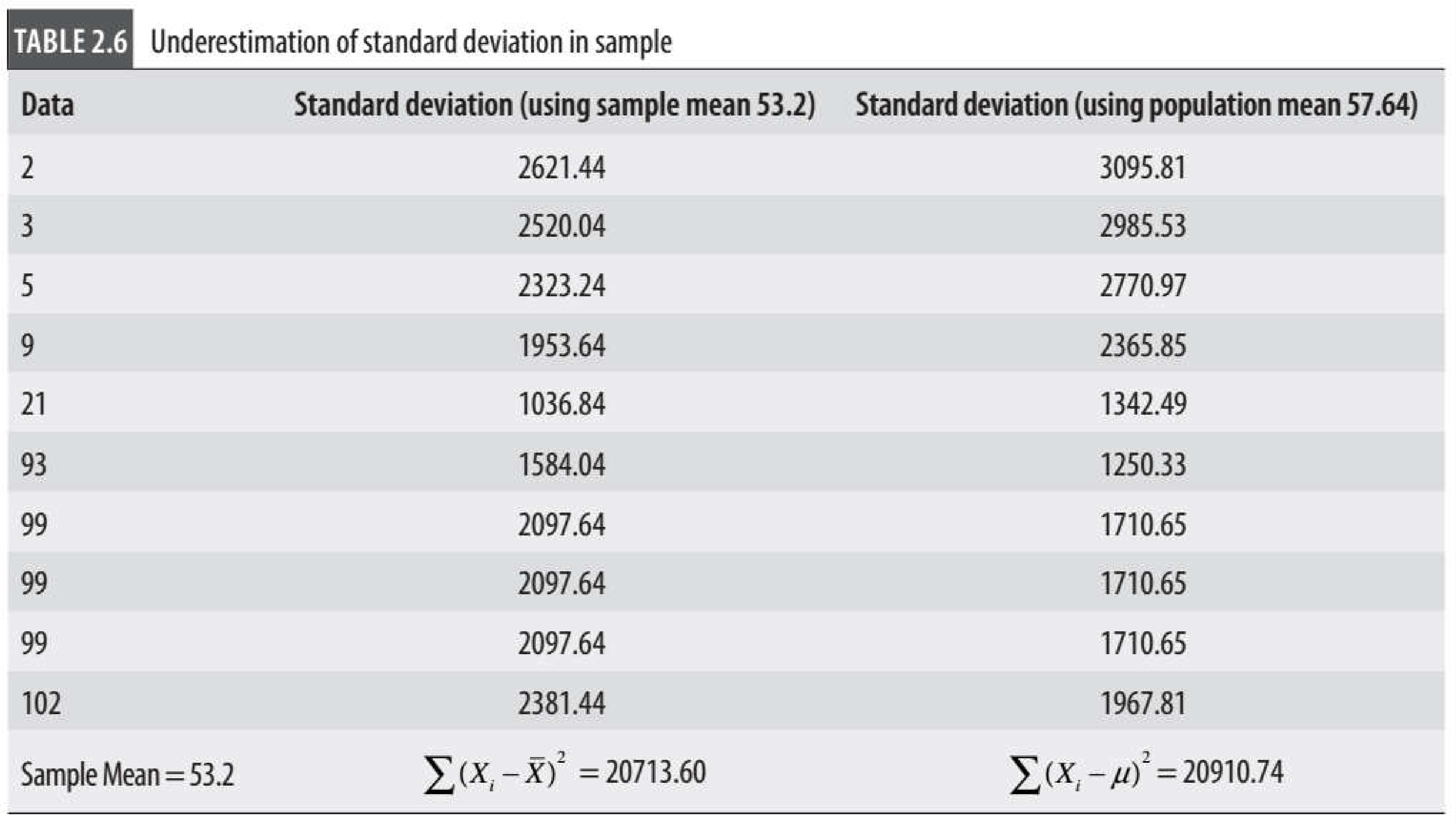


TABLE 2.6 Underestimation of standard deviation in sample

Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum(X_i - \bar{X})^2 = 20713.60$	$\sum(X_i - \mu)^2 = 20910.74$

Bessel Correction

- When we take a sample and estimate the mean from the sample, we tend to **underestimate the sum of squared deviations from the mean**.
- This is referred as **Downward Bias.**

- **Degrees of freedom** is equal to the number of independent variables in the model.
- For example, we can create any sample of size n with mean value of X-bar by randomly selecting (n − 1) values. We need to fix just one out of n values. Thus the number of independent variables in this case is (n − 1).

Degrees of Freedom

- Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated.

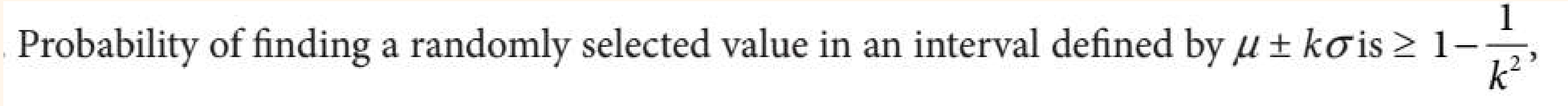
- If there are **n** observations in the sample and **k** parameters are estimated from the sample, then the degrees of freedom is **(n − k).**

- While using sample std. or variance, the value of X — is estimated from the sample. Thus the degrees of freedom is (n − 1).

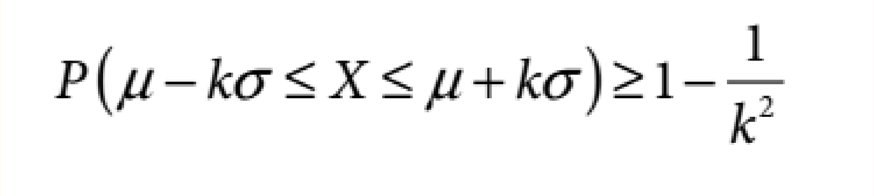
Prashant Sahu

Chebyshev’s Theorem

- Chebyshev’s theorem (also known as Chebyshev’s inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation.



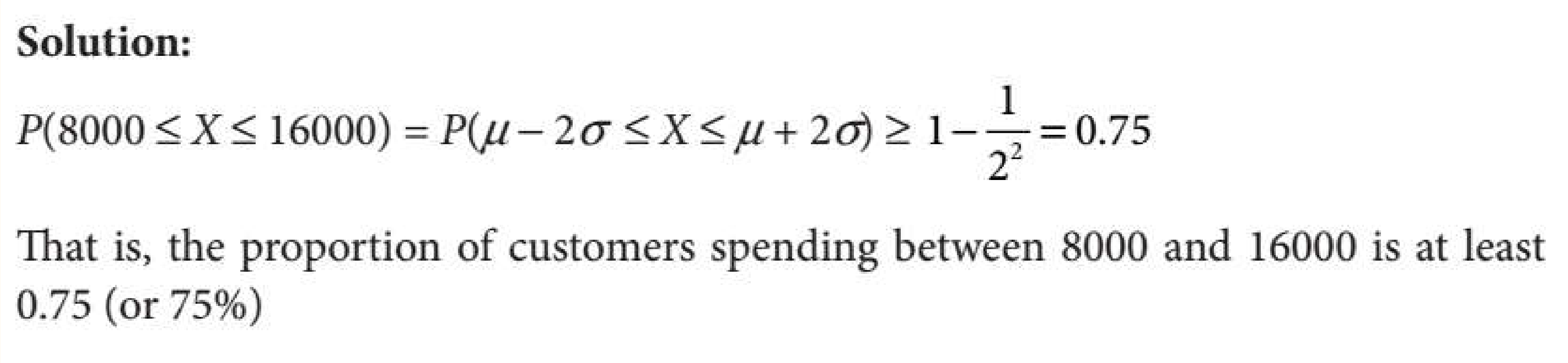
. Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is $\geq 1-\frac{1}{k^2}$,



$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Example

- Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000.



Solution:

$$P(8000 \le X \le 16000) = P(\mu - 2\sigma \le X \le \mu + 2\sigma) \ge 1 - \frac{1}{2^2} = 0.75$$

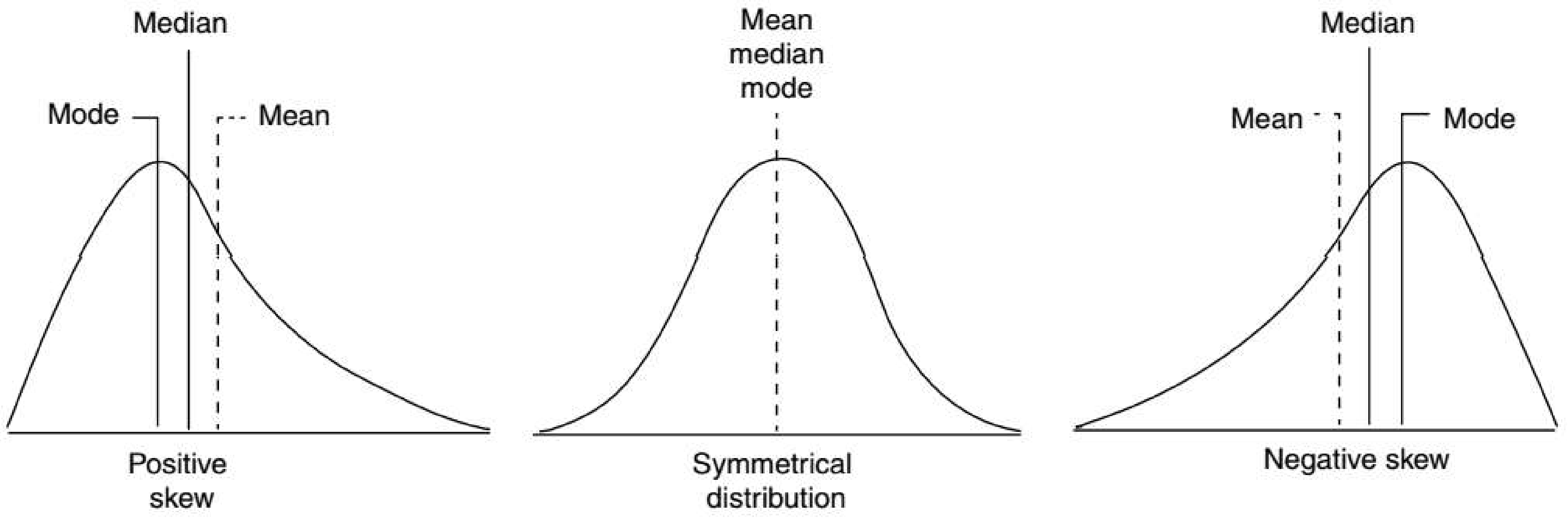
That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

Measures of Shape: Skewness and Kurtosis

- **Skewness is a measure of symmetry or lack of symmetry.**
- A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal.
- Measure of skewness can be used to identify whether the distribution is left skewed (longer tail on left side of the distribution) or right skewed (longer tail on the right side of the distribution).
- There are many different approaches to measuring skewness. **Pearson's moment coefficient of skewness** for a data set with n observations is given by:

$$g_1 = \frac{\sum_{i=1}^{n} (X_i - \mu)^3 / n}{\sigma^3}$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$



Prashant Sahu

Kurtosis

- Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light.

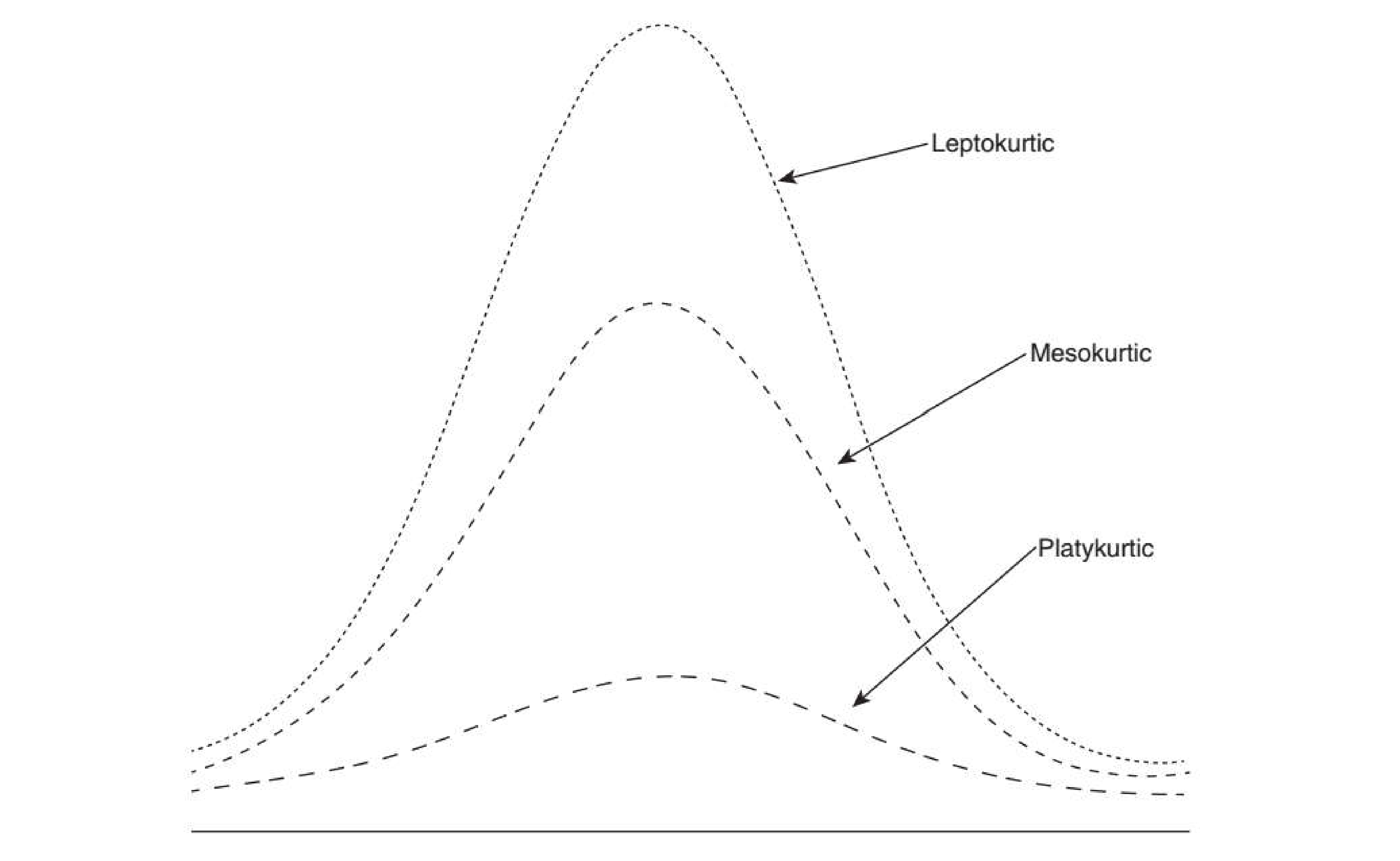
$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4}$$

- Kurtosis value of less than 3 is called **platykurtic** distribution and greater than 3 is called **leptokurtic** distribution.
- The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**).

Kurtosis

- Excess Kurtosis:

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4} - 3$$



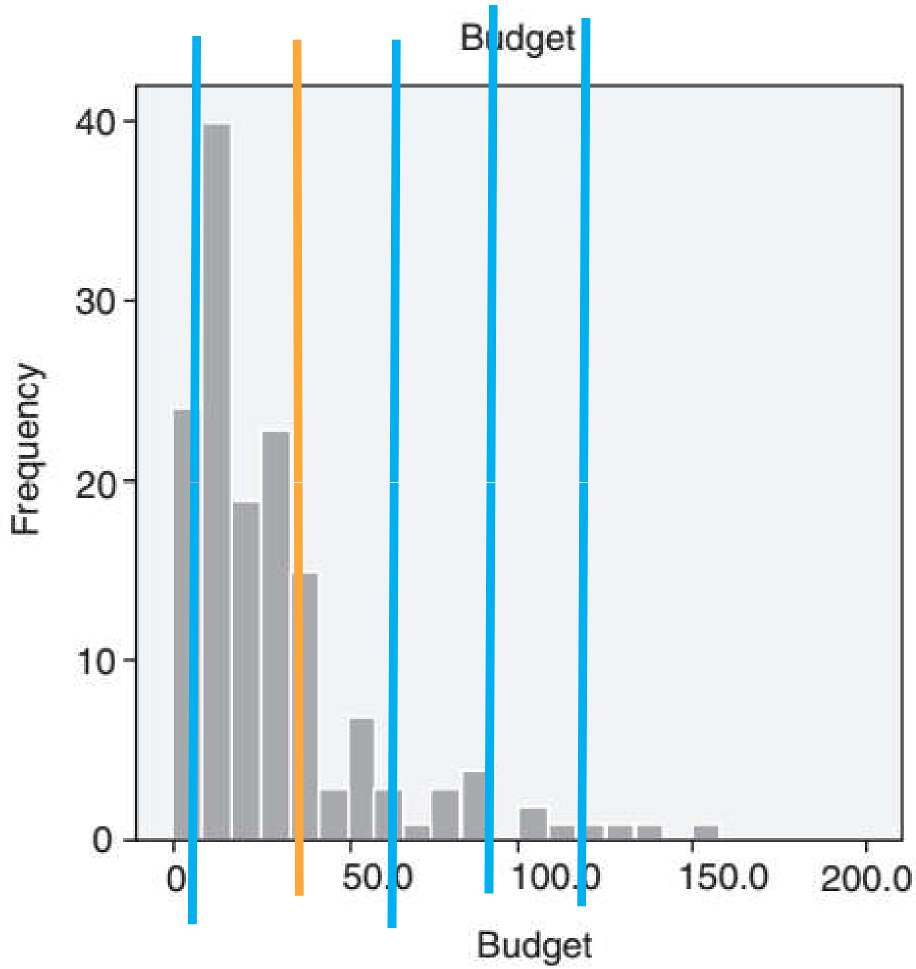
Prashant Sahu

DATA VISUALIZATION

Histogram

- Histograms are created for **continuous (numerical) data.**
- The following steps are used in constructing histograms:

1. Divide the data into finite number of non-overlapping and consecutive bins (intervals).
2. Count the number of observations from the data that fall under each bin (interval).
3. Create a frequency distribution (bin in the horizontal axis and frequency in the vertical axis) using the information obtained in steps 1 and 2.

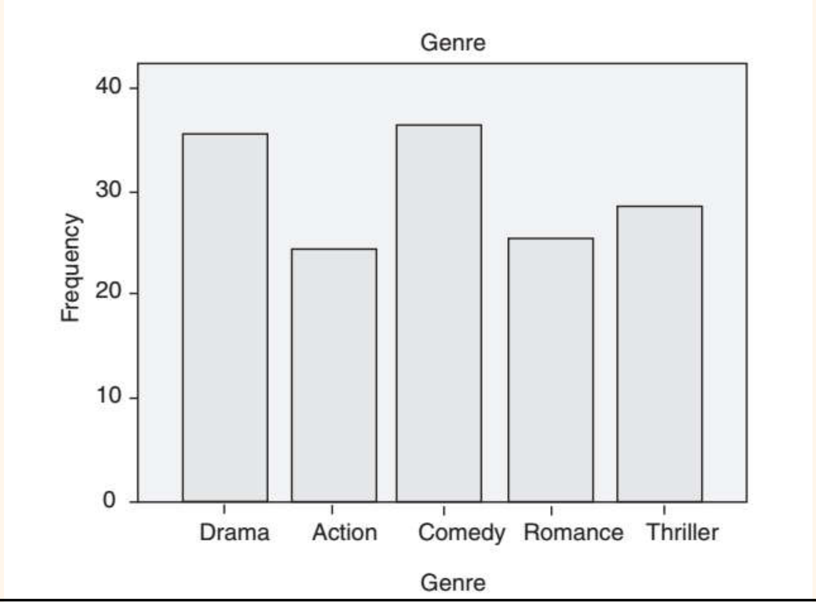


Mean = 29.43
Std. dev. = 28.25
N = 149

Prashant Sahu

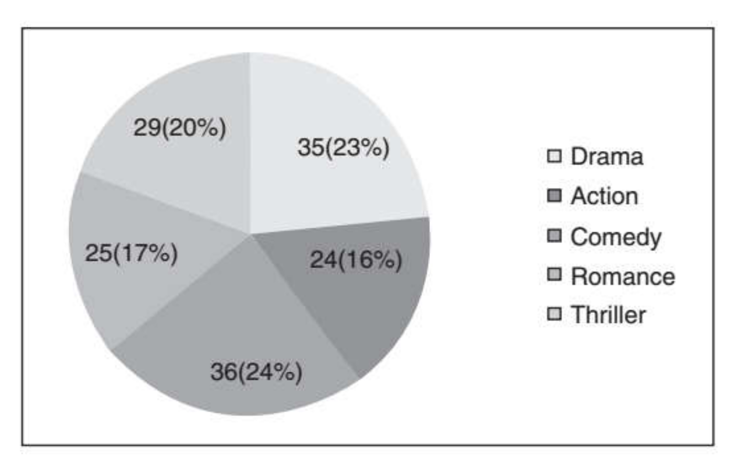
Bar Chart

- Bar chart is a frequency chart for **qualitative variable (or categorical variable).**
- Histograms cannot be used when the variable is qualitative.
- Bar chart can be used to assess the most-occurring and least-occurring categories within a data set.



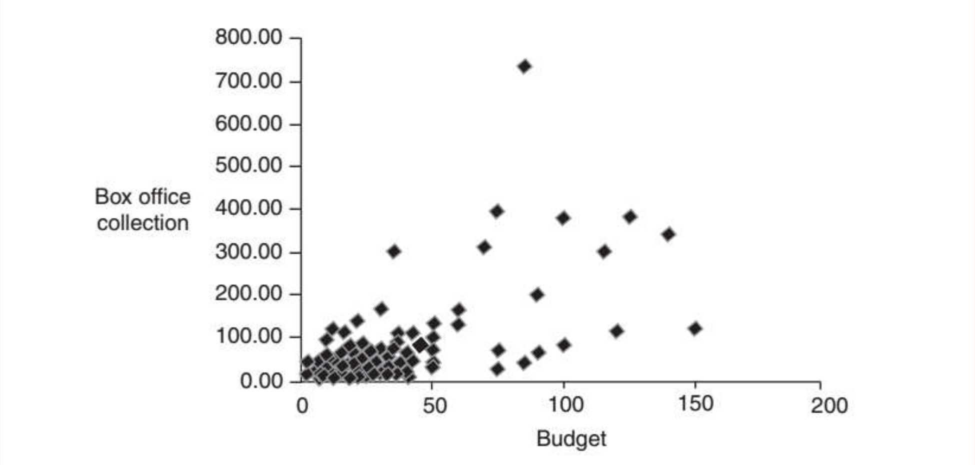
Pie Chart

- Pie chart is mainly used **for categorical data** and is a circular chart that displays the proportion of each category in the data set.
- Pie chart helps to visualize the **proportion (percentage)** of each category as sector of a circle.



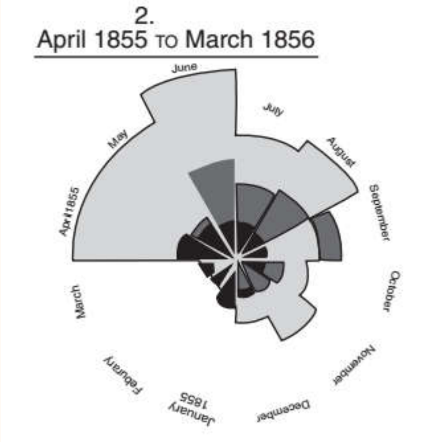
Scatter Plot

- Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables.
- The relationship could be linear or non-linear.
- Scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data.



Coxcomb Chart

- Coxcomb chart (also known as polar area chart or roses) is an extension of pie chart
- In a Coxcomb chart, each area represents the magnitude of the category.
- The main difference between the regular pie chart and coxcomb chart is that in the case of pie chart the radius of each sector is same, whereas, in coxcomb chart the radius of the sector is adjusted to create the magnitude of the area.

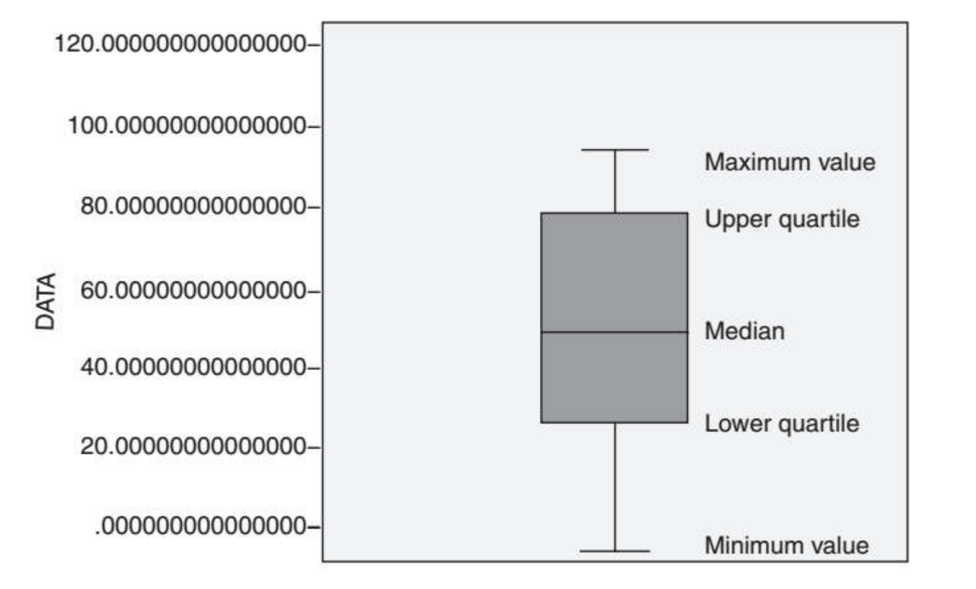


Box Plot (or Box and Whisker Plot)

- Box plot (aka Box and Whisker plot) is a graphical representation of numerical data that can be used to understand the variability of the data and the existence of outliers.
- Box plot is designed by identifying the following descriptive statistics:

1. Lower quartile (1st Quartile), median and upper quartile (3rd Quartile).
2. Lowest and highest value.
3. Inter-quartile range (IQR).

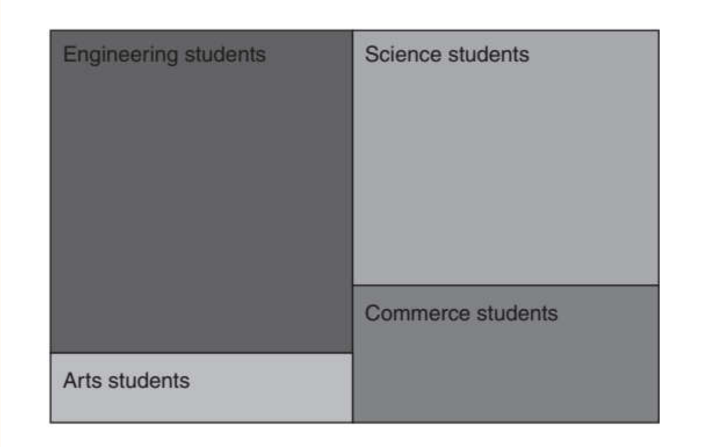
- The box plot is constructed using IQR, minimum and maximum values.



- The length of the box is equivalent to IQR. It is possible that the data may contain values beyond Q1 – 1.5 IQR and Q3 + 1.5 IQR.
- The whisker of the box plot extends till Q1 – 1.5 IQR (or minimum value) and Q3 + 1.5 IQR (or maximum value) Observations beyond these two limits are potential outliers.

Treemap

- Treemap is a hierarchical map made up of nested rectangles frequently used as part of business intelligence reports which helps organizations to understand the data hierarchically.
- To construct a treemap, the data should be hierarchical with several levels.
- The size of rectangle and colour are used for describing/differentiating the characteristics of the data.



Correlation Analysis – Day 3

Correlation is a statistical measure of an association relationship between two random variables.

- One of the challenging tasks in analytics, especially in predictive analytics, is identifying the variables or features that may be associated to the response variable or the outcome variable that is of interest to the data scientists.

Case Study : Telecom

For example, mobile service providers collect data on variables such as:

- call duration,
- number of calls,
- numbers to which the calls are made,
- number of calls received,
- the device that was used to make the call,
- location (and mobile tower that the phone was attached to),
- time between calls, last recharge (in case of pre-paid mobile services),
- recharge amount,
- service plan (in case of post-paid connection),
- number of messages sent, number of messages received,
- apps downloaded,
- time spent on surfing internet, and so on.

Telecom use-case >> Contd…

The idea behind collecting all these variables is to find answer to questions such as :

1. Which customer is likely to churn?
2. How to increase the revenue generated from a customer?
3. What is the customer lifetime value?
4. What is the best service plan for a customer?
5. What recommendations can be made to a customer?

Why is Correlation so important?

- Model building involves identifying the variables among thousands of variables (in analytics terminology this is called **variable selection or feature selection**) to build the model.
- Taking all the variables simultaneously to create a model can result in problems such as **multi-collinearity**, which can destabilize the model.
- Taking all Variables into modelling is also **time consuming** since most predictive analytics model development involves matrix operations such as matrix inverse calculation.
- *So, the knowledge of how different variables are related to one another is important in building analytical models*.

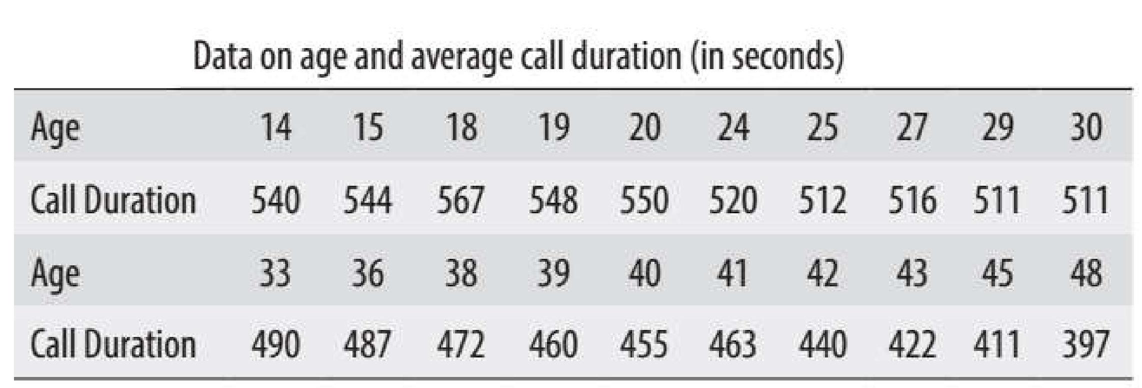
So then what exactly is Correlation?

- **Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient.**
- In others words, correlation is a measure of association between two variables.
- Correlation can assist the data scientists to choose the variables for model building that is used for solving an analytics problem.
- **Correlation is only an association relationship and not a causal relationship.**

Pearson Correlation Coefficient

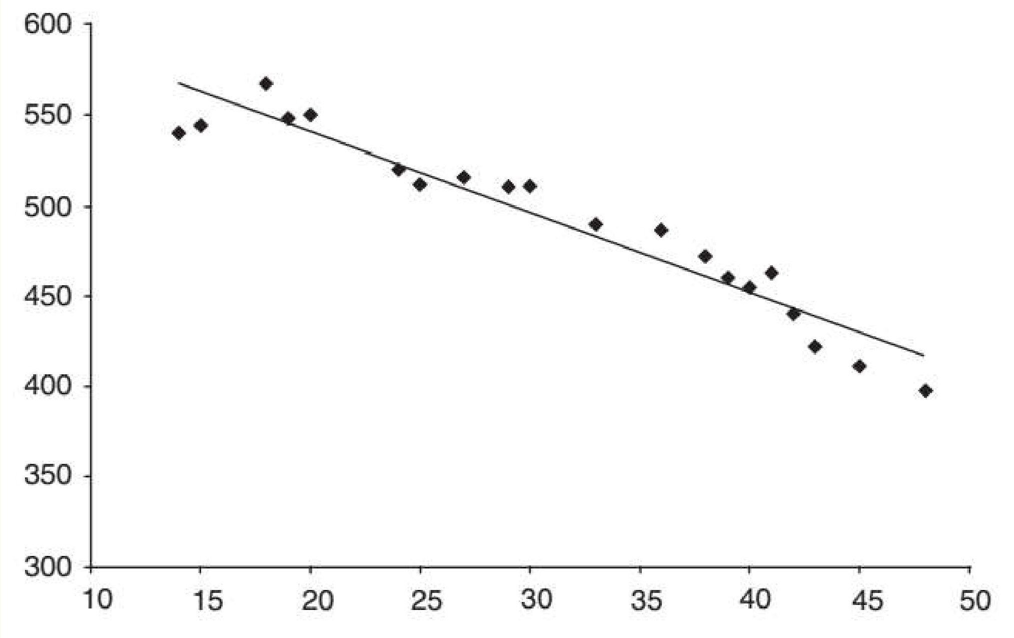
- Pearson product moment correlation (in short Pearson correlation) is used for measuring the **strength and direction of the linear relationship** between two continuous random variables X and Y.

- A simple approach for checking existence of association relationship is to draw a scatter plot.



Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

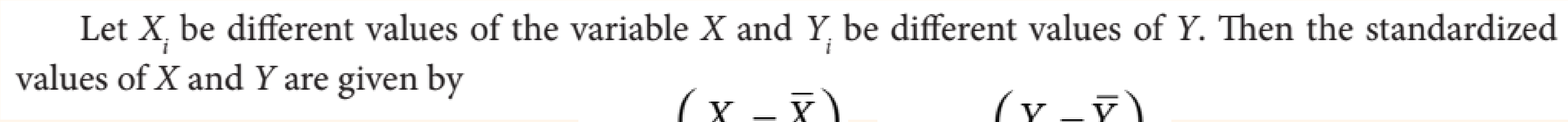


Association relationship between age and average call duration.

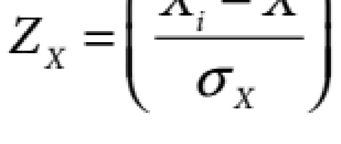
Prashant Sahu

Calculation of Pearson Coefficient

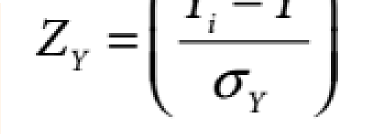
- Pearson product moment correlation is used when we are interested in finding linear relationship between two continuous random variables (that is, the variable should be either of ratio or interval scale).



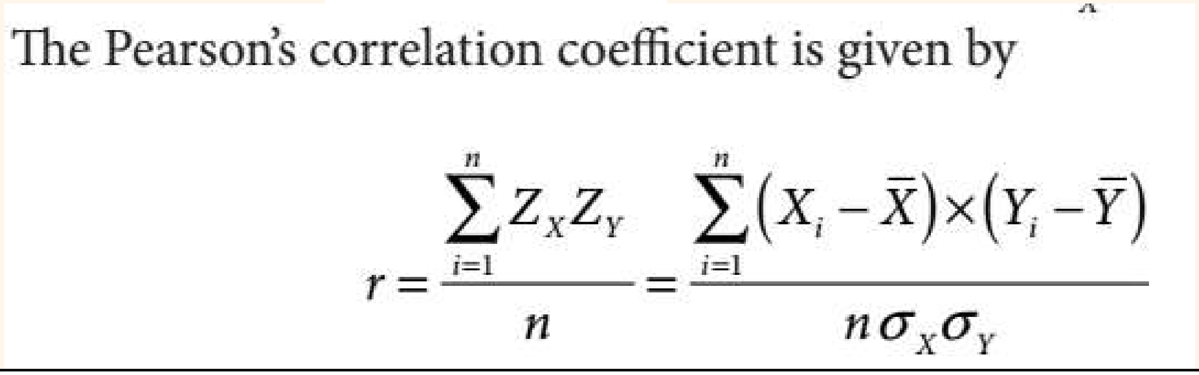
Let X_i be different values of the variable X and Y_i be different values of Y. Then the standardized values of X and Y are given by



$$Z_X = \left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

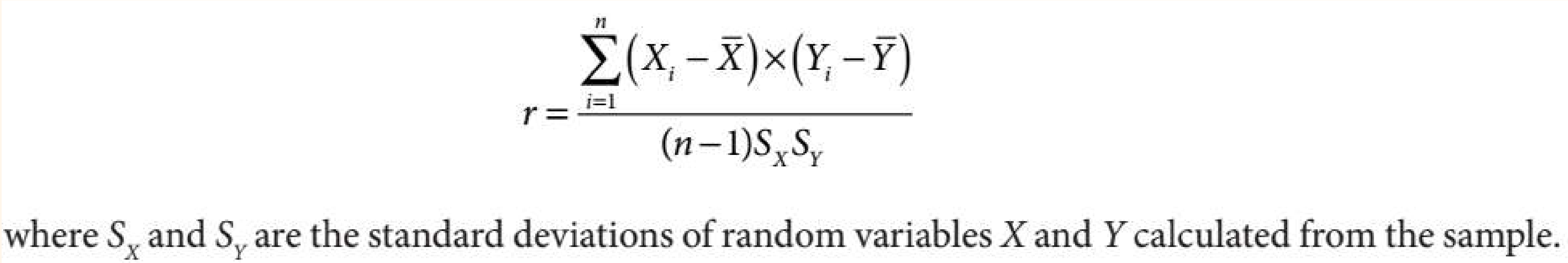


$$Z_Y = \left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right)$$



The Pearson's correlation coefficient is given by

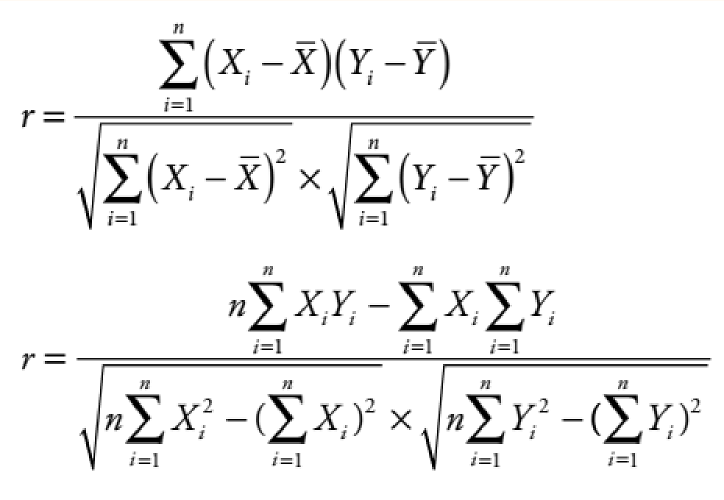
$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n\sigma_X \sigma_Y}$$



$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\times\left(Y_i - \bar{Y}\right)}{(n-1)S_X S_Y}$$

where S_X and S_Y are the standard deviations of random variables X and Y calculated from the sample.

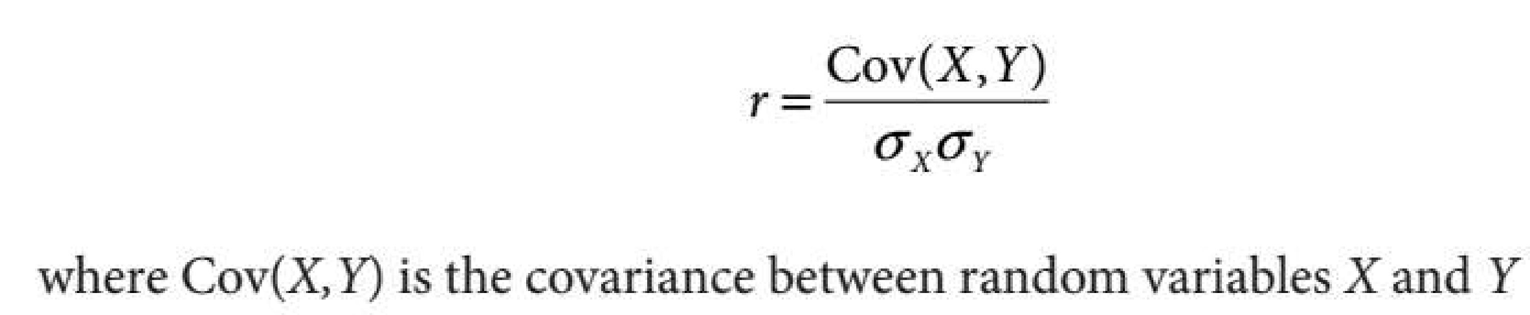
The value of Pearson’s correlation coefficient lies between −1 and +1.



$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \times \sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

Relation of Pearson Coeff with Covariance



$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\mathrm{Cov}(X,Y)$ is the covariance between random variables X and Y

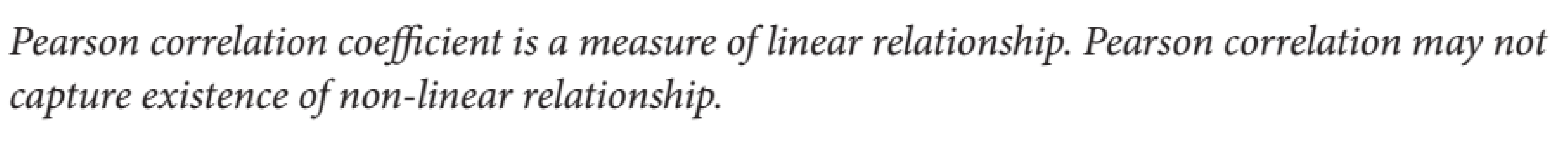
Properties of Pearson Correlation Coefficient

1. The value of correlation coefficient lies between −1 and +1. High absolute value of r, |r|, indicates strong relationship between the two variables.
2. Positive value of r indicates positive correlation (as value of X increases, the value of Y also increases) and negative value of r indicates negative correlation (as the value of X increases, the value of Y decreases).
3. The sign of correlation coefficient is same as the sign of covariance between the two random variables.

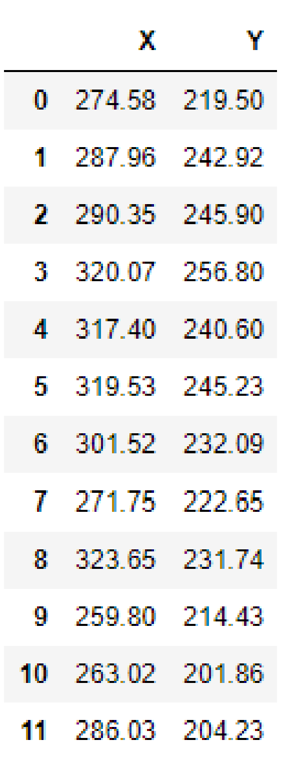
Properties of Pearson Correlation Coefficient

4. Mathematically, square of correlation coefficient is equal to the co-efficient of determination (R2) of the linear regression model,
that is r2 = R2.

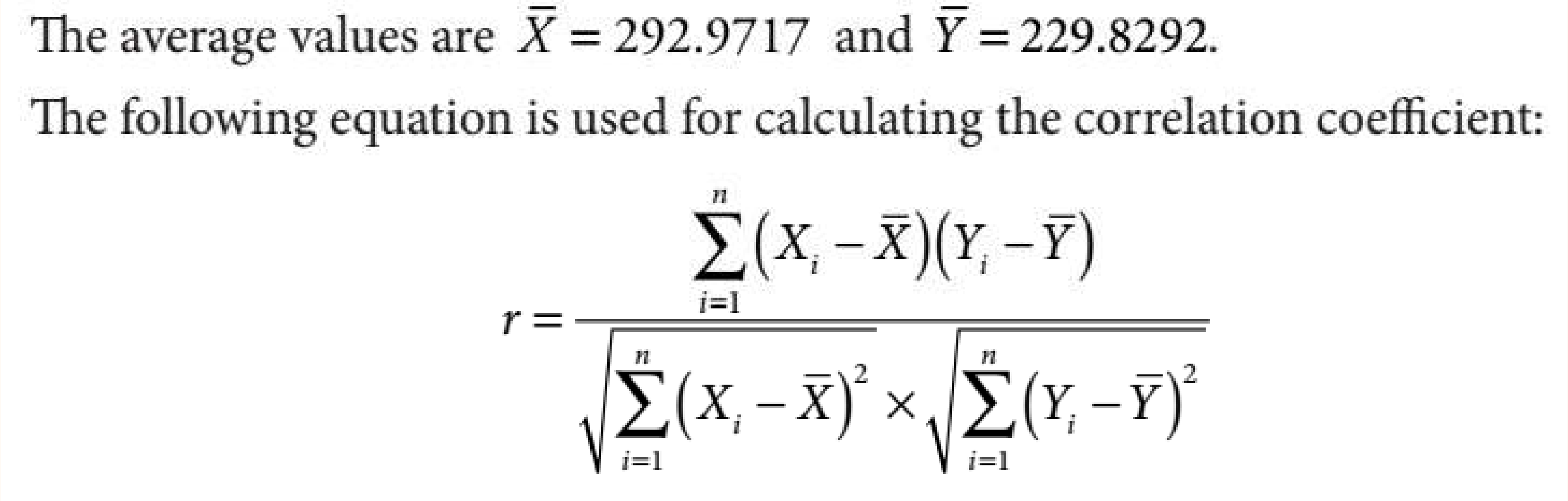
5. Pearson correlation coefficient value may be zero even when there is a strong non-linear relationship between variables X and Y. **Thus, low correlation coefficient value cannot be taken as an evidence of no relationship.**



Pearson correlation coefficient is a measure of linear relationship. Pearson correlation may not capture existence of non-linear relationship.



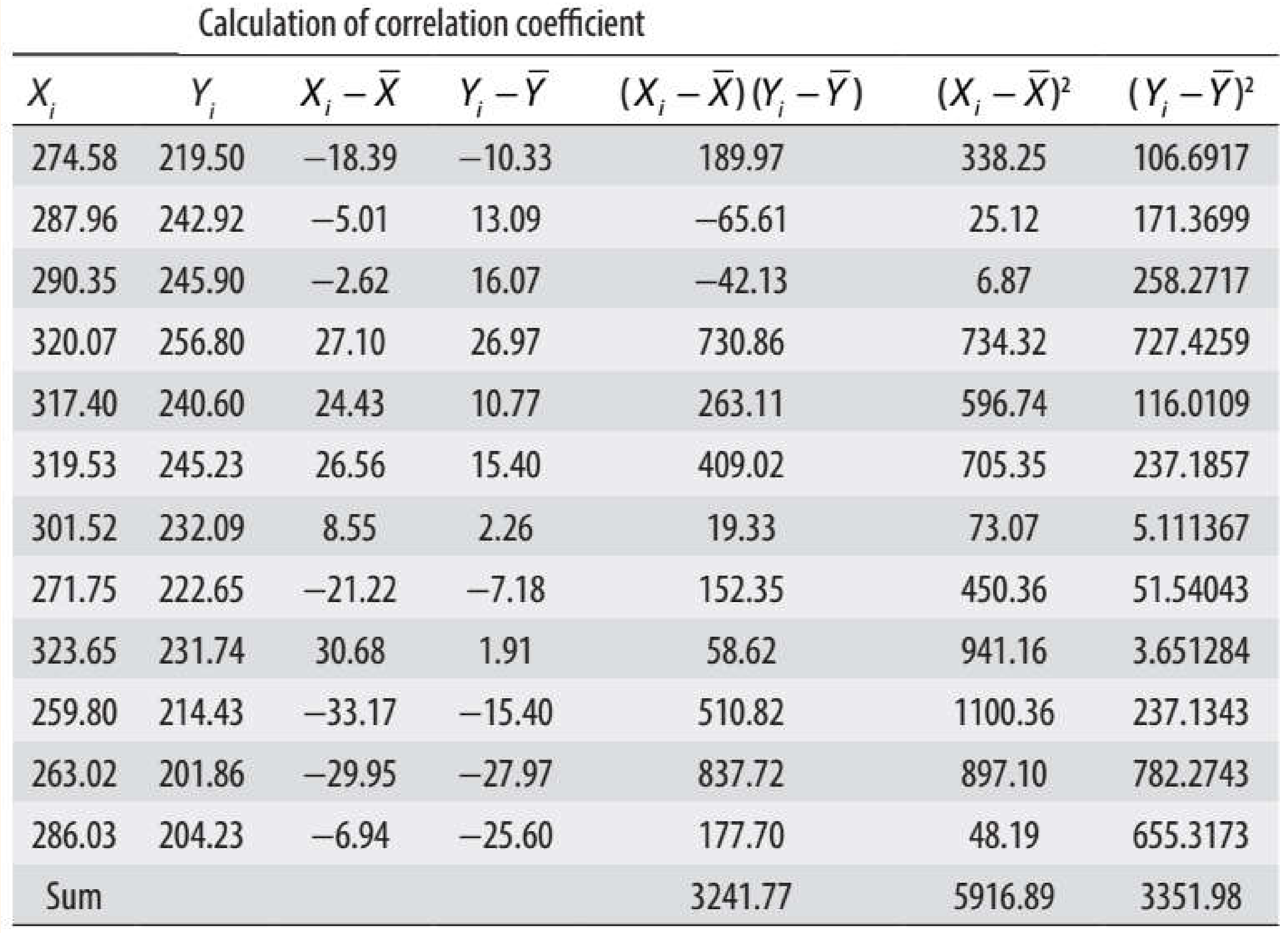
	X	Y
0	274.58	219.50
1	287.96	242.92
2	290.35	245.90
3	320.07	256.80
4	317.40	240.60
5	319.53	245.23
6	301.52	232.09
7	271.75	222.65
8	323.65	231.74
9	259.80	214.43
10	263.02	201.86
11	286.03	204.23



The average values are $\bar{X} = 292.9717$ and $\bar{Y} = 229.8292$.

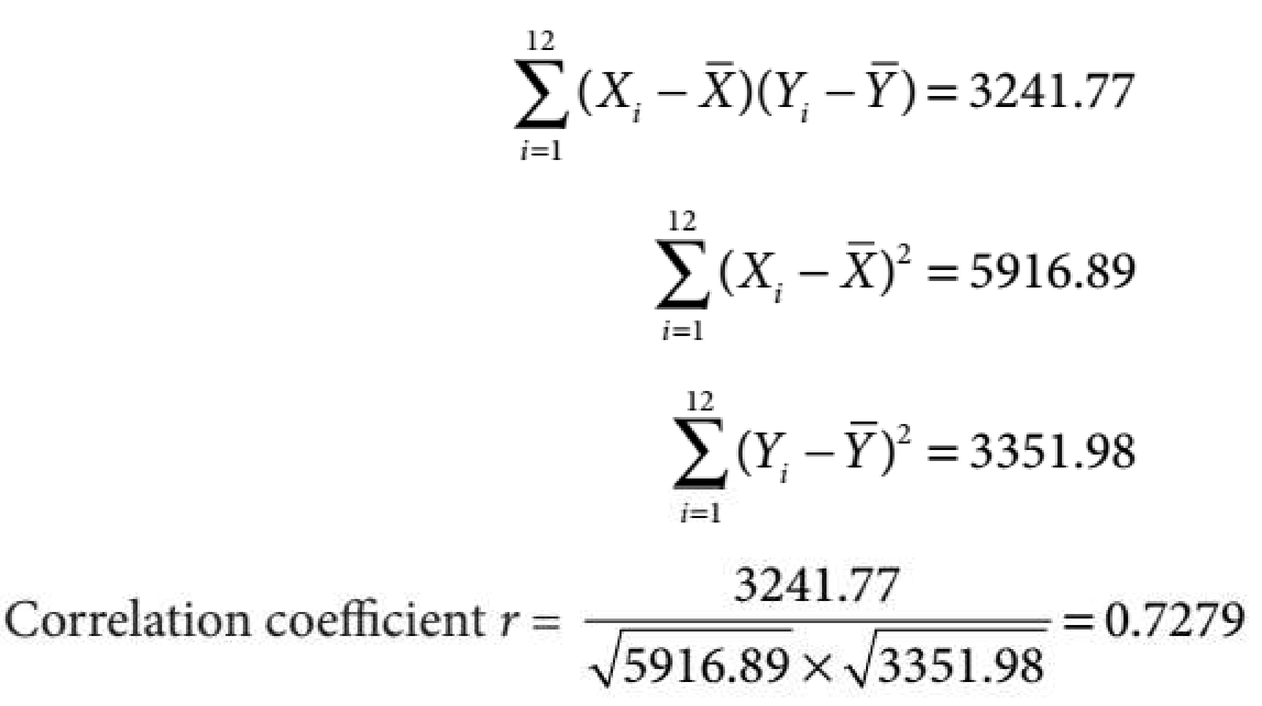
The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$



Calculation of correlation coefficient

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	−18.39	−10.33	189.97	338.25	106.6917
287.96	242.92	−5.01	13.09	−65.61	25.12	171.3699
290.35	245.90	−2.62	16.07	−42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	−21.22	−7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	−33.17	−15.40	510.82	1100.36	237.1343
263.02	201.86	−29.95	−27.97	837.72	897.10	782.2743
286.03	204.23	−6.94	−25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98



$$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$$

$$\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$$

$$\text{Correlation coefficient } r = \frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

In Microsoft Excel, CORREL(array 1, array 2) will give the Pearson product moment correlation value.

Spurious Correlation

- One of the major problem with correlation is the possibility of spurious correlation between two random variables which in many cases is caused due to some other latent variable (hidden variable) that influences both variables for which the correlation is calculated.

❑ Crime rate versus ice cream sale

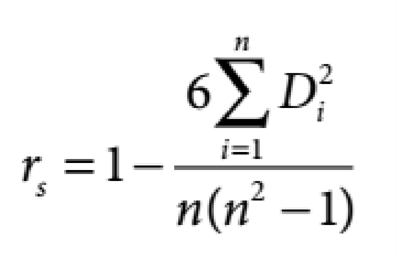
❑ Doctors and deaths

❑ Divorce rate in Maine and per capita consumption of margarine

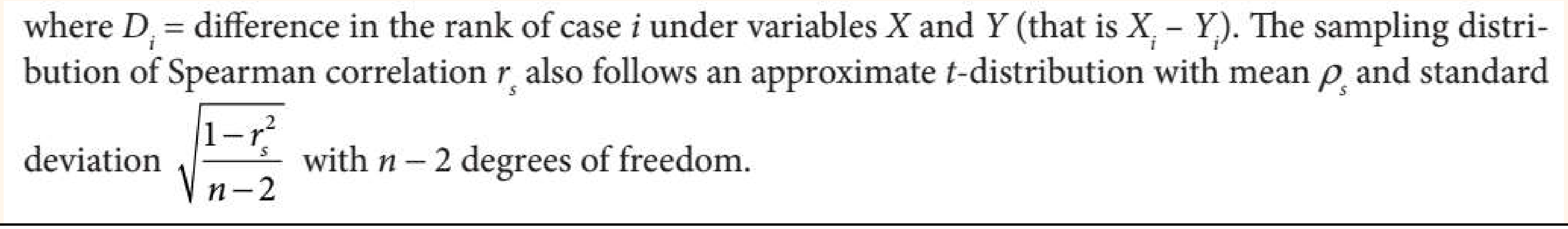
Spearman Rank Correlation

- Pearson correlation is appropriate when the random variables involved are both from either ratio scale or interval scale.

- **When both random variables are of ordinal scale, we use Spearman rank correlation** (also known as Spearman's rho denoted by ρ_s).



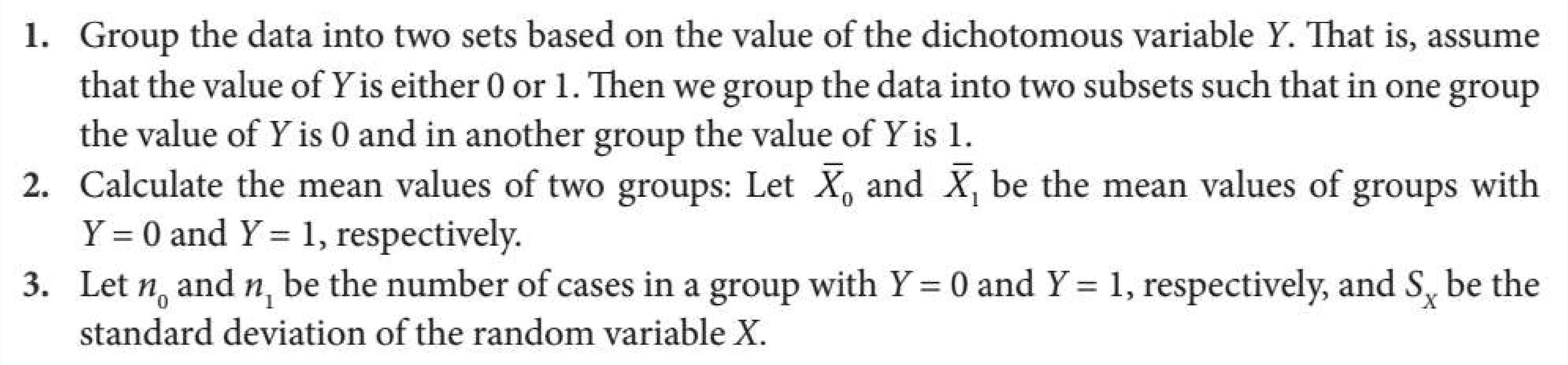
$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$



where D_i = difference in the rank of case i under variables X and Y (that is $X_i - Y_i$). The sampling distribution of Spearman correlation r_s also follows an approximate t-distribution with mean ρ_s and standard deviation $\sqrt{\frac{1-r_s^2}{n-2}}$ with $n - 2$ degrees of freedom.

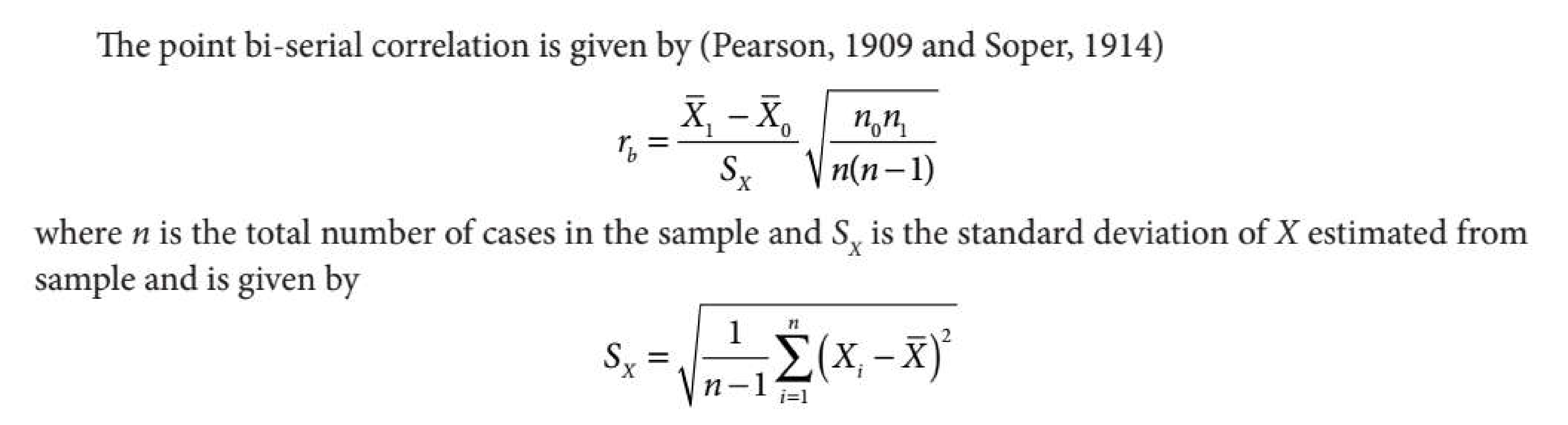
POINT BI-SERIAL CORRELATION

- Point bi-serial correlation is used when we are interested in finding correlation between a continuous random variable and a dichotomous (binary) random variable.
- Assume that the random variable X is a continuous random variable and Y is a dichotomous random variable.



1. Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1. Then we group the data into two subsets such that in one group the value of Y is 0 and in another group the value of Y is 1.
2. Calculate the mean values of two groups: Let $\bar{X}_0$ and $\bar{X}_1$ be the mean values of groups with $Y = 0$ and $Y = 1$, respectively.
3. Let n_0 and n_1 be the number of cases in a group with $Y = 0$ and $Y = 1$, respectively, and S_X be the standard deviation of the random variable X.

POINT BI-SERIAL CORRELATION



The point bi-serial correlation is given by (Pearson, 1909 and Soper, 1914)

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

where n is the total number of cases in the sample and S_X is the standard deviation of X estimated from sample and is given by

$$S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

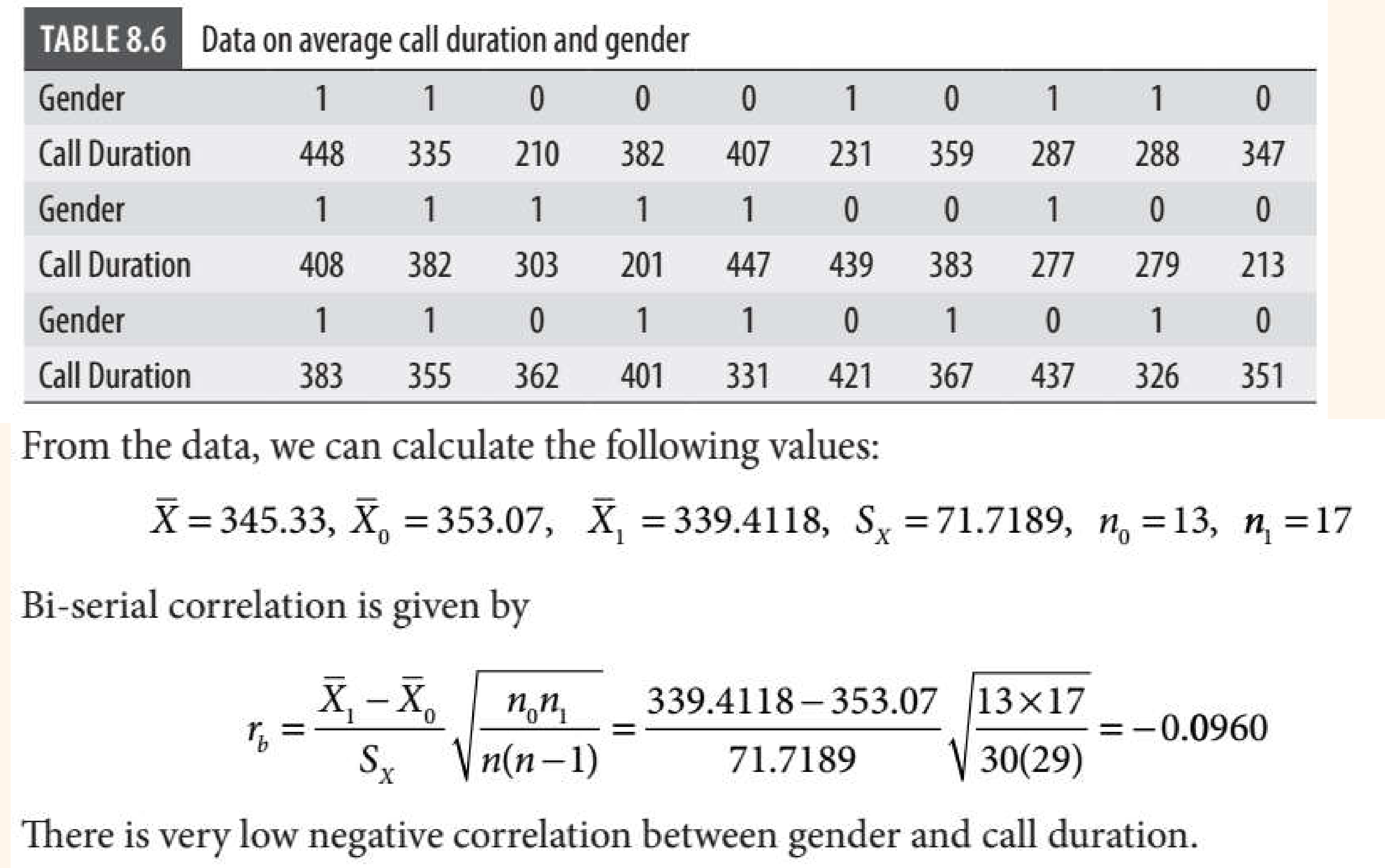


TABLE 8.6 Data on average call duration and gender

Gender	1	1	0	0	0	1	0	1	1	0
Call Duration	448	335	210	382	407	231	359	287	288	347
Gender	1	1	1	1	1	0	0	1	0	0
Call Duration	408	382	303	201	447	439	383	277	279	213
Gender	1	1	0	1	1	0	1	0	1	0
Call Duration	383	355	362	401	331	421	367	437	326	351

From the data, we can calculate the following values:

$$\bar{X} = 345.33,\ \bar{X}_0 = 353.07,\quad \bar{X}_1 = 339.4118,\ S_X = 71.7189,\ n_0 = 13,\ n_1 = 17$$

Bi-serial correlation is given by

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{339.4118 - 353.07}{71.7189}\sqrt{\frac{13 \times 17}{30(29)}} = -0.0960$$

There is very low negative correlation between gender and call duration.