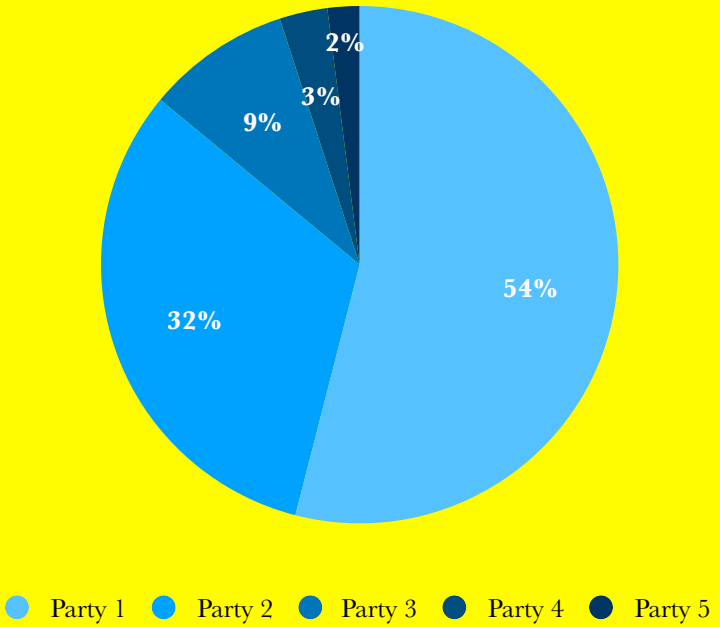
Sports Club A - Height Distribution

Observation	Age	Height	Sex
1	14	169	Male
2	17	156	Male
3	18	160	Female
4	17	165	Female
1400	15	177	Male

National Election Predictions



Basics of Statistics Part 1

Converting Data into Information

Instructor

Dr Sourish Das
Assistant Professor
Chennai Mathematical Institute
Common Wealth Rutherford Fellow
University of Southampton

What is Statistics?

Statistical methods can be used to find answers to the questions such as:

- How much data need to be collected and what kind of?
- How should we process, organize and summaries the data?
- How can we handle the uncertainty?
- How can we analyze the data and find some conclusions?
- How to assess the degree of strength of the conclusions?

Applications of Statistics

Statistics is the science of dealing with uncertain phenomenon and events. And uncertainty is almost everywhere in decision making.

Statistics in practice is applied successfully to study

- Health Care - The effectiveness of new medical drug
- Market Research - The perception on brand or a product to consumers
- Banking - Customer's ability to re-pay of a loan
- Advertising - The effectiveness of advertising to different channels

Statistical Processes

Broadly we can divide a full statistical analysis into:

Design: Planning and carrying out research studies.
What are the measures of a person we should consider to compute credit score?

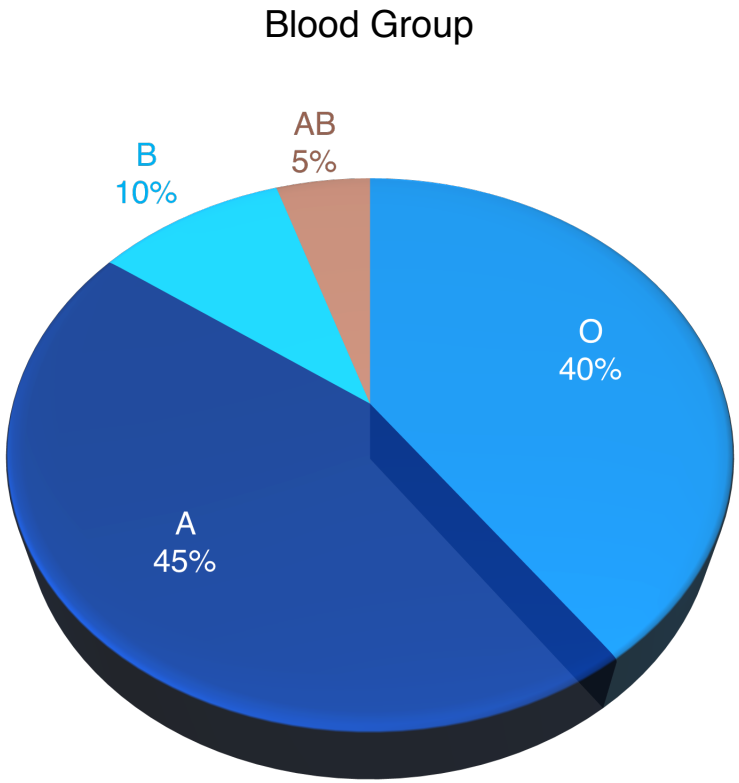
Description: Summarizing and exploring data.
Is all collected data useful for analysis?

Inference: Making predictions and generalizing about phenomena represented by the data.
Whether bank will disburse a loan or not, if yes then what will be the premium?

Descriptive Statistics

Descriptive statistics consist of methods for organizing and summarizing information (Weiss, 1999).

Blood Group	Frequency	Relative Frequency (%)
O	32	40
A	36	45
B	8	10
AB	4	5
Total	80	100



Descriptive Statistics

Parameters and Statistics

A parameter is an unknown numerical summary of the population.

A statistic is a known numerical summary of the sample which can be used to make inference about parameters.

(as per Agresti & Finlay, 1997)

Organization of Data

Variables

A characteristic or feature that varies from one subject to another is called a variable.

Qualitative Variables:

A variable that takes limited and usually a fixed number of values. Gender, eye color etc. are examples of qualitative or categorical variables.

Quantitative Variables:

Discrete variables.

Some variables such as number candidates enrolling for AI courses per day or number of children in family are the results of counting and thus these are discrete variables. As a definition, we can say that a variable is discrete if it has only a countable number of distinct possible values.

Continuous variables.

Quantities such as length, weight, or temperature can in principle be measured arbitrarily accurately, such a variable, called continuous.

Scales

The scale of the variable gives certain structure to the variable and also defines the meaning of the variable.

Qualitative Variables.
- Nominal scale
- Ordinal scale

Quantitative Variables.
- Interval scale
- Ratio scale

Scales - Qualitative Variables

Nominal Scale

Nominal scales are used for labeling variables, without any quantitative value. Here the categories of a qualitative variable are unordered. Examples can be

In which subject did you get highest marks in your graduation?

a) Artificial Intelligence
b) Database
c) Algorithm

Or

What was your mode of study?

a) Resident student
b) Commuter

Or

What was the color of your first car?

a) Black
b) Red
c) White
d) Blue

Scales - Qualitative Variables

Ordinal Scale

If the categories can be put in order, but the exact difference between two orders are not known we measure such categories on ordinal scale.
An example is the rating in a customer survey. Such as if any of your service providers ask you - how happy are you with their service and gives you options as below:
How happy are you with their service?

a) Very happy
b) Happy
c) OK
d) Not happy
e) Very unhappy

Here the order of the options is clear. But do you know what is the exact difference between 'Very happy' and 'Happy' or 'Happy' and 'OK'?
Do you know if the difference between 'Very happy' and 'Happy' is the same difference as the difference between 'Happy' and 'OK'?
You don't. No one knows the difference. We can only order them.

Scales - Quantitative Variables

Quantitative variables, whether discrete or continuous, are defined either on an interval scale or on a ratio scale.

Interval Scale:

Interval scale is a measure in which we know the exact difference between two values. Such as the difference between 100 degree Celsius and 90 degree Celsius. Here only the difference that is 10 degree matters. So the difference between 80 degree Celsius and 70 degree Celsius will be treated in the same way here as the difference between 100 degree Celsius and 90 degree Celsius. In this scale there is no true zero.

Scales - Quantitative Variables

Interval Scale - Explained:

Let us take the difference between 20 degree Celsius and 10 degree Celsius which is also 10. If we now convert 10 degree Celsius into Fahrenheit following the formula of conversion, we get 50 degree Fahrenheit. Following the same formula we get 20 degree Celsius is 68 degree Fahrenheit. Here we cannot say 68 degree Fahrenheit is twice as hot as 50 degree Fahrenheit obviously. The difference between 68 degree Fahrenheit and 50 degree Fahrenheit is 18 degree Fahrenheit. Now 30 degree Celsius is 86 degree Fahrenheit. Here the difference between 86 degree Fahrenheit and 68 degree Fahrenheit is exactly 18 degree Fahrenheit. Now you take 40 degree Celsius which is 104 degree Fahrenheit which is 86 degree Fahrenheit + 18 degree Fahrenheit. This tell us addition and subtraction is possible on interval scale but multiplication and division is not possible.

Scales - Quantitative Variables

Quantitative variables, whether discrete or continuous, are defined either on an interval scale or on a ratio scale.

Ratio Scale:

If we can compare both the difference between measurements of the variable and the ratio of the measurements meaningfully, then the quantitative variable is defined on ratio scale.
Example: The height of person is a ratio variable.

Dataset

Collecting the values of the variables for multiple people or subject generate data. Each individual piece of data is called an observation and the collection of all observations for particular variables is called a dataset or data matrix. Data set are the values of variables recorded for a set of sampling units.

Observation	Age	Height	Sex
1	45	169	Male
2	35	177	Male
3	24	160	Female
4	43	156	Female
5	65	170	Male
…	…	…	…

Describing Data by Tables and Graphs

For Qualitative Variable

Suppose we have thousands of observations in a dataset and there is one qualitative variable say marital status.

A way of describing this data is by its frequency distribution.

Observation	Age	Height	Sex
1	45	169	Male
2	35	177	Male
3	24	160	Female
4	43	156	Female
…	…	…	…
1400	65	170	Male

Describing Data by Tables and Graphs

Relative Frequency

Percentage of frequency for a class.

Relative frequency is calculated by dividing the frequency of the class by the total number of observations and multiplying the result by 100. The percentage of the class, expressed as a decimal, is referred to as the relative frequency of the class.

Sex	Frequency	Relative Frequency (%)
Male	735	??
Female	663	??
Trans	2	??
Total	1400	??

Describing Data by Tables and Graphs

Relative Frequency

Relative frequency for Male
= (735 / 1400) x 100
= 52.5

Sex	Frequency	Relative Frequency (%)
Male	735	52.5
Female	663	??
Trans	2	??
Total	1400	??

Describing Data by Tables and Graphs

Relative Frequency

Relative frequency for Male
= (735 / 1400) x 100
= 52.5

Relative frequency for Female
= (663 / 1400) x 100
= 47.4

Relative frequency for Transgender
= (2 / 1400) x 100
= 0.1

Sex	Frequency	Relative Frequency (%)
Male	735	52.5
Female	663	47.4
Trans	2	0.1
Total	1400	100

Describing Data by Tables and Graphs

Cumulative Frequency

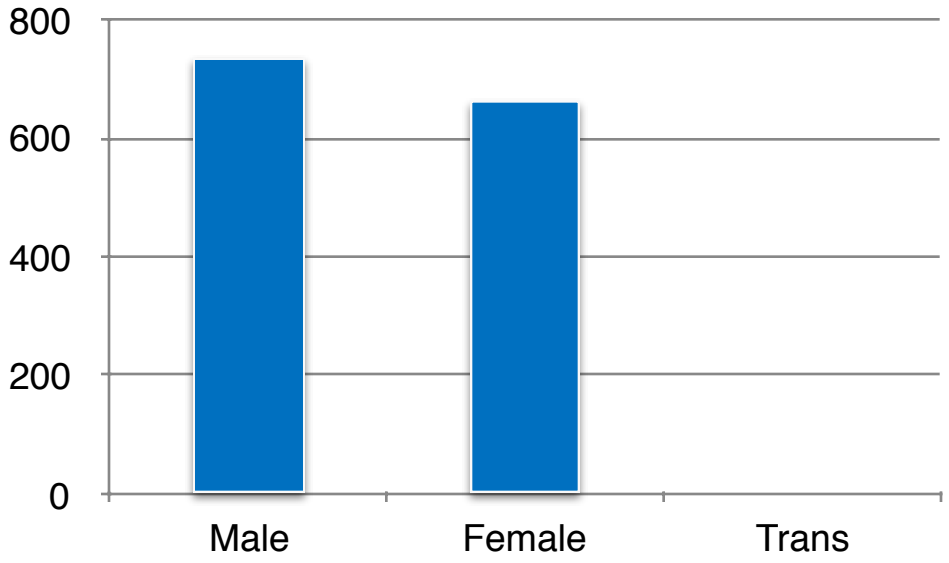
Cumulative frequency up to Female starting at Male is
= 735 + 663
= 1398

Cumulative frequency up to Transgender starting at Male is
= 1398 + 2
= 1400, which is the total sample size in this case

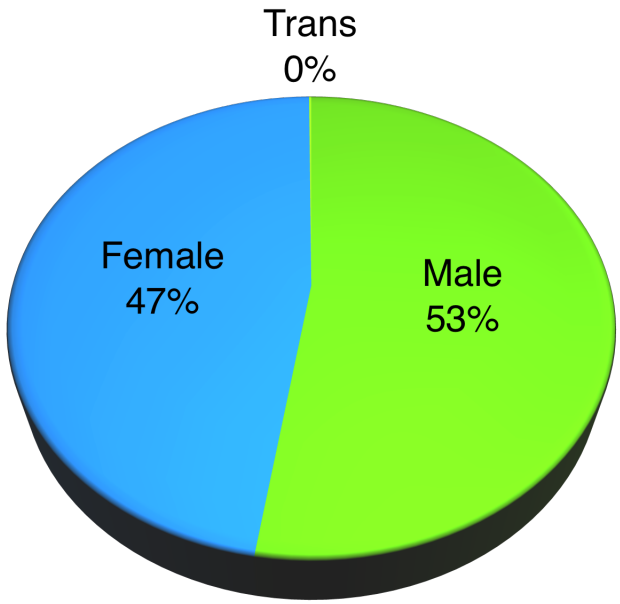
Sex	Frequency	Cumulative Frequency
Male	735	
Female	663	1398
Trans	2	1400

Describing Data by Tables and Graphs

Frequency



Relative Frequency

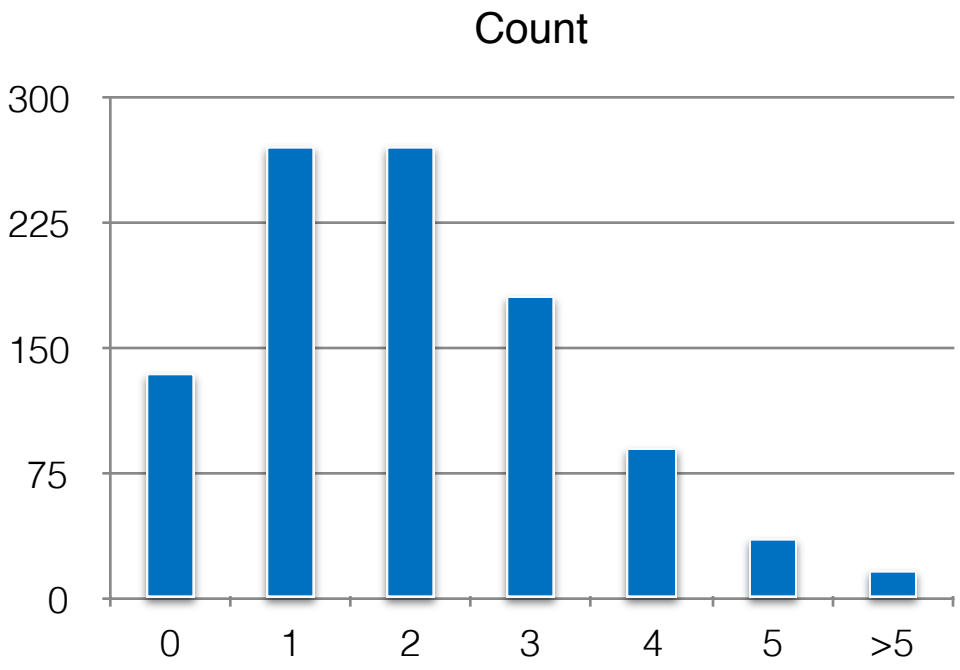


Describing Data by Tables and Graphs

For Discrete Quantitative Variable

The data of the discrete variable can be summarized in a same way as qualitative variables in a frequency table. In a place of the qualitative categories, we now list in a frequency table the distinct numerical measurements that appear in the discrete data set and then count their frequencies.

Errors (in a page)	No Page
0	135
1	271
2	271
3	180
4	90
5	36
> 5	17



Describing Data by Tables and Graphs

For Continuous Quantitative Variable

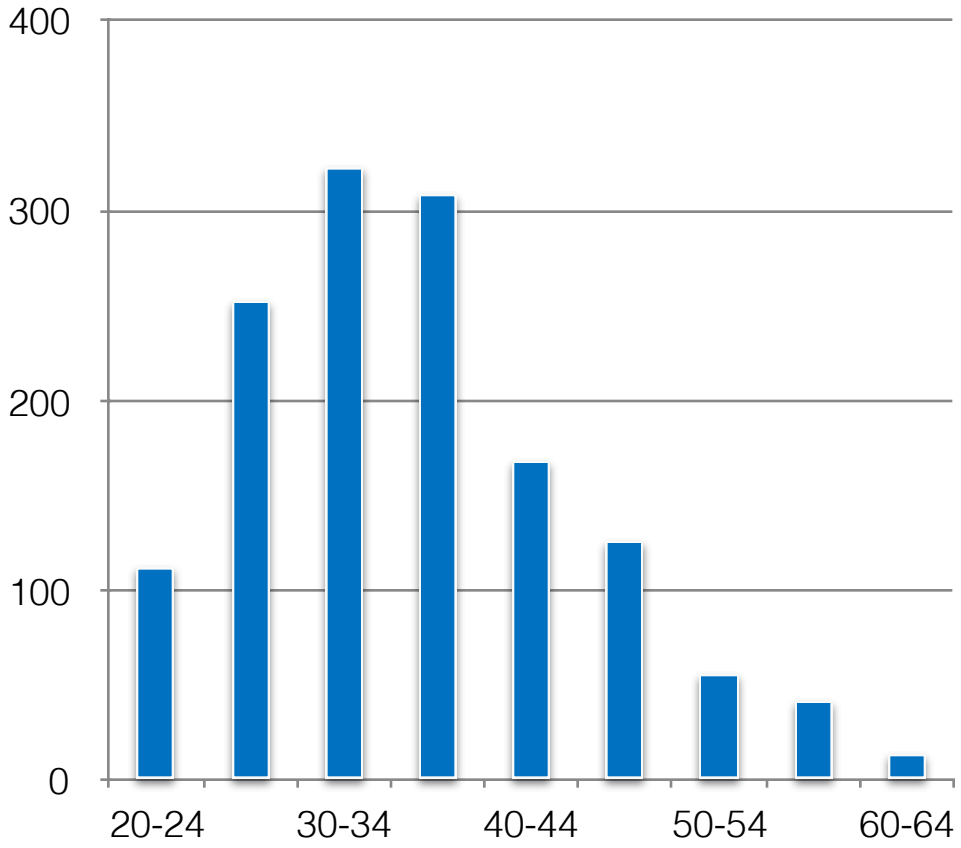
If the quantitative variable is the continuous variable, then the data must be grouped into classes (categories) before the table of frequencies can be formed.

The main steps in a process of grouping quantitative variable into classes are:

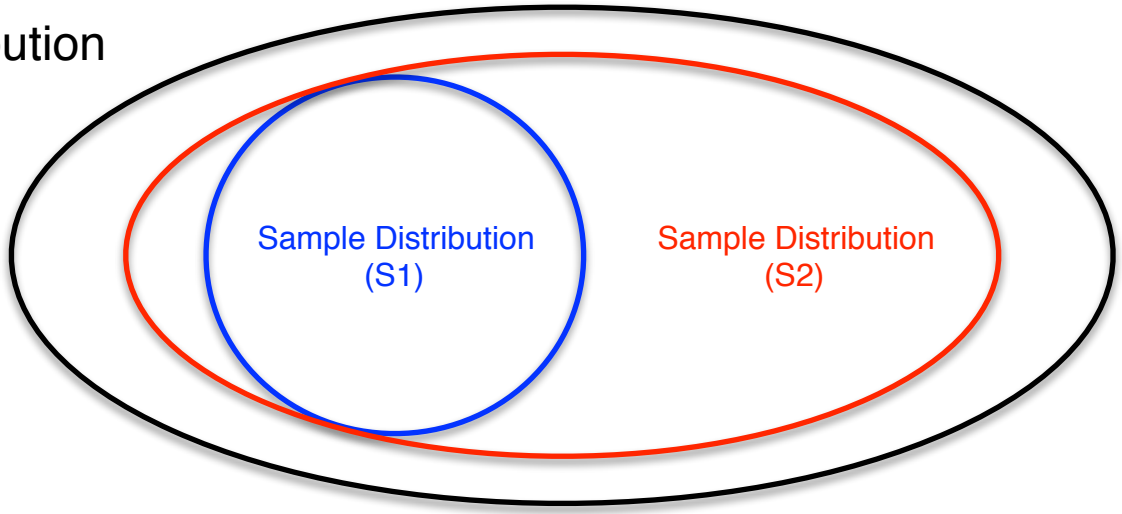
- Find the minimum and the maximum values variable have in the data set
- Choose intervals of equal length that cover the range between the minimum and the maximum without overlapping. These are called class intervals, and their end points are called class limits.
- Count the number of observations in the data that belongs to each class interval. The count in each class is the class frequency.
- Calculate the relative frequencies of each class by dividing the class frequency by the total number of observations in the data.

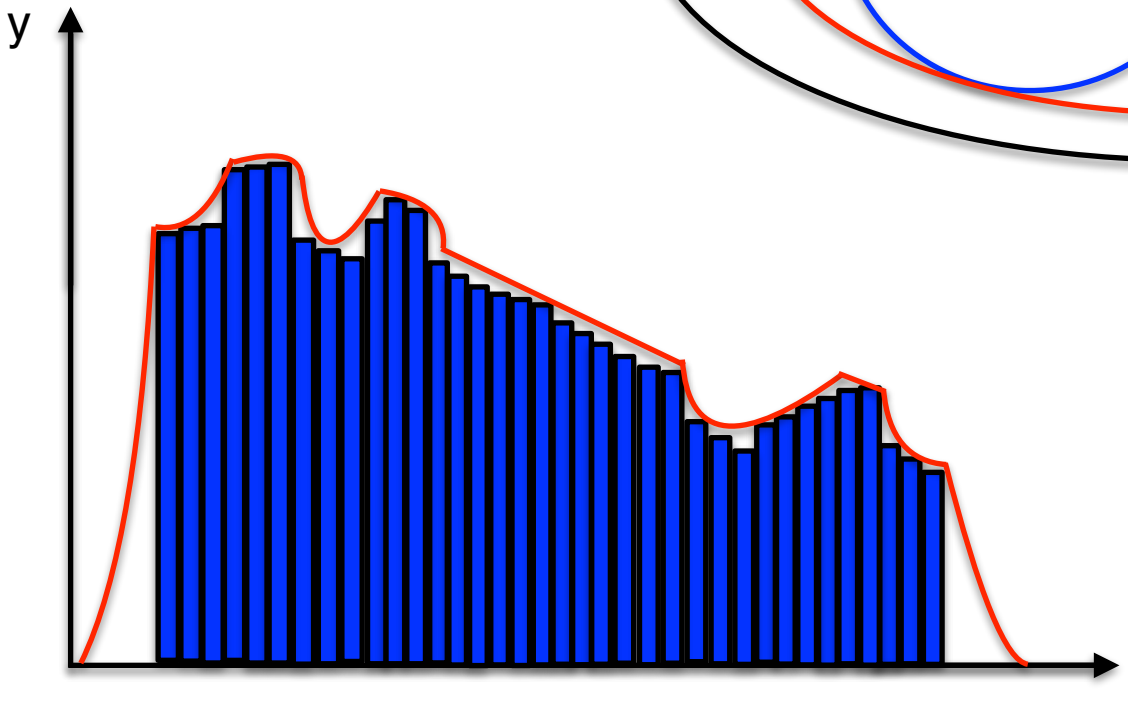
Describing Data by Tables and Graphs

Age Limit		Count
Lower	Upper	
20	24	112
25	29	252
30	34	322
35	39	308
40	44	168
45	49	126
50	54	56
55	59	42
60	64	14



Sample and Population Distributions





As the sample size increases and the intervals narrow down, the histogram of the sample distribution gradually becomes a curve.

Measures Of Central Tendency

Descriptive measures that indicate where the centre or the most typical value of the variable lies in collected set of measurements are called measures of central tendency. Measures of central tendency are often referred to as averages. The median and the mean apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data.

Measures Of Central Tendency: Mean

When people speak of taking an average, it is mean that they are most often referring to. The most commonly used measure of centre for quantitative variable is the sample mean.

The sample mean of the variable is the sum of observed values in a data divided by the number of observations.

If the sample size is n, then the mean of the variable x is

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Example 1: Salary (in Lakh) of 5 Data Scientists are: 7.6, 5.6, 9.3, 8.4, 6.1
Mean = 7.4

Example 2: Salary (in Lakh) of 6 Data Scientists are: 7.6, 5.6, 9.3, 8.4, 6.1, 20
Mean = 9.5

Measures Of Central Tendency: Median

The sample median of a quantitative variable is that value of the variable in a data set that divides the set of observed values in half.

- Arrange the observed values of variable in a data in increasing order.
- If the number of observation is odd, then the sample median is the observed value exactly in the middle of the ordered list.
- If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list.

Example 1: Salary (in Lakh) of 5 Data Scientists are: 5.6, 6.1, 7.6, 8.4, 9.3
Median = 7.6

Example 2: Salary (in Lakh) of 6 Data Scientists are: 5.6, 6.1, 7.6, 8.4, 9.3, 20
Median = 8.0

Measures Of Central Tendency: Mode

The sample mode of a qualitative or a discrete quantitative variable is that value of the variable which occurs with the greatest frequency in a data set.

- If the greatest frequency is 1 then the variable has no mode.
- If the greatest frequency is 2 or greater, then any value that occurs with that greatest frequency is called a sample mode of the variable.

Errors (in a page)	No Page
0	135
1	270
2	272
3	180
4	90
5	36
> 5	17

Mode

Which Measure to Use?

- The mode should be used when calculating measure of central tendency for the qualitative variable.
- When the variable is quantitative with symmetric distribution then the mean is proper measure of central tendency.
- In a case of quantitative variable with skewed distribution, the median is a good choice for the measure of central tendency.

Measures of Dispersion

- Measures of dispersion are used mostly only for quantitative variables.
- Another important aspect of a descriptive study of the variable is numerically measuring the extent of dispersion around the centre.
- Two data sets of the same variable may exhibit similar positions of centre but may be remarkably different with respect to dispersion.
- Two of the most frequently used measures of dispersion; the sample range and the sample standard deviation.

Measures of Dispersion: Range

The sample range of the variable is the difference between its maximum and minimum values in a data set: Range = Max – Min.

- The sample range of the variable is quite easy to compute.
- Only the largest and smallest values of the variable are considered; the other observed values are disregarded.
- The range cannot ever decrease, but can increase.

Example 1: Salary (in Lakh) of 5 Data Scientists are: 5.6, 6.1, 7.6, 8.4, 9.3
Range = 9.3 – 5.6 = 3.7

Example 2: Salary (in Lakh) of 6 Data Scientists are: 7.6, 5.6, 9.3, 8.4, 6.1, 20
Range = 20.0 – 5.6 = 14.6

Measures of Dispersion: Standard Deviation

The sample standard deviation is the most frequently used measure of dispersion, although it is not as easily understood as ranges. It can be considered as a kind of average of the deviations of observed values from the mean of the variable in question.

$$STD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Example 1: Salary (in Lakh) of 5 Data Scientists are: 5.6, 6.1, 7.6, 8.4, 9.3
Mean = ??, STD = ??

Measures of Dispersion: Standard Deviation

$$STD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Example 1: Salary (in Lakh) of 5 Data Scientists are: 5.6, 6.1, 7.6, 8.4, 9.3

Mean = (5.6 + 6.1 + 7.6 + 8.4 + 9.3) / 5 = 7.4
STD = Square root of (((5.6 - 7.4)^2 + (6.1 - 7.4)^2 + (7.6 - 7.4)^2 + (8.4 - 7.4)^2 + (9.3 - 7.4)^2) / 4)
= 1.55

References

Agresti, A. & Finlay, B., Statistical Methods for the Social Sciences, 3rd Edition. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., Introductory Statistical Analysis. Houghton Mifflin Company, 1974.

Johnson, R.A. & Bhattacharyya, G.K., Statistics: Principles and Methods, 2nd Edition. Wiley, 1992.

Weiss, N.A., Introductory Statistics. Addison Wesley, 1999.