1. Introduction

Welcome to the series of e-Learning module on Statistics.

At the end of this session, you will be able to:

- Explain what is a Histogram
- Explain the rules for making Histograms
- Explain different types of Histograms
- Explain what is a Frequency Polygon
- Explain what are the statistics derived from Histograms
- Explain what is a Box-and-Whisker Chart or Plot
- Explain how to prepare a Box-and-Whisker Plot
- Explain how to interpret a Box-and-Whisker Plot

In this module, we are going to coverthe concept of Histogram.

As you know, there are two types of natural scale graphs:

A) Time series graph
B) Frequency graphs

In frequency graphs, data expressed in terms of items or class intervals, and not in time.

There are 5 categories of frequency graphs. They are:

1. Histograms
2. Frequency Polygons
3. Frequency Curves
4. Ogives
5. Z-Charts

In today’s session, we are going to discuss two widely used types of frequency graphs, i.e. Histograms and Frequency Polygons.

Histogram provides visual representation of distribution of real-life data against variables other than time.

Let us take an example that shows the results of a final exam given to a hypothetical class of students.

A bar of a certain colour denotes each score range. This histogram if compared with those of classes from other years that received the same test from the same professor, then the conclusions drawn about intelligence changes among students and the improvement or decline of the professor's teaching ability.

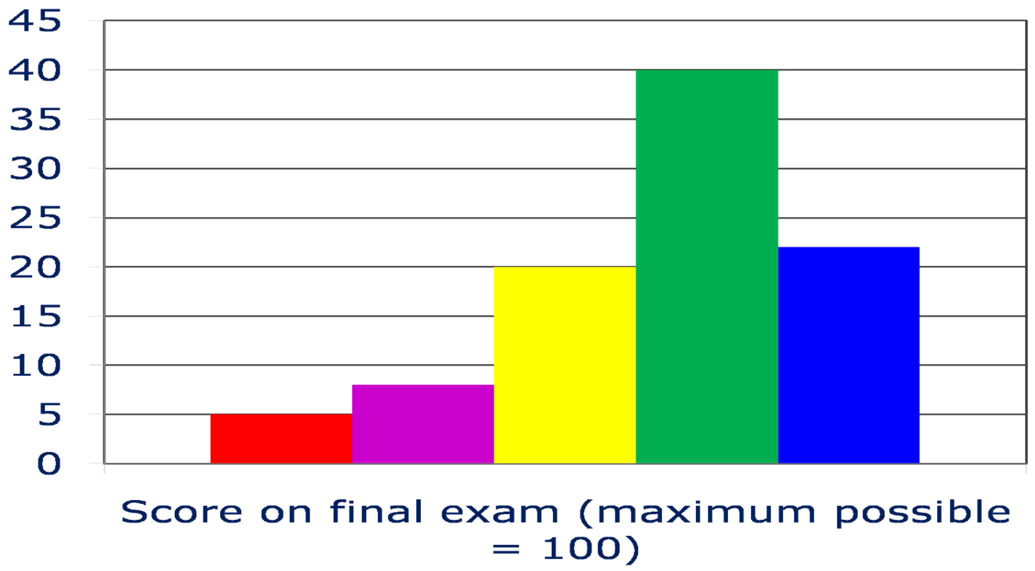


Figure 1

If this histogram compared with those of other classes in the same semester who had received the same final exam, but who had taken the course from different professors, one might draw conclusions about the relative competence of the professors.

Histograms are very useful for breaking out process data into regions or bins for determining frequencies of certain events or categories of data. Hence, they are considered as basic quality tools to analyze problems, causes and consequences.

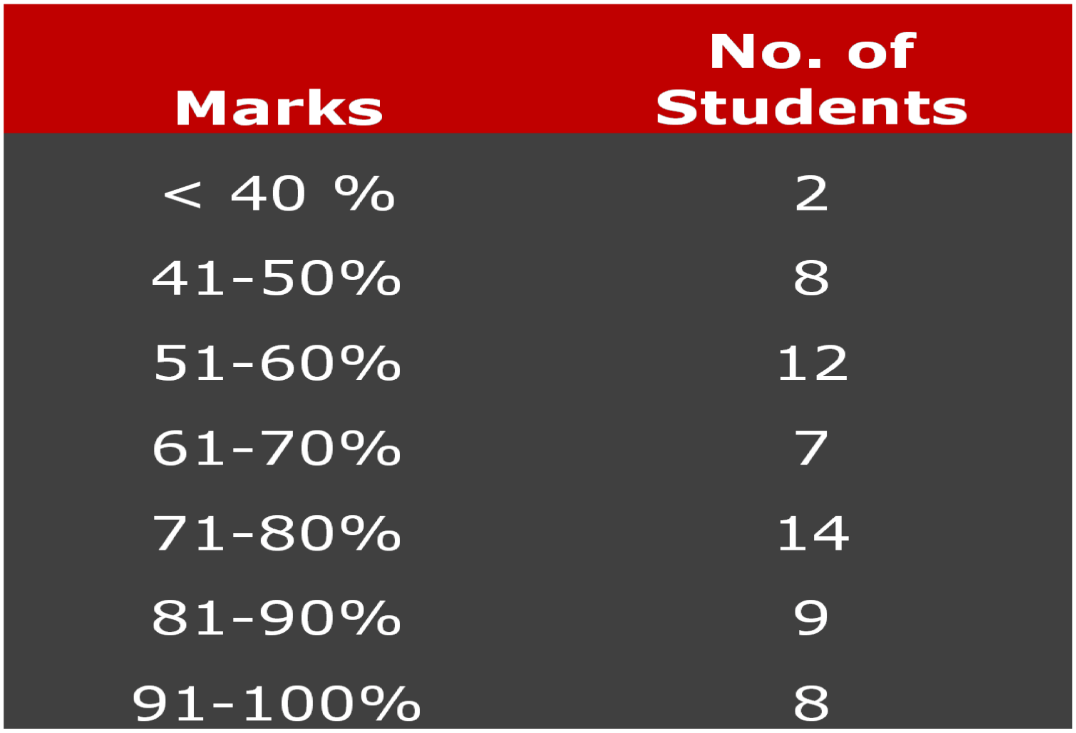
A histogram is a bar graph that shows how frequently data occur within certain ranges or intervals. The height of each bar gives the frequency in the respective interval.

In a histogram, the frequencies are shown in the form of rectangles, the base of which is the class interval. Another feature of this graph is that the rectangles are adjacent to each other without having any gap amongst them.

Items or class intervals are taken on the X-axis, whereas frequency is taken on the Y-axis. All the class intervals in a histogram will be equal.

Let us take an example to convert the distribution of percentage marks of the students into a histogram. The first column has marks of the students like <40%, 40-50%, 51-60% and so on and another column has the number of students like 2, 8, 12 and so on.

As you can see, the area of each rectangle is proportional to the frequency i.e. number of students. In addition, the total area of this figure comprising of 7 rectangles of varying sizes is equal to the total number of frequencies multiplied by the corresponding class intervals.



Marks	No. of Students
< 40 %	2
41-50%	8
51-60%	12
61-70%	7
71-80%	14
81-90%	9
91-100%	8

Figure 2

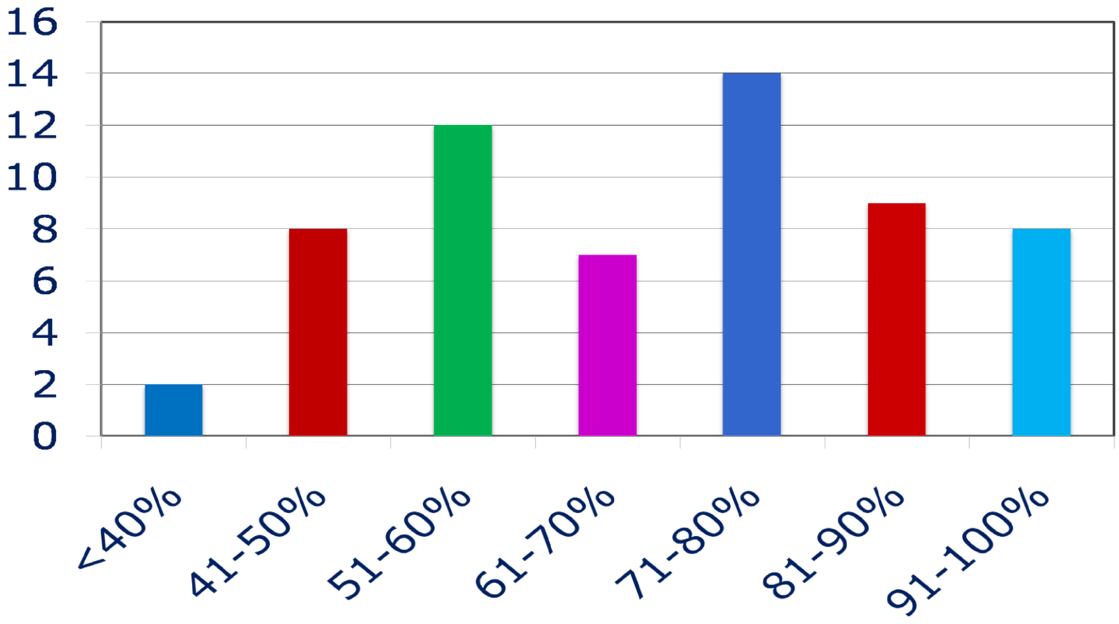


Figure 3

Histograms are widely used for quality control, especially in analysis of process effectiveness by comparing results to specification limits.

If you add the process specification limits to your Histogram, you can determine quickly whether the current process was able to produce "good" products.

Specification limits may take the form of length, weight, density and quantity of materials that has to be delivered. It can also take the form that is important for the product of a given process.

This histogram shows the defect density i.e. number of defects per 10,000 pieces of the product over a 6 months period.

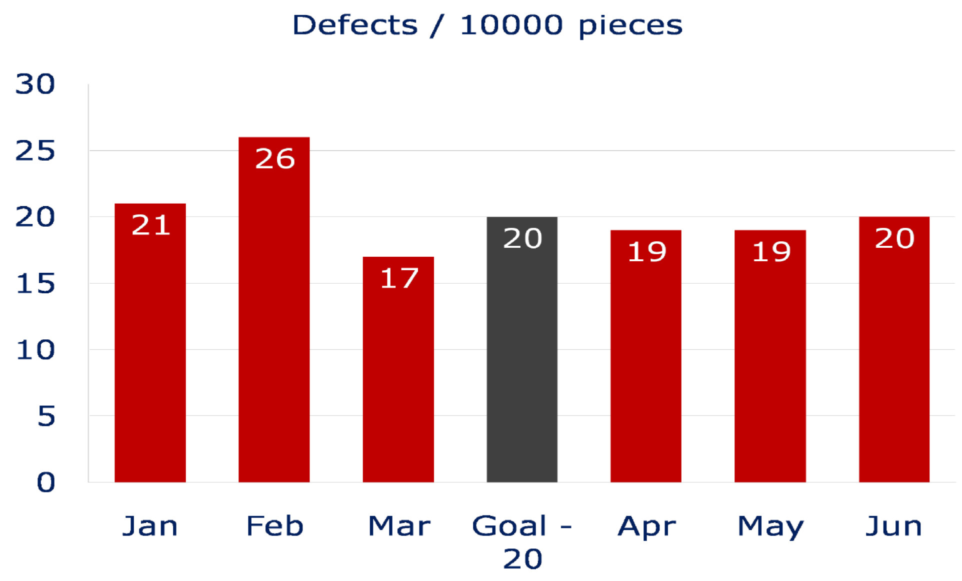


Figure 4

It tells us that more defects i.e. low quality happened during February and least defects i.e. high quality were delivered during March.

The quality during other months was close to the goal. Based on this, an investigation may be done to find out the reasons behind low quality in February.

Let us compare the Histogram with the line graph for the data having marks of the students on X-axis and number of students on Y-axis.

Apart from the distinctly different visual representation, there are two significant differences between Line Graphs and Histograms.

In a line graph, the vertical frequency lines are separate and unconnected to each other.

In Histogram, the area of each rectangle is proportional to the frequency whereas, the line graph does not allow us to do such measurements and comparisons.

2. Rules for making Histograms.

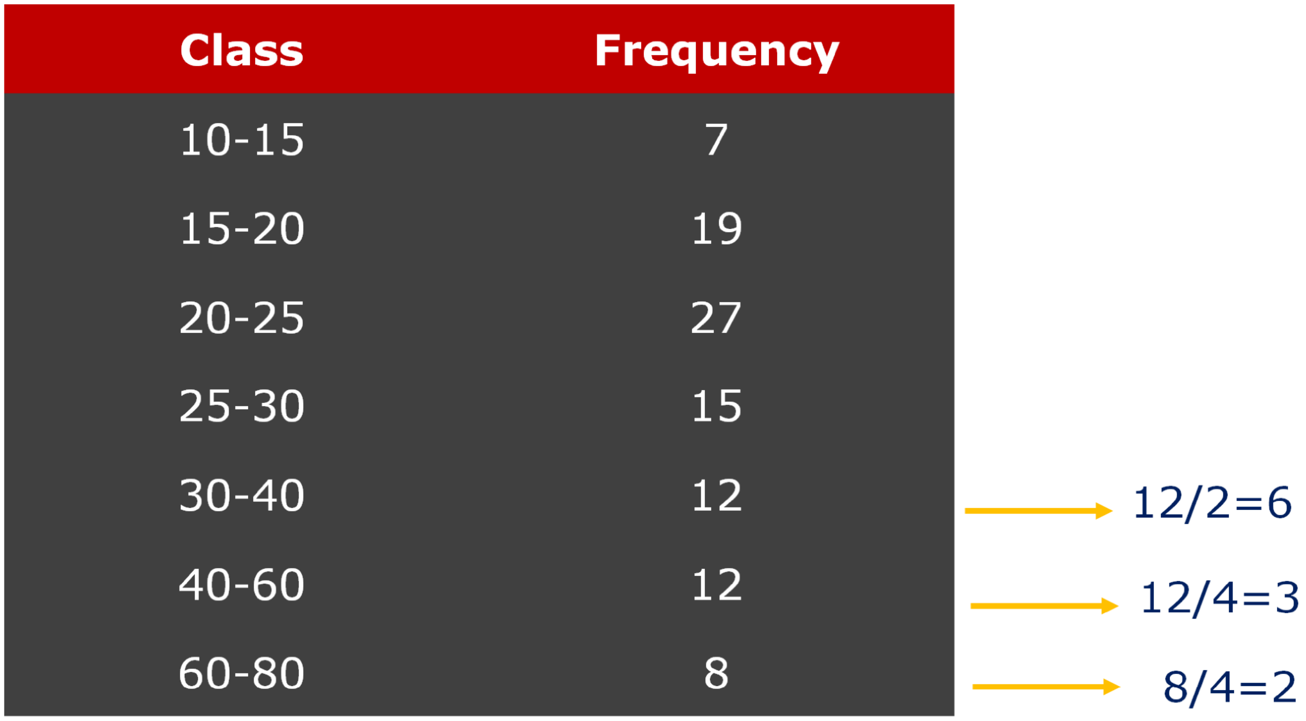
There are certain rules to be followed while making histograms.

Class intervals or items must be plotted on the X-axis and frequencies must be plotted on the Y-axis.

The series must be exclusive for all the frequency distribution graphs including Histograms. If it is in any other form, for example, inclusive, open-end or cumulative, it must be first converted into exclusive series before plotting the graph.

All the class intervals must be equal. In case it is not, we have to make it equal. This example graph has equal class intervals. However, it is not possible in real life and we often get data with unequal class intervals.

Now, let us take an example of unequal class interval. The table consists of two columns. The first column has the class interval like 10-15, 15-20, 20-25 and so on and another column has the frequency like 7, 19, 27 and so on.



Class	Frequency	
10-15	7	
15-20	19	
20-25	27	
25-30	15	
30-40	12	→ 12/2=6
40-60	12	→ 12/4=3
60-80	8	→ 8/4=2

Figure 5

Here the smallest class interval is of 5 units. Hence, we will convert all the intervals to this size. 30-40 class interval will be converted to 2 intervals of 5 and frequency will be divided by 2 i.e., 12/2=6. 40-60 and 60-80 class interval will be converted to 4 intervals and frequency will be divided by 4 i.e., 12/4=3 and 8/4=2.

The converted table with equal interval series will look like this.

The Histogram using the converted table with equal interval series will look like this.

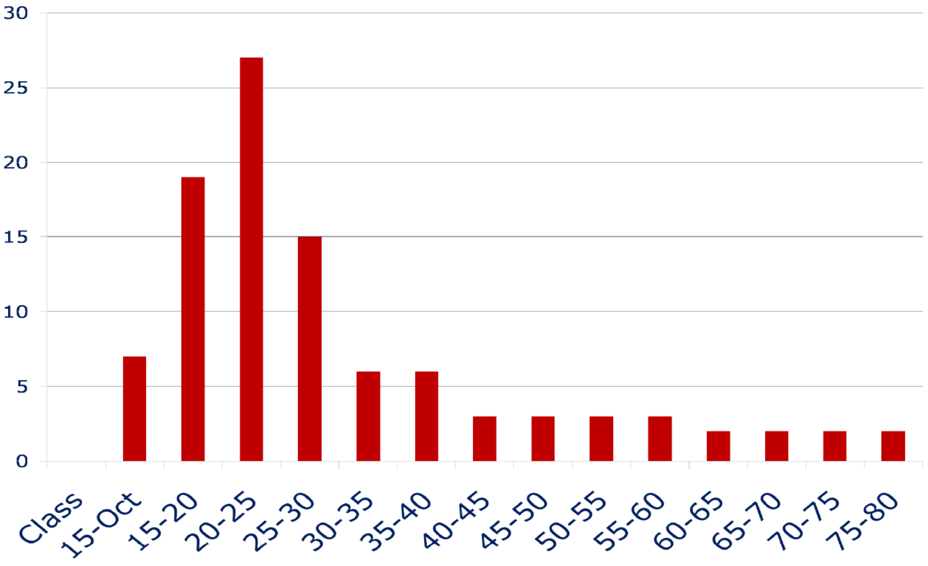


Figure 6

Histograms have evolved over a period to give us superior visual representation of data. In common language, they are also called as bar diagrams because the data is represented through rectangular bars.

Simple bar diagram shows one variable say frequency, for the same items called class frequency.

Now, let us take an example on simple bar diagram that shows the quantity of material produced from 3 factories. The table consists of two columns. The first column has the factories 1, 2 and 3 and another column has the output in tons like 1200, 1550 and 900.

Factories	Output (Tons)
1	1200
2	1550
3	900

Figure 7

This is how the simple bar diagram looks like.

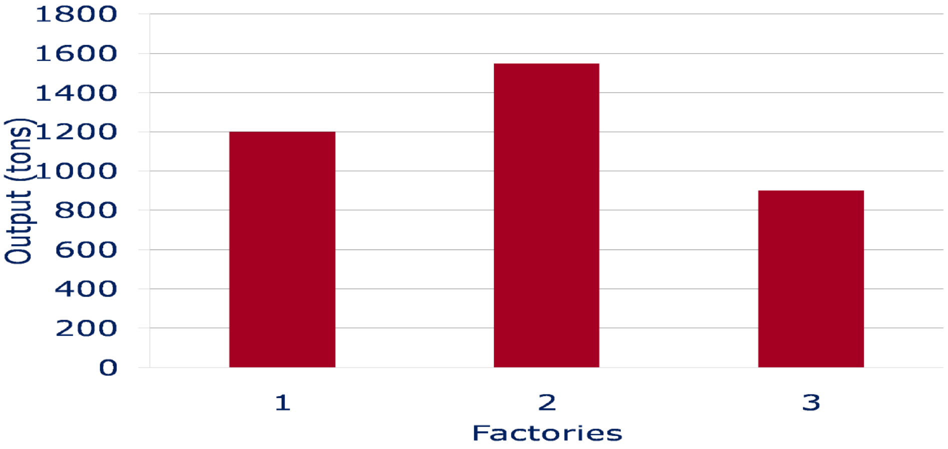


Figure 8

Multiple bar diagram shows two or more inter-related series of data against a common item say class frequency.

Now, let us take an example on multiple bar diagram that shows the export and import figures for each year. The table consists of three rows. The first row has years like 1, 2 and 3. The second row has export in lakhs like 550, 415 and 310. The third row has import in lakhs like 150, 150 and 210.

Year	1	2	3
Export (In Lakhs)	550	415	310
Import(In Lakhs)	150	150	210

Figure 9

This is how the multiple bar diagram looks like.

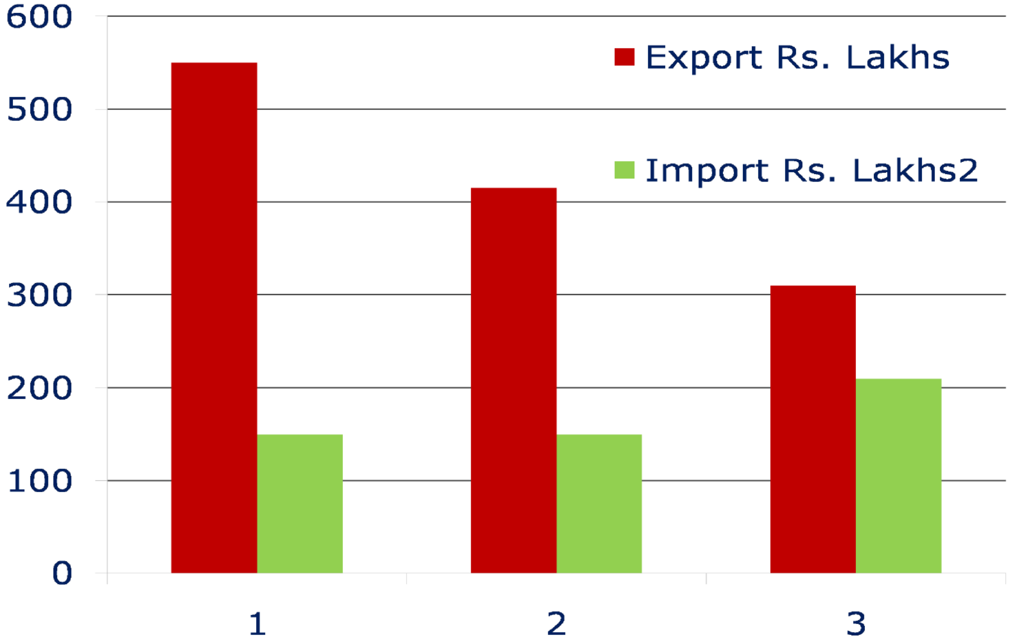
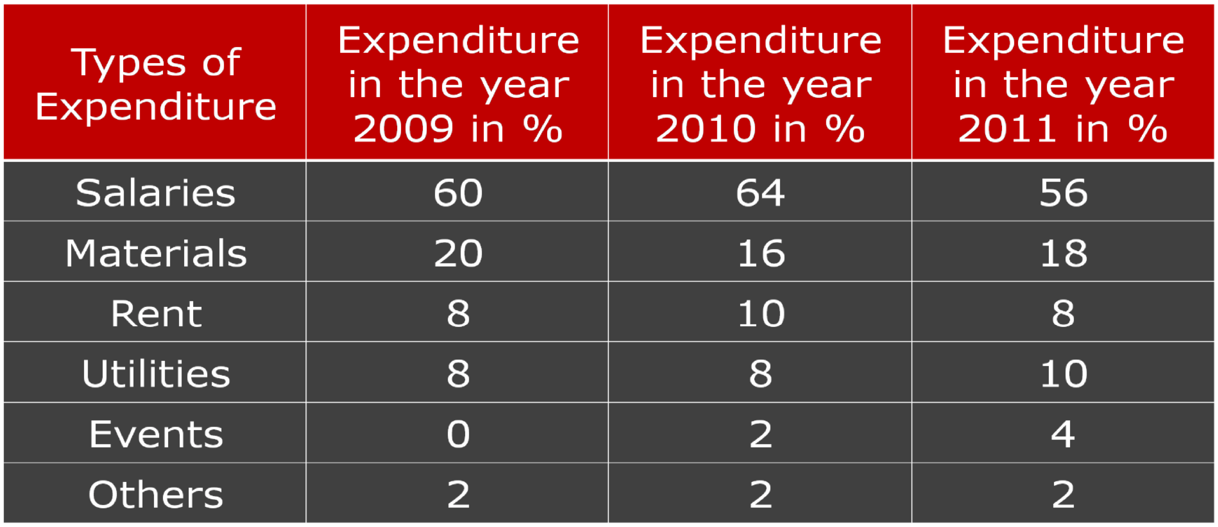


Figure 10

As the name implies, such diagram shows subdivisions of components in a single bar.

Now, let us take an example on sub-divided bar diagram that shows the different types of expenditure of an office over 3 years. The table consists of four columns. First column has types of expenditure like salaries, materials, rent and so on. Second column consists of the expenditure in the year 2009 in percentage like 60, 20, 8 and so on. Third column consists of the expenditure in the year 2010 in percentage like 64, 16 10 and so on. Fourth column consists of the expenditure in the year 2011 in percentage like 56, 18, 8 and so on.



Types of Expenditure	Expenditure in the year 2009 in %	Expenditure in the year 2010 in %	Expenditure in the year 2011 in %
Salaries	60	64	56
Materials	20	16	18
Rent	8	10	8
Utilities	8	8	10
Events	0	2	4
Others	2	2	2

Figure 11

This is how the sub-divided bar diagram looks like.

In addition, there are several other types of bar diagrams. Some of them are,

- Deviation bar diagrams
- Duo-directional bar diagrams
- Sliding bar diagrams
- Pyramid diagrams, etc.

3. Frequency Polygons

Now, we can discuss about frequency polygon. A polygon is a drawing consisting of several angles. A frequency polygon is created from a histogram when straight lines join the mid-points of the rectangles and the extremes are joined with the base.
As shown in the pictures, the histogram of students' distribution is converted to a frequency polygon.

As the histogram and frequency polygon are very similar graphs, you may wonder why it is necessary to have two similar graphs. It will be good for you to know their respective advantages.

Now, let us list the advantages of histogram.

1. Each rectangle shows distinctly separate class in the distribution.
2. The area of each rectangle in relation to all other rectangles shows the proportion of the total number of observations pertaining to that class.

Let us list the advantages of frequency polygon.

1. It is simpler to understand compared to the underlying histogram.
2. It shows a vivid outline of the data pattern.
3. As the number of classes and number of observations increase, the frequency polygon becomes smoother & makes it easier to interpret the data.

Histograms are used to calculate various statistics. Let us take a quick look at some of the most commonly used ones.

1. Mean is the average of all the values.
2. Minimum is the smallest value.
3. Maximum is the biggest value.
4. Standard Deviation is an expression of how widely spread the values are around the mean
5. Mean Class Width is the x-axis distance between the left and right edges of each bar in the histogram.
6. Number of Classes is the number of bars, including zero height bars in the histogram.
7. Skewness is zero, if the histogram is symmetrical. If the left hand tail is longer, skewness will be negative. If the right hand tail is longer, skewness will be positive.

Let us now take a look at Box-and-Whisker plots.

A box-and-whisker plot is a type of diagram depicting groups of numerical data through their five-number summaries : the sample minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and sample maximum.

As the name indicates, boxes are drawn to represent the data clusters (the quartiles and medians) and whiskers are drawn at either end of boxes to represent the minimum and maximum samples.

4. Box-and-Whisker Plots

Box-and-whisker plots are a descriptive statistics method, to show the concentration of data and the tilt of data (bias) in certain directions.

Box-and-whisker plots are used for understanding concentration patterns for a variety of purposes like region-wise grain production, compensation levels of employees, price points and so on.

Very often we come across the need to understand the concentration patterns of data points.

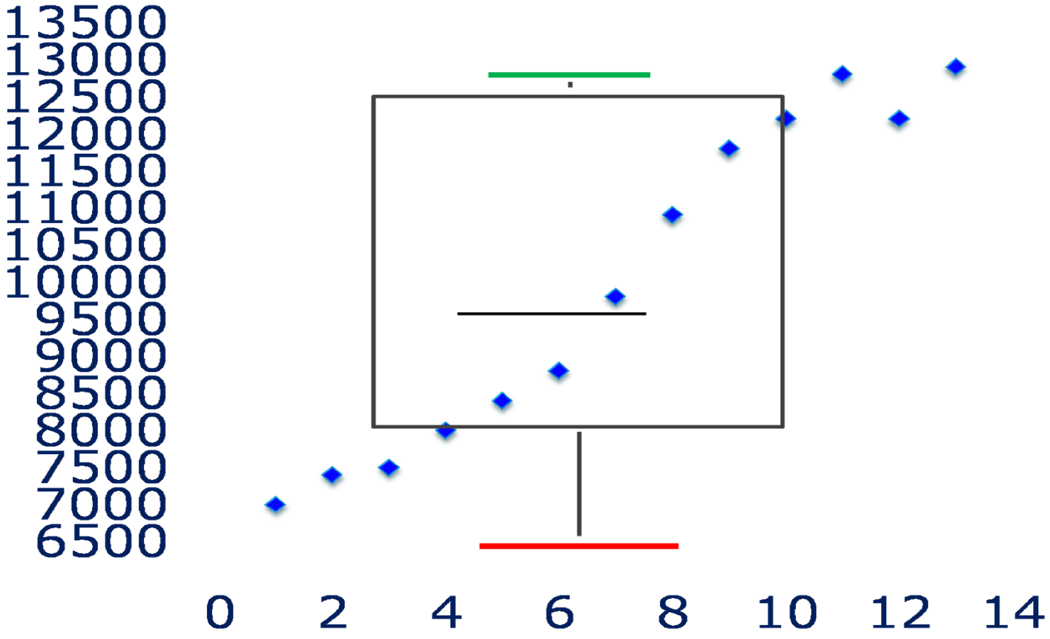


Figure 12

For example, a company is planning to do a compensation review. They need to know whether compensation for same type of employees is evenly distributed across a range, or whether there is a skew towards higher or lower side of the salary range.

A Box-and-Whisker plot is made of:

- One or more boxes, each representing a cluster of data
- Whiskers on either side of the boxes representing the minimum and maximum samples.

A typical box-and-whisker diagram will look like this:

In this diagram this is the box.

And these are the whiskers.

In a box-and-whisker plot, the concentration of data is divided into two quartiles based on medians.

The concentration is then visually depicted using boxes and line connectors. The line connectors are also called “whiskers”. This is how the name “Box-and-Whiskers” came into being.

In statistics, it is assumed that data points are clustered around some central value. The "box" in the plot contains, and thereby highlights, the middle half of these data points.
They are also known simply as “Box Plots”

Box plots are depicted using the five-member summaries of numerical data. These summaries are:

- The Smallest Observation
- Lower Quartile
- The Median
- Upper Quartile
- The Largest Observation

This is the smallest observation also known as the sample minimum.

This is the “lower quartile”, representing data points below the median.

This is the “Median” of data points.

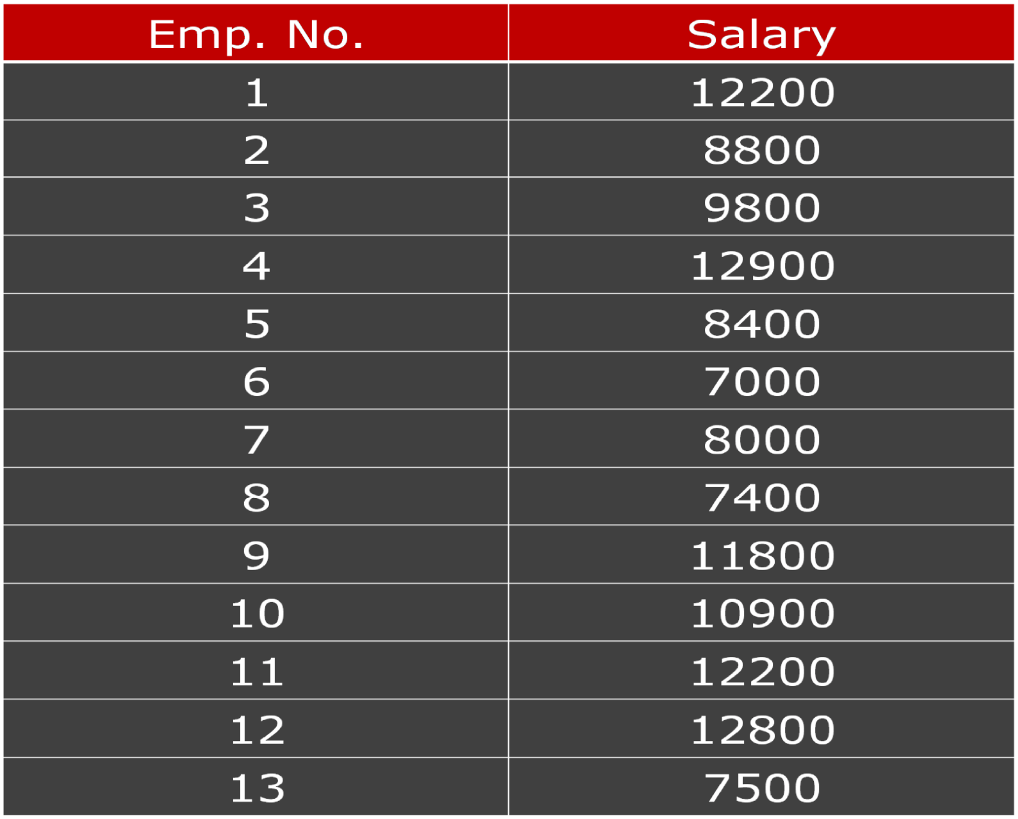
This is the “upper quartile”, the collection of data points above median.

This is the “Largest observation” also known as the “sample Maximum”.

5. Creating Box-and-Whisker Charts

There are several steps to be followed for creating box plots. Let us do a real-life exercise.

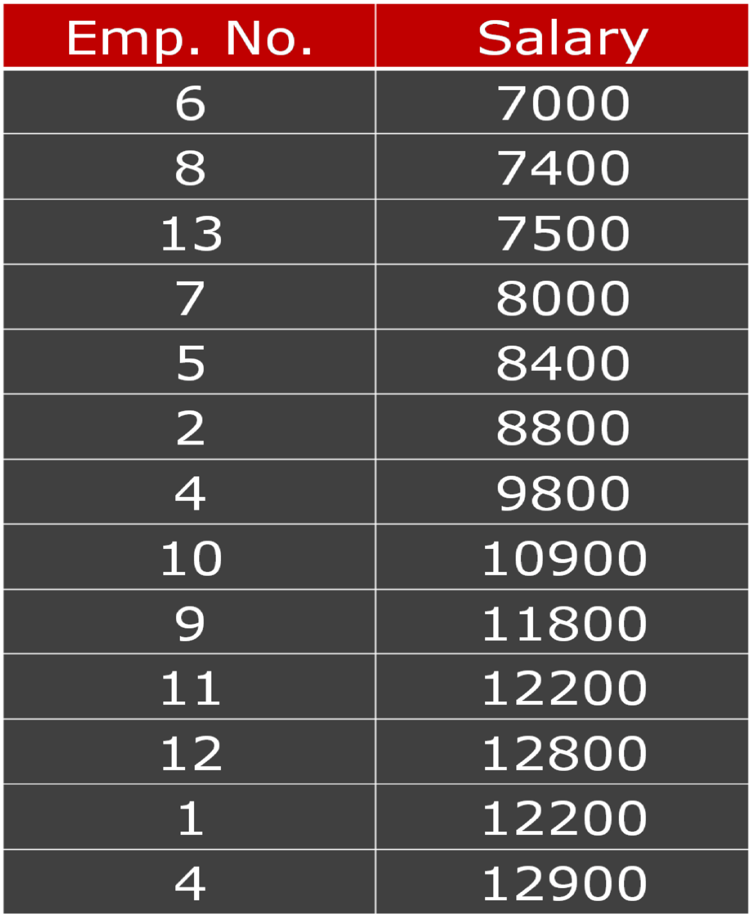
Let us take a look at the following data, depicting basic salaries of same class (grade / role) of employees of a company.



Emp. No.	Salary
1	12200
2	8800
3	9800
4	12900
5	8400
6	7000
7	8000
8	7400
9	11800
10	10900
11	12200
12	12800
13	7500

Figure 13

Now let us assume that the company is planning to do a salary revision exercise. Typical to all companies here also a range of salaries are being offered to the employees. In this case the range varies from Rs.7000 to Rs.12,900.
The company wants to know how the employees are distributed across the range i.e, whether they are evenly distributed with varying salaries or more concentrated in one range or towards lower or upper sides.



Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900

Figure 14

A box plot can help in this regard.

Since the salary data is in focus, first step is to sort the set of data by salaries.

Here is the sorted list in ascending order of salaries.

There are 13 values in this ordered list. The 7th will be the median. i.e, 9800.
Let us call this Q2.

The Median divides the data into two halves.

The Median for this half will be sum of midpoints (7500+8000) divided by 2 = 7750.

Let us call this Q1.

The Median for this half will be sum of midpoints (12200 + 12800) divided by 2 = 12500.

Let us call this Q3.

Let us now generate a scatter plot and mark the 3 medians (Q1, Q2, Q3) in it. The graph will now look like this.

Draw a box connecting Q1 and Q3, including all the data points which lie between them.

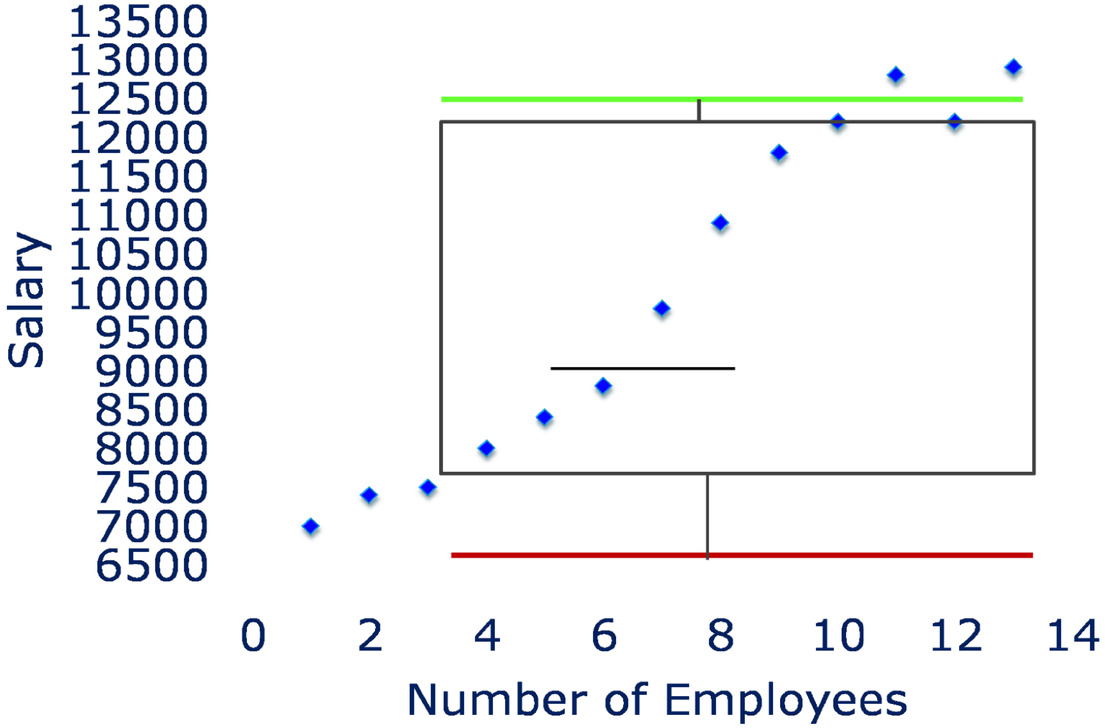


Figure 15

Now we can remove the top and bottom medians (Q1 and Q3) from this picture. The "Box" is ready.

Mark the smallest observation (7000) and highest observation (12900) in the chart.

Connect the smallest and highest observations to the box.
The "Box" and "Whiskers" are now done.

In this example, what does the box chart tell us?
The Plot tells us several things.
There is a larger concentration of data points in the upper quartile. This means that more employees are concentrated on the upper salary range.

There is a larger concentration of data points outside the box. This means quite a few employees are farther from the mid-point salary level".

Even distribution of salaries can be achieved by bringing up the salaries in the lowest quartile. This will also be the most economical approach for the company.

We have used a plotter chart in this example. However this is not mandatory, you can use any other type of chart. In fact it is not mandatory to show the data points within the box at all. All we need is to show the 5 summaries i.e, sample minimum, lower quartile, median, upper quartile and sample maximum.
In the example discussed earlier, we had prepared a vertical Box plot. However, boxes can be plotted vertically or horizontally as per your convenience.
If required, multiple boxes (each representing a unique data set related to same matter) can be represented in same graph, each with its own whiskers. See this example.

There is no colour code prescribed for Box plots. You can choose what works for you.
If required, multiple boxes (each representing a unique data set related to same matter) can be represented in same graph, each with its own whiskers. See this example.

Box-and-whisker plots are used for a variety of purposes in our daily lives.
Some examples are:

1. Policy making: To determine the most deserving segment to benefit from policies.
2. Financial decision making: To determine the most economical actions to achieve objectives.
3. Understanding of imbalances: For example: Social and Economic conditions across regions.

Here’s a summary of our learning in this session:

- Explain what is a Histogram
- Explain the rules for making Histograms
- Explain different types of Histograms
- Explain what is a Frequency Polygon
- Explain what are the statistics derived from Histograms
- Box-and-Whisker Plots
- Preparing Box-and-Whisker Plots (We also briefly discussed examples of using these diagrams for real-life purposes.)