1. Introduction to Statistics and Statistical methods

Welcome to the series on E-Learning module in Statistics and Statistical methods. In this module we are going to cover a basic yet important concept – Frequency Distribution. Frequency distribution forms the basis of many statistical analysis methods.
Our objective in today's session is to understand the concept of Frequency distribution (FD) and how useful it can be in our life. By the end of this session you will be able to understand and appreciate the following concepts:

- What is Frequency distribution
- What is the need for Frequency distribution in statistical analysis
- Steps involved in preparing a Frequency Distribution table
- Uses of Frequency Distribution

Statistical analysis involves not just numbers or graphs but also facilitates the process of simplifying data or numerical expressions to arrive at the final conclusion be it for National economy, budget, census, used by corporate houses for business forecast and planning, educational institutions, sports, tourism, etc. Statistics and statistical analysis are integral part of every nation's Planning Commission and Economic growth projections.

Statistical analysis is a mechanism which transforms a disorganized, meaningless data into organized, structured, meaningful and useful data which can be used effectively in planning and development activities. Frequency distribution helps in transforming the raw data into organized and important data for statistical analysis and inferences.

2. Frequency Distribution

Frequency distribution is a process of arranging data into sequences and groups or separating them into different parts according to certain chosen common characteristics. It shows us a summarized grouping of data & presents the data in a form that clearly describes a thematic representation of the data based on the selected characteristics.
For example:
•A Consumer goods company would like to know the sales performance of a certain product (say TV or Refrigerator or Washing machine) in a region for the current year
•Central Government would like to know the how different states are performing on the agricultural yield or milk production
•An academic institution would like to know the academic performance of their students in various exams
•A company would like to know it has performed year on year on parameters like revenue, profitability vis-à-vis its competitors

This way of compressing data brings more meaning to the data for better interpretation and effective decision making.

In order to make this data more meaningful it needs to be organized, classified and presented in a more meaningful format. In statistics we classify the data using various statistical series.

A series in statistics is defined as things or attributes arranged in some logical manner.

Statistical series are classified into –Individual series, Discrete series and Continuous series
Let us start with an example. The table shows marks obtained by 50 students in Statistics in their exam. Data in the above form is called 'Raw Data'. It is a form of data where all the values are shown separately and is called an Individual series.

Figure 1

Marks obtained by 50 students in Statistics									
42	60	51	42	57	49	37	22	42	48
45	60	47	63	53	42	38	41	25	45
43	46	39	45	60	36	40	50	52	30
64	58	61	46	42	41	32	17	38	28
31	46	46	49	29	45	45	53	52	50

A Raw Data is typically disorganized, of less use and does not give any valuable information since there is no significance, relationship or correlation of the data.

Typically when we think of marks of students, we would like to know what is the minimum and maximum marks, how many passed and how many failed, How many got distinction, what is the class average.

In order to make this data more meaningful it needs to be organized, classified and presented in a more meaningful format. Let us see how we can do this.

As a first step let us use a simple mathematical sorting technique to arrange the data in Ascending Order- where the data is arranged serially staring from a small value to a big value (smallest to largest). We can also arrange in Descending order where the data is arranged serially from a larger value to a small value (largest to smallest).
The process of arranging data in a certain order is called 'Arraying of data'.

Figure 2

Marks obtained by 50 students in Statistics									
17	22	25	28	29	30	31	32	36	37
38	38	39	40	41	41	42	42	42	42
42	43	45	45	45	45	45	46	46	46
46	47	48	49	49	50	50	51	52	52
53	53	57	58	60	60	60	61	63	64

In this example we have sorted the data in ascending order. From the sorted data we can now easily find the lowest (17) and highest (64) marks and clearly distinguishable. We have now improved the representation of data however the number of data points still remains 50.

Is there a way to reduce the number of data points and make the data set compact and more presentable? In the next step we will learn how to condense the data using a technique called frequency counting.

As a second step we will now rearrange the data to bring out meaningful information from the data.

- We create a 2 column table to represent the data.
- First column represents all the possible values of the variable (marks in this case)
- Second column represents a 'vertical bar' called Tally bar that is put against the number (value of the variable)
- We use 1 vertical bar for every occurrence of the number in the series.
- For the 5th occurrence of a number, we use a cross mark from right to left on the 4 vertical bars – see example of number 42 and 45 (circled in red)

Figure 3

Marks	Tally bar
17	|
22	|
25	|
28	|
29	|
30	|
31	|
32	|

Marks	Tally bar
36	|
37	|
38	||
39	|
40	|
41	||
42	𝍸
43	|

Marks	Tally bar
45	𝍸
46	||||
47	|
48	|
49	||
50	||
51	|
52	||

Marks	Tally bar
53	||
57	|
58	|
60	|||
61	|
63	|
64	|

The technique of putting a cross Tally mark at every 5th repetition of a data helps in counting the number of occurrences of the value in the end.
In the absence of a cross Tally mark, we will get continuous vertical bars (||||||||||...) which can be confusing to count, leading to mistakes in counting, data representation and data analysis.
Now we have achieved the objective of partially condensing the data from 50 data points to 31 data points using a simple Tally bar concept.

In the Third step, we introduce a new column 'Frequency' in the table called Frequency against each variable. This column contains a simple mathematical value of each variable representing how many times each number occurs in the data set (Frequency).

Figure 4

Marks	Tally bar	Frequency	Marks	Tally bar	Frequency	Marks	Tally bar	Frequency	Marks	Tally bar	Frequency
17	|	1	36	|	1	45	|卌	5	53	||	2
22	|	1	37	|	1	46	||||	4	57	|	1
25	|	1	38	||	2	47	|	1	58	|	1
28	|	1	39	|	1	48	|	1	60	|||	3
29	|	1	40	|	1	49	||	2	61	|	1
30	|	1	41	||	2	50	||	2	63	|	1
31	|	1	42	|卌	5	51	|	1	64	|	1
32	|	1	43	|	1	52	||	2			

This type of data representation is called Discreet or Ungrouped Frequency distribution where Marks are called the Variables (X) and number of students obtaining the marks as Frequency (f).
At times the identity of the units (students in our example) about whom a particular information is collected (marks in our example) is not relevant, nor is the order in which the observations occur, then the first real step of condensation consists in classifying the data into classes by dividing the entire range into suitable number of groups.

In this step we have used a simple technique to group the Marks column I nto data points in a range - each of interval 5 starting from the lowest value, without using any formula. The groups into which the values are classified is called “Class Intervals” – 17-21 or 22-26 is called Class Interval

Figure 5

Marks (x)	Frequency (f)
17-21	1
22-26	2
27-31	4
32-36	2
37-41	7
42-46	15
47-51	7
52-56	4
57-61	6
62-66	2

Class Interval Class Intervals are obtained by subtracting the largest and the smallest value of the data and dividing it by the needed magnitude of the class.

In the above data the largest value is 64 and the smallest value is 17 therefore the class interval will be for a magnitude of 5, **64-17 = 47/5 = 9.4.**

Thus the entire data can be represented in **10 groups.**

Length of the class interval is called “Magnitude of Classes” - 5 is the Magnitude of the classes in this example.

Class Intervals can also be obtained by subtracting the largest and the smallest value of the data and dividing it by the needed magnitude of the class.

For eg: in the given data the Largest value is 64 and the smallest value is 17 therefore the class interval will be for a magnitude of 5, 64-17= 47/5=9.4. Thus the entire data can be grouped in 10 groups

It can be grouped into magnitude of 10 then we need to calculate as follows 64-17 = 47/10 = 4.7 thus the entire data can be grouped into 5 groups or classes.

Figure 6

Marks (x)	Frequency (f)
17-26	3
27-36	6
37-46	18
47-56	15
57-66	8

The entire data can be grouped into 5 groups.

As you can see the number of data points has further reduced from 31 to 10 or 5 in the 2 tables.

3. Class Intervals & Class Limits

Groups into which the values of the variables are classified are known as classes or class intervals
The two values specifying the class are called the class limits. The larger value is called the upper class limit and the smaller value is called the lower class limit
In the above example , Class interval = 17- 21 , Magnitude = 5 ,Lower class limit = 17 , Upper class limit = 21

Figure 7

Marks (x)	Frequency (f)
17-21	1
22-26	2
27-31	4
32-36	2
37-41	7
42-46	15
47-51	7
52-56	4
57-61	6
62-66	2

Now we can see a more meaningful distribution of data showing how many students have scored marks in a certain range. It is clear from the table that maximum students (15 in this case) have scored marks in the range 42-46.

Having arrived at a Frequency distribution table, we will introduce a concept called Continuous series and 2 most frequently used types of continuous series called Inclusive Series and Exclusive series.

In the above table, if you notice, the lower limit of the second group (22) is different from upper limit of the previous group (21). We count the frequency of 21 in the first group. Such a series is called Inclusive Continuous series.
While dealing with a continuous variable it is not desirable to present the data into a grouped frequency distribution of this type because this classification does not take into consideration all the possible values (integral and fraction) in a specified range. For example, what happens if the marks is 21.5 – does

it get counted in first range or second range. In such situations we form a continuous class intervals without any gaps in data series.

Figure 8

Inclusive Continuous series.

Marks (x)	Frequency (f)
17-21	1
22-26	2
27-31	4
32-36	2
37-41	7
42-46	15
47-51	7
52-56	4
57-61	6
62-66	2

In the above table, the lower limit of the second group (22) is different from upper limit of the previous group (21). We count the frequency of 21 in the first group. Such a series is called **Inclusive Continuous series**.

Figure 9

Exclusive Continuous Series

Marks (x)	Frequency (f)
17-21	1
21-26	2
26-31	3

31-36	**2**
36-41	**6**
41-46	**13**
46-51	**10**
51-56	**5**
56-61	**5**
61-66	**3**

In the above table, the upper limit of first group (21) and lower limit of second group (21) are same but frequency counting of upper limit is included in the second group and not first group. Such a series is called Exclusive Continuous series.

Any value below 21 gets counted in first group and any value greater than or equal to 21 gets counted in the second group. Such a series is called Exclusive Continuous Series.
You can notice that the frequency distribution pattern changes depending on whether we choose Inclusive or Exclusive series method.

We have used graphical representation of the Inclusive and Exclusive table to have better representation of the data with Class intervals in X-axis and the frequency in Y-axis.
In Inclusive graph, there is a peak of the data group 42-46 with data frequency more than 10 where as in the Exclusive graph, there are 2 peaks of data more than 10 points for the range 41-45 and 46-51.

Figure 10

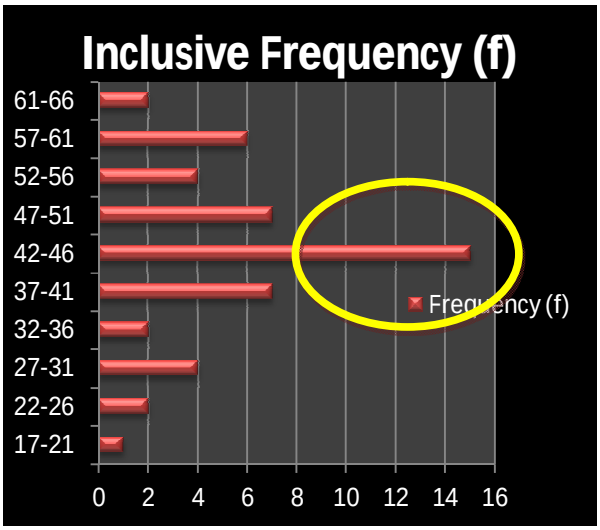
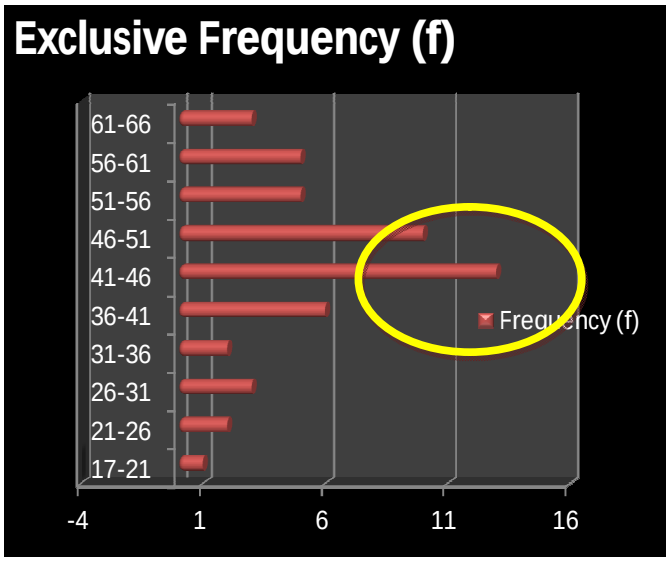


Figure 11



4. Cumulative Frequency Distribution

A frequency distribution simply tells us how frequently a particular value of the variable is occurring. However if we want to know the total number of observations getting a value less than or more than a particular value of the variable this frequency table fails to furnish the information as such.

This information can be obtained very conveniently from a 'cumulative frequency distribution table, which is a modification of the given frequency distribution and is obtained on successively adding the frequencies of the values of the variable according to a certain law.

The frequencies so obtained are called as 'cumulative frequencies" abbreviated as C.F. The law used are of 'less than' and 'more than' type giving rise to a 'less than cumulative frequency distribution' and ' a more than cumulative frequency distribution table'.

Figure 12

"Less Than" CF

Marks (x)	Frequency (f)	Less than C.F
17-21	1	1
21-26	2	1 +2 = 3
26-31	3	3 + 3 = 6
31-36	2	6 + 2 = 8
36-41	6	8 + 6 = 14
41-46	13	14 + 13 = 27
46-51	10	27 + 10 = 37
51-56	5	37 + 5 = 42
56-61	5	42 + 5 = 47
61-66	3	47 + 3 = 50

Marks (X)	Cumulative Frequency
< 21	1
< 26	3
< 31	6
< 36	8
< 41	14
< 46	27
< 51	37
< 56	42
< 61	47
< 66	50

We shall explain the construction of a "Less Than" distribution for the same example
The first column consists of the marks range (X), second column consists of frequency (f) and we introduce a third column which shows the "Less than" Cumulative frequency for the given sample. As you can see the Cumulative Frequency for the first range 17-21 is 1, which is obvious since there is only

1 data point for this range (value 17). The CF for the second range 21-26 is computed as the frequency for the range 21-26 + the frequency of the previous range 17-21. The Cumulative frequency for this range < 26 is 3 since there are 3 values that are less than 26. Proceeding in similar manner, we can populate the CF for all the ranges until the last range. The resultant CF is shown in a new table with the range values represented as "Less than X".

Figure 13

"More Than" CF

Marks (x)	Frequency (f)	More than C.f.
17-21	1	49+1 = 50
21-26	2	47+2 = 49
26-31	3	44+3 = 47
31-36	2	42+2 = 44
36-41	6	36+6 = 42
41-46	13	23+13 = 36
46-51	10	13+10 = 23
51-56	5	8+5 = 13
56-61	5	3+5 = 8
61-66	3	3

Marks (X)	Frequency
> 17	50
> 21	49
> 26	47
> 31	44
> 36	42
> 41	36
> 46	23
> 51	13
> 56	8
> 61	3

We shall next explain the construction of a "More Than" distribution for the same example
The first and second column remains the same as in the "Less than" scenario. The third column will show the CF for "More than" scenario. In a More than scenario, we start from the highest range. The Cumulative Frequency for the range 61-66 is the same as its Frequency. The CF for the range 56-61 is the sum of Frequency of this range (5) and the CF of the range 61-66 (3), which is equal to 8. The CF for the next lower range 51-56 is the Sum of the CF of the range 56-61 (8) and Frequency of the range 51-56 (5) – which is equal to 13. Proceeding in similar manner, we can populate the CF for all the ranges. The resultant CF is shown in a new table with the range values represented as "More than X".

Similarly we can also obtain a distribution table using Logical grouping of the data points depending on what kind of logic is needed for the business use. In the above example we have used the standard Grading system used by the University to rank the students – for ex Pass, Fail, Distinction, First class, Second Class, Third Class.

Figure 14

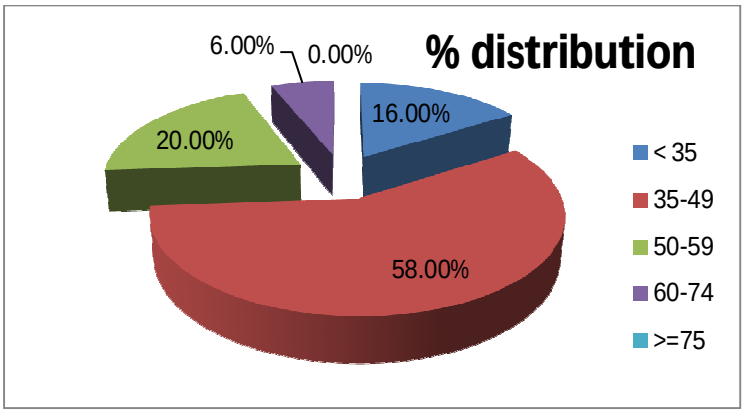
Marks (x)	Frequency (f)	% distribution
< 35 (fail)	8	16.00%
35-49 (3rd)	29	58.00%
50-59 (2nd)	10	20.00%
60-74 (1st)	3	6.00%
>=75 (Distinction	0	0.00%

Here we have classified the data range into more logical groups in order to define the Status of the exams and the Academic class obtained by the students. Here we use a simple logic for grading the students and classifying them as per the University Graduation Class or ranking.

- Marks < 35 is considered as Fail.
- Marks between 35 and 50 is considered to be 3rd class
- Marks between 50 and 59 is considered to be 2nd class
- Marks between 60 and 74 is considered to be 1st class
- Marks >= 75 is Distinction

From the above you can see that only 6% of the candidates have secured 1st class and none were able to secure Distinction. 16% of the students failed while 84% of the students passed in different classes. We can represent the above data as a Pie-chart – where each block of the chart represent a certain data point – we have used the percentage distribution of the students for the different Ranking categories.

Figure 15



5. Summary

- What is Frequency Distribution
- How to change a raw data to a FD table
- Steps involved in creating a FD table
  - Arraying, Condensing, Classifying, Grouping, Representing and Series representation
- Types of Frequency Distribution series
  - Inclusive and Exclusive
  - ‘Less than’ and ‘More than’ distribution