Learning Outcome

By the end of this lecture, you will be able to

- Reproduce a Simple Frequency Table (SFT)
- Reproduce a Grouped Frequency Table (GFT)

Introduction

Consider Ahmad, a Quality Engineer working in a manufacturing company. He is responsible for monitoring and improving the quality of the production process. He collects huge data from various production processes. However, analyzing huge data can be a complex and cumbersome task for him.

Summarization of data is essential for effective data analysis. This helps convert huge amount of data into simple tables that are easy to analyze and interpret. Frequency distribution table, such as given here, is one of the popular methods to summarize the data.

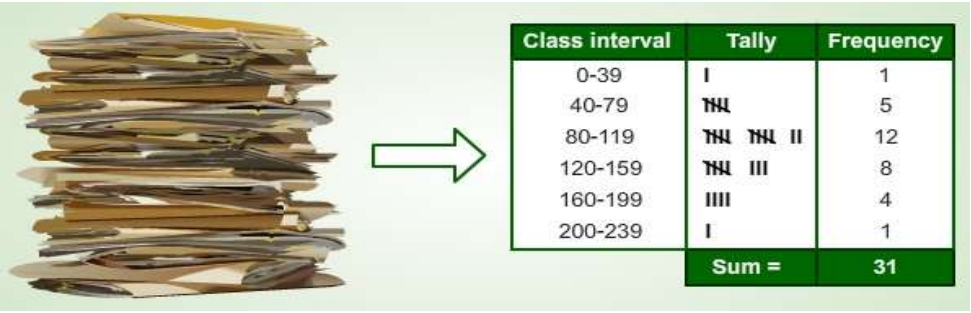
Sample frequency table

Number	Tally	Frequency
1	IIII	4
2	𝍸	5
3	III	3
4	III	3
5	II	2
6	III	3

In this lecture, you will learn to reproduce a Simple Frequency Table and a Grouped Frequency Table.

Frequency Distribution

A *frequency distribution* is a list or table used to summarize a large set of data.



Class interval	Tally	Frequency
0-39	I	1
40-79	𝍸	5
80-119	𝍸 𝍸 II	12
120-159	𝍸 III	8
160-199	IIII	4
200-239	I	1
	Sum =	31

Frequency table

It is often useful to classify the data into classes or categories and to determine the number of individuals belonging to each class, called the *class frequency*.

A tabular arrangement of data, by classes together with the corresponding frequencies, is called a *frequency distribution* or simply a *frequency table*.

Advantages of a Frequency Distribution

A frequency distribution is a way to summarize and *condense* the raw data in a limited number of *classes* (or categories). It shows the distribution of the data over the different classes for a pictorial interpretation of the data.

Some Definitions

Let's take a look at some definitions of a few terms of frequency table. Here is a sample frequency table.

Class	**Frequency**	**Cumulative frequency**	**Relative frequency**	**Cumulative relative frequency**
0.8 – 1.0	1	1	0.05	0.05
1.0 – 1.2	6	7	0.30	0.35
1.2 – 1.4	4	11	0.20	0.55
1.4 – 1. 6	6	17	0.30	0.85
1.6 – 1. 8	3	20	0.15	1.00
Total	20		1.00	

Class Width	The difference between the upper and lower class boundaries.
Frequency	The number of observations in a class or corresponding to a category.
Relative frequency	The ratio of the class frequency to the sample size. It is used to compare two similar frequency distributions with different sample sizes. $rf = \frac{Freqency\ of\ i^{th}\ class}{Total\ number\ of\ observations} = \frac{f_i}{\sum_{i=1}^{k} f_i} = \frac{f_i}{n}$ 'k' in the formula is the number of classes.
Cumulative frequency	The total frequency of all values less than or equal to the upper class boundary.
Relative cumulative frequency	The cumulative frequency divided by the total number of observations.

Frequency Distribution for Discrete or Categorized Data

If the variable of interest is discrete, (that has a small number of values) or categorical then the frequency distribution is, simply, constructed by counting the number of times each possible value (or category) occurs in the data set.

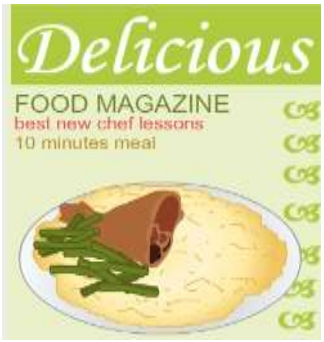
Blood groups

Frequency of blood groups in a population

Blood group	Frequency
O+	38
O –	9
A+	32
A –	6
B+	9
B –	2
AB+	3
AB –	1

Simple Frequency Table (SFT) for Discrete Data>Example

A restaurant-business magazine surveyed 16 people to determine how many times they had meals outside the house per week in 2008.



Food magazine

Here is the observed data:

3	0	0	1
1	2	2	0
0	2	1	0
2	3	4	2

Here are the steps to construct a frequency distribution and a relative frequency distribution for given data:

Step 1	List the possible values. The possible values listed in order of the discrete variable are 0,1,2,3 & 4
Step 2	Count the number of occurrences for each value.
Step 3	Determine the relative frequencies

International trips	**Frequency** f_i	**Relative frequency** $r.f$
0	5	0.3125
1	3	0.1875
2	5	0.3125
3	2	0.1250
4	1	0.0625
Total	**16**	**1.00**

The meaning of relative frequency (*r.f*):

0.3125: Just over 31% of those persons did NOT eat at all outside the house per week.

0.3125: Just over 31% of those persons ate outside the house twice per week.

Frequency Distribution for Continuous Data

If the variable of interest is continuous or discrete and has many possible outcomes, then it may not be possible to list all the values along with their corresponding frequencies. Instead, we need to summarize the data by grouping them in numerical classes or intervals in what is called a grouped frequency distribution. Here is an example of weight of students in a classroom.

Weights of students in a classroom

Weights (in kg)	Number of students
30.5 - 35.5	9
35.5 - 40.5	6
40.5 - 45.5	15
45.5 - 50.5	3
50.5 - 55.5	1
55.5 - 60.5	2
Total	36

Grouped Frequency Table (GFT) for Continuous Data

A group of sports-wear stores recorded the profits (in $1000's) from 20 different products as shown in the table here:

24	35	17	21	24	37	26	46	58	30
32	13	12	41	43	44	27	38	53	27

Here are the steps to construct a grouped frequency distribution:

Step 1:

Sort the data in ascending (or descending) order.

Step 2:

$$\textit{Find the Range} = \textit{Largest observation} - \textit{Smallest observation}$$

$$R = 58 - 12 = 46$$

Step 3:

Divide the range into equal-width classes. A simple rule of thumb for the number of classes is $k = \sqrt{n}$. Another criterion to determine how many classes to use is $2^k \geq n$

, where k = the number of classes and n = the sample size.

$$\Delta = \textit{Class width} \approx \frac{\textit{Range}}{\textit{Number of classes}} = \frac{R}{k}$$

Since we have a sample of size 20, the number of classes should be approximately $\sqrt{20} = 4.47 \approx 5$
In this case, the class width would be approximately $46/5 = 9.2 \approx 10$
(*always rounded up to the nearest significant digit*).
The smallest observation is 12.
The first class boundary can start at 12 or to the nearest multiple of 5, say at 10 (to guarantee that the smallest observation does not fall on the lower class boundary).
Therefore, the first class would be (10, 20].
The second class would be (20, 30], and so on.

Step 4:

For each class, compute the midpoint (*class mark*), denoted by (x_i)

$$\text{Midpoint } (x_i) = \frac{\textit{Upper boundary} + \textit{Lower boundary}}{2}$$

Step 5:

For each class, count the number of observations that fall in that class. That is to find the class frequency (*f*).

Step 6:

For each class, find the relative frequency (*rf*), the cumulative relative frequency (*crf*).

1. The classes **MUST** be *mutually exclusive* i.e., each value must fall in *only* one class.
2. The classes **MUST** be *exhaustive* i.e., contain all possible data values.

As a result, for the given profits, we have the following frequency distribution:

Class	**Tally**	***F***	**Midpoint** x_i	***cf***	***r f***	***rcf***
[10 , 20)	///	3	15	3	0.15	0.15
[20 , 30)	///// /	6	25	9	0.30	0.45
[30 , 40)	/////	5	35	14	0.25	0.70
[40 , 50)	////	4	45	18	0.20	0.90
[50 , 60)	//	2	55	20	0.10	1.00
Total		20			1.00	

Recap

In this lecture, you have learned that:

- A *frequency distribution* is a list or table used to summarize a large set of data
- A frequency distribution is a way to summarize and *condense* the raw data in a limited number of *classes*
- If the variable of interest is discrete with few possible outcomes or categorical then the frequency distribution is constructed by counting the number of times each possible value occurs in the data set
- If the variable of interest is continuous or discrete and has many possible outcomes, we need to summarize the data by grouping them in numerical classes or intervals in what is called a grouped frequency distribution